Ancrages et modèles dynamiques de la prosodie : application à la reconnaissance des émotions actées et spontanées

Fabien Ringeval

THESE DE DOCTORAT

Ancrages et modèles dynamiques de la prosodie : application à la reconnaissance des émotions actées et spontanées

présentée par

Fabien RINGEVAL

pour l'obtention du grade de

Docteur de l'Université Pierre et Marie Curie – Paris 6

Spécialité

Traitement du Signal

Travaux soutenus publiquement le 4 Avril 2011 devant le jury composé de

Rapporteurs	Hervé GLOTIN	Professeur	LSIS, UTLN – Toulon
	Yannis STYLIANOU	Professeur	MMI, UOC – Voutes
Examinateurs	Olivier ADAM	Professeur	LAM, UPMC – Paris 6
	Björn SCHULLER	Maître de Conférences	MMK, TUM – Munich
Invité	David COHEN	Professeur	ISIR, UPMC – Paris 6
Directeur	Jean-Luc ZARADER	Professeur	ISIR, UPMC – Paris 6
Encadrement	Mohamed CHETOUANI	Maître de Conférences	ISIR, UPMC – Paris 6

Résumé

La reconnaissance de l'état émotionnel d'un locuteur est une étape importante pour rendre la communication Homme-machine plus naturelle et conviviale. Nous étudions dans cette thèse la problématique du traitement automatique de la parole (TAP) orienté émotion sur des données actées et naturelles. L'étude des émotions spontanées a été effectuée en parallèle avec celles des troubles de la communication (TC), puisque ces troubles limitent les capacités d'interaction de l'enfant. Les techniques incluses dans les systèmes de TAP orienté émotion doivent reposer sur des paramètres robustes dans la description des corrélats de l'affect, mais aussi face aux contraintes liées au changement de locuteur et de contexte sémantique. Dans cet esprit, nos travaux ont exploité un ensemble de traitements automatiques pour effectuer la reconnaissance des émotions. Nous avons notamment identifié des points d'ancrage complémentaires de la parole (e.g., pseudo-phonèmes) pour extraire plusieurs types de paramètres (e.g., acoustique et prosodique) sur le signal. Des techniques de fusion ont aussi été employées pour estimer la contribution de ces approches dans la tâche de reconnaissance. De plus, un effort a été tout spécialement porté sur le développement de modèles *non-conventionnels* du rythme, puisque cette composante apparaît clairement comme étant sous modélisée dans les systèmes *état-de-l'art*. Les expériences effectuées dans cette thèse visent à démontrer la pertinence des points d'ancrage de la parole et des modèles du rythme pour identifier les paramètres corrélés aux émotions. L'étude des émotions prototypiques (i.e., actées) par les modèles *non-conventionnels* du rythme a, par exemple, permis de définir un continuum de valeurs représentant alors les classes d'émotions qui apparaissent selon la roue de Plutchik.

Les analyses portant sur les TC ont été effectuées en étroite collaboration avec des équipes de cliniciens et de chercheurs en TAP orienté émotion. Ces travaux ont eu pour but d'employer des méthodes automatiques (i.e., identification des points d'ancrage de la parole et extraction de paramètres prosodiques) pour caractériser les particularités associées aux types de TC étudiés, i.e., autisme, dysphasie et troubles envahissants du développement non-spécifiés (TED-NOS). Un groupe contrôle composé d'enfants à développement typique a aussi été étudié pour comparer les capacités prosodiques des sujets TC. Les résultats de cette étude sont prometteurs puisqu'ils ont montré que l'ensemble des sujets pathologiques pouvait être discriminé significativement des typiques, tout comme les différents groupes de TC, selon deux types d'épreuves distinctes : (i) *imitation de contours intonatifs* (tâche contrainte) et (ii) *production de parole affective spontanée* (tâche non-contrainte). De plus, les résultats fournis par une analyse automatique des données ont permis de retrouver les caractéristiques cliniques des groupes de TC.

Les techniques actuelles en TAP orienté émotion sont donc suffisamment matures pour s'affranchir des difficultés créées par l'étude de corpus contenant de la parole spontanée et/ou produite par des voix d'enfants. Par conséquent, la difficile mais au combien importante tâche « *d'humanisation* » des systèmes communicants peut être envisagée, puisque les machines peuvent avoir la capacité de percevoir de façon robuste l'affect dans des situations naturelles.

Abstract

Recognition of emotional state of a speaker is an important step in making the human-machine communication more natural and friendly. We study in this thesis the problem of emotion-oriented automatic speech processing (ASP) on both acted and natural data. The study of spontaneous emotions is conducted along with the ones having communication disorders which limit the development of the interaction's capabilities of a child. Techniques derived from emotion-oriented ASP must be based on robust parameters to describe the emotional correlates, and also face the constraints that are related to the change of speaker and semantic context. In this view, our work is based on the use of automated techniques to perform emotion recognition: we use many complementary anchors of speech (e.g., pseudo-phonemes) to extract different types of parameters from the signal (e.g., acoustic and prosodic), and also combine techniques to estimate their contributions in the recognition task. An effort has been done to focus on the development of new *unconventional* models of speech rhythm, since this component is not modeled clearly in the state-of-the-art emotion recognition systems. The experiments conducted in this thesis aim to demonstrate the relevance of using several anchor points of speech and their associated rhythmic patterns to identify the features that are correlated with emotions. The study of prototypical emotions has permitted to define a continuum which represents the emotional categories along with the emotional wheel of Plutchik.

The analysis of communication disorders are carried out in close collaboration with clinicians and researchers teams in emotion-oriented ASP. This work aims to use automated methods (i.e., identification of speech anchor points and extraction of prosodic features) to characterize the features that are associated to a given language impairment (LI), e.g., autism, dysphasia and pervasive developmental disorders non-otherwise specified (PDD-NOS). A control group of typically developing children is also used to compare the prosodic abilities of the LI subjects. The results we obtained in this study are very promising because they contributed significantly to discriminate all of the LI subjects from the typically developing children, and also discriminate the different groups of LI in two distinct type of events: (i) *imitation of intonation contours* (constrained task) and (ii) *production of spontaneous emotional speech* (unconstrained task). In addition, the results provided by an automatic analysis of these data also allowed retrieving the diagnostic criteria defined by clinicians on the different groups of LI children.

Current techniques in ASP can thus overcome the difficulties created by the study of spontaneous speech data produced by children voices. This opens the way for the difficult but so interesting task of how to make friendly and less "cold" communication systems that are currently available to us.

Table des matières

Chapitre 2 : Ancrages acoustiques de la parole

Chapitre 3 : Reconnaissance acoustique de la parole affective actée

Chapitre 4 : Reconnaissance prosodique de la parole affective actée

Chapitre 5 : Emotions et troubles de la communication

Table des figures

Liste des tableaux

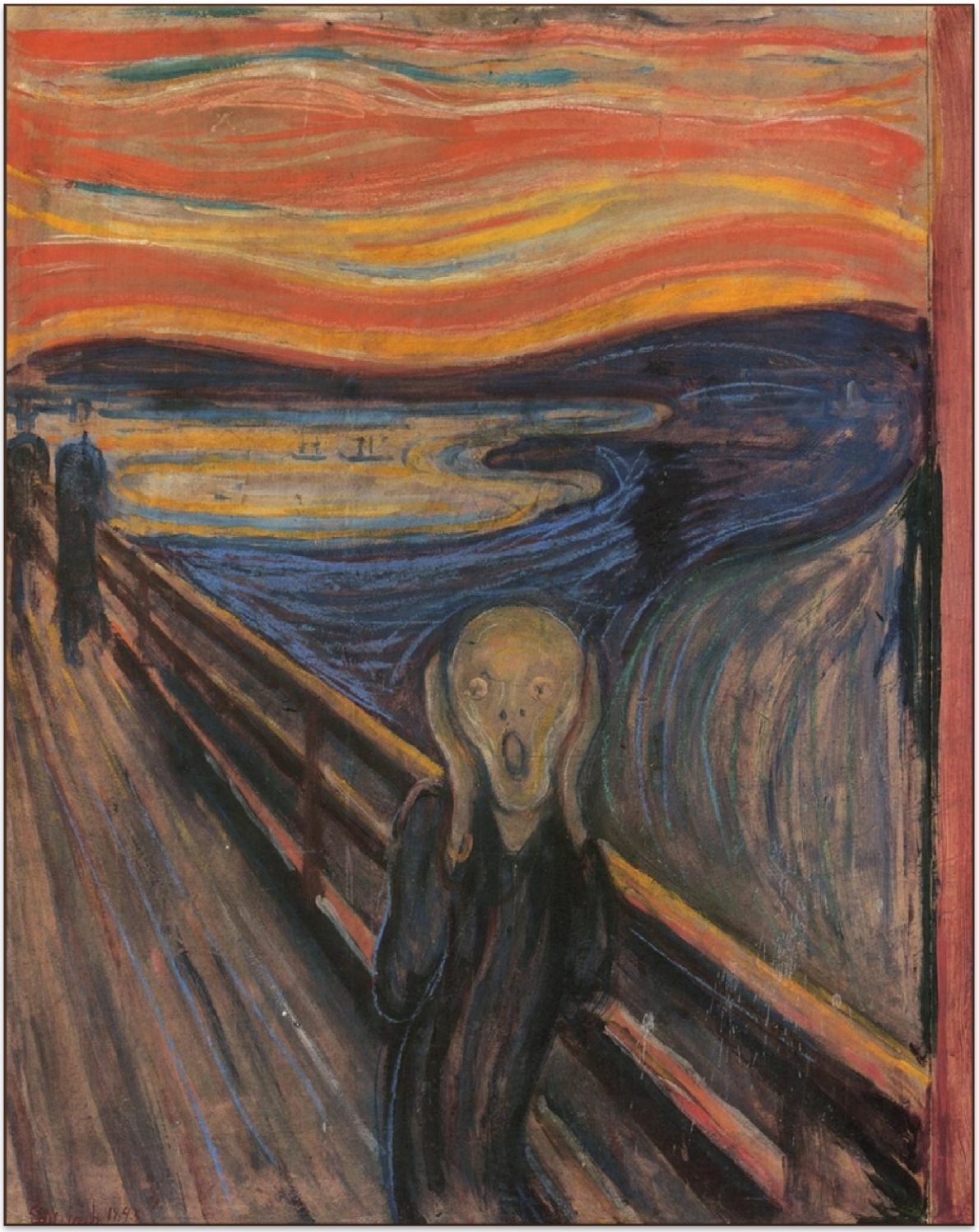
Chapitre 1

Introduction

Conscient de son intérêt pour la communication, l'Homme a appris très tôt à maîtriser ses facultés de langage. De nombreux systèmes de codification sont alors apparus de par les siècles. Des communautés variées de chercheurs se sont ensuite penchées sur l'étude des informations portées par le langage afin de comprendre leur origine et leur nature : (i) les philosophes et sociologues ont défini les types d'informations présentes dans la communication et leurs incidences sur les interactions sociales, (ii) les linguistes ont analysé les règles d'association des codes aux informations ; et les taxinomistes, les spécificités selon les langues, (iii) les neurosciences ont identifié les processus cognitifs par lesquels le traitement des informations s'opère, (iv) les biologistes et physiciens ont caractérisé le fonctionnement des organes dits de la parole et (v) les sciences computationnelles ont proposé des méthodes et des techniques permettant d'analyser de façon automatique les phénomènes observables dans le langage.

Le but affiché par ces dernières est de définir un ensemble de traitements permettant de constituer un dictionnaire des codes expliquant alors les associations entre les phénomènes observables et les informations. Cependant, compte tenu de l'extrême variabilité de ces données entre les individus (e.g., style personnel) et les groupes d'individus (e.g., facteurs sociaux-culturels), de nombreux verrous technologiques s'opposent à la faisabilité du traitement automatique de la parole (TAP). La limitation technologique est telle que l'emploi massif de traitements robustes pour l'analyse des signaux affectifs et sociaux ne s'est véritablement répandu qu'au cours des dix dernières années ; avec notoirement les travaux de Picard, Narayanan, et Schuller *et al.* pour les émotions ; et Pentland, Vinciarelli *et al.* pour les signaux sociaux.

CHAPITRE 1. INTRODUCTION



« Le cri » (« *Skrik* », 1893), tableau expressionniste de l'artiste Norvégien Edvard Munch symbolisant un Homme moderne emporté par une crise d'angoisse. La technique utilisée par ce peintre permet de véhiculer de façon remarquable un mélange complexe d'émotions représentant alors les pensées de l'individu du tableau. Cette œuvre picturale très célèbre a donné lieu à de nombreux détournements desservant alors des causes diverses et souvent revendicatrices d'un mal-être sociétal.

1. Motivations et contexte

Le 20ème siècle fut le témoin de l'apparition des technologies de la communication. Les possibilités d'interaction offertes à l'Homme ont ainsi été étoffées par la mise à disposition de moyens de communication de plus en plus sophistiqués : télégramme, télex, fax, téléphone, minitel, Internet, téléphonie mobile, messageries instantanées, réseaux sociaux, robotique et agents conversationnels animés. Les informations qui sont échangées par l'Homme à travers ces multiples canaux de communication incluent des caractéristiques propres aux interactions sociales. Puisque ces informations permettent de caractériser pleinement les individus lors de la communication, les sciences computationnelles se sont efforcées d'extraire les codes du langage (i.e., expliquant les informations) pour les intégrer dans les systèmes communicants de dernières générations, cf. Fig. 1.2. Ces systèmes doivent être capables de décoder les signaux sociaux de l'Homme de façon à pouvoir y répondre en retour par des comportements adaptés, rendant ainsi les interactions Homme-machine plus naturelles et conviviales.

Néanmoins, cette tâche « *d'humanisation* » des systèmes communicants est loin d'être évidente, puisque de nombreuses contraintes s'opposent à la caractérisation des informations de la communication. En effet, bien qu'un ensemble de codes fixant les règles d'encodage et de décodage ait été préétabli par l'Homme pour communiquer, la complexité et la variabilité de ces codes selon les individus est immense. De plus, les définitions apportées par les psychologues sur les associations codes – informations sont parfois ambigües.

1.1. Communication verbale et non-verbale

La base de la communication repose sur un processus de transfert d'informations entre deux entités vivantes. Elle fait ainsi intervenir une source (i.e., un émetteur) et un récepteur (i.e., un destinataire) à travers un canal de communication donné. Shannon a proposé un schéma illustrant certains paramètres de la communication, cf. Fig. 1.1 [SCH48][1]. Néanmoins, ce schéma souffre de certains manques puisqu'il ne prend pas en compte, par exemple, les possibilités d'interactions entre plusieurs individus ou groupes d'individus, et suppose le processus de communication linéaire alors que la rétroaction régule les interactions [ABR08][2]. De nombreux auteurs ont donc proposé d'inclure d'autres types d'informations pour décrire le processus de communication tels que le contexte [JAK60][3], la dimension perceptive, le contrôle et la communication de masse [GER56][4] ou encore le rôle du message dans l'interaction sociale [NEW53][5].

[1] C. Shannon et W. Warren, "A mathematical theory of communication", dans *The Bell Systems Techn. J.*, (reprinted with corrections), vol. 27, pp.379–423, (pp. 623–656), Jul. (Oct.) 1948.

[2] J. C. Abric, *Psychologie de la communication : théories et méthodes*, dans Armand Colin, 3ème Ed., Fev. 2008.

[3] R. Jakobson, *Closing statement: Linguistics and poetics*, dans T. Sebeok [Eds], *Style in Language*, 1960.

[4] G. Gerbner, "Toward a general model of communication", dans *Audio-Visual Comm. Review*, vol. 4, pp. 171–199, 1956.

[5] T. M. Newcomb, "An approach to the study of communicative acts", dans *Psychological Review*, vol. 60, pp. 393–404, 1953.

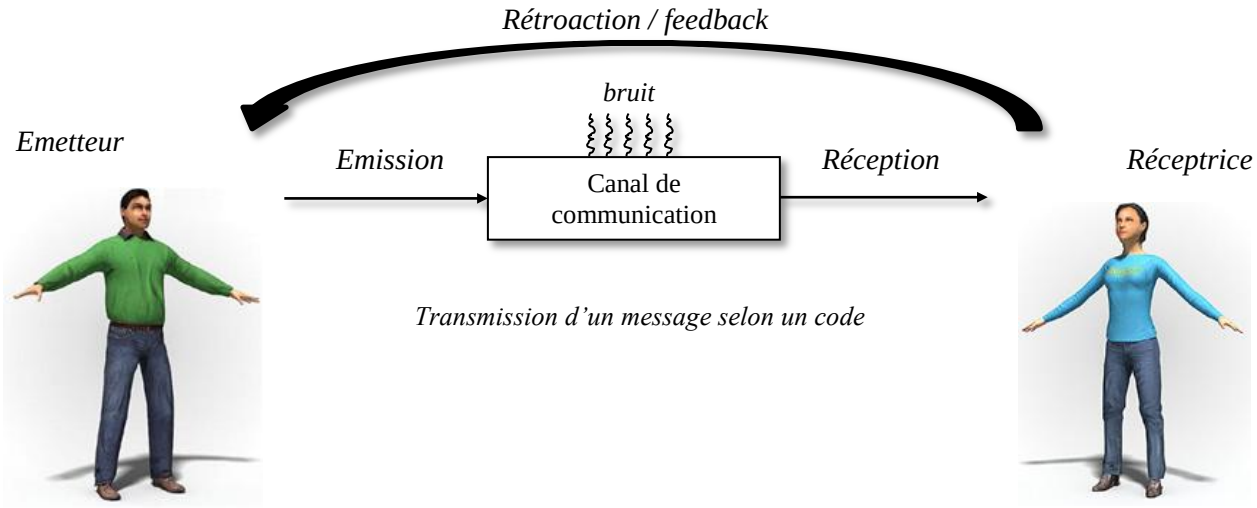


Fig. 1.1 Processus de communication selon Shannon [SCH48][1].

Comme les informations liées à la communication sont très nombreuses, il a été proposé de les distinguer en deux grandes catégories : (i) la communication *verbale* qui considère les informations propres au message et (ii) la communication *non-verbale* qui repose sur des signaux plus complexes et desservant alors les interactions sociales. Sur le plan ethnologique, la communication *verbale* correspond à un ensemble de sons émis dans le but d'établir un lien avec autrui. Au niveau de la linguistique, elle fait intervenir les différents niveaux du discours à travers le choix des mots et la structure des phrases. La prosodie est alors exploitée pour mettre en avant certaines parties du message (e.g., fonctionnalités grammaticales) ou introduire la pragmatique (e.g., question / déclaration). Alors que la communication *verbale* renvoie aux informations qui sont directement (e.g., mots, phrases) ou indirectement (e.g., prosodie – langage *para-verbal*) liées au message produit, la communication *non-verbale* intègre les informations extérieures à ce dernier et qui sont propres aux interactions sociales : attitudes, intentions ou émotions selon les modes d'expressions orales, faciales, gestuelles et posturales. La nature des informations *verbales* implique leur dépendance à la langue, ce qui n'est pas le cas des catégories principales d'informations *non-verbales* [ELF02][6]. Notons par ailleurs que ces signaux sont très dominants dans la communication orale [MEH67][7].

1.2. Traitement automatique de la parole

Le traitement automatique de la parole (TAP) a été abordé depuis l'existence des sciences de l'informatique et selon plusieurs niveaux d'analyses [FUR09][8] : (i) la perception et la reconnaissance de la parole, (ii) la reconnaissance du locuteur et (iii) les systèmes de questions / réponses. Les avancés dans le domaine du TAP ont permis de produire quantité de méthodes

[6] H. A. Elfenbein et N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis", dans *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, Mar. 2002.

[7] A. Mehrabian et S. R. Ferris, "Inference of attitude from nonverbal communication in two channels", dans *J. of Counseling Psycho.*, vol. 31, no. 3, pp. 248–252, Jun. 1967.

[8] S. Furui, "Selected topics from 40 years of research on speech and speaker recognition", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 1–8.

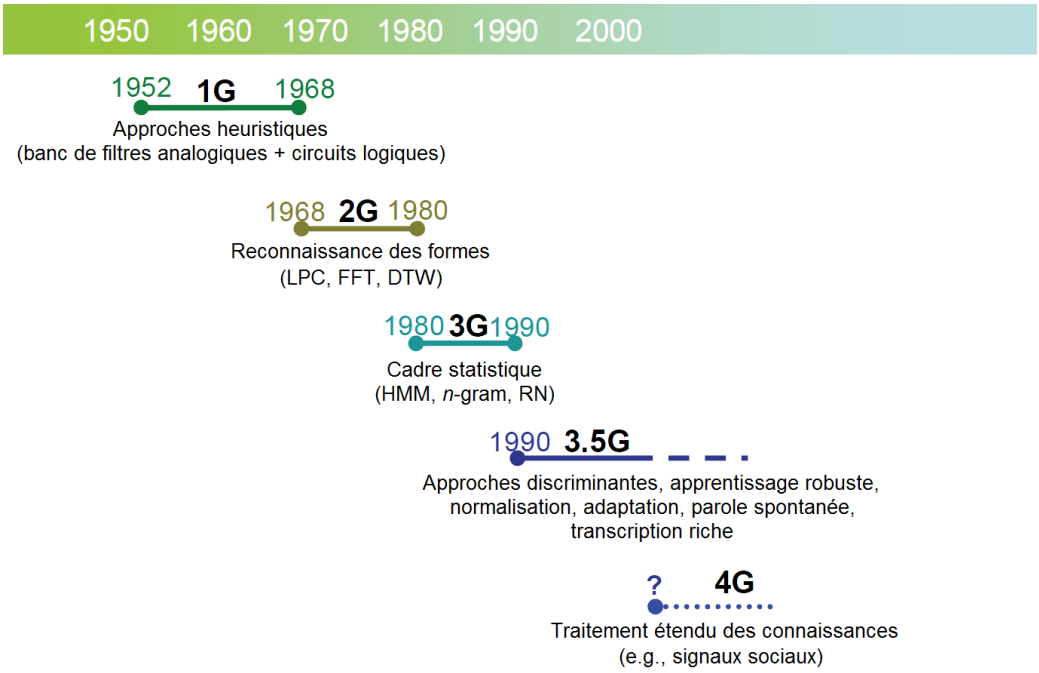


Fig. 1.2 Quatre générations de recherche en reconnaissance de la parole et du locuteur ; figure reproduite de [FUR05][9].

de modélisations célèbres telles que : (i) la prédiction linéaire [RAB78][10], (ii) la synthèse par chevauchement et ajout [ALL77][11], (iii) les coefficients de perception PLP [HER90][12] ou cepstraux selon l'échelle mel [DAV80][13] et (iv) les codecs audio tels que le codeur CELP [SCH-85b][14] utilisé en téléphonie mobile ou encore le populaire MP3 pour la parole et la musique. Le développement des techniques en extraction de caractéristiques a été parallèlement accompagné par des avancés dans le paradigme de la reconnaissance des formes. Les méthodes les plus connues sont alors les modèles de Markov cachés (HMMs), les mélanges de modèles Gaussiens (MMG), ou encore les machines à vecteur support (SVM) [DUD00][15]. Le TAP est donc une discipline de recherche très active et pour laquelle une large variété de progrès techniques significatifs a été accomplie durant les décennies précédentes. La Fig. e1.2 illustre les générations des systèmes de reconnaissance de la parole et du locuteur qui se sont succédées depuis 1952 [FUR05][9].

---

[9] S. Furui, "50 years of progress in speech and speaker recognition", dans proc. *SPECOM*, Patras, Greece, Oct. 17-19 2005, pp. 1–9.

[10] L. R. Rabiner et R. W. Schafer, *Digital Processing of Speech Signals*, dans Prentice Hall, Upper Saddle River, NJ, 1978.

[11] J. B. Allen et L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", dans *proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[12] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech", dans *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[13] S. B. Davis et P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", dans *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[14] M. R. Schroeder et B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates", dans proc. *ICASSP*, Tampa (FL), Mar. 26-29 1985, pp. 937–940.

[15] R. O. Duda, P. E. Hart et D. G. Stork, *Pattern classification*, 2nd Ed., New York: Wiley, 2000.

1.3. TAP orienté émotion

Les émotions ont un impact majeur sur de nombreux processus cognitifs. En effet, la littérature regorge d'études montrant que les émotions jouent un rôle essentiel non seulement dans l'intelligence et la créativité humaine, mais aussi dans la pensée rationnelle et la prise de décision [PIC97][16]. Les émotions ne sont donc pas en soi un luxe mais constituent un bagage prépondérant à la communication humaine. Par conséquent, les machines qui interagiront naturellement et intelligemment avec les humains auront besoin de la capacité de percevoir et d'exprimer, au minimum, de l'affect. Dans ce cens, le TAP orienté émotion (ou *Affective Computing*) est un domaine de recherche assez récent et qui porte sur la reconnaissance et la synthèse des émotions dans la parole, les expressions faciales, ou tout autre canal de communication biologique [PIC97]. L'objectif principal consiste à identifier les corrélats de l'affect présents dans ces canaux de communication. Concernant la reconnaissance des émotions, une des difficultés majeures réside à la fois dans la détermination des caractéristiques pertinentes et des classifieurs [VER06][17]. D'autres difficultés apparaissent dans la définition même des émotions et de leur annotation (cf. section 3) [DEV05][18]. Les méthodologies proposées par l'état-de-l'art en TAP orienté émotion sont présentées dans la section 4 de ce chapitre.

1.4. Traitement des signaux sociaux

Le traitement des signaux sociaux (SSP) est un domaine de recherche émergent qui consiste à analyser, interpréter et prédire les interactions sociales humaines ; [PEN07][19] et [VIN-09][20]. En son cœur, l'intelligence sociale vise à un usage adapté et une interprétation précise des signaux sociaux. Ces derniers sont continus et peuvent être associés à différents catégories de processus : comportementale, vocalique, chimique ou morphologique, [VIN09] et [PEN-07]. Ces processus sont utilisés pour transmettre des traits ou des états qui ne sont pas forcément perceptibles de l'extérieur [PAN08][21] ; ceux-ci pouvant être aussi divers que la condition de reproduction, les traits de personnalité, les émotions ou les attitudes à l'égard des objets ou des personnes. Les signaux sociaux permettent de communiquer toutes sortes d'informations dans les interactions quotidiennes, e.g., intérêt, empathie, hostilité, accord / désaccord, flirt, domination, supériorité, infériorité, etc. Une de leurs particularités réside dans le fait qu'ils

[16] R. W. Picard, *Affective Computing*, dans Perceptual computing section, technical report no. 321, Cambridge, Massachusetts, M.I.T. Media Lab., 1997.

[17] R. Ververidis et C. Kotropoulos, "Emotional speech recognition, features and method", dans *Speech Comm.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[18] L. Devillers, L. Vidrascu et L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection", dans *J. of Neural Net.*, vol. 18, no. 4, pp. 407–422, May 2005.

[19] A. Pentland, "Social signal processing", dans *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[20] A. Vinciarelli, M. Pantic et H. Bourlard, "Social signal processing: Survey of an emerging domain", dans *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[21] M. Pantic, A. Pentland, A. Nijholt et T. Huang, "Human-centred intelligent human-computer interaction (hci2): how far are we from attaining it ?", dans *Inter. J. of Autonomous and Adaptive Comm. Systems*, vol. 1, no. 2, pp. 168–187, Aug. 2008.

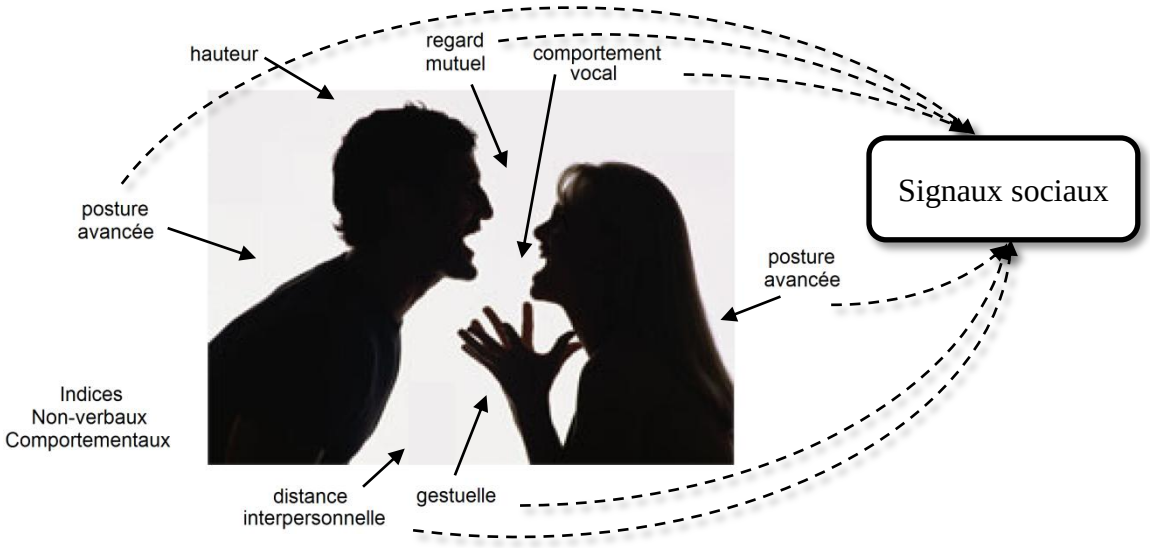


Fig. 1.3 Indices comportementaux et signaux sociaux ; figure reproduite de [VIN09][20].

peuvent prendre la forme de constellations complexes d'indices *non-verbaux* du comportement (e.g., expressions faciales, de la prosodie, des gestes, de la posture, etc.) accompagnant les interactions Homme-Homme ou Homme-machine, cf. Fig. 1.3.

L'intérêt de l'analyse des signaux sociaux est immense puisqu'elle permet de contribuer à la fois à la compréhension des informations véhiculées par les codes sociaux, mais également à la tâche « *d'humanisation* » des systèmes communicants. En effet, une interprétation précise de ces signaux par les machines leur permettrait de prendre en compte les paramètres sociaux présents dans les interactions Homme-machine. Toutefois, les signaux *non-verbaux* présentent une forte dynamique et une interdépendance qui complexifient la tâche de caractérisation des informations qu'ils véhiculent [ARG67][22]. Des études ont cependant montré qu'il est déjà possible d'identifier la personne dominante dans une conversation [ARA10][23] et [CAM09][24], ou encore de gérer les rétroactions dans les interactions Homme-machine [ALM09][25], cf. Fig. 1.4.

1.5. Enjeux théoriques et applicatifs

Les enjeux théoriques de cette thèse concernent à la fois le domaine du TAP orienté émotion et le SSP ; puisque ces dernières font parties des interactions sociales. Notre étude a eu pour objectif d'identifier les différents paramètres intervenant dans la communication orale des émotions. Nos travaux ont tout d'abord consisté à définir des méthodes permettant d'identifier automatiquement les supports temporels sur lesquels les informations sont ancrées dans

[22] M. Argyle, *The Psychology of Interpersonal Behaviour*, dans Penguin, 1967.

[23] O. Aran et D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversation", dans proc. *ICPR*, Istanbul, Turkey, Aug. 23-26 2010, pp. 3687–3690.

[24] N. Campbell, "An audio-visual approach to measuring discourse synchrony in multimodal conversation data", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 2159–2162.

[25] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J. C. Martin, S. Ondas, C. Pelachaud, J. Urbain et M. Yilmaz, "Generating robot/agent backchannels during a storytelling experiment", dans proc. *IEEE Inter. C. on Rob. and Automation*, Kobe, Japan, May 12-17 2009, pp. 2477–2482.

Fig. 1.4 Génération de signaux de rétroaction par : (i) un agent conversationnel animé (GRETA [ROS03][26]) et (ii) un jouet robotisé (AIBO Sony) pour l'accompagnement des interactions verbales.

la parole. Ainsi, nous qualifions de point d'ancrage ces supports puisqu'ils sont censés contenir les informations recherchées, e.g., l'affect. Nos travaux se sont ensuite penchés sur le développement de nouveaux paramètres du rythme puisque cette composante apparaît comme étant clairement sous-modélisée dans les systèmes *état-de-l'art*. Par ailleurs, la littérature fait apparaître une méconnaissance de l'importance des contributions apportées par les composantes prosodiques corrélées à l'affect, tout comme les points d'ancrage de ces informations dans la parole. Nous avons donc utilisé des techniques de fusion pour estimer ces données.

Les enjeux applicatifs de cette thèse concernent l'étude de paroles émotionnelles actées et naturelles. Les premières expériences portent tout d'abord sur les systèmes de détection automatique des points d'ancrage. Nous avons notamment estimé les scores en détection de pseudo-phonèmes et les corrélats phonétiques associés au centre de perception de la parole « *p-centre* » sur plusieurs types de corpus : (i) parole lue en qualité laboratoire (TIMIT) et téléphonique (NTIMIT) et (ii) parole affective actée selon plusieurs langues ; Allemand – corpus Berlin, Bas-que – corpus Aholab, et Hongrois – corpus Bute-TMI. Le corpus Berlin a ensuite servi à estimer la contribution apportée par plusieurs points d'ancrage de la parole et différents types de paramètres prosodiques pour la reconnaissance automatique d'émotion. Ces scores ont été comparés à ceux de l'état-de-l'art puisque ce corpus a été très largement exploité dans la littérature. Nous avons ensuite étudié le comportement des métriques *conventionnelles* et *non-conventionnelles* du rythme selon les classes d'émotion (analyse statistique).

Concernant l'étude des émotions spontanées, nous avons collaboré avec des cliniciens pour créer un corpus de parole permettant d'étudier, en parallèle des émotions, l'impact des troubles de la communication (TC) chez plusieurs groupes d'enfants. L'analyse de ce corpus a été effectuée avec les traitements automatiques qui ont été proposés pour l'étude des émotions actées. Nous avons ainsi exploité les systèmes de détection automatique des points d'ancrage de la parole et les étapes d'extraction de caractéristiques prosodiques, en particulier les modèles du rythme. Notons que l'identification des stratégies mises en œuvre par les sujets atteints de TC pour communiquer constitue un enjeu déterminant, puisque ces enfants sont bien souvent repliés sur eux-mêmes en raison de leurs difficultés à traiter (et exploiter) de façon convenable les codes sociaux qui sont malgré tout omniprésents.

[26] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio et B. De Carolis, "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent", dans *Inter. J. of Human-Computer Studies, Application of affective computing in Human–computer interaction*, vol. 59, no. 1-2, pp. 81–118, Jul. 2003.

Nous décrivons dans la section suivante les définitions et les fonctionnalités de la prosodie qui ont été identifiés par des courants de recherche variés. Les deux autres sections portent sur les émotions et présentent dans un premier temps, les définitions apportées par les psychologues sur l'affect et ses corrélats acoustiques, et dans un second temps, les méthodes de sollicitation permettant de constituer un corpus. Nous présentons ensuite un bref état-de-l'art dans le domaine du traitement automatique de la parole affective.

2. La prosodie, support des informations de la communication

Le terme « *prosodie* » est employé dans le vocabulaire de nombreuses communautés scientifiques et nécessite par conséquent d'être correctement défini pour en clarifier l'usage qui est effectué dans cette thèse.

2.1. Ensemble de définitions

Les définitions de la prosodie ont évolué au fil du temps. Toutefois, elles furent d'emblée liées à la musique, au chant et à la poésie chantée puisque le mot grec *προσωδία* (dont a hérité le latin avec le mot *prosodia*) signifie, selon le *Trésor de la langue française*, le « chant pour accompagner la lyre » et les « variations dans le niveau de la voix ». Les aspects musicaux de la prosodie sont assez contemporains puisque son usage ne s'est véritablement répandu qu'au milieu du 19ème siècle avec l'usage du verbe « prosodier ». La prosodie a été définie par les poètes de ce siècle comme la « manière de prononcer régulièrement dans les mots chaque syllabe prise à part et considérée en elle-même » [BAN72][27]. La notion de métrique a ensuite été renforcée par les linguistes puisqu'ils ont considéré la prosodie comme « l'ensemble des règles de versification qui concernent la quantité de voyelles, les faits accentuels et mélodiques » [MOU74][28]. La prosodie renvoie ainsi à « l'étude de phénomènes variés étrangers à la double articulation mais inséparables du discours comme la mélodie, l'intensité, la durée, etc. ». De nos jours, les définitions ne se limitent plus à la modalité parole et concernent plutôt l'ensemble des éléments multimodaux accompagnant le message. Il est ainsi courant de parler de prosodie faciale [GRA02b][29] ou encore gestuelle [LIM10][30] pour désigner les suppléments d'informations apportés au discours.

La fonction principale de la prosodie est avant tout une fonction d'assistance à l'encodage et au décodage des informations de la communication. Elle permet à un individu de structurer son discours et de véhiculer à travers le message produit des informations de niveau supérieur. Cette mise en valeur du message s'effectue selon un ensemble de règles annexes à la langue du locuteur, i.e., celles dont le non-respect ne modifie pas fortement la compréhension du discours [DIC00][31]. Elle fait intervenir, au niveau de la parole, des composantes musicales par

[27] T. de Banville, *Petit traité de poésie*, dans Chez l'écho de la Sorbonne, Paris, 1872.

[28] G. Mounin, *Dictionnaire de la Linguistique*, dans Paris: Presses Universitaires de France, 1974.

[29] H. P. Graf, E. Cosatto, V. Strom et F. J. Huang, "Visual prosody: Facial movements accompanying speech", dans proc. *5th AFGR*, Washington DC, USA, May 21 2002, pp. 381–386.

[30] F. Limousin et M. Blondel, "Prosodie et acquisition de la langue des signes française", dans *Lang., Interaction and Acquisition*, J. Benjamins Publishing Company, vol. 1, pp. 82–109, 2010.

la combinaison de variations (segmentales et suprasegmentales) de la fréquence fondamentale (f0), de l'énergie, de la qualité vocale et de la durée des sons produits [NOO99][32].

Les techniques servant à caractériser ces dimensions prosodiques sont présentées en détails dans le chapitre 4, section 2. Nous décrivons dans les paragraphes suivants les fonctionnalités de la prosodie dans la communication ainsi que les différentes étapes qui interviennent dans le processus d'encodage des informations dans la parole.

2.2. Fonctionnalités de la prosodie dans la communication

La prosodie intervient à tous les niveaux de la communication. Elle vise alors à construire le discours via le langage expressif à travers plusieurs niveaux de communication distincts : (i) *grammatical*, (ii) *pragmatique* et (iii) *affectif* [PAU05a][33].

2.2.1. Grammaticales

La prosodie *grammaticale* est utilisée pour introduire les informations syntaxiques dans les phrases [WAR96][34]. Par exemple, l'accent est exploité dans les langues dites à accent libre ou à tons pour signaler si un mot est utilisé comme un nom (***con**vict*) ou un verbe (*to con**vict***)[35]. Les contours du pitch signalent la fin des phrases et dénotent si elles correspondent, par exemple, à des questions (contour de pitch montant) ou à des affirmations (contour descendant). Les pauses permettent quant à elles, de structurer les énoncés et de réguler les tours de parole [SHR00][36]. Une pause pleine laisse ainsi le champ libre à l'interlocuteur pour prendre la parole alors qu'une pause verbalisée (*filled pause*) permet de conserver le tour de parole.

2.2.2. Pragmatiques

La prosodie *pragmatique* informe sur les intentions du locuteur, ou sur la hiérarchie des informations contenues dans l'énoncé [PAU05a]. Elle résulte de changements optionnels dans la façon de produire une phrase [LAN81][37] et contient donc les informations au-delà de celles véhiculées par la syntaxe de la phrase. L'accentuation est par exemple utilisée comme un moyen de mettre en valeur certains éléments de l'énoncé, e.g. : « **Je** vais terminer », opposition à quelqu'un d'autre ; « Je **vais** terminer », opposition à une action déjà accomplie ; et « Je vais **terminer** », opposition à une autre action ; exemples extraits de [QUA07b][38].

---

[31] A. Di Cristo, "Interpréter la prosodie", dans proc. *23èmes JEP*, Aussois, France, Jun. 19-23 2000, pp. 13–29.

[32] S. Nooteboom, "The prosody of speech: melody and rhythm", dans *The handbook of phonetic sciences*, W. J. Hardcastle and J. Laver [Eds], Hardcastle, Blackwell, Oxford, pp. 640-673, 1999.

[33] R. Paul, A. Augustyn, A. Klin et F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders" dans *J. of Autism and Develop. Disorders*, vol. 35, no. 2, pp. 205–220, Apr. 2005.

[34] P. Warren, "Parsing and prosody: An introduction", dans *Prosody and parcing*, East Sussex, UK: Psychology Press, pp. 1–16, 1996.

[35] Cette distinction entre nom et verbe par la prosodie est beaucoup moins valable sur le Français puisque l'unité préférentielle est alors la syllabe (e.g., « *Je viens de fermer la **porte*** » / « *Je **porte** une chemise* », ou encore avec les mots « *hache* », « *lance* », « *pique* », « *danse* », etc.).

[36] E. Shriberg, A. Stolcke, D. Hakkani-Tur et G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", dans *Speech Comm.*, vol. 32, no. 1-2, pp. 127–154, Sep. 2000.

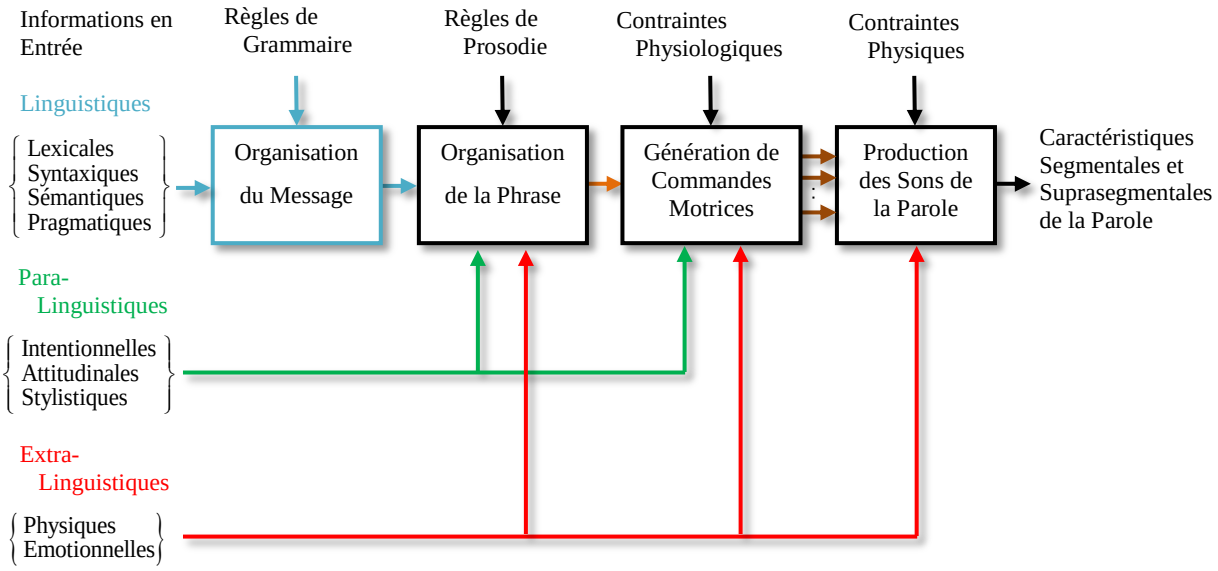


Fig. 1.5 Processus par lesquels des informations de types variés se manifestent dans les caractéristiques segmentales et suprasegmentales de la parole ; figure reproduite de [FUJ04][39].

2.2.3. Affective

La prosodie *affective* possède une fonction plus globale que celles desservies par les deux précédentes [PAU05a][33]. Elle exprime l'état général affectif d'un locuteur [WIN88][40] et comprend les changements de registre lorsque l'on parle à différents types d'interlocuteurs (e.g., nos pairs, de jeunes enfants ou des personnes de statut social plus élevé). Ses fonctionnalités sont donc : (i) extérieures au discours, (ii) concernent les intentions et les attitudes du locuteur face à ses semblables et (iii) ont pour objectif de desservir les interactions sociales.

2.3. Encodage des informations dans la parole

Les informations exprimées par la parole peuvent être décomposées en trois catégories : (i) *linguistiques*, (ii) *para-linguistique*, et (iii) *non-linguistique*. Bien que leurs frontières ne soient pas toujours très claires [FUJ04][41], cf. Fig. 1.5 ; (i) les informations *linguistiques* sont représentées par un ensemble fini et discret de symboles et de règles pour leurs combinaisons ; (ii) les informations *para-linguistiques* sont définies par celles qui ne peuvent être inférées par la partie écrite et qui sont délibérément ajoutées par le locuteur pour modifier ou supplémenter les informations *linguistiques*. Elles sont à la fois discrètes et continues, e.g., modalités discrètes de la phrase et continuum d'intentions ou d'attitudes du locuteur face au dis-

[37] D. van Lancker, D. Canter et D. Terbeek, "Disambiguation of ditropic sentences: Acoustic and phonetic cues" dans, *J. of Speech and Hearing Res.*vol. 24, no. 3, pp. 330–335, Sep. 1981.
[38] V. M. Quang, *Exploitation de la prosodie pour la segmentation et l'analyse automatique des signaux de parole*, Thèse de Doctorat, Institut National Polytechnique de Grenoble, 2007.
[39] H. Fujisaki, "Information, prosody, and modeling – with emphasis on tonal features of speech", dans *Speech Prosody*, Nara, Japan, Mar. 23-26 2004, invited paper.
[40] E. Winner, *The point of words: Children's understanding of metaphor and irony*, dans Cambridge, Harvard University Press, 1988.
[41] H. Fujisaki, "Information, prosody, and modeling – with emphasis on tonal features of speech", dans *Speech Prosody*, Nara, Japan, Mar. 23-26 2004, invited paper.

cours ; et (iii) les informations *non-linguistiques* concernent des facteurs tels que l'âge, le genre, les idiosyncrasies et les états physiques et affectifs du locuteur ; le contrôle sur ces composantes de la communication est alors généralement faible.

La relation entre ces trois types d'informations et la manifestation acoustique-phonétique de la prosodie, comme l'organisation des unités de la linguistique, est schématisée dans la Fig. 1.5. Ce schéma montre l'aspect complexe, multi-niveaux et multidimensionnel du processus d'encodage des informations dans la parole. Il explique également pourquoi il est très difficile de trouver une correspondance unique entre des caractéristiques physiques observables dans le signal de parole et les informations associées. L'identification des corrélats émotionnels de la voix, qui sont issus de déviances dans les règles de la grammaire et de la prosodie, constitue donc une tâche particulièrement difficile à mettre en œuvre.

3. L'Homme et ses émotions

Les études concernant la nature des émotions humaines sont très anciennes. Néanmoins, cette question reste un sujet continuel de débat et de recherche dans la psychologie moderne. De nombreux modèles ont été proposés par les psychologues pour les émotions. Les modèles discrets et assumant l'existence d'un faible nombre d'émotions basiques ont considérablement influencé les travaux de chercheurs issus de domaines de recherche variés, y compris celui des sciences computationnelles. Ces communautés se sont alors penchées sur l'étude de données contenant les « six grandes » émotions (*big six*). Afin d'apporter un cadre terminologique à notre étude, nous définissons et délimitons un ensemble de termes associés aux émotions dans la sous-section suivante. Nous présentons ensuite les théories modernes de l'affect et leurs racines historiques, ainsi que les propriétés d'encodage et de décodage acoustique des corrélats des émotions. Enfin, la dernière sous-section décrit les différentes possibilités qui existent pour solliciter des émotions chez un individu et constituer ainsi un corpus d'étude.

3.1. Terminologie

Une définition commune du terme *émotion* est nécessaire pour comparer les résultats et éviter des interprétations infondées. En effet, la façon par laquelle les émotions se définissent détermine le type de phénomène qui sera examiné par les chercheurs. Dans le cadre de cette thèse, et dans la continuité des travaux entrepris par la communauté du TAP orienté émotion [STE09][42], nous avons utilisé une définition du terme *émotion* qui a été proposée par Scherer [SCH00][43] : « des épisodes de changements coordonnés dans plusieurs composantes (incluant au moins l'activation neurophysiologique, l'expression motrice et les sensations subjectives, mais aussi probablement les actions tendancieuses et les processus cognitifs) en réponse à des évènements externes ou internes de signification majeure pour l'organisme ». Le déclenchement des évènements peut être produit par « le comportement des autres, un changement dans

[42] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, PhD thesis, University Friedrich-Alexander, Erlangen-Nuremberg, Germany, 2009.

[43] K. R. Scherer, "Psychological models of emotion", dans *The neuropsychology of emotion*, Oxford University Press, Oxford, New York, pp. 137–162, 2000.

une situation courante ou rencontrée avec un nouveau stimuli » [SCH00][43]. Bien qu'il existe de nombreux débats définitionnels, un consensus croissant peut être trouvé dans la littérature sur un ensemble de définitions du terme *émotion* (selon Scherer) :

(i) Les émotions sont de natures épisodiques et hautement distinctives [COW03][45]. L'idée sous-jacente est qu'un changement notable dans le fonctionnement de l'organisme est causé par le déclenchement d'évènements externes (e.g., le comportement des autres) ou internes (e.g., les pensées, mémoires, sensations). Les épisodes affectifs se manifestent pour une certaine durée sur laquelle l'état émotionnel est supposé s'amenuiser avec une intensité décroissante, et pour laquelle la disparition de l'état est plus difficile à détecter que la survenue.

(ii) Les émotions sont constituées de plusieurs composantes, incluant notamment la triade des émotions [SCH00] : (i) l'excitation physiologique (*arousal*), (ii) l'expression motrice et (iii) les sensations subjectives. D'autres composantes peuvent également être utilisées telles que les actions tendancieuses et les processus cognitifs impliqués dans l'évaluation des évènements suscités et la régulation des processus émotionnels en cours.

(iii) Les émotions sont d'une importance majeure pour l'organisme, puisque l'évaluation des évènements requis en respect de leur signification pour l'organisme détermine la réponse fonctionnelle de ce dernier (e.g., adaptation ou non à une situation donnée), et la nature des changements mentaux qui apparaissent durant l'épisode émotionnel.

(iv) Les épisodes émotionnels sont à caractère unique et requièrent une synchronisation et une interdépendance dans les changements effectués lors du traitement de leurs composantes.

Il existe divers termes décrivant le caractère épisodique et distinctif des émotions, chacun d'entre eux renvoyant alors à des implications théoriques spécifiques : *émotions primaires* [PLU84][46], *émotions basiques* [STE92][47], *émotions modales* [SCH94][48] ou *émotions aigües* [LAZ94][49]. Afin d'éviter des préjugés théoriques, Cowie *et al.* [COW01][50] utilisent le terme de Scherer *d'émotions pleines* (*full-blown*) [SCH99][51] comme un moyen neutre de référer aux épisodes qui seraient largement vus comme des exemples primaires de l'émotion. Avec cette définition, Scherer mentionne d'autres types de manifestations de l'affect que les émotions, cf. table 1.1. Ces phénomènes affectifs peuvent contrastés entre eux au moyen de caractéristiques conceptuelles, cf. table 1.2.

---

[44] Traduit de l'anglais : « episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism ».

[45] R. Cowie et R. R. Cornelius, "Describing the emotional states that are expressed in speech", dans *Speech Comm.*, vol. 40, no. 1-2, pp. 5–32, 2003.

[46] R. Plutchik, "Emotions: A general psychoevolutionary theory", dans K. R. Scherer and P. Ekman [Eds], *Approaches to Emotion*, Erlbaum, Hillsdale (NJ), 1984.

[47] N. Stein et K. Oatley, "Basic emotions: Theory and measurement", dans *Cognition and Emotion*, vol. 6, pp. 161–168, 1992.

Table 1.1 Brèves définitions de cinq état affectifs avec des exemples d'après [SCH00][43].

Etat affectif	Brève description	Exemples
Emotion	Episode relativement bref d'une réponse synchronisée de tous ou la plupart des sous-systèmes organiques en réponse à l'évaluation d'un évènement interne ou externe perçue comme étant de signification majeure.	Fâché, triste, joyeux, craintif, honteux, fier, ravi, désespéré
Humeur	Etat affectif diffus, plus prononcé que le changement dans le sentiment subjectif, de faible intensité mais de durée relativement longue, souvent sans cause apparente.	Gai, enjoué, irritable, distrait, déprimé, mélancolique
Positions interpersonnelles	Position affective prise envers une autre personne dans une interaction spécifique, colorant l'échange interpersonnel dans la situation.	Distant, froid, chaud, de soutien, méprisant
Attitudes	Relativement durables, croyances colorées affectivement, préférences et prédispositions envers les objets et les personnes.	Aimer, adorer, haïr, considérer, désirer
Traits personnels	Chargé émotionnellement, dispositions personnelles stables et comportement tendancieux, typiques pour une personne.	Nerveux, anxieux, morose, hostile, envieux, jaloux

Table 1.2 Délimitation des caractéristiques de différents états affectifs ; table extraite de [SCH03][52].

Etat affectif	intensité	durée	synchronisation	importance	rapidité de changement	impact comportemental
Emotion	++ – +++	+	+++	+++	+++	+++
Humeur	+ – ++	++	+	+	++	+
Positions interpersonnelles	+ – ++	+ – ++	+	++	+++	++
Attitudes	○ – ++	++ – +++	○	○	○ – +	+
Traits personnels	○ – +	+++	○	○	○	+

○ : faible ; + : moyen ; ++ : élevé ; +++ : très élevé ; – : indique une amplitude.

3.2. Théories des émotions

Les théories des émotions prédisent le nombre d'émotions différentes que l'on peut rencontrer, comment ces émotions sont différenciées, pour quelles raisons, et dans quelles situations elles sont produites. L'impact de ces théories sur les autres domaines de recherche comme les sciences computationnelles est donc extraordinairement élevé : (i) elles influencent le contenu des données issues des expériences, (ii) elles contrôlent la façon par laquelle ces données sont catégorisées en émotions et (iii) elles définissent, dans la synthèse des émotions, les instants auxquels un agent animé doit réagir par des signaux affectifs, ainsi que le choix de l'émotion la plus appropriée comme la façon par laquelle cette dernière doit être synthétisée.

[48] K. R. Scherer, "Towards a concept of modal emotions", dans P. Ekman and R. Davidson [Eds], *The Nature of Emotion: Fundamental Questions*, pp. 25–31, Oxford University Press, 1994.

[49] R. S. Lazarus, "The stable and the unstable in emotion. Fundamental questions", dans P. Ekman and R. Davidson [Eds], *The Nature of Emotion: Fundamental Questions*, pp. 79–85, Oxford University Press, 1994.

[50] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz et J. Taylor, "Emotion recognition in Human-computer interaction", dans *IEEE Signal Proc. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[51] K. R. Scherer, "Appraisal Theory", dans T. Dalgleish and M. Power, [Eds], *Handbook of Cognition and Emotion*, pp. 637–663, John Wiley, New York, 1999.

[52] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms", dans *Speech Comm.*", vol. 40, pp. 227–256, 2003.

La complexité des phénomènes produits par l'affect entraîne une certaine forme de concurrence dans les théories issues de la psychologie moderne des émotions. Compte tenu de la quantité de théories existantes, nous nous contenterons d'en résumer un échantillon [STE-09][42] ; [COR00][53] et [SCH00][43].

Platon (*théorie cognition-émotion*)

Plus de deux-mille ans auparavant, le philosophe Grec Platon (426-347 avant J.C.) suggéra que l'esprit avait une structure tripartie composée de trois domaines opposés : (i) la cognition, (ii) l'émotion et (iii) la motivation. Ce postulat a été sujet à une controverse quasi-constante dans la psychologie de l'émotion. 50 ans après, Aristote argumenta notamment en faveur de l'impossibilité d'une telle séparation puisqu'il existerait des interactions entre les niveaux du fonctionnement psychologique. Les débats ont récemment été relancés sous le nom de débat « *cognition-émotion* ».

Descartes (*théorie corps-esprit*)

L'écrivain, mathématicien et philosophe Français Descartes (1596-1650) a proposé de traiter simultanément les processus mentaux et physiologiques dans l'étude des émotions. Depuis ce temps, la relation entre les phénomènes mentaux et physiques est discutée de façon continuelle dans le débat « *corps-esprit* ».

Darwin (*théorie de l'évolution*)

Dans son livre intitulé The expression of the Emotions in Man and Animals [DAR72][54], Charles Darwin (1809-1882), naturaliste Anglais, a décrit les expressions faciales et les mouvements du corps qui surviennent avec les émotions chez l'Homme et les animaux au moyen d'une théorie de l'évolution. Selon cette théorie, les émotions sont liées aux modèles de survie qui ont évolué dans une espèce donnée pour résoudre certains problèmes rencontrés lors de son évolution. Elles doivent être ainsi comprises en termes de fonction de survie.

Dans la perspective Darwinienne, de nombreux chercheurs se sont efforcés de démontrer l'universalité des émotions, notamment pour certains types d'expressions faciales ; [EKM-69][55], [TOM84][56], [FRI86][57] et [IZA94][58]. Les trois dernières décennies ont vu ces auteurs collecter une quantité impressionnante d'évidences démontrant l'universalité d'un ensemble restreint d'expressions faciales des émotions. Ce nombre d'émotions varie néanmoins selon les études, cf. table 1.3. Ekman utilise par exemple un ensemble de six émotions faciales universelles souvent appelé les « six grandes » (*big six*) [EKM69], [COR96][59], cf. Fig. 1.6.

---

[53] R. R. Cornelius, "Theoretical approaches to emotion", dans proc. *ISCA Tut. and Res. W. on Speech and Emotion*, Newcastle, Northern Ireland, Sep. 5-7 2000, pp. 3–10, 2000.

[54] C. Darwin, *The Expression of Emotion in Man and Animals*, dans John Murray, London, 1872 (P. Ekman [Eds], Oxford University Press, 3th Ed., 1998).

[55] P. Ekman et W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding", dans *Semiotica*, vol. 1, pp. 49–98, 1969.

[56] S. S. Tomkins, "Affect theory", dans K. R. Scherer and P. Ekman [Eds], *Approaches to Emotion*, Erlbaum, Hillsdale, NJ, pp. 163–196, 1984.

[57] N. H. Frijda, *The emotion*, dans Cambridge University Press, 1986.

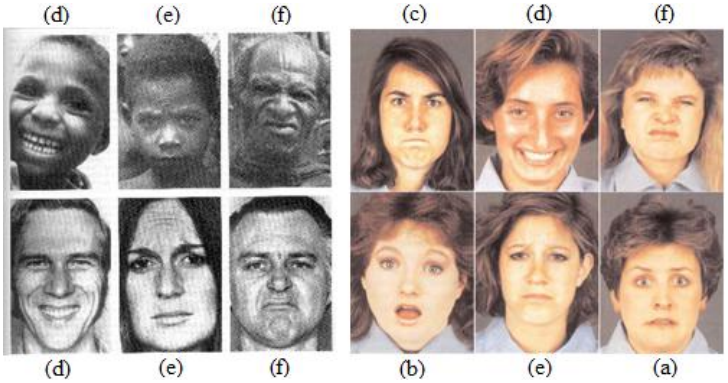


Fig. 1.6 Expressions faciales selon les différentes émotions primaires identifiées par Ekman : (a) : peur, (b) : surprise, (c) : colère, (d) joie, (e) : tristesse et (f) : dégoût ; images extraites de [EKM69][55].

James (*théorie de l'expérience*)

Dans son article de 1884 intitulé What is an emotion? [JAM84][60], William James (1842-1910), psychologue et philosophe Américain, a postulé que les émotions ne provoquent pas de changements physiologiques, et qu'elles correspondraient plutôt à la sensation liée aux changements corporels produits en situation affective. Ces changements sont causés par la perception d'un fait excitant. James illustre sa théorie de l'expérience en disant que « nous ressentons de la gêne parce que nous pleurons, de la colère parce que nous frappons, de la frayeur parce que nous tremblons, etc. » [JAM84], pp. 190. Ainsi, l'expérience des émotions implique l'existence de motifs uniques dans le jeu de réponses corporelles.

De nos jours, après plus d'un siècle de recherche, trois conclusions issues de la théorie de James peuvent être garanties selon Cornelius [COR00][53] : (i) la rétroaction engagée par les expressions faciales et les postures corporelles joue un rôle important dans les expériences émotionnelles [HOH66][61], [CHW88][62], (ii) les émotions peuvent être différenciées au niveau du système nerveux autonome [LEV92][63] et (iii) il existe des « programmes de l'affect » qui activent un certain nombre de systèmes expressifs, moteurs et expérimentaux. Concernant la parole, la prosodie est attendue comme étant l'un des systèmes activés par les « programmes de l'affect », et *vice versa*. Les paragraphes suivants décrivent les modèles de représentation des émotions fournies par la psychologie moderne.

3.3. Modèles de représentation des émotions

Puisque l'expérience émotionnelle dépend de nombreux facteurs contextuels et subjectifs, un nombre conséquent de termes compose le champ sémantique des émotions, cf. table 1.3.

[58] C. E. Izard, "Innate and universal facial expressions: Evidence from developmental and cross-cultural research", dans *Psychological Bulletin*, vol. 115, no. 2, pp. 288–299, 1994.
[59] R. R. Cornelius, *The science of emotion. Research and tradition in the psychology of emotion*, dans Prentice Hall, Upper Saddle River, NJ, 1996.
[60] W. James, "What is an emotion?", dans *Mind*, vol. 19, pp. 188-205, 1884.
[61] G. H. Hohmann, "Some effects of spinal cord lesions on experiences emotional feelings", dans *Psychophysiology*, vol. 3, pp. 143–156, 1996.
[62] K. Chwalisz, E. Diener et D. Gallagher, "Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured", dans *J. of Personality and Social Psycho.*, vol. 54, pp. 820–828, 1988.
[63] R. W. Levenson, "Autonomic nervous system differences among emotions", dans *Psycho. Sci.*, vol. 3, pp. 23–27, 1992.

Table 1.3 Catégorisation des émotions primaires selon des bases d'inclusion variées et proposées par différents auteurs ; table reprise de [ORT90][64].

Auteur	#	Emotions primaires	Bases d'inclusion
Plutchik	**8**	Acceptation, anticipation, colère, dégoût, joie, peur, tristesse, surprise	Relations aux processus adaptatifs biologiques
Arnold	**10**	Amour, aversion, colère, courage, dépression, désespoir, désir, haine, peur, tristesse	Relations aux actions tendancieuses
Ekman, Friesen et Ellsworth	**6**	Colère, dégoût, joie, peur, surprise, tristesse	Expressions faciales universelles
Fridja	**6**	Bonheur, désir, intérêt	Formes d'actions préparées
Gray	**4**	Anxiété, joie, rage, terreur	Câblés
Izard	**10**	Colère, culpabilité, dégoût, détresse, honte, intérêt, joie, mépris, peur, surprise	Câblés
James	**4**	Amour, grief, peur, rage	Participation du corps
Mc Dougall	**7**	Allégresse, colère, dégoût, étonnement, peur, soumission, tendresse	Relations aux instincts
Mowrer	**2**	Douleur, plaisir	Etat émotionnels non-appris
Oatley et Johnson-Laird	**5**	Colère, dégoût, anxiété, allégresse, tristesse	Ne requiert pas un contenu propositionnel
Paksepp	**4**	Espérance, panique, peur, rage	Câblés
Tomkins	**9**	Colère, intérêt, grief, dégoût, détresse, peur, joie, honte, surprise	Densité de l'activité neuronale
Watson	**3**	Amour, peur, rage	Câblés
Weiner et graham	**2**	Allégresse, tristesse	Attributs indépendants

un nombre conséquent de termes compose le champ sémantique des émotions, cf. table 1.3. Aussi, les psychologues ont cherché à définir des modèles permettant d'expliquer les traits caractéristiques des émotions. Toutefois, la nature discrète, idiosyncrasique et parfois mystérieuse de nos états affectifs a posé de nombreux problèmes aux psychologues lors de la définition des modèles de représentation des émotions.

3.3.1. Modèles catégoriels

Les catégories affectives issues du langage de tous les jours constituent l'une des façons les plus naturelles qu'il soit pour décrire les émotions [COW03][45]. Toutefois, les recherches qui ont été effectuées dans ce domaine ont révélé de nombreuses difficultés. La taille du lexique des émotions en est un exemple tout à fait remarquable. Pour l'Anglais, le *Semantic Atlas of Emotional Concepts* liste 558 mots à « connotation émotionnelle » [AVE75][65] et le Français comprendrait jusqu'à 950 mots décrivant les « émotions et les états psychologiques » [MAT05][66]. Dans les études qui se sont penchées sur les termes plus spécifiques aux émotions, le nombre de mots a été diminué de façon significative pour l'Anglais et varie alors

[64] A. Ortony et T. J. Turner, "What's basic about basic emotions?", dans *Psychological Review*, vol. 97, pp. 315–331, 1990.

[65] J. R. Averill, "A semantic atlas of emotional concepts", dans *JSAS Catalog of Selected Documents in Psycho.*, vol. 5, pp. 330, 1975.

autour d'une centaine : 107 [WHI89][67], 135 [STO87][68], 142 [PLU80][69], 196 [FEH84][70] et 213 [SHA-87][71]. Les études sur d'autres langues telles que l'Allemand et l'Italien ont montré des résultats comparables : 235 [SCH84][72] et 153 [ZAM98][73].

Cowie et Cornelius ont affirmé que les « langues ne multiplient généralement pas les termes sauf si cela s'avère nécessaire » [COW03][45]. L'existence des marqueurs linguistiques des émotions ne seraient donc pas apparus sans une certaine forme de nécessité liée à l'usage. Selon ces auteurs, 60 catégories seraient au minimum nécessaires pour décrire les émotions qui apparaissent dans la vie de tous les jours à une fréquence raisonnable [COW03]. Bien que toutes les nuances d'émotions distinguables par l'Homme ne puissent être couvertes par ces catégories, cela fait néanmoins apparaître un dilemme quant à leur représentativité. En effet, des conclusions fiables ne peuvent être atteintes qu'avec un nombre de données conséquent pour chaque catégorie d'émotion. Afin de contribuer au perfectionnement des systèmes de TAP orienté émotion, de telles données sont nécessaires pour prétendre à une analyse robuste en prenant en compte un ensemble plus représentatif des formes affectives observables chez un individu. Cowie propose une soixantaine de catégories [COW03] alors que les corpus actuellement disponibles pour la recherche n'en contiennent jamais plus d'une dizaine.

3.3.2. Modèles dimensionnels

Une alternative aux modèles catégoriels repose sur l'utilisation de modèles dimensionnels [DUF41][74]. La théorie sous-jacente est que les émotions peuvent être distinguées au moyen de certaines caractéristiques. Cette propriété est fondamentale pour le TAP orienté émotion.

Modèles unidimensionnels

Les modèles unidimensionnels considèrent qu'une seule dimension est suffisante pour différencier les émotions. Cette dimension pourrait être soit *l'activation*, i.e., le sentiment subjectif activé ou désactivé, soit *l'excitation* (*arousal*), i.e., le sentiment subjectif de la plaisance ou de la déplaisance. Selon la dimension *activation – excitation*, les différences majeures entre les émotions est le degré d'*excitation* qui varie alors dans de grandes proportions. Le

---

[66] Y. Y. Mathieu, "Annotation of emotions and feelings in texts", dans proc. *ACII*, Beijing, China, Oct. 22-24 2005, vol. 3784, pp. 350–357.

[67] C. Whissel, "The dictionary of affect in language", dans R. Plutchik and H. Kellerman, [Eds], *Emotion: Theory, Research and Experience*, vol. 4, *The Measurement of Emotion*, pp. 113–131, Academic Press, New York, 1989.

[68] C. Storm et T. Storm, "A taxonomic study of the vocabulary of emotion", dans *J. of Personality and Social Psycho.*, vol. 53, pp. 805–816, 1987.

[69] R. Plutchik, Emotion: *A Psychoevolutionary Synthesis*, dans Harper & Row, New York, 1980.

[70] B. Fehr et J. A. Russel, "Concept of emotion viewed from a prototype perspective", dans *J. of Experim. Psycho.*, vol. 113, pp. 464–486, 1984.

[71] P. Shaver, J. Schwartz, D. Kirson et C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach", dans *J. of Personality and Social Pscyho.*, vol. 52, pp. 1061–1086, 1987.

[72] K. R. Scherer, "Emotion as a multicomponent process: a model and some cross-cultural data", dans *R. of Personality and Social Pscyho.*, vol. 5, pp. 37–63, 1984.

[73] V. L. Zammuner, "Concepts of emotion: 'emotionness' and dimensional ratings of Italian words", dans *Cogn. and Emotion*, vol. 12, pp. 243–272, 1998.

[74] E. Duffy, "An explanation of 'emotional' phenomena without the use of the concept 'emotion'", dans *J. of General Pscyho.*, vol. 25, pp. 283–293, 1941.

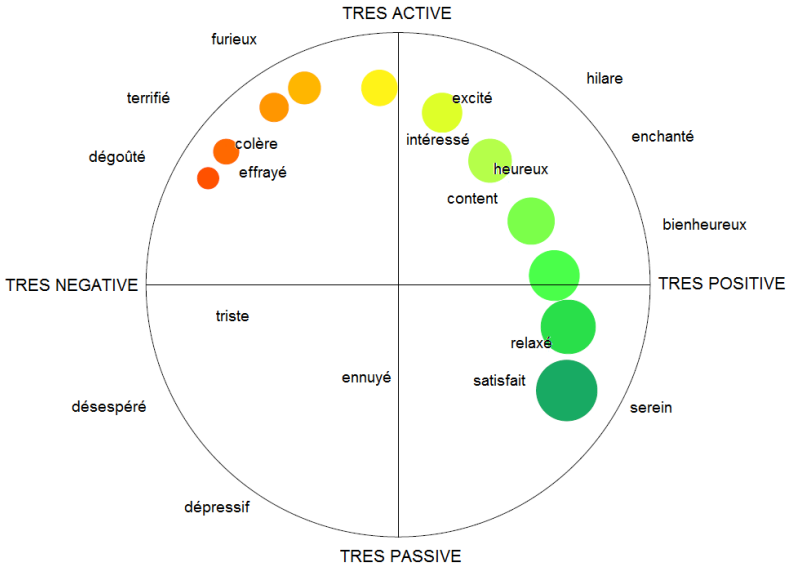


Fig. 1.7 *Feeltrace* : étiquetage dimensionnelle des émotions dans l'espace activation–évaluation ; figure reproduite de [COW01][50].

niveau d'*activation* peut être vu comme la force des prédispositions que présentent une personne pour entreprendre des actions telles que celles définies par James (théorie de l'expérience), ou encore celles qui conduisent à agir de certaines façons, e.g., les actions tendancieuses de Fridja [FRI86][57].

Modèles multidimensionnels

Dans les modèles multidimensionnels, les états émotionnels peuvent être représentés par des coordonnées dans un espace de faible dimension. Wundt a par exemple suggéré d'utiliser trois dimensions indépendantes : plaisance *vs.* déplaisance, repos *vs.* activité et relâchement *vs.* attention [WUN93][75]. Schlosberg démontra la pertinence des deux dimensions *activation* et *excitation* puisqu'elles permettent de représenter une quantité considérable d'informations sur les émotions [SCH54][76]. D'autres auteurs ont également contribué à la popularité des modèles bidimensionnels, e.g., Plutchik [PLU62][77] et Russel [RUS80][78]. Cowie *et al.* qualifient l'association des dimensions comme l'espace *activation – évaluation* [COW01][50]. Cet espace correspond à des émotions pleines réparties sur une forme grossièrement circulaire, cf. Fig. 1.7.

Le modèle bidimensionnel *activation – évaluation* fournit un moyen de décrire les états émotionnels qui est plus facilement exploitable que les mots. Toutefois, ceux qui sont associées à une émotion donnée peuvent être représentés en termes de positions dans l'espace bidimensionnel, et *vice versa*. Les similarités et les différences présentes entre les émotions peuvent s'exprimer au moyen d'une distance Euclidienne dans l'espace *activation – évaluation*. Cette propriété fut exploitée avec succès par la communauté du TAP orienté émotion. Notons toutefois que la réduction de l'espace des caractéristiques à deux dimensions est ac-

[75] W. Wundt, *Grundzüge der Physiologischen Psycho.*, dans Leipzig: Verlag von Wilhelm Engelmann, 4th revised Ed. [first published 1873], 1893.
[76] H. Schlosberg, "Three dimensions of emotion", dans *Psycho. R.*, vol. 61, pp. 81–88, 1954.
[77] R. Plutchik, *The Emotions: Facts, Theories, and a New Model*, dans Random House, New York, 1962.
[78] J. A. Russel, "A circumplex model of affect", dans *J. of Personality and Social Psycho.*, vol. 39, pp. 1161–1178, 1980.

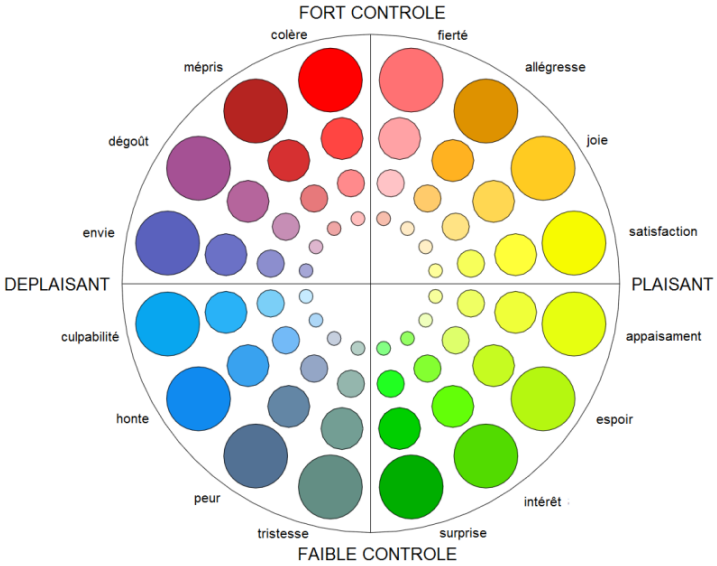


Fig. 1.8 Roue de l'émotion de Genève ; figure reproduite de [BAN05][79].

compagnée de quelques désagréments. Certaines émotions telles que la *Colère* et la *Peur* sont par exemple tellement proches dans l'espace des caractéristiques qu'il semble difficile de pouvoir les séparer convenablement. Ce problème peut être résolu par l'ajout de dimensions supplémentaires telles que le contrôle perçu ou l'inclinaison à l'engagement, mais il faut alors définir à partir de quel moment il apparaît raisonnable de s'arrêter dans le processus d'ajout de dimensions supplémentaires [CgyOW03][45].

La roue de l'émotion de Genève (*Geneva Emotion Wheel*) est un modèle qui combine à la fois les dimensions et les catégories d'informations affectives pour représenter les adjectifs associés [BAN05][79]. Les dimensions sous-jacentes sont : (i) le contrôle perçu (fort / faible) et (ii) la valence (négative / positive). Les prototypes issus de ce modèle diffèrent du modèle *Feeltrace* par le nombre de familles d'émotions et par leur ordre dans la roue (16 familles en l'occurrence contre 12 zones dans Feeltrace). Chaque famille d'émotions est composée de 4 membres d'intensité variable qui sont représentés par des cercles de teinte identique et par des tailles et des couleurs différentes, cf. Fig. 1.8. La taille et la saturation de la couleur des cercles augmentent alors avec l'intensité attribuée à un adjectif d'émotion.

Plutchik assume l'existence de huit émotions primaires à travers quatre paires d'émotions opposées : (i) *Joie vs. Tristesse*, (ii) *Acceptation vs. Dégoût*, (iii) *Peur vs. Colère* et (iv) *Surprise vs. Anticipation* [PLU80][69]. Dans ce modèle, les émotions opposées se neutralisent entres elles et sont arrangées dans des secteurs opposés sur un cercle d'émotions, cf. Fig. 1.9. Les mélanges d'émotions adjacentes (*dyades primaires*) peuvent produire des émotions plus complexes, e.g., l'émotion complexe du *Mépris* peut être obtenue en mélangeant les émotions primaires de la *Colère* et du *Dégoût*. Tous les mélanges des émotions présentes sur le cercle ne peuvent toutefois conduire à des catégories existantes. Ainsi, les mélanges des émotions qui sont plus largement séparées sur le cercle (*dyades secondaires* et *tertiaires*) sont plus rarement observables que celles qui sont à proximité immédiate.

[79] T. Bänziger, V. Tran, et K. R. Sherer, "The Geneva Emotion Wheel: A tool for the verbal report of emotion reactions", dans *ISRE*, Bari, Italy, Jul. 11-15, 2005, pp. 241–254.

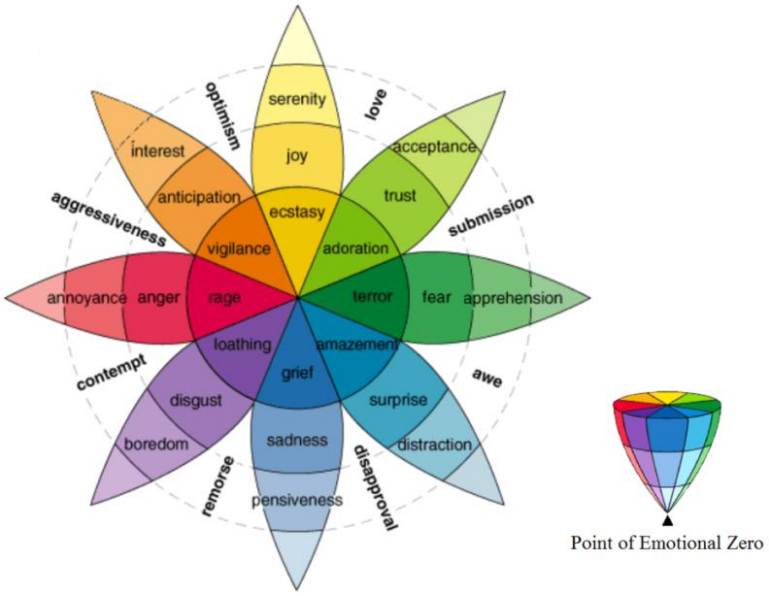


Fig. 1.9 Représentation des émotions à travers la roue de Plutchik : catégories d'émotions opposées et variation selon leur intensité ; figure reproduite de [PLU80][69].

3.4. Corrélats acoustiques de l'affect

Une revue exhaustive des travaux portant sur la caractérisation des corrélats acoustiques des émotions est hors de portée de cette section. Par conséquent, seules les tendances générales seront décrites dans les paragraphes suivants. Notre description s'appuie notamment sur les travaux de Murray et Arnott [MUR93][80] et de Scherer [SCH03][52]. Les recherches qui ont été conduites sur l'analyse des corrélats acoustique de l'affect peuvent se distinguer selon leur centre d'intérêt principal : le processus *d'encodage* ou de *décodage* des informations [BAN-01][81]. Les études qui concernent le processus *d'encodage* visent à décrire l'effet produit par différents états affectifs sur le système phonatoire. La caractérisation du processus de *décodage* des émotions repose quant à lui sur l'analyse des paramètres qui sont pris en compte par un individu pour caractériser une émotion donnée.

3.4.1. Encodage acoustique de l'émotion

La communication des émotions par la voix ne peut s'envisager que par l'utilisation appropriée de codes préétablis, puisqu'ils permettent de spécifier les correspondances entre un ensemble de caractéristiques physiques observables et une catégorie d'émotion donnée. Un grand nombre d'auteurs se sont penchés sur la tâche d'identification des codes de l'affect. Les méthodes ont reposé sur un ensemble de contextes variés permettant d'induire des émotions au moyen d'un contexte : (i) *naturel* (e.g., interactions spontanées), (ii) *semi-contrôlé* (e.g., tâches de type magicien d'Oz) ou (iii) *contrôlé* (e.g., discours affectif acté).

Le contexte *semi-contrôlé* permet d'observer des émotions associées à des modifications

[80] I. R. Murray et J. L. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion", dans *J. of Acous. Soc. of Amer.*, vol. 93, no. 2, pp. 1097–1198, 1993.

[81] T. Bänziger, D. Grandjean, P. J. Bernard, G. Klasmeyer et K. R. Scherer, "Prosodie de l'émotion : étude de l'encodage et du décodage", dans *Cahiers de Lingu. Française*, no. 23, pp. 11–37, 2001.

plus légères que celles induites par les deux autres contextes. L'utilisation de données actées a très souvent été préférée aux deux autres méthodes. Ce type de contexte présente en effet de nombreux avantages tels que : (i) le contrôle de l'influence du contenu linguistique sur la production des émotions et (ii) la possibilité d'avoir une large gamme d'états affectifs différents pour un même individu. Son usage très répandu a parfois été critiqué sur le fait que les émotions actées ne correspondraient pas, ou peu, à de véritables émotions. Mais il est néanmoins très peu probable que les émotions actées ne correspondent en rien aux expressions affectives qui surviennent dans les interactions sociales de la vie de tous les jours. Afin de paraître crédibles auprès de leurs auditeurs, les acteurs doivent en effet employer des codes expressifs qui seront interprétés comme authentiques par leurs auditeurs. En revanche, comme le processus utilisé est relativement artificiel, i.e., la cause du phénomène n'est pas vécue sur l'instant, l'émotion produite contient certainement une forme d'exagération. De plus, il est probable qu'une partie de la composante expressive habituellement présente dans les réactions physiologiques liées à l'affect, soit absente des enregistrements produits par les acteurs. Par conséquent les émotions actées correspondent à des émotions dites *prototypiques*.

Les corrélats acoustiques des émotions encodés dans la parole sont très souvent identifiés par des mesures prosodiques, i.e., le pitch, l'intensité, la qualité vocale et la durée [MUR93][80]. L'intonation est alors bien souvent considérée comme le marqueur acoustique principal des émotions [GUS01][82]. De nombreuses méthodes de stylisation des contours intonatifs ont été proposées dans la littérature, [HIR93][83], [SIL92][84] et [GRA98][85]. La plupart de ces méthodes reposent sur le principe de la stylisation par copie conforme (*close-copy stylisation*) pour traiter la fréquence fondamentale [MER04][86]. Des critiques ont toutefois été apportées sur les systèmes de codification de la prosodie, puisqu'ils réduisent le continuum de valeurs en un jeu élémentaire de symboles recouvrant alors qu'une petite partie de l'ensemble des phénomènes observables. De plus, les systèmes comme ToBI (*tones and break indices*) ne prennent pas en compte les aspects de la perception de la parole [WIG02][87].

Murray et Scherer ont préféré passer en revue les motifs acoustiques qui caractérisent l'expression vocale des émotions modales majeures ; *Colère*, *Peur*, *Joie*, *Tristesse* et *Dégoût* [MUR93], et *Excitation* / *Stress*, *Joie* / *Allégresse*, *Colère* / *Rage*, *Tristesse*, *Peur* / *Panique* et *Ennui* [SCH03][52]. Les résultats obtenus par ces deux études sont présentés dans les tables 1.4 et 1.5. Une grande partie de la cohérence de ces résultats est liée aux différents niveaux d'excitation de l'émotion cible. En effet, il a souvent été supposé que la voix ne peut marquer que l'excitation physiologique et ne permet pas de communiquer des différences qualitatives entre l'émotion, contrairement aux expressions du visage. Selon Scherer, la difficulté à démontrer

---

82 S. Gustafson-Capková, "*Emotions in speech: Tagset and acoustic correlates",* dans *Term paper in Speech Technology*, GSLT, Stockholm University, Depart. of Ling., Aut. 2001.

83 D. Hirst et R. Espesser, "Automatic modeling of fundamental frequency using a quadratic spline function", dans *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, vol. 15, pp. 75–85, 1993.

84 K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert et J. Hirschberg, "ToBI: A standard scheme for labeling prosody", dans proc. *2nd ICSLP*, Banff (AL), Canada, Oct. 13-16 1992, pp. 867–869.

85 E. Grabe, F. Nolan et K. Farrar, "IViE—A comparative transcription system for intonational variation in English", dans proc. *ICSLP*, Sydney, Australia, Nov. 30-Dec. 4 1998, pp. 1259–1262.

86 P. Mertens, "The Prosogram: 'Semi-automatic transcription of prosody based on a tonal perception model'", dans *Speech Prosody*, Nara, Japan, Mar. 23-26 2004, pp. 143–146.

87 C. W. Wightman, "ToBI or not ToBI ?", dans proc. *Speech Prosody*, Aix-en-provence, France Apr. 11-13 2002, pp. 25–29.

Table 1.4 Effets des émotions sur un ensemble de paramètres acoustiques : une revue synthétique des résultats présentés dans [SCH03][52].

Paramètres acoustiques	**Excit./ Stress**	**Joie/ Allégr.**	**Colère/ Rage**	**Trist.**	**Peur/ Panique**	**Ennui**
débit de parole et fluence						
nombre de syllabes par seconde	>	≥	<>	<	>	<
durée de la syllabe	<	≤	<>	>	<	>
durée des voyelles accentuées	≥	≥	>	≥	<	≥
nombre et durée des pauses	<	<	<	>	<>	>
durée relative des segments voisés			>		<>	
durée relative des segments non-voisés			<		<>	
source vocale (f0) et prosodie						
f0 : moyenne	>	>	>	<	>	≤
f0 : 5ème percentile	>	>	=	≤	>	≤
f0 : écart-type	>	>	>	<	>	<
f0 : amplitude	>	>	>	<	<>	≤
f0 : gradient des valeurs ↑ et ↓	>	≥	>	<		
f0 : amplitude et gradient de la chute finale	>	>	>	<	<>	≤
fréquence des syllabes accentuées	>	>	>	<	<>	≤
effort de la source vocale et type de phonation						
intensité : moyenne	>	≥	>	≤		≤
intensité : écart-type	>	>	>	<		<
intensité : gradient des valeurs ↑ et ↓	>	≥	>	<		≤
énergie spectrale dans les HF	>	>	>	<	<>	≤
pente spectrale	<	<	<	>	<>	>
effort du larynx		=	=	>	>	=
jitter		≥	≥		>	=
shimmer		≥	≥		>	=
rapport harmonique sur bruit HNR		≥	>	>	<	≤
Articulation – vitesse et précision						
formants – précision de la position	?	=	>	<	≤	≤
formants–largeur de bande	<		<	>		≥

<> : des cas à la fois d'augmentation et de diminution ont été reportés.

Table 1.5 Effets des émotions sur un ensemble de paramètres acoustiques : une revue synthétique des résultats présentés dans [MUR93][80].

Auteur	**Colère**	**Joie**	**Tristesse**	**Peur**	**Dégoût**
Débit de parole + Rapide, – Lent	Légèrement +	+ et –	Légèrement –	Beaucoup +	Enormément +
Moyenne du pitch + Elevée, – Basse	Enormément +	Beaucoup +	Légèrement –	Enormément +	Enormément +
Amplitude du pitch + Large, – Réduite	Beaucoup +	Beaucoup +	Légèrement –	Beaucoup +	Légèrement –
Intensité + Elevée, – Basse	+	+	–	Normale	–
Qualité vocale	Haletante, tons de poitrine	Haletante, hurlante	Résonante	Voisement irrégulier	Grommelant, tons de poitrine
Variations du pitch	Abruptes, sur les syllabes accentuées	Lisses, inflexions montantes	Inflexions descendantes	Normale	Larges inflexions descendantes terminales
Articulation	Tendue	Normale	Bredouillée	Précise	Normale

une différenciation qualitative des émotions à travers les paramètres prosodiques repose sur deux faits : (i) seul un nombre limité d'indices acoustiques, principalement la f0, l'énergie et le débit de parole, ont été étudiés alors que peu d'études ont analysé les paramètres fournis par les formants et la dynamique, i.e., le rythme. Il est donc possible que ces caractéristiques aient un impact sur des différences qualitatives des émotions, alors que la f0, l'énergie et le débit de parole fournissent plus des indications sur l'excitation. Et (ii) les écarts dans les études peuvent être causés par différentes formes d'une même émotion (e.g., *Colère* chaude, *Rage* explosive *vs. Colère* froide, modérée ou contrôlée). Par conséquent, les résultats présentés dans les tables 1.4 et 1.5 montrent qu'il existe plus des tendances générales pour certains corrélats acoustiques de certaines formes d'émotions que de véritables associations systématiques.

3.4.2. Décodage acoustique de l'émotion

Les recherches portant sur la reconnaissance des émotions communiquées par la voix ont été abordées à la fois par les psycho-acousticiens et les chercheurs en TAP. Comme les études produites par les sciences computationnelles sont décrites dans la section 4, nous décrivons dans les paragraphes suivants uniquement celles issues des tests de perception. Les émotions (souvent exprimées par des acteurs) sont plutôt bien identifiées par des auditeurs. Une revue relativement exhaustive de la littérature réalisée en 1989 par Scherer indique que le pourcentage de reconnaissance correcte moyen est d'environ 60% [SCH89][88]. Dans une revue plus récente et portant sur 11 études réalisées dans différents pays occidentaux, Scherer *et al.* rapportent un pourcentage de reconnaissance moyen de 62% avec des variations notables selon les émotions [SCH01a][89].

Peu d'études se sont intéressées aux processus de décodage des informations comparé à celles portant sur la reconnaissance des expressions vocales de l'affect [BAN01][81]. Celles qui ont été effectuées dans ce domaine ont utilisé plusieurs approches pour effectuer les tests en perception. Certaines approches ont quantifié les corrélations présentes entre les caractéristiques acoustiques des expressions et les attributs choisis par des auditeurs pour identifier les émotions. D'autres ont vérifié si les émotions restaient identifiables même si une partie de l'information disponible était manquante. La technique la plus fréquente consiste à filtrer les fréquences du signal de parole de façon à supprimer les informations relatives au timbre vocal et le contenu phonétique des expressions, tout en conservant l'essentiel des informations rythmiques et mélodiques du signal. Ces travaux ont montré qu'il reste possible d'identifier les émotions exprimées même lorsque l'on supprime certaines dimensions de l'information. L'émotion serait donc principalement communiquée par les aspects mélodiques et rythmiques de la voix [SCH85a][90].

Une autre technique consiste à manipuler certaines caractéristiques des expressions via des algorithmes de synthèse (ou resynthèse) de la parole. Cette approche serait l'une des plus

[88] K. R. Scherer, "Vocal correlates of emotion", dans H. Wagner and A. Manstead [Eds], *Handbook of psychophysiology: Emotion and social behavior*, Wiley and Sons Ltd, London, chap. 7, pp. 165–197, 1989.

[89] K. R. Scherer, R. Banse et H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures", dans *J. of Cross-Cultural Psycho.*, vol. 32, pp. 76–92, 2001.

[90] K. R. Scherer, S. Feldstein, R. N. Bond et R. Rosenthal, "Vocal cues to deception: A comparative channel approach", dans *J. of Psycholingu. Res.*, vol. 14, pp. 409–425, 1985.

prometteuses [BEL09][91] puisqu'elle permet de manipuler de façon séparée un large ensemble de paramètres vocaux sur lequel les effets des attributions émotionnelles peuvent être évalués. Ces études ont confirmé l'intervention de plusieurs dimensions acoustiques différentes dans le processus d'attribution émotionnelle. Parmi les dimensions manipulées, les variations dans le temps de la fréquence fondamentale (contour de la f0) semble jouer un rôle particulièrement important, cf. [BAN01][81] pour une revue des études portant sur l'intonation. Les caractéristiques dynamiques de la prosodie semblent ainsi jouer un rôle important dans la communication des états émotionnels.

4. La reconnaissance automatique des émotions par la parole

Notre travail porte sur la reconnaissance automatique des émotions et des états affectifs pour la modalité parole seule. Les sections suivantes présentent un résumé de l'état-de-l'art actuel dans ce domaine.

4.1. Etat de l'art

Les psychologues étudient depuis plus de 50 ans l'influence des émotions sur la parole en exploitant bien souvent des émotions prototypiques. Les sciences computationnelles se sont réellement impliquées dans la tâche de reconnaissance des émotions à partir des années 2000. Elles ont alors exploité des techniques issues des domaines du TAP et de la reconnaissance des formes pour extraire des caractéristiques sur un signal de parole et leur attribuer de façon automatique une étiquette d'émotion donnée, cf. sous-section 1.3. Les méthodes automatiques offrent l'avantage de traiter une très grande quantité de données dans un temps adéquat. De tels traitements entraîneraient par exemple un coût beaucoup plus important s'ils étaient effectués par l'Homme comparé à une machine. Les corpus de parole disponibles à ce jour sont bien souvent composés d'émotions actées pour un nombre relativement restreint d'émotions. La forte popularité des corpus de parole actée est enracinée dans les avantages intrinsèques à cette méthode de collecte des données : (i) les émotions sont représentées de façon intense par des expressions prototypiques, la recherche de corrélats et leur classification sont donc facilitées, (ii) les studios d'enregistrements permettent de limiter, de par la qualité des équipements, les problèmes liés au traitement de la parole bruitée ou avec écho, (iii) la répartition des données selon les émotions peut être équilibrée de façon à éviter d'utiliser des méthodes de sous- ou sur-échantillonnage des données, (iv) les données peuvent être collectées dans un temps relativement court et pour un coût relativement faible et (v) aucun étiquetage des émotions n'est nécessaire puisque ces dernières sont connues à l'avance.

Un tel corpus de parole est typiquement issu d'une lecture d'un texte non émotionnel. Le discours est donc non-interactif et l'analyse linguistique des émotions est rendue impossible. Il existe des corpus de parole émotionnelle actée pour de nombreuses langues, [VER03][92] et [SCH09a][93]. Trois corpus ont toutefois été principalement utilisés dans la littérature pour effectuer le TAP orienté émotion tels que : (i) le corpus Anglais *Emotional Prosody Speech and*

[91] G. Beller, *Analyse et modèle génératif de l'expressivité: Application à la musique et à l'interprétation musicale*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 2009.

Transcripts du Consortium des Données en Linguistique (LDC[94], sous droits d'auteur), (ii) le corpus *Danish Emotional Speech Corpus* [ENG96][95] et (iii) le corpus *Berlin Emotional Speech Database* [BUR05][96] (disponibles pour la recherche).

Un état-de-l'art des performances en reconnaissance montre qu'elles peuvent être très élevées sur la parole actée et que le discours spontané pose par contre beaucoup plus de problèmes [SCH09a][93]. Schuller *et al.* ont par exemple obtenu des scores de reconnaissance de 75% et de 88% sur les corpus de parole actée en langue Danoise (DES) et Allemande (Berlin) [SCH06b][97]. Ces résultats sont plutôt remarquables puisqu'ils sont supérieurs aux scores produits par les tests en perception effectués par l'Homme ; 67% pour DES et 84% pour Berlin. Ainsi, le traitement automatique des corrélats acoustiques produits par différents états affectifs sur un ensemble plutôt élevé de locuteurs (une dizaine) fournit des scores comparables à l'Homme.

Au-delà de l'intérêt général porté par le TAP orienté émotion, il existe un intérêt sans cesse croissant pour appliquer ces systèmes dans un ensemble varié de domaines tels que les centres d'appel téléphonique ou les systèmes communicants en général, e.g., robotique, jeux-vidéo, vidéo-surveillance, etc. Toutefois, peu d'études ont été effectuées dans l'analyse des émotions spontanées, et ces rares études ont bien souvent dû se limiter à un ensemble relativement restreint d'états affectifs différents. De plus, les émotions naturelles dépendent fortement du contexte apporté par le scénario choisi. Certaines études se sont par exemple penchées sur la détection d'un seul état affectif tel que *l'Ennui* et la *Frustration*, [ANG02][98] et [KAP07][99], la *Colère* [ARI07][100] la *Peur* [CLA06][101] ou le *Mamanais* [MAH10][102]. Toutefois, la création récente d'un corpus de parole spontanée produite par des enfants [STE09][42] a permis d'étudier l'impact d'un contexte naturel dans le TAP orienté émotion. Les résultats de ses travaux sont présentés dans le chapitre 5, sous-section 4.1.

Un très large ensemble de caractéristiques a été proposé jusqu'ici pour reconnaître les expressions affectives dans un signal de parole. Les caractéristiques peuvent être catégorisées en paramètres acoustiques et prosodiques (intonation, intensité, qualité vocale et rythme). Au-delà de ces paramètres, les caractéristiques linguistiques constituent une autre source d'infor-

---

92 D. Ververidis et C. Kotropoulos, "A review of emotional speech databases", dans 9th *Panhellenic C. in Informatics*, Thessaloniki, Greece, Nov. 1-23, 2003, pp. 560–574.

93 B. Schuller, *Emotion Recognition in the Next Generation: an Overview and Recent Development*, Tutoriel à *Interspeech*, Brighton, UK, Sep. 6-10 2009.

94 http://ldc.upenn.edu/Catalog.

95 I. S. Engbert et A. V. Hansen, "Documentation of the Danish Emotional Speech Database DES", dans *Tech. Rep. Center for PersonKommunikation*, Aalborg University Denmark, 1996.

96 F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier et B. Weiss, "A database of German emotional speech", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 1517–1520.

97 B. Schuller, D. Arsík, F.Wallhoff et G. Rigoll, "Emotion recognition in the noise applying large acoustic features sets", dans proc. *Speech Prosody*, Dresden, Germany, May. 2-5 2006.

98 J. Ang, R. Dhillon, E. Shriberg et A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in Human-computer dialog", dans proc. *7th ICSLP*, Denver (CO), USA, Sep. 16-20 2002, pp. 2037–2040.

99 A. Kapoor, W. Burleson et R. W. Picard, "Automatic prediction of frustration", dans *Inter. J. of Human-Computer Studies*, vol. 65, pp. 724–736, 2007.

100 Y Arimoto, S. Ohno et H. Iida, "Acoustic features of anger utterances during natural dialog", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2217–2220.

101 C. Clavel, *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales*, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, TELECOM Paris, 2007.

mations importante des émotions. Comme les deux précédents types de paramètres sont décrits dans le chapitre 3, section 2 (acoustique), et dans le chapitre 4, section 2 (prosodie), nous présentons dans les paragraphes suivants uniquement les études qui ont exploité des paramètres linguistiques pour effectuer la reconnaissance des émotions.

4.2. Informations linguistiques

Afin de préserver l'influence du contexte linguistique sur les émotions, la majorité des corpus de parole actée contiennent des expressions émotionnelles issues de la lecture d'un texte exempt de termes implicitement ou indirectement reliés aux émotions. Par conséquent, l'analyse des caractéristiques linguistiques a souvent été négligée. Cependant, elles peuvent fournir une source complémentaire d'indices émotionnels. Une première approche consiste à quantifier les probabilités $p(E_i \mid w_1 w_2 \dots w_N)$ d'une émotion E_i sachant la séquence formée par l'enchaînement d'une unité linguistique w_k dans une phrase donnée. Ces unités peuvent alors être très diversifiées, e.g., phonèmes, syllabes, mots, etc., cf. chapitre 2. Comme dans les modèles de langue pour la reconnaissance automatique de la parole [NOT01][103], les *n-grams* $p(E_i \mid w_{N-n+1} \dots w_N)$ sont utilisés pour réduire le contexte et des techniques de lissage sont employées pour traiter les cas non observés. Puisque les observations des *n-grams* dans les données sont très peu fréquentes avec $n \geq 3$, seuls les modèles *unigrams* [DEV03][104] ou *bi-grams* [POL00][105] semblent prometteurs. Certains auteurs ont proposé des améliorations de ces techniques. Lee a par exemple proposé de composer des *unigrams* avec uniquement les unités linguistiques qui ont été évaluées comme saillantes en regard des émotions [LEE02][106]. La quantité d'informations qu'un mot contient à propos d'une catégorie d'émotion donnée a ainsi été utilisée pour définir la saillance. Par ailleurs, le ratio de gain d'information (*information gain ratio* IGR) a été appliqué par Schuller pour sélectionner les mots les plus discriminants des émotions [SCH05][107]. Plusieurs alternatives au calcul de $\prod_w p(E_i \mid w)$ ont alors été considérées.

Une autre façon d'utiliser les informations lexicales repose sur les représentations en *sac-de-mots* (*bag-of-words*). Dans cette approche, chaque composante d'un vecteur caractéristique correspond à une entrée d'un lexique associé et contient la fréquence (absolue et relative) des mots respectifs dans l'énoncé étudié. La valeur absolue de la fréquence peut être pondérée avec la *fréquence inverse du document* [SAL88][108] et/ou le logarithme peut être calculé. Notons que les informations portées par les enchaînements des mots sont perdues dans les ap-

[102] A. Mahdhaoui, *Analyse des signaux sociaux pour la modélisation de l'interaction face-à-face*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 2010.
[103] E. Nöth, A. Batliner, H. Niemann, G. Stemmer, F. Gallwitz et J. Spilker, "Language models beyond word strings", dans proc. *ASRU*, Trento, Italy, Dec. 9-13 2001.
[104] L. Deviller, L. Lamel et I. Vasilescu, "Emotion detection in task-oriented spoken dialogs", dans proc. *4th ICME*, Baltimore (MD), USA, Jul. 6-9 2003, pp. 549–552.
[105] T. S. Polzin et A. Waibel, "Emotion-sensitive Human-computer interfaces", dans proc. *ISCA Tutorial and Res. W. on Speech and Emotion*, Newcastle, Northern Ireland, Sep. 5-7 2000, pp. 201–206.
[106] C. M. Lee, S. Narayanan et R. Pieraccini, "Combining acoustic and language information for emotion recognition", dans proc. *7th ICSLP*, Denver (CO), USA, Sep. 16-20 2002, pp. 873–876.
[107] B. Schuller, R. Jiménez Villar, G. Rigoll et M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition", dans proc. *ICASSP*, Philadelphia (PA), USA, Apr. 6-10 2005, pp. 325–328.

proches *sac-de-mots*. Des techniques ont été proposées pour réduire la taille du vecteur caractéristique, telles que : (i) l'exclusion de certains mots du champ d'analyse, (ii) le regroupement des mots selon des catégories [GUP07][109] et (iii) les analyses statistiques et les méthodes de sélection de caractéristiques permettant d'identifier les paramètres discriminants. Les mots peuvent également être regroupés en utilisant des caractéristiques de type *partie-de-parole* (*part-of-speech* – POS) [GUP07]. Bulut a ainsi pu différencier quatre émotions à travers les phrases en exploitant 13 types d'étiquettes POS [BUL07][110]. Enfin, une troisième méthode de traitement des informations lexicales consiste à identifier les mots clés des émotions au moyen des réseaux de croyance (*belief networks*). Un réseau de cinq niveaux réalisant un regroupement en super-mots, phrases, super-phrases et finalement aux émotions est suggéré dans [SCH04b][111].

5. Contribution à la recherche sur l'émotion

Les recherches actuelles dans le domaine du TAP orienté émotion montre que de nombreuses techniques d'extraction de caractéristiques et de classification peuvent être exploitées pour identifier automatiquement les états émotionnels. La performance de ces systèmes pour différencier divers états affectifs sur un ensemble de locuteurs est assez élevée, voire même comparable à celle de l'Homme. Cependant, la plupart de ces études ont analysé des données actées et donc associées à des émotions prototypiques. Les très bons résultats de classification obtenus sur ce type de données permettent d'élever le niveau de difficulté en se concentrant sur des émotions plus naturelles, comme celles qui peuvent être observées dans les scénarios de la vie de tous les jours. Les corpus de parole adéquat à la recherche sur les émotions spontanées sont toutefois rares. De plus, même si d'excellents scores ont été obtenus sur les corpus de parole actée, les méthodes qui ont été utilisées ne permettent pas d'identifier quels sont les types de paramètres les plus pertinents pour effectuer la reconnaissance des émotions.

La littérature montre qu'il existe, par exemple, un ensemble varié de support temporel et de paramètres acoustiques pouvant prétendre à la caractérisation des émotions. Alors qu'une très grande majorité des systèmes de l'état-de-l'art exploite un seul type de support temporel (typiquement les segments voisés) pour extraire des mesures de natures différentes (e.g., segmentale / acoustique, suprasegmentale / prosodie), et regroupées dans un unique vecteur de caractéristiques pour effectuer la reconnaissance. Par conséquent, les nombreux efforts qui ont été entrepris pour extraire un large éventail de caractéristiques prosodiques n'ont pas permis d'obtenir un véritable consensus quant à celles qui sont les plus pertinentes. De plus, les travaux qui ont été effectués dans ce domaine se sont focalisés sur l'analyse de caractéristiques

[108] G. Salton et C. Buckley, "Term-weighting approaches in automatic text retrieval", dans *Infor. Proc. and Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[109] P. Gupta et N. Rajput, "Two-stream emotion recognition for call-center monitoring", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2241–2244.

[110] M. Bulut, S. Lee et S. Narayanan, "Analysis of emotional speech prosody in terms of part of speech tags", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 626–629.

[111] B. Schuller, G. Rigoll et M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid Support Vector Machine-Belief Network architecture", dans proc. *ICASSP*, Montreal, Canada, May 17-21 2004, pp. 577–580.

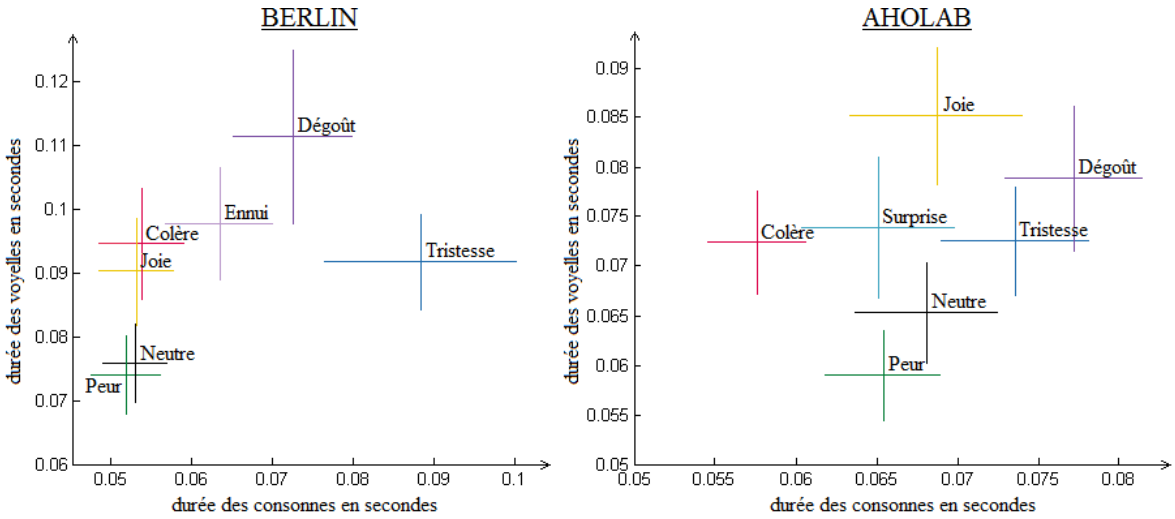


Fig. 1.10 Variations des mesures de durée des voyelles et des consonnes selon les émotions contenus dans les corpus Berlin (figure de gauche) et Aholab (figure de droite). La position de la croix dans l'espace des durées en détermine les valeurs moyennes, tandis que la hauteur et la largeur correspondent aux valeurs d'écart-type ; figure extraite de [RIN08c][112].

du pitch, de l'énergie et de la qualité vocale, alors que le rythme a été exclusivement modélisé par des mesures reposant sur le débit de parole ou sur la durée segmentale. La littérature montre néanmoins que la nature complexe du rythme ne peut être capturée par de telles mesures puisque beaucoup trop réductrices.

Par conséquent, nous proposons dans cette thèse d'exploiter diverses techniques de traitement du signal et de reconnaissance des formes pour définir un système de reconnaissance d'émotions qui se place dans une optique « *bien conçue* » plutôt que « *force brute*» [BAT-99][113]. Notre stratégie repose sur le principe de « diviser pour mieux régner » : nous combinons les informations fournies par des supports temporels et des paramètres complémentaires (e.g., voyelle / consones, acoustique / prosodie) pour caractériser les émotions. Cette approche permet de quantifier la contribution des différents paramètres intervenant dans la caractérisation des états affectifs de la parole au moyen des techniques de fusion d'informations. Ainsi, au lieu d'exploiter des segments définis de façon arbitraire (e.g., toutes les 500ms) ou de façon unique (e.g., segments voisés) pour extraire les caractéristiques du signal de parole, nous préférons exploiter différents points d'ancrages acoustiques complémentaires des informations (e.g., voyelle, consonne, syllabe, « *p-centre* », etc.). De nombreuses études ont en effet montré que la durée des phonèmes est liée aux variabilités affectives de la parole [LEE04][114], [BUL05][115], et [KIS10][116], et que ces variabilités peuvent également être dépendantes de la langue du locuteur [RIN08c][112] et [GOU10][117], cf. Fig. 1.10.

Ensuite, et au-delà du fait que la majorité des systèmes de reconnaissance issus de l'état-de-l'art en TAP orienté émotion ignore le contexte de production lors de l'étape d'extraction

[112] F. Ringeval et M. Chetouani, "A vowel based approach for acted emotion recognition", dans proc. *Interspeech*, Brisbane, Australia, Sep. 22-26 2008, pp. 2763–2766.
[113] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth et H. Niemann, "Prosodic feature evaluation: brute force or well designed?", dans proc. *14th ICPhS*, San Francisco, (CA), USA, Aug. 1999, pp. 2315–2318.
[114] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee et S. Narayanan, "Emotion recognition based on phoneme classes", dans proc. *Interspeech*, Jeju Island, Korea, Oct. 4-8 2004, pp. 205–211.

de caractéristiques, les méthodes proposées regroupent la plupart du temps des mesures de natures différentes dans un unique vecteur de caractéristiques. L'étape de classification des données est donc réalisée de façon incohérente puisqu'elles sont toutes traitées dans un unique cadre décisionnel lors de l'estimation des probabilités *a posteriori*. De nombreuses études ont cependant montré l'intérêt d'une modélisation différenciée des informations issues du cadre segmental, (e.g., les coefficients MFCC) et suprasegmental (i.e., la prosodie) pour le TAP orienté émotion [KIM07][118], [VLA07][119], [MAH09][120] et [SCH09c][121]. Nous avons donc défini dans cette thèse un système de reconnaissance pour chaque type de paramètres (i.e., acoustiques et prosodiques) et une fusion a été opérée entre ces composantes en détaillant l'analyse pour la prosodie, i.e., fusion des caractéristiques du pitch, de l'énergie, de la qualité vocale et du rythme. Notre approche permet donc de détailler l'analyse des corrélats de l'affect à travers plusieurs points d'ancrage des informations dans la parole et selon différents types de paramètres extraits de façon automatique sur le signal.

Enfin, la littérature montre un certain manque de techniques d'extraction de paramètres discriminants des émotions, en particulier pour les composantes du rythme qui sont jusqu'à présent sous-modélisées dans les systèmes. Mais caractériser le rythme de la parole n'est pas une tâche facile en soit, puisque la littérature suggère qu'il n'existe pas un rythme mais plutôt un ensemble de phénomènes rythmiques pouvant être corrélés entre eux. Partant de ce constat, nous proposons de nouvelles méthodes *non-conventionnelles* pour caractériser le rythme de la parole. La pertinence de ces paramètres pour caractériser les corrélats affectifs des émotions actées comme spontanées (i.e., naturelles) est démontrée dans cette thèse.

6. Structure de notre travail

Nous avons décrit dans ce premier chapitre d'*Introduction* le contexte et les enjeux à la fois théoriques et applicatifs de cette thèse. Nos travaux rejoignent alors celui du domaine du TAP orienté émotion et celui du SSP, cf. sous-section 1.1.5. Nous avons ensuite présenté les définitions et fonctionnalités associées à la prosodie, cf. sous-section 2.1 et 2.2. Cette dernière dessert alors toutes les fonctions du discours et permet de communiquer un large ensemble

[115] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. Lee, S. Lee, et S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 801–804.

[116] G. Kiss et J. van Santen, "Automated vocal emotion recognition using phoneme class specific features", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010 (pas de pagination).

[117] M. Goudbeek et M. Broersma, "Language specific effects of emotion on phoneme duration", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010.

[118] S. Kim, P. G. Georgiou, S. Lee, et S. Narayanan, "Real-time emotion detection system using speech: multimodal fusion of different timescale features", dans proc. *9th W. on MMSP*, Chania, Crete, Greece, Oct. 1-3 2007, pp. 48–51.

[119] B. Vlasenko, B. Schuller, A. Wendemuth et G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing", dans proc. *2nd Inter. C. on Affective Comp. and Intel. Interaction*, Lisbon, Portugal, pp. 139–147, Sep. 12-14 2007.

[120] A. Mahdhaoui, F. Ringeval et M. Chetouani, "Emotional speech characterization based on multi-features fusion for face-to-face interaction", dans proc. *3rd Inter. C. on Signals, Circuits and Systems*, Djerba, Tunisia, Nov. 6-8 2009, pp. 1–6.

[121] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll et A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances", dans proc. *W. on ASRU*, Merano, Italy, Dec. 13-17 2009, pp. 552–557.

d'informations dont le processus d'encodage dans la parole est particulièrement complexe, cf. sous-section 2.3. Nous avons décrit dans la sous-section 3.1 le terme *émotion* avec d'autres termes de ce contexte. Différentes théories proposées par les études en psychologie ont également été présentées, cf. sous-section 3.2, ainsi que des modèles de représentation des émotions, cf. sous-section 3.3. Les processus d'encodage et de décodage des informations affectives dans la parole, ainsi que les méthodes de constitution de corpus selon différents degrés de contrôle dans le processus d'induction des émotions ont ensuite été décrits. Nous avons enfin présenté un état-de-l'art dans le domaine du TAP orienté émotion, en spécifiant la description des systèmes à ceux qui ont exploité des informations linguistiques.

Le chapitre 2, *Ancrages acoustiques de la parole*, présente les différents types de support temporel sur lesquels peuvent être ancrées les informations affectives. L'un des défis du TAP orienté émotion consiste à identifier automatiquement et de façon robuste les points d'ancrages des informations affectives qui sont encodées dans la parole. Cette étape permet de faciliter par la suite l'extraction des caractéristiques. Nous présentons dans ce chapitre des méthodes permettant de localiser automatiquement des points d'ancrage de nature pseudo-phonétique (e.g., pseudo-voyelle et pseudo-consonne) et rythmique (e.g., « *p-centre* ») dans un signal de parole. Ces méthodes sont évaluées sur plusieurs corpus par le calcul des taux d'erreurs en détection de voyelles. Nous estimons aussi les corrélats des « *p-centres* » selon les autres points d'ancrage.

Le chapitre 3, *Reconnaissance acoustique de la parole affective actée*, exploite un système de reconnaissance pour étudier la pertinence de l'emploi de plusieurs points d'ancrage complémentaires de la parole pour l'étape d'extraction des caractéristiques acoustiques (e.g., MFCC). L'objectif consiste à évaluer l'influence de divers contextes de production du signal de parole sur les scores obtenus par le système de reconnaissance via les coefficients MFCC et à travers différentes configurations de tests. Des techniques de normalisation des données (e.g., *z-score*) à travers des informations telles que le genre, le locuteur et la phrase ont notamment été employées, et nous avons exploité plusieurs méthodes d'exploration de données pour effectuer les expériences en reconnaissance : (i) tests d'indépendance aux classes (CVS) et (ii) au locuteur (LOSO).

Le chapitre 4, *Reconnaissance prosodique de la parole affective actée*, introduit tout d'abord un ensemble de phénomènes associés au rythme qui justifient le besoin de méthodes de modélisations *non-conventionnelles*. Nous justifions aussi pourquoi le rythme pourrait être caractérisé par des mesures de la dynamique dans les autres composantes de la prosodie (i.e., pitch, énergie et qualité vocale), et dans quel sens ces paramètres peuvent être reliés aux émotions. Comme les paramètres de la prosodie se répartissent selon quatre composantes distinctes (i.e., pitch, énergie, qualité vocale et rythme), nous avons cherché à quantifier leur contribution dans la tâche de caractérisation des corrélats de l'affect. Enfin, nous étudions les variations des paramètres du rythme selon modèles *conventionnels* et *non-conventionnels* et à travers les catégories d'émotions. La pertinence des métriques du rythme pour la différenciation des émotions actées est ainsi démontrée sur de multiples aspects dans ce chapitre.

Le chapitre 5, *Emotions et troubles de la communication*, propose une application des techniques exploitées dans les chapitre précédents sur de la parole affective produite de façon naturelle. Nous avons aussi profité de collaborations étroites avec plusieurs équipes de clini-

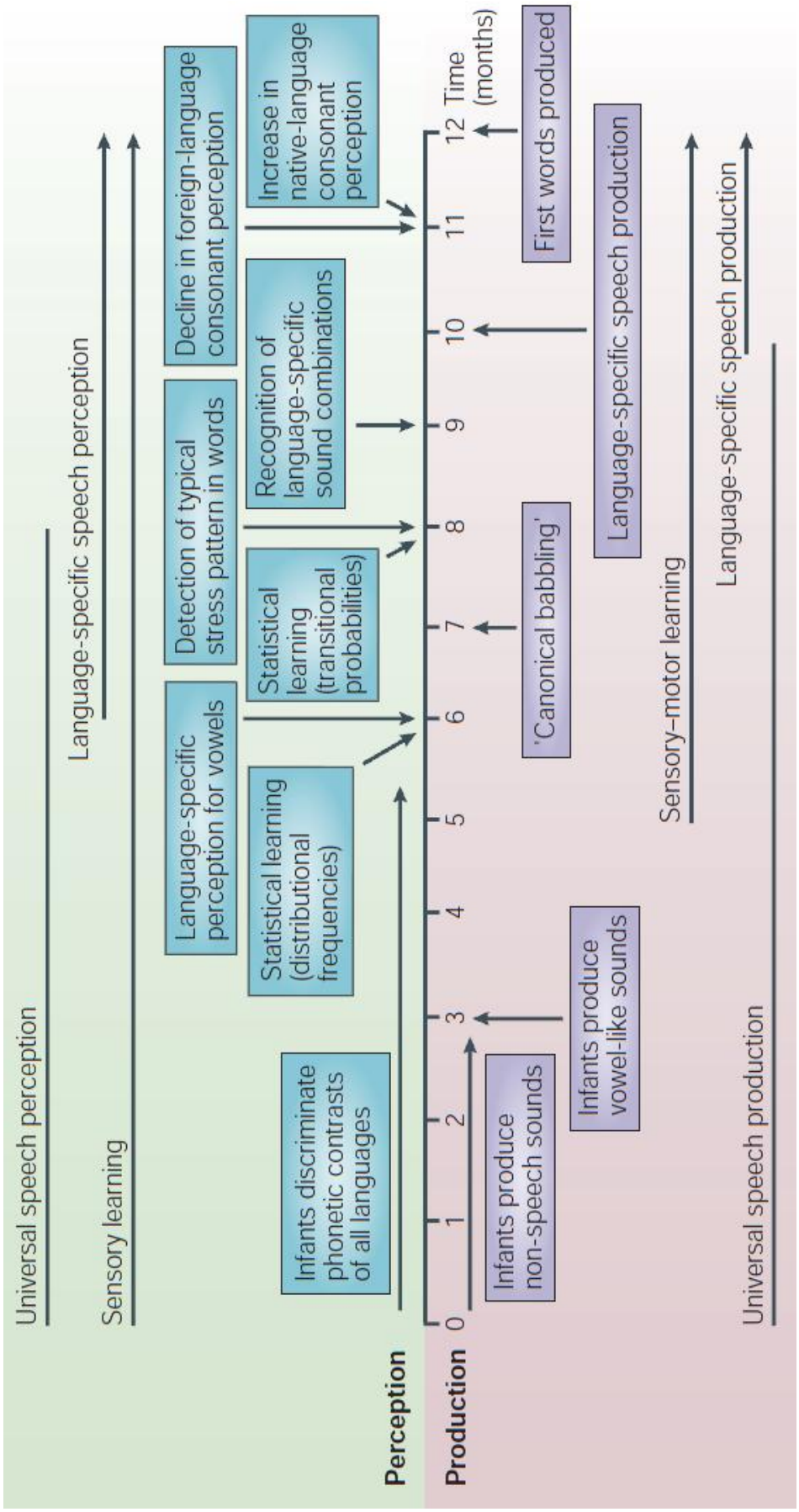
ciens pour associer l'étude des émotions avec celle des troubles de la communication (TC). Nous présentons donc dans l'introduction de ce chapitre les caractéristiques cliniques des TC que nous avons étudiés. Le protocole de collecte des données est ensuite présenté avec les épreuves. Ces dernières permettent de tester les sujets sur une large gamme de fonctionnalités prosodiques : (i) *imitation de l'intonation* (tâche contrainte) et (ii) *production de parole affective spontanée* (tâche non contrainte). Les expériences ont consisté à étudier la pertinence de l'usage des techniques de TAP orienté émotion pour caractériser les troubles de la prosodie présents chez des sujets TC. La prosodie de ces enfants a ainsi été comparée à celle de sujets à développement typique, et entre les sujets pathologiques en exploitant les points d'ancrages acoustiques du chapitre 2, et les paramètres prosodiques du chapitre 4. La pertinence des modèles du rythme pour caractériser les corrélats de l'affect encodés de façon spontanée par les enfants dans la parole est également démontrée dans ces expériences.

La thèse se termine par le chapitre 6 qui résume les différents aspects de la caractérisation des émotions par la parole actée et spontanée ainsi que les principaux résultats expérimentaux qui ont été obtenus dans cette thèse. Des perspectives sont ensuite proposées.

Chapitre 2

Ancrages acoustiques de la parole

Afin de caractériser de façon automatique les interactions sociales de l'Homme, il est nécessaire d'identifier au préalable les supports sur lesquels les informations sont encodées. Ces supports seront qualifiés d'ancrages car ils contiennent les caractéristiques permettant d'accéder aux informations. L'identification automatique des points d'ancrage des informations dans la parole n'est pas une tâche aisée puisque leur processus d'encodage est particulièrement complexe (cf. chapitre 1, sous-section 2.3). Les contraintes s'opposant à cette tâche sont alors liées aux nombreuses sources de variabilités du signal de parole qui rendent difficile la séparation des causes. La morphologie et l'étendue des modes d'utilisation du système phonatoire de l'Homme dépendent en effet d'un ensemble de facteurs difficilement maîtrisables, e.g., âge, genre, style, origine socio-culturelle, langue, attitude, état physique et affectif du locuteur. Ces facteurs introduisent des variabilités dans le signal de parole qui complexifient la tâche d'identification des points d'ancrage des informations véhiculées par ce dernier. L'un des défis du TAP est d'identifier automatiquement et de façon robuste ces points d'ancrages de la parole, puisque ils permettent d'extraire les caractéristiques servant à identifier les informations.



Ligne temporelle universelle représentant le développement du processus de production et de perception de la parole chez l'enfant de la naissance jusqu'au 12ème mois [KUH04]. Cette figure permet d'illustrer l'évolution du degré de complexité des différents types d'ancrages acoustiques exploités par l'enfant durant le processus de développement du langage. L'enfant commence ainsi par discriminer les contrastes phonétiques des langues durant les tous premiers mois de sa vie. La perception des voyelles spécifiques à la langue native apparaît ensuite à partir du 6ème mois. L'apprentissage des distributions et des transitions phonétiques propres à la langue commence dès lors. L'enfant arrive ainsi, dans le cadre d'un développement typique, à détecter des motifs accentuels dans les mots à partir du 8ème mois. Vient ensuite une spécialisation dans le développement des fonctionnalités langagières spécifiques à la langue native de l'enfant.

1. Introduction

Les sources de variabilité des ancrages acoustiques de la parole sont principalement liées à la complexité du système phonatoire de l'Homme. Le processus de phonation fait en effet appel à un ensemble d'organes dit « de la parole » dans lequel les caractéristiques physiques mises en jeu sont nombreuses : intensité du geste d'expulsion de l'air des poumons, mode de vibration des cordes vocales, vitesse, force et précision des gestes de la langue et des lèvres. La complexité du système phonatoire entraîne ainsi l'existence de marqueurs acoustiques des informations très variés. Toutefois, la diversité du langage liée à l'évolution de l'Homme a fait apparaître certaines spécificités sur les types de sons produits et leur agencement temporel dans la parole. La typologie rythmique des langues distingue par exemple les langues *syllabiques* (e.g., langues latines), des langues *accentuelles* (e.g., allemand, anglais, arabe et russe) et *moriques* (e.g., japonais, hongrois et tamoul). La production de la parole est ainsi supposée s'organiser sur la répétition d'unités semblables, comme la syllabe, le pied ou la more, chaque groupe de langue utilisant un seul type d'unité préférentiel.

Les sources de variabilités naturellement présentes dans le signal de parole complexifient la tâche d'identification des ancrages acoustiques. Cette tâche se doit cependant d'être robuste pour s'assurer que le contexte dans lequel les caractéristiques sont observées reste homogène à travers le support d'extraction utilisé. Cette propriété assimilable au principe du « diviser pour mieux régner » permet de rendre plus cohérente la tâche de reconnaissance. En effet, l'influence des sources de variabilité du signal de parole sur ses structures acoustiques est mieux maîtrisée si l'on exploite différents points d'observation. De plus, les performances en reconnaissance peuvent être améliorées en fusionnant les probabilités fournies par le système sur différents points d'ancrage complémentaires de la parole (e.g., voyelles et consonne).

2. Niveaux d'actualisation de la parole

Les niveaux d'actualisation des informations de la parole peuvent se différencier selon que l'on se place du côté de la production ou de la perception. Pour la phonation, les échelles sont de nature linguistique et définies par des caractéristiques d'ordre articulatoire (contexte segmental). Pour l'aspect perceptuel, les informations sont suprasegmentales : le pitch délimite par exemple les segments voisés et le rythme permet de régulariser l'impression perceptuelle laissée par la parole chez un locuteur. La Fig 2.1 illustre la diversité des informations pouvant être extraites sur le signal de parole via ses différents points d'ancrage.

2.1. Echelles linguistiques

Les échelles linguistiques sont hiérarchisées de la façon suivante : un message se décompose en phrases puis en mots. Chaque mot peut être segmenté en syllabe(s) et chaque syllabe est constituée de phonème(s). Les mores sont définies par le poids associé à la rime des syllabes qui peut être légère / lourde / extra-lourde selon le degré de complexité de la partie consonantique (coda). Le pied est une unité prosodique qui consiste en un groupement de syllabes ou de mores alternant alors un temps fort et un temps faible.

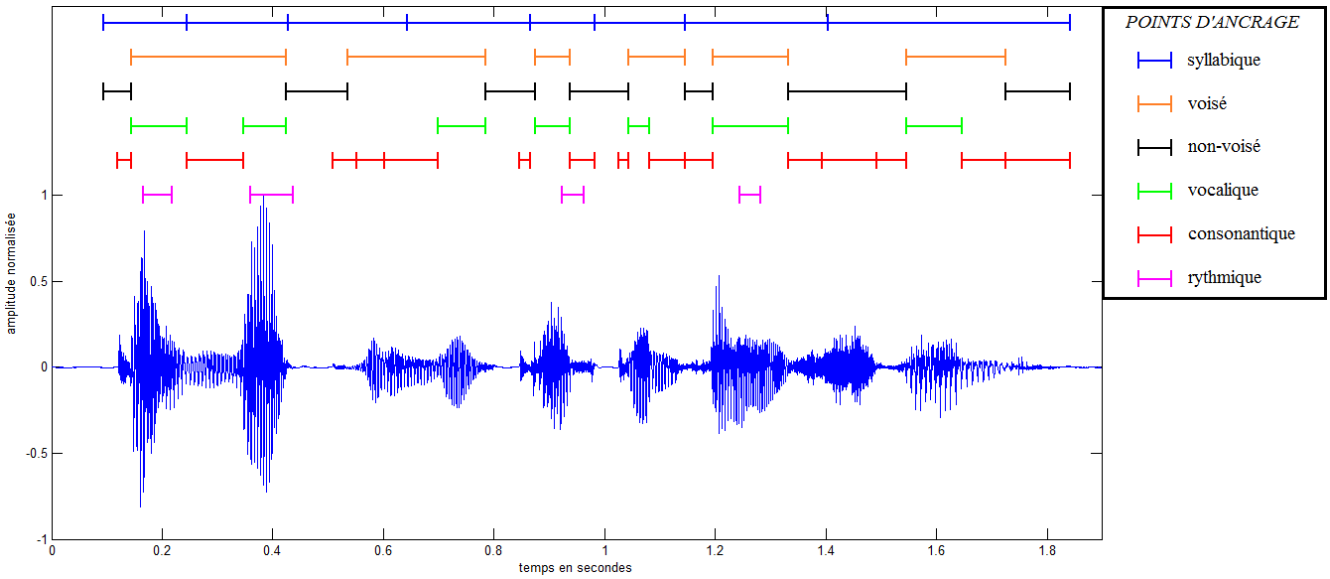


Fig. 2.1 Ancrages acoustiques et rythmiques de la parole illustrant la diversité des informations pouvant être extraites sur le signal. L'ancrage rythmique de la parole est obtenu par les « *p-centres* » identifiés au niveau 1 (cf. sous-section 3.3).

2.1.1 Les phonèmes

Le phonème est issu de la phonologie et correspond à la plus petite unité que l'on peut isoler dans le flux du discours. Les phonèmes sont identifiés selon le principe de la « paire minimale » : un phonème doit pouvoir opposer deux mots de significations différentes par sa seule différence (i.e. '**v**ie' - /v/ /i/ et '**p**ie' - /p/ /i/). Les phonèmes sont donc dépendants de la langue puisque leur distinction s'effectue selon le sens des mots. Cependant, un phonème peut correspondre à différents types de réalisations acoustiques alors appelées allophone.

L'Alphabet Phonétique International (API) recense les phonèmes des différentes langues. Il définit des classes ou macro-classes phonétiques pour lesquelles le regroupement s'effectue par des traits distinctifs d'ordre articulatoire, acoustique ou perceptif. Les voyelles sont par exemple catégorisées par leur degré d'aperture (e.g., fermée, moyenne ou ouverte), leur point d'articulation (e.g., antérieur, central ou postérieur) et par leur caractère arrondi ou non. Le degré d'aperture et le lieu d'articulation des voyelles sont observables sur le signal de parole via les deux premiers formants F_1 et F_2. Les formants correspondent en effet aux corrélats physiques des déviations réalisées par la partie supraglottique de l'appareil vocal sur l'onde acoustique pure (i.e., issue des cordes vocales, formants = harmoniques + déviations). Dans le plan formantique $F_2 - F_1$, les voyelles occupent un espace caractéristique alors qualifié de « triangle vocalique », cf. Fig. 2.2. Les consonnes se caractérisent quant à elles par le mode (e.g., nasale, occlusive, fricative, etc.) et le lieu d'articulation (e.g., labial, coronal, dorsal, laryngal, etc.), et par leur caractère sourd ou sonore selon le degré de voisement.

L'alphabet phonétique SAMPA[1] (*Speech Assessment Methods Phonetic Alphabet*) est très souvent utilisé pour transcrire phonétiquement des corpus de parole. Cet alphabet présente en effet la particularité de reposer sur des équivalents ASCII[2] des symboles de l'API ; ce qui facilite son utilisation. L'alphabet SAMPA permet d'illustrer les diversités phonétiques des

[1] L'alphabet SAMPA est accessible à l'adresse Internet suivante : http://www.phon.ucl.ac.uk/home/sampa/.

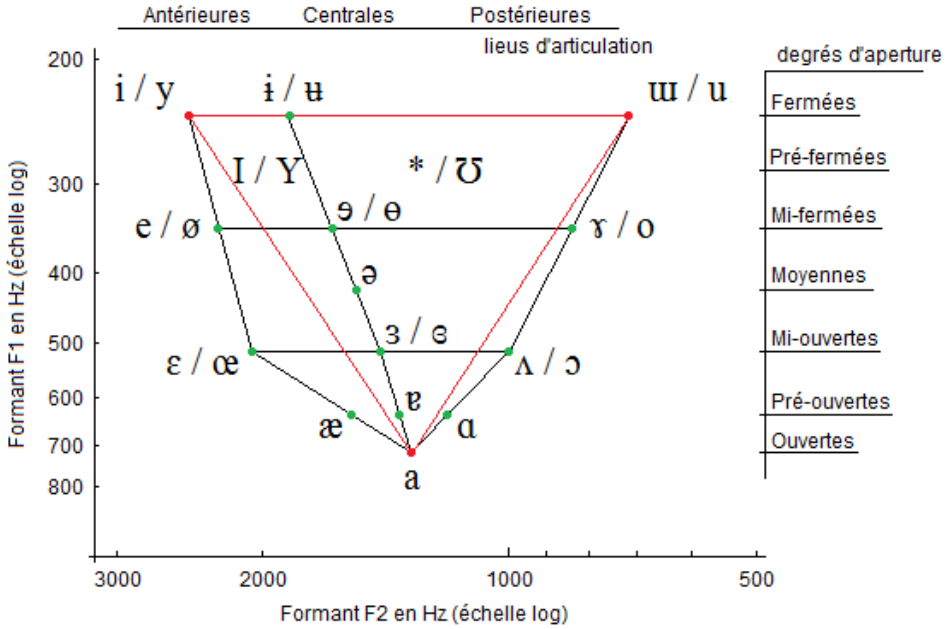


Fig. 2.2 Triangle vocalique (en rouge) de l'appareil vocal. Les phonèmes apparaissant par pair correspondent respectivement à leur version non-arrondie / arrondie. Les valeurs des fréquences sont données à titre indicatif et peuvent varier selon les caractéristiques du locuteur (e.g., genre, âge, idiosyncrasies, ...).

langues qui y sont représentées, cf. Fig. 2.3. Le Danois et l'Allemand apparaissent ainsi comme les langues étant phonétiquement les plus riches, le Croate et l'Espagnol comme étant les langues les moins pourvues en phonèmes. Des différences significatives peuvent également se manifester entre les dialectes d'une même langue. Grabe *et al.* ont ainsi montré que les variabilités rythmiques des dialectes de l'Anglais britannique pouvaient être au moins aussi importantes que celles trouvées sur différents groupes de langues [GRA00][3].

2.1.2 Les syllabes

La syllabe est une unité de la phonétique qui est composée d'un ensemble de voyelles et de consonnes se prononçant d'une seule émission de voix. La syllabe est considérée comme l'unité insécable du discours. Sa structure se compose d'une attaque et d'une rime. La rime est constituée d'un noyau (ou sommet) et d'une coda. Attaque et coda sont bien souvent facultatives et peuvent être formées d'une ou plusieurs consonnes. Le noyau est par contre obligatoirement composé d'une voyelle[4]. La structure prototypique d'une syllabe est la suivante : Consonne-Voyelle-Consonne, cf. Fig. 2.4.

Ce modèle de la syllabe est supposé universel [LAB05][5]. L'unité accentuable du discours est par définition la syllabe, ou bien une unité plus petite définie en référence à elle [GAR68][6]. Sur le plan acoustique, on considère traditionnellement qu'il y a autant de syllabes que de pics de sonorité dans une séquence donnée. Plusieurs versions de l'échelle de sonorité existent

[2] Abréviation désignant la norme d'encodage informatique des caractères alphanumériques de l'alphabet latin (*American Standard Code for Information Interchange*).

[3] E. Grabe, B. Post, F. Nolan et K. Farrar, "Pitch accent realization in four varieties of British English", dans *J. of Phonetic*, vol. 28, no. 2, pp. 161–185, Jun. 2000.

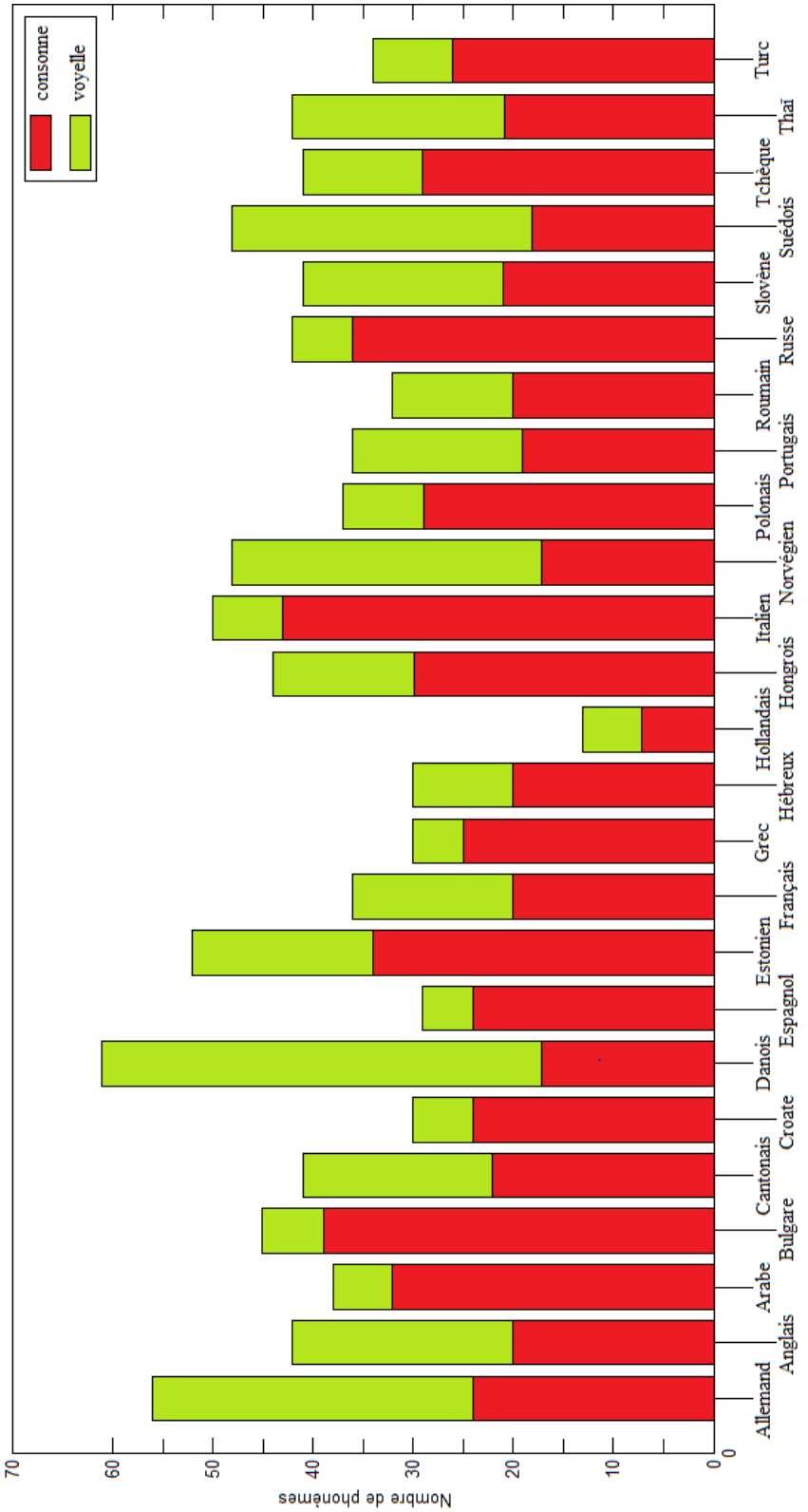


Fig. 2.3 Caractéristiques phonologiques des langues : nombre de phonèmes et répartitions en macro-classes phonétiques : voyelle / consonne.

[4] Des singularités s'opposent à l'universalité de cette définition. L'absence de noyau vocalique dans le dialecte imdlawn tashlhiyt du berbère [DEL85] et le bella coola (langue amérindienne du nord) [BAG91] en sont des exemples flagrants. Le noyau vocalique des structures syllabiques peut également provenir de structures consonantiques propices à un ancrage vocalique : l'interjection « psss » en Français, le mot « bottle » en Anglais, « haben » en Allemand, etc. Enfin, certains linguistes supposent que tout constituant syllabique peut être nul [KAY84].

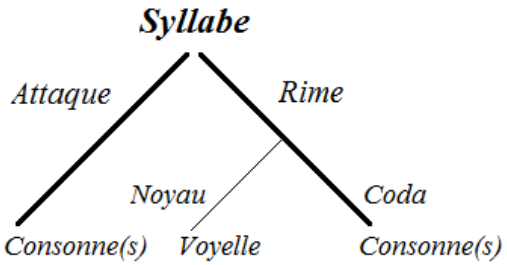


Fig. 2.4 Structure d'une syllabe prototypique : attaque (C) et rime (noyau – V et coda – C) ; figure extraite de [LAB05][5].

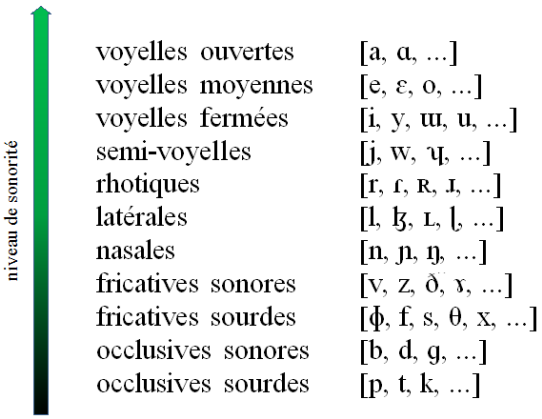


Fig. 2.5 L'échelle de sonorité ou d'audibilité phonétique ; figure extraite de [LAB05][5].

[ANG97][7]. L'échelle présentée par [LAB05][5] montre que le niveau sonore perçu croît au fur et à mesure que l'on s'approche du noyau de la syllabe prototypique, cf. Fig. 2.5. Notons que cette échelle reste valable pour les langues exemptes de noyaux vocaliques[4].

2.2. Echelles perceptuelles

Nous décrivons dans les sections suivantes les propriétés perceptuelles qui permettent de définir les segments voisés et le centre de perception de la parole.

2.2.1. Les segments voisés

Les segments voisés sont liés à la perception d'une structure spectrale particulière dans le flux de la parole. Cette structure acoustique provient de l'oscillation périodique des cordes vocales produite lors de la phonation des segments vocaliques. Par conséquent, les segments voisés se délimitent par les discontinuités du pitch et se situent entre le phonème et la syllabe, cf. Fig. 2.1. En raison de phénomènes coarticulatoires, le trait de voisement peut s'étendre sur plusieurs phonèmes. Comme toutes les voyelles sont par définition voisées, un segment voisé

[5] L. Labrune, "Autour de la syllabe : les constituants prosodiques mineurs en phonologie", dans *Phonétique et phonologie, approches contemporaines*, N. Nguyen, S. Wauquiers et J. Durand [Eds], Hermès, pp. 95–116, 2005.

[6] P. Garde, *L'accent*, dans Presses Universitaires de France [Eds], Paris, 1968.

[7] J.-P. Angoujard, *Théorie de la syllabe, rythme et qualité*, dans CNRS [Eds], Paris, 1997.

inclut donc le noyau vocalique des syllabes et une partie de l'attaque et/ou de la coda selon leurs propriétés de voisement. Son identification dans le flux de la parole est relativement aisée puisque seule la connaissance du pitch suffit, cf. Fig. 2.6. Toutefois, l'extraction de la fréquence fondamentale peut s'avérer problématique sur certains types de signaux de parole, e.g., parole bruitée, voix d'enfants et troubles de la parole, cf. Fig 5.7.

La relative facilité avec laquelle les segments voisés peuvent être extraits dans un signal de parole les ont rendu très populaires dans la communauté du TAP. Ils sont ainsi considérés comme le support d'extraction privilégié des caractéristiques acoustiques et prosodiques. Cependant, il existe de nombreux autres supports temporels de la parole dont leur pertinence pour la reconnaissance des émotions a été démontrée, tant sur le plan acoustique [LEE04][8] que prosodique [BRO07][9]. Les segments voisés ne sont donc pas l'unique support adéquat pour extraire les caractéristiques de l'affect. De plus, ces segments ne prennent pas en compte le contexte de production puisqu'ils incluent à la fois les informations issues du noyau syllabique et des consonnes voisées situées à proximité (attaque et/ou coda, selon le cas de figure). Bien que liés, ces segments sont de nature très différentes : le noyau vocalique a pour objectif d'apporter une proéminence acoustique (les voyelles sont situées au sommet de l'échelle de sonorité) alors que l'attaque et la coda apportent une information sur comment cette proéminence prend forme, et se termine, respectivement.

2.2.2. Le centre de perception

Des expériences ont montré que les locuteurs et les auditeurs n'utilisent pas l'attaque des syllabes comme support d'organisation temporelle principale de la parole. Au contraire, ils semblent se focaliser sur un point situé à l'intérieur de la syllabe et qui est perçu comme étant le « moment d'occurrence » de la syllabe [MAR81][10], [SCO93][11]. Cet instant a été qualifié de « *p-centre* » par des psychologues. Le phénomène du « *p-centre* » s'observe par exemple lorsque l'on demande à des individus de produire des séquences de syllabes alternées et isochrones, e.g. « *bad-sad-bad-sad…* » produites au rythme d'un métronome [BAR05][12]. Les locuteurs introduiront alors invariablement des déviations systématiques sur l'isochronie syllabique afin d'accomplir une certaine régularité perceptuelle [FOW79][13]. Dans l'exemple mentionné, la syllabe « *sad* » serait produite plus tôt en regard d'un espacement régulier.

---

[8] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee et S. Narayanan, "Emotion recognition based on phoneme classes", dans proc. *Interspeech*, Jeju Island, Korea, Oct. 4-8 2004, pp. 205–211.

[9] L. Bronakowski, K. Slot, J. Cichosz et J. Kim, "Application of Poincare map-based description of vowel pronunciation variability for emotion assessment in speech signal", dans *Int. Symp. on Inf. and Tech. Conv.*, Jeonju, Korea, Nov. 23-24 2007, pp. 175–178.

[10] S. Marcus, "Acoustic determinants of perceptual center (P-center) location", dans *Perception and Psychophysics*, vol. 30, pp. 247–256, Sep. 1981.

[11] S. K. Scott, *P-centers in speech: an acoustic analysis*, Thèse de doctorat, University College London, 1993.

[12] P. A. Barbosa, P. Arantes, A. Meireles et J. M. Vieira, "Abstractness in speech-metronome synchronisation: p-centres as cyclic attractors", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 1441–1444.

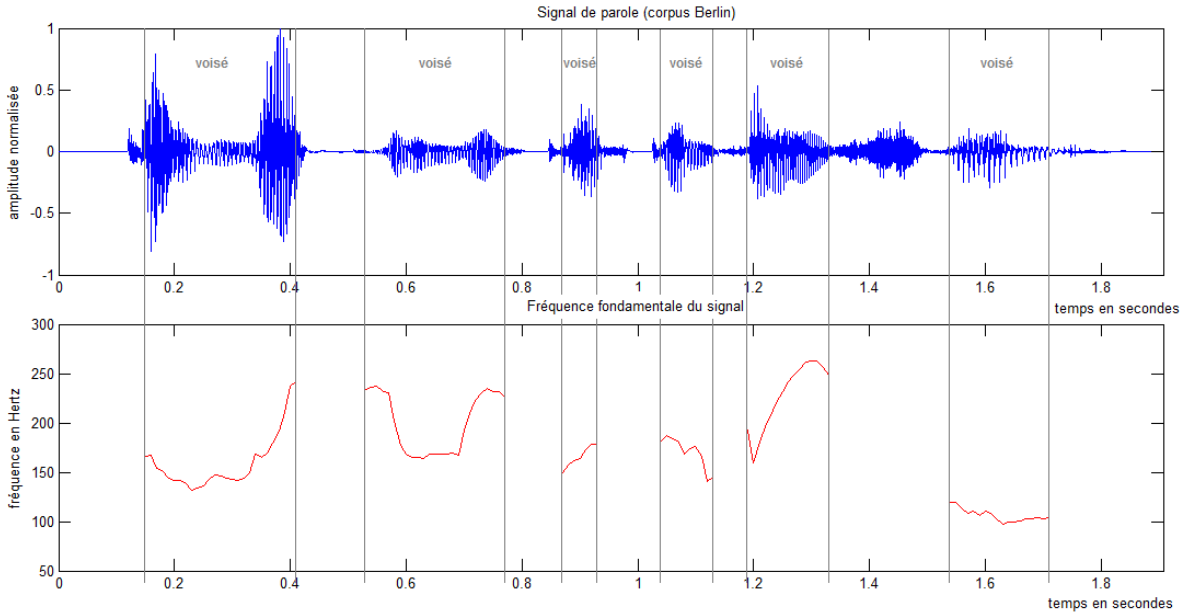


Fig. 2.6 Segmentation d'un signal de parole en segments voisés ; figure du haut : signal acoustique de parole ; figure du bas : fréquence fondamentale.

3. Identification automatique d'ancrages acoustiques

Contrairement aux études menées par les sciences linguistiques, où les unités utilisées peuvent fortement varier selon l'objet de l'étude, les sciences computationnelles utilisent des unités essentiellement basées sur la segmentation de la f0 (segments voisés) et de l'énergie (segments non-voisés). Ce choix s'explique principalement par des raisons d'ordre technique : même s'il existe de nombreux systèmes de reconnaissance de la parole, il est plus difficile de transcrire automatiquement et de façon robuste le signal acoustique en unités linguistiques (e.g., phonèmes, syllabes, mots), que d'en exploiter ses propriétés pour localiser certains types de constituant (e.g., segment voisés et non-voisés). En outre, les méthodes de reconnaissance automatique de la parole ne sont pas infaillibles, et les erreurs commises sont d'autant plus fréquentes lorsque la parole analysée s'écarte des chemins traditionnels tels que, par exemple, les contextes affectifs [ROT05][14], [SCH06a][15] et [STE10][16], et/ou pathologiques [GRI00][17].

Les propriétés acoustiques du signal de parole permettent d'accéder aux informations d'ordre macro-phonétique (e.g., voyelle et consonne). Ces points d'ancrage de la parole sont relativement simples à identifier puisque les macro-classes phonétiques présentent des carac-

[13] C. Fowler, "Perceptual centers", dans *Speech Production and Perception, Perception and Psychophysics*, vol. 25, pp. 375–388, 1979.

[14] M. Rotaru, D. J. Litman et K. Forbes-Riley, "Interactions between speech recognition problems and user emotions", dans proc. *Interspeech*, Lilsbon, Portugal, Sep. 4-8 2005, pp. 1–4.

[15] B. Schuller, J. Stadermann et G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation", dans *Speech Prosody*, Dresden, Germany, May 2-5 2006, paper 169.

[16] S. Steidl, A. Batliner, D. Seppi et B. Schuller, "On the impact of children's emotional speech on acoustic and language models", dans *EURASIP J. on Audio, Speech, and Music Proc.*, vol. 2010, article ID 783954, 2010.

[17] S. Griffin, L. Wilson, L. Clark, et S. McLeod, "Speech pathology applications of automatic speech recognition technology", dans L. Wilson & S. Hewat [Eds], proc. *8th Australian Inter. Conf. on Speech Sci. and Tech.*, Melbourne, Australia, Dec. 5-7 2000, pp. 362–366.

téristiques acoustiques de natures très différentes. Nous détaillons dans les paragraphes suivants un algorithme de segmentation du signal de parole en pseudo-phonèmes [OBR88][18] ainsi qu'une méthode permettant d'identifier les voyelles de par ces segments [PEL98][19].

3.1. Les pseudo-phonèmes

Les pseudo-phonèmes sont des unités de la parole qui ont été introduites par les sciences computationnelles. Ils reposent sur la notion de stationnarité du signal de parole (algorithme *divergence forward backward* – DFB [OBR88]). L'identification des pseudo-phonèmes est donc pertinente pour les voyelles puisque l'onde acoustique observée est stationnaire pour une durée largement supérieure à une trentaine de ms. Cette caractéristique est cependant moins valable pour les consonnes où la durée est parfois inférieure au critère de stationnarité (30ms). De plus, les formes acoustiques observées sur ces dernières sont bien souvent non-linéaires, ce qui peut avoir pour conséquence de rendre le signal de parole non-stationnaire[20]. La méthode DFB a été comparée avec de nombreuses autres méthodes de segmentation du signal de parole [OBR93][21]. Des expériences ont montré que la durée des segments délimités par le DFB est porteuse d'une information pertinente [OBR97][22].

La segmentation DFB est réalisée, via un modèle de type prédictif linéaire (LPC), par un seuillage de la divergence entropique entre deux fenêtres d'analyse. La méthode distingue ainsi trois types de segments :

- les segments quasi-stationnaires caractérisant la partie stable des phonèmes, en particulier la partie centrale des voyelles,
- les segments transitoires entre deux phonèmes,
- et les segments courts (durée inférieure à 20ms) appelés segments évènementiels liés à de brefs gestes articulatoires pouvant se chevaucher.

L'algorithme DFB émet l'hypothèse que le signal de parole est décrit par une suite de zones quasi-stationnaires caractérisées chacune par un modèle autorégressif gaussien (AR) [1]. Le principe de segmentation consiste à comparer les erreurs de prédiction e_n commises sur deux fenêtres d'analyse. La première fenêtre M_0 est de longueur fixe, alors que la seconde fenêtre M_1 est de longueur variable, cf. Fig. 2.7. La distance entre les erreurs de prédiction e_n est calculée par la divergence de Kullback-Leibler. Cette divergence correspond à la mesure

[18] R. André-Obrecht, "A new statistical approach for automatic speech segmentation", dans *IEEE Transaction on Acoustic Speech and Signal Processing*, vol. 36, no. 1, pp. 29–40, Jan. 1988.

[19] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques*, Thèse de doctorat, Université Paul Sabatier, Toulouse, 1998.

[20] Les phénomènes de non-linéarités constituent une raison nécessaire mais non suffisante pour induire de la non-stationnarité dans le signal de parole [CHE09a].

[21] R. André-Obrecht, *Segmentation et parole ?*, Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, 1993.

[22] R. André-Obrecht et B. Jacob, "Direct identification vs. correlated models to process acoustic and articulatory information in automatic speech recognition", dans proc. *ICASSP*, Munich, Germany, Apr. 21-24 1997, pp. 989–992.

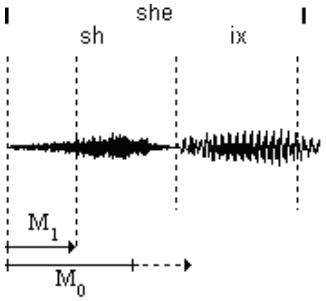


Fig. 2.7 Fenêtres d'analyse pour la segmentation DFB. Les traits en pointillés correspondent aux frontières détectées par l'algorithme.

d'entropie mutuelle w_k calculée entre deux lois de probabilité gaussiennes [1]. La mesure w_k est estimée pour chaque indice k correspondant à une incrémentation de la taille de la fenêtre M_1 [2]. La segmentation est ensuite réalisée par la détection d'un changement de pente dans la statistique W_n définie comme la somme cumulée des N estimations de l'entropie w_k [3].

$$\text{Modèle AR Gaussien} \quad \begin{cases} y_n = \sum a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases} \qquad [1]$$

avec, y_n le signal de parole, e_n un bruit blanc gaussien de variance σ_n^2 et a_i le vecteur de coefficients du modèle AR.

$$w_k = \frac{1}{2}\left[2\frac{e_0^k e_1^k}{\sigma_1^2} - \left(1 + \frac{\sigma_0^2}{\sigma_1^2}\right)\frac{e_0^k}{\sigma_0^2} + \left(1 - \frac{\sigma_0^2}{\sigma_1^2}\right)\right] \qquad [2]$$

$$\text{avec,} \qquad e_n^j = y_n - \sum a_i^j y_{n-i} \mid j = 0{,}1$$

$$W_n = \sum_{k=1}^{N} w_k \qquad [3]$$

Cette étape est réalisée dans les deux sens temporels, i.e., *forward* et *backward*, cf. Fig. 2.8. La phase de détection conduit à localiser des frontières acoustiques, qui sont liées à la divergence de Kullback-Leibler des modèles LPC du signal de parole. Comme ces frontières s'avèrent être proches de celles des phonèmes, les segments de parole identifiés par le DFB ont donc été qualifiés de pseudo-phonèmes. Les pseudo-phonèmes peuvent être catégorisés en macro-classes phonétiques (e.g., voyelle / consonne), cf. Fig. 2.9. Pour cela, le DFB localise tout d'abord les frontières acoustiques du signal de parole et les segments obtenus sont ensuite filtrés par un détecteur d'activité vocale. Dans le même temps, un deuxième système localise la présence des noyaux vocaliques dans le signal de parole, cf. étage du bas dans la Fig. 2.9. Les noyaux vocaliques, qui sont alors détectés, sont ensuite fusionnés avec les segments issus du détecteur d'activité vocale pour identifier les ancrages de type « *pseudo-voyelle* »[23]. Les segments issus du détecteur d'activité vocale et qui ne présentent pas de noyau vocalique sont alors considérés comme des ancrages de type « *pseudo-consonne* »[24].

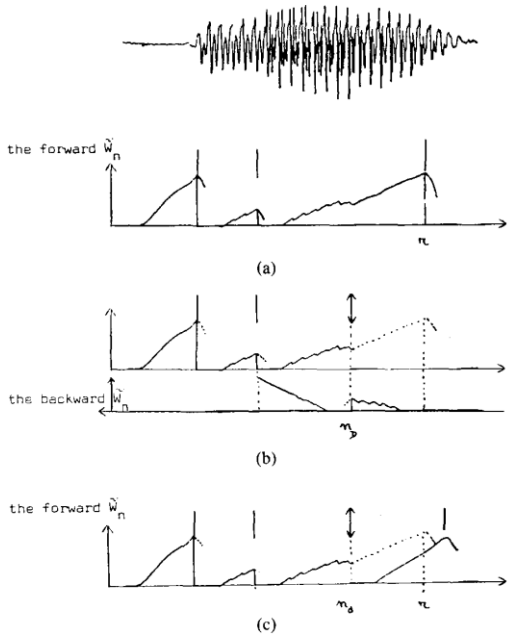


Fig. 2.8 Segmentation automatique d'un signal de parole par le DFB : après avoir détecté un nouveau saut à l'instant *r* (a) et si $r > L$ (L=2 dans cet exemple), le signal est traité dans le sens *backward* (b) ; si un nouveau saut est détecté à l'instant n_d le signal est traité depuis cet instant dans le sens *forward* (c) ; sinon, le traitement recommence depuis l'instant *r* ; figure extraite de [OBR88][18].

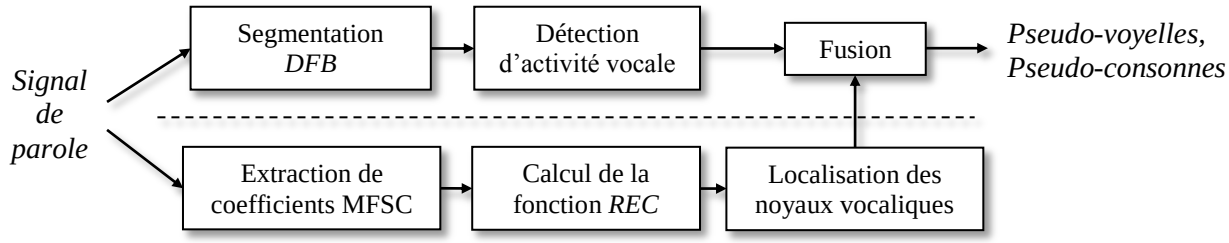


Fig. 2.9 Système de détection de pseudo-phonèmes dans un signal de parole.

Si l'on note N_s le nombre de segments issus de la segmentation DFB et $\{S_1, S_2, \dots, S_{N_s}\}$ la suite de ces segments, l'activité vocale[25] peut être définie par un seuil σ_a sur la variance des segments S_i [4]. Les segments dont la variance est inférieure à σ_a peuvent être vus comme du silence. La valeur de la constante α vaut typiquement 4. Afin de limiter d'éventuels effets de bord liés à des phénomènes transitoires (e.g., oscillation d'amortissement) ou à des décalages entre les frontières détectées et celles issues du signal d'origine, la variance est calculée sur une portion centrée S_i' des segments S_i, cf. Fig. 2.10.

Les voyelles sont identifiées par les proéminences de la dérive spectrale. Cette mesure est obtenue par la fonction « *reduced energy cumulating* » – REC [PEL98][19]. La fonction REC a

[23] Certains segments voisés peuvent présenter des pics dans la dérive spectrale qui sont considérés, à tort, comme étant des noyaux vocaliques ; cela est notamment le cas pour les semi-voyelles /l/ et les sonorantes /n/.

[24] Les pseudo-phonèmes consonantiques ont tendance à être sur-segmentés par l'algorithme DFB. Ceci est dû au fait que la modélisation LPC du signal de parole est adaptée aux voyelles, mais beaucoup moins pour les consonnes. Ces dernières présentent en effet de fortes non-linéarités conduisant à un spectre d'énergie relativement plat dans les hautes fréquences. Un spectre parfaitement plat est par définition non prédictible par les modèles AR.

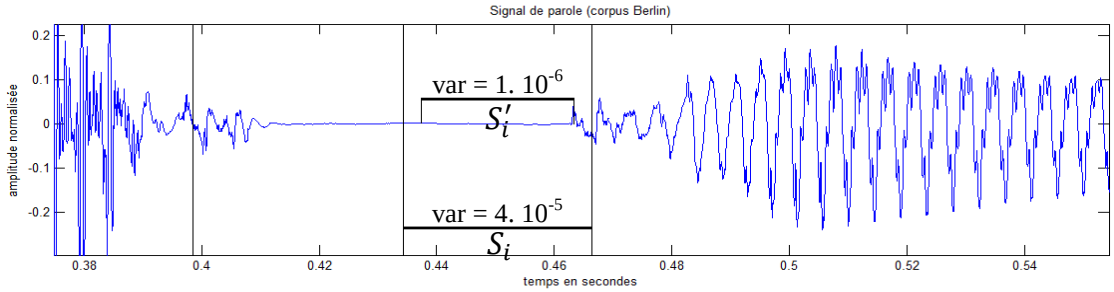


Fig. 2.10 Effets de bord lors du calcul de la variance des segments issus du DFB. L'estimation de la variance est nettement plus importante sur la partie entière du segment S_i que sur sa version tronquée S_i' (80 % de la durée).

$$\sigma_a = \alpha \min_i \big(var(S_i)\big) \qquad [4]$$

pour objectif de mesurer l'adéquation entre la distribution spectrale d'une trame du signal de parole et la structure formantique propre aux segments vocaliques. Ainsi, plus un son se rapproche d'une structure vocalique, plus la valeur de $REC(k)$ sera grande par rapport aux trames voisines. A *fortiori*, les maxima de cette fonction localisent les segments vocaliques contenus dans le signal de parole. L'estimation de la fonction REC nécessite de segmenter tout d'abord le signal de parole en trames, cf. Fig. 3.2. La durée des trames est alors fixée à l'approximation numérique (puissance de 2) du critère de stationnarité, i.e., 30ms → 32ms, et les trames se recouvrent de moitié ; ces dernières sont fournies toutes les 16ms. 24 coefficients d'énergie spectrale E_i sont ensuite extraits sur chaque trame k au moyen d'une échelle psycho-acoustique (mel, coefficients *mel frequency spectrum coding* – MFSC) et la fonction REC [5] cumule pour chaque trame k les valeurs d'énergie $E_i(k)$ qui sont supérieures à leur valeur moyenne $\bar{E}(k)$. Un poids α_i peut être attribué à chaque filtre utilisé dans la sommation pour tenir compte des caractéristiques de codage du signal de parole (e.g., réduction de la bande passante pour la parole téléphonique : 330 – 3400Hz). Afin d'apporter plus de confiance lors de la phase de détection, les valeurs issues de la sommation sont pondérées par la proportion d'énergie contenue dans les basses fréquences (rapport d'énergie entre les basses fréquences E_{BF} et le spectre total E_T) [5].

La phase de détection des noyaux vocaliques consiste à identifier les pics présents sur la dérive spectrale REC. Afin d'éliminer d'éventuels sommets parasites, la courbe REC est lissée par un filtre moyenneur via une fenêtre glissante de type SMA (*simple move average*) ; la longueur du filtre vaut trois trames, i.e., 48ms. Les maxima locaux sont ensuite filtrés par un seuil S_e sur la courbe REC (fixé expérimentalement au dixième de la valeur moyenne de la courbe REC) pour éliminer les segments peu énergétiques. Afin de renforcer la confiance apportée lors de la détection des voyelles, nous avons ajouté un test sur le coefficient d'auto-

[25] D'autres techniques existent pour détecter l'activité vocale. La plupart de ces méthodes reposent sur une segmentation du signal en trames et n'utilisent pas d'information contextuelle. Les caractéristiques qui y sont extraites sont : (i) les spectres d'énergie totale et en sous-bandes [WOO00], (ii) des mesures de divergence spectrale entre la parole et le bruit de fond [MAR02], (iii) le pitch [TUC92], (iv) le taux de passage par zéro [RAB75] et (v) des statistiques haut-niveaux [NEM01], [SOH99]. La règle de décision parole / non-parole peut reposer sur la distance Euclidienne [GOR06] ou les divergences d'Itakura-Saito et de Kullback-Leibler [RAM-04]. D'autres techniques exploitent des méthodes de classification telle que la logique floue [BER02], les machines à vecteur support [RAM06] ou encore les algorithmes génétiques [EST05].

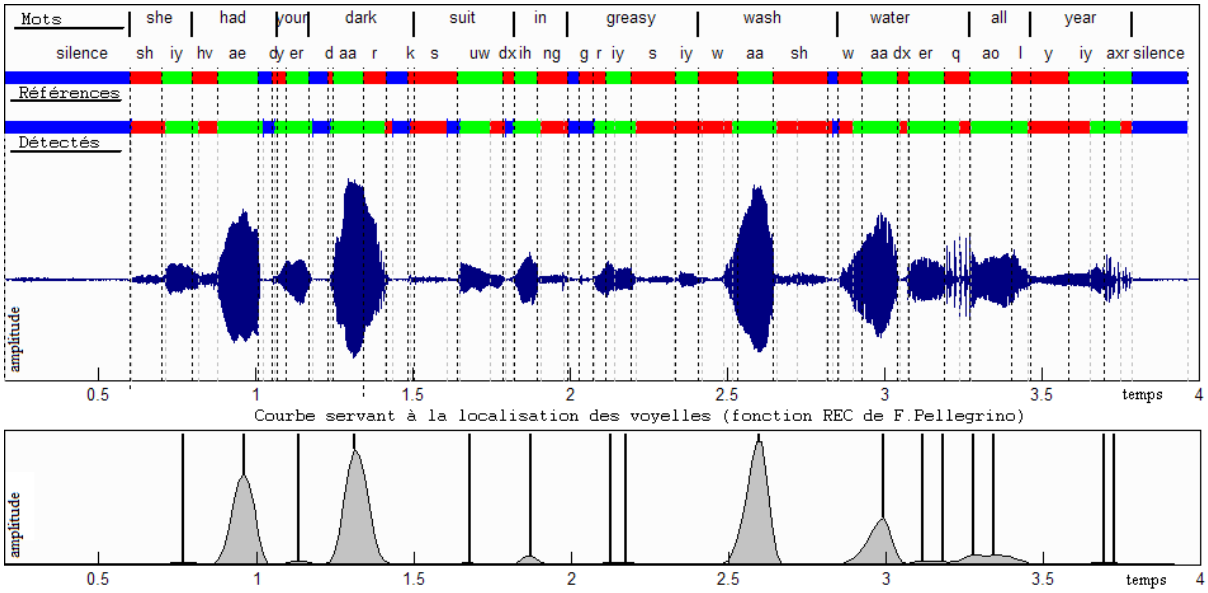


Fig. 2.11 Comparaison d'une segmentation phonétique manuelle *vs.* automatique (Références *vs.* Détectés). Les données sont issues du corpus *TIMIT* ; le code de couleur est le suivant : bleu – silence, rouge – consonne et vert – voyelle.

$$REC(k) = \frac{E_{BF}(k)}{E_T(k)} \sum_{i=1}^{24} \alpha_i \left(E_i(k) - \bar{E}(k)\right)^+ \qquad [5]$$

avec,

k :	indice de trame du signal de parole
$E_i(k)$:	énergie contenue dans le filtre numéro no. i
$\bar{E}(k)$:	énergie moyenne à travers tous les filtres Mel
$E_{BF}(k)$:	énergie moyenne contenue dans les filtres de fréquence inférieure à 1kHz
$E_T(k)$:	énergie totale contenue dans la trame k
α_i :	poids affecté au filtre numéro i (non utilisés dans cette étude, i.e., $\alpha_i = 1$)

et,

$$S_e = \frac{1}{10} mean(REC) \qquad S_a = 40\% \qquad [6]$$

corrélation du signal. La valeur moyenne de ce coefficient, qui est fourni toutes les 10ms par l'algorithme *Snack*, doit alors être supérieure à un seuil S_a fixé expérimentalement à 40% [6].

Les segments de parole issus du DFB sont considérés comme des « *pseudo-voyelles* » lorsqu'un ou plusieurs pics sont détectés simultanément sur la fonction REC. La détection doit toutefois être confortée par deux critères pour être validée : (i) le sommet principal des pics doit être supérieur au seuil S_e et (ii) la valeur moyenne du coefficient d'autocorrélation du segment correspondant doit être supérieur au seuil S_a. La Fig. 2.11 représente l'évolution de la fonction REC sur une phrase du corpus TIMIT : « She had your dark suit in greasy wash water all year ». On constate que la mesure de dérive spectrale détermine des lobes généralement centrés sur les voyelles et dont la hauteur est liée à l'énergie du signal.

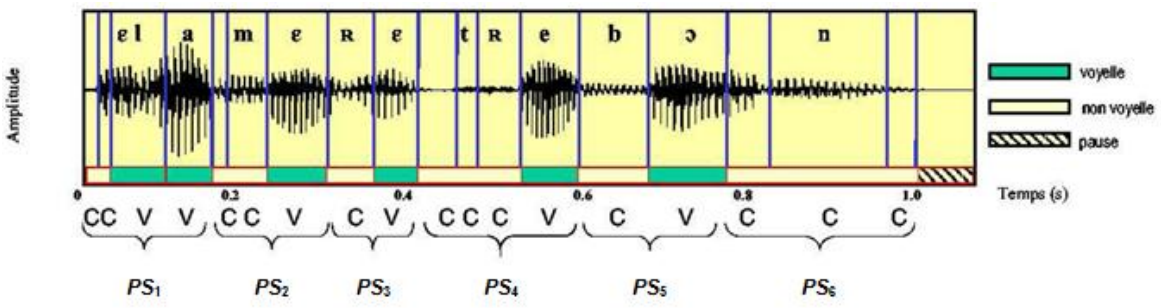


Fig. 2.12 Regroupement des pseudo-phonèmes (C et V) en pseudo-syllabes (*PS*) ; figure extraite de [ROU05][26].

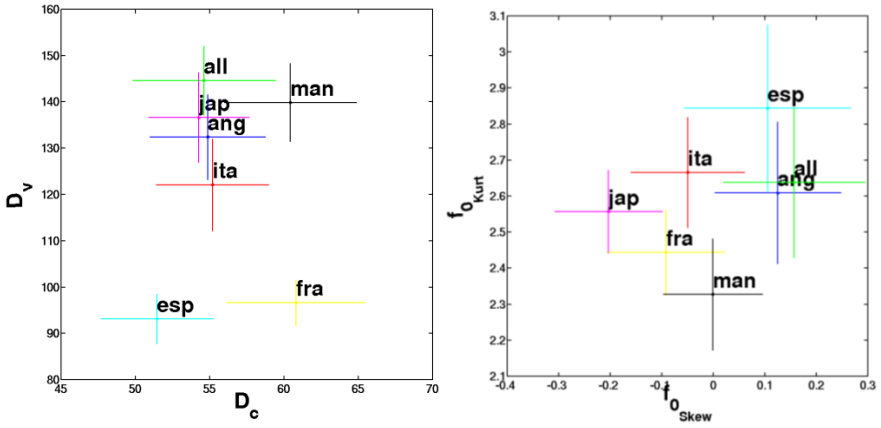


Fig. 2.13 Caractérisation des langues au moyen de la pseudo-syllabe ; all : Allemand ; ang : Anglais ; esp : Espagnol ; fra : Français ; ita : Italien ; jap ; Japonais ; man : Mandarin ; Dv : durée de la voyelle ; Dc durée totale des consonnes ; $f_{0\ skew}$: moment statistique d'ordre 3 ; $f_{0\ kurt}$: moment statistique d'ordre 4 ; figure extraite de [ROU05][26].

3.2. Les pseudo-syllabes

Une unité rythmique basée sur le regroupement des pseudo-phonèmes a été introduite dans [FAR01][27]. Il a notamment été proposé de regrouper les consonnes précédant les voyelles afin de créer des structures de type C^nV alors appelées « pseudo-syllabes », cf. Fig. 2.12. Cette unité a permis de discriminer différentes langues issues du corpus MULTEXT[28] au moyen de paramètres prosodiques tels que la durée et les moments statistiques d'ordre 3 et 4 du pitch, cf. Fig. 2.13, [ROU05][26]. Des modèles décrivant l'enchainement de ces classes ont alors été employés (multi-grammes).

3.3. Les « *p-centres* »

Il existe dans la parole un centre de perception appelé « *p-centre* », cf. sous-section 2.2.2.

[26] J. L. Rouas, *Caractérisation et identification automatique des langues*, Thèse de doctorat, 2005.

[27] J. Farinas et F. Pellegrino, "Automatic rhythm modeling for language identification", dans proc. *Eurospeech*, Aalborg, Denmark, Sep. 3-7 2001, pp. 2539–2542.

[28] Ce corpus contient du texte (5 phrases totalisant environ 20 secondes) qui a été lu par 10 locuteurs (5 hommes, 5 femmes) pour cinq langues différentes (Anglais, Allemand, Espagnol, Français et Italien).

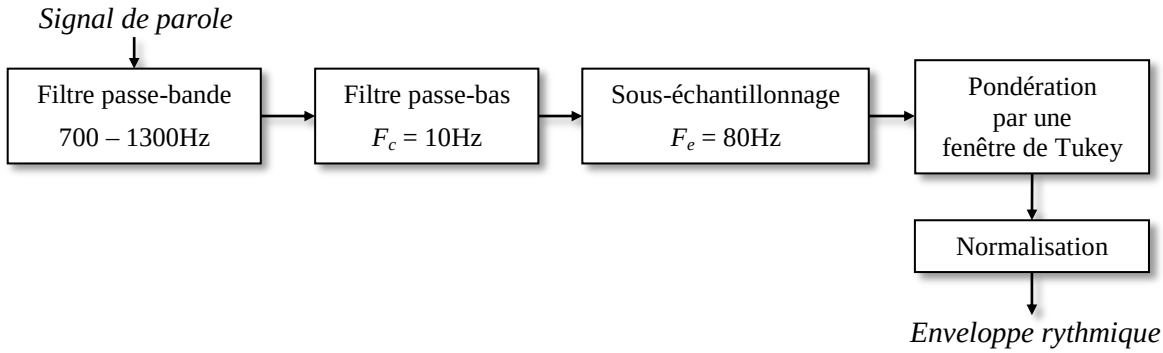


Fig. 2.14 Filtrages successifs réalisés pour extraire l'enveloppe rythmique d'un signal de parole.

Ce point désigne les instants rythmiques à la fois perçus par un locuteur et un auditeur. Une méthode permettant d'extraire l'enveloppe rythmique d'un signal de parole a été proposée par Tilsen *et al.* [TIL08b][29]. Les proéminences issues de cette enveloppe permettent de localiser les « *p-centres* ». Un ensemble de filtres numériques, supposés représenter les processus employés par l'Homme pour percevoir le rythme de la parole, est pour cela employé, cf. Fig. 2.14.

Un filtre de Butterworth est tout d'abord appliqué sur le signal de parole dans la bande de fréquence 700 – 1300 Hz ; cette bande de fréquence a été identifiée comme étant celle du « *p-centre* » [CUM98][30]. La seconde étape du filtrage consiste à extraire l'enveloppe du signal au moyen d'un passe-bas de type Butterworth (ordre 4 et fréquence de coupure égale à 10Hz). Le signal est ensuite sous-échantillonné à la fréquence de 80Hz et une correction de 45ms est appliquée pour le retard de phase introduit par les filtres, i.e., la somme des retards de phase des filtres dans leur bande-passante. L'enveloppe rythmique est enfin pondérée par une fenêtre de Tukey (r = 0.1) et normalisée par sa moyenne. L'exemple de la Fig. 2.15 montre que les proéminences rythmiques du signal de parole ne correspondent pas forcément avec celles de l'énergie acoustique.

L'enveloppe rythmique permet de localiser les centres de perception de la parole. Les ancrages « *p-centre* » peuvent par exemple être obtenus en définissant des seuils (e.g., 1/3, 1/4, 1/6) sur l'amplitude de l'enveloppe rythmique du signal de parole [RIN09][31]. Ces seuils permettent de représenter, de façon artificielle, différents niveaux de perception de la proéminence rythmique : niveau 1, seuil = 1/3 de l'amplitude, niveau 2, seuil = 1/4 de l'amplitude, et niveau 3, seuil = 1/6 de l'amplitude, cf. Fig. 2.16. Les « *p-centres* » détectés au niveau 1 sont ainsi situés au sommet de l'échelle de proéminence rythmique. Les deux autres niveaux intègrent quant à eux plus d'informations sur la structure de la proéminence, e.g., la dissymétrie présente dans l'avant-dernier segment « *p-centre* » de l'exemple de la Fig. 2.16 n'apparaît que sur le niveau 3.

[29] S. Tilsen et K. Johnson, "Low-frequency Fourier analysis of speech rhythm", dans *J. of Acoust. Soc. of Amer., Express Letters*, vol. 124, no. 2, pp. 34–39, Aug. 2008.

[30] F. Cummins et R. Port, "Rhythmic constraints on stress timing in English", dans *Journal of Phonetics*, vol. 26, pp. 145–171, 1998.

[31] F. Ringeval et M. Chetouani, "Hilbert-Huang transform for non-linear characterization of speech rhythm", dans proc. *NOLISP*, Vic, Spain, Jun. 25-27 2009.

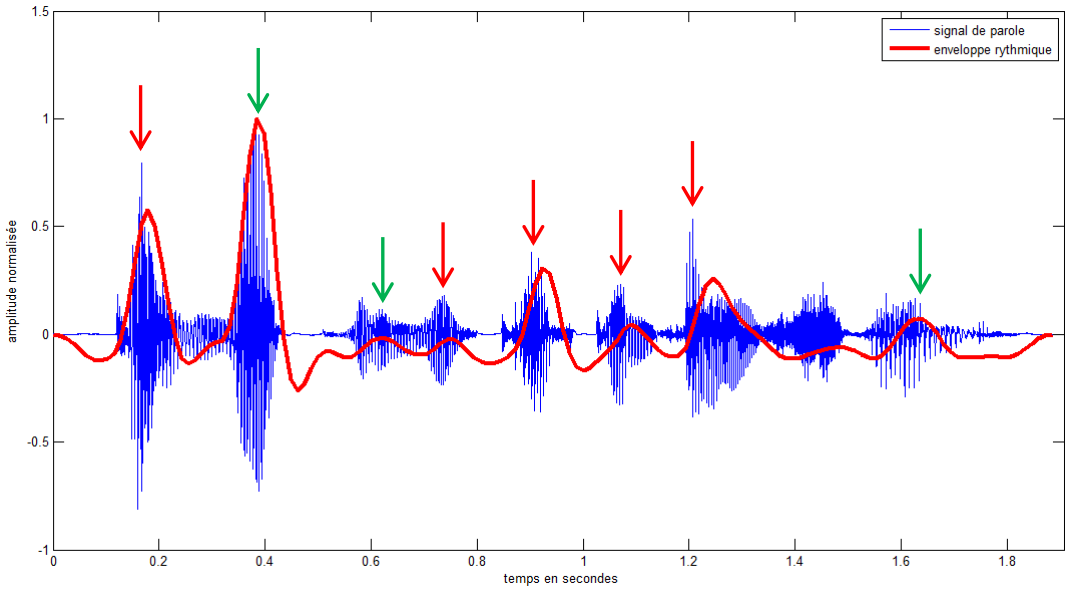


Fig. 2.15 Exemple d'enveloppe rythmique extraite sur un signal de parole.

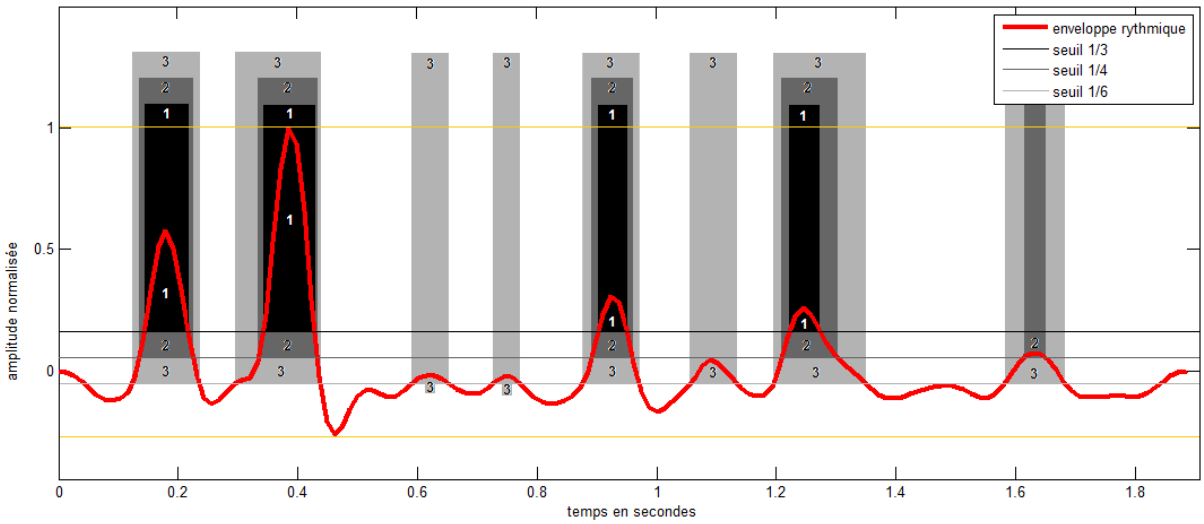


Fig. 2.16 Niveaux de perception des « *p-centres* » selon le degré de seuillage ; niveau 1, seuil = 1/3 de l'amplitude ; niveau 2, seuil = 1/4 de l'amplitude ; et niveau 3, seuil = 1/6 de l'amplitude.

4. Expérimentations

Cette section présente les résultats d'une étude visant à caractériser le comportement des systèmes précédents sur des données incluant différents types de parole (e.g., lue et affective). Les scores en détection de pseudo-phonèmes ainsi que les structures phonétiques (i.e., taux de recouvrement des voyelles et des consonnes) des ancrages rythmiques « *p-centres* » ont pu être calculés grâce aux transcriptions contenues dans les corpus. Ces derniers sont présentés dans la sous-section 4.1 et les résultats sont décrits dans les deux sous-sections suivantes.

4.1. Corpus de parole étudiés

Les expériences incluent l'analyse de corpus de parole lue en qualité laboratoire (corpus

Table 2.1 Répartition des locuteurs de la base TIMIT (TRAIN + TEST) selon les régions dialectales.

Région dialectale (U.S.A.)	Nombre de locuteurs	Durée totale en minutes
New England	51	25'
Nothern	104	52'
North Midland	104	51'
South Midland	102	52'
Southern	100	52'
New York City	48	24'
Western	102	51'
Army Brat (moved a lot)	35	16'

TIMIT) et téléphonique (corpus NTIMIT) pour différents dialectes de l'Anglais Américain, ainsi que des corpus de parole affective pour plusieurs langues : Allemand (corpus Berlin), Hongrois (corpus Bute-TMI) et Basque (corpus Aholab). Tous ces corpus contiennent des transcriptions phonétiques qui permettent de fournir des données cibles au système de détection des pseudo-phonèmes. Les performances de ce dernier peuvent donc être évaluées à partir des transcriptions phonétiques, tout comme les corrélats acoustiques associés aux « *p-centres* ».

4.1.1. Parole lue

Les corpus de parole dite « lue » contiennent peu de variabilités dans le signal de parole compte tenu de la nature contrainte de l'épreuve de lecture. Ce type de corpus permet de confronter les systèmes de détection des pseudo-phonèmes à un premier niveau de difficulté qui est plutôt bas si les enregistrements ne sont que faiblement bruités.

4.1.1.1. Corpus TIMIT

Le corpus TIMIT [GAR93][32] a été créé sous l'impulsion du *Defense Advanced Research Projects Agency* (DARPA) par le *Massachusetts Institute of Technology* (MIT), le *Stanford Research Institute* (SRI) et *Texas Instruments* (TI), et a été très utilisé en reconnaissance de la parole et du locuteur [CHE04][33]. Il contient 10 phrases qui ont été prononcées par 630 locuteurs (192 hommes et 438 femmes) issus de 8 régions dialectales des Etats-Unis, cf. table 2.1. Deux sous-ensembles de données sont disponibles pour chaque région : TRAIN et TEST. La fréquence d'échantillonnage est de 16kHz et le signal est codé sur 16 bits. La durée des enregistrements totalise environ 6h de parole lue. Les phrases sont de trois types : (i) 2 phrases communes à tous les locuteurs, (ii) 5 phrases avec un contexte phonétique imposé et (iii) 5 phrases avec un contexte phonétique libre. Les phrases ont été manuellement transcrites en phonèmes via un lexique composé de 52 phonèmes, i.e., 20 voyelles et 32 consonnes. L'intérêt du corpus TIMIT pour notre étude consiste à analyser l'impact des styles de parole associés aux régions dialectales sur les performances en détection de pseudo-phonèmes.

[32] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren et V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia (PA), U.S.A., 1993.

4.1.1.2. Corpus NTIMIT

Le corpus NTIMIT [JAN90][34] a été obtenu par le passage du corpus TIMIT à travers le réseau téléphonique terrestre. La bande passante est donc de 330Hz – 3400Hz avec un signal toujours échantillonné à 16kHz. Les configurations des appels téléphoniques varient. La moitié d'entre eux sont des appels locaux et donc de la même région, alors que les autres sont des appels longues distances. La base NTIMIT est également divisée en régions dialectales dans lesquelles deux sous-ensembles sont disponibles (TRAIN et TEST) pour chaque région. Une segmentation phonétique est également fournie. Le lexique et la répartition en locuteur selon les régions dialectales est identique au corpus TIMIT. L'intérêt principal de ce corpus pour notre étude réside dans la qualité téléphonique des signaux de parole. Ce type de qualité permet de confronter le système de détection des ancrages acoustiques de la parole à un second niveau de difficulté. En effet, il a été montré que les performances des systèmes de reconnaissance de la parole [MOR94][35] et du locuteur [CHE09a][36] sont diminuées lors du traitement du corpus NTIMIT.

4.1.2. Parole affective

Les corpus de parole affective peuvent être constitués par différentes approches, cf. chapitre 1, sous-section 3.4.1. Les variabilités présentes dans le signal de parole sont plus importantes lorsque le locuteur est emprunt à une émotion donnée qu'en situation non-émotionnelle [SCH86][37]. Des études ont ainsi montré que le taux de reconnaissance de la parole se dégrade lorsque cette dernière est chargée d'affect [ROT05][14], [SCH06a][15] et [STE10][16], et que la durée de production des phonèmes varie selon l'état émotionnel du locuteur, [BUL05][38] et [RIN08c][39] ; cf. Fig. 1.10. La tâche de détection des ancrages acoustiques de la parole sur les corpus affectifs est donc plus complexe par rapport au corpus contenant de la parole lue, puisque les caractéristiques acoustiques des phonèmes ont été modifiées pour introduire les informations liées à l'affect. Dans notre étude, ces corpus permettent d'étudier le système de détection des pseudo-phonèmes à travers les variabilités introduites par l'affect dans la parole.

---

[33] M. Chetouani., *Codage neuro-prédictif pour l'extraction de caractéristiques de signaux de parole*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris, 2004.

[34] C. Jankowski, A. Kalyanswamy, S. Basson et J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwith speech database", dans proc. *ICASSP*, Albuquerque (NM), USA, Apr. 3-6 1990, vol. 1, pp. 109–112.

[35] P. J. Moreno et R. M. Stern, "Sources of degradation of speech recognition in the telephone network", dans proc. *ICASSP*, Pittsburgh (PA), USA, Apr. 19-22 1994, vol. 1, pp. 109–112.

[36] M. Chetouani, M. Faundez-Zanuy, B. Gas et J. L. Zarader, "Investigation on LP-residual representations for speaker identification", dans *Pattern Recogn.*, vol. 42, no. 3, pp. 487–494, 2009.

[37] K. R. Scherer, "Vocal affect expression: a review and a model for future research", dans *Pychological Bulletin*, vol. 99, no. 2, pp. 143–165, Mar. 1986.

[38] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. Lee, S. Lee, et S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 801–804.

[39] F. Ringeval et M. Chetouani, "A vowel based approach for acted emotion recognition", dans proc. *Interspeech*, Brisbane, Australia, Sep. 22-26 2008, pp. 2763–2766.

Fig. 2.17 Un des acteurs durant l'enregistrement du corpus Berlin ; figure extraite de [BUR05][40].

Table 2.2 Styles de production des phrases du corpus Berlin.

Style de production	Durée totale en minutes
Colère	6'
Peur	2'
Joie	4'
Tristesse	4'
Dégoût	2'
Ennui	4'
Neutre	3'

4.1.2.1. Corpus Berlin

Le corpus Berlin [BUR05][40] est couramment utilisé pour la reconnaissance des émotions. Ce corpus contient 10 phrases (5 courtes et 5 longues) issues de discussions de tous les jours. Ces phrases ont été produites selon 6 styles affectifs différents par 10 acteurs Allemands (5 hommes et 5 femmes). Une attention particulière a été apportée par les auteurs de la base sur le style de parole des acteurs. Afin que les émotions jouées soient le plus naturelle possible, il leur a notamment été demandé d'adopter un style de production différent de celui utilisé lors de leurs représentations théâtrales. Les émotions jouées sont les suivantes : *Colère*, *Peur*, *Joie*, *Tristesse*, *Dégoût* et *Ennui* ; un style « *Neutre* », i.e., sans émotion particulière, est également fourni. Le corpus Berlin contient au total environ 25min. de parole affective, cf. table 2.2. Excepté pour l'émotion « *Ennui* », toutes les autres correspondent à des émotions pleines [COW03][41], cf. chapitre 1, sous-section 3.1. Le signal est échantillonné à la fréquence de 16kHz et est codé sur 16 bits avec un matériel de haute qualité (microphone *Sennheiser*, chambre anéchoïque, cf. Fig. 2.17). Plusieurs sessions espacées dans le temps ont été effectuées pour les enregistrements. 20 auditeurs ont participé à un test de perception afin d'évaluer l'aspect naturel des phrases jouées et les émotions reconnues. 535 phrases alors jugées comme naturelles à 60% minimum et dont l'émotion était reconnaissable à 80% minimum ont été conservées sur un total de 800 phrases disponibles (plusieurs versions existent).

[40] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier et B. Weiss, "A database of German emotional speech", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 1517–1520.

[41] R. Cowie et R. R. Cornelius, "Describing the emotional states that are expressed in speech", dans *Speech Comm.*, vol. 40, pp. 2–32, 2003.

Table 2.3 Styles de production des phrases du corpus Bute-TMI.

Style de production	Durée totale en minutes
Colère	4'
Peur	5'
Joie	5'
Tristesse	5'
Dégoût	5'
Surprise	5'
Nerveux	4'
Neutre	5'

Ces phrases ont été étiquetées avec l'alphabet SAMPA[1] de l'Allemand dans une transcription fine incluant les phonèmes, les syllabes et les mots. Le lexique est constitué de 37 phonèmes : 16 voyelles et 21 consonnes. Une analyse détaillée des transcriptions phonétiques contenues dans le corpus Berlin est donnée en annexe 1.

4.1.2.2. Corpus Bute-TMI

Le corpus Bute-TMI [TOT08][42] contient de la parole affective actée en langue Hongroise. 37 acteurs non-professionnels (i.e., issus de différents milieux universitaires, e.g., enseignant-chercheurs, étudiants, musiciens, avocats) ont prononcé 3 phrases à travers 7 styles affectifs : *Colère*, *Peur*, *Joie*, *Tristesse*, *Dégoût*, *Surprise*, *Nerveux* ; un style « *Neutre* » est également fourni, cf. table 2.3. Les phrases contiennent chacune un contexte phonétique différent. Le corpus Bute-TMI totalise environ 38min d'enregistrement de parole affective. Excepté pour l'émotion « *Nerveux* », toutes les autres correspondent à des émotions pleines [COW03][41]. Le signal est échantillonné à la fréquence de 44,1kHz et est codé sur 16 bits. Des tests de perception ont été effectués par les membres du laboratoire TMI afin de vérifier que les émotions produites soient correctement identifiables par l'Homme. Sur un total de 888 phrases, 728 ont été retenues et étiquetées en phonèmes. La segmentation a été réalisée en deux phases : (i) les alignements temporels des phonèmes avec les signaux de parole ont tout d'abord été fournis automatiquement par un système reposant sur des chaînes de Markov cachées (HMM) et (ii) la segmentation a ensuite été vérifiée manuellement pour éliminer les éventuelles erreurs produites par le système. Le lexique est composé de 27 phonèmes qui sont issus de l'alphabet SAMPA du Hongrois : 8 voyelles et 19 consonnes.

4.1.2.3. Corpus Aholab

Le corpus Aholab [SAR06][43] contient de la parole affective actée en langue Basque. Il est constitué d'un ensemble de 702 phrases qui ont été extraites via différentes sources (plus de 500 000 phrases fournies par des journaux Basques, des romans et d'autres sources littéraires)

[42] S. L. Tóth, D. Sztahó et K. Vicsi, "Speech emotion perception in human and machine", dans *LNCS*, A. Esposito, N. G. Bourbakis, N. Avouris and I. Hatzilygeroudis [Eds], *Verbal and Non-verbal features of Human-Human and Human-machine interaction*, Springer-Verlag, vol. 5024, pp. 213–224, 2008.

Fig. 2.18 Speakerine reproduisant les phrases du corpus Aholab durant l'une des sessions d'enregistrement ; figure extraite de [SAR06][43].

Table 2.4 Styles de production des phrases du corpus Aholab.

Style de production	Durée totale en minutes
Colère	58'
Peur	65'
Joie	78'
Tristesse	72'
Dégoût	71'
Surprise	59'
Neutre	69'

en conservant tant que possible la fréquence et la distribution des diphonies alors observées. Les phrases ont ensuite été reproduites par une speakerine radiophonique dans un studio semi-professionnel, cf. Fig. 2.18. Le signal est échantillonné à la fréquence de 48kHz et est codé sur 16 bits. Les styles affectifs incluent les six émotions pleines décrites par [COW03][41] : *Colère*, *Peur*, *Joie*, *Tristesse*, *Dégoût* et *Surprise* ; un style « *Neutre* » est également fourni. Le corpus Aholab totalise environ 8h d'enregistrements de parole affective et la durée des enregistrements varie selon le style de production, cf. table 2.4. Les phrases ont été transcrites manuellement en phonèmes au moyen de l'alphabet SAMPA[1] du Basque. Le lexique est composé de 35 phonèmes : 5 voyelles et 30 consonnes.

4.1.3. Récapitulatif

La table 2.5 présente un récapitulatif des données utilisées dans les expériences de ce chapitre. La table 2.6 détaille quant à elle le nombre de fichiers disponibles et le nombre d'unités extraites sur les corpus étudiés. Deux valeurs sont données pour le corpus Berlin. Cela est dû au fait que les labels phonétiques utilisés lors de la transcription de ce corpus ne sont pas tous identifiés par l'alphabet SAMPA de l'Allemand ; 62% des labels disponibles ne sont pas identifiés par cet alphabet, cf. annexe 1. Cependant, ces labels ont pu être caractérisés en voyelle / consonne par une analyse manuelle des exemples fournis par les signaux de parole. Ainsi, les deux valeurs données dans la table 2.6 sur le corpus Berlin correspondent aux deux configurations d'analyse des transcriptions : la première configuration inclut uniquement les données qui sont renseignées par l'alphabet SAMPA, alors que la seconde intègre les phonèmes manquants. Nous verrons dans les expériences qui vont suivre que cette inclusion est

[43] I. Saratxaga, E. Navas, I. Hernaez et I. Luengo, "Designing and recording an emotional speech database for corpus based speech synthesis in Basque", dans proc. *LREC*, Genoa, Italy, May 24-26 2006, pp. 2126–2129.

Table 2.5 Comparaison des caractéristiques principales des corpus de parole étudiés.

Caractéristique	TIMIT	NTIMIT	Berlin	Bute-TMI	Aholab
Parole	Lue	Lue	Affective	Affective	Affective
Qualité	Correcte	Téléphone	Correcte	Correcte	Correcte
Classes d'information	8 régions dialectales	8 régions dialectales	7 styles affectifs	8 styles affectifs	7 styles affectifs
Labels phonétiques	52 20 V / 32 C	52 20 V / 32 C	37 16 V / 21 C	27 8 V / 19 C	35 5 V / 30 C
Locuteur	630	630	10	37	1
Phrase	10	10	10	3	702
Durée	~ 6 heures	~ 6 heures	~ 25 minutes	~ 38 minutes	~ 8 heures

Table 2.6 Comparaison des caractéristiques des transcriptions issues des corpus de parole étudiés.

Corpus	TIMIT	NTIMIT	Berlin	Bute-TMI	Aholab
Fichiers	6 300	6 300	533	728	5 040
Phonèmes	192 571	192 571	15 226 / 17 298	26 543	277 949
Consonnes	114 154	114 154	10 072 / 10 861	16 768	141 312
Voyelles	78 417	78 417	5 154 / 6 437	9 775	136 637
P-phonèmes	280 721	126 983	21 786	25 472	306 908
P-consonnes	207 056	62 057	14 770	14 752	185 435
P-voyelles	73 665	64 926	7 016	10 720	121 473
Voisé	111 661	102 190	4 175	6 085	159 379
Non-voisé	64 047	50 416	5 147	10 646	105 217
« *p-centres* » 1	31 873	30 388	3 234	4378	69 609
« *p-centres* » 2	39 440	37 494	4 078	5545	84 726
« *p-centres* » 3	50 036	47 726	4 830	6769	87 874

Les deux valeurs données pour le corpus Berlin correspondent respectivement à une configuration correspondant soit (i) aux phonèmes identifiés par l'alphabet SAMPA[1] ou soit (ii) tous les phonèmes contenus dans les transcriptions.

nécessaire pour atteindre un taux d'erreur convenable en détection des pseudo-phonèmes. Toutefois, les scores en reconnaissance des émotions ne sont pas forcément meilleurs sur ces données, cf. annexe 1, section 4.

4.2. Détection automatique des pseudo-phonèmes

Le système de détection automatique des voyelles, cf. sous-section 3.1, a été testé par leurs auteurs sur de nombreuses bases de données. Ces tests ont permis d'en étudier sa robustesse face à différents types de situations : parole de qualité laboratoire, téléphonique, changement de locuteur et de langues [OBR93][21]. Il est apparu qu'une version robuste du système pouvait être utilisée quel que soit le type de signal étudié. Par ailleurs, d'excellents résultats ont été obtenus pour une configuration unique du système de détection.

Cette section a pour objectif de vérifier que la fiabilité du système de détection des pseudo-phonèmes est effective, notamment lorsque ce dernier est confronté à différents types

de corpus de parole affective. L'évaluation des performances repose sur le *Vowel Error Rate* (VER) et nécessite des données cibles, l'accès aux transcriptions phonétiques du signal de parole est donc (ici) indispensable. Le VER [7] regroupe les deux types d'erreur qui peuvent être commises lors de la détection automatique des points d'ancrages vocaliques : les non-détections et les insertions.

$$VER = 100.\left[\frac{N_{nondet} + N_{ins}}{N_{voy}}\right]\% \qquad [7]$$

avec, N_{nondet} le nombre de voyelles non détectées, N_{ins} le nombre de voyelles insérées (i.e., détectées de façon erronée), et N_{voy} le nombre de voyelles totales pouvant être détectées (i.e., issues des transcriptions).

Nous présentons dans la table 2.7 les résultats obtenus en détection de pseudo-phonèmes sur les corpus de parole étudiés [RIN08abc][44,45,39]. La configuration du système de détection employé est alors unique quel que soit le corpus. De plus, les phonèmes détectés par le système ne peuvent valider qu'une seule fois les segments issus des transcriptions. Au total, plus de 700k phonèmes, i.e., 400k consonnes et 300k voyelles, ont été testés pour une durée d'enregistrement supérieure à 20h de parole, cf. table 2.5 et 2.6. Le taux d'erreur CER des consonnes est calculé de façon identique au VER.

Les expériences montrent que le VER reste inférieur à 30% dans tous les cas de figure, excepté pour le corpus de parole affective Bute-TMI qui contient le plus grand nombre d'acteurs, qui plus est, non-professionnels. Les meilleurs résultats sont obtenus sur le corpus de parole lue TIMIT qui contient plus de 6h de parole pour 630 locuteurs différents. Les variabilités introduites dans le signal de parole par la qualité téléphonique ou le style affectif dégradent dans des proportions comparables les performances du VER par rapport au corpus TIMIT. Notons toutefois que les taux d'insertion des pseudo-phonèmes (V_{INS} et C_{INS}) sont plus faibles sur NTIMIT, et plus élevés pour les corpus de parole affective puisque les émotions introduisent des modifications dans les paramètres acoustiques. Comme le rapport signal sur bruit est mauvais sur la version bruitée du corpus TIMIT, le détecteur d'activité vocale a tendance à rejeter plus souvent les segments de faible énergie. De plus, l'introduction de bruit dans le signal modifie l'équilibre entre le taux de non-détection et le taux d'insertion dont la résultante est le taux d'erreur. Cette modification est d'autant plus flagrante sur les consonnes qui sont moins énergétiques que les voyelles ; le CER diminue de 36% entre le corpus TIMIT et sa version bruitée NTIMIT.

Les deux approches employées pour l'analyse des transcriptions phonétiques du corpus Berlin (i.e., ensemble de phonèmes limités à l'alphabet SAMPA[1] ou complet, cf. sous-section 4.1.3) font apparaître des différences significatives sur les taux d'erreurs : le VER est divisé par 2 entre les deux approches et le CER est diminué. Ces différences montrent que les trans-

[44] F. Ringeval et M. Chetouani, "Exploiting a vowel based approach for acted emotion recognition", dans *LNCS*, A. Esposito, N. G. Bourbakis, N. Avouris and I. Hatzilygeroudis [Eds], *Verbal and Nonverbal Features of Human-Human and Human-machine Interaction*, Springer Verlag, vol. 5042, pp. 243–254, 2008.

[45] F. Ringeval et M. Chetouani, "Une approche basée voyelle pour la reconnaissance automatique des émotions actées", dans proc. *JEP*, Avignon, France, Jun. 9-13 2008.

Table 2.7 Comparaison des résultats en détection de pseudo-phonèmes sur divers corpus de parole.

Taux	TIMIT	NTIMIT	Berlin	Bute-TMI	Aholab
V_{DET}	86.8	78.0	84.2 / 90.0	79.4	82.3
V_{INS}	7.1	4.8	52.0 / 19.0	11.7	6.6
VER	**20.3**	**26.9**	**67.8 / 29.0**	**32.3**	**24.3**
C_{DET}	87.5	42.2	78.9 / 80.5	71.7	75.8
C_{INS}	93.8	12.1	67.7 / 55.5	41.9	55.4
CER	**106**	**69.9**	**88.8 / 75.0**	**70.2**	**79.6**

V/C_{DET} : taux de détection en % de voyelle / consonne ; V/C_{INS} : taux d'insertion (…) ; V/C-ER : taux d'erreur (…) ; Les deux valeurs données pour le corpus Berlin correspondent : (i) aux phonèmes identifiés par l'alphabet SAMPA[1] ou (ii) à tous les phonèmes contenus dans les transcriptions.

Table 2.8 Performances en détection des pseudo-phonèmes (V et C) selon les styles de production du corpus Berlin.

Taux	Colère	Peur	Joie	Tristesse	Dégoût	Ennui	Neutre
V_{DET}	90.5	83.8	100	82.2	92.0	86.4	85.7
V_{INS}	19.0	22.1	14.3	17.8	18.4	23.5	21.4
VER	**28.6**	**38.2**	**14.3**	**35.6**	**26.4**	**37.0**	**35.7**
C_{DET}	86.0	88.9	85.7	90.0	81.5	86.7	82.1
C_{INS}	60.5	59.8	85.7	56.3	54.1	62.2	57.1
CER	**74.4**	**70.9**	**100**	**66.3**	**72.6**	**75.5**	**75.0**

V/C_{DET} : taux de détection en % des voyelles / consonnes ; V/C_{INS} : taux d'insertion (…) ; V/C-ER : taux d'erreur (…).

criptions phonétiques absentes de l'alphabet SAMPA[1] contiennent bien des informations vocaliques et consonantiques. La faible valeur de VER obtenue sur le corpus Aholab peut s'expliquer par le fait que les données ont été fournies par une seule locutrice, leur variabilité est donc beaucoup moins importante comparée aux autres corpus qui contiennent plus de locuteurs, cf. table 2.5. La valeur élevée de VER qui apparait sur le corpus Bute-TMI peut également s'expliquer par le fait que la parole a été produite par des acteurs non-professionnels.

Compte tenu des fortes non-linéarités présentes dans le signal de parole sur les segments consonantiques, les taux d'erreurs sont beaucoup plus élevés sur les consonnes comparés à ceux obtenus sur les voyelles. L'algorithme de segmentation DFB a en effet tendance à sur-segmenter les segments consonantiques ; le taux de détection est très correct mais égal voir inférieur au taux d'insertion. Une optimisation des paramètres a été effectuée sur chaque corpus afin d'évaluer l'amélioration des performances obtenues par rapport à une configuration unique du détecteur. Bien que de nombreux paramètres interviennent dans le processus de détection des pseudo-phonèmes tels que : (i) la taille des trames et le nombre de coefficient du modèle LPC utilisé dans le DFB, (ii) le seuil de détection d'activité vocale α, (iii) la taille des trames et nombre de coefficients MFSC utilisés dans la fonction REC et (iv) les seuils S_e et S_a employés pour détecter les voyelles, les améliorations apportées par les combinaisons optimales de ces paramètres se sont révélées minimes : le VER a été diminué d'environ 8% pour le corpus NTIMIT, de 5% pour TIMIT et de moins de 3% pour les corpus de parole affective ; les CER sont quant à eux soit restés inchangés, soit dégradés lors de l'optimisation des VER.

Table 2.9 Performances en détection des pseudo-phonèmes (V et C) selon les styles de production du corpus Bute-TMI.

Taux	Colère	Peur	Joie	Tristesse	Dégoût	Surprise	Nerveux	Neutre
V_{DET}	79.5	78.8	82.8	78.6	79.6	79.3	76.2	80.2
V_{INS}	9.0	15.1	8.9	13.9	16.5	10.9	10.2	10.2
VER	**29.6**	**36.3**	**26.1**	**35.3**	**36.9**	**31.6**	**34.0**	**30.0**
C_{DET}	77.8	70.9	73.4	70.0	70.6	72.2	70.4	68.6
C_{INS}	46.7	43.9	40.7	43.1	54.0	41.6	33.0	34.0
CER	**68.9**	**73.0**	**67.3**	**73.2**	**83.4**	**69.4**	**62.6**	**65.4**

V/C_{DET} : taux de détection en pourcentage des voyelles / consonnes ; V/C_{INS} : taux d'insertion en pourcentage des voyelles / consonnes ; V/C ER : taux d'erreur en pourcentage des voyelles / consonnes.

Table 2.10 Performances en détection des pseudo-phonèmes (V et C) selon les styles de production du corpus Aholab.

Taux	Colère	Peur	Joie	Tristesse	Dégoût	Surprise	Neutre
V_{DET}	80.9	78.7	85.7	84.3	82.2	82.4	81.9
V_{INS}	4.0	4.1	8.4	8.0	8.7	5.8	7.1
VER	**23.2**	**25.4**	**22.7**	**23.8**	**26.5**	**23.4**	**25.3**
C_{DET}	76.4	69.0	80.1	75.7	82.6	74.3	72.6
C_{INS}	48.6	36.3	68.2	54.2	78.2	49.7	52.6
CER	**72.1**	**67.4**	**88.1**	**78.5**	**95.7**	**75.4**	**80.0**

V/C_{DET} : taux de détection en pourcentage des voyelles / consonnes ; V/C_{INS} : taux d'insertion en pourcentage des voyelles / consonnes ; V/C ER : taux d'erreur en pourcentage des voyelles / consonnes.

L'influence des régions dialectales des corpus TIMIT et NTIMIT sur les taux d'erreurs VER et VER sont très minimes. Le système de détection des pseudo-phonèmes apparaît donc comme indépendant au changement de dialecte de l'anglais américain, que ce soit pour la parole lue en qualité laboratoire ou téléphonique. L'influence des variabilités apportées par l'affect sur le système est par contre plus conséquente que pour le changement de langue, cf. table 2.8-10. Ces variations sont flagrantes sur les corpus Berlin et Bute-TMI et moins importantes sur le corpus Aholab (une seule locutrice). Le taux d'erreur VER est par exemple très faible sur le corpus Berlin pour le style affectif « *Joie* » (le taux de détection vaut 100%) alors que la *Peur* amène un VER supérieur au double de celui obtenu sur l'émotion précédente ; cette émotion entraîne par ailleurs un taux d'erreur élevé en détection de voyelles sur l'ensemble des corpus affectifs étudiés. Notons que contrairement à ce que l'on pourrait attendre, le style « *Neutre* » ne conduit pas à un très bon taux d'erreur, ce dernier est même plutôt mauvais comparé aux styles affectifs issus des corpus Berlin et Bute-TMI, et dans la moyenne pour le corpus Aholab. Ces expériences montrent que le contexte affectif peut faciliter ou perturber la détection des pseudo-phonèmes selon l'émotion, et que le style « *Neutre* » contient un style de production différent de celui de la parole lue, puisque les taux d'erreurs sont bien meilleurs sur ce type de corpus (e.g., TIMIT et NTIMIT). Ce résultat conforte donc l'aspect imprécis ou incompris de la notion de neutralité.

Un critère de durée a été introduit dans le calcul des taux d'erreur pour évaluer plus précisément les performances du système de détection de pseudo-phonèmes. Cette analyse per-

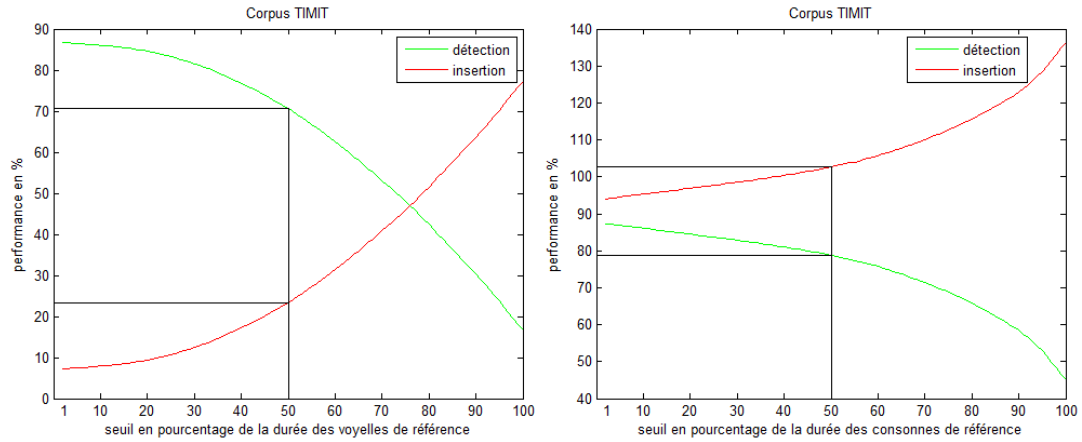


Fig. 2.19 Taux de détection et d'insertion des voyelles / consonnes du corpus TIMIT en fonction du seuil en pourcentage de la durée des segments de référence.

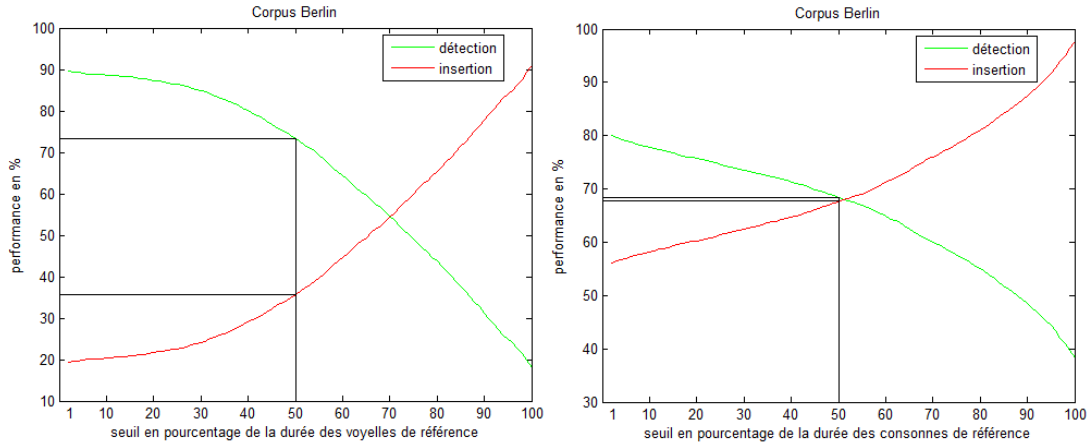


Fig. 2.20 Taux de détection et d'insertion des voyelles / consonnes du corpus Berlin en fonction du seuil en pourcentage de la durée des segments de référence.

met de connaître la proportion de durée des phonèmes qui ont été détectés par le système. Dans cette expérience, les pseudo-phonèmes doivent donc recouvrir une portion minimale des segments de référence pour être validés. Cette portion varie en pourcentage de la durée des transcriptions phonétiques. Les courbes obtenues pour les taux de détection et d'insertion sont présentées uniquement pour les corpus Berlin et TIMIT, cf. Fig. 2.19-20. Nous avons fixé pour l'analyse un seuil à 50% de la durée des segments phonétiques (valeur empirique). Les résultats de ces expériences montrent que le taux d'erreur augmente sensiblement lorsque la durée des voyelles détectées doit recouvrir au moins la moitié de celles issues des transcriptions : TIMIT, VER = +22% ; NTIMIT, +11% ; Berlin, +33% ; Bute-TMI, +33% et Aholab, +24%. Toutefois, les voyelles restent toutes détectées à plus de 70%, excepté pour le corpus Bute-TMI. Par ailleurs, la proportion de voyelles détectées à au moins 100% de la durée des segments de référence est relativement élevée : 1/3 pour le corpus NTIMIT, 1/4 pour les corpus Bute-TMI et Aholab, et 1/5 pour TIMIT et Berlin.

Nous présentons dans la table 2.11 une comparaison des résultats issus de la littérature en détection automatique de voyelles pour différents types de corpus de parole. Un astérisque est placé sur les corpus contenant de la qualité téléphonique (e.g., MULTEXT et NTIMIT). Les résultats que nous présentons ainsi que ceux qui ont été obtenus par F. Pellegrino *et al.* sont

Table 2.11 Comparaison des résultats obtenus par divers auteurs sur une tâche de détection de voyelles.

Référence	Corpus	Langue	Type de Parole	V_{DET}	V_{INS}	VER
Pfitzinger et al. [PFI96]	PhonDatII	Allemand	Lue	x	x	**12,9**
	Verbmobil	Allemand	spontanée	x	x	**21,0**
Fakotakis et al. [FAK96]	TIMIT TRAIN	Anglais	Lue	x	x	**32.0**
Pfau et Ruske [PFA98]	Verbmobil	Allemand	spontanée	x	x	**22,7**
Howitt [HOW00]	TIMIT TRAIN	Anglais	Lue	x	x	**29,5**
Pellegrino et al. [PEL98]	MULTEXT* DEV	Coréen	spontanée	93,1	15,3	**22,2**
		Espagnol	spontanée	93,4	9,9	**16,5**
		Français	spontanée	95,7	10,8	**15,1**
		Japonais	spontanée	94,3	9,6	**15,3**
		Vietnamien	spontanée	95,9	16,8	**20,9**
Ringeval et al. [RIN08ac]	TIMIT TRAIN+TEST	Anglais	lue	86,8	7,1	**20,3**
	NTIMIT* TRAIN+TEST	Anglais	lue	78,0	4,8	**26,9**
	Aholab	Basque	émotions actées	82,3	6,6	**24,3**
	Berlin	Allemand	émotions actées	90,0	19,1	**29,0**
	Bute-TMI	Hongrois	émotions actées	79.4	11.7	**32.3**

donnés pour une configuration unique du détecteur. Le système qui a été développé a permis de diminuer le VER de l'ordre de 34% par rapport aux études précédemment réalisées sur le corpus TIMIT (29.5% → 19.5% pour des paramètres optimisés). De plus, ces résultats ont été obtenus sur toutes les données du corpus TIMIT alors que ceux de la littérature n'ont été fournis que sur le sous-ensemble d'apprentissage (TRAIN). La table 2.11 montre aussi que la tâche de détection automatique des pseudo-voyelles est relativement indépendante de la langue et des conditions d'acquisition, i.e., parole lue, spontanée et téléphonique. Néanmoins, les performances en détection se dégradent de façon notable lorsque le système est confronté à de la parole affective (introduit des modifications dans le signal), ce qui est également le cas pour les systèmes de reconnaissance de la parole [ROT05][14], [SCH06a][15] et [STE10][16].

4.3. Corrélats phonétiques du « *p-centre* »

Une méthode a été récemment proposée pour extraire automatiquement l'enveloppe rythmique d'un signal de parole [TIL08b][29]. Cette méthode permet d'accéder au centre de perception de la parole alors appelé « *p-centre* », cf. sous-section 2.2.2. Différents niveaux de perception peuvent être utilisés pour localiser les proéminences du « *p-centre* », cf. sous-section 3.3. Afin de caractériser la nature acoustique de ces segments, nous avons calculé leur taux de recouvrement, en pourcentage de la durée, avec les autres unités de la parole, i.e., les phonèmes, les pseudo-phonèmes, et les segments voisés et non-voisés, cf. Fig. 2.21. Dans ces

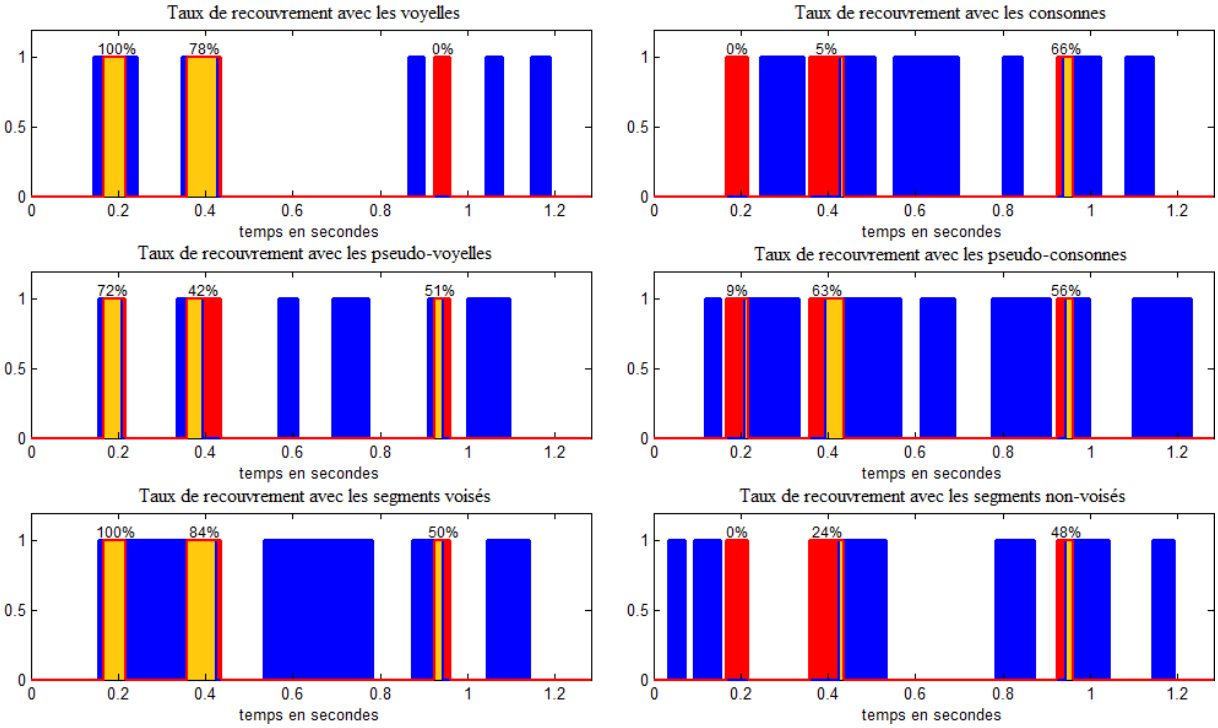


Fig. 2.21 Taux de recouvrement (en orange) des « *p-centres* » (en rouge, seuil = 1/3) avec les ancrages acoustiques (en bleu) extraits sur une phrase du corpus Berlin (groupe de phonèmes identifiés par SAMPA[1] pour les voyelles et les consonnes de référence).

expériences, les phonèmes sont fournis par les transcriptions contenues dans les données tandis que les pseudo-phonèmes sont détectés par le système présenté dans la sous-section 3.1, cf. Fig. 2.9. Les segments voisés sont obtenus quant à eux par la segmentation de la fréquence fondamentale (cf. Fig. 2.6) et les segments non-voisés par un seuillage de l'énergie.

Les tests ont été effectués pour différentes configurations d'analyse (e.g., changement de langue, de dialecte et de style affectif). Des mesures statistiques telles que la moyenne et l'écart-type ont été calculées pour étudier l'influence des catégories d'informations (e.g., région dialectale, émotion) sur les taux de recouvrement des « *p-centres* » avec les ancrages acoustiques. Ces mesures sont présentées à travers les différents niveaux de perception utilisés pour localiser les « *p-centres* », cf. table 2.12. Rappelons que le seuil le plus élevé, i.e., seuil = 1/3, conduit à se placer sur le sommet de la proéminence rythmique, alors que les deux autres seuils, i.e., 1/4 et 1/6, intègrent plus d'information sur la structure de la proéminence, cf. Fig. 2.16.

Les résultats montrent que les centres de perception de la parole sont essentiellement et uniformément composés de segments voisés ; la participation des segments non-voisés est anecdotique puisque le taux de recouvrement est strictement inférieur à 10%. Cela laisse donc supposer que les structures voisées de la parole correspondent au support préférentiel des ancrages rythmiques, et ce quel que soit le type de discours analysé ; parole lue et émotions actées en l'occurrence. Les résultats montrent également que la contribution des segments pseudo-phonétiques sur la structure acoustique des « *p-centres* » est similaire à celle des phonèmes, notamment si l'on considère le fait que l'algorithme DFB à tendance à sur-segmenter les consonnes ; notons l'exception du corpus NTIMIT pour lequel le niveau du rapport signal

Table 2.12 Taux de recouvrement des « *p-centres* » en % avec les autres types d'ancrage acoustique de la parole selon les différents corpus de parole étudiés.

Corpus	Seuil	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
TIMIT	**1/3**	**74 (15)**	14 (10)	64 (15)	28 (14)	**75 (23)**	2 (3)
	1/4	**74 (11)**	17 (9)	62 (12)	32 (13)	**79 (19)**	2 (3)
	1/6	**72 (9)**	21 (8)	59 (11)	38 (11)	**82 (16)**	4 (5)
NTIMIT	**1/3**	**72 (15)**	15 (11)	**79 (17)**	8 (10)	**70 (22)**	2 (3)
	1/4	**73 (12)**	17 (10)	**81 (13)**	10 (10)	**75 (18)**	2 (4)
	1/6	**71 (10)**	20 (8)	**80 (10)**	13 (9)	**77 (15)**	4(5)
Berlin	**1/3**	48 (20) / 68 (17)	18 (14) / 19 (14)	64 (15)	28 (14)	**70 (22)**	4 (7)
	1/4	44 (16) / 65 (15)	23 (13) / 24 (14)	62 (13)	33 (12)	**74 (18)**	6 (7)
	1/6	39 (13) / 61 (13)	29 (10) / 30 (12)	58 (10)	38 (10)	**75 (16)**	9 (8)
Bute-TMI	**1/3**	67 (16)	24 (13)	60 (18)	31 (17)	**76 (26)**	3 (7)
	1/4	66 (13)	28 (12)	60 (14)	34 (14)	**81 (18)**	5 (8)
	1/6	62 (12)	33 (11)	58 (12)	37 (12)	**81 (14)**	8 (9)
Aholab	**1/3**	64 (11)	30 (11)	**73 (11)**	22 (11)	**80 (20)**	1 (2)
	1/4	62 (10)	33 (10)	**72 (9)**	24 (9)	**82 (18)**	2 (3)
	1/6	59 (8)	36 (8)	69 (8)	29 (8)	**83 (17)**	3 (3)

[Valeur] (écart-type) ; les deux valeurs données pour le corpus Berlin correspondent respectivement à une configuration correspondant soit aux phonèmes identifiés par l'alphabet SAMPA[1] soit tous les phonèmes contenus dans les transcriptions.

sur bruit est faible. Les taux de recouvrement des « *p-centres* » avec les unités phonétiques montrent que les voyelles sont les sources principales des structures acoustiques des ancrages rythmiques. Les segments consonantiques interviennent de façon non négligeable, notamment comparé aux segments non-voisés, et de façon d'autant plus marquée lorsque la parole produite contient des émotions. Cela démontre ainsi l'existence d'un lien entre la structure acoustique des « *p-centres* » et la nature affective de la parole, lien qui peut être mis en corrélation avec les études qui ont montré les effets de l'affect dans le style de production de la parole [SCH86][37], [BUL05][38].

L'ensemble des résultats obtenus sur les taux de recouvrement des « *p-centres* » avec les autres types d'ancrages acoustiques de la parole doit tenir compte des caractéristiques de ces derniers, cf. table 2.13. Les consonnes sont par exemple bien plus nombreuses que les voyelles et cela est d'autant plus le cas pour les pseudo-phonèmes puisque les consonnes sont sur-segmentées par le DFB. Les segments non-voisés sont également plus nombreux que les segments voisés. Mais la durée des voyelles, des pseudo-voyelles et des segments voisés est, en contrepartie plus importante que celle de leur unité complémentaire, i.e., consonne, pseudo-consonne et segments non-voisés respectivement. Enfin, précisons que de nombreuses études ont montré que la durée des phonèmes est liée aux variabilités affectives de la parole [LEE04][8], [BUL05], [KIS10][46], [RIN08c][39] et [GOU10][47], cf. Fig. 1.10.

Les tables A.2.1-3 (en annexe) présentent donc les détails des taux de recouvrement des « *p-centres* » avec les autres types d'ancrage de la parole selon les catégories d'informations affectives. Les résultats montrent que les variations liées aux régions dialectales des corpus TIMIT et NTIMIT n'influencent pas la structure acoustique des ancrages rythmiques de la

Table 2.13 Caractéristiques des ancrages acoustiques extraits sur les corpus de parole étudiés : proportion en voyelle / consonne, pseudo-voyelle / pseudo-consonne et segments voisés / non voisés, et durée segmentale en millisecondes.

Corpus	TIMIT	NTIMIT	Berlin	Bute-TMI	Aholab
%V / %C	41 / 59	41 / 59	34 / 67 – 37 / 63	37 / 63	49 / 41
Durée voyelle	96.0 (48.7)	96.0 (48.7)	68.6 (38.9) – 88.4 (50.6)	74.5 (44.4)	70.6 (36.6)
Durée consonne	63.0 (36.5)	63.0 (36.5)	51.9 (36.8) – 59.6 (60.6)	66.8 (38.2)	66.9 (35.9)
% P-V / %P-C	26 / 74	51 / 49	32 / 68	42 / 58	40 / 60
Durée p-voyelle	82.9 (33.5)	126.8 (61.1)	78.6 (31.4)	84.0 (37.3)	81.0 (28.6)
Durée p-consonne	54.4 (27.5)	74.6 (43.1)	54.4 (36.9)	58.0 (29.8)	51.2 (22.8)
% SV / %SNV	37 / 63	33 / 67	45 / 55	63 / 27	40 / 60
Durée s. voisé	154 (163)	129 (150)	218 (193)	209 (149)	162 (199)
Durée s. non-voisé	89.8 (77.2)	51.6 (36.7)	100 (101)	42.5 (54.1)	91.6 (103)
Durée « *p-centres* » 1	97.6 (58.2)	95.3 (56.9)	76.8 (56.3)	82.3 (60.5)	59.6 (37.6)
Durée « *p-centres* » 2	109 (69.1)	106 (66.9)	91.0 (74.5)	93.9 (70.4)	75.7 (50.5)
Durée « *p-centres* » 3	127 (88.8)	121 (83.3)	122 (110)	115 (86.1)	112 (81.4)

[valeur moyenne] et (écart-type) ; les deux valeurs données pour le corpus Berlin correspondent respectivement à une configuration correspondant soit aux phonèmes identifiés par l'alphabet SAMPA[1] soit tous les phonèmes contenus dans les transcriptions.

parole, puisque les taux de recouvrement sont quasi-identiques à travers les régions (les tables ne sont donc pas données). En revanche, les écarts observés sur les corpus affectifs sont une nouvelle fois beaucoup plus significatifs, notamment sur les corpus Berlin (cf. table A.2.1) et Aholab (cf. table A.2.3). L'émotion de la *Peur* induit par exemple des structures acoustiques dans lesquelles les ancrages de type « consonne » et « non-voisé » sont majoritairement plus présents que sur les autres émotions. Ces différences sont moins flagrantes sur les ancrages de type « pseudo-consonne », puisque leur détection n'a pas été optimisée.

5. Conclusion

Les différents types d'ancrages acoustiques existants dans la parole ont été présentés. Ces segments servent de support à l'encodage (ou l'actualisation) des informations dans la parole et peuvent être de nature linguistique (e.g., phonèmes et syllabes) ou liés à des phénomènes de perception de la parole (e.g., segment voisés et non-voisés, pseudo-phonèmes et « *p-centre* ») [CHE09b][48]. L'identification robuste des supports sur lesquels sont encodées les informations affectives conditionne la tâche de reconnaissance de ces dernières. Les contraintes s'opposant à leur identification sont nombreuses et liées aux sources de variabilités naturelles du signal de parole : âge, genre, origine sociogéographique, état physique et affectif du locuteur, etc.

[46] G. Kiss et J. van Santen, "Automated vocal emotion recognition using phoneme class specific features", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010.

[47] M. Goudbeek et M. Broersma, "*Language specific effects of emotion on phoneme duration*", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010.

[48] M. Chetouani, A. Mahdhaoui et F. Ringeval, "Time-scale feature extractions for emotional speech characterization", dans *Cognitive Comp.*, Springer Verlag, vol. 1, no. 2, pp. 194–201, 2009.

Des méthodes permettant d'identifier automatiquement les points d'ancrage de la parole (e.g., vocalique, consonantique, et rythmique) ont été décrites dans ce chapitre. Ces méthodes ont été testées sur divers corpus, via les transcriptions phonétiques, pour étudier leur robustesse face à différentes configurations d'analyse. Les structures acoustiques associées aux « *p-centres* » ont aussi été caractérisées par les autres points d'ancrages issus des données, i.e., phonèmes, pseudo-phonèmes et segments voisés et non-voisés. Les expériences en détection de voyelles ont montré que le système amène un très bon VER sur le corpus de parole lue TIMIT ; les résultats issus de la littérature ont été nettement améliorés (VER = 19.5% contre 29.5%). Les variations des scores à travers les dialectes de l'Anglais Américain sont minimes. Alors que la qualité téléphonique (corpus NTIMIT) ou l'affect (corpus Berlin, Bute-TMI et Aholab) occasionnent une dégradation notable des performances du VER et du CER, même si le taux de détection des voyelles reste proche d'environ 80% sur tous les corpus.

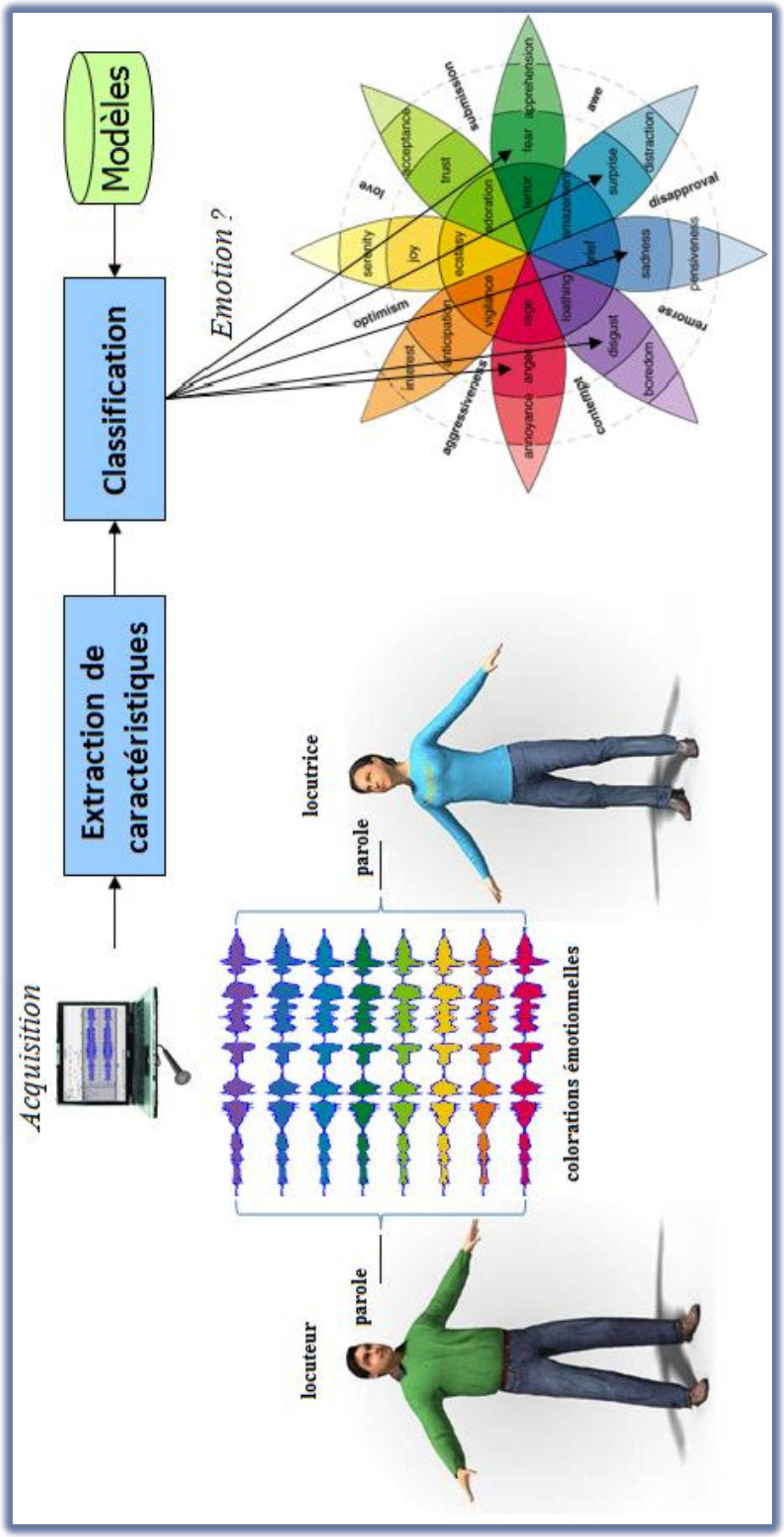
Les expériences conduites sur les structures acoustiques associées aux « *p-centres* » montrent que ces dernières sont clairement de nature voisée. Bien que les voyelles soient la source principale de ces structures, la participation des consonnes est loin d'être négligeable. En effet, les contributions consonantiques sont plus marquées sur les corpus de parole affective que lue, ce qui laisse supposer l'existence d'un lien entre la structure acoustique des « *p-centres* » et la nature affective de la parole. Notons qu'il a ainsi été montré la présence de nombreuses modifications dans le style de production lorsque ce dernier intervient en contexte émotionnel [SCH86][37].

Maintenant que nous avons déterminé l'influence des variabilités affectives de la parole sur les systèmes permettant d'en détecter les points d'ancrage, nous allons pouvoir exploiter ces derniers pour extraire des caractéristiques et effectuer la reconnaissance des émotions. Nos travaux portent sur l'étude de données produites par des acteurs adultes (cf. chapitre 3 – paramètres acoustiques et 4 – paramètres prosodiques), ou de façon spontanée par des enfants présentant des troubles de la communication, cf. chapitre 5 (paramètres prosodiques).

Chapitre 3

Reconnaissance acoustique de la parole affective actée

Les causes des phénomènes observables sur le signal de parole sont multiples et peuvent être corrélées entre elles, e.g., états physique et affectifs du locuteur, âge, genre, langue, origine sociogéographique, etc. L'influence de l'affect sur la tâche d'identification automatique des ancrages phonétiques et rythmiques de la parole a par exemple été évoquée dans le chapitre précédent. Les différents types de supports temporels que nous avons introduits vont maintenant servir à extraire les caractéristiques acoustiques corrélées à l'affect. L'objectif recherché est d'évaluer, via des techniques de fusion d'informations, la contribution de divers contextes d'analyse du signal de parole pour la reconnaissance des émotions. Notre approche inclut notamment : (i) des points d'ancrages complémentaires de la parole, e.g., voyelle et consonne, (ii) deux méthodes de classification, e.g., *k*-ppv et MMG, et (iii) deux types de décisions, e.g., « *phrase* » et « *segmental* ». Les modèles utilisés dans ce chapitre pour décrire les caractéristiques acoustiques du signal de parole reposent sur les coefficients cepstraux *mel frequency cepstrum coding* (MFCC). Cette approche s'inscrit donc dans la continuité des systèmes classiques de reconnaissance de la parole puisque les paramètres MFCC ont été originellement proposés pour ces systèmes. Cependant, comme ces données sont fortement dépendantes du locuteur, une étape de normalisation doit être effectuée pour limiter l'influence du style de parole sur les phénomènes associés aux expressions émotionnelles.



Représentation graphique des différentes étapes nécessaires à la reconnaissance des émotions de la parole.

1. Introduction

Les premières expériences portant sur la reconnaissance automatique de la parole datent des années cinquante [DAV52][1]. Un ensemble restreint de mots était alors reconnu au moyen d'un circuit électronique analogique. L'avènement des méthodes de calcul numérique dans les années quatre-vingt-dix a permis au domaine du TAP de prendre son véritable essor. Les travaux de [JEL75][2] et de [DAV80][3], particulièrement cités dans la littérature, marqueront cette époque. Les méthodes proposées possédaient l'architecture des systèmes de reconnaissance actuels : une étape d'acquisition du signal, suivie d'un module de codage (ou d'extraction de caractéristiques) et d'un classifieur, cf. Fig. 3.1. L'étape d'extraction de caractéristiques occupe une place fondamentale dans la chaîne du TAP. Les contributions de cette thèse portent essentiellement sur cette étape.

Fig. 3.1 Schéma générique d'un système de reconnaissance de la parole.

L'objectif du module d'extraction de caractéristiques consiste à rendre compatible les informations extraites avec la tâche de classification. Le classifieur, dont le rôle est d'attribuer une classe (ou une catégorie) à un ensemble de paramètres, doit en effet disposer de données en faible quantité et surtout, à fort pouvoir discriminant pour mener à bien sa tâche de classification. Comme le signal de parole contient beaucoup d'informations redondantes, ses données doivent être résumées au moyen de descripteurs adaptés à la tâche considérée, i.e., fournissant des descriptions discriminantes des catégories d'informations analysées.

Concernant la reconnaissance des émotions, de très nombreux systèmes utilisent comme vecteur principal de caractéristiques les coefficients cepstraux MFCC. Bien qu'originellement destinés à la reconnaissance de la parole, ces paramètres se sont vus adaptés avec succès à bien d'autres types de tâches telles que la reconnaissance du locuteur, de ses intentions ou encore de ses émotions. Les coefficients « passe-partout » MFCC incluent en effet une description à la fois compacte et robuste au bruit des caractéristiques spectrales du signal de parole. Nous présentons dans la section suivante les techniques de traitement du signal qui permettent d'extraire automatiquement les coefficients MFCC dans un signal de parole.

---

[1] K. H. Davis, R. Biddulph et S. Balashek, "Automatic recognition of spoken digits", dans *The J. of the Acoust. Soc. of Amer.*, vol. 24, pp. 637–642, 1952.

[2] F. Jelinek, L. Bahl et R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech", dans *IEEE Trans. on Information Theory*, vol. 21, p. 250–256, 1975.

[3] S. Davis et P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", dans *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

2. Modélisation acoustique de la parole

Les modèles acoustiques visent à caractériser, dans un cadre segmental, la distribution de l'énergie contenue dans le spectre du signal de parole. Le signal est pour cela préalablement segmenté en trames (cf. Fig. 3.2) pour une durée correspondant à l'approximation numérique du critère de stationnarité de la parole, i.e., 32ms. Un facteur de recouvrement de moitié est alors utilisé pour mieux représenter l'évolution du signal à travers ses trames.

Il existe de nombreuses méthodes pour caractériser la distribution spectrale du signal de parole. Une très forte majorité d'entre elles s'appuient sur le modèle source-filtre [FAN60][4]. Ce modèle présente le signal de parole comme la résultante du produit de convolution d'une source bimodale (i.e., présentant deux modes de fonctionnement) par un filtre représentant le conduit vocal [8]. La source est soit modélisée par un générateur d'impulsions dans le cas d'un son voisé, soit par un générateur de bruit blanc dans le cas d'un son non-voisé, cf. Fig. 3.3. Le conduit vocal est quant à lui modélisé par un filtre AR linéaire de type tout-pôle [8] :

$$s(k) = e(k) * a(k) \quad \text{et} \quad A(z) = \frac{1}{1 + \sum_{i=0}^{N} a_i z^{-i}} \qquad [8]$$

avec, e : signal d'excitation, s : signal de parole, a_i : coefficients du filtre AR, et N : ordre du filtre AR.

Les caractéristiques extraites dans les modèles de prédiction linéaire (e.g., codage LPC) correspondent aux coefficients a_i du filtre $A(z)$ modélisant le conduit vocal. Ces coefficients décrivent les valeurs des formants. Les pôles de la fonction de transfert modélisant le conduit vocal coïncident effectivement avec les maximums locaux du spectre. En raison de son faible coût calculatoire, le codage LPC a été très largement utilisé dans les applications nécessitant de transmettre en temps réel la parole, e.g., les télécommunications. Cependant, de nombreux travaux ont montré l'importance d'autres types de représentation spectrale que le filtre tout-pôle [KLE03][5]. Ces dernières passent par l'intégration de modèles perceptifs de la parole qui emploient, bien souvent, des techniques mathématiques permettant de séparer les informations fournies par la source et le conduit vocal.

Les coefficients MFCC font par exemple intervenir une étape de déconvolution entre la source et le filtre par un homomorphisme, i.e., un jeu d'écriture mathématique [9]. Cette opération transforme le produit de convolution en une simple somme, ce qui permet de filtrer par la suite les variabilités du signal produites par le conduit vocal dans l'objectif de caractériser celles issues de la source.

$$C(\tau) = DFT^{-1}[log(DFT[s(t)])] \qquad [9]$$

avec, s : signal de parole, $C(\tau)$: cepstre du signal, DFT : transformée discrète de Fourrier, DFT^{-1} : transformée discrète et inverse de Fourrier.

[4] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton [Eds], 1960.

[5] W. B. Kleijn, "Signal processing representations of speech", dans *IEICE Trans. on Information and Systems*, vol. E86-D, no. 3, pp. 359–376, Mar. 2003.

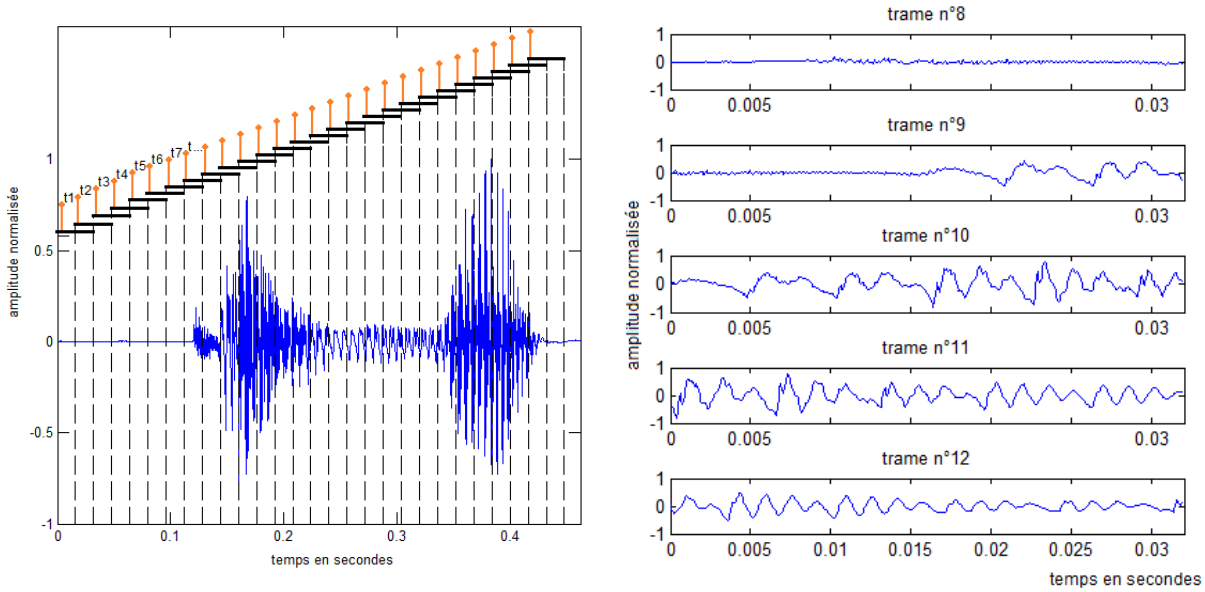


Fig. 3.2 Exemple de segmentation d'un signal de parole en trames.

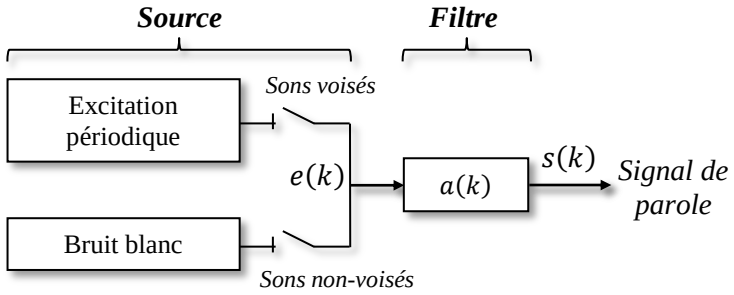
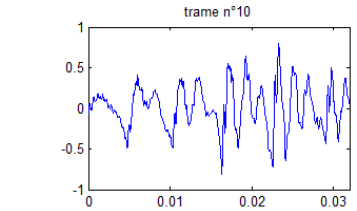


Fig. 3.3 Modèle source filtre du signal de parole.

L'intégration de cette étape dans le codage MFCC s'effectue conjointement avec une réduction de la dimension des données traitées. Le spectre en puissance du signal original est tout d'abord calculé puis analysé par un banc de filtres qui fournit un vecteur de coefficients correspondants à la moyenne des énergies spectrales issues de chaque filtre (première étape de réduction des données). Ces coefficients subissent ensuite une compression logarithmique et une transformation en cosinus discret (seconde étape de réduction des données). La Fig. 3.4 résume l'ensemble de ces étapes.

Le rôle du banc de filtre dans le codage MFCC est primordial car il permet d'une part de réduire fortement la quantité de données traitées, et d'intégrer d'autre part des connaissances en perception. Ainsi, les filtres utilisés pour le calcul des coefficients d'énergie spectrale sont répartis selon une échelle inspirée des propriétés psycho-acoustiques de l'oreille humaine. Cette échelle est graduée en Mel et correspond à une transformation non-linéaire de type logarithmique des fréquences f [10] :

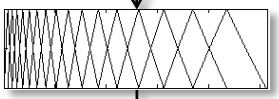
$$Mel(f) = 2595\ .\ log\left(1 + \frac{f}{700}\right) \qquad [10]$$



Trame d'un signal de parole

Spectre de puissance
$|DFT|^2$

Calcul du spectre de puissance



Filtrage perceptuel par un banc de filtres non-linéaires

Log

Compression logarithmique des coefficients d'énergie

DCT

Projection des coefficients par transformation en cosinus discret

Vecteur de caractéristiques MFCC

Fig. 3.4 Processus d'extraction des caractéristiques MFCC sur une trame de signal de parole.

La largeur de bande des filtres utilisés lors du codage varie également selon une échelle logarithmique. Cette échelle vise à reproduire l'effet de masque qui se manifeste lorsque deux sons purs de fréquences différentes sont perçus simultanément. L'influence réciproque de ces deux sonorités entraîne une dégradation du seuil d'audition, cf. Fig. 3.5. L'effet de masque, qui peut être total ou partiel, dépend des intensités et fréquences relatives des deux sonorités appelées son masqué et son masquant. En déterminant la bande de fréquence du son qui contribue au masquage du son pur, on obtient la largeur de bande dite bande critique. Cette bande correspond à l'écartement en fréquence nécessaire pour discriminer deux harmoniques issus d'un son périodique. L'échelle des $Barks$, qui détermine la largeur de bande des filtres du codeur MFCC, est définie par la largeur de la bande critique. Cette dernière vaut 100Hz jusqu'à 500Hz, et environ 20% de la fréquence fondamentale au-delà de cette limite.

Il est intéressant de noter l'importante réduction réalisée dans la quantité d'informations extraite lors du codage acoustique du signal de parole. En effet, le nombre de paramètres est généralement de 16 pour chaque trame extraite (i.e., toutes les 16ms), contre plus de 256 échantillons pour le signal d'origine échantillonné à la fréquence de F_e = 16kHz. Toutefois, une réduction encore bien plus importante de la quantité d'informations peut être obtenue par l'intermédiaire des caractéristiques prosodiques, puisqu'elles sont généralement fournies par un vecteur de paramètres calculés sur l'intégralité d'une phrase. De plus, les paramètres de la prosodie permettent de discriminer de nombreuses catégories d'informations issues de diverses tâches de TAP, e.g., reconnaissance du locuteur [REY03][6], de ses intentions [BRE02][7] ou de ses émotions [AUS05][8]. Ces paramètres seront donc étudiés dans le chapitre suivant.

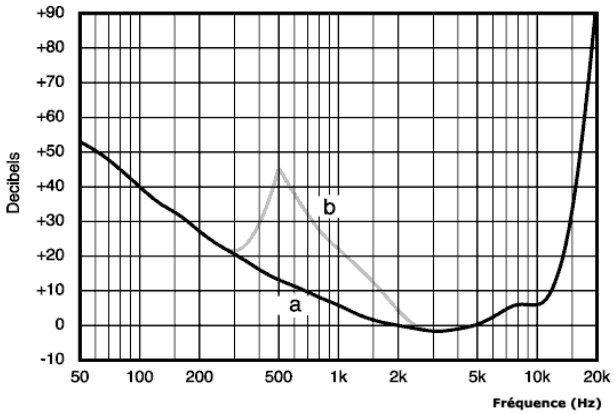


Fig. 3.5 Seuil d'audition selon les fréquences pour (a) aucun signal en entrée et (b) un signal fort de 70 dB SPL à 500 Hz ; figure extraite de [ZWI90][9].

De très nombreux travaux ont montré que les coefficients MFCC sont fortement liés au locuteur, puisque d'excellents scores de reconnaissance ont été obtenus sur de grandes bases de données pour différents contextes d'analyse, e.g., campagnes d'évaluations NIST. Bien que cette propriété soit pertinente pour les systèmes de reconnaissance du locuteur, elle l'est beaucoup moins pour la reconnaissance des émotions. L'influence des différentes sources de variabilité du signal de parole doit en effet être maîtrisée pour ne pas biaiser les mesures réalisées quant aux contributions de l'affect. Par conséquent, les systèmes de reconnaissance d'émotions qui exploitent les paramètres MFCC doivent normaliser ces données. L'étape de normalisation des caractéristiques acoustiques C_a est bien souvent réalisée par le calcul du *z-score* qui fait intervenir les moments statistiques d'ordre 1 et 2 des données [11].

$$C'_a = \frac{C_a - \mu}{\sigma} \qquad [11]$$

3. Système de reconnaissance

Nous présentons dans les paragraphes suivants le système qui a été développé pour la reconnaissance acoustique des émotions. Ce système s'appuie sur le codage MFCC qui a été présenté dans la section précédente. Un étalonnement des paramètres du codeur a été effectué afin d'identifier la configuration la plus pertinente pour le corpus Berlin. Cette étape a permis de faire ressortir la configuration suivante : des trames de 32ms ont été extraites toutes les 16ms avec une fenêtre de Hanning sur le signal de parole, et un jeu de 16 coefficients spectraux a été calculé sur chaque trame au moyen d'un banc de filtres triangulaires.

[6] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones et B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", dans proc. *ICASSP*, Hong Kong, China, pp. 784–787, 2003.

[7] C. Breazeal et L. Aryananda, "Recognition of affective communicative intent in robot-directed speech", dans *Autonomous Robots*, Kluwer Academic Publishers, vol. 12, no.1, pp. 83–104, Jan. 2002.

[8] A. Austermann, N. Esau, L. Kleinjohann et B. Kleinjohann, "Prosody based emotion recognition for MEXI", dans proc. *IROS*, Edmonton, Alberta, Canada, Aug. 2-6 2005, pp. 83–104.

[9] E. Zwicker et H. Fastl, *Psychoacoustics: Facts and models*, dans Springer-Verlag [Eds], Heidelberg, 1990.

3.1. Architecture

L'architecture du système proposé repose sur la fusion de différentes approches utilisées pour caractériser les coefficients MFCC, cf. Fig. 3.6. De nombreux travaux ont démontré l'intérêt des techniques de fusion d'informations dans les systèmes de reconnaissance [KUN-04][10]. La contribution des approches employées pour réaliser la reconnaissance des émotions peut par exemple être obtenue par une étape de fusion. Dans le cadre de notre étude, cette étape est importante puisque nous souhaitons quantifier les contributions apportées par différents points d'ancrage de la parole dans la tâche de reconnaissance d'émotions. De plus, les performances en fusion peuvent être améliorées si les informations exploitées conduisent à des descriptions complémentaires de l'affect. Toutefois, cette amélioration ne constitue pas le but principal de notre étude. Les différentes stratégies employées dans notre système de reconnaissance incluent : (i) une étape d'extraction des informations selon différents points d'ancrages acoustiques, (ii) des prises de décision tant au niveau de la phrase que du segment et (iii) l'emploi de deux types de classifieurs : *k*-ppv et MMG.

3.2. Décisions « *segmentale* » et « *phrase* »

Les techniques traditionnelles utilisées en reconnaissance d'émotion consistent à prendre une décision quant à l'émotion reconnue pour chaque trame t voisée du signal de parole. Des différences apparaissent toutefois sur la méthode à utiliser pour identifier l'émotion finale reconnue par le système, e.g., [VLA07][11] et [SHA07][12]. La décision quant à l'émotion reconnue sur chaque trame voisée du signal de parole s'écrit, dans le cadre bayésien, comme la probabilité *a posteriori* $p(E_i \mid C_a(t))$ de reconnaître l'émotion E_i ($i \in 1, 2, \ldots N_e$ émotions) sachant les caractéristiques acoustiques $C_a(t)$ observées sur la trame t analysée, i.e., les coefficients MFCC. La décision finale D_{phr} quant à l'émotion reconnue E^* est obtenue, dans le cadre d'une décision de type « *phrase* », ou « *turn* » [VLA07], par l'émotion $E^*_{D_{phr}}$ qui ressort majoritaire (fonction arg max) dans la moyenne des N_t vecteurs de probabilités *a posteriori* $p(E_i \mid C_a(t))$ fournit par l'étape de classification de chaque trame t voisée [12]. Cette approche correspond à une décision de type « *phrase* » puisque les émotions reconnues sur chaque trame du signal de parole sont moyennées à travers la phrase sans faire intervenir de coefficient de pondération. Par conséquent, ce type de décision permet d'obtenir un jugement général sur les caractéristiques acoustiques $C_a(t)$ observées à travers les trames t du signal de parole.

Afin de tenir compte de la durée de production des segments de parole, lors de la décision finale prise quant à l'émotion reconnue, Shami et Verhelst [SHA07][12] ont proposé d'intégrer

[10] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley & Sons Inc. [Eds], Hoboken, New Jersey, Jul. 2004.

[11] B. Vlasenko, B. Schuller, A. Wendemuth et G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing", dans proc. *2nd Inter. C. on Affective Comp. and Intel. Interaction*, Lisbon, Portugal, pp. 139–147, Sep. 12-14 2007.

[12] M. Shami et W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", dans *Speech Comm.*, vol. 49, no. 3, pp. 201–212, Mar. 2007.

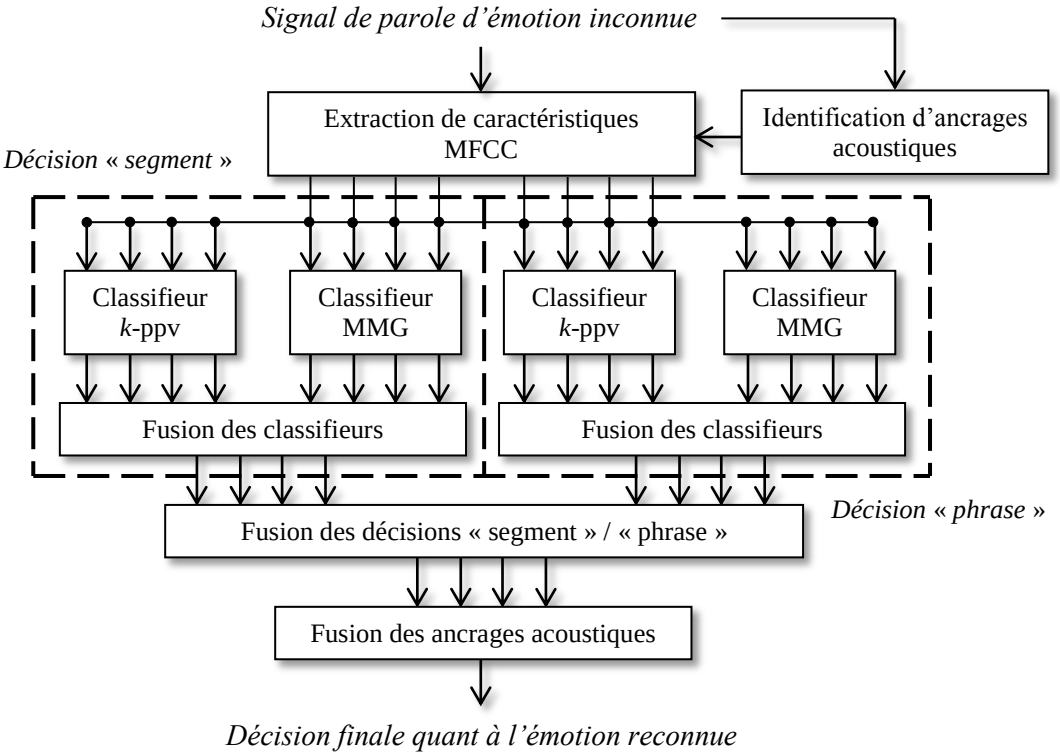


Fig. 3.6 Architecture du système de reconnaissance acoustique de la parole affective.

$$E^*_{D_{phr}} = \underset{1 \le i \le N_e}{\arg\max} \left\{ \frac{1}{N_t} \sum_{t=1}^{N_t} p\big(E_i \mid C_a(t)\big) \right\} \quad [12]$$

$$E^*_{D_{seg}} = \underset{1 \le i \le N_e}{\arg\max} \left\{ \sum_{t=1}^{N_t} Seg_x \,.\, p\big(E_i \mid C_a(t)\big) \right\} \quad [13]$$

les informations liées à la durée dans l'équation [12]. Ainsi, au lieu de considérer équiprobable les émotions reconnues sur les trames t, i.e., le coefficient $1/N_t$, les probabilités *a posteriori* sont pondérées par un coefficient variable Seg_x qui dépend de la durée du segment voisé x analysé [13]. En conséquence, l'approche « *segmentale* » accorde une importance aux probabilités qui est proportionnelle à la durée des segments voisés sur lesquelles elles ont été extraites. Une proéminence dans la durée d'un segment se traduit par une contribution plus importante de ces probabilités *a posteriori* associées lors du calcul de la décision finale $E^*_{D_{seg}}$.

3.3. Classifieurs

L'étape finale de tout système de reconnaissance (cf. Fig. 3.1) consiste à attribuer une catégorie d'informations, e.g., l'émotion E_i, à un ensemble de caractéristiques mesurables, e.g., les paramètres acoustiques $C_a(t)$. Cette étape de classification consiste à présenter au système des données inconnues sur lesquelles la décision s'opère soit : (i) par les informations collectées lors d'une phase d'apprentissage (approche *supervisée*) ou (ii) par des techniques de re-

groupement ou de classification automatique (*clustering*, approche *non-supervisée*). Ces techniques proviennent du domaine de la reconnaissance des formes (RDF) [DUD00][13] et font apparaître deux familles de classifieurs : *génératifs* et *discriminants*.

Les classifieurs *génératifs* (e.g., analyse linéaire discriminante, discriminant de Fisher, modèles de mélanges gaussiens, et classifieur bayésien naïf) estiment les probabilités conditionnelles $p(C_a(t)|E_i)$ par lesquelles celles dites *a posteriori* peuvent être obtenues au moyen de la règle de Bayes. Ces classifieurs exploitent alors des méthodes statistiques dans le but d'estimer l'appartenance aux classes des données testées, cf. Fig. 3.7.

Les classifieurs dits *discriminants* (e.g., perceptron et ses variantes multicouches – MLP, machines à vecteur support – SVM) cherchent, quant à eux, à définir le meilleur moyen de séparer les données à travers leur espace de représentation et selon leur classe associée, cf. Fig. 3.8. Des techniques sont notamment exploitées pour minimiser les erreurs en classification (e.g., critère MCE [JUA92][14]) et rétro-propager ces dernières sur les paramètres du système de reconnaissance (via des fonctions de coût) de façon à optimiser ses performances dans la tâche réalisée. On considère souvent que les classifieurs *génératifs* sont plus adaptés à l'analyse de corpus contenant peu de données. Les classifieurs *discriminants* reposent en effet, sur des méthodes d'optimisation des performances qui pourraient conduire de façon relativement aisée à de très bons résultats sur un ensemble de données plutôt restreint.

Nous décrivons dans les paragraphes suivants deux classifieurs qui sont couramment associés aux coefficients MFCC pour effectuer la reconnaissance des émotions : (i) l'algorithme des *k*-plus-proches-voisins (*k*-ppv) et (ii) les modèles de mélanges gaussiens (MMG). En raison de son faible degré de complexité calculatoire (distance L1 entre les paramètres de l'exemple testé et ceux issus de l'apprentissage), la méthode des *k*-ppv est très souvent utilisée pour obtenir une estimation à la fois rapide et fiable des performances d'un système de reconnaissance. Ce classifieur permet d'arriver à un compromis entre la complexité du système et le nombre de paramètres à traiter. Les MMG incluent quant à eux une modélisation statistique des données qui est adaptée à l'analyse des coefficients cepstraux MFCC, puisque ces données présentent une matrice de covariance qui a été diagonalisée lors du calcul de la DCT.

3.3.1. L'algorithme des *k*-plus-proches-voisins

La méthode des *k*-ppv consiste à comparer directement les caractéristiques mesurées au moyen d'une distance de norme L1. Cette distance est calculée entre le vecteur de paramètres extrait sur la phrase à tester et ceux qui ont été collectés lors de la phase d'apprentissage. Les informations de classe des exemples d'apprentissage correspondants aux k plus proches voisins de l'exemple testé sont ensuite analysées pour identifier celle qui est majoritaire. Ainsi, la probabilité *a posteriori* d'identifier l'émotion E_i sur le vecteur $C_a(t)$ s'écrit, au sens des *k*-ppv, comme le rapport du nombre de labels k_i ($i \in 1, 2, \ldots N_e$) alors retenus parmi les k plus proches voisins [DUD00][13] [14]. L'émotion reconnue E^* sur chaque trame t est obtenue par

[13] R. O. Duda, P. E. Hart et D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley & Sons Inc. [Eds], 2000.

[14] B. H. Juang et S. Katagiri, "Discriminative learning for minimum error classification", dans *IEEE Trans. on Signal Proces.*, vol. 40, pp. 3043–3054, Dec. 1992.

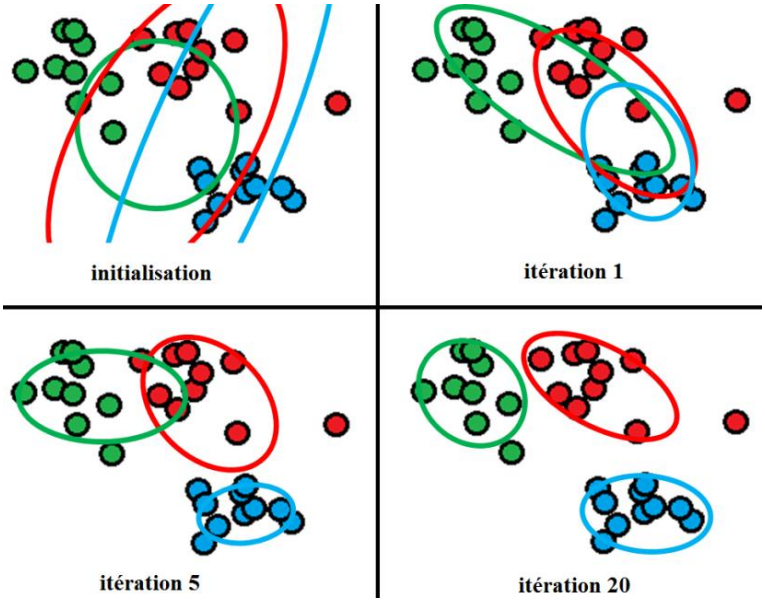


Fig. 3.7 Illustration des itérations réalisées par l'algorithme EM lors de l'estimation des paramètres du classifieur *génératif* MMG ; les ronds représentent trois classes distinctes (e.g., rouge, vert et cyan), tandis que les ellipses associées aux classes renseignent sur les paramètres (i.e., moyenne et écart-type) des modèles MMG estimés par l'algorithme EM ; figure extraite de [PRE05][15].

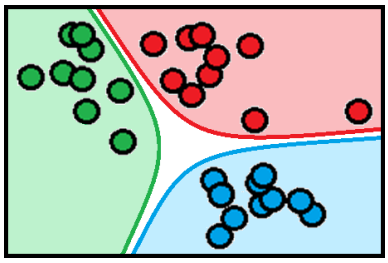


Fig. 3.8 Illustration des frontières de classes obtenues par le classifieur *discriminant* SVM dans la configuration « un contre tous ».

une fonction arg max sur les probabilités p_{k-ppv} [15]. La décision peut alors être prise dans le cadre d'une approche de type « *phrase* » [12], ou « *segmental* » [13].

La signification géométrique des classes retournées par la méthode des *k*-ppv correspond à la réunion des domaines d'influence des exemples de référence. La résolution spatiale des frontières est alors liée à la valeur de k : une faible valeur conduit à des frontières complexes tandis qu'une forte valeur lisse les frontières de classes, cf. Fig. 3.9. De nombreuses variantes de l'algorithme des *k*-ppv existent. L'une d'entre elles consiste à améliorer la précision de l'estimation des probabilités *a posteriori*. Dans cette variante, les k plus petites distances entre les exemples de référence et celui testé sont conservées pour chaque classe. Les k mesures de distance sont ensuite moyennées pour chaque classe et le vecteur obtenu est normalisé en un vecteur de probabilités. Cette variante de l'algorithme des *k*-ppv a été utilisée dans cette thèse de façon à avoir une précision adéquate pour les étapes de fusion. Enfin, notons qu'il est préférable de fixer une valeur de k impaire pour limiter la possibilité de votes non majoritaires.

[15] L. Prevost, *Cours de Reconnaissance des Formes*, Master 2 Sciences de l'Ingénieur, Université Pierre et Marie Curie, Paris 2005.

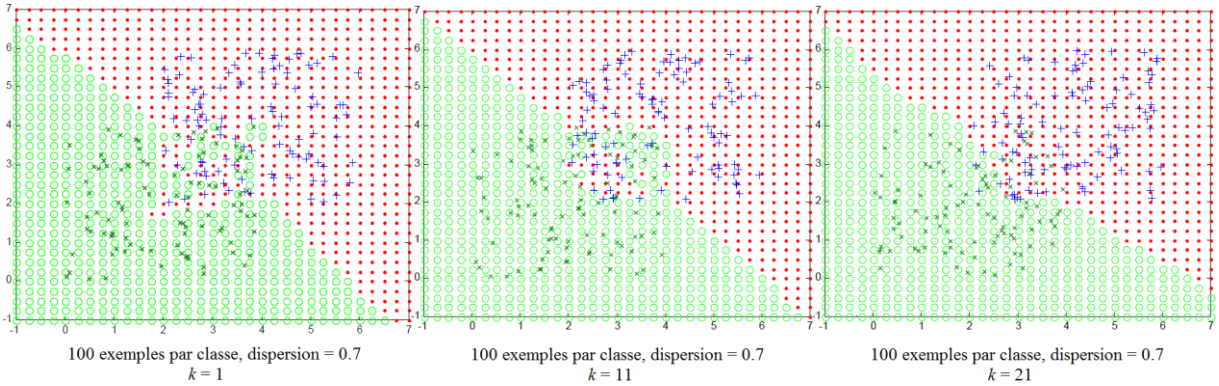


Fig. 3.9 Résolution spatiale des frontières servant à la classification de deux classes en fonction de la valeur de *k* ; figure extraite de [PRE05][15].

$$p_{k-ppv}\left(E_i \mid C_a(t)\right) = \frac{k_i}{k} \qquad [14]$$

$$E^* = \arg\max_{1 \le i \le N_e} \left(p_{k-ppv}\left(E_i \mid C_a(t)\right)\right) \qquad [15]$$

3.3.2. Les mélanges de modèles gaussiens

Les MMG caractérisent la distribution statistique d'un paramètre donné par une somme pondérée de M distributions gaussiennes multidimensionnelles, chacune définie par : (i) un vecteur moyen μ_i, (ii) une matrice de covariance Σ_i et (iii) un poids α_i. Le classifieur MMG est de type *génératif* puisqu'il permet de synthétiser des données ayant la même distribution statistique que celles qui ont été collectées dans la phase d'apprentissage. Il correspond à celui de l'état de l'art en reconnaissance acoustique des émotions.

Dans les MMG, chaque observation du signal de parole sur ses trames est définie par un vecteur de paramètres acoustiques de dimension d : $\varphi = (x_1, x_2, x_3, \dots, x_d)$; $d = 16$ coefficients MFCC. On note par $\Psi = \left(\varphi_1, \varphi_2, \varphi_3, \dots, \varphi_{n_p}\right)$ la suite des n_p vecteurs d'observations acoustiques extraits sur les trames. Pour chaque émotion, les paramètres μ_i, Σ_i et α_i du MMG sont appris à partir des observations Ψ et au moyen des algorithmes de quantification vectorielle *Linde-Buzo-Gray* (LBG) [LIN80][16] et de maximisation de l'espérance (EM) [DEM77][17], cf. Fig. 3.7. La probabilité d'observer φ_k sachant que l'émotion E_i est présente dans la trame analysée est définie par l'équation [16]. En faisant l'hypothèse que les observations φ_k sont indépendantes, la loi d'observation des vecteurs acoustiques $p(\Psi \mid E_i)$ peut s'écrire comme le produit des lois obtenues pour chaque vecteur acoustique φ_k [17]. L'émotion la plus probable E^* est définie, dans le cadre de l'approche bayésienne classique, par une fonction arg max sur les probabilités *a posteriori* $p(E_i \mid \Psi)$ [18]. La décision finale peut être prise selon l'approche « *phrase* » [12], ou « *segmentale* » [13].

[16] Y. Linde, A. Buzo et R. M. Gray, "An algorithm for Vector Quantizer design", dans *IEEE Trans. on Comm.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[17] A. Dempster, N. Laird et D. and Rubin, "Maximum likelihood from incomplete data via the EM algorithm", dans *J. of Acoust. Soc. of Amer.*, vol. 39, no. 1, pp. 1–38, 1977.

$$p_{MMG}(\varphi_k \mid E_i) = \sum_{j=1}^{Q_i} \frac{\alpha_j^i}{(2\pi)^{d/2}\sqrt{|\Sigma_j^i|}} \exp\left(-\frac{1}{2}(\varphi_k - \mu_j^i)^t (\Sigma_j^i)^{-1}(\varphi_k - \mu_j^i)\right) \quad [16]$$

avec, Q_i : le nombre de composantes du MMG, d : la dimension du vecteur de caractéristiques $\varphi_k$de la trame k et μ_j, Σ_j : paramètres de la loi gaussienne j.

$$p_{MMG}(\Psi \mid E_i) = \prod_{k=1}^{n_p} p(\varphi_k \mid E_i) \quad [17]$$

$$E^* = \arg\max_{1 \le i \le N_e}\left(p_{MMG}(E_i \mid \Psi)\right) \quad [18]$$

3.4. Fusion

L'architecture que nous proposons pour effectuer la reconnaissance des émotions fait apparaître différents étages de fusion, cf. Fig. 3.6. De nombreux travaux ont en effet montré l'importance de cette étape [MAH09][18], [POL09][19] et [LUE09][20]. L'intérêt de la fusion est double puisqu'elle permet non seulement de quantifier la contribution de plusieurs types d'informations exploitées dans la tâche de reconnaissance d'émotions, mais aussi d'améliorer les scores si les données fusionnées fournissent des descriptions complémentaires de l'affect. Toutefois, nous attachons ici plus d'importance à l'aspect qualitatif que quantificatif, des contributions apportées par les ancrages complémentaires de la parole, lors des étapes de fusion en reconnaissance d'émotions. Ces dernières s'opèrent sur trois niveaux dans l'architecture proposée, cf. Fig. 3.6 : (i) les classifieurs *k*-ppv et MMG, (ii) les décisions de type « *phrase* » et « *segmental* » et (iii) les points d'ancrages acoustiques complémentaires, e.g., segments voisés / non-voisés, phonèmes, pseudo-phonèmes et syllabes. Bien qu'il existe de nombreuses méthodes sophistiquées pour fusionner ces informations [KUN04][10] et [MON09][21], nous avons employé une somme pondérée par un coefficient α des probabilités *a posteriori* fournies par chaque étage du système de reconnaissance. Cette méthode présente l'avantage de quantifier directement la contribution des différents types d'informations exploitées par le système.

Les équations [19], [20] et [21] illustrent les fusions opérées sur les probabilités issues : (i) des classifieurs *k*-ppv et MMG (coefficient α_c) [19], (ii) des décisions de type « *segmentales* » et de type « *phrase* » (coefficient α_d) [20] et (iii) des ancrages acoustiques complé-

[18] A. Mahdhaoui, F. Ringeval et M. Chetouani, "Emotional speech characterization based on multi-features fusion for face-to-face interaction", dans proc. *3rd Inter. C. on Signals, Circuits and Systems*, Djerba, Tunisia, Nov. 6-8 2009, pp. 1–6.

[19] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner et F. Metze, "Emotion classification in children's speech using fusion of acoustic and linguistic features", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 340–343.

[20] I. Luengo, E. Navas et I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 332–335.

[21] E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy et J. Sole-Casals, "Maximum likelihood linear programming data fusion for speaker recognition", dans *Speech Comm.*, vol. 51, no. 9, pp. 820–830, Sep. 2009.

$$E^*_{F_c} = \arg\max_{1 \le i \le N_e} \left(\alpha_c * p_{k-ppv} + (1 - \alpha_c) * p_{MMG} \right) \quad [19]$$

$$E^*_{F_d} = \arg\max_{1 \le i \le N_e} \left(\alpha_d * p_{D_{seg}} + (1 - \alpha_d) * p_{D_{phr}} \right) \quad [20]$$

$$E^*_{F_a} = \arg\max_{1 \le i \le N_e} \left(\alpha_a * p_{voy} + (1 - \alpha_a) * p_{csn} \right) \quad [21]$$

mentaires, e.g., voyelles – voy et consonnes – csn, coefficient α_a [21]. Chaque étape de fusion retourne un nouvel ensemble de vecteurs de probabilités à l'étage suivant, cf. Fig. 3.6. L'étape finale réside dans la fusion des ancrages acoustiques complémentaires puisque nous souhaitons quantifier leur contribution respective dans la tâche de reconnaissance des émotions. Les coefficients de fusion α ont été obtenus par une étape de maximisation des performances. Ils représentent les poids attribués à chaque type d'informations exploitées lors de la phase de reconnaissance des émotions. Ces poids varient de 0 à 1 et ont été définis lors des mesures en généralisation (i.e., les tests) et non d'apprentissage. En effet, une optimisation des poids réalisée sur toutes les données, comme cela est traditionnellement effectué, ne permettrait pas d'obtenir de la même manière la contribution des informations qui sont exploitées dans la reconnaissance des émotions. Une telle approche supprimerait par exemple l'influence que peuvent avoir les méthodes d'exploration de données sur les contributions respectives des informations utilisées lors des étapes de fusion, i.e., les poids α.

3.5. Méthodes d'exploration de données

Les techniques d'exploration de données (ou *datamining*) ont montré que les mesures de performance des systèmes de reconnaissance sont fortement dépendantes de la configuration des tests effectués. Tester un système de reconnaissance sur un ensemble de données revient à estimer la loi d'association des exemples à leurs classes (*risque réel*). Toutefois, cette loi peut être modifiée par la configuration des tests réalisés pour évaluer les performances du système de reconnaissance (*risque empirique*). Ainsi, le *risque réel* correspond aux véritables performances du système de reconnaissance, alors que le *risque empirique* représente celles que l'on peut mesurer [DUD00][13].

La technique de validation statistique croisée (CV) permet de faire en sorte que le *risque empirique* approche le *risque réel* [LEI98][22] ; les performances peuvent alors être améliorées comme dégradées. Ces méthodes reposent sur le principe du « *leave one out* », i.e., « *mets en un* [type de données] *de côté* ». Dans les approches de type CV, le système de reconnaissance est testé dans un ensemble de configurations parcourant toutes les données du corpus analysé, et les données testées ne font pas partie des sous-ensembles d'apprentissage. La technique de CV nécessite donc de découper l'ensemble des données en partitions égales (*folds*) représentant chacune une configuration de test du système. Ainsi, dans le cadre d'une *n*-CV, l'ens-

[22] F. Leisch, L. C. Jain et K. Hornik, "Cross-validation with active pattern selection for neural network classifiers", dans *IEEE Trans. on Neural Net.*, vol. 9, no. 1, pp. 35–41, Jan. 1998.

Table 3.1 Partitionnement des données en *folds* pour les tests de validation statistique croisée.

Tests	*fold* #1	*fold* #2	*fold* #3	*fold* #4	*fold* #5	*fold* #6	*fold* #7	*fold* #8	*fold* #9	*fold* #10
#1	***TST***	APP	APP	APP	APP	APP	APP	APP	APP	APP
#2	APP	***TST***	APP	APP	APP	APP	APP	APP	APP	APP
#3	APP	APP	***TST***	APP	APP	APP	APP	APP	APP	APP
#4	APP	APP	APP	***TST***	APP	APP	APP	APP	APP	APP
#5	APP	APP	APP	APP	***TST***	APP	APP	APP	APP	APP
#6	APP	APP	APP	APP	APP	***TST***	APP	APP	APP	APP
#7	APP	APP	APP	APP	APP	APP	***TST***	APP	APP	APP
#8	APP	APP	APP	APP	APP	APP	APP	***TST***	APP	APP
#9	APP	APP	APP	APP	APP	APP	APP	APP	***TST***	APP
#10	APP	APP	APP	APP	APP	APP	APP	APP	APP	***TST***

APP : partition servant à l'apprentissage des modèles ; TST : partition utilisée pour les tests.

emble des données est segmenté en *n* partitions de tailles égales pour autant de tests à effectuer ; la valeur de *n* est très fréquemment fixée à 10. Pour chaque itération, une partition est réservée pour les tests tandis que les autres servent à l'apprentissage des modèles, cf. table 3.1. Pour le cas d'une validation statistique croisée stratifiée (CVS), les exemples d'apprentissage sont équitablement distribués dans les catégories d'informations, et pour chaque configuration de test. Cela permet de s'assurer que les classes majoritaires ne seront pas favorisées puisque mieux représentées lors de l'étape de reconnaissance. Si les données analysées présentent des classes majoritaires, ce qui est le cas du corpus Berlin, leur partitionnement doit s'effectuer pour chaque classe. Notons que ce dernier peut s'effectuer via un tirage aléatoire des exemples, ou à travers des d'informations *a priori* telles que l'identité du locuteur (*leave one speaker out*) ou le contenu linguistique prononcé (*leave one utterance out*).

4. Reconnaissance acoustique

Nous présentons dans cette section, un bref état de l'art des travaux en reconnaissance d'émotions qui ont été effectués sur le corpus Berlin. Nous donnons ensuite les résultats que nous avons obtenus sur ce corpus, avec le système qui a été présenté dans la section précédente. Rappelons que l'objectif de cette étude, est d'estimer les contributions apportées par différents types de supports d'extraction d'informations, i.e., les points d'ancrages de la parole, cf. chapitre 2, dans la tâche de reconnaissance des émotions.

Le choix du corpus Berlin réside dans le fait que nous souhaitons analyser les caractéristiques de l'affect sans à avoir d'incertitude quant à l'émotion produite, et donc étudiée. Ce type d'approche nécessite par conséquent d'avoir à disposition des données correspondant à des émotions actées. Notons que les corrélats acoustiques de l'affect produits par les acteurs sont moins naturels que ceux que l'on peut observer dans un contexte de production spontanée, cf. chapitre 1, sous-section 3.4. Le corpus Berlin permet donc d'identifier plus des émotions *prototypiques* que de véritables signatures acoustiques des expressions de l'affect. Un autre élément qui rentre en compte dans le choix du corpus Berlin réside dans la richesse des transcriptions phonétiques qui y sont fournies. Ces dernières permettent en effet de localiser sans difficultés les points d'ancrage linguistique du signal de parole utilisés dans le système.

4.1. Etat-de-l'art sur le corpus Berlin

Le corpus Berlin a été étudié par l'auteur du principe de la décision « *segmentale* » [SHA07][12]. La méthode proposée consiste à caractériser des segments voisés avec 2 paramètres acoustiques (e.g., variance et somme de la valeur absolue des dérivées des coefficients MFCC) et 10 paramètres prosodiques (e.g., valeur de durée segmentale, 6 mesures statistiques du pitch, et 3 de l'énergie). La reconnaissance a été effectuée avec un classifieur SVM et dans le cadre d'une CVS. Les performances qui ont été obtenues avec ce système sont de 66% pour la décision « *segmentale* ». L'algorithme des *k*-ppv fournit un score plus faible : 59%. Une décision de type « *phrase* » (approche implémentée dans le développement du chien robotisé *AIBO* de *Sony*)[23] permet d'améliorer toutefois les performances en reconnaissance : 76% pour les SVM et 68% pour les *k*-ppv, cf. table 3.2. Comme le nombre de paramètres exploités dans la méthode de classification SVM est relativement faible, les scores obtenus par ce classifieur sont comparables avec ceux issus de la méthode des *k*-ppv.

Vogt *et al.* ont proposé une méthode pour contourner le problème de dépendance des paramètres MFCC aux informations locuteur [VOG06][24]. Un système de détection du genre du locuteur est pour cela employé en amont de la phase de reconnaissance. Le score obtenu par le système de détection du genre sur le corpus Berlin est de 69% avec la valeur moyenne du pitch et de 90% avec un ensemble optimisé de 20 paramètres composés de mesures statistiques du pitch (1 paramètre), de l'énergie (2 paramètres), et des coefficients cepstraux MFCC (17 paramètres comprenant 16 coefficients d'énergie MFCC plus leur valeur moyenne). Un classifieur naïf bayésien a été utilisé à la fois pour la détection du genre et pour la reconnaissance des émotions. Les paramètres ont été calculés sur des mots [VOG05][25]. Ceux qui ont été retenus pour la tâche de détection du genre ont été utilisés pour la classification des émotions dans une configuration indépendante au genre. Le nombre de paramètres destiné au genre masculin est de 10 (2 pitch + 1 énergie + 17 MFCC), et de 20 pour le genre féminin (3 pitch + 2 énergie + 15 MFCC). Le score en reconnaissance des émotions du corpus Berlin est de 79% dans la configuration indépendante au genre, et atteint 84% lorsque les résultats obtenus par les deux systèmes sont fusionnés. Ce qui prouve l'intérêt d'une architecture adaptant le traitement des données selon les informations. Cependant, l'utilisation du système de détection du genre montre que les rares erreurs commises par ce dernier (e.g., 10% de mauvaises détections) sont fortement pénalisées lors de la phase de reconnaissance, puisque le score est alors comparable à celui issu de l'analyse « sans informations de genre », cf. table 3.3.

---

[23] Cette approche exploite 5 descripteurs bas niveaux (LLD) issus du pitch, de l'énergie pour ses composantes basses fréquences, hautes fréquences, et pour tout le spectre, et de la valeur moyenne des dérivées en valeur absolue de 10 coefficients MFCC. 3 séries de valeurs sont ensuite calculées à partir des 5 LLD (e.g., valeurs des minimas, maximas et durée séparant les maximums locaux issus de la courbe filtrée à 10 Hz). Enfin, 10 mesures statistiques sont calculées pour chaque série de valeurs parmi les 20 résultantes (5 LLD + 3x5 LLD). Un algorithme d'apprentissage supervisé est utilisé pour construire un classifieur adapté à la base de données étiquetée [OUD03].

[24] T. Vogt et E. André, "Improving automatic emotion recognition from speech via gender differentiation", dans proc. *LREC*, Genoa, Italy, May 24-26 2006, pp. 1123–1126.

[25] T. Vogt et E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", dans proc. *ICME*, Amsterdam, The Netherlands, Jul. 6-8 2005, pp. 474–477.

Table 3.2 Scores de reconnaissance d'émotions sur le corpus Berlin [SHA07][12].

Classifieur	D_{seq}	D_{phr}
***k*-ppv**	59%	68%
SVM	66%	**76%**

Décisions de type « *segmental* » D_{seg} et « *phrase* » D_{phr} (approche *AIBO* de *Sony*).

Table 3.3 Scores de reconnaissance d'émotions sur le corpus Berlin [VOG06][24].

Stratégie de reconnaissance	Score
sans informations	79%
avec l'information correcte	**84%**
avec l'information détectée	80%

Les stratégies de reconnaissance correspondent à l'utilisation des informations de genre.

Table 3.4 Scores de reconnaissance d'émotions sur le corpus Berlin [SCH06c][26].

Classifieur	sans opt.	avec opt.
***1*-ppv**	64%	76%
***k*-ppv**	68%	79%
MLP	85%	87%
SVM	85%	88%

Avec et sans optimisation des jeux de paramètres acoustiques et prosodiques.

Le meilleur score de reconnaissance qui a été obtenu sur le corpus Berlin est de 97% (!) [SCH06c][26]. Cet excellent score s'explique par le fait qu'il est le fruit d'une approche qui sur-décrit les données [BAT99][27]. Le système proposé par Schuller *et al*. [SCH06c] exploite un ensemble de 276 mesures du signal de parole réparties sur le plan acoustique, e.g., rapport harmoniques sur bruit : 3 paramètres, *MFCC* : 120 paramètres (8 mesures statistiques issues de 15 coefficients MFCC, *Fast Fourier Transform* : 17 paramètres, et taux de passage par zéro : 3 paramètres) et prosodique (mesures statistiques du pitch : 12 paramètres, énergie : 11 paramètres, durée des segments voisés et non-voisés : 5 paramètres, et formants : 105 paramètres). Tous ces paramètres ont été calculés sur des segments, qui ont été extraits de façon indépendante du contexte de production, cf. Fig. 3.10. Ces segments vont de la phrase (GTI) à des portions de durée fixe situées en début, milieu et fin de phrase (GATIR), ou encore des zones découpées toutes les 500 ms (ATI), ou regroupant le tiers des informations disponibles (GRTI). Les paramètres, alors extraits sur ces différents types de segments de parole, sont regroupés en un super-vecteur de caractéristiques, e.g., GATIR. Une étape de sélection des caractéristiques est ensuite opérée par le classifieur *discriminant* SVM dans une approche de type flottante : chaque paramètre est séquentiellement inséré puis éliminé dans le but d'en identifier le jeu conduisant aux meilleures performances en reconnaissance (méthode *sequential floating forward selection* – SFFS). Plusieurs classifieurs ont été testés sur deux configurations d'analyse des caractéristiques, i.e., avec ou non l'étape de sélection des caractéristiques effectuées en amont. Le score de reconnaissance obtenu avec les segments GTI (i.e.,

[26] B. Schuller et G. Rigoll, "Timing levels in segment-based speech emotion recognition", dans proc. *Interspeech*, Pittsburgh, (PA), Sep. 17-21 2006, pp. 1818–1821.

[27] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth et H. Niemann, "Prosodic feature evaluation: brute force or well designed?", dans proc. *14th ICPhS*, San Francisco, (CA), USA, Aug. 1999, pp. 2315–2318.

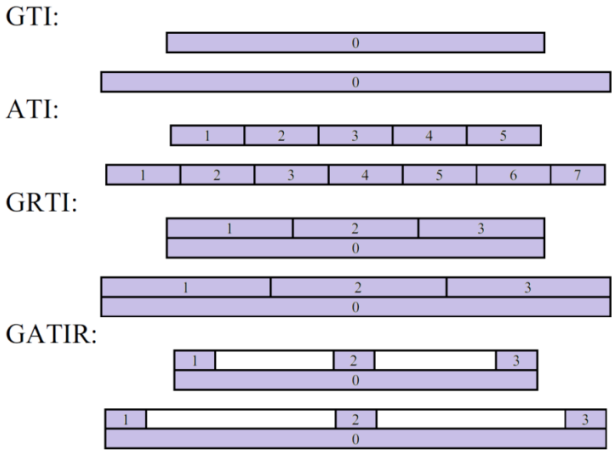


Fig. 3.10 Schémas de segmentation du signal de parole sans information de contexte (e.g., une phrase courte et une phrase longue). Les nombres indiqués renvoient à l'index de segment ; GTI : intervalle de temps global ; ATI : intervalles de temps absolus ; RTI : intervalles de temps relatifs ; GATIR : exemple de combinaison d'unités globale et segmentales ; figure extraite de [SCH06c][26].

toute la phrase) par un classifieur SVM et opérant sur les paramètres issus de l'étape de sélection SFFS est de 88% (85% sans optimisation). Le classifieur basé sur les *k*-ppv amène quant à lui, un score de 76% (64% sans optimisation), cf. table 3.4. Rappelons que l'algorithme des *k*-ppv permet d'obtenir un compromis entre la complexité du système de classification et la présence relativement faible des données à traiter.

Pour le détail, le score de 97% est obtenu avec le super-vecteur de caractéristiques issu de la configuration GRTI. Les scores obtenus dans ce type d'approche sont excellents puisque les informations sont sur-décrites. Cela reviendrait par exemple à optimiser le système de reconnaissance sur toutes les données fournies par tous les points d'ancrages acoustiques de la parole. L'amélioration des performances serait alors très probable, mais non pertinente, car due à une très grande redondance des informations exploitées par le système.

4.2. Stratégies de reconnaissance

Nous proposons dans cette thèse d'utiliser une stratégie de reconnaissance qui a pour objectif d'estimer la contribution des informations associées à la communication orale des émotions. Par exemple, au lieu de découper le signal de parole de façon arbitraire pour en extraire à l'aveugle les constituants acoustiques et prosodiques, nous proposons d'exploiter les points d'ancrages de la parole décrits dans le chapitre précédent. Notons que ces ancrages sont identifiés automatiquement au moyen de traitements relativement simples à mettre en œuvre (i.e., calcul de coefficients spectraux et cepstraux, de distances et de modèles statistiques), et dont leurs coûts calculatoires n'est pas pénalisant pour une contrainte telle que du temps-réel. Les systèmes de reconnaissance de la parole nécessitent quant à eux des méthodes plus complexes à mettre en œuvre, et donc plus coûteuses en temps de calcul. De plus, la quantité de données à traiter par le système est réduite de façon drastique lorsque l'on exploite des ancrages acoustiques, cf. Fig. 2.1. Cette réduction entraîne un gain de temps non négligeable pour les classifieurs *k*-ppv et MMG, puisqu'une distance est calculée entre chaque exemple testé et ceux issus de la phase d'apprentissage lors de l'estimation des probabilités *a posteriori*.

Ensuite, et au-delà du fait que la majorité des systèmes de reconnaissance issus de l'état-de-l'art ignore le contexte de production lors de l'étape d'extraction de caractéristiques, les méthodes proposées regroupent la plupart du temps des mesures de natures différentes (e.g., acoustique et prosodie) dans un unique vecteur de caractéristiques. L'étape de classification est ainsi réalisée sur un ensemble d'informations incohérentes puisque de nature différente. Par conséquent, de nombreuses études ont montré l'intérêt d'une modélisation différenciée des informations issues du cadre segmental et suprasegmental pour la reconnaissance des émotions [KIM07][28], [VLA07][11], [MAH09][18] et [SCH09c][29]. Dans nos expériences, nous séparons également le traitement des données issues de ces deux contextes : un système de reconnaissance est défini pour chaque type de mesure, i.e., acoustique et prosodique. La littérature montre que le classifieur MMG apparaît comme le plus pertinent pour une analyse segmentale (faible nombre de paramètres et matrices de covariance des MFCC diagonalisées), et le classifieur SVM pour les paramètres prosodiques en raison de leur pouvoir à traiter un grand ensemble de données non-linéairement séparables. La littérature a aussi montré que l'algorithme des *k*-ppv peut fournir des performances similaires aux deux autres classifieurs et ce quel que soit le type de décision employé, i.e., *segmental* ou *phrase*. Dans cette thèse, les classifieurs *k*-ppv et MMG ont été utilisés pour obtenir des scores compatibles avec l'analyse du corpus Berlin qui ne contient pas beaucoup de données, cf. table 2.4. La Fig. 3.6 décrit l'architecture du système qui a servi à effectuer les expériences en reconnaissance acoustique des émotions sur le corpus Berlin.

Nous rappelons que les transcriptions phonétiques contenues dans ce corpus ne sont pas toutes identifiées par l'alphabet SAMPA, cf. annexe 1. Nous avons donc réalisé une première analyse (décrite dans cette annexe) pour déterminer l'effet liée à l'introduction des données manquantes sur les scores en reconnaissance d'émotions. Une configuration élargie des transcriptions n'a pas permis d'apporter de résultats probants. Nous avons retenu, pour l'ensemble des expériences à venir sur ces types d'ancrage (chapitre 4 inclus), uniquement les données issues des transcriptions qui ont été identifiées par l'alphabet SAMPA. Les transcriptions syllabiques ont, quant à elles, été analysées sous deux aspects : les parties voisées et non-voisées des syllabes ont été identifiées de façon automatique dans un premier temps, et les structures de type : consonne – C, voyelle – V, consonne–voyelle – CV, voyelle-consonne – VC et consonne-voyelle-consonne – CVC – ont été extraites, via les transcriptions phonétiques dans un second temps. La table 3.5 reprend les caractéristiques principales de l'ensemble des points d'ancrages acoustiques et rythmiques qui ont été identifiés sur le corpus Berlin.

Après l'étape d'extraction de caractéristiques, les coefficients MFCC ont été normalisés par le calcul du *z-score* [11]. Plusieurs variantes de ce calcul ont été utilisées pour introduire des informations dans le système de reconnaissance : (i) *Raw* : aucune normalisation des données, (ii) *Z-tout* : normalisation sans information, (iii) *Z-genre* : normalisation des données selon le genre du locuteur, (iv) *Z-locuteur* : normalisation des données selon le locuteur et (v)

[28] S. Kim, P. G. Georgiou, S. Lee, et S. Narayanan, "Real-time emotion detection system using speech: multi-modal fusion of different timescale features", dans *9th W. on MMSP*, Chania, Crete, Greece, Oct. 1-3 2007, pp. 48–51.

[29] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll et A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances", dans proc. *W. on ASRU*, Merano, Italy, Dec. 13-17 2009, pp. 552–557.

Z-phrase : normalisation des données selon la phrase. L'approche utilisée pour inclure les informations *a priori* dans le système de reconnaissance permet ainsi d'adapter les données au système, plutôt que d'effectuer le contraire [VOG05][25].

Les méthodes d'exploration de données étudiées dans cette thèse pour la reconnaissance des émotions sont de deux types : (i) test d'indépendance aux classes, i.e., CVS et (ii) test d'indépendance au locuteur, i.e., LOSO. Les paragraphes suivants présentent les scores de reconnaissance à travers ces deux schémas d'analyse de données. Les résultats issus des tests LOSO sont généralement inférieurs à ceux obtenus en CVS. En effet, attribuer une émotion à un signal de parole produit par un individu inconnu, et dans un état affectif également inconnu, s'avère particulièrement délicat à réaliser, que ce soit pour une machine ou pour un être humain [ESP09][30]. Le score de reconnaissance d'un classifieur purement naïf (i.e., attribuant toujours la même étiquette correspondant à la classe majoritaire) sur le corpus Berlin est de 24% et les tests en perception amènent un score de 85% [PAE00][31]. Rappelons enfin que l'objectif principal des expériences qui vont suivre consiste à estimer la contribution des points d'ancrages acoustiques de la parole (cf. table 3.5) dans la tâche de reconnaissance des émotions. Par conséquent, seules les tables relatives à ces résultats seront données.

4.3. Tests en validation statistique croisée et stratifiée

Afin de faire ressortir les résultats que nous avons évalués comme probants parmi toutes les configurations étudiées, nous avons fixé pour l'analyse un seuil à 60% sur les scores issus de la CVS. Cette valeur équivaut à environ 90% du meilleur score obtenu par l'algorithme des *k*-ppv sur le corpus Berlin. Les paragraphes suivants décrivent les principaux résultats qui ont été obtenus sur la méthode d'exploration de données CVS. Cette méthode permet d'évaluer, dans un contexte d'indépendance aux classes, la pertinence liée à l'emploi d'un ensemble de points d'ancrages complémentaires de la parole pour extraire les caractéristiques acoustiques MFCC servant à effectuer la reconnaissance d'émotions.

4.3.1. Décisions « *segmentale* » et « *phrase* »

Les résultats montrent que les deux classifieurs employés (i.e., *k*-ppv et MMG) amènent des scores de reconnaissance très proches à travers les deux types de décisions « *segmentale* » et « *phrase* ». Les meilleurs scores sont sans surprise obtenus lorsque les informations locuteurs sont introduites dans le système de reconnaissance. Les autres types de normalisation ne montrent pas de réelles améliorations, pire, les résultats peuvent même être dégradés par rapport au traitement des données brutes, i.e., sans normalisation. Cela montre donc l'importance du type d'information exploité pour normaliser les données, comme cela a également été montré par Vogt *et al.* [VOG05][25].

[30] A. Esposito, M. Teresa Riviello et N. Bourbakis, "Cultural specific effects on the recognition of basic emotions: A study on Italian subjects", dans A. Holzinger and K. Miesenherger [Eds], *LNCS, USAB,* vol. 5889, pp. 135–148, 2009.

[31] A. Paeschke et W. Sendlmeier, "Prosodic characteristics of emotional speech: measurements of fundamental frequency movements", dans proc. *ISCA ITRW on Speech and Emotion*, Belfast, United-Kingdom, Sep. 5-7 2000, pp. 75–80.

Table 3.5 Caractéristiques des ancrages acoustiques contenus dans le corpus Berlin.

Ancrage acoustique	Quantité	Durée
voisé	4 175	218 $_{193}$
non-voisé	5 147	100 $_{101}$
voyelle	5 154	68.6 $_{38.9}$
p-voyelle	7 016	78.6 $_{31.4}$
voyelle courte	4 493	65.0 $_{35.0}$
voyelle longue	361	98.8 $_{59.6}$
voyelle diphtongue	300	86.7 $_{42.5}$
consonne	10 072	51.9 $_{36.8}$
p-consonne	14 770	54.4 $_{36.9}$
consonne plosive	3 668	33.4 $_{23.2}$
consonne fricative	3 071	51.6 $_{34.7}$
consonne sonorante	3 333	72.6 $_{39.8}$
syllabe	7 736	171 $_{87.0}$
syllabe voisée	4 147	209 $_{193}$
syllabe non-voisée	4 625	82.4 $_{67.9}$
syllabe partie C	705	114 $_{64.5}$
syllabe partie V	141	162 $_{58.1}$
syllabe partie CV	1 865	146 $_{64.6}$
syllabe partie VC	1 177	142 $_{57.5}$
syllabe partie CVC	3 848	202 $_{95.7}$
« p-centres » 1	3 234	76.8 $_{56.3}$
« p-centres » 2	4 078	91.0 $_{74.5}$
« p-centres » 3	4 830	122 $_{110}$

Durée en ms et dans le format suivant : [Valeur moyenne] $_{(\text{écart-type})}$; les groupe de phonèmes issus des transcriptions correspondent à ceux qui ont été validés par l'alphabet SAMPA.

Les ancrages qui apparaissent comme les plus pertinents pour extraire des caractéristiques acoustiques corrélées à l'affect sont : (i) les segments voisés (issus ou non des syllabes), (ii) les voyelles, (iii) les pseudo-voyelles et aussi (iv) les centres de perception de la parole « p-centres » (identifiés au niveau 3). Les meilleurs scores sont alors de : 66% pour les ancrages voisés (1-ppv, D_{seg}), 65% pour les voyelles (1-ppv, D_{seg}), 64% pour la partie voisée des syllabes (1-ppv, D_{seg}) et 62% pour les pseudo-voyelles (8 MMG, D_{seg} et D_{sup}) et les « *p-centres* » identifiés au niveau 3 (16 MMG, D_{seg}).

4.3.2. Fusion des classifieurs

Les résultats issus du premier étage de fusion montrent que les améliorations apportées par la fusion des classifieurs sont minimes dans tous les cas de figure, et sont essentiellement présentes sur les approches n'exploitant pas d'informations *a priori*, e.g., *Raw* et *Z-tout*. Les poids issus de la phase d'optimisation des scores montrent que le classifieur *k*-ppv amène, en grande majorité, les résultats les plus contributifs dans la tâche de classification des émotions.

4.3.3. Fusion des décisions « *segmentale* » et « *phrase* »

Les résultats de cette avant-dernière étape de fusion montrent que les décisions « *segmen-*

tale » et « *phrase* » ne permettent d'améliorer que très légèrement les scores issus de l'étage précédent. Bien que les contributions de chacune de ces décisions varient énormément selon les points d'ancrage de la parole et le type de normalisation, elles sont dans la globalité équivalentes (e.g., 5/5), avec une légère préférence pour les décisions de type « *segmental* ».

4.3.4. Fusion des ancrages acoustiques

La dernière étape de fusion consiste à exploiter les différents types d'ancrages acoustiques complémentaires (e.g., voyelle / consonne) qui ont été préalablement identifiés sur le signal de parole. Comme l'introduction des informations rythmiques apporterait de la redondance par rapport aux autres types d'ancrages, nous ne les avons pas exploités pour cette dernière étape de fusion. Les résultats présentés dans la table 3.6 montre que l'apport des ancrages complémentaires non-dominants en scores (e.g., non-voisé, consonne et pseudo-consonne) sur les ancrages dominants (e.g., voisé, voyelle et pseudo-voyelle) est relativement faible, ce qui peut paraître décevant de prime abord compte tenu de leur complémentarité présumée pour la description des caractéristiques acoustiques de la parole affective. Toutefois, la fusion des classes phonétiques permet d'approcher les performances obtenues par les macro-classes (e.g., voyelles et consonnes). Leur contribution varie alors selon le type de normalisation des données, cf. table 3.6. Ces résultats sont également valables pour les ancrages issus des classes de syllabes, i.e., V/C/CV/VC/CVC.

Alors que la fusion des paires d'ancrages complémentaires ne permet pas d'améliorer significativement les performances, les valeurs moyennes des scores obtenus cachent des résultats beaucoup plus intéressants. En effet, les matrices de confusion montrent que les points d'ancrages complémentaires de la parole conduisent à des taux de reconnaissance très différents selon les émotions. Ces matrices de confusion, trop nombreuses pour être toutes présentées, sont résumées dans la table 3.7. Cette table montre que la *Colère* est par exemple très bien reconnue par les ancrages phonétiques, mais bien moins pour les ancrages voisés, et de même pour la *Tristesse*. Les ancrages rythmiques amènent quant à eux, les meilleurs résultats sur le style « *Neutre* », bien que le score obtenu soit le plus faible d'entre tous. Ainsi, ce style amène non seulement de mauvais scores en détection de voyelles (cf. chapitre 2, sous-section 4.2), mais également en reconnaissance d'émotions pour l'approche CVS.

Le meilleur score obtenu par le système proposé pour effectuer la reconnaissance des émotions sur le plan acoustique est donc de 68% pour une CVS. Ce score est produit pour différents types de points d'ancrage de la parole. Il est aussi identique à celui issu des travaux de Shami [SHA07][12] et légèrement supérieur à celui obtenu par Schuler *et al.* avec le classifieur *k*-ppv et sans optimisation des paramètres [SCH06c][26]. Le système que nous avons proposé amène donc des performances comparables à celles de la littérature, notamment si l'on prend en compte le fait que nous avons exploité en tout et pour tout 16 paramètres acoustiques MFCC. De plus, la combinaison des meilleurs scores obtenus par la fusion des ancrages complémentaires selon les émotions est de 74% pour la configuration CVS. Une analyse spécifique aux émotions de ces derniers, permet donc d'améliorer significativement le score de reconnaissance en CVS.

Table 3.6 Scores en reconnaissance acoustique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires de la parole.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**63** $_{9/1}$	**63** $_{9/1}$	**64** $_{9/1}$	**68** $_{9/1}$	**62** $_{1/0}$
voyelle / consonne	**61** $_{5/5}$	**62** $_{5/5}$	**62** $_{7/3}$	**68** $_{7/3}$	**60** $_{1/0}$
p-voyelle / p-consonne	**62** $_{2/8}$	**61** $_{8/2}$	**62** $_{6/4}$	**67** $_{1/0}$	**60** $_{1/0}$
phonèmes voyelles	55 $_{7/1/3}$	53 $_{1/0/0}$	56 $_{8/1/0}$	**64** $_{8/0/2}$	54 $_{6/0/4}$
phonèmes consonnes	58 $_{0/1/8}$	57 $_{1/2/7}$	57 $_{1/1/8}$	**62** $_{1/2/7}$	58 $_{1/3/7}$
phonèmes V / C	59 $_{2/8}$	59 $_{0/1}$	59 $_{7/3}$	**65** $_{4/6}$	59 $_{6/4}$
syllabes voisées / non-voisées	**62** $_{9/1}$	**61** $_{9/1}$	**62** $_{7/3}$	**68** $_{7/3}$	**60** $_{7/3}$
syllabes V/C/CV/VC/CVC	58 $_{0/0/2/0/8}$	56 $_{1/0/2/2/5}$	56 $_{3/0/2/5/1}$	**61** $_{1/0/0/2/7}$	57 $_{1/1/3/1/5}$

[valeur en %] $_{\text{poids}}$ α_a ; décisions « *segmentale* » et « *phrase* » ; classifieurs *k*-ppv et MMG ; méthode CVS.

Table 3.7 Comparaison des scores en reconnaissance acoustique d'émotions sur le corpus Berlin selon les fusions des ancrages acoustiques complémentaires de la parole.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	79	**74**	59	74	61	48	81
voyelle / consonne	79	**74**	**65**	46	59	49	87
p-voyelle / p-consonne	79	68	52	63	**62**	52	87
phonèmes voyelles	77	69	**65**	46	54	43	85
phonèmes consonnes	72	72	38	72	49	39	94
phonèmes V / C	**87**	53	51	**78**	59	43	92
syllabes voisées / non-voisées	78	56	41	74	56	35	90
syllabes V/C/CV/VC/CVC	77	72	52	65	52	39	**95**
« p-centres » niveau 3	81	56	35	67	54	**56**	94

Valeur en % ; méthode CVS.

4.4. Tests d'indépendance locuteur

Afin de faire ressortir les résultats que nous évaluons comme probants parmi toutes les configurations étudiées, nous avons fixé pour l'analyse un seuil à 50% sur les scores de reconnaissance issus des tests LOSO. Cette valeur équivaut à environ 75% du meilleur score obtenu par l'algorithme des *k*-ppv sur le corpus Berlin. Ce seuil est plus bas que dans la configuration de CVS puisque la difficulté est supérieure. La méthode LOSO permet d'évaluer, dans un contexte d'indépendance au locuteur, la pertinence liée à l'emploi d'un ensemble de points d'ancrages acoustiques complémentaires de la parole. Cela, afin d'extraire les caractéristiques acoustiques MFCC servant à effectuer la reconnaissance d'émotions.

4.4.1. Décisions « *segmentale* » et « *phrase* »

Les résultats montrent que les deux classifieurs amènent des scores de reconnaissance très proches pour les décisions « *segmentale* » et « *phrase* ». Comparé à la méthode CVS, les différences obtenues avec la méthode LOSO sont minimes, excepté bien entendu les valeurs des scores qui sont dégradées d'environ 10 points de performance. Dans la méthode LOSO, la décision « *segmentale* » fournit à nouveau des scores qui sont supérieurs à ceux issus de

l'approche « *phrase* ». Cette amélioration est également plus marquée sur le classifieur *k*-ppv que MMG, et les meilleurs scores sont obtenus lors de l'introduction des informations locuteurs dans le système de reconnaissance. Les autres types de normalisation ne montrent pas de réelles améliorations, et les dégradations des scores par rapport au traitement des données brutes (i.e., sans normalisation) restent présentes. Cela montre encore une fois l'importance de la pertinence des informations exploitées pour normaliser les données, en particulier dans les tests d'indépendance au locuteur.

Les résultats les plus marquants sont portés par les ancrages rythmiques *« p-centre »*, unités que nous avons proposés d'exploiter pour extraire les informations de l'affect, cf. chapitre 2, sous-section 3.3 [RIN09][32]. Les scores en reconnaissance d'émotions obtenus par ces ancrages sont ainsi bien meilleurs comparés à tous les autres ; l'écart avec les segments voisés atteint presque les 10 points de performance sur la décision de type « *phrase* », prise à travers l'algorithme des *k*-ppv. Le centre de perception *« p-centre »* apparait donc comme un point d'ancrage des variabilités acoustiques apportées par l'affect, en particulier dans les tests d'indépendance au locuteur. De plus, ce résultat est conforté par le fait que les performances sont comparables à travers leurs trois niveaux d'identification, et que les proéminences du rythme (i.e., les « *p-centres* » identifiés au niveau 1) amènent les meilleurs résultats alors qu'elles fournissent moins de données.

4.4.2. Fusion des classifieurs et des décisions « *segmentale* » et « *phrase* »

Les résultats issus du premier étage de fusion montrent que les améliorations apportées par la fusion des deux classifieurs sont une nouvelle fois minimes et se manifestent sur les approches qui n'exploitent pas d'informations *a priori* dans l'étape de normalisation des données. Les poids issus de la phase d'optimisation des scores montrent que le classifieur *k*-ppv est encore une fois beaucoup plus contributif que celui basé sur les MMG. La fusion des décisions « *segmentale* » et « *phrase* » montre quant à elle que les meilleurs résultats sont dans la majorité portés par les décisions de type « *segmental* ».

4.4.3. Fusion des ancrages acoustiques

Les scores obtenus lors de la dernière étape de fusion sont donnés dans la table 3.8. Encore une fois, hormis les valeurs de scores en eux-mêmes, les résultats obtenus sont similaires à ceux de la méthode CVS : l'apport des principaux ancrages non-dominants en score (e.g., non-voisé, consonne et pseudo-consonne) sur leur complémentaire (e.g., voisé, voyelle et pseudo-voyelle) est extrêmement faible, et la fusion des ancrages phonétiques et syllabiques permet d'approcher les performances obtenues par leur ancrages respectifs d'ordre supérieur (e.g., macro-classes phonétiques et syllabiques). Notons que les valeurs moyennes des scores cachent une nouvelle fois des résultats qui confortent l'intérêt que nous portons aux points d'ancrage de la parole. En effet, les matrices de confusion issues du dernier étage du système de reconnaissance, montrent que les différents types d'ancrages analysés ne conduisent pas à

[32] F. Ringeval et M. Chetouani, "Hilbert-Huang transform for non-linear characterization of speech rhythm", dans proc. *NOLISP*, Vic, Spain, Jun. 25-27 2009.

Table 3.8 Scores en reconnaissance acoustique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires de la parole.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	49 $_{5/5}$	46 $_{7/3}$	48 $_{9/1}$	**54** $_{1/0}$	46 $_{4/6}$
voyelle / consonne	48 $_{0/1}$	49 $_{0/1}$	49 $_{0/1}$	**54** $_{7/3}$	48 $_{0/1}$
p-voyelle / p-consonne	**52** $_{9/1}$	**50** $_{1/0}$	**53** $_{1/0}$	**57** $_{9/1}$	49 $_{1/0}$
phonèmes voyelles	49 $_{9/1/0}$	46 $_{8/2/0}$	48 $_{9/1/0}$	**54** $_{1/0/0}$	47 $_{4/5/1}$
phonèmes consonnes	**50** $_{0/4/6}$	**50** $_{0/0/1}$	**50** $_{0/1/9}$	**53** $_{1/1/9}$	48 $_{0/0/1}$
phonèmes	**51** $_{5/5}$	**50** $_{5/5}$	**50** $_{0/1}$	**54** $_{1/0}$	49 $_{6/4}$
syllabes voisées / non-voisées	49 $_{1/0}$	48 $_{1/0}$	**51** $_{1/0}$	**55** $_{9/1}$	47 $_{1/0}$
syllabes V/C/CV/VC/CVC	49 $_{1/1/8/0/0}$	48 $_{0/5/4/1/1}$	**50** $_{1/3/5/1/1}$	**53** $_{4/1/3/0/2}$	47 $_{2/1/7/0/0}$

[valeur en %] $_{poids}$ α_a ; validation statistique LOSO.

Table 3.9 Comparaison des scores en reconnaissance d'émotions sur le corpus Berlin selon les fusions des ancrages acoustiques complémentaires de la parole.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	**99**	18	73	**37**	8	19	84
voyelle / consonne	**99**	26	62	13	14	39	76
p-voyelle / p-consonne	**99**	41	69	20	13	29	85
phonèmes voyelles	98	35	77	7	11	33	66
phonèmes consonnes	98	19	52	30	6	43	79
phonèmes	**99**	25	**78**	33	11	22	77
syllabes voisées / non-voisées	**99**	19	59	13	8	38	**90**
syllabes V/C/CV/VC/CVC	98	32	**78**	7	11	34	69
« p-centres » niveau 1	90	**44**	58	15	**18**	**61**	**90**

Méthode LOSO.

des taux de reconnaissances identiques à travers les émotions, cf. table 3.9. Bien que ces différences soient moins marquées que dans l'approche CVS. L'émotion « *Colère* » est par exemple la seule à être reconnue de façon quasi identique par l'ensemble des points d'ancrage de la parole[33]. Notons aussi que la moitié des styles affectifs étudiés, plus le « *Neutre* », sont bien mieux reconnus lorsque les paramètres MFCC sont calculés sur les proéminences du rythme de la parole comparé aux autres types d'ancrage.

5. Conclusion

Des méthodes permettant de modéliser les caractéristiques acoustiques du signal de parole ont été présentées, cf. section 2. La littérature montre que les coefficients MFCC semblent adaptés à de nombreuses tâches dédiées au TAP. Concernant la reconnaissance des émotions, une étape de normalisation de ces paramètres est requise pour limiter l'influence du loc-

[33] Les différences entre les scores issus de la méthode CVS et LOSO sur l'émotion « *Colère* » peuvent s'expliquer par le fait que cette émotion est de loin la plus représentée parmi toutes les autres ; 25% des données sont issues de cette émotion. L'influence des classes majoritairement représentées sur les scores est nulle dans la méthode CVS, mais présente dans la méthode LOSO. Raison pour laquelle l'émotion du *Dégoût* est bien mieux reconnue dans l'approche CVS que LOSO, puisque cette dernière est minoritaire en nombre d'exemples. De plus, les tables montrent que la *Colère* génère le plus grand nombre de mauvaises classifications par le système de reconnaissance.

uteur sur les caractéristiques acoustiques mesurables de ses expressions affectives. Une méthode basée sur le calcul du *z-score* a été proposée dans la littérature. Plusieurs variantes de ce calcul ont été utilisées dans cette thèse pour étudier l'impact de l'introduction d'informations *a priori* (e.g., genre, identité et type de phrase produite par le locuteur testé) dans le système de reconnaissance.

Le système que nous avons proposé repose sur la fusion de différentes approches pour caractériser les coefficients MFCC : (i) une étape d'extraction des informations selon différents types d'ancrages acoustiques, (ii) des prises de décision à l'échelle segmentale et de la phrase et (iii) l'emploi de deux types de classifieurs : *k*-ppv et MMG, cf. Fig. 3.6. La contribution de ces différentes approches dans la tâche de reconnaissance des émotions a été obtenue lors des étapes de fusion, i.e., par les poids associés à une combinaison linéaire des probabilités *a posteriori* fusionnées. Les expériences ont été réalisées à travers deux méthodes d'exploration de données : la première méthode (CVS) permet de tester l'indépendance du système de reconnaissance aux classes, alors que la deuxième méthode (LOSO) teste l'indépendance au locuteur (difficulté supérieure).

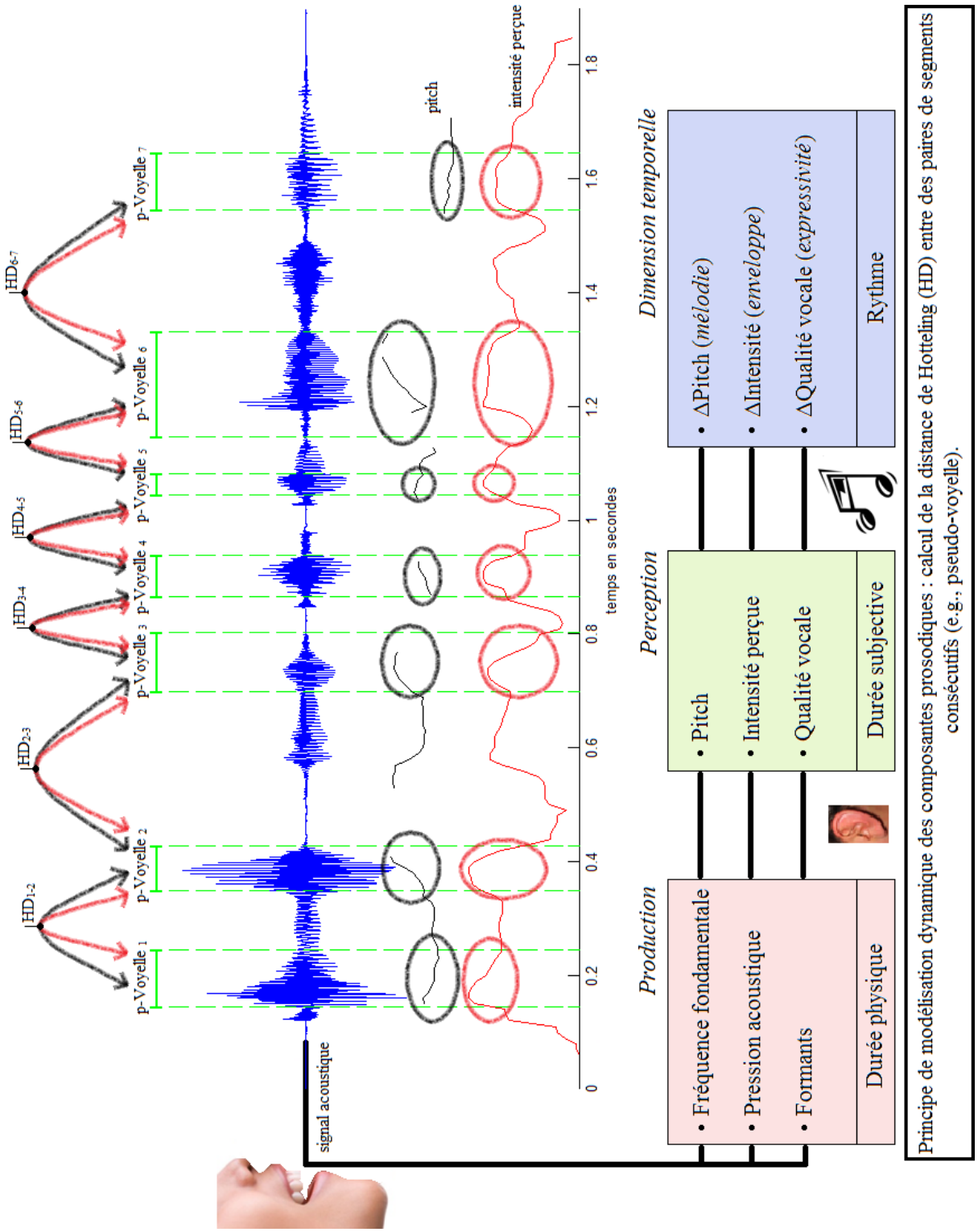
Les résultats montrent que l'introduction des informations locuteur dans le système de reconnaissance est pertinente puisque les scores s'en trouvent significativement améliorés. Néanmoins, les performances sont souvent dégradées sur les autres normalisations. Les deux types de classifieurs étudiés amènent des performances comparables sur l'ensemble des configurations d'analyse. Les décisions « *segmentales* » produisent des scores légèrement supérieurs à ceux issus des décisions de type « *phrase* », le gain apporté par leur contribution respective lors de la fusion est toutefois minime. De même, l'apport de la fusion des différents points d'ancrages de la parole sur les scores de reconnaissance est relativement faible. Les meilleurs scores obtenus en reconnaissance d'émotions avec le système proposé sont de 68% pour la méthode CVS (score obtenu par différents types d'ancrage), et de 59% pour la méthode LOSO (score obtenu par l'ancrage rythmique « *p-centre* »). Ces résultats sont comparables avec ceux issus de la littérature, puisqu'ils ont été uniquement obtenus au moyen d'un ensemble de 16 paramètres MFCC. Une analyse détaillée des scores fait apparaître des différences importantes dans les valeurs obtenues selon les différents points d'ancrages de la parole. Les performances atteignent ainsi jusqu'à 74% en CVS et 66% pour l'évaluation LOSO lorsque l'on combine les scores obtenus par les ancrages amenant les meilleurs résultats selon les émotions. Les scores en reconnaissance d'émotions peuvent donc être nettement améliorés en spécifiant l'analyse selon plusieurs types d'ancrages acoustiques de la parole.

Cependant, et bien que les coefficients MFCC apparaissent adaptés à un grand nombre de tâche du TAP, reconnaissance d'émotions comprises, il existe d'autres types de mesures du signal de parole qui semblent *a priori* plus adéquats pour effectuer la reconnaissance des émotions. Les paramètres MFCC sont en effet dépendants du cadre segmental qu'ils imposent au moment de leur extraction. Certains auteurs ont ainsi proposé des moyens de détourner cette contrainte (e.g., calcul des dérivés, ou de mesure statistique), mais les données qui en résultent ne peuvent pas véritablement prétendre à caractériser l'aspect suprasegmental de la parole puisqu'elles sont issues d'un contexte segmental. La prosodie, qui supporte les informations verbales et non-verbales du discours dans un contexte suprasegmental, se prêtent donc plus volontiers à l'étude des émotions que les paramètres « passe-partout » MFCC, cf. chapitre 4.

Chapitre 4

Reconnaissance prosodique de la parole affective actée

La prosodie permet à un locuteur de moduler et de rehausser le sens apporté aux informations de la communication. Ces informations peuvent être de nature *grammaticale*, *pragmatique*, ou *affective* et sont portées dans le discours par des variations paramétriques de composantes acoustiques telles que le pitch / intonation, l'énergie / modulation d'intensité, les formants / modulation de la qualité vocale et la durée / rythme. Tous les systèmes de reconnaissance actuels s'appuient sur ces paramètres de la prosodie pour caractériser les informations du discours, e.g., modalité de la phrase, attitude et intention face au discours, style du locuteur, émotions, etc. Cependant, et bien que de nombreux efforts aient été entrepris pour extraire un large éventail de caractéristiques prosodiques permettant d'analyser les corrélats acoustiques de l'affect, il n'existe pas de véritable consensus quant à celles qui sont les plus pertinentes. De plus, les travaux qui ont été effectués se sont focalisés sur l'analyse de caractéristiques du pitch, de l'énergie et de la qualité vocale, alors que le rythme est trop souvent modélisé par des mesures exploitant uniquement le débit de parole ou la durée segmentale. Toutefois, caractériser la composante rythmique de la parole n'est pas en soi une tâche facile puisque la littérature montre qu'il n'existe pas un rythme mais plutôt des rythmes. Partant de ce constat et en suivant les préceptes posés par un grand nombre d'auteurs, nous proposons des méthodes *non-conventionnelles* pour définir de nouvelles métriques du rythme de la parole. Ces méthodes s'appuient sur les points d'ancrages acoustiques de la parole et incluent : (i) le calcul de mesures spectrales sur l'enveloppe (analyses de Fourier basses fréquences – Tilsen), (ii) l'estimation de l'enveloppe et de la fréquence instantanées (transformée d'Hilbert-Huang), (iii) les indices de variabilité prosodique (extension du PVI de Grabe et Low aux composantes prosodiques) et (iv) les distances prosodiques (distance de Hotteling, *nouveau*). La pertinence de ces mesures pour caractériser les émotions prototypiques, i.e., actées, est démontrée dans ce chapitre sur de multiples aspects.



Principe de modélisation dynamique des composantes prosodiques : calcul de la distance de Hotteling (HD) entre des paires de segments consécutifs (e.g., pseudo-voyelle).

1. Notions fondamentales sur le rythme

Parmi les composantes physiques autour desquelles la prosodie de la parole s'articule, il est admis que le rythme n'est pas un signal en soi, contrairement au pitch et à l'énergie. En effet, le rythme est associé aux informations perçues lors de l'alternance ou la répétition d'évènements espacés dans le temps. Ces évènements peuvent être de nature très variable : (i) biologique (e.g., alimentaire, cardiaque, respiratoire, etc.) [ROB89][1], (ii) corporelle (e.g., chorégraphie) ou (iii) de la parole [CUM02][2]. Le rythme renvoie ainsi à la notion de mouvement (ou dynamique) dans la perception de divers types de phénomènes. Comme ces phénomènes sont très variés, que leur intrication est manifeste pour la parole [CUM09][3] et [TIL08a][4], et que les mécanismes de perception de l'Homme, sont d'autre part très complexes [ZWI90][5], définir concrètement et simplement les caractéristiques du rythme s'avère particulièrement difficile, même si l'on en restreint le champ d'analyse à celui de la musique [JON87][6]. Les mesures du rythme de la parole ne peuvent pas, par exemple, se limiter aux mesures du débit puisque ce dernier en est tout simplement qu'une composante [SCH04a][7], [MEI08][8], [DEL08][9]. Nous présentons dans les paragraphes suivants, un ensemble de phénomènes rythmiques qui ont été identifiés dans la littérature.

1.1. Dualité forme / structure

Un des éléments fondamental qui doit être pris en compte dans l'analyse du rythme concerne la dualité entre forme et structure [FRA56][10]. En effet, sa définition renvoie à la notion d'alternance dans le temps de phénomènes perceptuels. Mais, est-ce pour autant la variation dans la forme de ces phénomènes qui crée l'impression de rythme (approche *locale*) ? Ou, est-ce la structure issue de l'agencement temporel de ces formes (approche *globale*) ? Autrement dit, doit-on mesurer les intervalles de durées séparant les phénomènes perceptibles de la parole pour en caractériser les aspects rythmiques ? Ou bien, doit-on chercher à mesurer les différences entre leurs formes respectives ? La réponse à cette question n'est pas évidente. Cependant, nous sommes convaincus qu'il faut tenir compte de ces dualités pour analyser le ry-

[1] L. Robert, *Les horloges biologiques*, dans Paris: Flammarion, 1989.
[2] F. Cummins, "Speech rhythm and rhythmic taxonomy", dans proc. *Speech Prosody*, Aix-en-Provence, France, 11–13 Apr. 2002, pp. 121–126.
[3] F. Cummins, "Rhythm as entrainment: The case of synchronous speech", dans *J. of Phonetics*, vol. 37, no. 1, pp. 16–28, Jan. 2009.
[4] S. Tilsen, "Multitimescale dynamical interactions between speech rhythm and gesture", dans *Cogn. Sci.*, vol. 33, pp. 839–879, Jul. 2009.
[5] E. Zwicker et H. Fastl, *Psychoacoustics: facts and models*, dans Springer-Verlag, Heidelberg, 1990.
[6] M. R. Jones, "Dynamic pattern structure in music: recent theory and research", dans *Perception & psychophysics*, Psychonomic Society, Austin (TX), USA, [Eds], vol. 41, no. 6, pp. 621–634, Jun. 1987.
[7] D. Schreuder et D. Gilbers, "The influence of speech rate on rhythm patterns", dans Gilbers, D., Schreuder, M. and N. Knevel [Eds], *On the Boundaries of Phonology and Phonetics*, pp. 183–202, 2004.
[8] A. Meireles et P. A. Barbosa, "Speech rate effects on speech rhythm", dans proc. *Speech Prosody*, Campinas, Brasil, May 6-9, 2008, pp. 327–330.
[9] V. Dellwo, "The role of speech rate in perceiving speech rhythm", dans proc. *Speech Prosody*, Campinas, Brasil, May 6-9, 2008, pp. 375–378.
[10] P. Fraisse (préface de A. Michotte), *Les structures rythmiques : étude psychologique*, dans publications universitaires de Louvain, Louvain, 1956.

thme puisqu'elles sont, par définition, intimement liées. La difficulté repose donc sur la question de savoir comment lier les caractéristiques de forme et de structure, des phénomènes créant le rythme. Un début de réponse peut être apporté par les études en psychologie.

1.2. Aspects psychologiques

De nombreuses études en psychologie ont cherché à définir un hypothétique mécanisme physiologique qui serait à même de réguler une horloge perceptuelle interne [FRI90][11]. Parmi les nombreux candidats qui ont été examinés (e.g., rythmes cardiaque, respiratoire, alpha issus du cortex cérébral et cellulaires issues de mesures autonomes), l'existence d'une horloge interne spécifique à la régulation du « sens temporel perçu » n'a pu être établie pour les animaux comme pour l'Homme [ROB89][1]. Selon Friedman, cette impossibilité serait liée au fait qu'aucune des horloges examinées n'a montré un continuum dans le lien établi avec la perception du temps [FRI90]. Des psychologues ont en effet montré qu'il existe de nombreux cas de distorsions dans la perception temporelle chez l'Homme, e.g., [KEL00][12] :

1. Les activités intéressantes réduisent l'estimation du temps passé ;
2. Un grand nombre de tâches effectuées dans une période donnée rallonge l'estimation du temps écoulé ;
3. Une période de temps est perçue plus longue par un sujet s'il connaît à l'avance la période de temps qu'il va devoir estimer ;
4. La perception des intervalles temporels de la musique dépendent partiellement de la distance tonale entre les sons.

1.3. Phénomènes psycho-acoustiques

Des études en psycho-acoustique ont montré qu'il existe des phénomènes de subjectivité dans la perception de la durée des sons de la parole [ZWI90][5]. La durée *subjective* perçue (1 *dura* = un son pur de fréquence 1kHz et produit à un niveau SPL de 60dB pendant 1 seconde) décroit moins rapidement que la durée *physique* du son produit pour des valeurs inférieures à 100ms ; et sont équivalentes pour des valeurs de durées supérieures à ce seuil, cf. Fig. 4.1. Des différences apparaissent également lorsque l'on compare la durée d'une pause et la durée des deux sonorités adjacentes, cf. Fig. 4.2. Lors de ces expériences, la durée physique du son a été modifiée par l'examinateur et le sujet a eu pour consigne de modifier la durée de la pause de façon à ce qu'elle soit perçue comme identique à celle des sonorités adjacentes. Différents types de sonorités ont été utilisés : (i) un bruit blanc, (ii) un son pur à 200Hz et (iii) à 3.2kHz. Les résultats obtenus sont très marquants puisque les durées de la pause qui ont été estimées par les sujets lors de la tâche (durée *subjective*) sont très nettement supérieures à la durée *physique* de cette dernière, cf. Fig. 4.2.

[11] W. Friedman, *About time. Inventing the fourth dimension*, dans Cambridge: MIT Press, 1990.
[12] B. Zellner Keller, et E. Keller, The chaotic nature of speech rhythm: hints for fluency in the language acquisition process, dans Delcloque, Ph., Holland, V.M. [Eds], *Integrating Speech Technology in Language Learning*, Swets & Zeitlinger, in press, 2000.

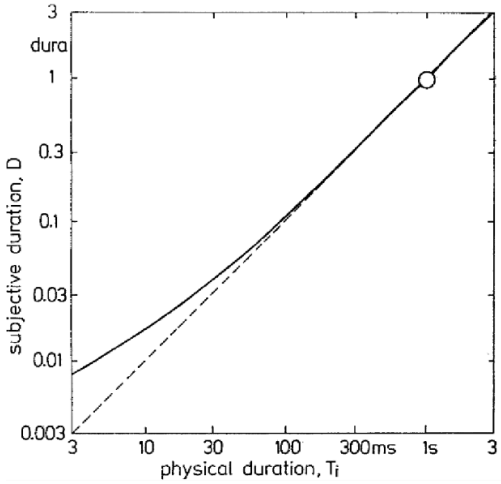


Fig. 4.1 Durée subjective en fonction de la durée physique d'un son pur de fréquence 1kHz et produit à un niveau de 60dB SPL ; figure extraite de [ZWI90][5].

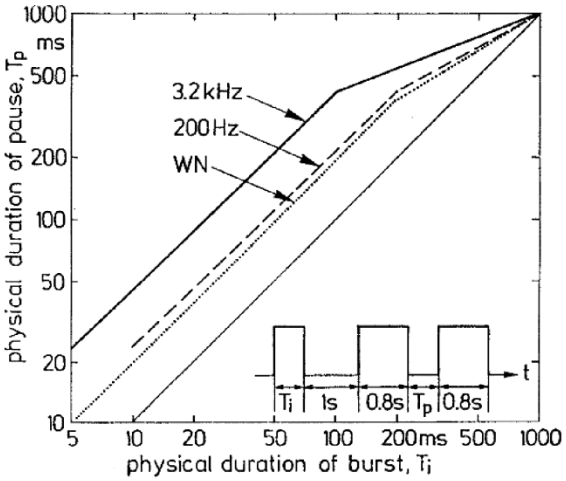


Fig. 4.2 Comparaison des durées subjectives produites par une pause selon différents types de sonorités adjacentes : WN : bruit blanc et sons purs à 200Hz et 3.2kHz ; figure extraite de [ZWI90][5].

1.4. Taxinomie de la parole

La majorité des études, qui ont été menées sur le rythme de la parole, ont été guidées par un esprit taxinomiste, i.e., dans le but d'effectuer la classification des langues. Les premiers travaux reposant sur l'idée que les langues peuvent être différenciées selon leurs propriétés rythmiques (e.g., *syllabique* / *accentuelle*) remontent au moins à ceux de Pike *et al.* [PIK43][13]. Il avait alors été supposé l'existence d'un espacement régulier des unités préférentielles de la parole selon les langues (e.g., accent, syllabe ou more) [ABE67][14]. Cette théorie de *l'isochronie* a été rejetée en masse par de nombreux auteurs puisqu'aucune régularité n'a pu être observée entre les unités spécifiques à chaque langue [BOL68][15], [LEH77][16] et [ROA82][17]. Après une discussion détaillée de cette controverse, Campbell a démontré qu'un modèle hiérarchique complexe du rythme, i.e., incluant des composantes à la fois segmentales, syllabiques et de niveaux supérieurs, est nécessaire [CAM92][18]. De plus, les considérations visant un unique paramètre temporel ou l'utilisation d'effets d'allongements compensatoires à l'intérieur ou entre les syllabes seraient inadéquates.

Les propriétés temporelles des unités vocaliques et consonantiques ont été exploitées par la suite pour analyser le rythme de la parole. Les résultats obtenus ont permis d'argumenter en faveur de l'existence d'un continuum entre les langues accentuelles et syllabiques [DAU83][19]. L'apparition de nouvelles métriques du rythme dans les années 2000 a apporté un net regain d'intérêt dans la communauté taxinomiste [GIB01][20]. Les auteurs de ces métriques (Scott, Gi-

[13] K. L. Pike, *The intonation of American English*, dans University of Michigan Press, Ann Arbor, 1945.
[14] D. Abercrombie, *Elements of general phonetics*, dans Edinburgh University Press, Mar. 1967.
[15] D. Bolinger, *Aspects of language*, dans Harcourt, Brace and World, New York, 1968.
[16] I. Lehiste, "Isochrony reconsidered", dans *J. of Phonetics*, vol. 5, no. 3, pp. 253–263, Mar. 1977.
[17] P. Roach, *On the distinction between 'stress-timed' and 'syllable-timed' languages*, dans D. Crystal [Eds], Linguistic Controversies, Arnold, London, 1982.
[18] W. N. Campbell, *Multi-level speech timing control*. Ph.D. thesis, University Sussex, UK, 1992.
[19] R. M., Dauer, "Stress-timing and syllable-timing reanalysed", dans *J. of Phonetics*, vol. 11, pp. 51–62, 1983.

bbon, Ramus, Low & Grabbe *et al.*) ont partagé une approche commune et reposant sur des considérations phonétiques, pour caractériser le continuum rythmique entre langues syllabiques et accentuelles. Ils ont ainsi pu tester la pertinence et le pouvoir discriminant d'une telle typologie des langues, en utilisant des mesures de variabilité et des distributions statistiques des intervalles, cf. Fig. 2.13.

1.5. Ancrages acoustiques

Le centre de perception (cf. chapitre 2, sous-section 2.2.2) permet d'accéder aux corrélats acoustiques du rythme de la parole (cf. chapitre 2, sous-section 3.3). Les expériences qui ont été réalisées dans le chapitre 2 ont permis d'identifier ces corrélats. Les résultats ont alors montré que le rythme est essentiellement constitué d'ancrages voisés et que les voyelles en sont la source principale. Les consonnes participent de façon non-négligeable à la structuration acoustique des « *p-centres* » ; d'autant plus lorsque la parole est chargée d'affect.

1.6. Nature chaotique

Keller a exploité les similarités existantes entre les préceptes de la théorie du chaos et les définitions associées au rythme de la parole (i.e., mouvement dynamique) pour en définir une métrique [KEL00][12]. Cette théorie permet de représenter la dynamique du signal de parole à travers un espace des phases composé de deux paramètres : (i) l'état du système à un instant donné et (ii) son évolution dans le temps (vélocité). Les expériences ont consisté à comparer les graphiques issus de l'espace des phases pour différents types de signaux, cf. Fig. 4.3. Les résultats montrent que les deux espaces des phases associés aux syllabes présentent un ensemble d'attracteurs (trois d'après l'auteur) ; contrairement aux signaux artificiels qui n'en présentent aucun. La différenciation entre les deux groupes de syllabes a été réalisée par le calcul de l'exponentielle de Hurst. Les résultats montrent des différences entre les deux styles de production des syllabes : style lent, H-exp = 0.033 et style rapide, H-exp = 0.099. Toutefois, la significativité de ces différences n'a pu être calculée puisque l'exponentielle de Hurst fait apparaître des valeurs d'écart-type qui sont largement supérieures aux données d'origine.

1.7. Nécessité d'une modélisation non-conventionnelle

Les paragraphes précédents ont montré combien il est difficile d'avoir une définition précise du rythme de la parole, puisqu'il existe de nombreux inventaires terminologiques et conceptuels [EVA86][21], [KEL00][12] et [DOU08][22]. Toutefois, un ensemble de phénomènes a pu être identifié à travers les études mentionnées tels que : (i) la dualité entre forme et structure, (ii) les distorsions temporelles, (iii) la subjectivité de la durée, (iv) des ancrages préférentiels selon les langues et (v) une nature chaotique. Ces résultats permettent de constituer une base méthodologique dans l'objectif de définir de nouveaux modèles *non-conventionnels* du rythme.

[20] D. Gibbon et U. Gut, "Measuring speech rhythm", dans proc. *Eurospeech*, Aalborg, Denmark, 3-7 Sep. 2001, pp. 95-98.

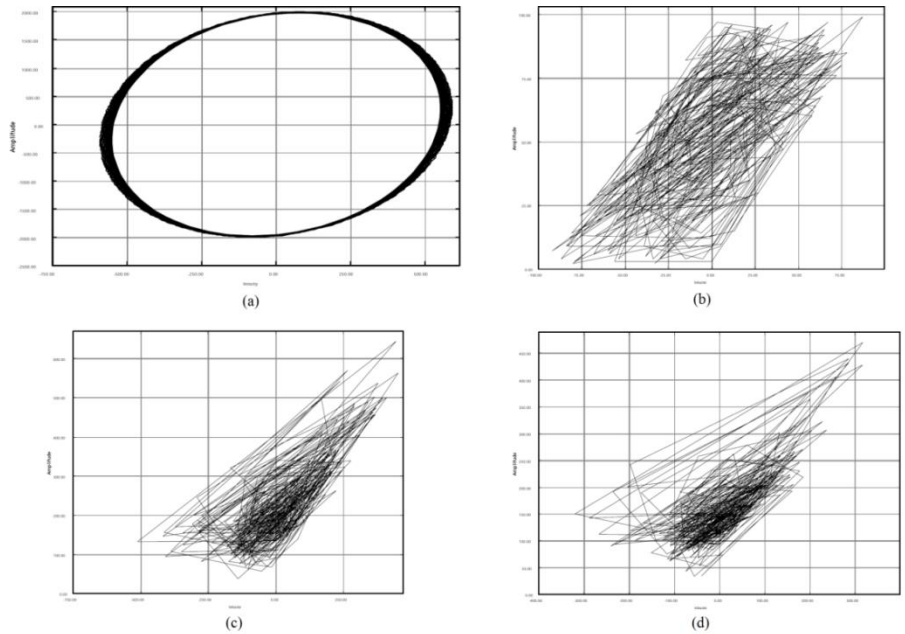


Fig. 4.3 Comparaison de l'espace des phases pour différents types de signaux : (a) onde sinusoïdale non amortie ; (b) bruit blanc ; (c) durées de syllabes produites à un débit lent ; et (d) durées de syllabes produites à un débit rapide ; figure extraite de [KEL00][12].

2. Modélisations prosodiques de la parole affective

L'étude des corrélats acoustiques de l'affect a été motivée ces dernières années par trois tendances majeures : (i) la tendance portée par l'analyse de données plus naturelles et plus proches de la vie réelle, (ii) la tendance visant à prendre en compte des catégories affectives pas seulement *prototypique* mais des émotions au sens large et (iii) la tendance cherchant à exploiter un large ensemble de caractéristiques, pouvant résulter en un super-vecteur composé d'une centaine, voire de milliers de paramètres, alors utilisés pour l'étape de classification [SCH07b][23]. Le challenge ultime de la reconnaissance des émotions est de trouver l'ensemble de caractéristiques à la fois discriminantes et indépendantes entres elles. Cet ensemble a été qualifié de « *Saint Graal* » par Batliner tant il a suscité l'intérêt et continue à motiver la communauté de chercheurs en TAP orienté émotion.

La tâche d'extraction automatique des corrélats affectifs présents dans le signal de parole est cependant loin d'être aisée, puisque les phénomènes observables sont dus à des causes multiples pouvant être corrélées entre elles, cf. Fig. 1.5. Par conséquent, l'existence d'un ensemble de caractéristiques prosodiques ayant le pouvoir de discriminer à 100% toutes les émotions rencontrées dans la communication humaine n'est clairement pas envisageable ; d'autant plus que des contradictions apparaissent déjà dans les théories des émotions, cf. cha-

[21] J. R. Evans, et M. Clynes, *Rhythm in psychological, linguistic, and musical processes*, dans Springfield, Charles C. Thomas, 1986.

[22] *L'Homme et ses rythmes*, séminaire dirigé par C. Doumet et A. W. Lasowski, Université Paris 8, Vincennes, Saint-Denis à l'ENS de la rue d'Ulm, France, 2006-2008.

[23] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, A. Noam, L. Kessous, et V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals", dans *Interspeech*, Antwerp, Belgium, 27-31 Aug. 2007, pp. 2253–2256.

pitre 1, section 3. Le « *Saint Graal* » serait donc (encore une fois ?) une illusion à fort pouvoir d'attrait. Néanmoins, notons que des scores de reconnaissance honorables (~70%) peuvent être atteints sur des corpus de parole contenant plusieurs classes d'émotions produites de façon spontanées par des enfants, cf. chapitre 5, sous-section 4.1. De plus, les expressions de l'affect ne se limitent pas à la modalité parole seule, mais se retrouvent également dans les expressions faciales et gestuelles [BUS04][24]. Toutefois, l'extraction de ces paramètres prosodiques semble plus difficile à réaliser que sur le signal de parole [VIN09][25].

Nous décrivons dans les paragraphes suivants, les paramètres qui ont été utilisés pour effectuer la reconnaissance prosodique des émotions sur le corpus Berlin. L'accent a été tout spécialement porté sur le développement de nouveaux paramètres du rythme puisque cette composante apparaît clairement comme étant sous-modélisée dans les systèmes *état-de-l'art*. Les méthodes couramment employées en caractérisation statique et dynamique de la prosodie à travers ses descripteurs bas-niveaux (LLDs) sont également présentées dans cette section.

2.1. Descripteurs bas-niveaux de la prosodie

Les paramètres prosodiques reposent sur des LLDs suprasegmentaux, contrairement aux coefficients acoustiques MFCC qui sont issus d'un contexte segmental, cf. chapitre 3, section 2. Les LLDs de la prosodie visent à extraire du signal de parole ses composantes perceptuelles telles que : (i) le pitch, (ii) l'énergie, (iii) la qualité vocale et (iv) le rythme.

2.1.1. Pitch

Le pitch correspond à la perception de la fréquence fondamentale f0 (e.g., échelle en demi-tons et non en Hz). Les méthodes d'estimation de la f0 reposent sur le calcul de l'autocorrélation du signal de parole, cf. Fig. 4.4. La position relative des minimums locaux présents sur le signal d'autocorrélation permet d'estimer par exemple la valeur de la f0. Les algorithmes populaires tels que *Snack* et *Praat* reposent quant à eux sur des méthodes de programmation plus complexes : les résidus d'un modèle de prédiction linéaire sont exploités pour estimer de façon optimale les valeurs de la f0 et les probabilités de voisement (méthode *Entropic Signal Processing System*– ESPS) [SEC83][26]. Cette méthode ayant montré des performances suffisamment bonnes [KIM06][27], elle est aujourd'hui quasi exclusivement employée par la communauté pour extraire la f0 dans le signal de parole. Nous avons également utilisé l'algorithme *Snack* dans les expériences de cette thèse pour estimer les valeurs de la f0 toutes les 10ms.

[24] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, S. Lee, U. Neumann, et S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information", dans proc. *6th ICMI*, State College (PA), USA, Oct. 13-15 2004, pp. 205–211.

[25] A. Vinciarelli, M. Pantic et H. Bourlard, "Social signal processing: Survey of an emerging domain", dans *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[26] B. Secrest et G. Doddington, "An integrated pitch tracking algorithm for speech systems", dans proc. *ICASSP*, Boston (MA), USA, Apr. 14-16 1983.

[27] C. Kim, K. D. Seo et W. Sung, "A robust formant extraction algorithm combining spectral peak picking and root polishing", dans *EURASIP J. on Applied Signal Proc.*, vol. 2006, article id. 67960, 2006.

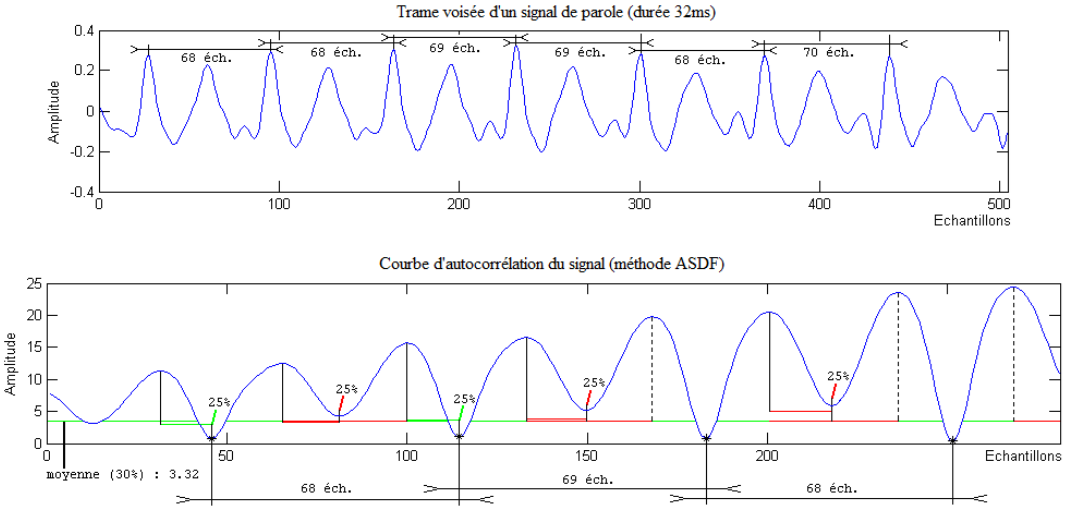


Fig. 4.4 Estimation de la f0 d'un signal de parole par la moyenne des écarts entre les minimums locaux retenus par un algorithme de détection sur la courbe ASDF (*average square difference function*) ; système développé lors du stage de Master 2 à l'UPMC.

2.1.2. Energie acoustique

Les caractéristiques d'énergie acoustique renvoient aux variations de pression produites par l'onde de phonation. Le calcul de l'énergie d'un signal de parole repose bien souvent sur la définition du traitement du signal, i.e., le calcul de l'intégrale des échantillons élevés au carré. Bien qu'il ait été montré l'existence de phénomènes liés à la perception de l'intensité des sons en psycho-acoustique [ZWI90][5], ces modèles sont néanmoins difficiles à exploiter puisqu'ils sont définis pour un ensemble de cas particuliers, cf. Fig. 3.5. Nous avons donc utilisé la définition issue du traitement du signal pour extraire les valeurs d'énergie sur le signal de parole. Elles ont été estimées toutes les 10ms et en dB par l'algorithme *Snack*.

2.1.3. Qualité vocale

Les qualités vocales correspondent à différents types de styles de production de la parole tels que : (i) la voix modale ; « *Neutre* », (ii) chuchotée, (iii) soufflée, (iv) craquée, (v) hachée ou (vi) de type falsetto [KEL05][28]. Scherer considère que les paramètres issus de la qualité vocale pourraient être un support préférentiel pour la différenciation vocale des émotions [SCH89][29], en particulier pour des variations subtiles de l'état affectif du locuteur [GOB03][30], [SCH01b][31]. Néanmoins, les paramètres issus de l'état-de-l'art en reconnaissance d'émotions n'ont pas permis de rejoindre les attentes produites par les études précédentes [STE09][32].

[28] E. Keller, "The analysis of voice quality in speech processing", dans *LNCS*, Springer-Verlag, vol. 3445, pp. 54–73, 2005.

[29] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance", dans H. L. Wagner and A. S. R. Manstead, [Eds], *Handbook of Social Psychophysiology*, John Wiley and Sons Ltd., chap. 7, pp. 165–197, London, 1989.

[30] C. Gobl et A. Ní Chasaide, "The role of voice quality in communication emotion, mood and attitude", dans *Speech Comm.*, vol. 40, no. 1-2, pp. 189–212, Apr. 2003.

[31] M. Schröder, "Emotional speech synthesis: A review", dans proc. *Interspeech*, Aalborg, Denmark, Sep. 3-7 2001, vol. 1, pp. 561–564.

La qualité vocale peut être caractérisée par la forme du signal acoustique produite par les formants ; le modèle source-filtre de Fant considère en effet la production, comme le résultat de la convolution d'une source par un filtre représentant le conduit vocal, cf. Fig. 3.3. Rappelons que nous exploitons dans cette thèse différents points d'ancrage de la parole pour extraire les caractéristiques servant à effectuer la reconnaissance des émotions. Comme ces points d'ancrages se différencient naturellement selon les valeurs des deux premiers formants F_1 et F_2 pour les voyelles (cf. Fig. 2.2), nous avons choisi d'exploiter ces données comme LLDs de la qualité vocale. Les valeurs de fréquence (Hertz) et d'énergie (dB) ont été estimées toutes les 10ms par l'algorithme *Snack*, ce qui a fourni 4 LLDs de la qualité vocale.

Nous proposons également d'exploiter une métrique qui permet de définir un continuum de valeurs entre des styles vocaux *hyper-articulé* (e.g., large triangle dans le plan $F_2 - F_1$) et *hypo-articulé* (e.g., faible triangle dans le plan $F_2 - F_1$). Ce $5^{ème}$ LLD est estimé par la surface du polygone créée par les valeurs des deux premiers formants F_1 et F_2 sur un segment de parole donné. Le calcul de l'aire du polygone s'effectue par la fonction *polyarea* de *Maltab* (somme des produits entre les deux séries de valeurs). Enfin, nous avons défini un $6^{ème}$ LLD de la qualité vocale par la mesure de dérive spectrale REC. Cette mesure a été utilisée pour détecter les voyelles dans un signal de parole et représente l'harmonicité spectrale du signal telle que définie par l'équation [5], cf. chapitre 2, sous-section 3.1.

2.1.4. Prétraitements du pitch, de l'énergie et des formants

Les valeurs brutes de la f0, de l'énergie et des formants doivent être prétraitées pour tenir compte des propriétés en perception de la parole (e.g., échelle logarithmique en fréquence et en amplitude, lissage des variations temporelles non-perceptibles) et des changements dans les conditions d'enregistrement (e.g., gain du microphone, distance au locuteur pour les valeurs d'énergie). Les valeurs du pitch ont été obtenues par le calcul donné dans l'équation [22] ; échelle 12 demi-tons ; fréquence référence égale à 55Hz (note A_1 en musique). Les valeurs d'énergie du signal et des formants (en dB) ont quant à elles été normalisées sur chaque phrase par la valeur maximale en valeur absolue. La dynamique de ces signaux évoluent donc dans l'intervalle : [-1 ; +1]. Un filtrage médian constitué d'une fenêtre glissante travaillant sur 3 échantillons a été employé pour éliminer certains artefacts unitaires présents sur la f0, l'énergie ou les valeurs de formants. Ce filtrage permet de lisser les variations non perceptibles dans les données tout en les adaptant à une modélisation statistique.

$$Pitch = \frac{12}{log(2)} log\left(\frac{f0}{55}\right) \qquad [22]$$

2.2. Techniques de modélisation du rythme

Nous décrivons dans les paragraphes suivants plusieurs techniques qui ont été proposées dans la littérature pour caractériser les informations du rythme de la parole. Nous séparons la description de ces techniques selon les approches *conventionnelles* et *non-conventionnelles*.

[32] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, PhD thesis, University Friedrich-Alexander, Erlangen-Nuremberg, Germany, 2009.

2.2.1. Modèles conventionnels

Les modèles *conventionnels* du rythme reposent sur des mesures de la durée segmentale ou de la quantité de segments présents selon un point d'ancrage donné (e.g., segments voisés, voyelles, consonnes, etc.). Le débit de parole est par exemple très souvent utilisé comme unique paramètre du rythme dans les systèmes de reconnaissance d'émotions [ANG02][33], [YIL04][34], bien que de nombreuses études aient montré qu'il en est seulement une composante [SCH04a][8], [MEI08][9], [DEL08][10]. De plus, il existe de nombreuses métriques du rythme dont l'intérêt pour la caractérisation des corrélats de l'affect demande à être démontré. Nous présentons dans les paragraphes suivants, quatre métriques qui ont été proposées par différents auteurs au moyen de considérations phonétiques : (i) les phénomènes de réduction (%V et ΔC), (ii) les coefficients de variations ($Varco_C$ et $Varco_V$), (iii) la périodicité (mécanismes oscillatoires) et (iv) les variabilités inter-segmentales (*r*PVI et *n*PVI). Les caractéristiques de ces métriques sont expliquées dans la table 4.1.

Les phénomènes de réduction vocalique et consonantique

Ramus *et al.* ont proposé une mesure du rythme basée sur le pourcentage d'intervalles vocaliques (%V) et l'écart-type des intervalles consonantiques (ΔC) dans l'objectif de quantifier un continuum rythmique entre les langues *syllabiques* et *accentuelles* [RAM99][35]. Ces mesures ont occasionné de nombreux débats pour savoir si elles permettaient ou non de représenter le rythme des langues selon cette topologie [CUM02][2] et [GRA02a][36]. Grabe et Low n'ont pu, par exemple, reproduire les mêmes résultats de l'étude de Ramus *et al.*[GRA02a]. Ce dernier a alors supposé que le manque de contrôle dans le débit de la parole a pu conduire à des résultats différents dans [GRA02a] ; [RAM02][37]. En conclusion, les paramètres %V et ΔC ne seraient pertinents que pour l'étude de corpus dont le débit de parole ne serait que strictement contrôlé. Dans une tentative de valider cette hypothèse, Dellwo *et al.* ont reproduit les expériences en demandant à des individus de lire un texte à différentes vitesses de lecture. Les résultats obtenus par ces auteurs n'ont pu que partiellement valider la pertinence des mesures des phénomènes de réduction %V et ΔC pour caractériser le rythme des langues [DEL03][38].

Les phénomènes de compensation (mécanismes oscillatoires)

Brady *et al.* ont cherché à démontrer l'existence d'une horloge interne synchronisant les

[33] J. Ang, R. Dhillon, E. Schriberg et A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in Human-computer dialog", dans proc. *Interspeech*, *7th ICSLP*, Denver (CO), USA, 16-20 Sep. 2002, pp. 67–79.

[34] S. Yildrim, M. Buhut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee et S. Narayanan, "An acoustic study of emotions expressed in speech", dans proc. *Interspeech*, *8th ICSLP*, Jeju Island, Korea, 4-8 Oct. 2004.

[35] F. Ramus, M. Nespor, et J. Mehler, "Correlates of linguistic rhythm in the speech signal", dans *Cognition*, vol. 73, no. 3, pp. 265–292, Dec. 1999.

[36] E. Grabe et E. L. Low, "Durational variability in speech and the rhythm class hypothesis", dans C. Gussenhoven & N. Warner [Eds], *Papers in laboratory phonology VII*, The Hague: Mouton de Gruyter, vol. 7, pp. 515–546, 2002.

[37] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives", dans proc. *Speech Prosody*, B. Bel and I. Marlin [Eds], Aix-en-Provence, France, Apr. 11-13 2002, pp. 115–120.

[38] V. Dellwo et P. Wagner, "Relations between language rhythm and speech rate", dans proc. *15th ICPhS*, Barcelona, Spain, Aug. 3-9 2003, pp. 471–474.

Table 4.1 Résumé des caractéristiques des métriques conventionnelles du rythme de la parole.

Métrique	**Paramètre**	**Calcul**	**Domaine d'application**	**Avantage(s)**	**Inconvénient(s)**
%V, ΔC	pourcent V et delta C	proportion et écart-type des intervalles vocaliques et consonantiques	phénomènes de réduction *mesure globale*	calcul très simple à mettre en œuvre, nécessite au min. 2 unités pour **%V**	dépend du débit, ne prend pas en compte les phénomènes locaux
Varco	coefficient de variation	rapport de l'écart-type sur la moyenne des intervalles	phénomènes de réduction *mesure globale*	prend en compte les moments statistiques d'ordre 1 et 2	mêmes inconvénients que **%V** et **ΔC**
$\bar{R}$	périodicité moyenne	écart-type de la distribution statistique circulaire des intervalles	phénomènes de compensation *mesure globale*	fait ressortir des irrégularités dans les intervalles de durée	de même que **%V** et **ΔC**, nécessite au min. 3 unités
RR	rhythm ratio	rapport de durée entre deux intervalles consécutifs	phénomènes de variations dans les intervalles *mesure locale*	étudie les enchainements à court terme des intervalles	dépend du débit, nécessite au min. 3 unités
***r*PVI**	raw pairwise variability index	différence absolue entre deux intervalle consécutifs	de même que **RR** *mesure locale*	de même que **RR**	de même que **RR**
***n*PVI**	normalized pairwise variability index	de même que pour *r*PVI mais avec une normalisation au débit	de même que **RR** *mesure locale*	de même que **RR**, prend en compte le débit	nécessite au min. 3 unités

processus cognitifs de la parole pour le Japonais ; notons que cette horloge est sujette à la controverse, cf. sous-section 1.2. Une de leurs récentes études [BRA06][39], conduite sur les phénomènes de compensation des mores en relation avec les intervalles de durée séparant les attaques des syllabes voisées du Japonais, les ont amené à considérer ces unités comme des cibles importantes pour une horloge servant de référence, ou un mécanisme de planification de la parole. Le phénomène de compensation temporelle a été évalué au moyen de mesures statistiques circulaires. Un système permettant d'identifier automatiquement les attaques des syllabes voisées sur un signal de parole a tout d'abord été développé (filtrage du signal), cf. Fig. 4.5. Une fois ces segments identifiés, une onde sinusoïdale a ensuite été générée avec une période fixée à la valeur moyenne des intervalles séparant les segments. Leur position respective dans le temps correspond ainsi à une valeur de phase dans la sinusoïde générée. La périodicité des intervalles séparant les segments est quantifiée par la moyenne des *n* écarts de phase θ_i issus de la projection des données dans un espace de représentation circulaire [23], cf. Fig. 4.6. Cette mesure équivaut à celle de l'écart-type. Elle a notamment permis de montrer que : (i) les durées des segments de parole reliés à la more en Japonais dépendent de la durée des autres segments et que (ii) ces compensations déjoueraient n'importe quel oscillateur adaptatif qui supposerait une isochronie relativement bruitée dans les attaques des syllabes.

[39] M. C. Brady et R. F. Port, "Quantifying vowel onset periodicity in Japanese", dans proc. *16th ICPhS*, Saarbrücken, Germany, Aug. 6-10 2006, pp. 337–342.

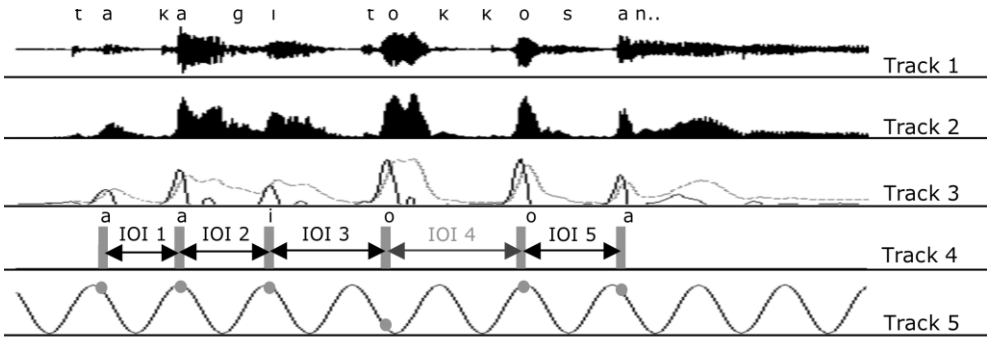


Fig. 4.5 Estimation des intervalles *Inter-Onset-Intervals* entres les attaques de syllabes voisées ; *Track* 1 : signal de parole analysé ; *Track* 2 : enveloppe du signal filtré par un passe-bas ; *Track* 3 : signal de détection des *onsets* (taux de changements positifs dans le signal filtré) ; *Track* 4 : attaques des syllabes voisées détectées par les pics issus du signal de la *Track* 3 ; *Track* 5 : onde sinusoïdale de période égale à la moyenne des IOI ; figure extraite de [BRA06][39].

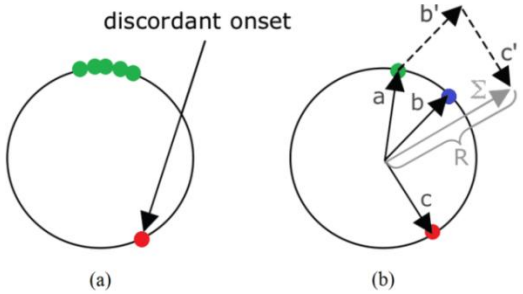


Fig. 4.6 (a) Estimation des phases correspondant aux intervalles *Inter-Onset-Intervals* entres les attaques des syllabes voisées ; et (b) calcul de leur périodicité [23] ; figure reproduite de [BRA06][39].

$$R^2 = \left(\sum_i sin(2\pi\theta_i)\right)^2 + \left(\sum_i cos(2\pi\theta_i)\right)^2$$

$$\bar{R} = \frac{R}{n} \qquad [23]$$

Les coefficients de variation de la durée segmentale et inter-segmentale

Le coefficient de variation Varco se définit par le rapport entre le deuxième et le premier moment statistique d'une distribution donnée, et a été utilisé sur des intervalles vocaliques et consonantiques par [DEL06][40]. La combinaison de cette mesure avec celle du %V a permis de discriminer des langues *syllabiques* (e.g., Anglais et Allemand) et *accentuelles* (e.g., Espagnol et Français), mais ces différences ont été aussi importantes sur les dialectes de ces langues.

Les variabilités intra- et inter-segmentales

Toujours dans une optique d'effectuer la taxinomie des langues par le rythme, Grabe et Low ont proposé de mesurer la variabilité temporelle de *N* paires d'intervalles phonétiques I_k successifs [24], cf. Fig. 4.7, plutôt que de calculer directement la statistique, i.e., mesure glo-

[40] V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC", dans proc. *Lang. and Lang. Proc., 38th Ling. Colloq.*, Piliscsaba, Hungary, Sep. 6-8 2006, pp. 231–241.

bale vs. locale [GRA02a][36]. Une normalisation au débit a été proposée (*n*PVI). Ces mesures ont permis de conforter la théorie des classes rythmiques des langues exposée précédemment.

$$rPVI = \frac{1}{N-1}\sum_{k=1}^{N-1}|I_k - I_{k+1}|$$
$$nPVI = \frac{2}{N-1}\sum_{k=1}^{N-1}\frac{|I_k - I_{k+1}|}{I_k + I_{k+1}} \quad [24]$$

Les signaux RR

D'autres études taxinomistes ont suggéré que la comparaison des intervalles de durée entre les unités phonétiques pouvait être réalisée, non pas par le calcul de la différence entre les valeurs, mais par leur rapport [25] [GIB01][20]. La mesure *RR* (*rhythm ratio*) a fournie des résultats proches du *n*PVI sur des corpus de langue.

$$RR = \frac{1}{N-1}\sum_{k=1}^{N-1}\frac{I_k}{I_{k+1}} \quad [25]$$

Autres méthodes

Après une revue détaillée des définitions associées au rythme et des méthodes proposées dans la littérature pour en quantifier les corrélats dans la musique, F. Gouyon propose dans sa thèse d'exploiter des métriques plus axées sur le traitement du signal [GOU05][41]. Les techniques étudiées sont les suivantes : (i) l'autocorrélation de signaux générés par un filtre passe-bande sur le signal acoustique et par des impulsions de Dirac dont l'espacement représente les intervalles de durée entre des évènements musicaux, (ii) la transformée discrète de Fourier (TFD) sur ces mêmes signaux et (iii) la TFD du signal de parole après un filtrage en peigne. Ces techniques semblent prometteuses puisqu'elles ont permis de fournir des résultats en accord avec les autres modèles de la littérature sur différents corpus de données.

2.2.2. Modèles non-conventionnels

Les modèles conventionnels du rythme (cf. sous-section 2.1.5) n'ont pas réussi à valider de façon significative la théorie de l'existence d'un continuum rythmique entre les groupes de langues décrites dans la littérature, i.e., *accentuelle* et *syllabique*. Par conséquent, les débats continuent d'être alimentés aujourd'hui par de nombreuses études, e.g., [ARV10][42], [FEN10][43] et [SZC10][44]. D'autres travaux effectués dans l'analyse du rythme de la musique, ont proposé d'élargir les définitions des métriques issues des études taxinomistes [SMI00][45]. Ces auteurs ont par exemple, considéré que les structures du rythme pouvaient être générées par la dyna-

[41] F.Gouyon, *A Computational Approach to Rhythm Description: Audio Features for the Computation of Rhythm Periodicity Functions and Their Use in Tempo Induction and Music Content Processing*, PhD thesis, University Pompeu Fabra, Barcelona, Spain, 2005.

[42] A. Arvaniti et T. Ross, "Rhythm classes and speech perception", dans proc. *Speech Prosody*, Chicago (IL), USA, May 11-14, 2010, paper id. 100173:1-4.

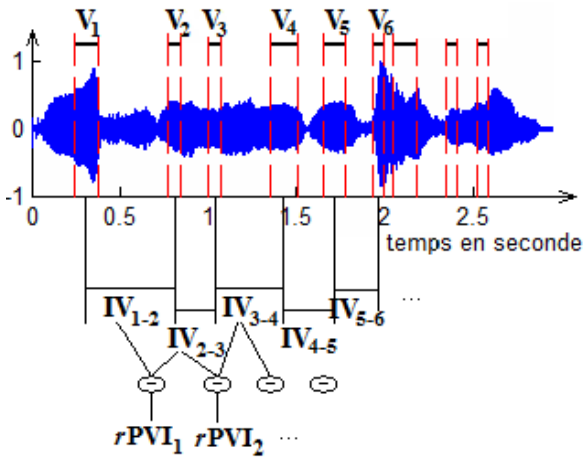


Fig. 4.7 Estimation de la mesure *r*PVI sur un signal de parole comprenant 9 voyelles en tout ; IV : intervalle vocalique ; *r*PVI : *raw*-pairwise variability index.

mique prosodique, i.e., les successions de motifs formés par les variations du pitch, de l'intensité et/ou de la qualité vocale. En effet, il a été montré que la plupart des compositeurs, utilisent ces dynamiques pour transmettre des émotions dans leurs œuvres [BEL09][46]. De plus, la musique des populations Arabes contient des structures rythmiques (uniques au monde) qui reposent sur des motifs mélodiques improvisés ; technique du « maqām » – مقام TOU03][47].

Les études conduites par Lerdahl *et al.* supportent également l'idée d'exploiter le pitch pour définir une métrique du rythme, puisqu'ils ont distingué les accents de la musique en : (i) groupes *métriques*, (ii) de *phénomènes* et (iii) de *structures* [LER96][48] ; (i) les accents *métriques* apparaissent lorsqu'un accent est situé à l'intérieur d'un motif métrique (répétition régulière d'accents) ; (ii) les accents de *phénomènes* existent au niveau de la surface musicale et mettent alors en avant, un moment unique qui active ce que les musiciens appellent une syncope[49] ; et (iii) les accents de *structures* sont définis comme « un accent causé par des points de mélodie ou d'harmonicité produisant un moment d'importance dans une phrase ou une section – spécifiquement par la cadence, le but du mouvement tonal » [LER96]. Ces auteurs ont aussi noté des particularités prosodiques liées aux temps fort du rythme puisque ces derniers tombent : (i) à des changements importants ou de faibles valeurs dans le pitch, (ii) à des changements dans les harmoniques ou (iii) à des changements dans les cadences.

Les caractéristiques du rythme peuvent donc reposer sur la dynamique du pitch. Comme les autres composantes de la prosodie, i.e., l'énergie et la qualité vocale, participent de façon active à la structuration des informations du discours, cf. chapitre 1, section 2, nous supposons

[43] G. Fenk-Oczlon et A. Fenk, "Measuring basic tempo across languages and some implications for speech rhythm", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010, pp. 1537–1540.
[44] B. Szczepek Reed "Speech rhythm across turn transitions in cross-cultural talk-in-interaction", dans *J. of Pragmatics*, vol. 42, no. 4, pp. 1037–1059, Apr. 2010.
[45] L. M. Smith, *A Multiresolution Time-Frequency Analysis and Interpretation of Musical Rhythm*, PhD thesis, University of Western Australia, 2000.
[46] G. Beller, *Analyse et modèle génératif de l'expressivité: Application à la musique et à l'interprétation musicale*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 2009.
[47] H. H. Touma, *La musique des Arabes*, Amadeus Press, 2003.
[48] F. Lerdahl et R. Jackendoff, *A generative theory of tonal music*, dans MIT Press, Mass, 1996.
[49] Prolongation d'un temps faible sur un temps fort.

que leur dynamique puisse aussi être à l'origine d'un phénomène rythmique. Caractériser la dynamique des composantes prosodiques à travers différents ancrages de la parole, tout en tenant compte des paramètres temporels, pourrait constituer une solution élégante à la problématique de la dualité entre forme et structure du rythme, cf. sous-section 1.1. Enfin, notons que des études ont montré que le rythme pourrait être à l'origine des émotions perçues par un auditeur lors de l'écoute d'une œuvre musicale [BEL09][46]. Et comme la prosodie repose avant tout sur les aspects musicaux de la parole [DIC04][50], nous supposons que le rythme peut être également corrélé à l'affect dans la communication orale. De nombreux travaux ont en effet démontré l'existence de liens étroits et situés à différents niveaux entre la parole et la musique [SNO31][51] et [LEE99b][52].

Nous décrivons dans les paragraphes suivants, quatre modèles *non-conventionnels* du rythme que nous avons développé pour étudier les corrélats de l'affect : (i) transformée de Fourier sur l'enveloppe rythmique du signal de parole « *p-centre* » [TIL08b][53] pour calculer des caractéristiques spectrale (e.g., entropie, fréquence moyenne et barycentre), (ii) transformée d'Hilbert-Huang (THH) pour estimer l'amplitude et la fréquence instantanées de signaux ré-échantillonnés et définies par des intervalles de durée. Les deux autres modèles intègrent quant à eux, les trois composantes de la prosodie telles que le pitch, l'énergie et la qualité vocale : (iii) calcul dérivé du PVI pour quantifier le changement dans le coefficient de variation (Varco) de chaque composante prosodique à travers des paires de segments consécutifs et (iv) distance de Hotteling pour calculer la dynamique à travers toutes les composantes prosodiques tout en prenant éventuellement en compte les corrélations. Pour compléter la description des modèles *non-conventionnels* du rythme, nous présentons une mesure qui a été proposée par Galves *et al.* et qui repose sur un calcul direct de l'entropie spectrale [GAL02][54].

La table 4.2 présente les caractéristiques des principales méthodes de modélisation *non-conventionnelles* du rythme de la parole. Elle montre qu'un large ensemble de phénomène est couvert par ces méthodes : (i) la dynamique à court terme dans le spectre du signal de parole, (ii) la dynamique dans l'enveloppe rythmique « *p-centre* », (iii) les dynamiques instantanées dans la fréquence et l'enveloppe estimées sur une unité donnée et (iv) la dynamique des composantes prosodiques autres que la durée et quantifiées à travers des paires d'unités consécutives. Ces modèles sont présentés en détails dans les paragraphes qui suivent.

Analyse basses fréquences de Fourier

Tilsen *et al.* ont proposé une méthode pour extraire l'enveloppe rythmique d'un signal de parole, cf. chapitre 2, sous-section 3.3 ; [TIL08b]. Un signal basse fréquence de *Fe* égale à 80Hz est calculé sur le signal de parole par différentes étapes de filtrages qui sont supposées

[50] A. Di Cristo, "La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions", dans *Travaux interdisciplinaires du Laboratoire Parole et Language d'Aix-en-Provence*, no. 23, pp. 67–211, 2004.

[51] W. B. Snow, "Audible frequency ranges of music, speech and noise", dans *J. of the Acous. Soc. Amer.*, Jul. 1931.

[52] T. van Leeuwen, *Speech, Music, Sound*, Palgrave Macmillan [Eds], Oct. 1999.

[53] S. Tilsen et K. Johnson, "Low-frequency Fourier analysis of speech rhythm", dans *J. of Acoust. Soc. of Amer.*, Express Letters, vol. 124, no. 2, pp. 34–39, Aug. 2008.

[54] A. Galves Jesus, J. Garcia, D. Duarte et C. Galves, "Sonority as a basis for rhythm class discrimination", dans proc. *Speech Prosody*, Aix-en-Provence, France, Apr. 11-13 2002, pp. 11–13.

Table 4.2 Résumé des caractéristiques des métriques non-conventionnelles du rythme de la parole.

Métrique	Calcul	Domaine d'application	Avantages	Inconvénients
Mesure de sonorité	divergence de Kullback-Leibler sur des coefficients spectraux et entre des trames consécutives	variations à court terme dans le spectre du signal de parole	ne nécessite pas de segmentation en voyelle / consonne	ignore le contexte de production, dépend du cadre segmental
Analyse basses fréquences de Fourier	entropie, barycentre et fréquence moyenne spectrale calculés sur la TF du « *p-centre* »	variations à long terme dans l'amplitude du signal « *p-centre* »	de même que précédemment, prend en compte les aspects en perception du rythme	ne prend pas en compte les phénomènes locaux
Fréquence et amplitude instantanées	THH sur des signaux créés par des intervalles de durée séparant une unité donnée	enveloppe et fréquence instantanées d'une unité donnée	fournit beaucoup de valeurs pour décrire l'enveloppe et la fréquence du rythme	nécessite au min. 3 unités, estimation couteuse en temps de calcul
Variabilité prosodique	calcul du *r*PVI sur le coefficient de variation d'un LLD prosodique et normalisation au débit	variations dans la dispersion d'un LLD à travers des paires d'unités consécutives	intègre les informations d'un LLD prosodique, prend en compte le débit	nécessite au min. 3 unités et la présence d'un LLD prosodique
Distance prosodique	distance de Hotteling à travers les LLDs et entre des paires d'unités consécutives, normalisation au débit et inclut les corrélations des LLDs	variations dans la distribution des LLDs à travers des paires d'unités consécutives	intègre les informations de tous les LLDs prosodiques, prend en compte le débit et les inter-corrélations	de même que précédemment et peut requérir la présence de plusieurs LLDs selon la config.

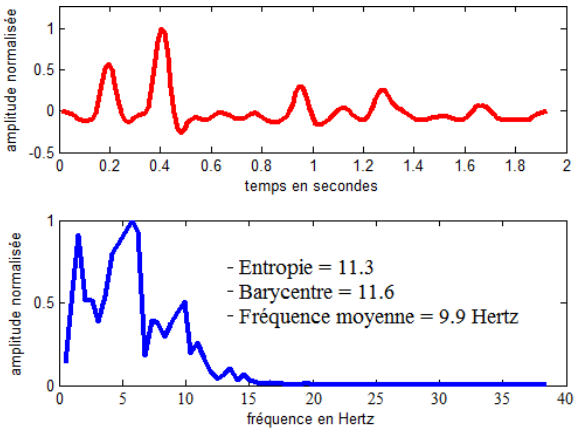
représenter les processus de perception du rythme. Comme la forme d'onde de ce signal est plutôt stationnaire, nous pouvons exploiter la transformée de Fourier pour estimer les valeurs d'entropie, de barycentre et la fréquence moyenne de l'enveloppe rythmique du signal, cf. Fig. 4.8 [26]. Ces paramètres permettent de décrire de façon globale, la structure rythmique contenue dans un signal de parole à travers la courbe décrivant les valeurs du « *p-centre* », cf. chapitre 2, sous-section 2.2.2.

Fréquence et amplitude instantanées

Nous avons proposé dans [RIN09][55] d'utiliser la transformée d'Hilbert-Huang (THH) pour extraire les composantes rythmiques de la parole [HUA98][56]. Des signaux SUI (*Speech Unit Intervals*) ont pour cela été générés avec les intervalles de durée séparant les segments consécutifs et issus d'un même ancrage acoustique, cf. Fig. 4.9. Puisque le nombre de segment disponible par phrase est relativement faible (souvent en dessous d'une dizaine), nous avons dû sur-échantillonner les signaux SUI avant de calculer la THH. Nous avons notamment utilisé des splines cubiques avec une fréquence d'échantillonnage *Fe* égale à 32Hz. Nous avons

[55] F. Ringeval et M. Chetouani, "Hilbert-Huang transform for non-linear characterization of speech rhythm", dans proc. *NOLISP*, Vic, Spain, Jun. 25-27 2009.

[56] N. Huang, Z. Shen, S. Long, *et al.* : "The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis", dans proc. *R. Soc. London*, ser. A, vol. 454, pp. 903–995, Mar. 1998.



$$Entropie = -\sum TF.log_2(TF)$$

$$Barycentre = \frac{Fe}{N}\frac{\sum(1:N)TF}{\sum TF} \qquad [26]$$

$$Fréquence\ moyenne = \frac{Fe}{N}\frac{\sum(1:N)(TF^2)}{\sum TF^2}$$

Fig. 4.8 Figure du haut : signal rythmique basse fréquences issue d'un signal de parole ; Figure du bas : spectre fréquentiel du signal rythmique.

choisi cette valeur pour qu'elle soit en accord avec le plus faible intervalle de durée qui puisse être présent dans nos données, ce qui est le cas des ancrages phonétiques et dont l'amplitude fréquentielle varie de 1Hz à 16Hz [DRU94][57]. La valeur de *Fe* que nous avons choisi correspond donc au double du plus grand écartement fréquentiel pouvant être observé entre deux segments consécutifs, i.e., 16Hz. La THH permet alors de fournir des données fiables sur les signaux SUI générés à partir des points d'ancrages acoustiques étudiés dans cette thèse, cf. chapitre 2. Notons qu'une plus grande valeur de *Fe* risquerait d'apporter des artefacts en raison d'un sur-échantillonnage trop important des signaux SUI, d'autant plus que des erreurs apparaissent déjà avec une valeur de *Fe* fixée à 32Hz, cf. Fréquence instantanée, Fig. 4.9.

La première étape de la THH consiste en une décomposition par mode empirique (EMD). La méthode EMD est une approche conduite par les données, et dans laquelle une série de valeurs $x(t)$ est décomposée en un ensemble fini d'oscillations individuelles caractéristiques appelées fonctions à mode intrinsèque (IMFs) [HUA98]. Les IMFs sont extraites à travers une représentation locale du signal $x(t)$, qui est considéré comme étant issu de la somme d'une composante oscillante $d(t)$ – partie hautes fréquences – et d'une tendance locale $m(t)$ – partie basses fréquences. Les IMFs sont itérativement obtenues par un processus de tamisage jusqu'à ce que deux conditions soient satisfaites : (i) une moyenne nulle et (ii) un nombre identique d'extrema et de passages par zéro, ou une différence de un. Le signal $x(t)$ est alors représenté par la somme de *N* IMFs d_k et des composantes résiduelles finales r_k [27].

Puisqu'un nombre trop important d'itérations peut conduire à une sur-décomposition du signal, Flandrin *et al.* ont proposé un nouveau critère pour stopper le processus de tamisage inclut dans l'EMD [RIL03][58]. Ce dernier est alors itéré tant qu'une fonction d'évaluation σ reste en-dessous d'un seuil θ_1 pour une fraction (1-α) de la durée totale et en dessous d'un deuxième seuil θ_2 pour la fraction temporelle restante. La fonction d'évaluation σ est définie par le rapport entre la moyenne $m(t)$ et la composante oscillante $d(t)$. Cette approche permet d'assurer de façon globale, de petites fluctuations dans la moyenne tout en tenant compte des

[57] R. Drullman, J. M. Festen et R. Plomp, "Effect of temporal envelope smearing on speech reception", dans *J. of the Acous. Soc. of Amer.*, vol. 95, pp. 1053–1064, 1994.

[58] G. Riling, P. Flandrin et P. Gonçalvès, "On empirical mode decomposition and its algorithms", dans proc. 6th *IEEE-EURASIP W.* on *NSIP*, Grado, Italy, Jun. 8-11, 2003.

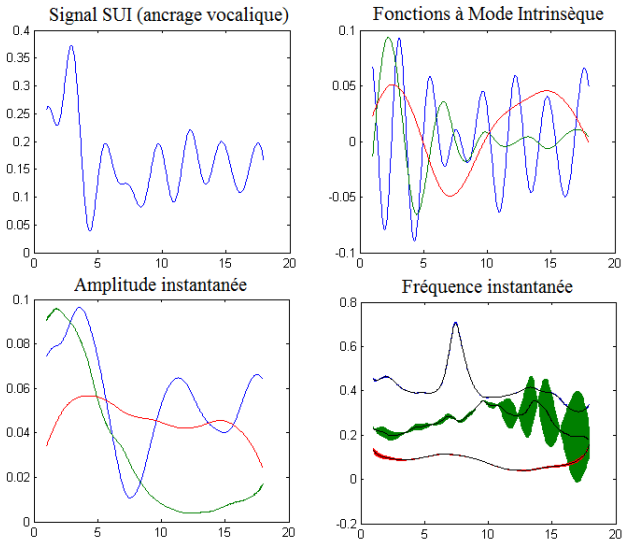


Fig. 4.9 Extraction de l'amplitude et de la fréquence instantanées sur un signal SUI par la THH ; figure extraite de [RIN09][55].

$$x(t) = \sum_{k=1}^{N} d_k(t) + r_k(t) \qquad [27]$$

possibles excursions locales plus larges. Les valeurs des coefficients sont fixées par défaut à : α = 0.05, θ_1 = 0.05 et θ_2 = 10. La méthode EMD peut être résumée telle que suit [HUA98][56] :

1. Extraire tous les extrema de $x(t)$;
2. Interpoler entre les minima (resp. maxima) pour obtenir deux enveloppes : $e_{min}(t)$ et $e_{max}(t)$;
3. Calculer la moyenne : $m(t) = \big(e_{min}(t) + e_{max}(t)\big)/2$;
4. Extraire le détail: $d(t) = x(t) - m(t)$;
5. Itérer sur le détail tant que les conditions du tamisage ne sont pas satisfaites.

La seconde étape de la THH est la transformée d'Hilbert. En partant du constat que les IMFs sont des signaux à bandes de fréquences limitées, Huang *et al.* ont proposé d'appliquer la transformée d'Hilbert (TH) sur les IMFs [HUA98]. Cette transformée permet d'extraire à la fois la fréquence instantanée et l'enveloppe temporelle du signal réel $x(t)$ dans le domaine temps-fréquence. Elle présente également l'avantage de traiter des signaux non-linéaires et non-stationnaires tels que ceux du rythme. La TH de $x(t)$ est donnée par l'équation [28]. Le signal analytique de $x(t)$ peut être définie par $z(t)$ [29]. L'enveloppe du signal $x(t)$ est alors fourni par l'amplitude $a(t)$ de $z(t)$, tandis que la fréquence instantanée est obtenue par la dérivée de la phase [30].

Des précautions doivent être prises en compte lors du calcul de la THH sur les signaux SUI. En effet, des valeurs de fréquences négatives peuvent apparaître en début et fin du signal généré en raison d'observations unitaires et finies, cf. Fig. 4.9. Afin de résoudre ce problème, nous avons introduit une série de zéros au début et à la fin de la série étudiée de façon à dé-

$$y(t) = \frac{1}{\pi} p.v. \int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} d\tau \qquad [28]$$

avec $\boldsymbol{p.v.}$ désignant la valeur principale de Cauchy.

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)}$$
$$\text{avec,} \quad a(t) = \sqrt{x^2(t) + y^2(t)}\,;\ \theta(t) = \arctan\left(\frac{y(t)}{x(t)}\right) \qquad [29]$$

$$f(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt} \qquad [30]$$

placer le problème de fréquences négatives sur ces données qui seront supprimées par la suite. Enfin, notons que des fluctuations locales dans la phase apparaissent à cause du sur-échantillonnage des signaux SUI. Ces fluctuations entraînent par conséquence, des variations beaucoup plus importantes dans les valeurs de fréquences instantanées puisqu'elles sont issues du calcul d'une dérivée [30]. Cinq passes d'un filtre moyenneur travaillant sur 3 échantillons ont permis de réduire jusqu'au possible ces erreurs.

L'ensemble de caractéristiques extraites par cette méthode se compose de mesures du spectre du signal SUI (entropie, barycentre et fréquence moyenne [26]) et de la THH de ce signal : amplitudes [29] et fréquences instantanées [30] issues des 3 premières IMFs et de la somme de ces IMFs, ainsi que la fréquence moyenne (*MNF*) obtenue par le calcul proposé par [XIE06][60] [31]. Notons que la THH a déjà été utilisé avec succès par notre équipe pour effectuer la reconnaissance des émotions au moyen de divers types de signaux physiologiques [ZON09][61].

$$WMIF(i) = \frac{\sum_{j=1}^{N} f_i(j)\, a_i^2(j)}{\sum_{j=1}^{N} a_i^2(j)} \quad ; \quad MNF = \frac{\sum_{i=1}^{N} \|a_i\|\, WMIF(i)}{\sum_{i=1}^{N} \|a_i\|} \qquad [31]$$

Variabilité prosodique

Dans l'optique de caractériser le rythme à travers la dynamique de la prosodie (i.e., pitch, énergie et qualité vocale), nous avons proposé dans [RIN08d][61] d'exploiter une extension du PVI [GRA02a][36] dans laquelle nous remplaçons la mesure d'intervalle de durée par celle de la dispersion relative d'un LLD prosodique. Les coefficients de variation c_v sont calculés par le rapport entre l'écart-type σ et la moyenne μ des données présentes sur chaque segment de parole. Nous avons aussi inclus un facteur de normalisation pour tenir compte des durées d_k et

[59] C. Zong et M. Chetouani, Hilbert-Huang transform based physiological signals analysis for emotion recognition, dans proc. *ISSPIT*, Ajman, UAE, Dec. 14-17 2009, pp. 334-339.

[60] H. Xie et Z. Wang, "Mean frequency derived via hilbert-huang transform with application to fatigue emg signal analysis", dans *Computer Methods and Programs in Biomedicine*, vol. 82, no. 2, pp. 114-120, May 2006.

[61] F. Ringeval, M. Chetouani, D. Sztahó et K. Vicsi, "Automatic prosodic disorders analysis for impaired communication children", dans proc. *1st W. on Child, Computer and Interaction*, Chaniaa, Greece, Oct. 2008.

$$P\text{-}PVI = \frac{1}{N-1}\sum_{k=1}^{N-1}\frac{d_k d_{k+1} I_k}{d_k + d_{k+1} + I_k}\left|c_{v_k} - c_{v_{k+1}}\right| \quad [32]$$

$$\text{avec,} \quad c_v = \frac{\sigma}{\mu}$$

d_{k+1} et de l'intervalle I_k séparant les paires de segments consécutifs qui sont exploités dans la métrique du P-PVI [32].

La valeur de ce paramètre est nulle si la dispersion du LLD prosodique mesuré est identique sur deux segments consécutifs de parole, ce qui correspond à une certaine forme de monotonie dans la composante prosodique. Dans le cas contraire, les valeurs dépendent de façon proportionnelle à l'importance du changement présent dans la dispersion statistique du LLD entre les segments de parole. Ainsi, un maximum définit une proéminence dans la dispersion statistique du LLD prosodique. Notons que les valeurs du P-PVI dépendent également de la durée des segments comparés et de l'intervalle qui les sépare. Par conséquent, la valeur est d'autant plus élevée que les durées d_k et d_{k+1} des segments comparés est longue, et de même pour celle de l'intervalle I_k séparant ces segments. Ces effets sont alors cumulatifs.

La mesure du P-PVI permet donc de caractériser les changements dans la dispersion des données d'un LLD prosodique et à travers des paires de segments consécutifs. Toutefois, il pourrait s'avérer pertinent de définir une métrique du rythme qui compare d'un seul bloc plusieurs composantes prosodiques tout en prenant en compte les paramètres de durée ainsi que les corrélations entre les LLDs. Cette requête peut être satisfaite par la distance de Hotteling.

Distance prosodique

La distance de Hotteling (HD) est une mesure qui permet de comparer la distribution statistique de deux ensembles de données par un calcul similaire à la distance de Mahalanobis. Elle fait notamment intervenir un facteur de normalisation par la durée d des deux segments analysés [33]. Nous avons appliqué ce calcul sur les LLDs extraits à travers des ancrages acoustiques donnés de façon à calculer la dynamique prosodique. Le principe de notre méthode est illustré dans la figure d'introduction de ce chapitre. Les valeurs fournies par la HD sont nulles lorsque les distributions statistiques des LLDs sont identiques sur les paires de segments consécutifs comparées, et positives dans tous les autres cas de figure. Les valeurs varient proportionnellement selon l'importance du changement présent dans la distribution statistique des LLDs et selon la durée des segments comparés. Toutefois, comme la durée de l'intervalle I_{ij} qui sépare les deux segments *i* et *j* n'est pas incluse dans le calcul [33], nous proposons d'inclure cette donnée dans la métrique que nous appelons distance prosodique de Hotteling (PHD) [34]. Deux techniques différentes ont été employées pour définir la matrice d'autocorrélation Σ : (i) la première consistait à remplir la diagonale par les écarts-types des données issues de chacune des *N* LLDs prosodiques [35], et la seconde technique exploitait toutes les valeurs issues de la matrice de covariance des données [36].

La fonction A-PHD se comporte donc de façon similaire à la HD, si ce n'est qu'elle inclut en plus dans le facteur de normalisation k, la durée séparant les paires de segments consécu-

$$HD_{ij} = \frac{d_i d_j}{d_i + d_j}\left[\left(\mu_i - \mu_j\right)^T \Sigma_{i\cup j}^{-1}\left(\mu_i - \mu_j\right)\right] \quad [33]$$

avec, $i \cup j$ l'union des données issues de deux ancrages consécutifs i et j, d_i et d_j la durée respective de ces segments, et $\Sigma_{i\cup j}^{-1}$ la matrice de covariance inverse.

$$PHD_{ij} = k\left[\left(\mu_i - \mu_j\right)^T \Sigma_{i\cup j}^{-1}\left(\mu_i - \mu_j\right)\right] \quad [34]$$

$$A\text{-}PHD_{ij} = k\left[\begin{pmatrix} \mu_i^1 - \mu_j^1 \\ \dots \\ \mu_i^N - \mu_j^N \end{pmatrix}^T \begin{pmatrix} \sigma_{i\cup j}^1 & 0 & 0 \\ 0 & \sigma_{i\cup j}^x & 0 \\ 0 & 0 & \sigma_{i\cup j}^N \end{pmatrix} \begin{pmatrix} \mu_i^1 - \mu_j^1 \\ \dots \\ \mu_i^N - \mu_j^N \end{pmatrix}\right] \quad [35]$$

$$I\text{-}PHD_{ij} = k\left[\begin{pmatrix} \mu_i^1 - \mu_j^1 \\ \dots \\ \mu_i^N - \mu_j^N \end{pmatrix}^T \begin{pmatrix} \sigma_{i\cup j}^1 & \sigma_{i\cup j}^{1,x} & \sigma_{i\cup j}^{1,N} \\ \sigma_{i\cup j}^{x,1} & \sigma_{i\cup j}^x & \sigma_{i\cup j}^{x,N} \\ \sigma_{i\cup j}^{N,1} & \sigma_{i\cup j}^{N,x} & \sigma_{i\cup j}^N \end{pmatrix} \begin{pmatrix} \mu_i^1 - \mu_j^1 \\ \dots \\ \mu_i^N - \mu_j^N \end{pmatrix}\right] \quad [36]$$

$$\text{avec,} \qquad k = \frac{d_i * d_j * I_{ij}}{d_i + d_j + I_{ij}}$$

tifs dans le facteur de normalisation. Ainsi, elle prend en compte (de façon indépendante) les données de plusieurs composantes prosodiques dans le calcul de la métrique du rythme. La mesure A-PHD fournit donc des valeurs qui sont proportionnellement liées à la somme des variations observées dans la distribution statistique de chaque LLD de la prosodie et entre des paires de segments consécutifs. Les valeurs sont élevées si la distribution d'un des LLD varie fortement à travers les segments (et d'autant plus si leur durée et leur intervalle est grand), et maximales si tous les paramètres présentent de très grandes variations (les contributions de chaque LLD sont sommées). Enfin, la mesure I-PHD se comporte de la même façon que la mesure A-PHD, hormis le fait qu'elle prend aussi en compte les corrélations entre les composantes prosodiques.

Comme nous l'avons précisé dans la table 4.2, certaines méthodes de caractérisation du rythme nécessitent la présence d'un certain nombre de segments de parole pour être calculées. Cette contrainte est d'autant plus flagrante pour les mesures P-PVI, HD et PHD puisqu'elles requièrent d'avoir en plus des valeurs pour les LLD prosodiques. La f0 et les formants ne sont pas toujours présents dans les données, contrairement aux valeurs d'énergie. Ainsi, nous avons dû restreindre les données fournies par nos LLDs aux instants voisés pour calculer la mesure I-PHD. Par ailleurs, nous avons réalisé une interpolation linéaire entre les valeurs du LLD de la qualité vocale (i.e., aire formantique) sur les paires de segment consécutif de façon à obtenir un échantillon toutes les 10ms, comme pour le pitch et l'énergie. Fait à noter, les mesures issues de la HD sont très peu coûteuses en temps de calcul et peuvent donc être estimées en temps réel. D'ailleurs, de précédents travaux portant sur la génération de signaux de rétroaction par des systèmes interactifs ont déjà utilisé cette mesure [ALM09][62].

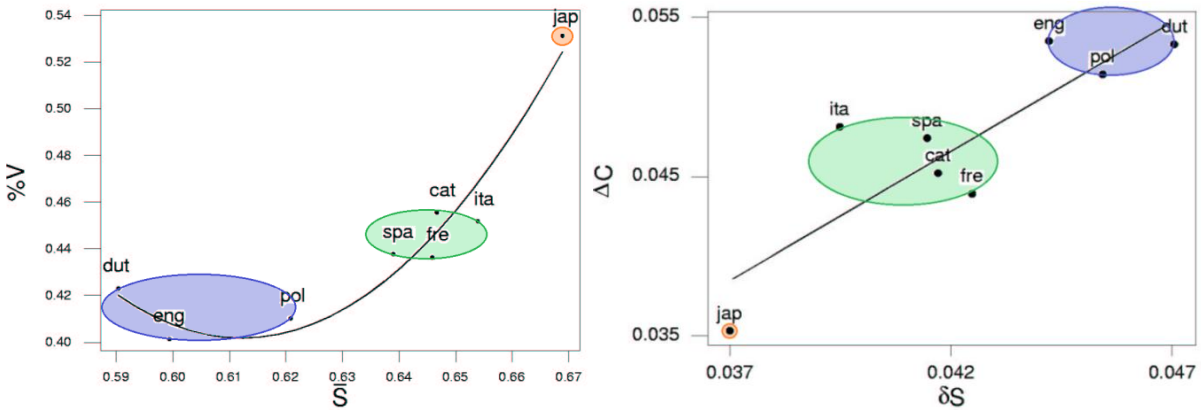


Fig. 4.10 Discrimination des langues au moyen des paramètres proposés par Grabe (figure en haut à gauche) et corrélations avec les paramètres de Ramus (figure en haut à droite et en bas) ; *jap* : Japonais ; *ita* : Italien ; *spa* : Espagnol ; *cat* : Catalan ; *fre* : Français ; *eng* : Anglais ; *pol* : Polonais ; et *dut* : Allemand ; le code couleur correspond au suivant : bleu, langues accentuels ; vert, langues syllabiques ; et orange, langue morique ; figure reproduite de [GAL02][54].

Mesure de sonorité

Galves *et. al* ont considéré que la discrimination des langues par le rythme ne pourrait reposer sur une distinction fine entre les voyelles et consonnes [GAL02][54], puisque les nouveaux nés sont capables de distinguer ces langues via un signal filtré à 400Hz [MEH96][63]. La différenciation des langues par le rythme serait donc plutôt grossière chez ces derniers et s'effectuerait, selon les auteurs, par la perception de l'intensité par opposition à ce qu'ils définissent comme l'obstruence, i.e., de faibles écarts dans la dynamique spectrale [GAL02]. Ces derniers ont donc proposé de définir une mesure de la sonorité. Cette mesure prend des valeurs comprises entre [0 ; 1], et proches de 1 pour des zones comportant des motifs réguliers caractéristiques des portions sonores (e.g., les segments voisés), et proches de 0 pour les zones définies comme obstruentes.

La fonction de sonorité repose sur la divergence de Kullback-Leibler qui est estimée entre deux vecteurs de probabilités. Ces vecteurs sont fournis par le spectre de puissance d'un signal de parole, après une étape de normalisation des coefficients de Fourrier $c_t(i)$ [37], fréquence i et instant t. Les motifs du rythme sont ensuite capturés à travers les valeurs de probabilité $p_t(i)$ par le calcul de la divergence de Kullback-Leibler $h(p_t \mid p_{t-1})$ sur deux trames consécutives du signal de parole [38]. Comme l'entropie relative est toujours positive (inégalité de Jensen) et proche de 0 quand les mesures sont similaires, la fonction de sonorité $s(t)$ a donc été définie par le calcul [39]. Elle est caractérisée par la valeur moyenne de ses valeurs et de celles issues de la dérivée à l'ordre 1 [40]. L'intérêt de la fonction $s(t)$ repose sur le fait que son estimation ne nécessite pas de segmentation en voyelles / consonnes. De plus, les mesures $\bar{S}$ et δS ont pu être corrélées avec celles proposées par Ramus *et al.*, i.e., %V et ΔC, cf. Fig. 4.10.

---

[62] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J. C. Martin, S. Ondas, C. Pelachaud, J. Urbain et M. Yilmaz, "Generating robot/agent backchannels during a storytelling experiment", dans proc. *IEEE Inter. C. on Rob. and Automation*, Kobe, Japan, May 12-17 2009, pp. 2477–2482.

[63] J. Mehler, E. Dupoux, T. Nazzi, et G. Dehaene-Lambertz, "Coping with linguistic diversity: the infant's viewpoint", dans *Signal to syntax: bootstrapping from speech to grammar in early acquisition*, J.L. Morgan and K. Demuth [Eds], Erlbaum, Mahwah (NJ), USA, 1996.

$$p_t(i) = \frac{c_t(i)^2}{\sum_f c_t(f)^2} \quad [37]$$

$$h(p_t \mid p_{t-1}) = \sum_i p_t(i) \log\left(\frac{p_t(i)}{p_{t-1}(i)}\right) \quad [38]$$

$$s(t) = 1 - \min\left(1, \frac{1}{27} \sum_{u=t-4}^{t+4} \sum_{i=1}^{3} h(p_u \mid p_{u-i})\right) \quad [39]$$

$$\bar{S} = \frac{1}{T}\sum_{t=1}^{T} s(t)\ ,\ \delta S = \frac{1}{T}\sum_{t=1}^{T} |s(t) - s(t-1)| \quad [40]$$

2.3. Reconnaissance statique / dynamique

La littérature [SCH09b][64] fait apparaître deux types d'approche pour effectuer la reconnaissance des émotions à travers les LLDs de la prosodie : (i) la méthode *statique* exploitant un ensemble de mesures statistiques et (ii) la méthode *dynamique* dans laquelle un HMM est optimisé pour traiter directement les LLDs. Comme les résultats sont généralement comparables voire meilleurs pour la méthode *statique* [SCH09b], nous avons uniquement utilisé cette méthode dans les expériences à venir. Toutefois, la méthode *dynamique* des HMM a été employée dans le chapitre 5 pour caractériser des profils intonatifs sur différents types de phrase (e.g., question / affirmation). De plus, notons que les métriques P-PVI, HD et PHD du rythme correspondent à des mesures dynamiques de la prosodie.

Les mesures statistiques qui ont été utilisées pour effectuer la caractérisation *statique* des LLDs prosodiques, sont données dans la table 4.3. Cet ensemble de mesures est très varié et contient non seulement les traditionnelles (e.g., le maximum, le minimum, les quatre premiers moments statistiques et les quartiles), mais aussi des coefficients de perturbations (e.g., *jitter* pour le pitch, et *shimmer* pour l'énergie), des caractéristiques dérivées du modèle *Rise-Fall-Connection* [TAY94][65] (e.g., les positions relatives des valeurs minimales et maximales) et celles issues des systèmes de détection de question (e.g., la proportion / moyenne de valeurs montantes / descendantes) [QUA07a][66]. Ces paramètres ont été utilisés tels qu'ils sont définis dans la littérature, excepté pour les coefficients de perturbation vu qu'il existe quantité de modèles. Après en avoir essayé plusieurs, nous avons retenu celui qui consiste à moyenner les écarts entre les valeurs d'interpolation de la courbe du LLD prosodique et la courbe d'origine. L'ordre du polynôme qui a servi à interpoler les données varie de 1 à 3 selon la disponibilité des données (au moins *n*+1 valeurs différentes pour un polynôme d'ordre *n*).

[64] B. Schuller, S. Steidl, et A. Batliner, "The Interspeech 2009 emotion challenge", dans proc. *Interspeech*, Brighton, United-Kingdom, 2009.

[65] P. Taylor, "The Rise/Fall/Connection model of intonation", dans *Speech Communication*, vol. 15, issue 1-2, pp. 169-186, Oct. 1994.

Table 4.3 Ensemble de 27 mesures statistiques utilisées pour la modélisation statique de la prosodie.

Mesure	Description
Max	Valeur du maximum
RPmax	Position relative du maximum
Min	Valeur du minimum
RPmin	Position relative du minimum
RP_AD	Différence absolue entre RPmax et RPmin
Range_n	Amplitude divisée par RP_AD
Mean	Valeur moyenne
STD	Valeur d'écart-type
Skewness	Moment statistique du 3ème ordre
Kurtosis	Moment statistique du 4ème ordre
Q1	Valeur du premier quartile
Median	Valeur médiane
Q3	Valeur du troisième quartile
IQR	Ecart inter quartiles
IQR_STD_AD	Différence absolue entre IQR et STD
Jitter / Shimmer	Coefficient de perturbation du pitch / énergie
Slope	Premier coefficient de la pente de régression
OnV	Valeur d'onset (i.e., valeur de départ)
TaV	Valeur cible (i.e., valeur centrale)
OfV	Valeur d'offset (i.e., valeur de fin)
TaVOnV_AD	Différence absolue entre TaV et OnV
OfVOnV_AD	Différence absolue entre OfV et OnV
OfVTaV_AD	Différence absolue entre OfV et TaV
%↑	Proportion de valeurs montantes
%↓	Proportion de valeurs descendantes
μ↑	Valeur moyenne des écarts montants
μ↓	Valeur moyenne des écarts descendants

3. Système de reconnaissance

Nous proposons d'utiliser dans cette thèse, une stratégie de reconnaissance des émotions qui repose sur le célèbre principe de « diviser pour mieux régner » [BAT99][67]. Par conséquent, l'objectif des expériences de ce chapitre consiste à estimer (pour différents points d'ancrages complémentaires de la parole), les contributions apportées par les LLDs de la prosodie dans la reconnaissance des émotions. Nous fusionnons pour cela, les informations apportées par les différentes composantes prosodiques (i.e., pitch, énergie, qualité vocale et rythme), selon les points d'ancrages acoustiques complémentaires de la parole (e.g., voyelle, consonne), cf. Fig. 4.11. Le principe de fusion utilisé dans cette thèse est décrit dans la sous-section 3.4 du cha-

[66] V. M. Quang, L. Besacier et E. Castelli, "Automatic question detection: prosodic-lexical features and crosslingual experiments", dans proc. Interspeech ICSLP, Antwerp, Belgium, Aug. 27–31 2007, pp. 2257–2260.

[67] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth et H. Niemann, "Prosodic feature evaluation: brute force or well designed?", dans proc. *14th ICPhS*, San Francisco, (CA), USA, Aug. 1999, pp. 2315–2318.

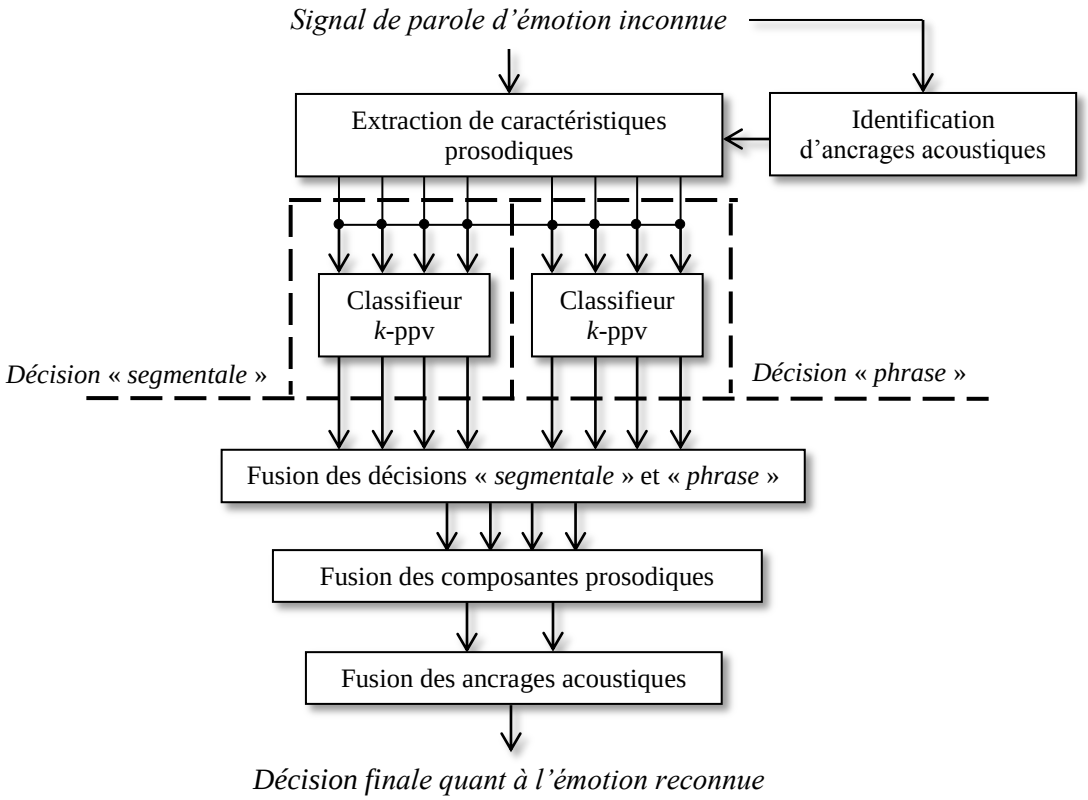


Fig. 4.11 Architecture du système de reconnaissance prosodique de la parole affective.

pitre 3. La contribution des composantes acoustiques (MFCC) et prosodiques (mesures statistiques des LLDs) en reconnaissance ont également été estimées en fusionnant leurs vecteurs de probabilités respectifs. Nous avons aussi porté une attention particulière aux modèles du rythme qui ont été présentés dans la section précédente. En effet, l'objectif des expériences est non seulement d'étudier la contribution des différentes composantes de la prosodie en reconnaissance d'émotions, mais aussi de s'assurer que les modèles du rythme que nous avons proposés sont également pertinents. Une première preuve pourrait par exemple être fournie au moyen d'un algorithme de sélection de caractéristiques qui retient les paramètres les plus discriminants en regard des classes d'émotion (e.g., *bottom-up*). Une analyse statistique des paramètres du rythme selon les classes d'émotion pourrait fournir une seconde preuve de leur pertinence.

3.1. Architecture

L'architecture du système employé dans ce chapitre est quasiment identique à celle du chapitre 3, cf. sous-section 3.1, si ce n'est que nous avons exploité qu'un seul type de classifieur (*k*-ppv) et qu'aucune fusion n'a donc été effectuée à ce niveau-là. Les décisions « *segmentale* » et « *phrase* » occasionnent une variation dans le nombre de LLDs pouvant être calculés pour les paramètres du rythme, cf. table 4.4. Ces mesures sont bien souvent fournies de façon unitaire pour chaque segment de parole. Ainsi, 1172 mesures prosodiques ont pu être calculées pour la décision « *phrase* » contre seulement 288 pour celle « *segmentale* », cf. table 4.5. Notons que le développement intensif des modèles *non-conventionnels* du rythme a con-

Table 4.4 Nombre de mesures disponibles sur les LLDs du rythme selon le type de décision.

LLD rythmique	*phrase*	*segmentale*
D_{seg} ; D_{intra} ; D_{inter}	81	3
RR_{intra} ; RR_{inter}	54	2
$rPVI_{intra/inter}$; $nPVI_{inter}$	81	3
prcSU	1	1
$\bar{R}$	1	1
$VarCo_{intra/inter}$	2	2
HD ; *PHD* ; *A/I-PHD*	594	22
$entropie_{lf}$; $barycentre_{lf}$; f_moy_{lf}	3	3
$module_{emd}$; $freq_{emd}$; f_moy_{emd}	55	55
$entropie_{SUI}$; $barycentre_{SUI}$; f_moy_{SUI}	3	3
P-PVI	81	3

Table 4.5 Nombre de mesures statistiques utilisées pour la reconnaissance statique de la prosodie selon les composantes et le type de décision.

Décision	Pitch	Energie	Qualité vocale	Rythme	Total
phrase	108	108	243	713	**1172**
segmentale	30	30	139	89	**288**

duit à un déséquilibre dans le taux de représentativité des dimensions prosodiques. De façon à limiter ce déséquilibre, nous avons distribué certaines mesures du rythme (P-PVI, HD et PHD) calculé sur le pitch, l'énergie ou la qualité vocale dans leurs composantes respectives.

Les mesures statistiques du pitch et de l'énergie reposent directement sur leurs valeurs, i.e., nous n'avons pas calculé les dérivées Δ et ΔΔ pour laisser les mesures de dynamique aux modèles du rythme. La qualité vocale est représentée par 6 LLDs qui sont fournis par les valeurs des deux premiers formants (fréquence et énergie), de la mesure d'harmonicité REC et de la valeur d'aire formantique, cf. sous-section 2.1.3. Les LLDs du rythme sont plus nombreux et incluent : (i) la durée segmentale, (ii) les intervalles de durée *inter*, (iii) les variations de durée entre segments successifs *intra* et (iv) les métriques *conventionnelles* et *non-conventionnelles* (excepté la mesure de sonorité).

3.2. L'approche bottom-up

Afin d'identifier les paramètres prosodiques qui seront corrélés aux émotions, nous avons utilisé une approche de type *bottom-up* pour effectuer la reconnaissance des émotions. Cette approche traite les données en deux étapes. La première étape consiste à évaluer la pertinence de chaque paramètre en évaluant leur score de reconnaissance respectif. Les paramètres sont ensuite triés par ordre décroissant selon leur valeur de score. La seconde étape consiste ensuite à réaliser une boucle sur les *N* paramètres qui sont successivement insérés dans un super-vecteur de caractéristiques. Un test de reconnaissance est effectué à chaque itération sur le super-vecteur et seuls les paramètres qui permettent d'améliorer les scores sont conservés. Nous avons préféré l'approche *bottom-up* plutôt qu'une méthode *flottante* (combinaison d'une approche *montante* et *descendante*, cf. chapitre 3, sous-section 4.1), puisque son temps de

calcul n'est pas compatible avec les analyses que nous souhaitons effectuer sur les nombreux points d'ancrage de la parole ; complexité en $O(n!)$ contre $O(n)$ pour l'approche *bottom-up*.

3.3. Stratégies de reconnaissance

Nous avons employé deux stratégies pour effectuer la reconnaissance des émotions par la prosodie et selon les points d'ancrages acoustiques. La première stratégie exploite un super-vecteur de caractéristiques pour chaque composante prosodique et des techniques de fusion sont employées pour estimer leurs contributions respectives dans la tâche de reconnaissance. Alors que la seconde stratégie utilise toutes les données disponibles pour définir le super-vecteur. Avant d'effectuer les tests en reconnaissance prosodique, les LLDs ont été normalisés par la même méthode que pour les coefficients MFCC, i.e., par le calcul du *z-score* [11] et selon cinq types de normalisation : (i) *Raw* : aucune normalisation des données, (ii) *Z-tout* : normalisation sans information, (iii) *Z-genre* : normalisation des données selon le genre du locuteur, (iv) *Z-locuteur* : normalisation des données selon le locuteur et (v) *Z-phrase* : normalisation des données selon la phrase prononcée. Cette étape permet d'étudier la pertinence du type d'informations exploitées pour normaliser les données.

3.4. Méthodes d'exploration de données

Les méthodes d'exploration de données utilisées dans les expériences de ce chapitre sont en tous points identiques (i.e., mêmes listes de fichiers) à celles utilisées pour les MFCC, cf. chapitre 3, sous-section 3.5. Nous avons ainsi exploité : (i) la méthode de *cross-validation stratifiée* (CVS) qui permet d'effectuer un test d'indépendance aux classes d'émotions, et (ii) la méthode du *leave-one-speaker-out* (LOSO) qui teste quant à elle, l'indépendance locuteur.

4. Reconnaissance prosodique

Les paragraphes suivants présentent les résultats qui ont été obtenus en reconnaissance prosodique des émotions du corpus Berlin. Comme le nombre de données disponibles peut fortement varier selon le point d'ancrage acoustique analysé, nous avons inséré trois niveaux de vérifications dans le système pour s'assurer de la pertinence des résultats : (i) toutes les émotions doivent être représentées dans les données d'apprentissage, (ii) le taux de présence (ou calculabilité) des paramètres doit être supérieur à 75% de l'ensemble des phrases disponibles et (iii) le système doit pouvoir retourner des vecteurs de probabilités sur au moins 75% des phrases disponibles pour un paramètre. Tous ces éléments ont dû être vérifiés simultanément lors des tests, pour valider les résultats produits par le système de reconnaissance. Le contexte suprasegmentale conduit à des rejets dans les tests effectués sur les points d'ancrage qui sont faiblement représentés dans les données (e.g., voyelles longues). Ces rejets sont indiqués dans les tables des scores par le symbole D.I., i.e., données indisponibles. Enfin, rappelons que l'objectif principal des expériences qui vont suivre consiste à estimer la contribution des points d'ancrages acoustiques de la parole et des paramètres extraits sur ces derniers, notamment les modèles du rythme, dans la tâche de reconnaissance automatique d'émotions.

4.1. Tests en cross-validation croisée et stratifiée

Afin de faire ressortir les résultats que nous avons évalués comme probants parmi toutes les configurations étudiées, nous avons fixé pour l'analyse un seuil à 60% sur les scores issus de la cross-validation stratifiée (CVS). Ce seuil équivaut à ~90% du meilleur score obtenu par l'algorithme des *k*-ppv sur le corpus Berlin. Les paragraphes suivants décrivent les résultats qui ont été obtenus dans un contexte d'indépendance aux classes (méthode CVS).

4.1.1. Approche par composante

Les étapes de normalisations, i.e., *Z-tout*, *Z-genre*, Z-locuteur et *Z-phrase*, produisent une amélioration des scores plus significative sur les paramètres prosodiques que sur les coefficients MFCC. Cette amélioration est principalement portée par l'emploi des informations linguistiques pour les paramètres du rythme, ce qui montre l'existence d'un lien entre ces deux types d'informations en regard des émotions actées. Les meilleurs scores sont dans une immense majorité produits par l'approche « *phrase* » et sont fournis par le pitch, ce qui confirme sa position dominante dans la communication des expressions affectives de la voix [GUS-01][68] (63%, segments voisés, normalisation « *Z-locuteur* »). Les autres composantes prosodiques qui viennent dans l'ordre des meilleurs scores sont : la qualité vocale (57%, ancrage sonorante, normalisation « *Z-genre* »), le rythme (53%, ancrage voisé, normalisation « *Z-phrase* ») et l'énergie (48%, ancrages voisés, toutes les normalisations).

Les résultats issus de la fusion des décisions « *segmentale* » et « *phrase* » montrent que les améliorations des scores sont alors quasi-nulles pour les quatre composantes de la prosodie et que l'approche « *phrase* » est très dominante par rapport à celle « *segmentale* ». La fusion des composantes prosodiques permet quant à elle d'améliorer significativement les scores par rapport à ceux obtenus sur le pitch (amélioration moyenne) : *Raw* : +21%, *Z-tout* : +22%, *Z-genre* : +17%, *Z-locuteur* : +16% et *Z-phrase* : +30%. Le meilleur score est alors obtenu par les syllabes voisées (73%, normalisation « *Z-locuteur* »), vient ensuite les segments voisés (72%, même normalisation), les syllabes (71%, […]), les pseudo-voyelles (70%, […]), etc., cf. table 4.6. Ces dernières fournissent des scores qui sont alors largement supérieurs à ceux fournis par les voyelles issues des transcriptions phonétiques. Les ancrages consonantique et rythmique sont ensuite les seuls à produire des scores supérieurs à 60%.

La Fig. 4.12 illustre les poids retournés par la fusion des composantes prosodiques pour les normalisations qui ont conduit aux meilleurs scores. Les résultats montrent que les configurations de fusion les plus pertinentes varient fortement selon les ancrages acoustiques, même si certains d'entre eux sont de la même famille phonétique, e.g., classe vs. macro-classe. De nombreuses particularités apparaissent ainsi selon les ancrages : (i) l'importance du pitch, de la qualité vocale et du rythme sont équivalentes pour les ancrages voisés tandis que celle de l'énergie est moindre, (ii) le poids attribué au rythme est supérieur à 50% pour les ancrages non-voisés et les syllabes voisées (meilleur score de reconnaissance), et même supérieur à 80% pour les syllabes non-voisées, mais il est négligeable lorsque l'on considère les

[68] S. Gustafson-Capková, "*Emotions in speech: Tagset and acoustic correlates",* dans *Term paper in Speech Technology*, GSLT, Stockholm University, Depart. of Ling., Aut. 2001.

Table 4.6 Scores en reconnaissance prosodique des émotions obtenus par la fusion des composantes prosodiques.

Ancrage acoustique	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé	**61** 5/1/2/2	**67** 3/0/5/2	**70** 3/1/3/3	**72** 3/1/3/3	**67** 3/0/3/4
non-voisé	55 0/2/3/5	58 0/4/1/5	58 0/3/3/5	**60** 0/2/3/5	**62** 0/3/1/6
voyelle	52 3/1/2/4	55 2/1/4/3	**60** 3/0/5/2	**63** 4/1/4/1	**60** 2/0/3/5
p-voyelle	56 3/2/1/4	59 3/1/4/2	**67** 4/0/4/2	**70** 3/0/5/1	**62** 4/0/4/2
voyelle courte	49 4/0/2/4	54 4/0/4/3	60 5/0/4/0	**65** 4/0/6/1	55 2/0/5/3
voyelle longue	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
voyelle diphtongue	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
consonne	56 3/3/1/3	59 4/1/3/2	54 3/3/1/3	**60** 4/2/2/2	59 3/2/2/3
p-consonne	53 2/6/2/0	55 3/5/0/1	**63** 4/5/1/1	**65** 3/5/0/2	**61** 1/5/0/4
consonne plosive	48 0/3/4/2	42 0/4/1/5	41 0/4/4/3	44 0/7/0/3	41 0/4/1/6
consonne fricative	36 0/5/0/5	36 0/5/5/0	40 0/1/8/0	44 1/3/5/1	40 0/4/4/3
consonne sonorante	53 3/4/4/0	54 2/2/6/0	**61** 1/0/9/0	58 4/1/5/0	**61** 2/2/5/2
syllabe	59 3/3/1/3	**60** 4/2/3/1	**68** 5/0/3/3	**71** 4/2/3/1	**63** 3/2/4/2
syllabe voisée	**63** 4/2/0/4	**63** 4/1/4/1	**72** 5/1/3/2	**73** 3/0/3/5	**70** 3/0/3/4
syllabe non-voisée	42 0/4/4/3	53 0/4/1/5	54 0/3/2/5	53 0/3/1/6	58 0/1/1/7
syllabe partie C	32 0/1/7/1	34 0/0/1/0	34 0/0/1/0	35 3/5/0/2	42 4/1/0/5
syllabe partie V	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
syllabe partie CV	33 2/6/2/0	32 0/9/1/0	35 4/6/0/0	36 6/2/3/0	33 4/4/2/0
syllabe partie VC	34 5/1/4/0	34 1/8/0/0	40 8/0/2/0	41 7/0/2/1	38 3/0/0/7
syllabe partie CVC	51 5/2/3/0	50 4/1/4/0	55 5/1/4/0	53 4/2/3/0	54 4/1/4/1
« p-centres » 1	44 2/2/2/4	45 4/1/4/1	53 6/0/4/0	51 3/0/4/3	50 2/2/4/2
« p-centres » 2	50 5/0/2/2	54 2/1/5/1	56 2/0/6/2	**63** 2/2/3/3	58 2/1/4/3
« p-centres » 3	55 3/2/1/4	57 4/0/3/2	**63** 4/0/5/2	**65** 3/2/4/1	**63** 1/2/4/3

[valeur en %] $_{\text{poids}}$ α_p : pitch / énergie / qualité vocale / rythme ; approche composante ; méthode CVS.

Table 4.7 Scores en reconnaissance prosodique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**64** 3/7	**69** 6/4	**74** 6/4	**75** 5/5	**72** 7/3
voyelle / consonne	59 2/8	**63** 3/7	**65** 7/3	**67** 6/4	**64** 4/6
p-voyelle / p-consonne	59 3/7	**63** 4/6	**71** 5/5	**73** 5/5	**65** 3/7
phonèmes voyelles	49 1/0/0	54 1/0/0	**60** 1/0/0	**65** 1/0/0	55 1/0/0
phonèmes consonnes	**63** 3/2/4	59 5/0/5	**63** 4/0/6	**62** 3/3/4	**61** 0/0/1
phonèmes	**64** 1/9	**63** 4/6	**69** 2/8	**70** 3/7	**61** 4/6
syllabes voisées / non-voisées	**64** 8/2	**64** 6/4	**72** 1/0	**73** 1/0	**70** 1/0
syllabes V/C/CV/VC/CVC	51 1/0/1/2/7	50 0/1/0/1/8	55 0/0/1/1/8	53 1/0/0/0/9	54 1/0/0/0/9

[valeur en %] $_{\text{poids}}$ α_a ; approche composante ; méthode CVS.

deux, i.e., les syllabes et (iv) la qualité vocale intervient de façon quasi-exclusive sur les sonorantes, cf. Fig. 4.12. Nous pouvons résumer ces résultats en disant que le pitch apparaît en moyenne comme la composante la plus importante dans la caractérisation des émotions en CVS ; vient ensuite la qualité vocale, le rythme et l'énergie. Notons que l'ordre de ces composantes prosodiques reste inchangé comparé à une analyse séparée des meilleurs scores, cf. premier paragraphe de cette sous-section.

La table 4.7 présente les résultats issus de la dernière étape de fusion, cf. Fig. 4.11. La

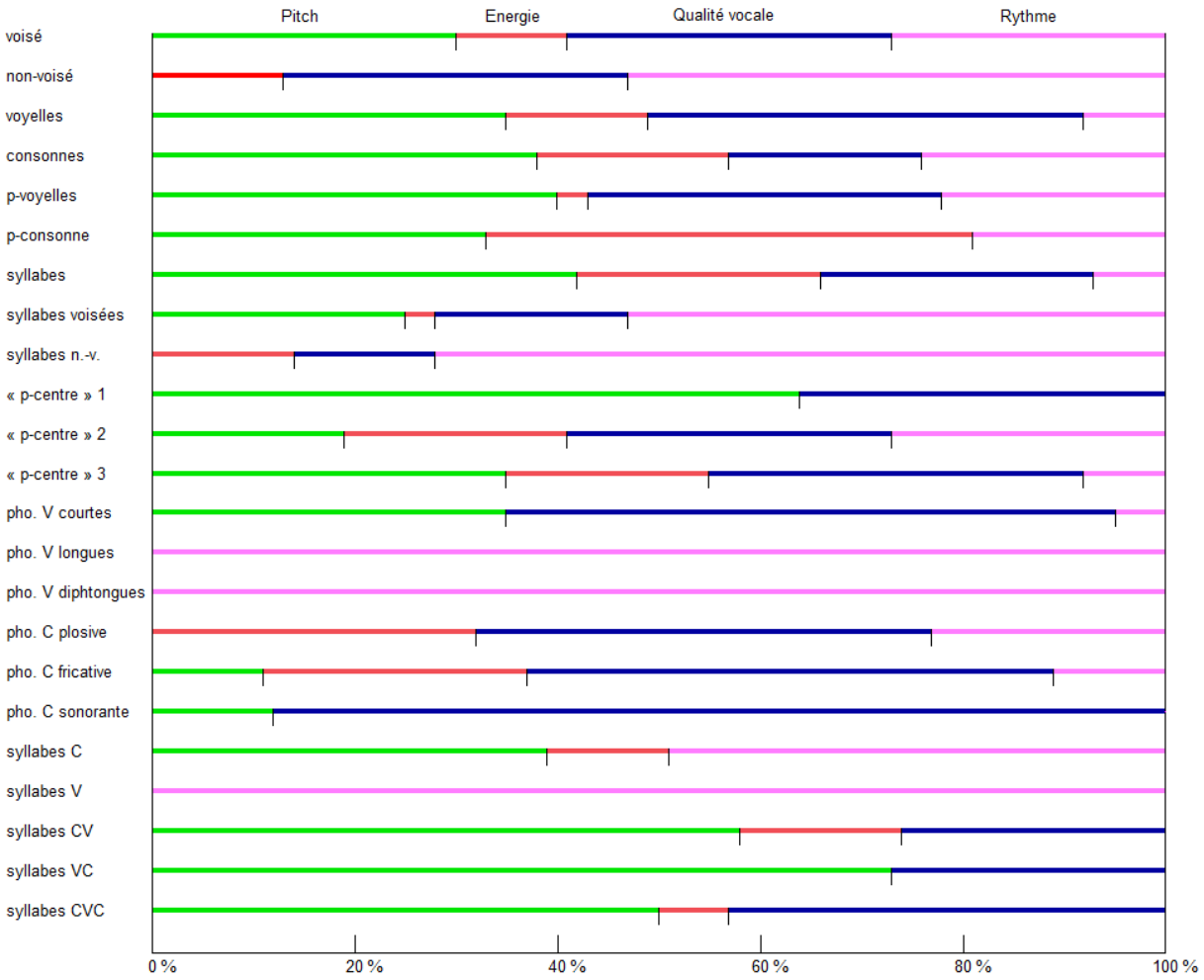


Fig. 4.12 Poids de fusion des composantes prosodiques selon les ancrages ; méthode CVS.

Table 4.8 Comparaison des scores en reconnaissance prosodique d'émotions sur le corpus Berlin selon les ancrages acoustiques complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	**81**	75	70	57	61	82	81
voyelle / consonne	66	66	64	50	52	87	77
p-voyelle / p-consonne	67	66	73	**65**	**69**	**89**	76
phonèmes voyelles	79	53	44	54	41	54	56
phonèmes consonnes	75	37	57	54	55	67	79
phonèmes	76	63	59	59	61	81	84
syllabes voisées / non-voisées	75	**82**	58	33	52	87	**89**
syllabes V/C/CV/VC/CVC	60	69	53	0	61	62	42
« p-centres » niveau 3	69	65	**79**	36	54	68	69

Approche composante ; méthode CVS.

contribution apportée par les différents points d'ancrages complémentaires de la parole sur les scores de reconnaissance est notable, ce qui n'était pas vraiment le cas pour l'analyse acoustique. Celle fournie par les ancrages voisés et vocaliques sur leur ancrage complémentaire (i.e., non-voisés et consonantiques) est par exemple équitable en moyenne. Une fusion de toutes les classes phonétiques montre que les ancrages consonantiques sont préférés à cette

échelle ; le score obtenu dépasse alors celui fourni par la fusion des macro-classes phonétiques (voyelle / consonne), ce qui n'est pas le cas des groupes de syllabe (V/CV/VC/CVC).

Les valeurs moyennes des scores cachent de fortes disparités selon les émotions et les ancrages acoustiques, cf. table 4.8. En effet, aucune émotion n'est uniformément reconnue par les différents segments de parole étudiés. Notons toutefois que les pseudo-phonèmes se démarquent favorablement des autres ancrages au niveau des scores sur la quasi-majorité des émotions (3/7) et que les « *p-centres* » conduisent au meilleur score pour *l'Ennui*.

4.1.2. Approche globale

Rappelons que cette approche exploite toutes les mesures prosodiques pour effectuer la reconnaissance des émotions et que le nombre de paramètres varie selon le type de décision, cf. table 4.5. Les résultats montrent que la décision « *phrase* » fournie encore une fois les meilleurs résultats (74%, ancrage voisé, normalisation *Z-locuteur*). Quelques différences apparaissent toutefois dans les scores selon les points d'ancrages. L'amélioration apportée par la normalisation des données selon le locuteur est très élevée puisqu'elle vaut en moyenne +31% et atteint même un maximum de +51% sur les ancrages « *p-centres* » identifiés au niveau 3.

Afin de démontrer la pertinence des caractéristiques du rythme pour la reconnaissance des émotions, nous avons analysé les paramètres qui ont été retenus par la méthode *bottom-up* pour chaque type d'ancrage. Toutefois, les tables ne sont pas présentées dans ce document puisqu'elles sont beaucoup trop nombreuses. Leur analyse montre que les paramètres fournis par les modèles *non-conventionnels* du rythme sont très souvent identifiés comme étant corrélés à l'affect, notamment pour les mesures issues de la distance de Hotteling (i.e., HD, PHD, A-PHD et I-PHD) qui apparaissent notamment comme particulièrement robustes puisqu'elles représentent environ la moitié de l'ensemble des paramètres retenus par la méthode *bottom-up*. Nous devons toutefois pondérer ce résultat, par le fait que les paramètres du rythme sont bien plus représentés en regard des autres composantes prosodiques. Notons aussi que la métrique du P-PVI apparaît de temps à autre (i.e., selon les points d'ancrage) dans la liste des paramètres discriminants. Par conséquent, les mesures de dynamique prosodique sont pertinentes pour quantifier les corrélats rythmiques des émotions actées du corpus Berlin dans le cadre d'une CVS, en particulier lorsque l'on utilise la métrique HD.

La table 4.9 présente les résultats obtenus par la fusion des points d'ancrages complémentaires de la parole. Des différences apparaissent par rapport aux résultats issus de l'approche reposant sur la fusion des composantes : excepté pour la fusion des macro-classes phonétiques dans laquelle les poids sont équivalents, la contribution des points d'ancrages complémentaires penche en faveur des segments vocaliques (e.g., ancrages voisés et p-voyelle). Les poids de fusion des informations phonétiques montrent ainsi une préférence notable pour les sonorantes et la fusion globale une contribution équivalente entre les macro-classes, i.e., voyelle et consonne. Notons que les valeurs moyennes des scores cachent de fortes disparités selon les émotions et les ancrages, puisqu'aucune émotion n'est uniformément reconnue par les différents points d'ancrage, cf. table 4.10. De plus, il existe une grande variabilité dans les résultats comparés à ceux fournis par l'approche en fusion des composantes (cf. table 4.8), puisque seule l'émotion *Joie* est la mieux reconnue par un même ancrage acoustique.

Table 4.9 Scores en reconnaissance prosodique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**61** 5/5	**66** 6/4	**71** 9/1	**74** 1/0	**72** 7/3
voyelle / consonne	55 1/9	**58** 9/1	**65** 5/5	**70** 5/5	**66** 6/4
p-voyelle / p-consonne	53 1/9	**61** 1/9	**72** 6/4	**72** 9/1	**69** 5/5
phonèmes voyelles	48 1/0/0	55 1/0/0	**68** 1/0/0	**68** 1/0/0	**62** 1/0/0
phonèmes consonnes	**51** 5/2/3	59 2/1/7	**67** 2/2/6	**66** 1/0/9	**62** 4/1/6
phonèmes	**56** 1/9	**64** 3/7	**69** 2/8	**67** 6/4	**65** 0/1
syllabes voisées / non-voisées	**64** 5/5	**68** 7/3	**75** 6/4	**72** 1/0	**73** 1/0
syllabes V/C/CV/VC/CVC	45 1/0/0/3/7	49 0/0/1/1/8	57 1/0/0/1/8	**60** 1/0/0/1/8	54 1/0/0/1/8

[valeur en %] $_{\text{poids}}$ α_a ; approche globale ; méthode CVS.

Table 4.10 Comparaison des scores en reconnaissance prosodique d'émotions sur le corpus Berlin selon les paires d'ancrages acoustiques complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	77	**84**	65	59	38	81	**90**
voyelle / consonne	70	75	53	48	59	**96**	81
p-voyelle / p-consonne	74	65	**70**	54	**61**	90	85
phonèmes voyelles	70	51	49	46	44	65	60
phonèmes consonnes	64	57	60	43	51	80	63
phonèmes	**83**	64	57	46	55	78	74
syllabes voisées / non-voisées	60	69	60	52	56	85	74
syllabes V/C/CV/VC/CVC	61	68	33	**63**	45	82	68
« p-centres » niveau 3	78	54	62	**63**	51	67	76

Approche globale ; méthode CVS.

4.2. Tests d'indépendance locuteur

Afin de faire ressortir les résultats que nous évaluons comme probants, nous avons fixé pour l'analyse un seuil à 50% sur les scores de reconnaissance issus des tests LOSO. Ce seuil est plus bas que dans la configuration en CVS puisque la difficulté est supérieure. Pour rappel, nous testons l'indépendance au locuteur des paramètres prosodiques corrélés à l'affect.

4.2.1. Approche par composante

Bien que la difficulté soit supérieure en LOSO, les scores en reconnaissance prosodique sont proches de ceux fournis par la CVS ; ce qui n'était pas le cas pour l'analyse acoustique. Un certain nombre de points communs peut d'autre part, être identifié entre ces deux méthodes d'exploration de données : (i) les différentes normalisations étudiées produisent une amélioration significative des scores, (ii) cette amélioration est tout spécialement portée par l'emploi des informations linguistiques pour les paramètres du rythme, (iii) les meilleurs scores sont quasiment tous issus de l'approche « *phrase* » et (iv) le pitch correspond à la composante prosodique la plus pertinente pour caractériser les corrélats de l'affect (63%, syllabe et syllabe voisée, normalisation *Z-locuteur*). Les scores obtenus par la qualité vocale en LOSO sont très proches de ceux du pitch (60%, p-voyelle, normalisation *Z-locuteur*) et supérieurs à ceux de la CVS. L'ordre d'importance des composantes corrélées à l'affect reste ainsi

Table 4.11 Scores en reconnaissance prosodique des émotions obtenus par la fusion des composantes prosodiques.

Ancrage acoustique	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé	**60** 2/3/1/4	**65** 2/1/5/2	**68** 3/1/5/0	**70** 5/2/3/0	**61** 4/0/6/0
non-voisé	51 0/8/1/2	54 0/5/4/2	53 0/4/3/3	**53** 0/5/3/2	**60** 0/3/2/5
voyelle	56 3/1/3/3	**61** 3/0/4/3	**69** 3/0/4/4	**64** 3/1/4/2	**66** 3/1/3/4
p-voyelle	**68** 4/1/2/2	**67** 2/1/5/2	**71** 5/0/3/2	**74** 3/1/4/3	**69** 3/0/3/5
voyelle courte	52 3/2/4/1	58 3/0/6/0	56 3/0/6/0	55 3/2/4/0	**63** 2/1/4/3
voyelle longue	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
voyelle diphtongue	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
consonne	**62** 2/1/1/7	**64** 4/1/2/3	**66** 7/1/0/2	**70** 8/0/1/1	69 3/0/1/6
p-consonne	59 4/3/1/3	**62** 3/3/2/2	**63** 6/2/2/0	**65** 7/1/0/1	**66** 2/4/2/3
consonne plosive	42 1/7/2/1	42 0/4/5/1	46 2/4/3/1	49 1/2/4/3	40 1/3/5/1
consonne fricative	36 1/9/0/1	36 3/2/0/6	44 1/3/1/5	46 1/2/1/6	44 1/5/1/3
consonne sonorante	56 3/2/6/0	59 3/1/6/1	58 6/0/4/0	**63** 3/2/5/0	**65** 3/1/3/3
syllabe	**67** 4/1/1/4	**68** 3/1/4/3	**71** 3/1/2/4	**73** 4/2/2/2	**68** 4/1/4/1
syllabe voisée	**64** 3/1/1/5	**66** 2/2/4/2	**69** 4/3/2/2	**71** 3/0/5/2	**66** 3/0/3/4
syllabe non-voisée	52 0/6/1/3	54 0/4/2/4	53 0/3/4/3	58 0/3/2/4	59 0/2/1/8
syllabe partie C	45 3/2/0/5	45 3/2/0/5	40 1/3/3/3	39 0/7/3/0	35 3/5/3/0
syllabe partie V	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1	24 0/0/0/1
syllabe partie CV	39 4/3/3/0	39 4/3/3/0	39 9/1/0/0	42 9/1/0/0	38 2/1/3/4
syllabe partie VC	38 3/2/2/4	38 2/2/1/5	43 5/1/2/2	47 3/1/1/5	42 2/4/2/3
syllabe partie CVC	57 4/3/2/2	55 2/3/4/1	57 5/1/3/0	57 8/1/1/0	62 2/3/1/4
« p-centres » 1	56 2/3/4/0	57 2/3/4/0	59 4/0/6/0	**61** 1/0/8/0	57 2/3/4/1
« p-centres » 2	53 2/4/4/0	**63** 3/2/3/2	**64** 4/2/4/0	**64** 3/1/5/0	**62** 2/2/4/2
« p-centres » 3	58 5/2/3/0	**60** 3/1/7/0	**63** 5/0/4/0	**66** 3/1/4/2	**62** 3/1/3/2

[valeur en %] $_{\text{poids}}$ α_p: pitch / énergie / qualité vocale / énergie ; méthode LOSO.

inchangé puisque vient ensuite le rythme (55%, ancrage consonne, normalisation *Z-phrase*) et l'énergie (50%, ancrages non-voisés, normalisation *Z-locuteur*).

Les résultats issus de la fusion des décisions « *segmentale* » et « *phrase* » montrent que les améliorations des scores sont encore une fois quasi-nulles pour les quatre composantes de la prosodie et que l'approche « *phrase* » reste dominante par rapport à celle « *segmentale* ». La fusion des composantes prosodiques permet aussi d'améliorer significativement les scores en comparaison de ceux fournis par le pitch : *Raw* : +25%, *Z-tout* : +26%, *Z-genre* : +16%, *Z-locuteur* : +17% et *Z-phrase* : +30%, cf. table 4.11. Le meilleur score est alors obtenu par l'ancrage des p-voyelles (74%, normalisation *Z-locuteur*) et les syllabes amènent un score très proche (73%, même normalisation) tout comme les consonnes (70%, …) et les segments voisés (70%, …). Les poids associés à la fusion des composantes prosodiques montrent une nouvelle fois l'importance de la normalisation des paramètres du rythme selon les phrases (i.e., *Z-phrase*), puisqu'ils sont bien souvent supérieurs aux autres configurations, cf. table 4.11. Ce résultat est aussi valable pour la CVS, cf. table 4.6.

La Fig. 4.13 illustre les poids obtenus par la fusion des composantes prosodiques et pour les normalisations produisant les meilleurs scores. Bien que certains types d'ancrages soient relativement proches sur le plan acoustique (e.g., segments voisés, voyelles, p-voyelle et syllabes voisées, ou segments non-voisés, consonnes, p-consonne et syllabes non-voisées), les configurations de fusion les plus appropriées varient, une nouvelle fois, fortement selon ces

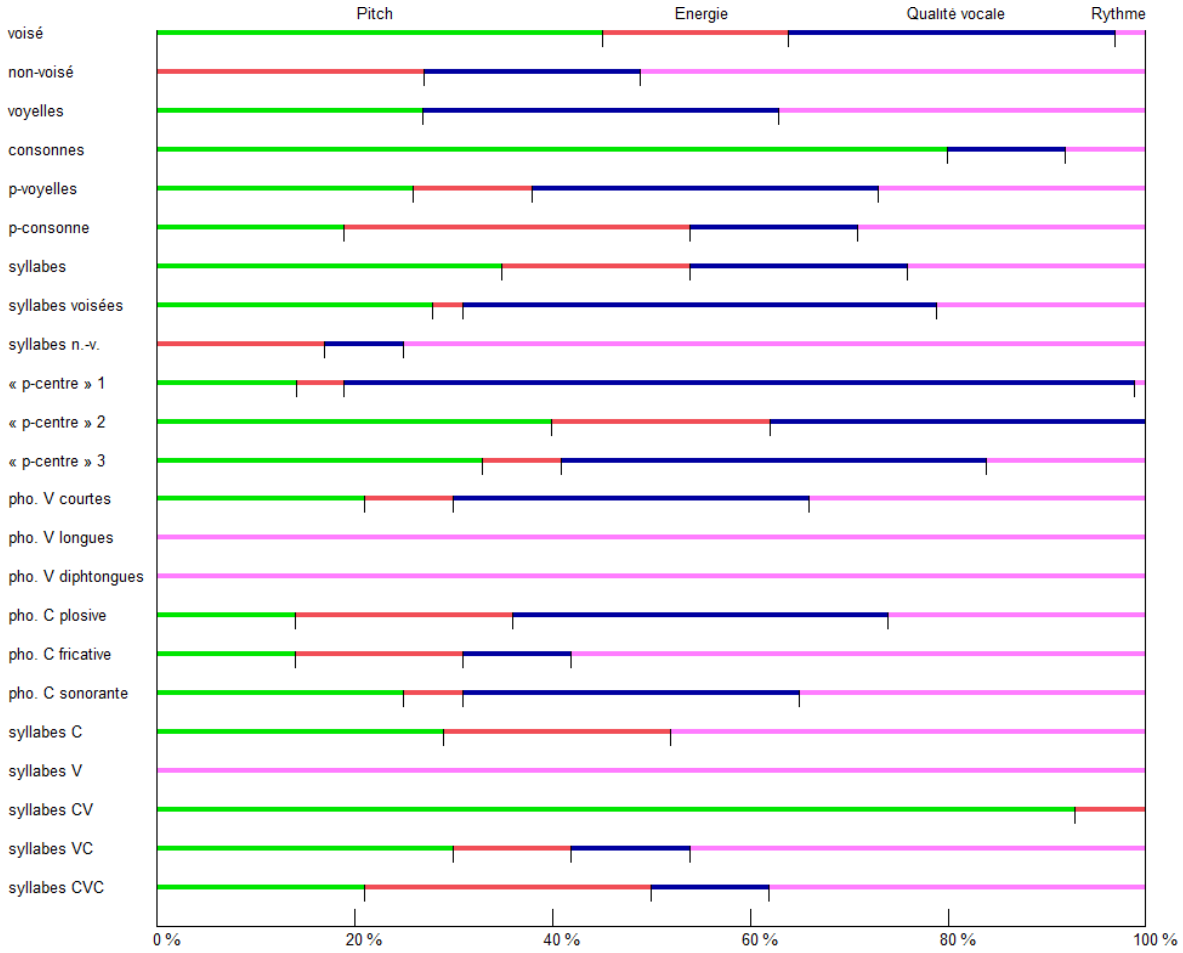


Fig. 4.13 Poids de fusion des composantes prosodiques selon les ancrages ; méthode LOSO.

Table 4.12 Scores en reconnaissance prosodique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**60** 7/3	**65** 7/3	**68** 1/0	**70** 1/0	**65** 5/5
voyelle / consonne	**64** 0/1	**66** 2/8	**72** 4/6	**72** 2/8	**73** 3/7
p-voyelle / p-consonne	**68** 8/2	**68** 7/3	**71** 1/0	**74** 1/0	**71** 7/3
phonèmes voyelles	51 1/0/0	51 1/0/0	54 1/0/0	53 1/0/0	55 1/0/0
phonèmes consonnes	54 1/0/9	55 1/0/9	59 0/4/6	**63** 0/1/9	**63** 2/1/7
phonèmes	57 5/5	**61** 3/7	**62** 6/4	**67** 4/6	**66** 7/3
syllabes voisées / non-voisées	**62** 7/3	**65** 8/2	**68** 1/0	**71** 9/1	**67** 5/5
syllabes V/C/CV/VC/CVC	55 1/0/0/0/9	55 1/0/0/0/9	55 0/0/1/0/9	56 0/0/2/1/7	59 1/0/0/0/9

[valeur en %] $_{\text{poids}}$ α_a ; approche composante ; méthode LOSO.

derniers. Des différences notables apparaissent toutefois entre les résultats issus des méthodes CVS et LOSO : l'importance du rythme est plus élevée dans la méthode LOSO alors que celle de la qualité vocale reste en moyenne inchangée. De plus, le pitch et l'énergie sont globalement moins contributifs comparés aux résultats issus de la CVS ; bien qu'ils conduisent également aux meilleurs scores.

Les résultats issus de la fusion des points d'ancrages acoustiques complémentaires sont donnés dans la table 4.12. Notons que les écarts entre les scores issus de la méthode CVS et

Table 4.13 Comparaison des scores en reconnaissance prosodique d'émotions sur le corpus Berlin selon les ancrages acoustiques complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	91	72	69	33	**37**	82	**77**
voyelle / consonne	94	63	**79**	**46**	32	84	74
p-voyelle / p-consonne	97	71	**79**	39	**37**	86	77
phonèmes voyelles	84	46	53	7	25	70	40
phonèmes consonnes	90	59	74	15	28	67	65
phonèmes	**95**	**74**	74	9	31	89	**82**
syllabes voisé / non-voisé	94	68	73	13	28	81	66
syllabes V/C/CV/VC/CVC	90	35	69	11	20	68	53
« p-centres » niveau 3	91	70	74	9	13	**92**	73

Approche composante ; méthode LOSO.

LOSO sont beaucoup moins importants dans l'analyse prosodique (-5% en moyenne) que dans l'analyse acoustique (-10% en moyenne), et que les contributions des points d'ancrages acoustiques complémentaires sont équivalentes à celles obtenues par l'approche globale en CVS. Les valeurs moyennes des scores font également apparaître de fortes disparités selon les émotions et les points d'ancrage de la parole, cf. table 4.13. Les ancrages « phonèmes » produisent alors les meilleurs scores sur le plus grand nombre d'émotions (3/7). Notons encore une fois, l'extrême variabilité des résultats comparé aux analyses précédentes (cf. tables 4.8 et 4.10) et le fait que les émotions *Dégoût* et *Joie* produisent des scores en LOSO qui sont relativement bas comparés à ceux obtenus en CVS.

4.2.2. Approche globale

Rappelons que cette approche exploite toutes les données des composantes prosodiques et que le nombre de paramètres disponibles varie selon le type de décision, cf. table 4.5. Les résultats obtenus par les décisions « *phrase* » et « *segmentale* » montrent que la première fournit à nouveau les meilleurs résultats (77%, ancrage p-voyelle, normalisation *Z-locuteur*) et que l'amélioration apportée par la normalisation des données selon le locuteur est importante et comparable à celle obtenue en CVS (moyenne = +29%, maximum = +48% pour l'ancrage des p-voyelles). Bien que les valeurs de scores soient proches entre les décisions « *segmentale* » et « *phrase* », leur contribution dans la fusion est quasi-nulle. Notons que le meilleur score est produit par les pseudo-voyelles (77%) et qu'il est supérieur à celui obtenu en CVS alors que la difficulté est censée être moindre avec cette méthode. Par ailleurs, la fusion des ancrages acoustiques ne permet d'améliorer que de très peu, les scores puisqu'ils sont déjà très élevés, cf. table 4.14.

Afin de démontrer la pertinence des caractéristiques du rythme pour la reconnaissance des émotions, nous avons étudié les paramètres qui ont été retenus par l'approche *bottom-up* pour chaque type d'ancrage. Les résultats sont alors très satisfaisants puisque les métriques *non-conventionnelles* du rythme ressortent une nouvelle fois comme pertinentes pour caractériser les corrélats prosodiques de l'affect, en particulier les mesures fournies par la distance de Hotteling, i.e., HD, PHD, A-PHD et I-PHD puisqu'elles représentent environ la moitié de l'ensemble des paramètres identifiés par l'algorithme. Puisque la métrique du P-PVI fait éga-

Table 4.14 Scores en reconnaissance prosodique des émotions sur le corpus Berlin obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	57 5/5	**70** 5/5	**71** 1/0	**73** 5/5	**70** 1/0
voyelle / consonne	**61** 1/9	**67** 6/4	**72** 1/9	**77** 6/4	**70** 4/6
p-voyelle / p-consonne	53 1/9	**71** 1/0	**73** 8/2	**77** 1/0	**72** 8/2
phonèmes voyelles	48 1/0/0	**63** 1/0/0	**69** 1/0/0	**73** 1/0/0	**62** 1/0/0
phonèmes consonnes	51 5/2/3	**60** 0/1/9	**59** 0/1/9	**64** 0/0/1	**61** 0/1/9
Phonèmes	56 1/9	**63** 3/7	**68** 7/3	**71** 6/4	**64** 0/1
syllabes voisé / non-voisé	**64** 5/5	**70** 1/0	**72** 1/0	**73** 9/1	**74** 1/0
syllabes V/C/CV/VC/CVC	48 2/0/0/0/8	54 0/0/5/0/5	58 1/0/0/1/8	59 0/0/1/1/8	56 1/0/0/0/9

[valeur en %] $_{\text{poids}}$ α_a ; approche globale ; méthode LOSO.

Table 4.15 Comparaison des scores en reconnaissance prosodique d'émotions sur le corpus Berlin selon les paires d'ancrages complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	91	72	73	33	21	**94**	**87**
voyelle / consonne	90	**76**	81	**57**	38	**94**	81
p-voyelle / p-consonne	90	69	**85**	50	**52**	89	82
phonèmes voyelles	92	49	67	20	24	72	60
phonèmes consonnes	89	47	68	46	23	63	66
phonèmes	**94**	72	78	41	32	84	79
syllabes voisé / non-voisé	90	57	79	41	30	87	68
syllabes V/C/CV/VC/CVC	67	43	60	39	48	77	60
« p-centres » niveau 3	93	49	77	22	28	78	65

Approche globale ; méthode LOSO.

lement partie des paramètres discriminants des émotions, les mesures de dynamique prosodique s'avèrent donc appropriées pour quantifier les corrélats rythmiques des émotions actées, et ce, quel que soit le type d'ancrage ou la méthode d'exploration de données choisie. Ces résultats valident donc notre hypothèse quant à l'intérêt d'exploiter les composantes de la prosodie pour caractériser les corrélats rythmiques des émotions (actées). De façon générale, tous les paramètres du rythme que nous avons utilisé, se sont révélés à un moment ou un autre, adéquat dans la reconnaissance des émotions. Notons toutefois que ces derniers sont majoritairement représentés en regard des autres composantes prosodiques, cf. table 4.5.

L'analyse détaillée des scores montre que la *Colère* est très bien reconnue par tous les points d'ancrage étudiés et que les fusions des paires complémentaires conduisent aux meilleurs résultats sur la quasi-totalité des émotions (6/7), cf. table 4.15. De plus, les émotions *Dégoût* et *Joie* amènent des scores plus élevés que ceux fournis par la fusion des composantes prosodiques, cf. table 4.13.

4.3. Fusion acoustique / prosodique

Les résultats issus de la fusion des informations acoustiques et prosodiques sont présentés dans cette section. Les vecteurs de probabilités retournés par l'étage de fusion des décisions prises sur les MFCC (cf. chapitre 3, Fig. 3.6) ont été fusionnés avec ceux issus de l'étage de

fusion des composantes prosodiques (ou des décisions pour l'approche globale), cf. Fig. 4.11. Les résultats sont présentés à travers les deux méthodes d'exploration de données : CVS et LOSO. Afin de faire ressortir ceux que nous évaluons comme probants, nous fixons pour l'analyse un seuil à 70% sur les scores de reconnaissance. Ce seuil est le plus élevé de tous puisque les données exploitées sont déjà issues de plusieurs phases de fusion.

4.3.1. Tests en cross-validation croisée et stratifiée

Bien que les meilleurs scores obtenus par la fusion des composantes prosodiques soient manifestement supérieurs à ceux fournis par les MFCC (74% contre 68%), la contribution de ces deux types d'informations dans la tâche de reconnaissance est globalement équivalente et penche même légèrement en faveur des coefficients MFCC. Les meilleurs scores sont quasi identiques entre les deux approches : « composante » et « globale ». Les résultats issus de la fusion des points d'ancrages acoustiques complémentaires selon ces deux approches sont donnés en annexe dans les tables A.4.1 et 4.2, et les scores de chaque émotion dans les tables additionnelles A.4.1.2 et A.4.2.2. La contribution des ancrages complémentaires montre une forte dominance des segments voisés sur les non-voisés alors que celle des classes et macro-classes phonétiques sont plus équitables. Notons que le meilleur score de chaque émotion varie selon le point d'ancrage considéré, cf. tables A.4.1.2 et A.4.2.2.

4.3.2. Tests d'indépendance locuteur

Contrairement à la méthode de CVS, les contributions des composantes acoustiques et prosodiques en LOSO penchent très nettement en faveur de ces dernières, voire même de façon quasi-exclusive lorsque l'on restreint l'analyse aux meilleurs scores. La différence entre les scores obtenus sur ces deux types de données est trop importante pour améliorer les performances en reconnaissance ; les meilleurs scores des MFCC plafonnent à 59% contre 77% pour la prosodie en LOSO. Les résultats issus de la fusion des ancrages acoustiques complémentaires selon les approches *composante* et *globale* sont donnés (en annexe) dans les tables A.4.3 et A.4.4. L'approche globale fournit des scores qui sont supérieurs à la fusion des composantes prosodiques. Les poids de fusion des ancrages acoustiques montrent que la contribution des segments voisés et des pseudo-voyelles sur leur ancrage complémentaire est très dominante, alors que celle des classes phonétiques et syllabiques est encore une fois plus contrastée. Notons que les meilleurs scores varient selon les émotions et sont portés par plusieurs types d'ancrages acoustiques de la parole, cf. tables A.4.3.2 et A.4.4.2.

4.4. Analyse des paramètres du rythme

Cette section détaille les paramètres du rythme selon les catégories d'émotions contenues dans le corpus de parole actée Berlin. Nous avons calculé les deux premiers moments statistiques des paramètres issus des modèles *conventionnels* et *non-conventionnels* du rythme à travers les classes d'émotions. Aucune normalisation n'a été effectuée sur les valeurs qui sont données dans la table 4.16 de façon à présenter tel quel le comportement des métriques.

Table 4.16 Analyse statistique des modèles *conventionnels* et *non-conventionnels* du rythme selon les classes d'émotions prototypiques.

Modèles du rythme	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
%V	30 $_{4.4}$	35 $_{5.7}$	35 $_{5.5}$	31 $_{4.6}$	32 $_{4.4}$	36 $_{4.9}$	32 $_{5.3}$
ΔC (ms)	48 $_{10}$	59 $_{21}$	74 $_{22}$	64 $_{21}$	54 $_{14}$	60 $_{19}$	101 $_{57}$
Varco	37 $_{9.8}$	41 $_{15}$	40 $_{13}$	51 $_{16}$	40 $_{10}$	37 $_{8.5}$	79 $_{33}$
$\overline{R}$ ($.10^{-2}$)	32 $_{15}$	34 $_{17}$	34 $_{14}$	33 $_{13}$	35 $_{15}$	31 $_{18}$	31 $_{17}$
RR ($.10^{-1}$)	12 $_{7.5}$	12 $_{9.0}$	12 $_{8.2}$	13 $_{12}$	12 $_{7.4}$	12 $_{7.2}$	16 $_{21}$
***r*PVI** ($.10^{-3}$)	74 $_{66}$	75 $_{82}$	89 $_{84}$	122 $_{127}$	78 $_{68}$	73 $_{56}$	214 $_{323}$
***n*PVI** ($.10^{-3}$)	42 $_{31}$	45 $_{33}$	46 $_{33}$	54 $_{39}$	44 $_{31}$	44 $_{31}$	61 $_{45}$
F_{moy} **« *p-centre* »** (Hertz $.10^{-1}$)	78 $_{13}$	88 $_{13}$	71 $_{13}$	73 $_{12}$	82 $_{12}$	89 $_{12}$	65 $_{14}$
***MNF* de la THH** (Hertz.10^{-1})	79 $_{14}$	81 $_{16}$	81 $_{16}$	78 $_{15}$	80 $_{13}$	83 $_{15}$	76 $_{15}$
PHD (pitch) ($.10^{-4}$)	55 $_{47}$	42 $_{45}$	75 $_{75}$	70 $_{73}$	57 $_{52}$	53 $_{50}$	69 $_{79}$
PHD (énergie) ($.10^{-4}$)	38 $_{42}$	27 $_{29}$	46 $_{57}$	46 $_{52}$	33 $_{36}$	29 $_{35}$	45 $_{55}$
PHD (qualité vocale) ($.10^{-4}$)	55 $_{25}$	49 $_{24}$	76 $_{42}$	66 $_{35}$	55 $_{26}$	56 $_{32}$	80 $_{49}$
A-PHD (toutes) ($.10^{-3}$)	15 $_{9.2}$	12 $_{8.6}$	20 $_{15}$	19 $_{13}$	15 $_{9.3}$	14 $_{10}$	20 $_{16}$
I-PHD (toutes) ($.10^{-3}$)	63 $_{156}$	41 $_{39}$	79 $_{85}$	67 $_{89}$	54 $_{73}$	55 $_{60}$	63 $_{57}$

[Valeur moyenne] $_{\text{écart-type}}$; ancrage des pseudo-voyelles sauf pour la mesure ΔC qui est calculée sur les pseudo-consonnes.

Notons que les deux méthodes d'estimation de la fréquence moyenne (i.e., F_{moy} issue du « *p-centre* » – traitement spectrale suite à un filtrage statique des données – et *MNF* issue de la THH des SUI formés par les pseudo-voyelles – traitement statistique des données via un filtrage adaptatif) amènent des valeurs relativement proches entres elles, même si la THH entraîne une plus grande dispersion des valeurs. Nous ne donnons pas les mesures fournies par les P-PVI car l'écart-type de ces données est bien souvent supérieur à la valeur moyenne, tout comme cela est également le cas pour les *r*PVI et *n*PVI. Notons que la *Tristesse* produit bien souvent des maximums dans les mesures statistiques calculées sur les modèles *conventionnels* du rythme, et de façon beaucoup moins fréquente pour les modèles *non-conventionnels*, cf. table 4.16. Les modèles *conventionnels* amènent des résultats assez intéressants avec les mesures %V et ΔC, cf. (a) Fig. 4.14. La plupart des émotions sont alors plutôt bien séparées dans le plan formé par ces deux paramètres. Les modèles *non-conventionnels* que nous avons proposés donnent quant à eux, des résultats beaucoup plus intéressants, puisqu'ils permettent non seulement de séparer de façon plus manifeste les émotions dans le plan formé par certaines métriques, cf. (b) et (c) Fig. 4.14, mais aussi de définir un continuum de valeurs entre les catégories d'émotions qui apparaissent dans l'ordre défini par la roue de Plutchik [PLU80][70], cf. (d) Fig. 4.14. Les émotions *Ennui* et *Dégoût* qui correspondent à des zones connexes dans la roue sont alors très proches dans l'espace des caractéristiques rythmiques, (b) notamment (c), ce qui n'est pas le cas pour les paramètres *conventionnels*, cf. (a) Fig. 4.14. De plus les catégories d'émotions successives dans la roue sont séparées par des distances relativement constantes dans le plan formé par les paramètres *non-conventionnels* du rythme. Les corrélations entre ces paramètres et les catégories d'émotions sont donc immenses, puisque les mesures permettent de définir un continuum relativement régulier entre les catégories d'émotions qui apparaissent dans l'ordre de la roue de Plutchik.

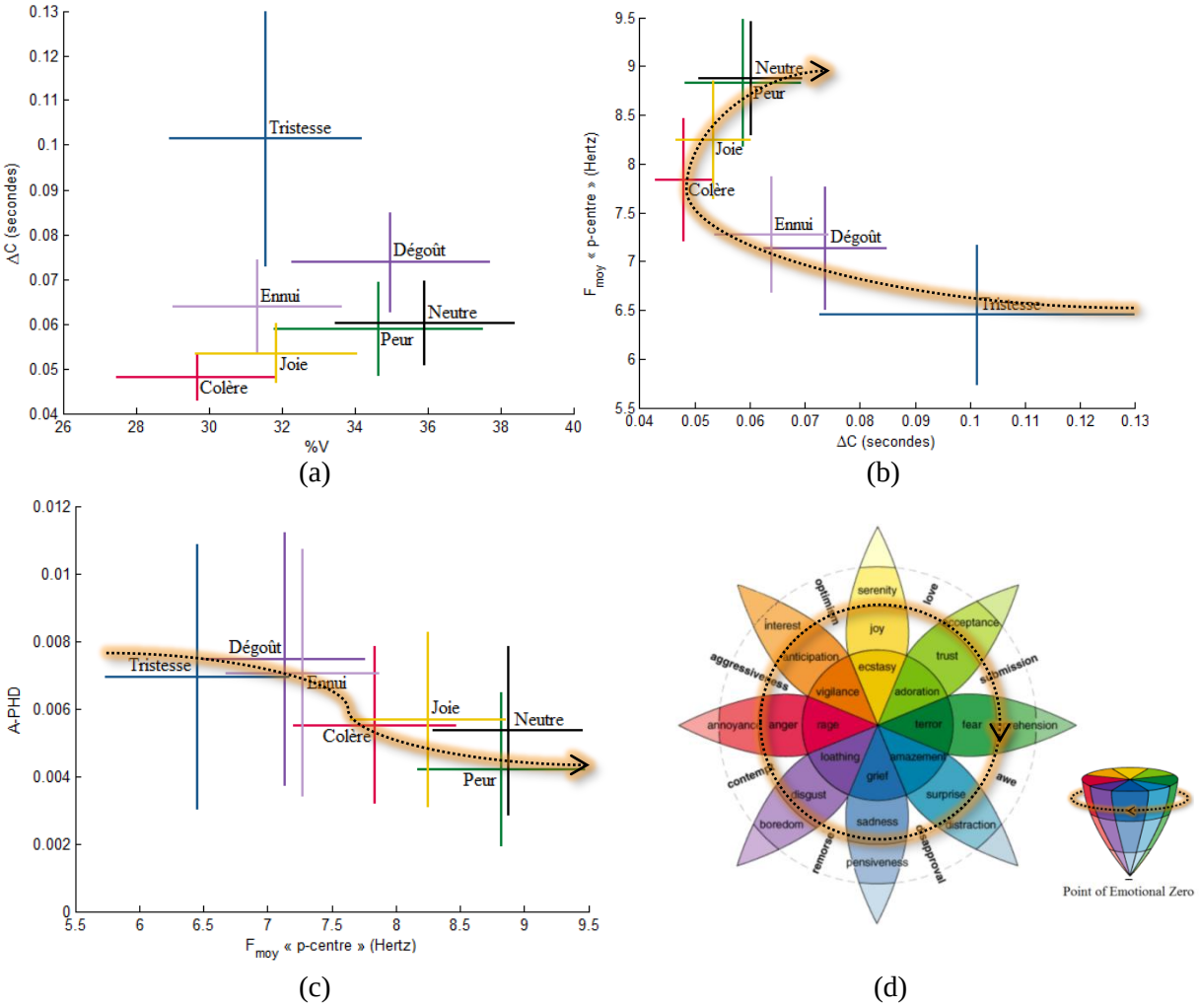


Fig. 4.14 Variations des mesures issues des modèles du rythme *conventionnels* (a), *mixtes* (b) *et non-conventionnels* (c) selon les catégories d'émotions. La position de la croix dans l'espace des paramètres en détermine les valeurs moyennes, tandis que la hauteur et la largeur correspondent aux valeurs d'écart-type ; (d) roue des émotions de Plutchik [PLU80][69].

5. Conclusion

Nous avons présenté différentes théories du rythme dans l'introduction de ce chapitre. Cette première partie a montré que le rythme véhicule des phénomènes complexes dont leur caractérisation ne peut reposer sur des mesures simples telles que le débit, puisque ce dernier en est tout simplement qu'une composante. Comme les phénomènes du rythme peuvent être à l'origine des émotions procurées par la musique, nous avons proposé (comme d'autres auteurs l'ont fait auparavant) de faire le lien entre les propriétés de la musique et de celles de la parole. En effet, le rythme apparaît clairement comme sous-modélisé dans les systèmes issus de l'état de l'art en reconnaissance d'émotions. Nous avons donc développé des métriques *non-conventionnelles* pour capturer les phénomènes du rythme de la parole. Différentes techniques ont alors été exploitées : (i) les *mesures spectrales* sur l'enveloppe estimée par la méthode de Tilsen, (ii) *l'enveloppe* et la *fréquence instantanées* calculées au moyen de la THH, (iii) les

[69] R. Plutchik, Emotion: *A Psychoevolutionary Synthesis*, dans Harper & Row, New York, 1980.

indices de *variabilité prosodique par paire* (extension du PVI de Grabe et Low aux composantes prosodiques) et (iv) les *distances prosodiques inter-segmentales* (HD, A-PHD et I-PHD). Les expériences de ce chapitre consistent à étudier le système de reconnaissance d'émotions selon différents types de caractéristiques prosodiques et d'ancrages acoustiques de la parole.

Les scores obtenus en reconnaissance prosodique sont nettement supérieurs à ceux fournis par les données acoustiques (MFCC), cf. chapitre 3. La fusion des composantes acoustique et prosodique montre une légère domination des MFCC en CVS alors que celle de la prosodie est beaucoup plus importante en LOSO (indépendance locuteur). Les deux approches utilisées pour fusionner les composantes prosodiques, i.e., par composante ou globale (tous les paramètres), amènent des résultats assez similaires. Le pitch apparaît alors comme la composante la plus pertinente, vient ensuite la qualité vocale, le rythme et l'énergie. Des différences apparaissent toutefois selon le support d'extraction des paramètres, i.e., les ancrages acoustiques. La fusion de ces derniers indique que les ancrages voisés sont nettement privilégiés aux non-voisés, contrairement aux classes et macro-classes phonétiques dont la contribution est très variable selon les configurations d'analyse CVS et LOSO. Une analyse détaillée des résultats montre que le meilleur score de chaque émotion est très souvent obtenu par différents types d'ancrage acoustique et que ces derniers varient aussi selon les approches. Notons que les ancrages pseudo-phonétiques amènent de très bons résultats puisque les émotions ont été reconnues à plus de 75% en LOSO. Cependant, ce type de tests entraîne une dégradation des performances sur les émotions *Joie* et *Dégoût* ; les meilleurs scores sont alors inférieurs à 60%. Ce qui n'est pas le cas de la *Colère* qui est reconnue à plus de 90% par tous les ancrages acoustiques de la parole.

Les expériences de ce chapitre ont donc montré que les ancrages acoustiques de la parole peuvent conduire à des résultats très différents dans la tâche d'identification des corrélats prosodiques des émotions. Notre étude a également démontré la pertinence de l'emploi de métriques *non-conventionnelles* du rythme pour caractériser les corrélats affectifs présents dans la parole. Les expériences ont, par exemple, montré que la composante rythmique joue un rôle important dans la communication des émotions, puisqu'une analyse de la liste des paramètres discriminants (i.e., identifiés par l'approche *bottom-up*) fait apparaître très fréquemment les métriques *non-conventionnels* du rythme selon : (i) les tests d'indépendance aux classes et au locuteur et (ii) les points d'ancrages acoustiques de la parole. De plus, certains paramètres permettent de définir un continuum de valeurs relativement uniforme entre les catégories d'émotions qui apparaissent dans l'ordre de la roue de Plutchik, cf. Fig. 4.14. La pertinence des modèles *non-conventionnels* du rythme pour caractériser les composantes affectives de la parole est ainsi démontrée de façon directe (variations statistiques) et indirecte (variations dans les poids de fusion en reconnaissance d'émotions) dans ce chapitre.

Les émotions *prototypiques* qui sont contenues dans le corpus Berlin peuvent donc être caractérisées de façon robuste par les techniques actuelles du TAP. Par conséquent, la difficulté peut être augmentée pour étudier des corpus contenant des émotions plus larges telles que celles produites de façon spontanée. Cette analyse est nécessaire pour confirmer la validité des techniques utilisées dans les systèmes de reconnaissance d'émotions prototypiques. Le chapitre qui suit propose d'effectuer cette analyse en parallèle avec celle des troubles de la

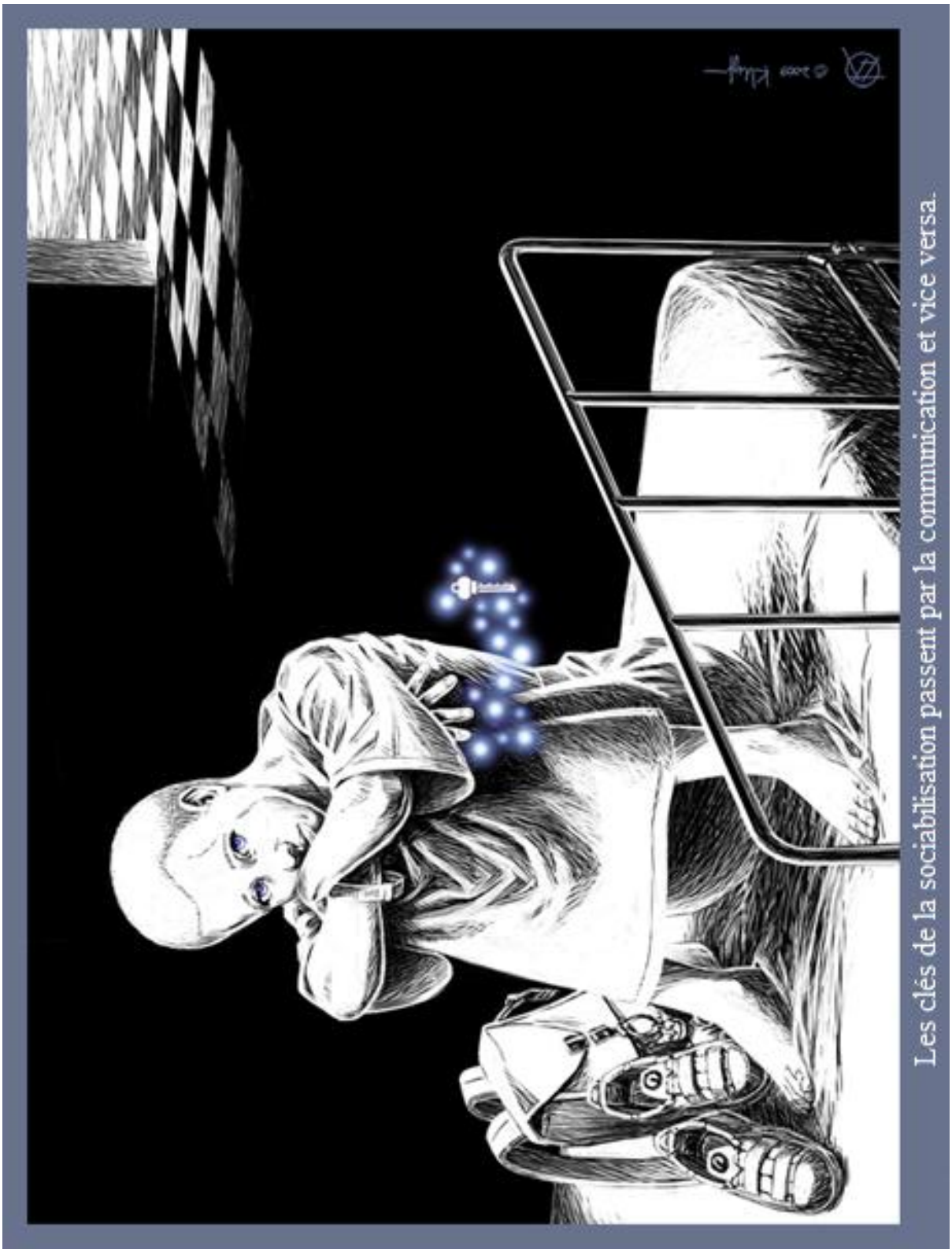
communication chez divers groupes d'enfants; (i) développement typique (groupe contrôle) ; (ii) autisme ; (iii) dysharmonie ; et (iv) dysphasie.

Chapitre 5

Emotions et troubles de la communication

Des études en psychologie ont montré que les émotions occupent une place fondamentale dans la communication humaine. L'échange des informations par le langage nécessite cependant de connaître un ensemble de règles qui ont été préétablies. Ces codes permettent de relier les manifestations acoustiques de la parole et les significations qui y sont associées. L'acquisition et l'usage correct de tels codes dans la parole jouent un rôle essentiel dans le développement de l'intersubjectivité et des capacités d'interactions sociales des enfants. Cette étape cruciale du développement, est supposée être fonctionnelle lors des premières étapes de la vie. Or le développement de l'enfant peut parfois s'écarter du cadre typique. Les troubles de la communication (TC) affectent ainsi un large ensemble d'individus qui présentent des difficultés à créer des liens avec autrui. Par exemple, le manque d'appétence à la communication et à l'interaction des sujets atteints de *troubles envahissants du développement* (TED, e.g., autisme) laisse bien souvent ces enfants repliées sur eux-mêmes, sans possibilités de comprendre les codes sociaux qui sont pourtant omniprésents. Identifier clairement et indubitablement, les particularités de la prosodie des sujets atteints de TC constitue donc un enjeu déterminant. Leur connaissance permettrait d'améliorer la prise en charge orthophonique des sujets, puisque les paramètres sur lesquels agir auront été dès lors mieux identifiés. Les diagnostics pourraient aussi être améliorés.

Ce chapitre présente un protocole qui a été utilisé pour caractériser les troubles dans la prosodie chez des sujets atteints de divers TC. Nous avons pour cela collaboré avec plusieurs groupes de cliniciens. Les expériences que nous avons réalisées reposent sur deux types d'épreuves : la première est plutôt basique et permet de vérifier si les sujets ont la capacité d'exploiter les fonctionnalités *grammaticales* de la prosodie (i.e., imitation du contour intonatif selon la modalité de la phrase), alors que la seconde épreuve est plus complexe et porte sur les fonctionnalités *affectives* de la prosodie (production de parole affective spontanée via le récit d'une histoire imagée). L'analyse des données issues de ces épreuves a été effectuée au moyen des traitements automatiques qui ont été proposés pour l'étude des émotions prototypiques. Les résultats montrent qu'il est tout à fait possible de retrouver, de façon automatique et significative, les particularités prosodiques des TC.


Les clés de la sociabilisation passent par la communication et vice versa.

1. Introduction

La parole est une forme d'onde complexe qui véhicule un grand nombre d'informations utiles pour la communication entre personnes et les interactions de type Homme-machine. Pour s'exprimer, un locuteur ne produit pas seulement un message composé d'informations textuelles, mais il ou elle transmet également un large éventail d'informations qui modulent et rehaussent le sens du message produit [ANA09][1]. Ce complément d'information est porté par la prosodie, cf. chapitre 1, section 2. Afin de communiquer proprement par le langage, la connaissance des règles qui ont été préétablies est nécessaire, puisqu'elles relient les manifestations acoustiques de la parole à un ensemble varié d'informations, e.g., intention, attitude, sentiment, émotion, etc. L'acquisition et l'utilisation correcte de ces codes jouent un rôle essentiel dans le développement de l'intersubjectivité et des capacités d'interactions sociales des enfants. Cette étape est supposée être effective dans les premières étapes de la vie d'un enfant dans le cas d'un développement typique, cf. Fig. d'introduction du chapitre 2 [KUH04][2]. Cependant, le développement de l'enfant peut parfois s'écarter du cadre typique.

Nous présentons dans les paragraphes suivants plusieurs troubles de la communication (TC) que nous avons étudiés : (i) les troubles autistiques (TA), (ii) les troubles envahissants du développement non-spécifiés (TED-NOS) et (iii) les troubles spécifiques du langage (TSL). Nous détaillons aussi la prévalence des différents types de TC chez les populations infantiles et nous présentons ensuite un bref état-de-l'art des études portant sur les fonctionnalités prosodiques *grammaticales* et *affectives* de ces sujets. Enfin, les principales méthodes servant à caractériser des troubles dans la prosodie sont présentées dans la section suivante et nous clôturons cette section par les objectifs de notre étude.

1.1. Troubles envahissant du développement et dysphasie

Leo Kanner présente en 1943 des travaux portants sur un groupe composé de 11 enfants qu'il décrit alors comme étant « venus au monde avec l'incapacité biologique innée de développer le contact affectif usuel avec autrui, tout comme d'autres viennent au monde avec d'autres handicaps physiques ou intellectuels innés » [KAN43][3]. L'année suivante, Hans Asperger [ASP44][4] étudie un groupe d'enfants qui présentent des troubles similaires à ceux décrits par Kanner (syndrome d'Asperger – SA). En 1979, Wing et Gould conduisent, dans un but épidémiologique et classificatoire, une étude plus poussée sur les troubles sévères dans les interactions sociales et les anomalies associées chez l'enfant [WIN79][5]. Un ensemble de troubles, qualifiée de *triade autistique*, a ainsi été identifié : (i) troubles des interactions sociales réciproques, (ii) comportements répétitifs et stéréotypés et (iii) troubles de la communication (TC) et de l'imagination. Cette triade constitue, aujourd'hui[6], la base fondamentale des critères diagnostiques cliniques des TED [DSM94][7].

---

[1] S. Ananthakrishnan et S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition", dans *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 17, no. 1, pp. 138–149, Jan. 2009.

[2] P. K. Kuhl, "Early language acquisition: cracking the speech code," dans *Nature Reviews Neuroscience*, vol. 5, pp. 831–843, Nov. 2004.

[3] L. Kanner, "Autistic disturbances of affective contact", dans *Nervous child*, vol. 2, pp. 217–250, 1943.

1.1.1. Le trouble autistique (TA)

Le trouble autistique est un trouble envahissant du développement (TED) qui inclut ceux de la triade identifiée par Wing *et al.* [WIN79][5]. Le diagnostic de l'autisme s'effectue selon les critères du DSM IV[7]. Dans la plupart des cas, une analyse des aptitudes cognitives du sujet est associée. Celles-ci montrent que : (i) le retard mental du TA varie de léger à profond, (ii) le profil est habituellement hétérogène quel que soit le niveau général d'intelligence et (iii) les capacités verbales sont typiquement plus faibles que les capacités non-verbales.

1.1.2. Les troubles envahissants du développement non spécifiés (TED-NOS)

Selon le DSM IV, le diagnostic de cette catégorie inclut des tableaux cliniques qui diffèrent de ceux du TA par : (i) un âge de début plus tardif, (ii) une symptomatologie atypique ou sous le seuil, ou (iii) l'ensemble de ces caractéristiques. Contrairement aux sujets TA et TSL, la catégorie des TED-NOS rassemble un groupe hétérogène d'enfants présentant des profils cliniques variés et pour lesquels il n'existe pas vraiment de critères diagnostics bien définis [FOM03][10]. Les cliniciens considèrent qu'il s'agit, bien souvent, d'un diagnostic par défaut en l'absence de l'ensemble des symptômes décrits pour le TA [VOL05][8]. Notons toutefois que d'autres termes ont été proposés entre temps, pour limiter l'hétérogénéité des troubles rencontrés chez les TED-NOS, e.g., dysharmonie, ou troubles développementaux complexes et multiples [VOL05].

1.1.3. La dysphasie ou les troubles spécifiques du langage (TSL)

La dysphasie se définit par l'existence d'un déficit durable des performances verbales, et significatif en regard des normes établies pour l'âge. Cependant, cette condition ne peut pas être liée à : (i) un déficit auditif, (ii) une malformation des organes phonatoires, (iii) une insuffisance intellectuelle, (iv) une lésion cérébrale acquise au cours de l'enfance, (v) un TED ou (vi) une carence grave affective ou éducative [GER93][9]. Ainsi, la dysphasie correspond à une altération qui intervient uniquement dans le développement des fonctions langagières. Cette dernière implique notamment un échec dans l'acquisition typique du langage expressif et/ou réceptif.

---

4 H. Asperger, "Autistic psychopathy in childhood", (Translation and annotation by U. Frith of the original paper), dans *Autism and Asperger Syndrome*, U. Frith [Eds], Cambridge, England: Cambridge University Press (1991, pp. 37–92), 1944.

5 L. Wing et J. Gould, "Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification", dans *J. of Autism and Developmental Disorders*, vol. 9, no. 1, pp. 11–29, 1979.

6 Notons que la communauté clinique organise régulièrement des colloques et autres conférences qui permettent de contribuer à une amélioration continuelle des définitions et des descriptions apportées sur les TC. Nous avons ainsi présenté les premiers résultats issus de nos travaux sur les émotions lors des ateliers de réflexion prospective *PIRSTEC*, no. 8, intitulés "*Autisme et Prosodie, quelles implications possibles ?*", Oct. 2009.

7 American Psychiatric Association, *Diagnostic and Statistical Manual of mental disorders*. 4th Ed., Washington, DC, 1994.

8 F. Volkmar, "Handbook of Autism and Pervasive Developmental Disorder", dans John Wiley and Sons [Eds], New jersey: Hoboken, 2005.

9 C. L. Gerard, *L'enfant dysphasique*, 1ère Ed. : Paris, 1993.

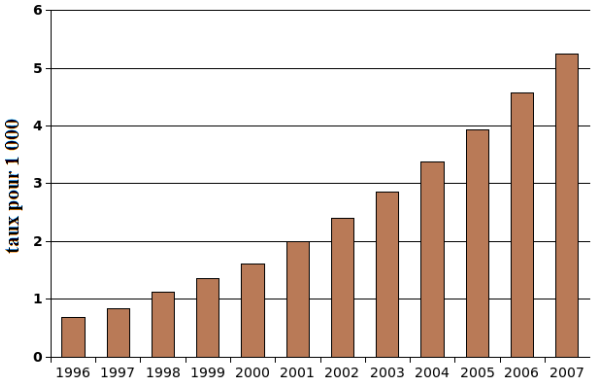


Fig. 5.1 Evolution des statistiques du gouvernement américain sur la prévalence de l'autisme dans la population infantile entre 1996 et 2007.

1.1.4. Prévalence des sujets TED et TSL

Les données de prévalence des enfants atteints de TED semblent avoir été sous-estimées durant ces dernières décennies. Les études réalisées dans la population française ont montré que le taux de prévalence des TED était situé autour des 1.6 pour 1 000 en 1992 [FOM92][10], puis 2 pour 1 000 en 1997 [FOM97][11] et 6 pour 1 000 en 2003 [FOM03][12]. Les statistiques issues du gouvernement américain montrent ainsi que le taux de prévalence de l'autisme augmente chaque année de l'ordre de 10 à 17%, cf. Fig. 5.1. De nos jours, la plupart des études internationales s'accordent à dire que la prévalence des TED se situerait autour des 1 pour 100. Ramené à la population française (~70M d'habitants), il y aurait environ 700 000 enfants présentant des troubles apparentés à l'autisme en France. L'autisme ne constitue donc pas un cas rare mais, au contraire, une véritable priorité de santé publique compte tenu de la difficulté que présentent ces enfants à s'insérer dans la société.

En 2002, le Ministère de l'Education Nationale estimait à environ 5%, les enfants concernés par les TSL pris dans leur ensemble, i.e., aussi bien oral qu'écrit [PLA02][13]. En 2004, Verloes et Excoffier évoquent des chiffres inférieurs en introduisant toutefois une fourchette : de 0,5% à 1% des enfants d'âge scolaire seraient atteints de TSL [VER04][14]. Au niveau international, le DSM IV[7] différencie : (i) les *troubles expressifs purs*, qui concernent 3 à 5 % de la population infantile et (ii) les *troubles mixtes* (expressifs et réceptifs) qui toucheraient 3% de cette même population. Les TSL constituent donc un phénomène de santé publique qui, comme pour l'autisme, est loin d'être un cas marginal en regard des statistiques fournies par les études effectuées sur le plan national comme international.

---

[10] E. Fombonne et C. du Mazaubrun, "Prevalence of infantile autism in 4 French regions", dans *Social Psychiatry and Psychiatric Epidemiology*, vol. 27, pp. 203–210, 1992.

[11] E. Fombonne, C. du Mazaubrun, C. Cans et H. Grandjean, "Autism and associated medical disorders in a French epidemiological survey", dans *J. of the Amer. Acad. of Child and Adolesc. Psychiatry*, vol. 36, pp. 1561–1569, 1997.

[12] E. Fombonne, "Epidemiological surveys of autism and other pervasive developmental disorders: an update", dans *J. of Autism and Develop. Disorders*, vol. 33, no. 4, pp. 365–382, 2003.

[13] M. Plaza, D. Chauvin, O. Lanthier, M. T. Rigoard, J. Roustit, M. P. Thibault, et M. Touzin, "Validation longitudinale d'un outil de dépistage des troubles du langage écrit. Etude d'une cohorte d'enfants dépistés en fin de CP et réévalués en fin de CE1", dans *Glossa*, vol. 81, pp. 22–33, 2002.

[14] A. Veroloes et E. Excoffier, "Dysphasie : aspects génétiques", dans C. L. Gerard et V. Brun [Eds], *Les dysphasies Rencontres en rééducation*, Paris : MASSON, pp. 17–22, 2004.

1.2. La prosodie dans les troubles de la communication (TC)

La plupart des enfants présentant des TC peuvent également montrer des troubles de la prosodie. La prosodie des sujets autistes est par exemple connue pour être différente de celle de leurs pairs, ce qui ajoute une barrière supplémentaire sur leur intégration sociale [ALL-92][15]. De plus, la barrière de communication liée à la prosodie, est bien souvent persistante, alors que les autres compétences du langage peuvent s'améliorer dans le temps ; [MCC03][16] et [PAU-05b][17]. L'incapacité à exploiter les fonctionnalités de la prosodie pour communiquer, contribue donc aux troubles du langage, de la communication et de l'interaction sociale. La prosodie atypique chez les individus souffrant de TC est alors devenue un sujet de recherche à part entière. Il est apparu qu'une sensibilisation à la prosodie sous-tend les compétences linguistiques et qu'une carence peut affecter à la fois le développement du langage et de l'interaction sociale de l'enfant.

Nous présentons dans les paragraphes suivants les principales caractéristiques qui ont été décrites dans la littérature sur les troubles de la prosodie chez les sujets atteints de TED ou de TSL. Nous spécifions les descriptions aux travaux portant sur l'*intonation* et l'*affect* puisque notre étude porte uniquement sur ces deux fonctionnalités prosodiques.

1.2.1. La prosodie dans les troubles envahissants du développement (TED)

La prosodie atypique a été identifiée comme une caractéristique centrale des individus atteints d'autisme [KAN43][3]. Les différences observées incluent : (i) une intonation monotone, (ii) des déficits en matière de contrôle du pitch et de l'intensité et (iii) une qualité vocale dite « concernée ». Notons que de nombreuses études ont tenté, i.e., sans parvenir à un véritable consensus, de définir les caractéristiques prosodiques des sujets atteints de troubles apparentés au spectre autistique ; cf. [WEL03][18] pour une revue de ces caractéristiques.

Fonctionnalités intonatives

Concernant la production de contours intonatifs et les tâches d'imitation des sujets TA, les résultats sont contradictoires. Dans une tâche de lecture à voix haute, il a, par exemple, été observé que les sujets ne différenciaient pas les questions des déclarations : toutes les phrases étaient produites avec un unique schéma de type déclaratif [FOS99][19]. De meilleures performances ont toutefois été obtenues dans une tâche (contrainte) d'imitation. Les auteurs ont donc conclu que *les sujets TA peuvent produire des structures intonatives, bien qu'ils ne sa-*

[15] D. A. Allen et I. Rapin, "Autistic children are also dysphasic", dans *Neurobiology of infantile autism*, H. Naruse and E. M. Ornitz [Eds], Amsterdam: Excerpta Medica, pp. 157–168, 1992.
[16] J. McCann et S. Peppé, "Prosody in autism: a critical review", dans *Inter. J. of Lang. and Comm. Disorders*, vol. 38, no. 4, pp. 325–350, Oct.-Dec. 2003.
[17] R. Paul, L. Shriberg, J. Mc Sweeny, D. Ciccheti, A. Klin et F. Volkmar, "Brief Report : relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders", dans *J. of Autism and Develop. Disorders,* vol. 35, no. 6, pp. 861–869, Dec. 2005.
[18] B. Wells et S. Peppé, "Intonation abilities of children with speech and language impairments", dans *J. of Speech, Lang. and Hearing Research*, vol. 46, pp. 5–20, Feb. 2003.
[19] S. Fosnot et S. Jun, "Prosodic characteristics in children with stuttering or autism during reading and imitation", dans proc. *14th Annual Congress of Phonetic Sc.*, San Francisco (CL), Aug. 1-7 1999, pp. 103–115.

chent pas convenablement les exploiter et qu'ils n'en comprennent pas leur valeur communicative. Paul *et al.* [PAU08][20] ont ainsi analysé les capacités à reproduire l'accent dans une tâche d'imitation de suites de syllabes sans sens chez divers sujets atteints de TC ; TA, SA et TED-NOS. Les mesures instrumentales ont révélé des différences faibles mais significatives entre les sujets TC et DT. Toutefois, une autre étude effectuée par ces mêmes auteurs, n'a pas permis de trouver une différence significative, entre des enfants TA et DT dans la façon d'utiliser l'intonation pour distinguer les questions des déclarations [PAU05a][21]. Ce résultat apparaît donc en contradiction avec ceux fournis par les études précédentes.

Fonctionnalités affectives

L'étude de Paul *et al.* [PAU05a] comprenait aussi une tâche de *production* du mamanais[22] et d'autres styles de parole adressée à un adulte. Aucune différence significative n'est apparue sur cette tâche entre les sujets autistes et les typiques. Le protocole comprenait également une épreuve qui consistait à produire des énoncés dans un style *calme* ou *énervé*. Une nouvelle fois, aucune différence significative n'a été retrouvée entre les deux groupes. Concernant l'aspect de la *perception*, Rutherford *et al.* [RUT02][23] ont évalué la capacité à interpréter le contenu émotionnel de 40 phrases chez 19 adultes à haut niveau de fonctionnement ou atteints du SA. Les résultats indiquent que ces deux groupes sont moins performants que les sujets contrôles. Paul *et al.* [PAU05a] n'ont, quant à eux, pu observer de différence significative entre les sujets TA et ceux présentant un DT dans la discrimination du mamanais et des autres styles de parole (y compris les styles *calme* et *énervé*) adressés à l'adulte.

Ainsi, les études qui ont cherché à comparer les stratégies employées par les sujets TED et à DT pour communiquer via les fonctionnalités *grammaticales* et *affectives* de la prosodie n'ont pas réussi à trouver des différences significatives entres les groupes de sujets, puisque des contradictions apparaissent dans les études. Cependant, les sujets TED sont supposés présenter des difficultés importantes dans le traitement des informations *pragmatiques* et *affectives*, puisqu'elles ont été identifiées comme une caractéristique centrale des sujets TA [KAN-43][3], et sont supposées être encore plus importantes pour les TED-NOS ; [DSM94][7].

1.2.2. La prosodie dans les troubles spécifiques du langage

Une théorie, celle de l'amorçage prosodique et phonologique, suggère que le traitement prosodique effectué en entrée par l'enfant serait un facteur clé lui permettant de décoder les structures syntaxiques et d'en faciliter l'acquisition [MOR96][24]. Il est donc probable que les troubles du langage qui sont présents dans la dysphasie proviennent d'une difficulté à traiter

[20] R. Paul, N. Bianchi, A. Agustyn, A. Klin et F. Volkmar, "Production of syllable stress in speakers with autism spectrum disorders", dans *Research in Autism Spectrum Disorders*, vol. 2, pp. 110–124, Jan.-Mar. 2008.

[21] R. Paul, A. Augustyn, A. Klin et F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders" dans *J. of Autism and Develop. Disorders*, vol. 35, no. 2, pp. 205–220, 2005.

[22] Ce langage correspond au style de parole adopté par à un adulte lorsqu'il s'adresse à un enfant en bas-âge. Les particularités reposent essentiellement sur un registre élevé de la voix et une intonation exagérée des phrases.

[23] M. D. Rutherford, S. Baron-Cohen et S. Wheelwright, "Reading the mind in the voice: a study with normal adults and adults with Asperger syndrome and high functioning autism", dans *J. of Autism and Develop. Disorders*, vol. 32, no. 3, pp. 189–194, Jun. 2002.

convenablement les caractéristiques de la prosodie, telles que l'intonation et/ou le rythme. Les enfants dysphasiques pourraient ainsi présenter des déficits prosodiques en compréhension, en production ou sur les deux versants, et ces déficits seraient uniquement causés par leurs troubles spécifiques de la parole et du langage [WEL03][18].

Fonctionnalités intonatives

L'intonation n'a été que très peu étudiée chez les enfants atteints de TSL [WEL03]. Alors que certaines études ont conclu que les TSL n'ont pas de déficit significatif et que l'usage de l'intonation ne serait pas pénalisée par leurs troubles du langage, [SNO98a][25] et [MAR09][26], d'autres ont montré des déficits relativement faibles mais significatifs entres les TSL et les DT [WEL03], [HAR89][27] et [SAM03][28]. Il a par exemple été montré que les sujets TSL produisent moins de contours congruents que les enfants à DT [WEL03]. Les auteurs ont alors supposé que « *les dysphasiques comprennent le contexte pragmatique, mais n'arrivent pas à sélectionner le contour intonatif correspondant* ». D'autres résultats ont ainsi montré que les TSL ont des capacités plus faibles à imiter les contours prosodiques que des enfants à DT [MEU-97][30] et [WEL03]. Cependant, il a été montré, de façon contradictoire, que les sujets TSL présentent les mêmes capacités d'imitation de l'intonation que des enfants à DT [MAR09].

Fonctionnalités affectives

L'étude de Wells et Peppé [WEL03] ne montre aucune différence significative entre les enfants à TSL et les typiques sur la capacité à exprimer l'affect. Ces résultats ont également été observés dans l'étude de van der Meulen *et al.* [MEU97][29] qui comprenait une tâche de reconnaissance des phrases émotionnelles préenregistrées par un comédien ; *émotions Peur, Joie, Colère* et *Tristesse.* Ces résultats sont conformes avec les critères diagnostics des TSL qui énoncent l'absence de déficiences dans le traitement des informations pragmatiques.

1.3. Evaluations de troubles dans la prosodie

La littérature montre qu'il n'existe que très peu de consensus dans les descriptions des troubles de la prosodie des enfants atteints de TA, de TED-NOS et de TSL. Les chercheurs et les cliniciens partagent ainsi l'objectif ambitieux, d'évaluer les compétences prosodiques des enfants atteints de TC au moyen de tests appropriés. Le but principal étant de déterminer les

[24] J. Morgan et K. Demuth, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, dans J. Morgan and K. Demuth [Eds], Mahwah, NJ: Erlbaum, 1996.

[25] D. Snow, "Prosodic markers of syntactic boundaries in the speech of 4-year-old children with normal and disordered language development", dans *J. of Speech, Lang. and Hearing Research,* vol. 41, pp. 1158–1170, Oct. 1998.

[26] C. R. Marshall, S. Harcourt Brown, F. Ramus et H. J. K. Van der Lely, "The link between prosody and language skills in children with SLI et/or dyslexia", dans *Inter. J. of Lang. and Comm. Disorders,* vol. 44, no. 4, pp. 466–488, Jul. 2009.

[27] P. Hargrove et C. P. Sheran, "The use of stress by language impaired children", dans *J. of Comm. Disorders,* vol. 22, no. 5, pp. 361–373, Oct. 1989.

[28] C. Samuelsson, C. Scocco et U. Nettelbladt, "Towards assessment of prosodic abilities in Swedish children with language impairment", dans *Logopedics Phoniatrics Vocology*, vol. 28, no. 4, pp. 156–166, Oct. 2003.

[29] S. van der Meulen et P. Janssen, "Prosodic abilities in children with Specific Language Impairment", dans *J. of Comm. Disorders,* vol. 30, pp. 155–170, May-Jun. 1997.

caractéristiques prosodiques spécifiques aux types de TC, afin d'en améliorer le diagnostic et la prise en charge orthophonique.

Nous proposons d'utiliser dans cette étude, des traitements et des méthodes automatiques pour évaluer les fonctionnalités prosodiques *grammaticales* (i.e., reproduction de la modalité de la phrase) et *affectives* (i.e., narration spontanée d'une histoire imagée et chargée d'affect) chez des enfants atteints de divers types de TC : (i) TA, (ii) TED-NOS et (iii) TSL. Avant de présenter cette étude, nous décrivons tout d'abord les méthodes qui ont été proposées dans la littérature pour évaluer, de façon manuelle ou automatique, les caractéristiques de la prosodie chez un sujet donné.

1.3.1. Méthodes manuelles

Les procédures manuelles d'évaluation de la prosodie, comme celles venant des Etats-Unis : [PAU05a][21] et [SCH90][30], le test Britannique : PROP [CRY82][31], la méthode Suédoise : [SAM03][28] et le PEPS-C : [MAR08][32], nécessitent des jugements d'experts pour évaluer les compétences prosodiques de l'enfant. La prosodie peut ainsi être évaluée en enregistrant un échantillon de parole et en s'accordant sur les fonctions et les formes communicatives alors identifiées. Cette méthode requiert en conséquence une transcription experte des informations prosodiques. De plus, l'aspect non-contraint des enregistrements implique nécessairement des formes variées de prosodie selon les locuteurs, ce qui complique l'analyse des données.

Par conséquent, les fonctionnalités principales de la prosodie sont évaluées dans un cadre contraint pour le PEPS-C [MAR08]. Dans cette méthode, le programme d'évaluation affiche des images (cf. Fig. 5.2) sur un écran d'ordinateur à la fois comme stimuli pour les énoncés expressifs (*sortie*) et comme choix de réponse en perception des stimulus acoustiques joués par l'ordinateur (*entrée*). Deux réponses sont possibles pour chaque objet proposé lors de l'évaluation en *entrée*. La demande sur la mémoire auditive des sujets est donc très limitée. Cette limitation crée un biais dans le calcul du score, mais ce dernier est réduit par le nombre plutôt élevé d'objets disponibles pour chaque tâche. Pour l'évaluation en *sortie*, l'examinateur doit juger si la prosodie des phrases produites par les enfants correspond au stimulus de chaque tâche. Les options de scores données à l'examinateur sont catégorisées en deux ou trois possibilités afin de juger l'imitation : *correcte* / *incorrecte*, ou *bonne* / *moyenne* / *mauvaise*. Cette procédure ne nécessite pas un niveau d'expertise très élevé puisque le nombre de catégories disponibles pour juger la production de la prosodie est faible. Néanmoins, on peut se demander si la richesse des informations portées par la prosodie peut être évaluée de façon aussi restrictive. De façon alternative, utiliser un plus grand nombre de catégories pourrait rendre difficile le choix de celle jugée par l'examinateur comme étant la plus appropriée parmi toutes les autres. Le compromis semble donc difficile à trouver dans les procédures d'évaluation manuelle de la prosodie.

---

[30] L. D. Schriberg, J. Kwiatkowski, et C. Rasmussen, *The Prosody-Voice Screening Profile*. Tuscon, AZ: Communication Skill Builders, 1990.

[31] D. Crystal, *Profiling Linguist. Disability*. Edward Arnold, London, 1982.

[32] P. Martínez-Castilla et S. Peppé, "Developing a test of prosodic ability for speakers of Iberian-Spanish", dans *Speech Comm.*, vol. 50, no. 11-12, pp. 900–915, Mar. 2008.

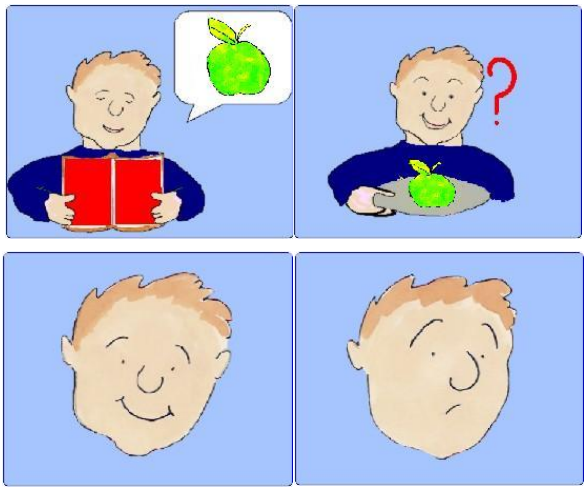


Fig. 5.2 Exemples d'images utilisées dans le PEPS-C pour évaluer la production et la perception des fonctionnalités *grammaticales* (images du haut) et *affectives* (images du bas) de la prosodie ; images extraites de [MAR08][32].

1.3.2. Méthodes automatiques

Des études récentes ont proposé des systèmes automatisés permettant d'évaluer la production de la prosodie [SAN09][33], les troubles de la parole [MAI09][34], et de l'alphabétisation précoce [BLA09][35]. De multiples défis doivent être rencontrés par ces systèmes lors de la caractérisation des particularités prosodiques associées aux types de TC étudiés. Ces défis concernent essentiellement l'étape d'extraction de caractéristiques, puisque la prosodie de la parole est associée à de nombreuses composantes (e.g., pitch, intensité, qualité vocale et rythme) qui sont toutes incluses dans l'onde acoustique. De plus, ces paramètres présentent une forte variabilité qui est due à des facteurs contextuels (e.g., perturbations pendant l'enregistrement) ou liés aux idiosyncrasies du locuteur (e.g., affect [LEE05][36] et style de parole [LAA97][37]). Les caractéristiques acoustiques, lexicales et linguistiques du discours des enfants sont également corrélées avec l'âge et le genre, pour de la parole sollicitée comme spontanée [POT07][38]. De plus, extraire la fréquence fondamentale (f0) sur le signal de parole produit par un enfant s'avère plus délicat à effectuer que pour un adulte. Cette difficulté est encore plus élevée pour l'extraction des formants puisque leur énergie est censée être moindre comparée à celle du fondamental [ROD09][39].

[33] J. P. H. van Santen, E. T. Prud'hommeaux et L. M. Black, "Automated assessment of prosody production", dans *Speech Comm.*, vol. 51, no. 11, pp. 1082–1097, Nov. 2009.

[34] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster et E. Nöth, "PEAKS – A system for the automatic evaluation of voice and speech disorder", dans *Speech Comm.*, vol. 51, no. 5, pp. 425–437, May 2009.

[35] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee et S. Narayanan, "Automatic pronunciation verification of english letter-names for early literacy assessment of preliterate children", dans proc. *ICASSP*, Taipei, Taiwan, Apr. 19-24 2009, pp. 4861–4864.

[36] C. M. Lee et S. Narayanan, "Toward detecting emotions in spoken dialogs", dans *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 2, pp. 293–303, Feb. 2005.

[37] Gitta P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and read speaking style", dans *Speech Comm.*, vol. 22, pp. 43–65, Mar. 1997.

[38] A. Potamianos et S. Narayanan, "A review of the acoustic and linguistic properties of children's speech", dans proc. *IEEE 9th W. on Multimedia Signal Process.*, Chania, Greece, Oct. 23 2007, pp. 22–25.

Puisque caractériser la prosodie de la parole est difficile, en particulier dans un contexte de production spontanée et pour des voix d'enfant, van Santen et al. [SAN09][33] ont défini six principes de conception : (i) des *méthodes très contraignantes* visant à réduire les variations non-voulues dans les données et pouvant être causées par des facteurs externes à la procédure d'évaluation, (ii) une *conception de type « paires prosodiques minimales »* pour au moins une tâche afin d'étudier le contraste prosodique, (iii) des *caractéristiques acoustiques robustes* qui, dans l'idéal, détectent automatiquement les tours de parole, les erreurs de pitch et les évènements extérieurs à la tâche [OLL10][40], (iv) une *fusion des caractéristiques pertinentes* pour déterminer leur importance dans la caractérisation des troubles de la prosodie, (v) des *caractéristiques globales et dynamiques* pour mesurer des contrastes spécifiques dans la prosodie et (vi) des *techniques de paramètres-libres* dans lesquelles les algorithmes sont soit basés sur des faits établis (e.g., le phénomène d'allongement en fin de phrase), soit développés dans des analyses exploratoires sur un ensemble de données distinct dont les caractéristiques seraient très différentes des données principales des locuteurs.

Le système proposé par van Santen *et al.* [SAN09] évalue la prosodie sur ses fonctions *grammaticales* (e.g., accent lexical et frontière de phrase), *pragmatiques* (e.g., focus et style) et *affectives* (e.g., émotions). Les scores sont estimés par des humains et une machine au moyen de caractéristiques spectrales, de la f0 et de la durée. Il a été constaté dans presque toutes les tâches que les résultats fournis par la machine étaient corrélés avec la moyenne des jugements portés par l'homme et de la même façon que pour les valeurs de corrélation des scores inter-juges. Des résultats similaires ont été obtenus avec le système appelé PEAKS [MAI09][37] et dans lequel des chaînes de Markov cachées (HMM) sont exploitées pour évaluer des troubles dans la parole et dans la voix chez des sujets atteints de divers conditions pathologiques (e.g., ablation du larynx, fente labiale ou palatine). Des résultats satisfaisants ont aussi été obtenus sur des sujets TA et TSL en bas-âge (10 – 48 mois), puisqu'ils ont été significativement différenciés des DT par une analyse automatisée de la prosodie sur des enregistrements naturels (système portatif) [OLL10].

Ainsi, les méthodes automatiques visant à caractériser les troubles de la parole et de la prosodie sont en mesure de fournir des scores comparables à ceux obtenus par des jugements humains experts, notamment lorsque le système tend à inclure les prérequis mentionnés par van Santen *et al.* [SAN09]. De plus, ces méthodes permettent de dépasser les limitations dues à une évaluation catégorielle des caractéristiques prosodiques, puisque les paramètres pertinents sont identifiés par des traitements automatiques. Ainsi, l'analyse des particularités prosodiques d'un sujet donné peut s'effectuer avec une précision et une fiabilité (ou objectivité) qui ne peut pas être atteinte par une évaluation humaine.

---

[39] W. R. Rodríguez et E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool", dans proc. *ISCA Inter. W. on Speech and Language Techn. in Educ.*, Wroxall Abbey Estate, UK, Sep. 3-5 2009.

[40] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel et S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development", dans *PNAS of the USA*, Jun. 2010, Appendix, pp. 16–21.

1.4. Objectifs de notre étude

L'objectif principal de notre étude consiste à étudier (en étroite collaboration avec des cliniciens), l'intérêt de l'emploi de techniques de traitement du signal et de reconnaissance des formes, pour analyser les caractéristiques prosodiques de sujets atteints de divers types de TC. Un tel système doit pouvoir différencier les patients atteints de TC des enfants à DT, puisque les troubles dans la prosodie est une caractéristique clinique connue des TED, cf. sous-section 1.2. Les traitements automatiques utilisés pour identifier les particularités prosodiques associées aux types de TC, permettent de surmonter les difficultés créées par des évaluations catégorisées et par les biais introduits par des jugements subjectifs. En effet, les corrélats de la prosodie sont : (i) beaucoup trop nombreux pour être entièrement catégorisés par l'Homme, et (ii) ne peuvent pas être jugées de façon fiable par les humains qui ont une opinion subjective [KEN96][41] dont la variabilité inter-juge est également problématique ; les préjugés et les incohérences dans le jugement perceptif ont été listés [TVE69][42], et les caractéristiques pertinentes pour la caractérisation de la prosodie ont été définies [PEN07][43], [SCH07b][44]. Cependant, malgré les récents progrès réalisés dans la recherche d'un large éventail de caractéristiques prosodiques, il n'existe pas de réel consensus sur les paramètres qui sont les plus pertinents. Raison pour laquelle nous avons proposé des techniques d'extraction de caractéristiques qui reposent sur différents ancrages acoustiques de la parole et qui incluent des modèles *non-conventionnels* du rythme.

Les travaux que nous allons présenter dans les sections suivantes ont été réalisés en étroite collaboration avec différentes équipes de chercheurs et de cliniciens. Ces travaux ont notamment fait l'objet d'une récente publication [RIN10][45]. Pour l'aspect clinique, nous avons tout spécialement collaboré avec les services de pédopsychiatrie du Pr. D. Cohen à l'Hôpital de la Pitié-Salpêtrière, et du Pr. B. Golse à l'Hôpital Enfants-malades Necker (représenté par le Dr. L. Robel). La psychologue Dr. M. Plaza a participé à cette étude, ainsi que l'étudiante J. Demouy dans le cadre de la réalisation de son mémoire d'orthophonie. Nous avons aussi profité de rencontres effectuées dans le cadre de l'action européenne 2102 du COST pour nouer des contacts avec une équipe de chercheurs en TAP de l'université de Budapest, Hongrie (équipe TMI dirigée par K. Vicsi). Cette équipe est spécialisée dans la modélisation dynamique de la prosodie. Je profite de ce paragraphe pour remercier toutes les personnes qui ont participé à notre projet, et sans lesquelles, nous n'aurions pu mener à bien les études qui sont présentées dans ce chapitre.

---

[41] R. D. Kent, "Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders", dans *Amer. J. of Speech-Lang. Pathology*, vol. 5, no. 3, pp. 7–23, Aug. 1996.

[42] A. Tversky, "Intransitivity of preferences", dans *Psychological Review*, vol. 76, pp. 31–48, Jan. 1969.

[43] A. Pentland, "Social Signal Processing", dans *IEEE Signal Proc. Magazine*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[44] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous et V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals", dans proc. *Interspeech ICSLP*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2253–2256.

[45] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, et M. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children", dans *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. PP, no. 4, Oct. 2010.

Nos travaux portent sur la caractérisation des fonctionnalités *grammaticale* (i.e., modalités de phrase) et *affective* de la prosodie (i.e., émotions) de sujets natifs Français et atteints de divers TC (e.g., TA, TED-NOS et TSL), cf. sous-section 1.1. Nous avons cherché à exploiter les techniques proposées dans cette thèse, pour comparer les particularités prosodiques entre les différents groupes de sujets et selon deux types de tâches bien distinctes : (i) *imitation de contours intonatifs* (tâche contrainte), et (ii) *production de parole affective spontanée* (tâche non contrainte). Cette étude a été en partie financée par la fondation Orange pour l'autisme dans le cadre d'un mécénat d'entreprise[46].

2. Recrutement et enregistrement des sujets

Le recrutement et les évaluations cliniques des sujets ont été réalisés par des pédopsychiatres spécialisés dans les TC. Nous décrivons dans cette section, les caractéristiques cliniques et sociodémographiques des sujets, ainsi que les épreuves et le protocole de passation qui a été utilisé pour collecter les données de notre étude.

2.1. Recrutement de sujets atteints de troubles de la communication

Trente-cinq sujets monolingues (Français) âgés de 6 à 18 ans ont été recrutés dans deux départements universitaires de psychiatrie de l'enfant et de l'adolescent à Paris, France : (i) *Université Pierre et Marie Curie*, Hôpital de la Pitié-Salpêtrière et (ii) *Université René Descartes*, Hôpital Necker. Ces sujets ont été diagnostiqués comme TA, TED-NOS ou TSL selon les critères du DSM IV[7]. Les diagnostics TA et TED-NOS ont été attribués selon les scores obtenus par les patients sur l'ADI-R [LOR94][47] et la CARS [SCH80][48]. L'ADI-R, pour *autism diagnosis interview-revised*, est un entretien diagnostique parental qui détaille le développement de l'enfant et permet de rechercher les signes du syndrome autistique vers l'âge de 5 ans. La CARS, pour *child autism rating scale*, est une échelle clinique qui évalue la sévérité de la symptomatologie autistique. Les caractéristiques sociodémographiques et cliniques des sujets sont résumées dans la table 5.1. Tous les sujets TC ont reçu une évaluation psychométrique, pour laquelle, aucun des sujets pathologiques n'a montré un retard mental (QI > 70).

2.2. Evaluation du langage oral des sujets pathologiques (ELO)

Les sujets ont reçu une évaluation portant sur : (i) le *vocabulaire réceptif*, (ii) le *vocabulaire expressif* et (iii) la *répétition de mots* ; batterie de tests ELO [KHO01][49]. Comme ces tests sont dédiés aux enfants de 3 à 11 ans et que notre étude comporte de nombreux sujets âgés de plus de 11 ans, les scores ont été ajustés en niveaux de sévérité ; un effet plafond trop

[46] Site internet de la fondation : http://www.orange.com/fr_FR/mecenat/fondation/sante/.

[47] C. Lord, M. Rutter et A. Le Couteur, "Autism Diagnostic Interview-Revised: A revision version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders", dans *J. of Autism and Develop. Disorders*, vol. 24, no. 5, pp. 659–685, Oct. 1994.

[48] E. Schopler, R. Reichler, R. Devellis et K. Daly, "Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS)", dans *J. of Autism and Develop. Disorders*, vol. 10, no. 1, pp. 91–103, 1980.

Table 5.1 Caractéristiques sociodémographiques et cliniques des sujets recrutés.

Caractéristique	TA	TED-NOS	TSL
Age	9.8 $_{3.5}$	9.8 $_{2.2}$	9.8 $_{3.9}$
#Garçon – #Fille	10 – 2	9 – 1	10 – 3
Scores ADI-R			
Déficit social	21.1 $_{5.8}$	12.7 $_{7.8}$	N.P.
Communication	19.3 $_{5.2}$	8.5 $_{6.4}$	N.P.
Intérêt respectif	6.4 $_{2.4}$	2.0 $_{1.6}$	N.P.
Total	50.7 $_{12.8}$	25.7 $_{15.4}$	N.P.
Scores CARS	33.2 $_{15.4}$	22.3 $_{5.4}$	N.P.

[Valeur moyenne] $_{(\text{écart-type})}$ pour les valeurs d'âge et de score ; TA : troubles autistiques ; TED-NOS : troubles envahissants du développement non-spécifiés ; TSL : troubles spécifiques du langage ; ADI-R : interview révisée du diagnostic de l'autisme [LOR94][47] ; CARS : échelle d'évaluation de l'autisme [SCH80][48] ; N.P. : non pertinent.

Table 5.2 Niveau de sévérité dans les compétences basiques de langage des sujets pathologiques selon les tâches d'ELO [KHO01][49].

Tâche d'ELO	TA	TED-NOS	TSL
Vocabulaire réceptif	2.4 $_{1.6}$	1.9 $_{1.5}$	1.9 $_{1.0}$
Vocabulaire expressif	2.0 $_{1.8}$	1.2 $_{1.8}$	1.4 $_{1.0}$
Répétition de mots	2.9 $_{1.5}$	2.7 $_{1.4}$	3.5 $_{0.7}$

TA : troubles autistiques ; TED-NOS : troubles envahissants du développement non-spécifiés ; TSL : troubles spécifiques du langage (dysphasie) ; les valeurs de scores sont données dans le format suivant : [Valeur moyenne] $_{(\text{écart-type})}$.

important aurait été obtenu sur les sujets ayant des troubles du langage. Ainsi, il a été déterminé pour chaque sujet la correspondance entre l'âge pour chaque score et l'écart entre « l'âge verbal » et « l'âge chronologique ». La différence a ensuite été convertie en niveaux de sévérité via une échelle de Likert ; « 0 » pour le niveau attendu à l'âge chronologique, « 1 » pour un écart de 1 an à partir du niveau attendu à l'âge chronologique, …, et « 4 » pour 4 ans ou plus d'écart. Les 3 groupes de TC ont montré un retard équivalent à 1 ou 2 ans par rapport à leur âge chronologique sur les tâches de *vocabulaire*, cf. table 5.2. Ils ont également présenté des difficultés similaires mais plus importantes dans la tâche de *répétitions* de mots, qui exige des compétences phonologiques.

2.3. Recrutement des sujets contrôles

Un groupe contrôle composé d'enfants monolingues (N = 70), appariés en âge chronologique (âge moyen = 9.8 ans, écart type = 3.3 ans) avec un ratio de 2 enfants à DT pour 1 enfant TC, a été recruté dans le lycée privé Hermitage, Maisons-Laffitte, Hauts-de-Seine. Le recrutement des enfants a été réalisé sans difficulté grâce à la sagacité des responsables de ce lycée. Les parents d'élèves ont rempli les autorisations pour enregistrer les enfants. Il a été vérifié qu'aucun sujet « contrôle » n'a eu d'antécédent en trouble de la parole, du langage, de l'audition, ou de quelconques problèmes liés à l'apprentissage. La population témoin n'a pas été évaluée avec les batteries de test ELO puisqu'elle représente la norme.

[49] A. Khomsi, *Evaluation du Langage Oral*. Paris: ECPA, 2001.

2.4. Epreuves de notre étude

Nous avons conçu pour notre étude, deux épreuves courtes et à caractère ludique : (i) *imitation de contours intonatifs* représentants différents types de phrases (e.g., déclarative, interrogative) et (ii) *production de parole affective spontanée* (narration d'une histoire imagée). La première épreuve permet d'évaluer dans un cadre contraint les capacités basiques en compréhension et en production de la prosodie. La nécessité de cette épreuve basique s'impose avant de faire appel à des notions plus complexes telles que l'affect. La seconde épreuve repose sur le récit d'une histoire imagée sollicitant (de façon spontanée) de la parole affective. Notons que nos épreuves étudient d'une seule traite, les caractéristiques en production et en perception de la prosodie, puisque la littérature a montré des résultats contradictoires avec une analyse séparée de ces deux capacités chez les sujets atteints de TC, cf. sous-section 1.2.

2.4.1. Contrainte : « *imitation de contours intonatifs* »

Les tâches d'imitation sont couramment effectuées par les patients atteints de TC, même pour ceux présentant des TA (du moment qu'ils ont accès au langage oral) [NAD02][50]. Chez un patient, cette capacité peut être utilisée pour tester le domaine prosodique sans limitations dues à un handicap linguistique. Les tâches d'imitation introduisent toutefois un biais dans les données en raison d'un manque de spontanéité lors de la production. En conséquence, les contours intonatifs qui ont été reproduits par les sujets de notre étude peuvent être assez éloignés des stimulus d'origine. Mais l'influence de ce biais peut être minimisée par la stratégie de reconnaissance adoptée pour caractériser la prosodie de l'intonation, cf. sous-section 3.2.4.

Présentation du matériel

Suivant les règles de la prosodie du français, 26 phrases représentant différents types de modalité (cf. table 5.3) et quatre types de profil intonatif (cf. Fig. 5.3) ont été définis pour la tâche d'imitation. Afin de faciliter la reproductibilité et d'éviter une demande trop importante au niveau cognitif, les phrases choisies sont phonétiquement faciles à reproduire et relativement courtes. Ces phrases, qui ont servi de stimuli pour l'épreuve, ont été enregistrées par nos soins au moyen du logiciel *Wavesurfer* [SJO00][51]. Ce dernier a permis de vérifier que le contour intonatif des phrases correspondait bien au profil intonatif de leur groupe, cf. Fig. 5.3.

Protocole de passation

Les phrases ont été proposées aux sujets dans un ordre aléatoire pour limiter la perception des groupes intonatifs à travers les phrases. Un seul fichier « son » a ainsi été généré à partir des 26 phrases en les concaténant après un tri aléatoire. Un blanc de 5 secondes a été inséré

[50] J. Nadel, "Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism", dans *The imitative mind: Development, evolution and brain bases*, A. N. Meltzoff and W. Prinz [Eds], Cambridge University Press, pp. 2–14, 2002.

[51] K. Sjöleter et J. Beskow, "WaveSurfer - an open source speech tool", dans proc. *6th ICSLP*, vol. 4, Beijing, China, Oct. 16-20 2000, pp. 464–467. Disponible à l'adresse suivante : http://www.speech.kth.se/wavesurfer/.

entre chaque phrase, afin d'envisager une éventuelle continuité dans la tâche. La consigne a été la suivante : « Maintenant tu vas entendre des phrases. Tu dois les répéter exactement de

Table 5.3 Matériel de parole utilisé pour la tâche d'*imitation de contours intonatifs*.

Intonation	Modalité	Phrase
Descendante ***affirmations***	Déclarative, affirmative	"David a mangé un croissant."
		"Je viens d'arriver de l'école."
	Déclarative, négative	"Cette maison ne me plaît pas du tout."
		"Il n'est pas encore l'heure."
	Déclarative, dubitatif	"Je ne suis pas sûr de pouvoir le faire."
		"Il me semble qu'il ne soit pas encore prêt."
	Exclamatoire, emphatique	"C'est Rémy qui va être content."
		"C'est ainsi que vont les choses."
Tombante ***questions / affirmations***	Interrogative	"Où se tient-il ?"
		"Comment vas-tu ?"
	Interrogative, ordre réduit	"Pouvez-vous passer à mon bureau ?"
		"Pourriez-vous nous accorder un instant ?"
	Exclamatoire	"Oh non, je ne te le donnerais pas."
		"Comme je suis content !"
	Impérative, ordre/conseil	"Ne l'abîme pas !"
		"Dis-moi la vérité !"
Flottante ***affirmations***	Déclarative	"Anna viendra avec toi."
		"Je suis très content que tu sois venu."
	Exclamatoire	"J'aime les crêpes au chocolat."
		"Il n'aime pas le sucre en poudre."
Montante ***questions***	Interrogative, questions courtes	"Qui ?"
		"Un croissant ?"
		"Pardon ?"
		"A l'intérieur ?"
		"Ah bon ?"
		"Quoi ?"

la même manière que tu vas les entendre. ». Plusieurs écoutes ont été proposées aux enfants qui ont semblé éprouver des difficultés, mais aucun retour quant à leurs performances n'a été donné par l'examinateur. La consigne a néanmoins due être adaptée pour certains sujets qui ont eu le sentiment d'être évalués : « Tu sais, il n'y a pas de bonnes ni de mauvaises réponses. Il faut juste que tu répètes les phrases que tu vas entendre. ». Les enregistrements issus de l'épreuve d'*imitation des contours intonatifs* totalisent une durée de 4h16min pour les sujets à DT, 58min pour les TA, 39min pour les TED-NOS, et 53min pour les TSL.

Prétraitements des données en vue de leur traitement automatique

Comme les systèmes de reconnaissance sont relativement sensibles aux bruits présents dans les enregistrements, les données fournies par les passations ont été soigneusement contrôlées. Nous avons notamment exclu les phrases qui contenaient : (i) un faux-départ, (ii) des répétitions, (iii) des bruits issus de l'environnement et/ou (iv) de la parole non liée à la tâche[52]. Si bien que nous avons conservé, au total, 2 772 phrases correspondant à 1h de parole pour effectuer l'analyse, cf. table 5.4. Cette phase de prétraitement a notamment diminué d'un facteur 7 la durée totale des enregistrements.

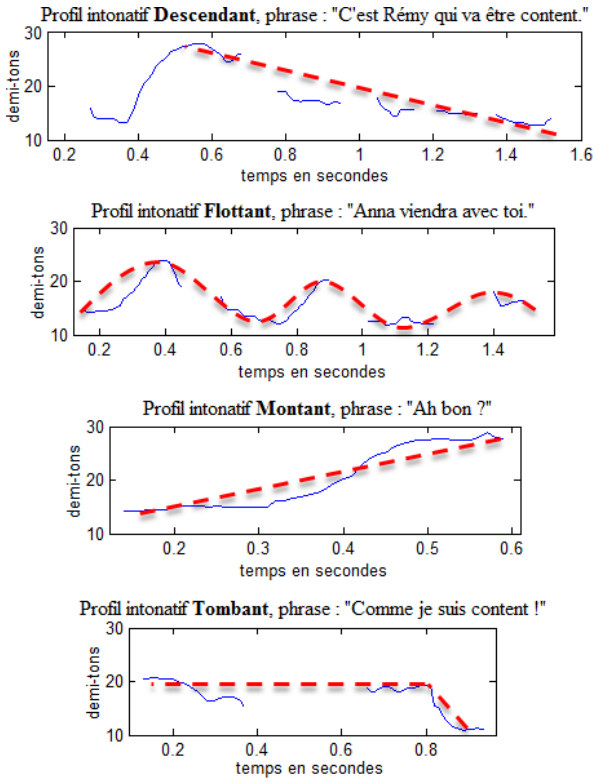


Fig. 5.3 Profils intonatifs selon le contour du pitch : (i) profil *descendant*, (ii) profil *flottant*, (iii) profil *montant*, et (iv) profil *tombant* ou « *à accents* » ; les valeurs estimées du pitch sont représentées par des lignes en trait continu, et les profils intonatifs par des lignes en pointillées.

Table 5.4 Quantité de phrases disponibles selon les groupes pour l'analyse de la tâche d'*imitation de contours intonatifs*.

Intonation	REF	DT	TA	TED-NOS	TSL
Descendante	8	580	95	71	103
Tombante	8	578	94	76	104
Flottante	4	291	48	40	52
Montante	6	432	70	60	78
Toutes	**26**	**1 881**	**307**	**247**	**337**

REF : stimuli ; DT : sujets contrôles ; TA : sujets autistes ; TED-NOS : sujets dysharmoniques ; et TSL : sujets dysphasiques.

2.4.2. Non-contrainte : « *production de parole affective spontanée* »

La deuxième épreuve de notre protocole est de nature non contrainte et porte sur l'évaluation de la capacité des enfants à structurer spontanément les dimensions affectives de la parole par la prosodie. Cette épreuve est de nature non contrainte car nous souhaitons étudier le comportement naturel de l'enfant lors de la description de stimuli implicitement affectifs,

[52] Le choix d'une segmentation manuelle plutôt qu'automatique des phrases s'explique par notre volonté de s'assurer de la qualité des données exploitées pour effectuer les TAP. La tâche de segmentation serait plutôt aisée à réaliser puisque les stimuli d'origine sont présents dans les enregistrements et que l'ordre des phrases est connu à l'avance. De plus, la détection des faux-départs, des répétitions et des bruits environnementaux peut également être effectuée automatiquement, [YIL09][66] et [OLL10][40]. Enfin, la détection automatique de la parole non liée à la tâche pourrait être effectuée à travers un système de reconnaissance de la parole.

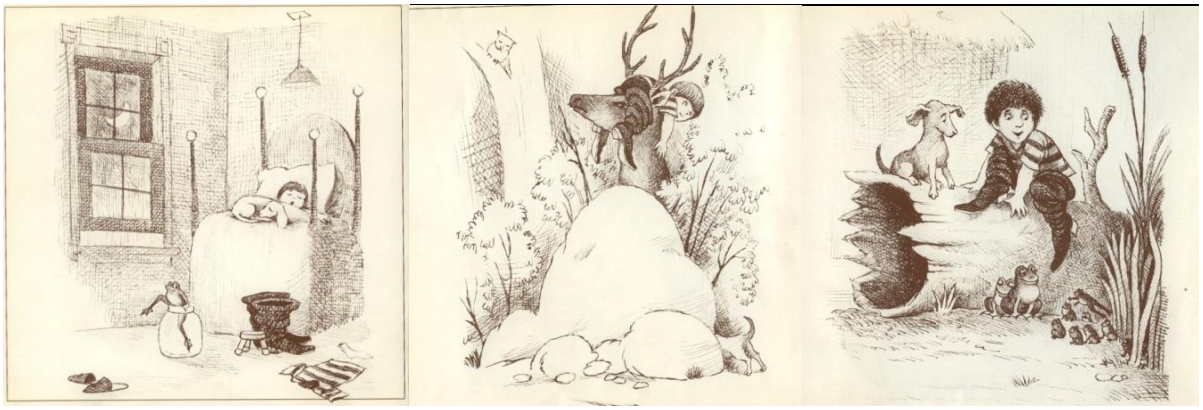


Fig. 5.4 Extraits du livre « *Frog where are you ?* » qui a été utilisé lors de l'épreuve des émotions.

i.e., contenus dans les images d'une histoire. Ainsi, les descriptions fournies par les sujets lors de la tâche contiennent non seulement les émotions communiquées par ces derniers, mais aussi, celles qui ont été perçues à travers les images de l'histoire ; comme cela était le cas pour la première épreuve, i.e., *entrée* et *sortie* des fonctionnalités prosodiques analysées d'un seul bloc. Enfin, notons que les expériences portant sur les émotions ne sont pas dépendantes de la langue (contrairement à la première épreuve), puisqu'elles reposent sur la description d'une histoire au moyen d'un livre contenant uniquement des images.

Présentation du matériel

La seconde épreuve consistait en un récit d'une histoire imagée [MAY69][53]. Cette dernière contient des stimulus affectifs qui ont été catégorisés par une pédopsychiatre en quatre niveaux de valence émotionnelle : *positive*, *neutre* (i.e., sans valence émotionnelle particulière), *négative* et *ambivalent* (i.e., contenant plusieurs catégories de valence), cf. table 5.5. Le choix d'un livre reposant uniquement sur des images, permet de s'assurer que le discours des sujets ne sera pas influencé par des indications écrites. Notons que des études cliniques ont déjà été conduites sur le livre que nous avons choisi [MAY69] ; [THU93][54].

Ce dernier met en scène, les aventures d'un petit garçon parti à la recherche de sa grenouille qui s'est échappée dans la nuit, cf. Fig. 5.4. L'ensemble des péripéties rencontrées par les personnages permet de susciter des descriptions spontanées qui peuvent être chargées d'émot-ions. Notons toutefois, l'hypothèse forte qui est faite dans cette épreuve puisque nous supposons que le style de parole (i.e., les caractéristiques prosodiques) produit lors du récit spontané des images est corrélé avec les catégories de valence émotionnelle définies par la pédopsychiatre. Or, les images associées aux catégories d'émotions ne représentent pas des archétypes picturaux de l'affect, mais plutôt une histoire dans laquelle un jeune garçon vit tantôt des évènements *positifs*, tantôt des évènements *négatifs*. Les stimulus sont donc purement liés à la pragmatique et les émotions peuvent en conséquence être mal comprises (ou mal interprétées) par les enfants.

[53] M. Mayer, *Frog where are you?*, dans New York: Dial Books for young readers, 1969.

[54] C. Thurber et H. Tager-Flusberg, "Pauses in the narratives produced by autistic, mentally retarded, and normal children as an index of cognitive demand", dans *J. of Autism and Develop. Disorders*, vol. 23, no. 2, pp. 309–322, 1993.

Table 5.5 Catégorisation selon le degré de valence affective des images contenues dans le livre « *Frog where are you ?* » [MAY69][53].

Page	Valence affective	Page	Valence affective
1	neutre	16	négative
2	neutre	17	négative
3	négative	18	négative
4	négative	19	négative
5	neutre	20–21	négative
6	négative	22	négative
7	ambivalente (pos. / neg.)	23	positive
8–9	neutre	24	ambivalente (neu. / pos.)
10	neutre	25	ambivalente (neu. / pos.)
11	négative	26	positive
12–13	négative	27	positive
14–15	négative	28–29	positive

Protocole de passation

Les images de l'histoire ont toutes été scannées et présentées, une par une, aux enfants sous la forme d'un diaporama affiché en plein écran sur un ordinateur portable. La consigne a été la suivante : « Je vais te montrer une histoire en images sur l'ordinateur. Tu vas me raconter l'histoire comme si tu la racontais à quelqu'un qui ne la connaissait pas, comme à ton frère ou à ta sœur par exemple ». Par la suite et de manière à ne pas influencer les productions des sujets, aucun étayage n'a été fourni. Seules des relances neutres de type « et puis ? » ont été proposées par l'examinateur pour étayer le discours ou que le sujet confirme le passage à l'image suivante. Le défilement des images a été contrôlé de façon à laisser le patient exprimer tout ce qu'il avait à dire sur l'image. Certains sujets autistes ont souhaité faire défiler eux-mêmes les images, ce qui a été autorisé par l'examinateur. Les données issues de l'épreuve de *production de parole spontanée affective* totalisent 7h38min d'enregistrements pour les sujets à DT, 1h35min pour les TA, 1h12min pour les TED-NOS, et 1h56min pour les TSL.

Prétraitements des données en vue de leur traitement automatique

Pour les mêmes raisons que celles évoquées dans la première épreuve (i.e., sensibilité des systèmes de reconnaissance aux perturbations présentes dans les enregistrements), les données issues de l'épreuve des émotions ont été soigneusement contrôlées. De plus, une segmentation des données en groupes de souffle (cf. Fig. 5.5) fut nécessaire, puisque la parole produite dans cette épreuve est de nature spontanée. Cette segmentation vise à définir un cadre uniforme pour l'analyse. Les perturbations rencontrées dans les enregistrements de cette deuxième épreuve ont été, de toute évidence, beaucoup plus nombreuses que dans la première épreuve puisque de nature contrainte. Toutefois, les types de perturbations rencontrées restent les mêmes : (i) faux-départs, (ii) hésitations, (iii) bruits issus de l'environnement et (iv) parole non liée à la tâche. Le prétraitement des données, que nous avons voulu (une nouvelle fois) manuel, pour minimiser les biais apportés par les perturbations dans le système de reconnaissance, n'a pas été une mince affaire. Il a fallu en effet prendre des décisions qui n'ont pas tou-

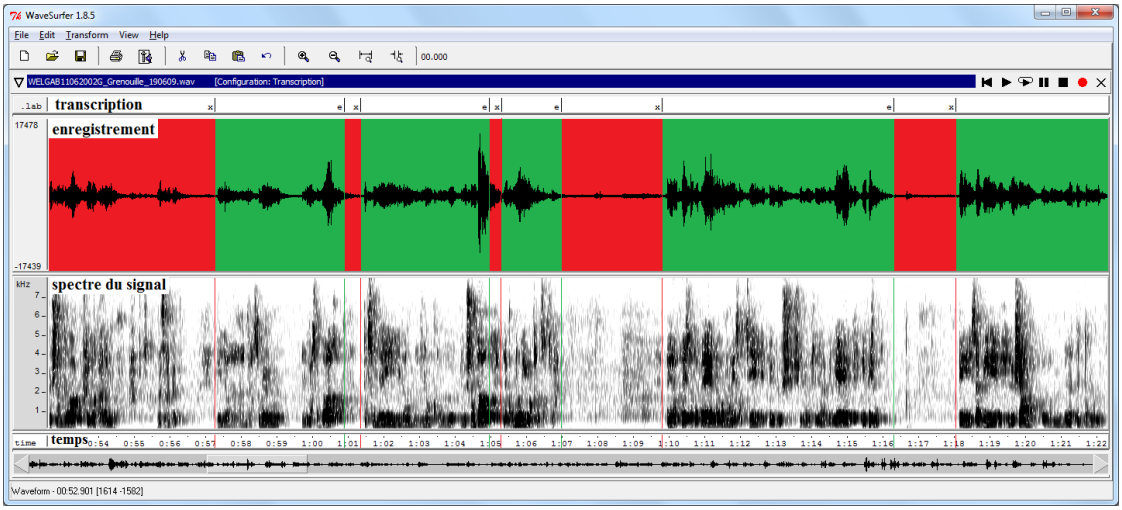


Fig. 5.5 Segmentation des enregistrements en groupes de souffle à l'aide du logiciel *Wavesurfer*[51] ; en vert les instants de parole correspondant à la production du sujet enregistré ; en rouge les autres instants non retenus par la segmentation.

Table 5.6 Quantité de groupes de souffle disponibles pour l'analyse de la tâche de *production de parole spontanée affective*.

Valence	DT	TA	TED-NOS	TSL
Positive	597	99	118	184
Neutre	926	151	126	238
Négative	2050	339	283	535
Ambivalente	370	63	59	99
Toutes	**3943**	**652**	**586**	**1048**

DT : sujets contrôles ; TA : sujets autistes ; TED-NOS : sujets dysharmoniques ; et TSL : sujets dysphasiques.

jours été évidentes quant au rejet ou non des données bruitées issues des enregistrements (plus de 10h de données à traiter). Trois passes d'écoute et de segmentation ont été nécessaires pour s'assurer de l'homogénéité de la segmentation des données en groupes de souffle, cf. Fig. 5.5. Après la phase de segmentation, la durée totale des données retenues représente 4h30min de parole, sur un total de plus de 10h d'enregistrement (facteur de réduction des données égale à ½). La table 5.6 donne le nombre de groupes de souffle qui ont été segmentés dans les enregistrements selon les sujets et les catégories d'émotions.

2.5. Passation des épreuves

Les passations des épreuves ont été effectuées par une étudiante en orthophonie [DEM-08][55]. Le matériel qui a été utilisé pour effectuer les passations est le suivant : (i) un ordinateur portable standard, (ii) un microphone *Logitech USB Desktop* de qualité correcte et (iii) un logiciel informatique *Audacity* pour contrôler une éventuelle saturation du micro. Chaque pas-

[55] J. Demouy, "*Caractéristiques prosodiques des enfants et adolescents autistes, dysharmoniques, dysphasiques et sans pathologie*", Mémoire de l'école d'orthophonie de la Pitié-Salpêtrière, Université Pierre et Marie Curie, Paris 6, 2009.

sation s'est déroulée de la même manière : les 3 groupes pathologiques ont passé la batterie de tests ELO et les deux épreuves que nous venons de présenter. Les sujets ont toujours été placés sur la gauche de l'examinateur et face à l'écran. Les épreuves ont à chaque fois été proposées dans le même ordre : un récit de l'histoire en image (épreuve de *production spontanée de parole affective*) et une répétition des phrases présentant des profils intonatifs variés (épreuve *d'imitation de contours intonatifs*). Les sujets ont été informés à l'avance du nombre d'épreuves qui allait leur être proposé, et le visionnage d'un petit dessin animé[56] en fin de protocole leur était promis en guise de récompense. Un bureau calme a été mis à disposition dans les 2 services de pédopsychiatrie de façon à ce que l'acquisition des données puisse être effectuée dans un environnement peu bruyant. Une salle au calme a également été réservée pour les enregistrements au lycée, au collège et en école primaire à l'Ermitage pour l'enregistrement des sujets contrôles.

3. Reconnaissance automatique de l'intonation

Cette section présente les méthodes et les résultats qui ont été obtenus sur l'épreuve *d'imitation des contours intonatifs*, cf. sous-section 2.4.1. L'analyse des caractéristiques prosodiques repose sur l'emploi de systèmes de reconnaissance statique et dynamique de la prosodie, cf. chapitre 4, sous-section 2.3. Les performances des sujets dans la tâche ont été comparées au moyen d'une fusion des informations issues des systèmes de reconnaissance statique et dynamique de l'intonation, comme l'a suggéré [SAN09][33]. Avant de décrire ces systèmes et les stratégies qui ont été utilisées dans notre étude, nous présentons tout d'abord un bref état-de-l'art dans le domaine de la reconnaissance automatique de l'intonation.

3.1. Etat de l'art

Les techniques qui sont utilisées pour caractériser l'intonation devraient *a priori* reposer uniquement sur des caractéristiques du pitch, puisque les catégories intonatives à identifier sont définies par des formes spécifiques au contour intonatif. Toutefois, les systèmes issus de la littérature ont montré que l'inclusion d'autres types d'informations telles que l'énergie et le rythme est nécessaire pour atteindre des scores de reconnaissance convenables [ANA08][57], [ROS09][58]. En l'occurrence, la détection du mamanais, qui est un langage spécifiquement caractérisé par des valeurs élevées et des variations prononcées du pitch, nécessite d'autres caractéristiques que celles provenant du pitch pour atteindre des scores de reconnaissance satisfaisants [MAH08][59].

[56] Promotion publicitaire du film « *L'âge de glace 3* » en libre accès sur Internet.

[57] S. Ananthakrishnan et S. Narayanan, "Fine-grained pitch accent and boundary tones labeling with parametric f0 features", dans proc. *ICASSP*, Las Vegas (NV), USA, Mar. 30 - Apr. 4, 2008, pp. 4545–4548.

[58] A. Rosenberg et J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level", dans proc. *Human Lang. Tech.: The 2009 Ann. C. of the North Amer. Chapter of the Assoc. for Comp. Ling.*, Boulder (CO), USA, May 31 - Jun. 5, 2009, pp. 81–84.

[59] A. Mahdhaoui, M. Chetouani et C. Zong, "Motherese detection based on segmental and supra-segmental features", dans proc. *ICPR*, Tampa (FL), Dec. 8-11 2008, pp. 1–4.

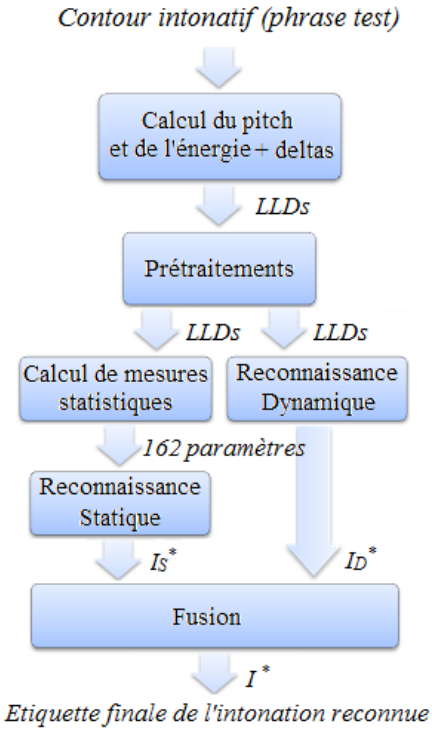


Fig. 5.6 Schéma du système de reconnaissance de l'intonation.

Ananthakrishnan *et al.* [ANA08][57] ont proposé un système qui utilise des caractéristiques issues du modèle *connexion-monte-tombe* du pitch (RFC [TAY94][60]) avec un classifieur prosodique de type *n-grams* pour étiqueter 4 types d'accents du pitch. Un score de reconnaissance de 56% a été atteint par ce système sur le corpus *Boston University Radio News Corpus* (BURNC) qui comprend 3h de parole de qualité radiophonique lue par 6 adultes. Rosenberg *et al.* ont, quant à eux, comparé le pouvoir discriminant de différents points d'ancrage de la parole, e.g., les voyelles, les syllabes et les mots dans l'analyse des indicateurs acoustiques de l'accent du pitch [ROS09][58]. Les mesures prosodiques ont été effectuées à travers des LLD qui ont été caractérisés par des modèles de régression logistique. Les ancrages linguistiques reposants sur des mots ont fourni un score de 83% sur le corpus BURNC. Dans un système proposé par Szaszák et al. [SZA09][61], un classificateur basé sur des chaînes de Markov cachées (HMM), a été conçu dans le but d'évaluer les performances en production de l'intonation dans une tâche d'apprentissage de langue, chez des sujets atteints de troubles de l'audition. Ce système a été utilisé pour catégoriser cinq classes associées à différents types de contours intonatifs. Les résultats ont été comparés à ceux fournis par des tests subjectifs. La classification automatique a fourni un taux de reconnaissance de 52% et l'évaluation subjective a produit un score de 69%. Le système de reconnaissance dynamique de la prosodie, i.e., basé sur des modèles HMM, a été réutilisé dans cette étude.

[60] P. Taylor, The Rise/Fall/Connection model of intonation, dans *Speech Communication*, vol. 15, no. 1-2, pp. 169-186, Oct. 1994.

[61] G. Szaszák, D. Sztahó et K. Vicsi, "Automatic intonation classification for speech training systems", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 1899–1902.

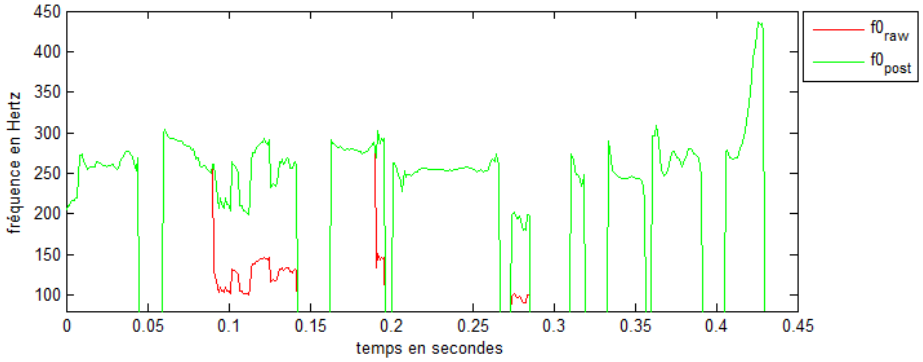


Fig. 5.7 Prétraitement de la f0 par un filtre de type anti saut d'octave (cf. annexe 4) sur une phrase reproduite par un enfant ; en rouge, les valeurs de la fréquence fondamentale avant traitement ($f0_{raw}$), en vert, après traitement ($f0_{post}$).

3.2. Système de reconnaissance du contour intonatif

La chaîne de traitements que nous proposons pour le contour intonatif inclut les étapes d'extraction d'informations prosodiques et de classification, cf. Fig. 5.6. Comme les contours de notre étude ont été fournis par l'imitation de phrases préenregistrées, la phrase a été utilisée comme unité de traitement par le système. Cette unité correspond à l'instant où un enfant imite une phrase. Ainsi, notre étude ne repose pas sur de la parole spontanée ou lue, mais sur de la parole contrainte dans laquelle un certain degré de spontanéité peut être trouvé selon l'enfant.

L'étape d'extraction de caractéristiques repose sur des LLDs du pitch et de l'énergie qui ont été estimés par l'algorithme *Snack* [SJO00][51]. La fréquence fondamentale a été calculée par la méthode ESPS toutes les 10ms. Les étapes de prétraitement comprenaient un filtre anti saut d'octave pour réduire les erreurs d'estimation de la f0 sur les voix d'enfants, cf. Fig. 5.7. Après cette étape de filtrage, les valeurs de la f0 ont été linéairement extrapolées à travers les segments non voisés (durée inférieure ou égale à 250ms, valeur empirique). Un filtre moyenneur et travaillant sur 11 échantillons (i.e., 110ms) a ensuite été appliqué sur la f0 et les valeurs d'énergie pour lisser leurs variations abruptes. Nous avons aussi normalisé ces valeurs pour réduire les variabilités produites par les locuteurs et les conditions d'enregistrement : les valeurs de la f0 ont été divisées par la valeur moyenne de toutes les trames voisées, et l'énergie a été normalisée à 0dB. Enfin, les dérivées du premier ordre et du second ordre (Δ et $\Delta\Delta$) ont été calculées sur les caractéristiques du pitch et de l'énergie, de sorte que six LLDs prosodiques soient extraits en tout sur chaque phrase dans le système de reconnaissance.

Comme les correspondances entre le type de la phrase et la prosodie sont spécifiques à la langue, l'intonation elle-même a été traitée par les systèmes de reconnaissance. Cette dernière a été reconnue séparément par les approches statique et dynamique, cf. Fig. 5.6. Notons que l'approche statique nécessite le calcul de mesures statistiques sur les LLDs prosodiques alors que l'approche dynamique est optimisée pour les traiter directement. Comme ces deux approches emploient des stratégies différentes pour reconnaitre l'intonation, nous avons supposé qu'elles pourraient fournir des résultats complémentaires. Par conséquent, les probabilités issues de chaque système de reconnaissance ont été fusionnées pour obtenir le profil intonatif final reconnu par le système. Enfin, nous avons exploité une méthode d'exploration de don-

nées reposant sur une CVS pour effectuer les expériences, ce qui permet de réduire l'influence du partitionnement des données lors des phases d'apprentissage et de test, tout en s'assurant que les intonations faiblement représentées ne soient pas désavantagées [DUD00][62], cf. chapitre 3, sous-section 3.5.

3.2.1. Classification statique du contour intonatif

Ce système correspond à celui de l'état-de-l'art en prenant une décision sur une phrase à l'aide de mesures statistiques des LLDs (pitch, énergie et leurs dérivées Δ et ΔΔ) concaténées dans un super-vecteur. Nous avons utilisé un ensemble de 27 mesures statistiques pour caractériser l'intonation à travers les LLDs, cf. table 4.3. Le pouvoir discriminant de ces paramètres a été évalué par l'algorithme RELIEF-F [ROB03][63]. Cet algorithme repose sur le calcul des entropies *a priori* et *a posteriori* des paramètres prosodiques selon les contours intonatifs. Il a été utilisé pour initialiser une méthode de sélection de caractéristiques de type montante (i.e., *bottom-up*) pour l'étape de classification : les caractéristiques prosodiques ont été successivement insérées dans le super-vecteur selon leur pouvoir discriminant estimé par l'algorithme RELIEF-F, et nous n'avons retenu que celles qui ont créé une amélioration dans les performances en classification. Cette méthode permet d'identifier les paramètres les plus pertinents pour décrire l'intonation.

L'algorithme des *k*-plus-proches-voisins (cf. chapitre 3, sous-section 3.3.1) a été utilisé pour effectuer la classification des caractéristiques ; la valeur de *k* vaut trois. La méthode des *k*-ppv estime par le maximum de vraisemblance les probabilités *a posteriori* de reconnaître un contour intonatif I_n (n = 1, 2, ... N_i classes intonatives) sur une phrase testée S. Ce calcul nécessite d'identifier les classes k_n issues des données d'apprentissage qui contiennent les caractéristiques prosodiques les plus proches de celles extraites sur la phrase testée S. L'intonation reconnue par la modélisation statique I_S^* est issue d'une fonction arg max sur les estimations des probabilités *a posteriori* $p_{stat}(I_n \mid S)$ [41].

$$p_{stat}(I_n \mid S) = \frac{k_n}{k}$$
$$I_S^* = \underset{1 \le n \le N_i}{\arg\max}\left(p_{stat}(I_n \mid S)\right) \qquad [41]$$

3.2.2. Classification dynamique du contour intonatif

La reconnaissance dynamique utilise des chaînes de Markov cachées (HMM) qui traitent directement les LLDs prosodiques pour effectuer la classification du profil intonatif. Les paramètres de position et de durée d'unités intonatives représentant alors différents états du contour prosodique sont pour cela exploités, cf. Fig. 5.8. Les distributions statistiques des LLDs ont été estimées par des MMG, cf. chapitre 3, sous-section 3.3.2, et les vecteurs

[62] R. O. Duda, P. E. Hart et D. G. Stork, *Pattern classification*. 2nd Ed. New York: Wiley, 2000.
[63] M. Robnik et I. Konenko, "Theoretical and empirical analysis of ReliefF and RReliefF", dans *Mach. Learn. J.*, vol. 53, pp. 23–69, Oct.-Nov. 2003.

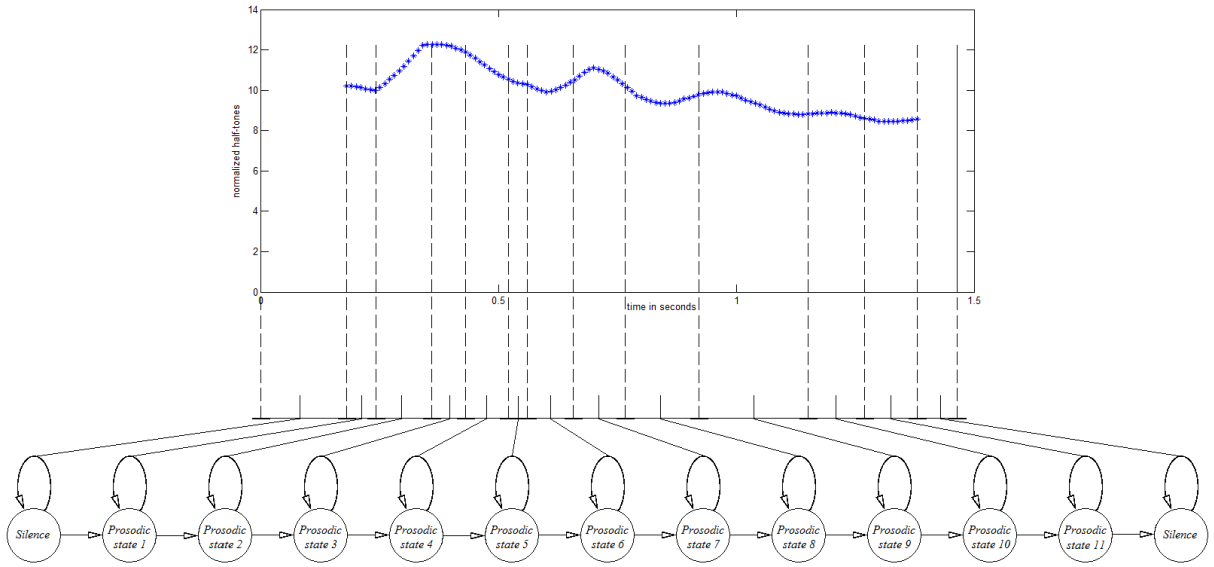


Fig. 5.8 Principe de la modélisation HMM du contour intonatif du pitch extrait sur une phrase.

$$I_D^* = \arg\max_{1 \le n \le N_i} \left(p_{dyn}(I_n \mid S) \right) \quad [42]$$

$$p_{dyn}(I_n \mid S) = \frac{p(S \mid I_n) * p(I_n)}{p(S)} \quad [43]$$

d'observation ont été de dimension 6, i.e., égal au nombre de LLDs utilisés. Puisque la durée des phrases varie selon les contours intonatifs (cf. table 5.3), un nombre fixe ou variable d'états selon la durée des phrases a été utilisé pour configurer le système de reconnaissance dynamique. Un nombre fixe de 11 états modélisés par un MMG composé de 8 gaussiennes a produit les meilleures performances en reconnaissance de l'intonation sur un corpus de parole Hongroise [SZA09][61] (optimisation empirique des paramètres). Comme les intonations de notre étude sont identiques à celles de [SZA09], la même configuration a pu être exploitée pour le français. L'intonation reconnue par la classification dynamique a été obtenue par une fonction arg max sur les estimations des probabilités *a posteriori* $p_{dyn}(I_n \mid S)$ [42]. L'estimation de $p_{dyn}(I_n \mid S)$ a été décomposée de la même manière que dans la reconnaissance de la parole, i.e., selon la règle de Bayes : $p(S \mid I_n)$ spécifie la probabilité des observations prosodiques extraites sur la phrase testée S, $p(I_n)$ est la probabilité associée à l'intonation, et $p(S)$ celle associée à la phrase [43].

3.2.3. Fusion des classifieurs

Puisque les classifieurs statiques et dynamiques fournissent des informations différentes en utilisant des processus distincts pour caractériser l'intonation, leur combinaison devrait améliorer les performances en reconnaissance. Bien que de nombreuses techniques sophistiquées existent pour fusioner les informations, e.g., [KUN03][64] et [MON09][65], nous avons utilisé une somme pondérée par un coefficient statique α ($0 \le \alpha \le 1$) des probabilités

$p_{stat}(I_n \mid S)$ et $p_{dyn}(I_n \mid S)$ [44]. Cette approche permet d'estimer la contribution directe des informations dans la reconnaissance de l'intonation, comme cela a été recommandé par [SAN09][33]. Les poids attribués aux informations lors de l'étape de fusion ont été estimés lors de la phase de test en reconnaissance de l'intonation. Ces poids reflètent ainsi la contribution directe des systèmes statique et dynamique, et non pas celle de l'apprentissage (indirecte). Afin d'évaluer la similarité entre ces deux classifieurs, nous avons calculé la statistique Q [YIL09][66] [45]. Cette mesure prend des valeurs comprises entre [-1 ; +1] et plus les valeurs sont proches de 0, plus les classifeurs sont vus comme dissemblables. Par exemple, $Q_{stat,dyn} = 0$ correspond à une dissimilitude totale entre les deux classifieurs, i.e., ils retournent à chaque fois la même classe associée à une phrase test. La statistique Q a été utilisée par Yildirim *et al.* pour évaluer la complémentarité des informations audiovisuelles dans la détection de la disfluence[67] sur la parole spontanée d'un enfant [YIL09].

$$I^* = \arg\max_{1 \le n \le N_i} \left(\alpha * p_{stat}(I_n \mid S) + (1 - \alpha) * p_{dyn}(I_n \mid S) \right) \quad [44]$$

$$Q_{stat,dyn} = \frac{N^{00}N^{11} - N^{01}N^{10}}{N^{00}N^{11} + N^{01}N^{10}} \quad [45]$$

avec, N^{00} le nombre de fois où les deux classifieurs donnent une réponse erronée, N^{11} le nombre de fois où les deux classifieurs donnent une réponse correcte, N^{01} le nombre de fois où le premier classifieur donnent une réponse correcte et le second une réponse fausse, et N^{01} le nombre de fois où le premier classifieur donnent une réponse fausse et le second une réponse correcte.

3.2.4. Stratégies de reconnaissance

Les systèmes ont tout d'abord été utilisés sur les données du groupe contrôle pour définir les scores « cibles » en reconnaissance des contours intonatifs. Les phrases qui ont été produites par les enfants à DT ont été testées dans le cadre d'une CVS et les probabilités *a posteriori* retournées par les classifieurs statique et dynamique ont été fusionnées comme indiqué dans [44]. Les compétences prosodiques des sujets pathologiques en reproduction de l'intonation ont ensuite été analysées en testant leur contours intonatifs alors que ceux produits par les typiques ont été appris par le système de reconnaissance, cf. Fig. 5.9. Le schéma de reconnaissance employé pour caractériser l'intonation des DT a ainsi été croisé avec celui des pathologiques lors des tests effectués sur les phrases produites par ces derniers. Les 10

[64] L. Kuncheva et C. Whitaker, "Measure of diversity in classifier ensembles", dans *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003.

[65] E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy et J. Sole-Casals, "Maximum likelihood linear programming data fusion for speaker recognition", dans *Speech Comm.*, vol. 51, no. 9, pp. 820–830, Sep. 2009.

[66] S. Yildirim et S. Narayanan, "Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information", dans *IEEE Trans. on Audio Speech and Lang. Proc.*, vol. 17, no. 1, pp. 2–12, Jan. 2009.

[67] Le terme de disfluence, employé pour décrire les trébuchements de la parole dans les énoncés de français parlé standard est un anglicisme du terme "*speech disfluency*". Le terme de diffluence désigne quant à lui les troubles du discours *et* de la pensée.

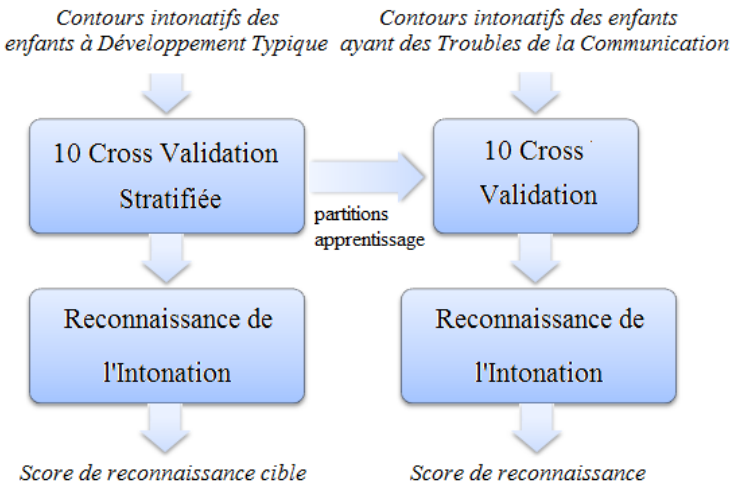


Fig. 5.9 Stratégies de reconnaissance du contour intonatif.

partitions de test des données des sujets TC ont donc été traitées 10 fois, i.e., avec chaque partition d'apprentissage des DT. L'ensemble des paramètres discriminants qui ont été identifiés (approche statique) par la méthode de reconnaissance *bottom-up* sur les DT a été utilisé pour caractériser les contours produits par les TC. Le poids optimal pour la fusion des classifieurs a, quant à lui, été estimé pour chaque groupe, i.e., DT, TA, TED-NOS et TSL. Cela permet de faire ressortir d'éventuelles différences entre les groupes dans la contribution des deux systèmes de reconnaissance du contour intonatif : statique et dynamique .

3.3. Résultats expérimentaux

Les analyses effectuées sur l'épreuve *d'imitation des contours intonatifs* ont été divisées en deux étapes : (i) une analyse statistique de la durée des phrases et (ii) une utilisation des systèmes de reconnaissance qui ont été décrits dans les paragraphes précédents. Les scores de reconnaissance obtenus par les enfants à DT sont considérés comme des valeurs cibles pour les sujets atteints de TC. Notons que la stratégie de reconnaissance proposée exploite les caractéristiques des sujets DT pour reconnaître l'intonation des sujets pathologiques, cf. Fig. 5.9. En d'autres termes, le biais introduit par les enfants à DT dans la tâche d'imitation a été inclus dans la configuration du système de reconnaissance. Tout écart significatif par rapport à ce biais sera considéré dans cette étude comme lié à une déficience dans les compétences prosodiques *grammaticales* des sujets étudiés, ou du moins, à une carence dans les capacités à imiter un contour intonatif. Notons que la stratégie de reconnaissance employée montre qu'un apprentissage des modèles sur les données des enfants à DT influence, *a priori*, les scores de reconnaissance sur les intonations produites par les sujets atteints de TC ; comparé notamment à un apprentissage des modèles qui aurait été effectué sur leurs propres données. Cependant, et de façon *a posteriori*, cette configuration n'a pas montré de réelles différences dans les performances en reconnaissance de l'intonation comparé à des modèles appris sur les données des sujets à DT.

Une méthode non-paramétrique a été utilisée pour effectuer la comparaison statistique des données entre les groupes d'enfants, i.e., une *p*-valeur a été estimée par la méthode de

Table 5.7 Mesures statistiques de la durée des phrases qui ont été reproduites par les sujets à DT.

Intonation	REF	DT
Descendante	1.7 $_{0.3}$	1.7 $_{0.6}$
Tombante	1.2 $_{0.3}$	1.3 $_{1.4}$
Flottante	1.6 $_{0.2}$	1.6 $_{0.4}$
Montante	0.7 $_{0.2}$	0.5 $_{0.2}$

[Moyenne] $_{(\text{écart-type})}$; REF : phrases de référence (stimuli) ; et DT : développement typique.

Kruskal-Wallis. Cette méthode présente l'avantage de ne pas faire d'hypothèse quant à la distribution statistique des données comparées. La *p*-valeur fournie par ce test statistique correspond à la probabilité que les données comparées soient issues de la même population ; $p < 0.05$ est souvent utilisée comme une hypothèse alternative pour laquelle il y a moins de 5% de chance que les données soient issues d'une population identique.

3.3.1. Sujets à développement typique

Durée de la phrase

Les résultats montrent que les patrons temporels des phrases ont été conservés pour tous les groupes intonatifs lorsque les phrases ont été reproduites par les sujets à DT (p=0.9), cf. table 5.7. En conséquent, les imitations des contours intonatifs par les sujets typiques ont été effectuées sur le même modèle temporel que celui des stimulus.

Reconnaissance de l'intonation

Les scores de reconnaissance des contours intonatifs des enfants à DT sont donnés dans la table 5.8. A titre de comparaison, nous avons calculé la performance d'un classifieur naïf, qui attribue toujours l'étiquette de l'intonation la plus représentée (e.g., *Descendante*) à une phrase donnée ; le score chance vaut 31%. Les statistiques Q (cf. sous-section 3.2.3) ont été calculées pour évaluer la similarité entre les classificateurs lors de la tâche. Le système produit un taux de reconnaissance moyen de 70%, soit plus du double du score issu de la chance, pour les 73 sujets à DT âgés de 6 à 18 ans. Ce score est égal à la valeur moyenne de ceux qui ont été obtenus par d'autres auteurs sur le même type de tâche (reconnaissance du contour intonatif), mais sur de la parole adulte et pour 6 locuteurs [ANA08][57], [ROS09][58]. L'effet de l'âge sur les systèmes de TAP a été montré comme étant un sérieux facteur de perturbation, notamment lorsque les données incluent des voix d'enfants [ELE04][68].

La statistique Q montre que les systèmes de reconnaissance statique et dynamique apparaissent comme étant similaires pour l'intonation *Flottante* (Q = 0.6754), même si le score obtenu par l'approche dynamique est nettement supérieure à celui de la méthode statique, cf. table 5.8. Notons que certains contours ont été, soit mieux reconnus par l'approche statique, soit par la méthode dynamique, et que la fusion a permis d'en améliorer les performances. La complémentarité des deux approches employées pour reconnaître l'intonation est donc réelle, même si leur fusion montre que le meilleur score est plutôt porté par l'approche statique, cf. Fig. 5.10. Bien que les deux intonations *Montante* et *Flottante* aient été très bien reconnues par le système, les intonations *Descendante* et *Tombante* ont fourni des

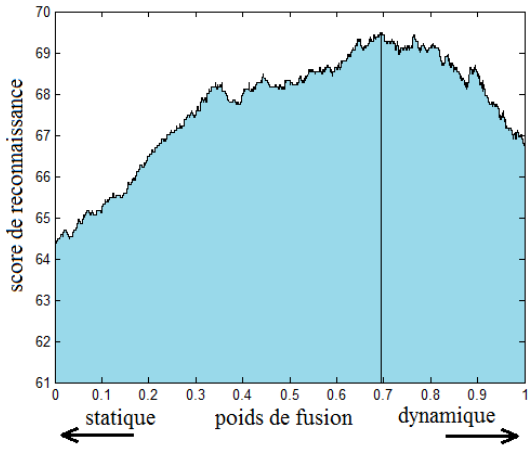


Fig. 5.10 Scores de reconnaissance des sujets à DT en fonction du poids de fusion des classifieurs.

Table 5.8 Performances en reconnaissance de l'intonation basée sur une modélisation statique, dynamique, et sur la fusion des deux pour les sujets à DT.

Intonation	Statique	Dynamique	Fusion	$Q_{stat,\ dyn}$
Descendante	61	55	64	0.1688
Tombante	55	48	55	0.3830
Flottante	49	71	72	0.6754
Montante	93	95	95	0.2716
Toutes	**67**	**64**	**70**	**0.4166**

Table 5.9 Matrice de confusion en reconnaissance de l'intonation pour les sujets contrôles.

Intonations	Descendante	Tombante	Flottante	Montante	Score %
Descendante	**377**	58	151	2	**64**
Tombante	104	**320**	116	46	55
Flottante	50	33	**212**	0	**72**
Montante	2	16	4	**416**	**95**
					70

scores assez faibles. Le plus faible d'entre eux est alors obtenu sur l'intonation *Tombante*, qui est représentée par des phrases associées à de nombreuses modalités ambigües (e.g., question / ordre / conseils, etc.) par rapport aux autres contours intonatifs, cf. table 5.3.

Comme l'intonation *Flottante* a une tendance descendante, elle a été confondue avec les intonations *Descendante* et *Tombante* mais jamais avec celle *Montante*, cf. table 5.9. De plus, cette intonation apparaît comme étant très différentes des autres car elle a été très bien reconnue et seulement confondue avec l'intonation *Flottante*. Les mauvaises classifications sont nombreuses sur le groupe intonatif *Flottant* (score le plus faible), et principalement portées par les intonations *Descendante* et *Flottante*.

L'ensemble des caractéristiques prosodiques estimées comme pertinentes, i.e., issues de la méthode bottom-up utilisée pour la classification statique (cf. sous-section 3.2.1), est quasi exclusivement constitué des dérivés Δ et ΔΔ, cf. table 5.10. Les caractéristiques provenant du pitch sont plus nombreuses que celles de l'énergie, ce qui peut être dû au fait que nous avons

[68] D. Elenius et M. Blomberg, "Comparing speech recognition for adults and children", dans proc. *FONETIK*, Stockholm, Sweden, May 26-28 2004, pp. 105–108.

Table 5.10 Ensemble de caractéristiques prosodiques pertinentes identifiées par la reconnaissance statique de l'intonation produite par les sujets à DT.

Pitch	Energie
B – RPmax	Δ – IQR
Δ – Q1	Δ – Shimmer
Δ – Q3	Δ – Slope
Δ – Jitter	Δ – TaV
Δ – Slope	Δ – TaVOnV_AD
Δ – OfVTaV_AD	ΔΔ – RPmax
ΔΔ – RPmin	ΔΔ – RPmin
ΔΔ – RP_AD	ΔΔ – Q3
ΔΔ – STD	ΔΔ – OnV
ΔΔ – Q1	ΔΔ – TaV
ΔΔ – Median	ΔΔ – OfVOnV_AD
ΔΔ – Q3	
ΔΔ – IQR	
ΔΔ – Jitter	
ΔΔ – OnV	
ΔΔ – OfVOnV	

B : donnée brute, i.e., pas de calcul de dérivé ; Δ : dérivée du 1er ordre ; et ΔΔ : dérivée du 2nd ordre (Δ et ΔΔ sont tous les deux des descripteurs dynamiques).

porté exclusivement notre attention sur le contour du pitch lors des enregistrements des phrases, cf. sous-section 2.4.1. Environ la moitié de l'ensemble des caractéristiques jugées comme pertinentes pour la reconnaissance de l'intonation comprennent des mesures issues des systèmes de détection de question, i.e., les valeurs ou les différences entre les valeurs en *début* / *milieu* / *fin* de phrase et les positions relatives des extrema dans la phrase. Les autres paramètres sont composés de mesures statistiques traditionnelles de la prosodie telles que les valeurs de quartiles, du coefficient de régression et d'écart-type.

3.3.2. Sujets pathologiques

Durée de la phrase

Tous les intonations qui ont été reproduites par les sujets TC, apparaissent comme très différentes de celles des enfants à DT lorsque l'on compare la durée des phrase (p<0.05) : la durée a été allongée de 30% pour les trois premières intonations et de plus de 60% pour le contour *Montant*, cf. table 5.11. De plus, le groupe composé d'enfants atteints de TSL a produit, de façon significative, des phrases beaucoup plus longues que tous les autres groupes d'enfants, sauf pour l'intonation *Montante*.

Reconnaissance de l'intonation

Les scores de reconnaissance des sujets TC ont été très proches de ceux des typiques et similaires entre les groupes de TC pour l'intonation *Descendante*, alors que toutes les autres intonations ont été significativement différentes (p<0.05) entre les enfants DT et TC, cf. table 5.12. Toutefois, le système a produit de très bons taux de reconnaissance sur l'intonation

Table 5.11 Mesures statistiques de la durée des phrases selon les groupes.

Intonation	REF	DT	TA	TED-NOS	TSL
Descendante	1.7 $_{0.3}$	1.7 $_{0.6}$	2.2 $_{0.9}^{*T,L}$	2.2 $_{0.8}^{*T,L}$	2.4 $_{0.9}^{*T,A,N}$
Tombante	1.2 $_{0.3}$	1.3 $_{1.4}$	1.6 $_{0.6}^{*T,L}$	1.7 $_{0.8}^{*T,L}$	1.8 $_{0.8}^{*T,A,N}$
Flottante	1.6 $_{0.2}$	1.6 $_{0.4}$	2.1 $_{0.7}^{*T,L}$	2.1 $_{0.5}^{*T,L}$	2.4 $_{1.0}^{*T,A,N}$
Montante	0.7 $_{0.2}$	0.5 $_{0.2}$	0.9 $_{0.3}^{*T}$	0.9 $_{0.3}^{*T}$	0.8 $_{0.2}^{*T}$

[Moyenne] $_{(\text{écart-type})}$; * = p<0.05 : l'hypothèse alternative est vraie lorsque l'on compare les données entre les groupes d'enfants ; REF : phrases de référence (stimuli) ; DT (T) : développement typique ; TA (A) ; troubles autistiques ; TED-NOS (N) : troubles envahissants du développement non-spécifiés ; et TSL (L) : troubles spécifiques du langage.

Table 5.12 Performances en reconnaissance de l'intonation basée sur la fusion des classifieurs.

Intonation	DT	TA	TED-NOS	TSL
Descendante	64	64	**70**	63
Tombante	55	35^{*T}	45^{*T}	39^{*T}
Flottante	**72**	48^{*T}	40^{*T}	31^{*T}
Montante	**95**	57^{*T,L}	48^{*T,L}	**81**^{*T,A,N}
Toutes	**70**	56^{*T}	53^{*T}	58^{*T}

* = p<0.05 : l'hypothèse alternative est vraie lorsque l'on compare les données entre les groupes d'enfants ; DT (T) : développement typique ; TA (A) ; troubles autistiques ; TED-NOS (N) : troubles envahissants du développement non-spécifiés ; et TSL (L) : troubles spécifiques du langage.

Table 5.13 Statistique Q entre les classifieurs statique et dynamique selon les groupes.

Mesure	DT	TA	TED-NOS	TSL
$Q_{stat,dyn}$	0.4166	0.6539	0.4521	0.5542

DT : développement typique ; TA ; troubles autistiques ; TED-NOS : troubles envahissants du développement non-spécifiés ; et TSL : troubles spécifiques du langage.

Montante pour les enfants TSL et DT alors que les résultats obtenus sur les deux autres groupes de TC, i.e., TA et TED-NOS ont été significativement moins bons (p<0.05).

Les résultats montrent que les contributions des deux classifieurs sont similaires pour tous les groupes pathologiques, mais opposées à celles obtenues pour les enfants à DT : statique, $\alpha = 0.1$, cf. Fig. 5.11. Les valeurs de la statistique Q entre les classifieurs statique et dynamique ont été plus élevées pour les TC que les enfants à DT, ce qui montre que ces méthodes ont été moins dissemblables que sur les enfants à DT, cf. table 5.13. Les erreurs de jugements rendus par le système de reconnaissance sur les sujets TC ont été à peu près similaires à celles observées sur les DT, cf. table 5.14-16. Pour tous les TC, l'intonation *Flottante* a été exclusivement confondue avec celles *Descendante* et *Tombante*.

3.4. Discussion des résultats

Cette première expérience consistait à utiliser des techniques de traitement du signal et de reconnaissance des formes pour comparer les capacités prosodiques de sujets atteints de TC avec celles d'enfants à DT dans une tâche d'imitation de l'intonation. Une série de 26 phrases comprenant des déclarations et des questions (cf. table 5.3) selon quatre types de profil inton-

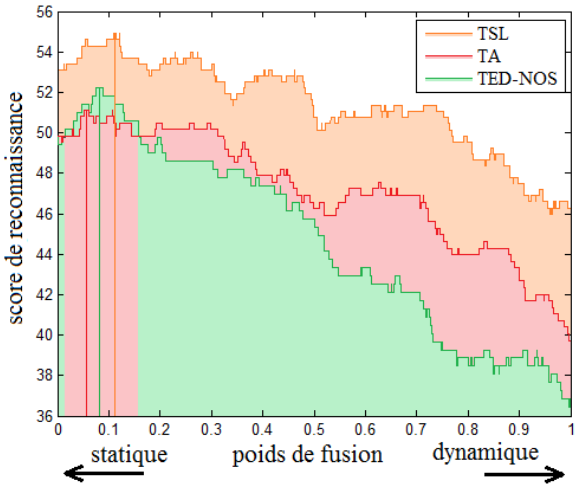


Fig. 5.11 Scores de reconnaissance des sujets TC en fonction du poids de fusion des classifieurs.

Table 5.14 Matrice de confusion en reconnaissance de l'intonation pour les sujets TA.

Intonations	Descendante	Tombante	Flottante	Montante	Score %
Descendante	**61**	14	20	0	**64**
Tombante	**39**	33	20	2	35
Flottante	16	9	**23**	0	48
Montante	5	23	2	**40**	57
					56

Table 5.15 Matrice de confusion en reconnaissance de l'intonation pour les sujets TED-NOS.

Intonations	Descendante	Tombante	Flottante	Montante	Score %
Descendante	**50**	5	16	0	**70**
Tombante	29	**34**	13	0	45
Flottante	**18**	8	16	0	40
Montante	8	19	4	**29**	48
					53

Table 5.16 Matrice de confusion en reconnaissance de l'intonation pour les sujets TSL.

Intonations	Descendante	Tombante	Flottante	Montante	Score %
Descendante	**65**	22	15	1	**63**
Tombante	**47**	41	16	0	39
Flottante	**20**	16	16	0	31
Montante	3	10	2	**63**	**81**
					58

atif (cf. Fig. 5.3) a été utilisée pour la tâche d'imitation de l'intonation. Nous avons ensuite recueilli manuellement 2 772 phrases à partir des enregistrements des enfants. Enfin, deux systèmes ont été fusionnées pour reconnaître l'intonation : reconnaissance statique (mesures statistiques puis k-ppv) et dynamique (modèles HMM).

Le système a bien fonctionné pour les enfants à DT sauf dans le cas de l'intonation *Tombante* qui présente un nombre trop élevé de modalités ambigües : *question* / *ordre* / *conseil*, etc. L'approche de reconnaissance statique a fourni une liste de 27 caractéristiques qui sont quasi exclusivement représentées par des descr-ipteurs dynamiques : Δ et ΔΔ.

Concernant les sujets présentant des TC (i.e., TA, TED-NOS et TSL), l'évaluation de leurs compétences prosodiques dans la tâche d'imitation de l'intonation a montré que la durée de leur phrase suffisait à les différencier significativement des TC : l'augmentation a été de 30% pour les trois premières intonations et de plus de 60% pour l'intonation *Montante*. Les discussions de ces résultats avec les cliniciens, ont permis d'aboutir à un lien avec l'hypothèse que les contours intonatifs montants sont plus difficiles à produire que les contours descendants chez les enfants atteints de TC [SNO98a][25]. De plus, un lien peut également être effectué avec les résultats issus des études qui ont montré l'existence de troubles dans la prosodie chez des sujets atteints de : (i) TSL ; [WEL03][18], [HAR89][27], [SAM03][28] et [MEU97][29], (ii) TA ; [FOS99][19], [MCC07][69] et [LEN08][70] et (iii) TED-NOS ; [PAU08][21], bien que certains résultats contradictoires aient été trouvés pour les sujets TSL dans [MAR09][27].

Le système de reconnaissance a montré que la meilleure approche de fusion a favorisé l'approche dynamique pour reconnaître l'intonation des TC, i.e., utilisation directe des LLDs par les HMM, alors que celle des DT a favorisé l'approche statique, i.e., mesures statistiques des LLDs combinées à un classifieur *k*-ppv. Par conséquent, les contours intonatifs n'ont pas été produits de la même manière par ces deux groupes : les sujets DT ont utilisé la dynamique du contour intonatif pour transmettre la modalité de la phrase, et les TC des particularités prosodiques spécifiques aux groupes intonatifs. Toutes les intonations ont été significativement différentes entre les enfants à DT et les TC au niveau des scores de reconnaissance ($p<0.05$), excepté pour le profil *Descendant*. De plus, les enfants atteints de TSL et les sujets typiques ont obtenus de très bons taux de reconnaissance pour l'intonation *Montante* alors que les résultats issus des deux groupes TA et TED-NOS ont été nettement moins bons. Les cliniciens considèrent ce résultat cohérent avec les études qui ont montré que les enfants atteints de TED ont plus de difficulté à imiter les questions que des énoncés [FOS99], ainsi que des éléments prosodiques courts et longs [MCC07], [PAU08]. De plus, comme la pragmatique est présente dans l'intonation *Montante*, il n'est pas si surprenant que de telles différences aient été trouvées dans les scores de reconnaissance entre les sujets TSL et les TEDs (i.e., TA et TED-NOS), puisque ces derniers présentent des déficits pragmatiques dans la communication, alors que la prosodie des TSL est épargnée de ces troubles. Notons que l'intonation *Montante* inclut des phrases très courtes par rapport aux autres groupes intonatifs (moitié de la durée), et que les TSL ne peuvent donc pas être désavantagés par rapport aux autres sujets [WEL03].

4. Caractérisation automatique de la valence affective

Cette section présente les méthodes et les résultats qui ont été obtenus sur l'épreuve de *production de parole affective spontanée* (cf. sous-section 2.4.2). Nous donnons tout d'abord un bref état-de-l'art dans le domaine de la reconnaissance automatique des émotions pour la parole produite par des enfants de façon spontanée.

[69] J. McCann, S. Peppé, F. Gibbon, A. O'hare, et M. Rutherford, "Prosody and its relationship to language in school-aged children with high-functioning autism", dans *Inter. J. of Lang. and Comm. Disorders*, vol. 42, no. 6, pp. 682–702, Nov.-Dec. 2007.

[70] M. T. Le Normand, S. Boushaba et A. Lacheret-Dujour, "Prosodic disturbances in autistic children speaking French", dans proc. *Speech Prosody*, Campinas, Brazil, May 6–9 2008, pp. 195–198.

4.1. Etat de l'art en reconnaissance automatique d'émotions

De très nombreux travaux ont montré que l'affect et la parole spontanée influencent les systèmes de TAP ; et d'autant plus, lorsque la parole est produite par des enfants [SCH08][71]. Concernant la tâche de reconnaissance automatique des émotions, très peu d'études se sont confrontées à la parole spontanée produite par des enfants, notamment si l'on regarde la quantité de travaux réalisés sur la parole actée. Ce constat peut s'expliquer par la difficulté liée à la tâche de collecte des données. Néanmoins, l'édition 2009 de la conférence *Interspeech* a vu apparaître la première campagne d'ampleur internationale portant sur la reconnaissance des émotions spontanées produites par des enfants [SCH09b][72], complétant ainsi les campagnes d'évaluations existantes pour la parole (NIST) et pour la musique (MIREX).

Le corpus qui a été étudié (FAU) est constitué d'émotions spontanées produites par des enfants dans le cadre d'interaction avec un jouet robotisé et commercialisé par Sony sous le nom d'AIBO, cf. Fig. 5.12. Le corpus FAU consiste en 9h de parole produite par 51 enfants Allemands âgés de 10 à 13 ans lors d'interactions spontanées avec le jouet robotisé AIBO. Les enregistrements ont été obtenus par le microphone présent à l'extrémité de la tête du jouet qui était ainsi dirigé vers l'enfant. Différents styles de production affectif ont ensuite été annotés sur chaque mot pour éviter les variations pouvant apparaître dans une même phrase (tâche non contrainte). 5 catégories affectives ont été retenues pour les tests : *Colère*, *Emphatique*, *Neutre*, *Positif* et *Autres*. Une configuration en deux classes selon la valence a également été testée : *Négative* et *Non-négative*. L'indépendance du locuteur a été assurée lors des tests en exploitant les données issues d'une école (OHM, 13 garçons et 13 filles) pour effectuer l'apprentissage des modèles et les données de l'autre école (MONT, 8 garçons et 17 filles) ont été utilisées pour réaliser la reconnaissance des émotions. Comme les instances ne sont pas équitablement distribuées à travers les classes, la balance a été réalisée par une technique d'exploration de données qui consiste à sur-échantillonner les données manquantes et sous-échantillonner celles en surnombre (méthode *synthetic minority over-sampling technique* – SMOTE). Cette technique a montré des performances supérieures à la technique de CVS sur différents types de données [CHA02][73].

Les LLDs prosodiques qui ont été extraits sur le corpus FAU concernent cinq types de composante acoustique : (i) *ZCR* : taux de passage par zéro, (ii) *RMS* : énergie, (iii) *F0* : fréquence fondamentale, (iv) *HNR* : rapport harmonique sur bruit et (v) *MFCC* : 12 coefficients cepstaux répartis selon l'échelle mel. Deux types de système de reconnaissance d'émotions ont été mis en œuvre pour effectuer le traitement des LLDs : (i) un système statique et (ii) un système dynamique. Les 16 LLDs prosodiques ont été caractérisés par un ensemble de 12 mesures statistiques pour l'approche statique : (i) 4 premiers moments, (ii) valeur et position relative des extrema dans les mots, (iii) amplitude et (iv) coefficients et erreur moyenne des valeurs en régression linéaire. La reconnaissance statique a été effectuée par un classifieur SVM (apprentissage par optimisation minimale séquentielle – méthode SMO, noyau linéaire

[71] B. Schuller, A. Batliner, S. Steidl et D. Seppi, "Does affect affect automatic recognition of children's speech?", dans proc. *1st W. on Child, Computer and Interaction*, Chaniá, Greece, Oct. 23 2008.

[72] B. Schuller, S. Steidl, et A. Batliner, "The Interspeech 2009 emotion challenge", dans proc. *Interspeech*, Brighton, United-Kingdom, Sep. 6-10 2009.

Fig. 5.12 Interaction entre des enfants et le jouet AIBO développé et commercialisé par Sony ; image extraite du site http://robotics.youngester.com/2009/10/robot-recession-in-japan.html.

Table 5.17 Scores (en %) obtenus en reconnaissance d'émotions spontanées produites par des enfants lors d'interactions avec le jouet AIBO (corpus FAU).

Test	Dyn. 1 E.	Dyn. 3 E.	Dyn. 5 E.	Stat.
2 classes	72	58	65	**73**
5 classes	51	35	37	**66**

et classification en 1 *vs.* 1), sur l'ensemble des 192 paramètres. La reconnaissance dynamique a été directement réalisée sur les LLDs par un HMM pour divers nombre d'états (1, 3, et 5), et un mélange de 2 Gaussiennes. Les résultats montrent que l'approche dynamique fournit les meilleurs résultats dans les deux configurations d'analyse, cf. table 5.17 ; c'est également le cas pour nos résultats en reconnaissance de l'intonation sur les sujets DT. Un score de 73% a pu être atteint dans la tâche de reconnaissance des deux catégories opposées de valence affective, i.e., *Négative* et *Non-négative* ; les 5 classes ont amenées un taux de reconnaissance de 66%.

Notons que d'autres travaux ont également été réalisés sur la détection automatique d'attitudes telles que la frustration, la politesse, et l'ennui dans la parole spontanée produite par des enfants [ANG02][74], [ARU01][75] et [YIL11][76].

4.2. Système de caractérisation de la valence affective

Cette sous-section présente le système qui a été utilisé pour étudier les données fournies par l'épreuve de *production de parole affective spontanée*, cf. sous-section 2.4.2. L'étude des caractéristiques prosodiques repose sur l'emploi d'un ensemble de descripteurs plus vaste que

[73] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", dans *J. of Artificial Intel. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[74] J. Ang, R. Dhillon, E. Schriberg et A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in Human-computer dialog", dans proc. *Interspeech, 7th ICSLP*, Denver (CO), USA, Sep. 16-20 2002, pp. 67–79.

[75] S. Arunachalam, D. Gould, E. Andersen, D. Byrd et S. Narayanan, "Politeness and frustration language in child-computer interactions", dans proc. *Eurospeech*, Aalborg, Denmark, Sep. 3-7 2001, pp. 2675–2678.

[76] S. Yildirim, S. Narayanan et A. Potamianos, "Detecting emotional state of a child in a conversational computer game", dans *Computer Speech and Lang.*, vol. 25, no. 1, pp. 29–44, Jan. 2011.

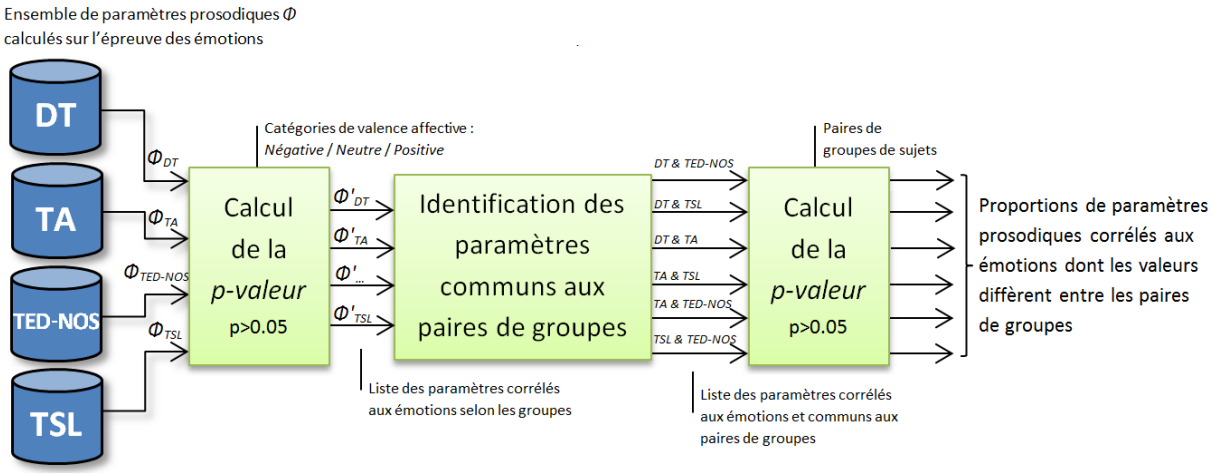


Fig. 5.13 Système d'analyse des paramètres prosodiques de l'épreuve de « *production de parole affective spontanée* ».

lors de la première épreuve, puisque l'affect est une information beaucoup plus complexe à identifier que les profils intonatifs. Les caractéristiques prosodiques qui ont été calculées couvrent les quatre composantes de la prosodie : (i) intonation, (ii) intensité, (iii) qualité vocale et (iv) rythme. L'identification des paramètres par lesquels les enfants peuvent communiquer la valence affective, a été effectuée au moyen d'une analyse statistique. Notre étude a consisté à comparer les groupes de sujets à travers leur ensemble de paramètres communs et discriminants des trois catégories de valence affective, cf. Fig. 5.13. Nous avons exploité pour cela différents points d'ancrages de la parole (cf. chapitre 2), ainsi que les paramètres prosodiques du chapitre 4 (cf. sous-sections 2.1 et 2.2), en particulier les modèles *conventionnels* et *non-conventionnels* du rythme. Nous avons également inclus dans l'analyse les paramètres liés à la proportion et la durée de production des groupes de souffle. Afin de tenir compte d'éventuels effets spécifiques au locuteur ou au genre, nous avons normalisé les paramètres selon chaque locuteur par le calcul du *z-score* (cf. chapitre 3, section 2). L'étude des émotions prototypiques a montré que cette méthode est appropriée, cf. chapitre 4, sous-sections 4.1, 4.2 et 4.3.

Après avoir vérifié que le taux de calculabilité des paramètres prosodiques était supérieur à 75%, nous avons estimé leur pouvoir discriminant selon les catégories de valence affective pour chaque groupe de sujets ; méthode non-paramétrique de Kruskal-Wallis. Un deuxième test statistique a ensuite été effectué entre les groupes de sujets, pour caractériser d'éventuelles différences sur les valeurs de paramètres discriminants et communs entre les paires de groupes, cf. Fig. 5.13. Les motivations sous-jacentes à ces analyses sont les suivantes : (i) est-ce que les enfants utilisent des paramètres prosodiques pour communiquer les corrélats associés à la valence affective des images (malgré l'hypothèse forte qui est faite) ? et, si oui, (ii) est-ce que les groupes de sujets utilisent des codes similaires pour communiquer les émotions ? Le système que nous proposons permet d'apporter des réponses à ces questions au moyen de techniques issues du TAP, telles que, l'identification automatique des ancrages acoustiques de la parole qui définit à quels instants calculer les paramètres, et l'extraction des caractéristiques telles que le pitch, l'énergie, la qualité vocale et le rythme. Les tests statistiques permettent ensuite d'identifier les corrélats prosodiques des émotions selon les groupes, et de comparer leurs codes à travers leur jeu de paramètres communs obtenus par la première analyse.

Table 5.18 Mesures statistiques de la proportion de groupes de souffle produits par les sujets selon la valence affective des images de l'histoire.

Valence	DT	TA	TED-NOS	TSL
Négative	27 $_{3}$	27 $_{4}$	24 $_{3}$	27 $_{5}$
Neutre	**30 $_{5}$**	**31 $_{12}$**	27 $_{6}$	**28 $_{7}$**
Positive	23 $_{5}$	22 $_{8}$	**30 $_{6}$**	27 $_{5}$
Ambivalente	20 $_{4}$	20 $_{4}$	19 $_{6}$	18 $_{5}$

[Moyenne] $_{(\text{écart-type})}$. DT : développement typique ; TA : troubles autistiques ; TED-NOS : troubles envahissants du développement non-spécifiés ; et TSL : troubles spécifiques du langage.

4.3. Résultats expérimentaux

Les résultats obtenus par l'analyse des données issues de l'épreuve de *production de parole affective spontanée* sont présentés à travers différents types de paramètres. Nous décrivons tout d'abord les résultats concernant la proportion en production de groupes de souffle selon les catégories affectives, et nous présentons ensuite ceux issus de la durée des phrases et des paramètres prosodiques.

4.3.1. Proportion des groupes de souffle

La table 5.18 présente les données associées à la proportion des groupes de souffle produits selon les catégories de la valence affective. Bien que nous présentons à titre indicatif, les données issues de la valence ambivalente, aucune analyse n'a été effectuée par la suite sur ces valeurs puisqu'elles correspondent à des catégories disparates d'émotions, cf. table 5.5. Les résultats de la table 5.18 montrent que tous les groupes de sujets produisent un nombre de groupes de souffle dont les proportions selon les catégories de la valence affective sont, de façon globale, équitablement réparties entre ces dernières et équivalentes entre les groupes.

4.3.2. Durée de production des groupes de souffle

Les valeurs de durée des groupes de souffle font apparaître de nombreuses différences significatives entre les groupes de sujets, cf. table 5.19. Contrairement à l'épreuve *d'imitation des contours intonatifs* dans laquelle les sujets TC ont produit des phrases dont la durée était significativement plus longue que celles des sujets à DT (cf. table 5.11), le cadre spontané de la deuxième épreuve amène tout le contraire : les durées des groupes de souffle produits par les TC sont toutes très inférieures ($p < 0.05$) à celles des DT et ces valeurs sont différentes ($p < 0.05$) entre les sujets TED-NOS et les TA sur les catégories opposées de valence. De plus, la durée de production des groupes de souffle varie significativement entre les catégories opposées de valence affective pour les sujets TA, et également avec la catégorie *Neutre* pour les TED-NOS, ce qui n'est pas le cas des DT et TSL.

4.3.3. Mesures prosodiques

Nous présentons dans les paragraphes suivants, les résultats issus des mesures proso-

Table 5.19 Mesures statistiques de la durée de production des groupes de souffle selon la valence affective des images de l'histoire.

Valence	DT	TA	TED-NOS	TSL
Négative	$2.8_{1.4}^{*A,N,L}$	$^{*+}2.1_{1.0}^{*T,N}$	$^{*+,=}2.3_{1.1}^{*T,A}$	$2.2_{1.1}^{*T}$
Neutre	$2.9_{1.5}^{*A,N,L}$	$2.3_{1.2}^{*T}$	$^{*-}2.5_{1.4}^{*T}$	$2.3_{1.2}^{*T}$
Positive	$2.8_{1.4}^{*A,N,L}$	$^{*-}2.4_{1.2}^{*T,N}$	$^{*-}2.1_{1.1}^{*T,A}$	$2.2_{1.3}^{*T}$

[Moyenne] $_{(\text{écart-type})}$; DT (T) : développement typique ; TA (A) : troubles autistiques ; TED-NOS (N) : troubles envahissants du développement non-spécifiés ; et TSL (L) : troubles spécifiques du langage. ; + : valence positive ; = : valence neutre ; - : valence négative ; * : $p < 0.05$.

diques, cf. chapitre 4, sous-sections 2.1 et 2.2. L'analyse de ces paramètres s'effectue en deux temps (cf. Fig. 5.13) : (i) la première étape consiste à identifier les paramètres permettant de discriminer les trois catégories de valences émotionnelles (*Positive*, *Neutre* et *Négative*) pour chaque groupe de sujet. Cela permet notamment d'estimer, selon les groupes de sujets, la contribution des composantes prosodiques qui interviennent dans la production spontanée des corrélats de l'affect ; (ii) la deuxième étape consiste à estimer, la significativité des différences qui peuvent apparaître entre les groupes à travers les valeurs de paramètres communs et identifiés comme corrélés aux catégories de valence affective. Cette deuxième étape permet d'étudier si les différents groupes de sujets utilisent de la même façon les codes de la prosodie pour communiquer la valence affective lors de l'épreuve de narration de l'histoire.

La Fig. 5.14 présente les résultats issus de la première étape d'analyse. Elle illustre les contributions des composantes prosodiques dans la communication de la valence affective à travers plusieurs ancrages acoustiques de la parole et selon les groupes de sujets, cf. chapitre 2. L'ordre d'importance de ces contributions est alors très différent de celui observé sur les émotions prototypiques (cf. Fig. 4.12-13) : la qualité vocale correspond à la contribution la plus importante (~35%) alors que celle des trois autres composantes sont équivalentes (~22%). Les différences observées entre les groupes de sujets dans la contribution des composantes prosodiques ne sont que rarement significatives ; cela est le cas entre les sujets TA et TSL sur le pitch et le rythme et pour les ancrages rythmiques « *p-centre* » identifiés au niveau 2. Notons que les listes des paramètres discriminant sont beaucoup trop nombreuses pour être présentées. Toutefois, une analyse de ces tables a permis de constater que les modèles *non-conventionnels* du rythme (en particulier les mesures de dynamique prosodique HD, A-PHD et I-PHD) se sont une nouvelle fois révélés pertinents dans l'analyse des corrélats des émotions, puisqu'ils ont souvent été présents dans la liste des paramètres ayant permis de différencier significativement les catégories de valence affective.

La seconde étape (cf. Fig. 5.13) consiste à comparer les styles de production des groupes de sujets à travers les valeurs fournies par leurs jeux communs de paramètres prosodiques, discriminant alors les catégories de valence affective. Ainsi, après avoir identifié les paires de groupes présentant un jeu de paramètres communs, nous avons effectué des tests statistiques pour comparer leurs valeurs à travers chaque paramètre. Nous avons ainsi pu estimer la proportion en pourcentage de paramètres prosodiques disponibles (i.e., discriminant les valences affectives et communs aux deux groupes comparés) pour laquelle des différences significatives sont apparues dans les valeurs et entre les groupes de sujets comparés. Ces tests ont été effectués séparément sur les données fournies par les trois catégories de valence de façon à

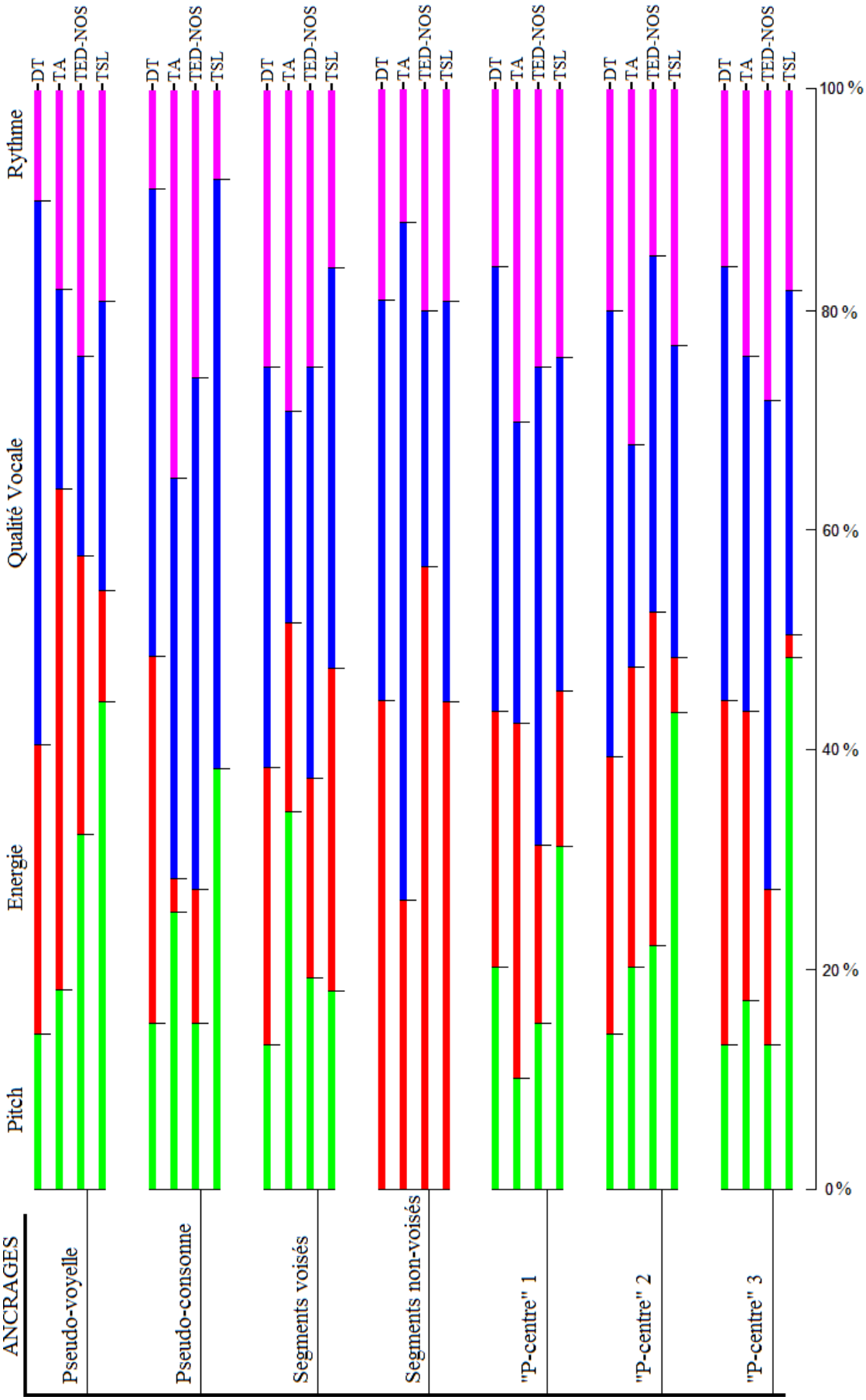


Fig. 5.14 Contribution des composantes prosodiques dans la communication de la valence affective selon les points d'ancrage et les groupes de sujets.

Table 5.20.1 Proportions de paramètres prosodiques corrélés à l'affect et communs entre les groupes de sujets sur lesquels des différences significatives sont apparus entre ces derniers.

Composante prosodique	DT vs. TA	DT vs. TED-NOS	DT vs. TSL	TA vs. TED-NOS	TA vs. TSL	TED-NOS vs. TSL
Pitch	10 / 10 / 10	0 / 0 / 0	30 / 10 / 50	10 / 10 / 0	10 / 0 / 0	0 / 0 / 0
Energie	0 / 0 / 0	50 / 50 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
Qualité Vocale	29 / 9 / 18	12 / 9 / 6	9 / 9 / 15	0 / 0 / 0	3 / 3 / 3	3 / 9 / 3
Rythme	14 / 7 / 0	14 / 7 / 14	0 / 0 / 0	14 / 14 / 21	14 / 21 / 0	0 / 0 / 0
Toutes	17 / 8 / 12	12 / 8 / 7	10 / 7 / 17	5 / 5 / 5	7 / 7 / 2	2 / 5 / 2

Les valeurs sont données pour les trois catégories de valence : *Négative* / *Neutre* / *Positive* ; ancrage pseudo-voyelle ; DT : développement typique ; TA : troubles autistiques ; TED-NOS : troubles envahissants du développement non-spécifiés ; et TSL : troubles spécifiques du langage.

Table 5.20.2 ; Ancrage voisé.

Composante prosodique	DT vs. TA	DT vs. TED-NOS	DT vs. TSL	TA vs. TED-NOS	TA vs. TSL	TED-NOS vs. TSL
Pitch	23 / 0 / 0	0 / 8 / 0	23 / 15 / 31	8 / 0 / 0	0 / 0 / 0	0 / 8 / 0
Energie	44 / 0 / 6	0 / 0 / 0	13 / 19 / 0	6 / 0 / 19	0 / 0 / 0	6 / 0 / 0
Qualité Vocale	11 / 8 / 6	8 / 6 / 8	26 / 11 / 15	9 / 0 / 6	2 / 2 / 2	6 / 2 / 6
Rythme	40 / 6 / 2	2 / 1 / 3	7 / 6 / 5	4 / 4 / 3	0 / 3 / 1	1 / 1 / 1
Toutes	31 / 5 / 3	3 / 3 / 3	14 / 9 / 9	6 / 2 / 5	1 / 2 / 1	3 / 2 / 2

assurer une certaine homogénéité dans l'analyse statistique. Les résultats sont présentés pour différents points d'ancrage de la parole qui ont été automatiquement identifiés, cf. tables 5.20. Ces tables montrent que les sujets atteints de TC utilisent des codes prosodiques qui sont quasi systématiquement différents de ceux des sujets à DT, puisqu'une proportion non négligeable de paramètres discriminants (la valence affective) et communs aux deux groupes (e.g., TA *vs.* DT) sont significativement différents lorsque l'on compare leurs valeurs. Les résultats les plus spectaculaires sont fournis par les ancrages « *p-centre* » pour lesquels la proportion de ces paramètres peut atteindre jusqu'à 75% de ceux disponibles (e.g., TA *vs.* DT, émotions *Positives*), cf. table 5.20.4. Notons que nous avons également pu observer des différences significatives entre les sujets pathologiques, en particulier entre les TA et les TED-NOS.

Enfin, nous avons repris l'analyse des modèles *conventionnels* et *non-conventionnels* du rythme à travers les pseudo-phonèmes (cf. chapitre 4, sous-section 4.4) et selon les catégories d'émotions et les groupes d'enfants, cf. tables A.5.1-4 et Fig. A.5.1-3 en annexe. Les résultats montrent que les émotions produites par les sujets à DT sont très bien séparées dans le plan formé par les modèles *conventionnels* %V et ΔC du rythme, cf. Fig. A.5.1. Les émotions *Négatives* entraînent alors des minima, et les émotions *Positives* des maxima. Le groupe des TED-NOS montre que même si les émotions peuvent être assez bien séparées dans le plan %V – ΔC, la position de la valence *Négative* est inversée avec celle du « *Neutre* », comparé aux résultats obtenus par les DT. Alors que les deux autres groupes de sujets obtiennent des résultats très différents : les groupes TA et TSL amènent des valeurs qui sont beaucoup plus proches entre les émotions. De plus, les valeurs sont nettement plus faibles pour les TSL. Ces résultats se retrouvent aussi sur certains paramètres *non-conventionnels* du rythme, cf. Fig. A.5.2-3.

Table 5.20.3 ; Ancrage non-voisé.

Composante prosodique	DT vs. TA	DT vs. TED-NOS	DT vs. TSL	TA vs. TED-NOS	TA vs. TSL	TED-NOS vs. TSL
Energie	0 / 0 / 5	30 / 20 / 10	25 / 5 / 5	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
Qualité Vocale	12 / 24 / 16	12 / 0 / 12	16 / 8 / 0	0 / 4 / 4	0 / 4 / 12	0 / 4 / 0
Rythme	2 / 2 / 0	28 / 2 / 5	21 / 9 / 2	0 / 0 / 2	2 / 0 / 0	0 / 0 / 0
Toutes	5 / 8 / 6	22 / 6 / 9	23 / 8 / 2	0 / 2 / 3	2 / 2 / 4	0 / 1 / 0

Table 5.20.4 Ancrage « p-centre » 3.

Composante prosodique	DT vs. TA	DT vs. TED-NOS	DT vs. TSL	TA vs. TED-NOS	TA vs. TSL	TED-NOS vs. TSL
Pitch	0 / 0 / 0	20 / 30 / 20	20 / 10 / 40	10 / 0 / 0	0 / 0 / 0	10 / 20 / 0
Energie	71 / 29 / 29	14 / 0 / 29	0 / 0 / 0	0 / 0 / 14	0 / 0 / 0	0 / 0 / 0
Qualité Vocale	18 / 10 / 6	12 / 18 / 14	10 / 8 / 8	12 / 6 / 6	0 / 2 / 4	4 / 0 / 0
Rythme	22 / 7 / 20	20 / 13 / 26	11 / 9 / 11	4 / 0 / 9	4 / 0 / 0	2 / 2 / 4
Toutes	21 / 9 / 12	16 / 16 / 20	11 / 8 / 12	8 / 3 / 7	2 / 1 / 2	4 / 3 / 2

4.3.4. Discussion des résultats

Cette deuxième partie porte sur l'étude des données issues de l'épreuve de *production de parole affective spontanée*. Nous avons exploité pour cela un livre [MAY69][53] présentant une histoire illustrée par une série de 24 images, cf. table 5.5. Les évènements associés aux images ont été catégorisés en quatre classes affectives selon la valence, et il a été demandé aux sujets de décrire spontanément l'histoire à travers les images du livre. 6229 groupes de souffle ont ensuite été segmentées manuellement à partir des enregistrements, cf. table 5.6.

Bien que les résultats de la littérature soient relativement fournis, ils ne permettent pas pour autant de répondre significativement aux questions concernant un usage typique ou atypique de la prosodie pour communiquer les émotions chez les sujets atteints de TC, du moins, les résultats sont contradictoires selon les études et selon les aspects de la perception et de la production, cf. sous-section 1.2. Nous avons donc préféré utiliser une approche qui permet de traiter d'un seul bloc les capacités en perception et en production de l'enfant, e.g., description spontanée des indices émotionnels perçus par les sujets. L'analyse de ces données a été effectuée avec les traitements automatiques qui ont été proposés dans les chapitres précédents : (i) identification automatique d'ancrages acoustiques et (ii) extraction de paramètres prosodiques avec les modèles *conventionnels* et *non-conventionnels* pour le rythme.

Cette analyse a tout d'abord montré que la répartition des groupes de souffle selon les catégories de valence affective ne varie que très peu selon les groupes de sujet, i.e., DT, TA, TED-NOS et TSL, et que cette distribution privilégie les évènements catégorisés comme « *Neutre* », cf. table 5.19. Les données ont ensuite montré que tous les sujets TC produisent des groupes de souffle qui sont significativement plus courts que ceux des enfants à DT, cf. table 5.19. Ces résultats s'opposent à ceux de la première épreuve qui consistait à imiter des phrases (tâche contrainte). Les psychologues, qui ont participé à cette étude, supposent que la différence de résultats, entre le discours spontanée et pour la répétition de phrase, est certainement liée au traitement psychique des émotions, indépendamment des mécanismes de reco-

nnaissance et de reproduction ; nos deux épreuves sont donc complémentaires et permettent d'analyser plus précisément les mécanismes en jeux. Par ailleurs, cette différence peut s'expliquer par les troubles associés aux TC. En effet, les difficultés à produire de la parole peuvent être la cause d'une limitation dans la durée de production des sujets TSL ; [WEL-03][18], [HAR89][27] et [SAM03][28]. Concernant les TEDs, leur difficulté à traiter les informations pragmatiques pourrait limiter leur durée de production, puisqu'ils n'ont pas vraiment la possibilité de comprendre et de retransmettre les informations associées aux évènements présents dans les images [MCC07][69] et [PAU05a][21].

Notre étude a ensuite consisté à identifier les paramètres prosodiques qui permettent de discriminer les trois catégories de valences émotionnelles selon chaque groupe de sujet, cf. Fig. 5.14. Cette étape a permis d'estimer la contribution des composantes prosodiques qui interviennent dans la production spontanée des corrélats émotionnels de la voix. Les résultats montrent que l'ordre d'importance de ces contributions est similaire à tous les groupes de sujets et très différent de celui observé sur les émotions prototypiques (cf. Fig. 4.12-13) : la qualité vocale apparaît comme la contribution la plus importante (~35%) alors que celles des trois autres composantes sont globalement équivalentes (~22%), i.e., pitch, énergie et rythme.

Une autre analyse consistait à isoler les paramètres discriminants communs aux paires de groupes de sujets, puis comparer leurs valeurs selon les trois catégories de valence. Les résultats ont montré que les sujets TC encodent la prosodie d'une façon qui est quasi systématiquement différente de celle des sujets typiques, cf. tables 5.20. En effet, une proportion non-négligeable de paramètres discriminants et communs aux paires de groupes comparés (i.e., TC *vs.* DT) a entraîné des différences significatives sur de nombreux ancrages acoustiques. Ainsi, même si les sujets TC arrivent à exploiter la prosodie pour transmettre des émotions via des paramètres identiques aux DT, l'encodage réalisé (i.e., les valeurs des paramètres) se trouve significativement différent de celui des typiques.

Enfin, l'étude des modèles *conventionnels* et *non-conventionnels* du rythme a permis de mettre en lumière certaines particularités dans le traitement des émotions chez les sujets TED-NOS vis-à-vis des autres sujets atteints de TC, cf. Fig. A.5.1-3. Ces modèles suggèrent par exemple que les sujets TED-NOS ont tendance à surjouer les émotions, puisque le « *Neutre* » conduit à des minima dans les valeurs, alors qu'il est situé entre les deux catégories opposées de valence affective pour les sujets à DT, cf. Fig. A.5.1. Les sujets TA et TSL montrent, quant à eux, des valeurs très proches entres les émotions, ce qui suggère une absence de traitements dédiés aux émotions.

5. Conclusion

La première partie de ce chapitre a été consacrée à la description des TC telles que les TED et les TSL. Les diagnostics cliniques et la prévalence de ces troubles chez les populations infantiles ont alors été présentés. Nous avons ensuite donné un bref état-de-l'art des études qui se sont consacrées aux compétences de sujets TC dans la production et la perception des fonctionnalités prosodiques, e.g., *grammaticale, pragmatique* et *affective*. Ces études, qui reposent sur une analyse manuelle des données, ont montré des résultats contradictoires concernant les compétences des sujets TC comparées à celles des DT. L'utilisation des mé-

thodes dérivées du TAP et de la reconnaissance des formes ont alors été proposées pour combler les lacunes produites par les jugements humaines, e.g., subjectivité du jugement et évaluations catégorielles.

L'étude que nous avons conduite dans ce chapitre a été effectuée en étroite collaboration avec des cliniciens et des psychologues. Il a été envisagé la possibilité d'utiliser un système automatisé pour évaluer les capacités prosodiques d'un enfant, notamment sur les aspects *grammaticaux*, et *affectifs*. Ces tâches, qui sont traditionnellement administrées par des orthophonistes, sont effectuées dans cette thèse au moyen de techniques issues du domaine du TAP et de la reconnaissance des formes. Elles incluent notamment : (i) une étape d'identification automatique de plusieurs points d'ancrages acoustiques complémentaires de la parole, (ii) une étape d'extraction de LLDs prosodiques selon ses différentes composantes et (iii) une méthode pour identifier les paramètres qui sont à la fois corrélés à un ensemble de catégories d'informations, e.g., l'intonation et les émotions spontanées dans ce chapitre, et communs à plusieurs groupes de sujets.

Nous avons ensuite étudié la pertinence d'un tel système dans une tâche visant à comparer les fonctionnalités prosodiques de divers groupes d'enfants, i.e., DT et TC (TA, TED-NOS, et TSL) selon deux épreuves distinctes : (i) *imitation du contour intonatif* (contrainte) et (ii) *production de parole affective spontanée* (non-contrainte). Une étude des données fournies par la première épreuve montre que les enfants atteints de TC sont capables d'atteindre des performances comparables à celles des DT sur le contour intonatif « *Descendant* ». Toutefois, les performances n'ont pas été portées par le même type de contribution prosodique entre les TC et les DT, puisque une approche dynamique, i.e., basée sur des HMM, a été plus contributive pour la reconnaissance des intonations des sujets typiques, alors que l'approche statique a produit de meilleurs résultats pour tous les sujets pathologiques. Notons que les modèles HMM requièrent une certaine régularité dans la structure prosodique. Alors que les modèles statistiques reposent uniquement sur des particularités dans les contours intonatifs.

Les résultats amenés par la seconde épreuve ont montré que tous nos sujets ont globalement réparti de la même façon leur production spontanée selon les catégories de valence affective. Concernant la durée de production des phrases (ou groupes de souffle), le cadre spontané de la deuxième épreuve amène des résultats opposés à ceux de la première épreuve : les durées des phrases produites par les TC sont toutes très inférieures à celles des DT, de plus, ces valeurs sont significativement différentes entre les sujets TED-NOS et les TA. L'ordre d'importance des contributions apportées par les composantes prosodiques est similaire à tous les groupes et différent de celui observé sur les émotions prototypiques : la qualité vocale satisfait une contribution qui est la plus importante de toutes alors que celles des trois autres sont plutôt équivalentes. Une analyse détaillée des paramètres fait apparaître de nombreuses différences entre les groupes. En effet, les sujets TC utilisent la prosodie d'une façon qui est souvent significativement différente de celle des DT. Notons que nous avons également pu observer des différences significatives entre les sujets TA et les TED-NOS. Enfin, les modèles *conventionnels* et *non-conventionnels* du rythme suggèrent que les TED-NOS ont tendance à surjouer les émotions, au contraire des sujets TA et TSL pour lesquels les valeurs ne montrent pas de variations significatives entre les catégories associées à la valence affective.

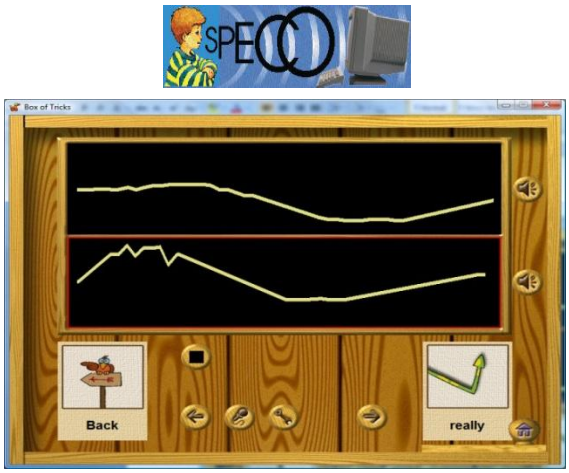


Fig. 5.15 Illustration d'une des tâches proposée par le logiciel éducatif SPECO ; profil intonatif servant de stimuli (image du haut) ; et contour intonatif produit par un enfant (image du bas).

Ainsi, les résultats produits par les systèmes proposés dans ce chapitre permettent de caractériser de façon significative les particularités prosodiques des sujets TC, notamment sur les fonctionnalités *grammaticales* et *affectives*. Par conséquent, ces systèmes pourraient être utilisés pour adapter les protocoles de remédiation prosodiques servant à améliorer les capacités de communication et d'interaction sociale des sujets atteints de TC, puisque les paramètres sur lesquels agir ont été identifiés. La technologie proposée pourrait, par exemple, être intégrée dans un système entièrement automatisé exploitable par des orthophonistes. L'étape d'acquisition de données serait effectuée manuellement par le clinicien alors que les données de référence, i.e., fournies par les enfants à DT, auraient déjà été collectées et mises à disposition pour réaliser l'apprentissage des modèles prosodiques inclus dans les systèmes. Notons que les profils intonatifs et les phrases qui ont été utilisés dans la première épreuve de cette étude sont dépendants de la langue, ces données devront donc être adaptées pour effectuer une étude de l'intonation sur des langues autres que le Français.

Enfin, nous devons préciser qu'il existe déjà des logiciels éducatifs ludiques qui proposent d'améliorer les compétences de l'enfant sur différents aspects du langage, et via des interactions de type *bio-feedback* (e.g., *Speech Corrector*, SPECO[77], Fig. 5.15). Des expériences ont ainsi montré l'impact de ce type d'interaction sur des TC puisque les performances des sujets ont été meilleures dans la reproduction de l'intonation avec un affichage et une écoute simultanée du contour intonatif comparé à une écoute seule [JAM76][78] et [DEB83][79]. Ces systèmes pourraient être améliorés en incluant d'autres types de paramètres prosodiques, tels que les modèles *conventionnels* (statistique phonétique) et *non-conventionnels* du rythme (e.g., dynamique prosodique), puisque nous avons démontré leur pertinence tant sur l'étude des émotions prototypiques (cf. chapitre 4, sous-section 4.4) que spontanées (cf. sous-section 4.3).

[77] K. Vicsi, *A multimedia multilingual teaching and training system for speech handicapped children*, Final Annual Report, SPECO-977126, University of Technology and Economics, Departements of Telecommunications and Telematics, 09.1998 - 08.2001; http://alpha.tmit.bme.hu/speech/speco/index.html.

[78] K. de Bot "Visual feedback of intonation: effectiveness and induced practice behavior", dans *Lang. and Speech*, vol. 26, no. 4, pp. 331–350, Oct.-Dec. 1983.

[79] E. James, "The acquisition of prosodic features of speech using a speech visualizer", dans *Inter. R. of Applied Ling.*, vol. 14, pp. 227–243, 1976.

Chapitre 6

Conclusions et perspectives

Cette thèse a présenté la problématique du TAP orienté émotion dans un cadre acté et spontané (troubles de la communication). Nous avons ainsi décrit dans le *Chapitre 1* le terme *émotion* avec d'autres termes du même contexte. Différentes théories des émotions produites par les études en psychologie de l'émotion ont été présentées. Ces théories ont un fort impact sur les processus d'étiquetage de données émotionnelles. Ce chapitre décrit également les divers processus d'encodage et de décodage des informations affectives dans la parole, et les méthodes de constitution de corpus selon différents degrés de contrôle dans l'induction des émotions. Ces techniques influencent la construction et l'analyse des corpus de parole émotionnelle.

Les différents types d'ancrages acoustiques existants dans la parole ont été présentés dans le *Chapitre 2*. Ces ancrages servent de support à l'encodage (ou l'actualisation) des informations dans la parole et peuvent être de nature linguistique (e.g., phonèmes et syllabes) ou liés à des phénomènes de perception (e.g., segment voisés et non-voisés, pseudo-phonèmes et « *p-centre* »). L'identification robuste des supports d'encodage des informations dans la parole conditionne la tâche de reconnaissance de ces dernières. Des méthodes permettant d'identifier automatiquement différents types d'ancrages acoustiques (e.g., vocalique, consonantique, et rythmique) ont été décrites. Ces méthodes ont été testées avec succès sur plusieurs corpus incluant différentes configurations d'analyse : changement de locuteur, de langue et de style de production. Les expériences en détection de voyelles ont notamment montrées que les résultats issus de la littérature sur le corpus TIMIT ont été très nettement améliorés (VER = 19.5% contre 29.5%). Les structures acoustiques associées aux « *p-centres* » ont par ailleurs été caractérisées via les autres types d'ancrages acoustiques (e.g., vocalique et consonantique). Ces expériences ont montré que (i) les structures acoustiques associées aux « *p-centre* » sont clairement de nature voisée, et (ii) que les consonnes participent de façon non négligeable à leur structuration acoustique, et ce d'autant plus lorsque la parole est chargée d'émotion.

Des méthodes permettant de modéliser les composantes acoustiques du signal de parole ont été présentées dans le *chapitre 3*. La littérature montre que les coefficients cepstraux MFCC sont adaptés à de nombreuses tâches issues du TAP. Concernant la reconnaissance des

émotions, une étape de normalisation de ces paramètres est requise pour diminuer l'influence du locuteur sur les caractéristiques acoustiques mesurables de l'affect. Le système de reconnaissance que nous avons proposé repose sur la fusion de différentes approches pour caractériser les coefficients MFCC. La contribution de ces approches dans la tâche de reconnaissance des émotions a été estimée lors des étapes de fusion (i.e., les poids associés à une combinaison linéaire des informations fusionnées). Les expériences ont montré que l'introduction des informations locuteur dans le système de reconnaissance est pertinente puisque les scores s'en trouvent significativement améliorés. Les performances se dégradent néanmoins sur les autres types de normalisation. L'apport de la fusion des différents points d'ancrages de la parole sur les scores de reconnaissance est relativement faible. Les meilleurs scores obtenus en reconnaissance d'émotions sur le corpus Berlin et avec les 16 coefficients MFCC sont de 68% pour la méthode CVS, et de 59% pour la méthode LOSO. Bien que ces résultats ne soient pas très élevés, ils sont néanmoins comparables avec ceux issus de la littérature pour une configuration similaire du système de reconnaissance.

Nous avons présenté dans l'introduction du *Chapitre 4* différentes théories associées aux définitions du rythme, montrant ainsi la complexité des phénomènes véhiculés par ce dernier. La caractérisation de ces phénomènes ne peut se limiter à des mesures statistiques telles que %V et ΔC, ou encore sur le débit. Nous avons montré, aux travers des études effectuées sur le sujet, que le pitch peut, par exemple, être à l'origine de multiples phénomènes associés au rythme. Comme le rythme apparaît clairement comme sous-modélisé dans les systèmes issus de l'état de l'art en TAP orienté émotion, nous avons développé de nouvelles mesures du rythme. Différentes techniques de degré de complexité variable ont ainsi été exploitées pour quantifier ces phénomènes dans le signal de parole. La pertinence des paramètres du rythme pour la reconnaissance des émotions a été démontrée dans les expériences qui ont été réalisées dans ce chapitre. Une analyse détaillée des résultats montre que le meilleur score de chaque émotion est très souvent obtenu par différents ancrages acoustiques de la parole et que ces derniers varient également selon les analyses ; le score moyen est alors de 78% en CVS et 72% en LOSO. Ce qui démontre la pertinence d'effectuer l'étape d'extraction de caractéristiques et de classification selon plusieurs points d'ancrages complémentaires de la parole.

L'introduction du *chapitre 5* fut consacrée à la description des TC telles que les TED et les TSL. Les diagnostics cliniques et la prévalence de ces troubles chez les populations infantiles ont alors été présentés. Nous avons ensuite donné un bref état-de-l'art des études qui se sont consacrées aux compétences de sujets TC dans la production et la perception des fonctionnalités prosodiques, e.g., *grammaticale, pragmatique* et *affective*. Ces études, qui reposent sur une analyse manuelle des données, ont montré des résultats contradictoires concernant les compétences des sujets TC comparées à celles des DT. L'utilisation des méthodes dérivées du TAP et de la reconnaissance des formes ont alors été proposées pour combler les lacunes produites par les jugements humaines, e.g., subjectivité du jugement et évaluations catégorielles. L'étude que nous avons conduite dans ce chapitre a été effectuée en étroite collaboration avec des cliniciens et des psychologues. Il a été envisagé la possibilité d'utiliser un système automatisé pour évaluer les capacités prosodiques d'un enfant, notamment sur les aspects *grammaticaux*, et *affectifs*. Ces tâches, qui sont traditionnellement administrées par des orthophonistes, sont effectuées dans cette thèse au moyen de techniques issues du domaine du TAP et

de la reconnaissance des formes. Elles incluent notamment : (i) une étape d'identification automatique de plusieurs points d'ancrages acoustiques complémentaires de la parole, (ii) une étape d'extraction de LLDs prosodiques selon ses différentes composantes et (iii) une méthode pour identifier les paramètres qui sont à la fois corrélés à un ensemble de catégories d'informations, e.g., l'intonation et les émotions spontanées dans ce chapitre, et communs à plusieurs groupes de sujets. Les résultats que nous avons obtenus par une analyse automatique des données collectées selon deux types d'épreuves distinctes (*grammaticale* – contrainte, *affective* – non-contrainte) ont permis de retrouver de façon significative ceux issus des diagnostiques cliniques associés au type de TC étudié. Dans la mesure où les systèmes proposés dans cette étude reposent sur un traitement automatique du signal de parole, leur intérêt pour le diagnostic des TC à travers la prosodie est ainsi pleinement justifié. De plus, ces systèmes pourraient être intégrés dans un logiciel qui serait exploité par des orthophonistes dans le but d'utiliser des protocoles de remédiation de troubles dans la prosodie adaptés aux sujets. Cela servirait ainsi à améliorer à la fois les capacités de communication et d'interactions sociales des enfants atteints de TC.

Les perspectives portées par cette thèse sont multiples et se situent à la fois au niveau théorique et applicatif. Concernant les aspects théoriques, nous pensons qu'il pourrait être pertinent d'exploiter les liens existant entre des ancrages acoustiques complémentaires de la parole pour caractériser la prosodie. Les meilleurs scores en reconnaissance d'émotions sont par exemple souvent obtenus lorsque les données du rythme sont normalisées selon les informations linguistiques, ce qui montre l'importance de la prise en compte du contexte lors de l'étape d'extraction de caractéristiques. Par ailleurs, les consonnes pourraient être associées à un détecteur spécifique, plutôt que d'être considérées comme des segments de parole de type « non-voyelle », i.e., pseudo-consonne. Dans l'optique de caractériser les aspects dynamiques de la communication, i.e., les variations de styles expressifs liées aux interactions humaines, il apparaît nécessaire de définir de nouvelles techniques d'extraction de caractéristiques. Ces dernières pourraient alors reposées sur des informations multimodales (e.g., parole, visage et geste) et multidimensionnelles (e.g., composantes prosodiques) dans lesquelles des paramètres de synchronie *intra* et *inter* locuteur pourraient être calculés. Nous avons par exemple présenté des travaux préliminaires exploitant le détecteur de voyelle avec un détecteur de mouvements labiaux pour caractériser la synchronisation du flux audio-visuelle dans le corpus de parole lue VidTIMIT [ABE09]. Les résultats obtenus par notre système se sont alors révélés conformes à ceux de la littérature, i.e., une maximisation de l'information mutuelle lors d'un léger décalage temporel des flux audiovisuels.

Concernant les perspectives applicatives, il pourrait être intéressant d'étudier l'impact des modèles *conventionnels* et *non-conventionnels* du rythme présentés dans cette thèse sur une tâche de caractérisation du continuum rythmique des langues. Des corrélations pourraient sûrement apparaître entre les paramètres du rythme et les groupes de langues définies dans la littérature. Enfin, une autre perspective applicative de cette thèse concerne l'intégration des techniques proposées, notamment l'identification automatique des ancrages acoustiques de la parole et des paramètres prosodiques associés, dans un système permettant d'aider un clinicien à effectuer des diagnostics et/ou des tâches de remédiation orthophonique sur des sujets atteints de TC.

Annexe 1

Etude détaillée du corpus Berlin

Cette première annexe présente une étude détaillée des transcriptions fournies par le corpus Berlin. L'alphabet SAMPA de l'allemand qui a servi à effectuer les transcriptions est tout d'abord présenté. Nous étudions ensuite ces transcriptions selon les symboles fournis par l'alphabet SAMPA. Cette étude a permis de mettre en lumière certaines variabilités présentes dans les transcriptions phonétiques. Nous avons évalués l'impact de ces variabilités sur la reconnaissance des émotions en utilisant un système similaire à celui employé dans le chapitre 3 (16 MFCC en CV). Cet impact s'avère minime puisque les scores ont été soit dégradés, soit améliorés mais de façon non significative. Nous clôturons enfin l'étude des transcriptions contenues dans le corpus Berlin par une analyse statistique de la durée des phonèmes et des macro-classes phonétiques définies par l'alphabet SAMPA selon les catégories d'émotions.

1. Analyse des transcriptions définies par SAMPA

Nous présentons ci-dessous l'alphabet SAMPA[1] pour l'Allemand ainsi que les différentes classes phonétiques qu'il introduit pour les voyelles et les consonnes. Les transcriptions phonétiques fournies par le corpus Berlin sont ensuite détaillées selon les classes définies par l'alphabet SAMPA.

1.1. L'alphabet SAMPA de l'Allemand

Les voyelles se répartissent en trois classes dans l'alphabet SAMPA de l'Allemand : (i) *vérifiées* (i.e., courtes), (ii) *libres* (i.e., longues) et (iii) les phonèmes correspondant aux réalisations voisées du phonème /r/, cf. table A.1.1. D'après les auteurs de cet alphabet, il existerait une distinction importante entre les voyelles courtes et longues, puisque ces dernières présenteraient une durée à peu près deux fois plus longue que les voyelles courtes. La troisième classe des voyelles inclut les réalisations voisées du /r/ qui sont représentées par un /6/ lorsqu'elles sont fusionnées avec le schwa /@/. Ce dernier est par définition non accentué et est toujours précédé d'une voyelle accentuée, produisant ainsi un ensemble de diphtongues additionnelles. 35 phonèmes « *voyelle* » sont définis par l'alphabet SAMPA.

Le système consonantique allemand comprend entre 19 et 21 phonèmes occlusifs (selon si l'on inclut ou non les sons périphériques empruntés aux langues étrangères) et 5 *sonorantes*, cf. table A.1.2. Les occlusifs comprennent 6 *plosives* (7 si l'on considère le phénomène acoustique correspondant à l'instant de fermeture glottique), 3 (ou 4) *affriquées* et 10 *fricatives*. Comme pour l'Anglais, les phonèmes occlusifs sont différenciés selon leur voisement. Toutefois, la nature périodique du signal de parole n'est pas le seul critère utilisé pour contraster les segments consonantiques entre eux. La durée et l'intensité des sons produits peuvent aussi servir à effectuer la classification ; e.g., lenis et fortis. 26 phonèmes « *consonne* » sont définis par l'alphabet SAMPA.

1.2. Transcriptions contenues dans le corpus Berlin

Le corpus de parole émotionnelle Berlin comporte différentes tires de transcriptions réparties dans plusieurs fichiers. Les fichiers *.lablaut* contiennent par exemple les phonèmes transcrits selon l'alphabet SAMPA ainsi que des informations sur la qualité vocale et le niveau d'accentuation des voyelles. Alors que les syllabes sont données dans les fichiers *.silb*. Comme plusieurs types d'informations sont rassemblés dans les fichiers *.lablaut*, nous avons réalisé un programme pour les identifier automatiquement et les séparer dans un fichier respectif : *.pho* (phonèmes), *.dia* (diacritiques de la qualité vocale) et *.mrk* (marqueurs d'accentuation). Cela permet de faciliter le traitement automatique des données. Un astérisque a été inséré sur les segments ne contentant pas de symbole de transcription. Nous analyserons dans la section suivante les informations qui sont attribuées à ces segments, e.g., silence, prolongement du segment précédent, ou autre type d'information.

[1] http://www.phon.ucl.ac.uk/home/sampa/german.htm.

Table A.1.1 Alphabet des classes phonétiques « *voyelle* » selon l'alphabet SAMPA de l'Allemand.

Classe	Phonème	Mot	Transcription
Courte	@[2]	bitte	"bIt@
	0	Trotz	tr0ts
	9	plötzlich	"pl9tslIC
	A	Satz	zats
	E	Gesetz	g@"zEts
	I	Sitz	zIts
	U	Schutz	SUts
	Y	hübsch	hYpS
Longue	2:	blöd	bl2:t
	a:	Tat	ta:t
	e:	Beet	be:t
	i:	Lied	li:t
	o:	rot	ro:t
	u:	Blut	blu:t
	y:	süß	zy:s
	E:	spät	SpE:t
	aI[3]	Eis	aIs
	aU[6]	Haus	haUs
	0Y[6]	Kreuz	kr0Yts
Diphtongue du schwa	6	besser	"bEs6
	i:6	Tier	ti:6
	I6	Wirt	vI6t
	y:6	Tür	ty:6
	Y6	Türke	"tY6k@
	e:6	schwer	Sve:6
	E6	Berg	bE6k
	E:6	Bär	bE:6
	2:6	Föhr	f2:6
	96	Wörter	"v96t6
	a:6	Haar	ha:6
	a6	hart	ha6t
	u:6	Kur	ku:6
	U6	kurz	kU6ts
	o:6	Ohr	o:6
	06	dort	d06t

Afin de répertorier les symboles utilisés dans les transcriptions phonétiques du corpus Berlin, nous avons créé un script permettant d'obtenir l'alphabet de ces labels. La table A.1.4

[2] La catégorie du schwa n'est pas définie dans le document de référence de SAMPA. Néanmoins nous pouvons supposer que ce dernier appartient au groupe des voyelles courtes compte tenu de ses caractéristiques en production (réalisation brève).

[3] Diphtongues.

Table A.1.2 Alphabet des classes phonétiques « *consonne* » selon l'alphabet SAMPA de l'Allemand.

Classe	Phonème	Mot	Transcription
Plosive	b	Bein	baIn
	d	Deich	daIC
	g	Gunst	gUnst
	k	Kunst	kUnst
	p	Pein	paIn
	t	Teich	taIC
	?[4]	Verein	fE6" ?aIn
Affriquée	pf	Pfahl	pfa:l
	ts	Zahl	tsa:l
	tS	deutsch	d0YtS
	dZ[5]	Dschungel	"dZUN=l
Fricative	f	fast	Fast
	h	Hand	Hant
	j	Jahr	ja:6
	s	Tasse	"tas@
	v	was	Vas
	x	Buch	bu:x
	z	Hase	"ha:z@
	C	sicher	"zIC6
	S	waschen	"vaS=n
	Z	Genie	Ze"ni:
Sonorante	m[6]	mein	maIn
	n[3]	nein	naIn
	N[3]	Ding	dIN
	l[7]	Leim	laIm
	R[4]	Reim	RaIm

Table A.1.3 Exemple d'identification et de recodage des labels issus du fichier *03a02Wb.lablaut*.

Fichier existant *03a02Wb.lablaut*	**Fichier généré** *03a02Wb.pho*	**Fichier généré** *03a02Wb.dia*
0.000000 -1 (	0.000000 0.076557 (	0.000000 0.076557 *
0.076557 -1 _d-sth	0.076557 0.108014 _d	0.076557 0.108014 -sth
0.108014 -1 !d	0.108014 0.116844 !d	0.108014 0.116844 *
0.116844 -1 +sth	0.116844 0.121259 *	0.116844 0.121259 +sth
0.121259 -1 a+hoe	0.121259 0.204591 a	0.121259 0.204591 +hoe
… … …	… … …	… … …
2.020969 -1 .	2.020969 2.020969 .	2.020969 2.020969 *

[4] Phonème correspondant à un instant de fermeture glottique.

[5] Phonème issu de l'emprunt de mots étrangers à l'Allemand.

[6] Phonèmes nasalisés.

[7] Phonèmes liquides.

Table A.1.4 Classe et macro-classe phonétiques des transcriptions du corpus Berlin identifiées par l'alphabet SAMPA de l'Allemand.

Phonème	M.-C.	Classe	Phonème	M.-C.	Classe	Phonème	M.C.	Classe
!	N.R.	N.R.	I@	N.R.	N.R.	e@	N.R.	N.R.
!b	N.R.	N.R.	J	N.R.	N.R.	f	consonne	fricative
!d	N.R.	N.R.	L	N.R.	N.R.	g	consonne	plosive
!g	N.R.	N.R.	N	consonne	sonorante	gas	N.R.	N.R.
!k	N.R.	N.R.	NN	N.R.	N.R.	gasp	N.R.	N.R.
!p	N.R.	N.R.	O	voyelle	courte	h	consonne	fricative
!t	N.R.	N.R.	OU	N.R.	N.R.	hoe	N.R.	N.R.
(	N.R.	N.R.	OY	voyelle	longue	i	N.R.	N.R.
(gef	N.R.	N.R.	R	consonne	sonorante	i6	N.R.	N.R.
*	N.R.	N.R.	S	consonne	fricative	i@	N.R.	N.R.
.	N.R.	N.R.	U	voyelle	courte	ia	N.R.	N.R.
4	N.R.	N.R.	V	N.R.	N.R.	j	consonne	fricative
6	voyelle	diphtongue	X	N.R.	N.R.	k	consonne	plosive
8	N.R.	N.R.	Y	voyelle	courte	kasp	N.R.	N.R.
9	voyelle	courte	Y6	voyelle	diphtongue	l	consonne	sonorante
?	consonne	plosive	_b	N.R.	N.R.	m	consonne	sonorante
@	voyelle	courte	_d	N.R.	N.R.	mm	N.R.	N.R.
@6	N.R.	N.R.	_g	N.R.	N.R.	n	consonne	sonorante
@E	N.R.	N.R.	_k	N.R.	N.R.	nn	N.R.	N.R.
@O	N.R.	N.R.	_p	N.R.	N.R.	o	N.R.	N.R.
@a	N.R.	N.R.	_t	N.R.	N.R.	p	consonne	plosive
A	N.R.	N.R.	a	voyelle	courte	pasp	N.R.	N.R.
AO	N.R.	N.R.	a6	voyelle	diphtongue	s	consonne	fricative
AU	N.R.	N.R.	aI	voyelle	longue	ss	N.R.	N.R.
B	N.R.	N.R.	aO	N.R.	N.R.	t	consonne	plosive
C	consonne	fricative	aU	voyelle	longue	tasp	N.R.	N.R.
D	N.R.	N.R.	ao	N.R.	N.R.	u	N.R.	N.R.
E	voyelle	courte	b	consonne	plosive	v	consonne	fricative
E6	voyelle	diphtongue	basp	N.R.	N.R.	vv	N.R.	N.R.
G	N.R.	N.R.	d	consonne	plosive	x	consonne	fricative
H	N.R.	N.R.	dasp	N.R.	N.R.	y	N.R.	N.R.
I	voyelle	courte	dsap	N.R.	N.R.	z	consonne	fricative
I6	voyelle	diphtongue	e	N.R.	N.R.			

M.-C. : macro-classe phonétique ; N.R. : label non-renseigné par l'alphabet SAMPA.

répertorie les labels et précise le type (e.g., *voyelle*, *consonne*) et la catégorie (e.g., *plosive*, *fricative*) des informations qu'ils véhiculent d'après l'alphabet SAMPA. Notons que les renseignements fournis par cet alphabet ne permettent de caractériser qu'une petite partie (38%) des labels utilisés dans les transcriptions phonétiques du corpus Berlin. La table A.1.5 présente une analyse statistique des durées des macro-classes phonétiques *voyelle* et *consonne* qui ont été identifiées par l'alphabet SAMPA dans les transcriptions, cf. table A.1.4. Les différences annoncées quant aux durées des voyelles selon les classes c*ourte* et l*ongue* peuvent s'observer dans les données de la table A.1.5. Les classes consonantiques font également apparaître des variations importantes au niveau de la durée : les *plosives* sont en moyenne plus courtes que les *fricatives*, qui sont elles-mêmes plus courtes que les *sonorantes*.

Table A.1.5 Caractéristiques des classes et macro-classes phonétiques des transcriptions du corpus Berlin qui sont identifiées par l'alphabet SAMPA de l'allemand.

Groupe SAMPA	Classe phonétique	Effectif	Durée en *ms*	
			μ	**σ**
Voyelle	**Courte**	4493	65.0	35.0
	Longue	362	98.6	59.7
	Diphtongue du schwa	300	86.8	42.5
Consonne	**Plosive**	3668	33.4	23.2
	Fricative	3071	51.7	34.7
	Sonorante	3334	72.7	39.9

Table A.1.6 Caractéristiques des classes phonétiques « *voyelle* » fournies par les transcriptions du corpus Berlin.

Phonème	Classe	Effectif	Durée en *ms*		Phonème	Classe	Effectif	Durée en *ms*	
			μ	**σ**				**μ**	**σ**
9	courte	58	58.9	33.2	ao	*longue ?*	5	132.7	85.3
@	courte	378	50.5	18.9	aO	*longue ?*	8	86.8	35.9
a	courte	1493	82.8	45.3	AO	*longue ?*	6	54.2	39.6
E	courte	314	65.9	28.5	AU	*longue ?*	6	78.3	54.2
I	courte	1649	51.6	21.0	OU	*longue ?*	2	84.5	48.8
O	courte	182	72.8	25.9	6	diphtongue du schwa	259	85.2	42.2
U	courte	323	62.0	24.1	i6	*diphtongue du schwa ?*	84	103.4	71.3
Y	courte	109	70.2	30.3	a6	diphtongue du schwa	17	113.4	52.3
e	*courte ?*	438	105.6	50.6	E6	diphtongue du schwa	16	91.4	34.2
i	*courte ?*	532	59.2	30.2	I6	diphtongue du schwa	7	69.5	28.6
o	*courte ?*	204	72.2	41.0	Y6	diphtongue du schwa	1	79.7	0
u	*courte ?*	52	110.6	39.3	@6	*diphtongue du schwa ?*	12	70.8	45.1
y	*courte ?*	55	66.9	40.1	@a	*diphtongue du schwa ?*	1	48.7	0
A	*courte ?*	1	97.0	x	@E	*diphtongue du schwa ?*	2	68.6	1.5
aI	longue	209	120.8	62.7	@O	*diphtongue du schwa ?*	1	47.8	0
aU	longue	101	75.7	40.6	e@	*diphtongue du schwa ?*	3	192.7	49.7
OY	longue	52	54.3	29.7	i@	*diphtongue du schwa ?*	3	74.1	6.1
ia	*longue ?*	3	117.0	104.9	I@	*diphtongue du schwa ?*	3	63.3	5.3

2. Analyse des transcriptions phonétiques manquantes

Comme nous l'avons décrit précédemment, seulement 38% des labels contenus dans les transcriptions phonétiques du corpus Berlin sont identifiés par l'alphabet SAMPA. Nous nous sommes donc posé la question de savoir à quel(s) type(s) d'informations pouvai(en)t correspondre les 62% autres labels restant. Une analyse manuelle montre que certains de ces labels correspondent par exemple à la fin d'une phrase « . » ou à une pause « (». Toutefois, tous les labels ne peuvent être identifiés de cette façon. Pour cela, nous avons tout d'abord cherché à séparer les labels selon les deux macro-classes phonétiques *voyelle* et *consonne*. Nous avons notamment utilisé les spectres des signaux de parole associés aux labels, puisque les voyelles ont une structure spectrale qualifiée de formantique alors que les consonnes sont associées à un spectre relativement plat dans les hautes-fréquences. Notons que le système de détection

Table A.1.7 Caractéristiques des classes phonétiques « *consonne* » fournies par les transcriptions du corpus Berlin.

Phonème	Classe	Effectif	Durée en *ms*		Phonème	Classe	Effectif	Durée en *ms*	
			μ	σ				μ	σ
!	(erreur) = !b, …	6	16.2	8.0	**tasp**	?plosive?	361	32.1	31.8
!b	plosive	301	11.5	9.5	**f**	fricative	301	44.6	31.4
!d	plosive	874	9.6	5.8	**h**	fricative	37	40.2	25.4
!g	plosive	383	13.9	12.5	**j**	fricative	88	29.6	13.1
!k	plosive	266	20.9	21.7	**s**	fricative	867	52.4	38.4
!p	plosive	299	12.9	9.0	**v**	fricative	640	43.7	20.4
!t	plosive	1061	15.0	14.3	**x**	fricative	242	57.7	30.1
?	plosive	76	24.2	21.8	**C**	fricative	217	38.5	28.6
b	plosive	477	33.9	20.4	**4**	?fricative?	176	44.4	20.5
d	plosive	770	26.7	17.1	**8**	?fricative?	2	84.2	26.4
g	plosive	488	32.2	19.1	**ss**	?fricative?	1	20.7	0
k	plosive	286	48.4	34.4	**vv**	?fricative?	1	178.2	0
p	plosive	306	46.4	31.8	**z**	?fricative?	399	60.7	26.1
t	plosive	1275	31.8	20.4	**H**	?fricative?	51	56.4	19.6
B	?plosive?	84	38.3	16.4	**S**	?fricative?	288	75.0	53.5
D	?plosive?	53	24.7	9.7	**X**	?fricative?	149	51.2	22.7
G	?plosive?	12	35.2	13.0	**l**	sonorante	442	54.3	24.9
J	?plosive?	1	19.1	0	**m**	sonorante	917	78.5	40.0
V	?plosive?	83	41.1	13.8	**n**	sonorante	1549	68.6	36.6
basp	?plosive?	14	16.3	5.4	**N**	sonorante	406	96.5	50.8
dasp	?plosive?	153	14.0	7.0	**R**	sonorante	26	57.1	19.0
dsap	(erreur) = dasp	1	22.9	0	**mm**	?sonorante?	1	107.4	0
gas	(erreur) = gasp	1	12.9	0	**nn**	?sonorante?	5	107.9	24.7
gasp	?plosive?	69	17.6	11.6	**L**	?sonorante?	28	30.8	15.7
kasp	?plosive?	174	62.4	40.0	**NN**	?sonorante?	5	140.7	68.0
pasp	?plosive?	159	32.0	21.1					

Table A.1.8 Caractéristiques des classes phonétiques fournies par les transcriptions du corpus Berlin.

Groupe ortho	Classe phonétique	Effectif	Durée en *ms*	
			mean	std
Voyelle	**Courte**	5788	68.2	38.2
	Longue	392	97.9	59.7
	Diphtongue du schwa	409	89.9	50.8
Consonne	**Plosive**	8033	25.5	23.1
	Fricative	3459	51.4	33.6
	Sonorante	3379	72.5	40.0

de pseudo-phonèmes pourrait être utilisé afin d'accomplir cette tâche. Mais seules les voyelles sont fournies par un détecteur spécifique dans ce système. Nous avons donc préféré une analyse manuelle des labels, plutôt qu'automatique, pour identifier leur macro-classe phonétique la plus probable.

Après cette première étape, nous avons tenté de séparer les labels selon leur hypothétique classe phonétique. Nous avons ainsi utilisé les propriétés acoustiques des segments pour iden-

tifier le type d'informations qu'ils véhiculent, e.g., consonnes : (i) *plosive* : bruit impulsionnel, (ii) *fricative* : durée plus longue que les plosives et spectre plat dans les hautes-fréquences et (iii) *sonorante* : spectre à tendance formantique dû au voisement partiel du phonème. Concernant les voyelles, nous avons considéré que toutes celles qui s'écrivaient avec une seule lettre font partie de la catégorie des voyelles *courtes* ; ce qui est le cas des données issues de l'alphabet SAMPA. Les autres classes sont définies par la présence ou non du schwa /@/ (ou /6/ pour la version voisée du /r/ fusionnée avec le schwa) dans la diphtongue.

Les tables A.1.6 et A.1.7 donnent les listes de tous les paramètres issus des transcriptions du corpus Berlin pour les voyelles et les consonnes, respectivement. Les classes des phonèmes qui ne sont pas identifiées par l'alphabet SAMPA sont données avec un point d'interrogation pour indiquer le fait que les associations proposées ne sont pas certaines. Une analyse statistique de la durée des classes phonétiques montre que les données ne varient que très peu par rapport à l'analyse précédente, i.e., celle qui repose sur les transcriptions qui ont été identifiées par l'alphabet SAMPA, cf., table A.1.8. Nous pouvons donc considérer que les hypothèses qui ont été prises pour le regroupement des phonèmes non identifiés par SAMPA sont correctes au niveau des caractéristiques de durée.

3. Cas de figures particuliers

Outre le fait que très peu de phonèmes sont identifiés par l'alphabet SAMPA, d'autres problèmes ont également été rencontrés dans les transcriptions. Tout d'abord, bon nombre de segments ne possèdent pas de labels. Ce type de cas est problématique puisque aucun élément ne nous permet de savoir comment traiter ces segments. Faut-il, par exemple, (i) les considérer comme du silence ?, (ii) les fusionner avec le segment précédent ? ou (iii) recopier le label du segment précédent ? Ainsi, il n'y a pas d'idée fixe pour traiter le cas de l'astérisque : certains labels correspondent tantôt à du silence, tantôt à de la parole, cf. Fig. A.1.1. Nous avons aussi noté des cas de séparation à l'intérieur d'une diphtongue, cf. Fig. A.1.2. De plus les labels identifiés par l'alphabet SAMPA ne sont pas toujours homogènes : un même phonème peut par exemple correspondre à l'attaque ou au noyau d'une syllabe, cf. Fig. A.1.3.

4. Reconnaissance acoustique des émotions selon les groupes

Les sections précédentes ont montré que de nombreux labels présents dans les transcriptions du corpus Berlin ne sont pas identifiés par l'alphabet SAMPA. Toutefois, certaines considérations ont permis de définir de façon *a priori* les classes et macro-classes phonétiques associées aux labels manquants dans l'alphabet SAMPA. Une analyse statistique a montré que la durée des classes phonétiques a été conservée par rapport au premier jeu de données, i.e., « phonèmes identifiés par SAMPA » *vs.* « toutes les transcriptions ».

Cette dernière section vise à estimer l'impact qu'ont ces deux groupes de données sur les scores en reconnaissance d'émotions. Nous avons pour cela défini un système de reconnaissance qui correspond à celui de état-de-l'art: nous regroupons les caractéristiques acoustiques (16 MFCC calculés toutes les 16 ms et sur une fenêtre de 32 ms) et les mesures statistiques (e.g., valeur max, min, moyenne et écart-type) des LLDs prosodiques tels que le pitch,

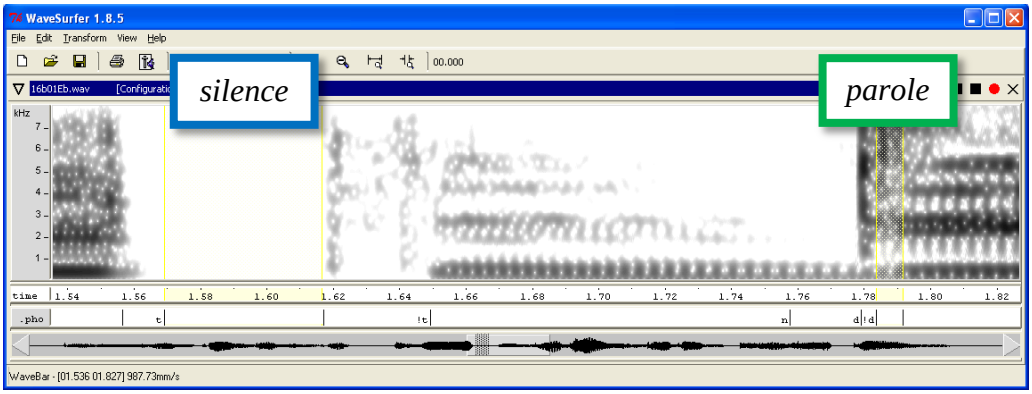


Figure A.1.1 Illustration de la variabilité des segments non labélisés, e.g. *silence* vs. *parole*.

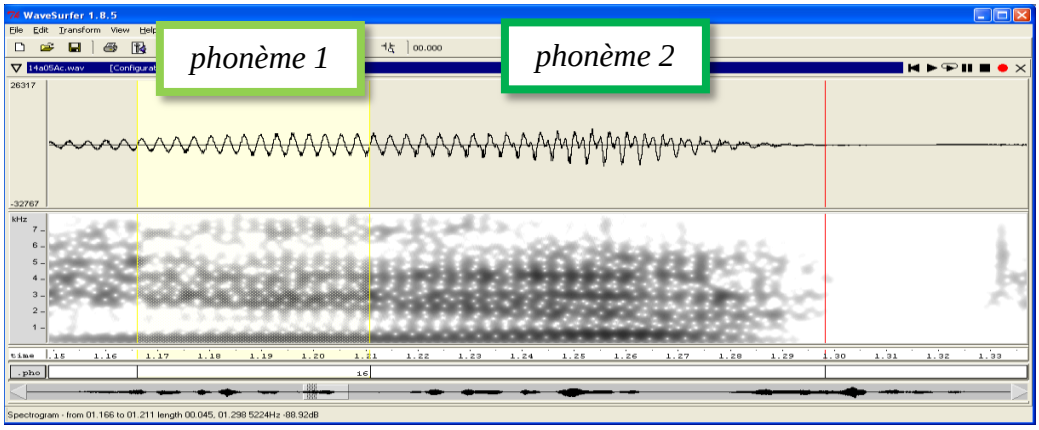


Figure A.1.2 Illustration de la variabilité des segments non labélisés, e.g. deuxième phonème d'une diphtongue.

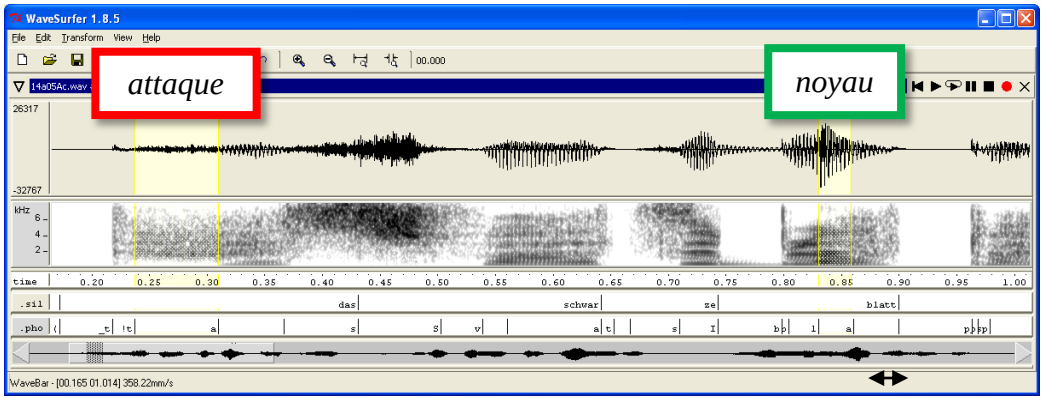


Figure A.1.3 Illustration de la variabilité du phonème /*a*/, e.g. *attaque* vs. *noyau*.

l'énergie et la durée des segments voisés. Nous avons ensuite utilisé un classifieur *k*-ppv (*k* = 5) pour attribuer une étiquette d'émotion au super-vecteur qui est composé en tout de 28 paramètres (16 acoustiques et 12 prosodiques). La décision était de type « *segmentale* » et dans un schéma de *cross-validation* (cf. chapitre 3). La table A.1.9 donne les résultats obtenus par ce système selon les classes phonétiques et les groupes de transcriptions : SAMPA (phonèmes identifiés par cet alphabet) et POST (phonèmes incluant aussi ceux issus qui ne sont par identifiés par SAMPA, cf. section 2). Ces résultats montrent que l'amélioration des scores de re-

Table A.1.9 Comparaison des scores en reconnaissance acoustique d'émotions selon les classes phonétiques et les groupes de transcription.

MFCC Raw	Classe phonétique	Groupe SAMPA	Groupe POST
Voyelle	**Courte**	70.0	70.4
	Longue	56.5	50.0
	Diphtongue du schwa	52.0	48.6
	Toutes	**73.4**	72.8
Consonne	**Plosive**	53.6	58.4
	Fricative	51.8	51.4
	Sonorante	73.6	**76.2**
	Toutes	68.2	68.4

connaissance n'est pas significative lorsque l'on inclut les transcriptions absentes du premier groupe dans le second. Raison pour laquelle nous avons préféré conserver les données issues du groupe SAMPA pour effectuer les expériences en reconnaissance d'émotions dans les chapitres précédents. De plus, nous ne sommes pas parfaitement certains que les classes et macro-classes que nous avons attribuées aux transcriptions absentes de l'alphabet SAMPA soient tout à fait correctes.

Annexe 2

Tables du chapitre 2

Cette annexe contient les tables associées aux résultats des expériences portant sur la caractérisation des corrélats phonétiques associés aux ancrages rythmiques de la parole, cf. chapitre 2.

Table A.2.1.1 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Berlin ; niveau de **perception 1 (seuil 1/3)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	54 (20) / 71 (18)	13 (11) / 14 (11)	61 (14)	31 (14)	64 (23)	4 (6)
Peur	39 (19) / 57 (18)	**24 (13) / 25 (15)**	58 (17)	**32 (17)**	66 (23)	**7 (11)**
Joie	50 (18) / 71 (15)	13 (10) / 12 (10)	63 (14)	29 (15)	68 (24)	3 (6)
Tristesse	35 (15) / 61 (16)	27 (16) / 30 (16)	64 (12)	30 (12)	74 (17)	9 (8)
Dégoût	50 (18) / 69 (16)	16 (10) / 17 (12)	66 (13)	27 (13)	78 (16)	3 (4)
Ennui	50 (23) / 75 (16)	16 (14) / 16 (15)	70 (14)	22 (14)	73 (23)	1 (3)
Neutre	49 (16) / 64 (16)	24 (14) / 22 (14)	66 (14)	27 (13)	75 (20)	5 (6)

[valeur moyenne] (écart-type) ; les deux valeurs données pour les ancrages issus des transcriptions phonétiques correspondent respectivement à une configuration correspondant soit aux phonèmes identifiés par l'alphabet SAMPA soit tous les phonèmes contenus dans les transcriptions.

Table A.2.1.2 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Berlin ; niveau de **perception 2 (seuil 1/4)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	52 (16) / 71 (15)	17 (9) / 17 (10)	60 (12)	35 (12)	70 (19)	4 (6)
Peur	38 (13) / 57 (14)	**28 (12) / 30 (13)**	57 (15)	**36 (13)**	69 (21)	**10 (10)**
Joie	47 (16) / 70 (14)	17 (10) / 17 (9)	62 (12)	32 (13)	72 (19)	4 (5)
Tristesse	32 (11) / 56 (12)	33 (12) / 36 (13)	61 (11)	35 (10)	76 (11)	13 (9)
Dégoût	47 (16) / 69 (14)	22 (10) / 22 (11)	63 (11)	33 (10)	79 (14)	5 (5)
Ennui	45 (18) / 69 (15)	21 (14) / 21 (14)	69 (13)	26 (12)	80 (17)	2 (3)
Neutre	40 (10) / 60 (12)	31 (13) / 29 (13)	63 (11)	32 (11)	76 (16)	7 (6)

[valeur moyenne] (écart-type).

Table A.2.1.3 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Berlin ; niveau de **perception 3 (seuil 1/6)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	47 (13) / 67 (13)	22 (8) / 22 (9)	58 (9)	39 (9)	73 (16)	6 (6)
Peur	35 (10) / 56 (9)	**32 (9) / 35 (10)**	56 (12)	**40 (11)**	70 (16)	**13 (9)**
Joie	43 (12) / 67 (11)	24 (10) / 24 (10)	59 (10)	37 (9)	74 (16)	6 (5)
Tristesse	28 (8) / 50 (9)	37 (8) / 43 (9)	55 (9)	41 (8)	75 (11)	20 (9)
Dégoût	40 (12) / 63 (11)	28 (9) / 30 (10)	58 (9)	40 (8)	79 (14)	9 (7)
Ennui	39 (11) / 64 (13)	27 (10) / 29 (12)	63 (11)	33 (11)	79 (17)	5 (5)
Neutre	36 (8) / 55 (10)	36 (8) / 35 (9)	58 (9)	38 (9)	76 (15)	10 (7)

[valeur moyenne] (écart-type).

Table A.2.2.1 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Bute-TMI ; niveau de **perception 1 (seuil 1/3)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	71 (14)	21 (12)	58 (17)	33 (18)	80 (19)	4 (5)
Peur	66 (15)	**27 (14)**	62 (18)	**28 (15)**	76 (21)	**5 (11)**
Joie	68 (16)	25 (14)	57 (18)	33 (98)	79 (23)	3 (4)
Tristesse	69 (16)	22 (13)	61 (16)	29 (16)	74 (28)	3 (7)
Dégoût	70 (17)	17 (11)	61 (19)	30 (18)	71 (33)	3 (6)
Surprise	64 (16)	26 (12)	16 (16)	31 (15)	75 (25)	3 (6)
Nerveux	65 (11)	29 (12)	58 (17)	34 (18)	77 (23)	5 (8)
Neutre	66 (18)	24 (12)	60 (18)	31 (17)	74 (29)	2 (6)

[valeur moyenne] (écart-type).

Table A.2.2.2 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Bute-TMI ; niveau de **perception 2 (seuil 1/4)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	69 (13)	25 (10)	58 (14)	36 (15)	80 (19)	5 (7)
Peur	62 (14)	**32 (13)**	62 (13)	**33 (14)**	80 (14)	**7 (11)**
Joie	66 (14)	29 (12)	59 (14)	35 (14)	83 (16)	4 (5)
Tristesse	69 (13)	25 (11)	61 (14)	32 (13)	82 (19)	4 (8)
Dégoût	72 (15)	22 (11)	59 (15)	34 (15)	81 (23)	4 (8)
Surprise	64 (12)	29 (11)	60 (12)	35 (13)	82 (17)	4 (7)
Nerveux	62 (10)	34 (10)	59 (15)	36 (15)	81 (17)	7 (10)
Neutre	67 (11)	28 (10)	60 (14)	33 (14)	83 (15)	4 (6)

[valeur moyenne] (écart-type).

Table A.2.2.3 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Bute-TMI ; niveau de **perception 3 (seuil 1/6)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	65 (10)	31 (10)	55 (11)	40 (11)	82 (11)	8 (8)
Peur	59 (13)	**36 (12)**	59(11)	**36 (11)**	80 (14)	**9 (11)**
Joie	63 (11)	33 (11)	58 (12)	37 (12)	81 (13)	7 (6)
Tristesse	65 (11)	31 (10)	60 (13)	35 (13)	84 (12)	6 (10)
Dégoût	68 (13)	27 (11)	59 (13)	37 (13)	81 (20)	8 (10)
Surprise	61 (11)	34 (10)	58 (10)	37 (10)	81 (14)	7 (8)
Nerveux	57 (10)	39 (9)	57 (11)	38 (12)	79 (15)	10 (10)
Neutre	62 (10)	34 (10)	59 (11)	35 (11)	82 (14)	7 (9)

[valeur moyenne] (écart-type).

Table A.2.3.1 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Aholab ; niveau de **perception 1 (seuil 1/3)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	64 (9)	30 (9)	75 (11)	20 (11)	78 (24)	1 (1)
Peur	54 (10)	**41 (10)**	71 (10)	**25 (10)**	83 (17)	**3 (4)**
Joie	67 (11)	25 (9)	76 (11)	19 (9)	78 (19)	1 (2)
Tristesse	63 (10)	32 (10)	72 (10)	23 (10)	83 (19)	2 (2)
Dégoût	69 (10)	25 (10)	69 (13)	26 (12)	82 (21)	1 (3)
Surprise	62 (11)	29 (10)	75 (11)	19 (10)	76 (21)	1 (1)
Neutre	69 (10)	26 (9)	74 (12)	21 (10)	82 (20)	1 (1)

[valeur moyenne] (écart-type).

Table A.2.3.2 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Aholab ; niveau de **perception 2 (seuil 1/4)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	63 (8)	32 (8)	73 (9)	23 (9)	80 (22)	1 (1)
Peur	52 (8)	**44 (8)**	70 (8)	**26 (8)**	84 (16)	**5 (4)**
Joie	66 (9)	28 (8)	75 (9)	22 (8)	80 (18)	2 (2)
Tristesse	60 (8)	36 (8)	71 (8)	25 (8)	85 (16)	2 (2)
Dégoût	67 (9)	28 (8)	68 (10)	29 (10)	86 (17)	2 (3)
Surprise	61 (9)	32 (8)	75 (9)	22 (9)	79 (19)	1 (1)
Neutre	65 (9)	30 (8)	73 (10)	23 (9)	83 (19)	1 (1)

[valeur moyenne] (écart-type).

Table A.2.3.3 Taux de recouvrement des « *p-centres* » en % avec les ancrages acoustiques de la parole selon les styles de production du corpus Aholab ; niveau de **perception 3 (seuil 1/6)**.

Emotion	V_{REF}	C_{REF}	V_{DET}	C_{DET}	S_{VOI}	S_{NVO}
Colère	61 (7)	34 (6)	70 (8)	28 (8)	81 (21)	2 (2)
Peur	51 (7)	**45 (7)**	67 (7)	**30 (7)**	83 (13)	**7 (4)**
Joie	63 (8)	31 (7)	71 (7)	26 (7)	81 (16)	3 (2)
Tristesse	57 (7)	40 (7)	69 (7)	29 (7)	85 (15)	3 (3)
Dégoût	63 (8)	34 (7)	64 (8)	34 (9)	88 (13)	4 (4)
Surprise	59 (8)	35 (7)	72 (7)	26 (7)	80 (19)	1 (1)
Neutre	60 (7)	36 (7)	71 (8)	27 (7)	82 (18)	2 (2)

[valeur moyenne] (écart-type).

Annexe 3

Tables du chapitre 4

Nous présentons dans cette annexe les tables qui accompagnent la description des résultats issus des expériences du chapitre 4 : reconnaissance prosodique de la parole affective actée.

Table A.4.1.1 Scores en reconnaissance d'émotions obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**70** 7/3	**75** 7/3	**76** 7/3	**78** 9/1	**74** 8/2
voyelle / consonne	64 6/4	**71** 4/6	**70** 8/2	**70** 6/4	**70** 6/4
p-voyelle / p-consonne	65 5/5	**70** 6/4	**75** 4/6	**78** 4/6	**71** 1/9
phonèmes voyelles	59 9/0/1	54 8/1/2	62 8/0/1	**70** 9/0/1	56 5/0/5
phonèmes consonnes	61 1/3/6	65 3/3/4	67 3/0/6	68 3/3/4	69 0/2/7
Phonèmes	65 3/7	67 1/9	69 1/9	**75** 4/6	**70** 2/8
syllabes voisées / non-voisées	**71** 7/3	**71** 2/8	**74** 7/3	**77** 4/6	**73** 6/4
syllabes V/C/CV/VC/CVC	60 0/0/1/0/9	56 0/0/3/1/6	55 1/0/5/3/2	58 1/0/1/0/8	64 2/0/0/2/6

[valeur en %] $_{\text{poids}}\,\alpha_a$: acoustique / prosodique ; approche « *composante* » ; méthode CVS.

Table A.4.1.2 Comparaison des scores en reconnaissance d'émotions selon les ancrages acoustiques complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	85	72	63	67	76	75	**95**
voyelle / consonne	72	74	63	54	58	**84**	79
p-voyelle / p-consonne	79	**78**	**70**	**80**	**79**	77	82
phonèmes voyelles	79	74	59	70	61	66	77
phonèmes consonnes	77	59	49	**80**	58	63	94
phonèmes	**89**	76	72	67	72	65	89
syllabes voisées / non-voisées	84	72	60	**80**	65	71	90
syllabes V/C/CV/VC/CVC	72	81	52	0	63	53	55
« p-centres » niveau 3	84	65	42	72	56	66	92

Fusion acoustique / prosodique ; approche « *composante* » ; méthode CVS.

Table A.4.2.1 Scores en reconnaissance d'émotions obtenus par la fusion des ancrages acoustiques complémentaires.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	**73** 6/4	**72** 6/4	**77** 7/3	**78** 9/1	**77** 6/4
voyelle / consonne	67 2/8	67 8/2	**72** 8/2	**77** 7/3	**74** 4/6
p-voyelle / p-consonne	65 2/8	69 5/5	**75** 5/5	**80** 5/5	**74** 5/5
phonèmes voyelles	55 9/0/1	60 1/0/0	66 1/0/0	**71** 1/0/0	66 9/0/1
phonèmes consonnes	63 3/2/5	67 0/6/4	69 2/2/6	**72** 2/2/6	69 1/1/8
Phonèmes	63 0/1	**70** 3/7	**72** 5/6	**76** 3/7	**72** 3/7
syllabes voisées / non-voisées	**70** 2/8	**75** 4/6	**75** 7/3	**78** 4/6	**74** 6/4
syllabes V/C/CV/VC/CVC	62 1/0/2/0/8	62 1/0/1/0/7	63 1/0/1/1/6	68 1/0/0/2/7	67 1/0/2/1/6

[valeur en %] $_{\text{poids}}\,\alpha_a$: acoustique / prosodique ; approche « *globale* » ; méthode CVS.

Table A.4.2.2 Comparaison des scores en reconnaissance d'émotions selon les paires d'ancrages complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	83	**85**	**74**	76	54	77	**95**
voyelle / consonne	81	82	72	61	70	77	92
p-voyelle / p-consonne	**86**	74	73	**80**	**74**	76	91
phonèmes voyelles	73	76	68	64	58	68	85
phonèmes consonnes	70	76	53	78	59	78	95
phonèmes	80	85	63	76	58	**80**	92
syllabes voisées / non-voisées	87	81	73	63	62	72	92
syllabes V/C/CV/VC/CVC	80	75	35	**80**	49	65	92
« p-centres » niveau 3	85	63	60	74	58	68	89

Fusion acoustique / prosodique ; approche « *globale* » ; méthode CVS.

Table A.4.3.1 Scores en reconnaissance d'émotions obtenus par la fusion des ancrages.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	60 $_{6/4}$	65 $_{6/4}$	59 $_{4/6}$	**70 $_{1/0}$**	66 $_{6/4}$
voyelle / consonne	61 $_{6/4}$	66 $_{2/8}$	**70 $_{9/1}$**	**71 $_{2/8}$**	67 $_{0/1}$
p-voyelle / p-consonne	68 $_{8/2}$	67 $_{6/4}$	**71 $_{1/0}$**	**74 $_{1/0}$**	66 $_{0/1}$
phonèmes voyelles	55 $_{9/1/0}$	55 $_{7/1/2}$	49 $_{8/2/1}$	61 $_{1/0/0}$	59 $_{9/1/0}$
phonèmes consonnes	55 $_{3/1/7}$	57 $_{1/0/9}$	60 $_{0/4/6}$	64 $_{1/0/0}$	62 $_{1/1/9}$
phonèmes	57 $_{9/2}$	60 $_{4/6}$	60 $_{0/1}$	67 $_{3/7}$	65 $_{2/8}$
syllabes voisées / non-voisées	63 $_{8/2}$	66 $_{1/0}$	68 $_{1/0}$	**71 $_{9/1}$**	67 $_{5/5}$
syllabes V/C/CV/VC/CVC	57 $_{1/0/0/0/9}$	53 $_{4/0/2/0/4}$	56 $_{0/3/0/0/7}$	59 $_{3/2/1/0/5}$	61 $_{0/3/1/0/6}$

Fusion acoustique / prosodique ; approche « *composante* » ; méthode LOSO ; [valeur en %] $_{\text{poids}}$ α_a.

Table A.4.3.2 Comparaison des scores en reconnaissance d'émotions selon les paires d'ancrages complémentaires.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	91	72	69	33	**37**	82	77
voyelle / consonne	96	62	78	**43**	28	80	81
p-voyelle / p-consonne	97	71	79	39	**37**	86	77
phonèmes voyelles	94	47	69	2	23	77	63
phonèmes consonnes	96	56	79	17	17	72	65
phonèmes	**98**	65	**81**	11	20	82	66
syllabes voisées / non-voisées	96	**74**	74	7	32	89	84
syllabes V/C/CV/VC/CVC	**98**	37	67	17	4	61	**85**
« p-centres » niveau 3	92	69	75	9	13	**92**	74

Fusion acoustique / prosodique ; approche « *composante* » ; méthode LOSO.

Table A.4.4.1 Scores en reconnaissance d'émotions obtenus par la fusion des ancrages.

Fusion des ancrages	Raw	Z-tout	Z-genre	Z-locuteur	Z-phrase
voisé / non-voisé	59 6/4	69 3/7	**71 8/2**	**73 1/0**	**70 4/6**
voyelle / consonne	62 2/8	68 8/2	**72 1/9**	**77 3/7**	**70 4/6**
p-voyelle / p-consonne	64 6/4	**72 9/1**	**74 9/1**	**78 9/1**	**74 6/4**
phonèmes voyelles	52 9/1/0	60 1/0/0	65 8/1/2	68 1/0/0	60 1/0/0
phonèmes consonnes	56 1/2/6	63 0/0/1	64 0//01	67 0/1/9	62 1/0/0
phonèmes	57 5/5	64 5/5	68 8/2	**72 2/8**	65 1/9
syllabes voisées / non-voisées	65 4/6	**70 1/0**	**72 1/0**	**72 8/2**	**74 1/0**
syllabes V/C/CV/VC/CVC	56 1/0/4/0/5	62 1/1/1/0/8	64 1/0/2/0/7	62 2/0/0/0/8	62 0/1/0/0/9

Fusion acoustique / prosodique ; approche « *globale* » ; méthode LOSO ; [valeur en %] $_{\text{poids}}$ α_a.

Table A.4.4.2 Comparaison des scores en reconnaissance d'émotions selon les paires d'ancrages complémentaires ; fusion acoustique / prosodique ; approche « *globale* » ; méthode LOSO.

Fusion des ancrages	Colère	Peur	Ennui	Dégoût	Joie	Neutre	Tristesse
voisé / non-voisé	96	**76**	70	46	25	82	87
voyelle / consonne	91	69	81	**57**	**41**	**91**	85
p-voyelle / p-consonne	91	69	**85**	50	52	89	82
phonèmes voyelles	91	56	83	22	25	81	81
phonèmes consonnes	96	51	74	41	15	68	**89**
phonèmes	95	68	**85**	35	21	80	87
syllabes voisées / non-voisées	97	71	80	39	15	85	87
syllabes V/C/CV/VC/CVC	**98**	43	69	39	10	62	82
« p-centres » niveau 3	**98**	43	69	39	10	62	82

Annexe 4

Annexe du chapitre 5

Cette annexe présente l'algorithme qui a été utilisé dans le chapitre 5 pour corriger les erreurs d'estimation de la f0 sur les voix d'enfants. Nous présentons ensuite les tables et les figures qui accompagnent la description des résultats fournis par l'épreuve de *production de parole affective spontanée*.

1. Description du filtre anti saut d'octave

Ce filtre a été développé par György Szaszák, laboratoire LSA, BUTE-TMIT, Budapest, Hongrie, dans le cadre de nos collaborations pour l'étude des troubles de la communication (TC). Le filtrage proposé pour supprimer les sauts d'octave pouvant être présents dans la f0 s'effectue en plusieurs étapes. Le format des données traitées par le filtre correspond à celui fournis par l'algorithme Snack pour la f0 (i.e., estimée toutes les 10 ms). Un segment voisé doit être composé d'au minimum 3 trames voisées pour être traité par le système. La méthode de filtrage est de nature heuristique et se compose de trois passes successives qui sont décrites ci-dessous. La Fig. 5.7 présente un exemple des résultats obtenus par ce filtre sur une phrase produite par un enfant.

- **Etape 1 (Initialisation de la première passe)**
  But : obtenir une référence.
  La première passe consiste à extraire des mesures statistiques sur les données de la f0 de façon à définir des références. Ainsi, le minimum et le maximum global sont tout d'abord estimés sur la f0 avec la valeur moyenne issue des 25 premières trames voisées qui servira de *référence* initiale pour l'algorithme.

- **Etape 2 (Première passe)**
  But : vérification trame-par-trame.
  Les régions non-voisées (i.e., f0 = 0) sont ignorées. Chaque valeur de la f0 est analysée pour vérifier si le double est supérieur au maximum global ou si la moitié est inférieure au minimum global. Si tel est le cas, la valeur de f0 est acceptée et aucune vérification supplémentaire n'est effectuée. Sinon, une seconde passe est effectuée pour tester l'hypothèse d'une erreur de facteur ½ ou 2 de la valeur :

  - **Etape 2.1**
    But : vérification d'une erreur de type facteur ½.
    - si $f0 < C_1 * référence$, alors une erreur de type ½ est supposée présente et la valeur *f0* est donc multipliée par 2 ; $C_1 = 0.55$.
    - si $C_1 * référence < f0 < C_2 * référence$, alors une erreur de type ½ est probable et nécessite une étape de vérification supplémentaire ; $C_2 = 0.65$:
      - la valeur moyenne *M* de la f0 est calculée sur les cinq trames voisées suivant celle testée, ensuite, si $M \geq C_3 * référence$, la valeur initiale *f0* est doublée, sinon elle reste inchangée. Elle reste également inchangée si *M* n'est pas calculable (e.g., dernière trame voisée) ; $C_3 = 0.60$.

  - **Etape 2.2**
    But : vérification d'une erreur de type facteur 2.
    - si $f0 > C_4 * reference$, alors une erreur de type doublement est probable et une seconde étape de vérification est requise avant de changer la valeur *f0* ; $C_4 = 1.80$.

- la valeur moyenne *M* de la f0 est calculée sur les cinq trames voisées suivant celle testée, ensuite, si $M \leq C_5$**référence*, la valeur initiale *f0* est divisée par 2, sinon elle reste inchangée ; de même si *M* n'est pas calculable ; $C_5 = 1.80$.

- **Etape 3 (Deuxième passe)**
  But : mise à jour de la référence puis étape 2.
  La valeur *référence* est de nouveau calculée comme indiqué dans l'étape 1. Cependant, cette valeur est maintenant estimée sur les cinq premières trames contenant des valeurs fiables de la f0, i.e., celles qui n'ont pas conduit à un test supplémentaire lors de l'étape 2. Cette deuxième étape est ensuite reproduite avec la nouvelle valeur *référence*.

- **Etape 4 (Troisième passe)**
  But : vérification segment-par-segment.
  Des erreurs locales dans l'estimation de la f0 et entraînant des pics dans sa courbe, ne peuvent être corrigées lors des deux précédentes passes. Afin de vérifier la continuité de la f0 dans les segments voisés, une dernière passe a été effectuée pour chaque segment sur les données : si une valeur $f_0(i)$ montre une variation abrupte avec la trame suivante, i.e., si $|f_0(i)/f_0(i+1)| > C_6$, toutes les valeurs délimitées par le changement abrupte sont modifiées en fonction du type d'erreur détecté, i.e., de type facteur ½ ou 2 par rapport à la *référence*.

2. Tables et figures en annexe du chapitre 5

Table A.5.1 Analyse statistique des modèles *conventionnels* et *non-conventionnels* du rythme selon les classes d'émotions spontanées.

Modèles du rythme	***Négative***	***Neutre***	***Positive***
%V	39 $_{4.7}$	42 $_{4.2}$	49 $_{8.9}$
ΔC (ms)	106 $_{27}$	134 $_{34}$	144 $_{51}$
Varco	43 $_{11}$	53 $_{13}$	41 $_{6.5}$
$\overline{R}$ (.10^{-2})	30 $_{18}$	24 $_{16}$	33 $_{12}$
RR (.10^{-1})	12 $_{9.3}$	14 $_{13}$	12 $_{6.6}$
***r*PVI** (.10^{-2})	11 $_{9.8}$	13 $_{13}$	11 $_{8.2}$
***n*PVI** (.10^{-2})	47 $_{33}$	53 $_{35}$	46 $_{28}$
F_{moy} **« *p-centre* »** (Hertz .10^{-1})	63 $_{14}$	59 $_{6.2}$	64 $_{10}$
***MNF* de la THH** (Hertz .10^{-1})	84 $_{15}$	74 $_{8.0}$	86 $_{4.8}$
PHD (pitch) (.10^{-3})	12 $_{15}$	13 $_{12}$	13 $_{19}$
PHD (énergie) (.10^{-3})	6.5 $_{8.1}$	6.9 $_{7.9}$	12 $_{24}$
PHD (qualité vocale) (.10^{-3})	12 $_{8.1}$	13 $_{7.6}$	16 $_{14}$
A-PHD (toutes) (.10^{-3})	31 $_{25}$	33 $_{24}$	41 $_{49}$
I-PHD (toutes) (.10^{-2})	12 $_{11}$	13 $_{12}$	12 $_{10}$

[Valeur moyenne] $_{\text{écart-type}}$; Ancrage des pseudo-voyelles sauf pour la mesure ΔC qui est calculée sur les pseudo-consonnes ; * : $p < 0.05$; Groupe des sujets DT.

Table A.5.2 ; Groupe des sujets TA

Modèles du rythme	***Négative***	***Neutre***	***Positive***
%V	42 $_{6.7}$	44 $_{4.7}$	41 $_{12}$
ΔC (ms)	114 $_{33}$	129 $_{34}$	139 $_{69}$
Varco	48 $_{14}$	48 $_{9.2}$	42 $_{16}$
$\overline{R}$ (.10^{-2})	36 $_{15}$	37 $_{16}$	44 $_{24}$
RR (.10^{-1})	13 $_{11}$	13 $_{9.6}$	13 $_{11}$
***r*PVI** (.10^{-2})	13 $_{10}$	13 $_{11}$	14 $_{12}$
***n*PVI** (.10^{-2})	55 $_{35}$	54 $_{34}$	57 $_{33}$
F_{moy} **« *p-centre* »** (Hertz .10^{-1})	75 $_{12}$	72 $_{19}$	68 $_{13}$
***MNF* de la THH** (Hertz .10^{-1})	81 $_{13}$	82 $_{18}$	89 $_{8.9}$
PHD (pitch) (.10^{-3})	14 $_{15}$	14 $_{20}$	17 $_{19}$
PHD (énergie) (.10^{-4})	80 $_{98}$	77 $_{100}$	88 $_{96}$
PHD (qualité vocale) (.10^{-3})	12 $_{7.4}$	13 $_{9.3}$	13 $_{7.9}$
A-PHD (toutes) (.10^{-3})	35 $_{28}$	35 $_{29}$	41 $_{37}$
I-PHD (toutes) (.10^{-2})	12 $_{15}$	13 $_{19}$	16 $_{24}$

Table A.5.3 ; Groupe des sujets TED-NOS.

Modèles du rythme	***Négative***	***Neutre***	***Positive***
%V	45 $_{8.8}$	41 $_{7.1}$	49 $_{9.8}$
ΔC (ms)	130 $_{66}$	109 $_{34}$	162 $_{69}$
Varco	47 $_{19}$	52 $_{19}$	62 $_{21}$
$\overline{R}$ (.10^{-2})	35 $_{23}$	45 $_{20}$	41 $_{19}$
RR (.10^{-1})	13 $_{11}$	13 $_{9.6}$	17 $_{21}$
***r*PVI** (.10^{-2})	13 $_{15}$	12 $_{13}$	16 $_{17}$
***n*PVI** (.10^{-2})	49 $_{37}$	49 $_{36}$	63 $_{46}$
F_{moy} **« *p-centre* »** (Hertz .10^{-1})	68 $_{22}$	72 $_{15}$	64 $_{11}$
***MNF* de la THH** (Hertz .10^{-1})	83 $_{15}$	68 $_{12}$	88 $_{12}$
PHD (pitch) (.10^{-3})	15 $_{17}$	14 $_{16}$	11 $_{13}$
PHD (énergie) (.10^{-4})	85 $_{114}$	78 $_{98}$	71 $_{105}$
PHD (qualité vocale) (.10^{-4})	12 $_{9.1}$	11 $_{8.0}$	10 $_{7.3}$
A-PHD (toutes) (.10^{-3})	35 $_{31}$	38 $_{37}$	32 $_{29}$
I-PHD (toutes) (.10^{-2})	14 $_{21}$	12 $_{12}$	11 $_{13}$

Table A.5.4 ; Groupe des sujets TSL.

Modèles du rythme	***Négative***	***Neutre***	***Positive***
%V	40 $_{5.7}$	40 $_{5.7}$	41 $_{5.0}$
ΔC (ms)	106 $_{42}$	95 $_{31}$	93 $_{27}$
Varco	48 $_{12}$	42 $_{15}$	44 $_{15}$
$\overline{R}$ (.10^{-2})	30 $_{21}$	26 $_{19}$	39 $_{17}$
RR (.10^{-1})	12 $_{9.6}$	13 $_{8.8}$	12 $_{7.7}$
***r*PVI** (.10^{-2})	10 $_{9.6}$	11 $_{9.3}$	9.6 $_{9.6}$
***n*PVI** (.10^{-2})	47 $_{36}$	48 $_{34}$	45 $_{34}$
F_{moy} **« *p-centre* »** (Hertz .10^{-1})	69 $_{13}$	77 $_{17}$	80 $_{23}$
***MNF* de la THH** (Hertz.10^{-1})	75 $_{11}$	75 $_{19}$	77 $_{13}$
PHD (pitch) (.10^{-3})	9.2 $_{11}$	8.9 $_{10}$	11 $_{12}$
PHD (énergie) (.10^{-4})	59 $_{88}$	68 $_{97}$	59 $_{77}$
PHD (qualité vocale) (.10^{-4})	99 $_{71}$	102 $_{69}$	90 $_{45}$
A-PHD (toutes) (.10^{-3})	26 $_{24}$	27 $_{22}$	26 $_{24}$
I-PHD (toutes) (.10^{-2})	12 $_{37}$	8.7 $_{8.1}$	9.2 $_{9.3}$

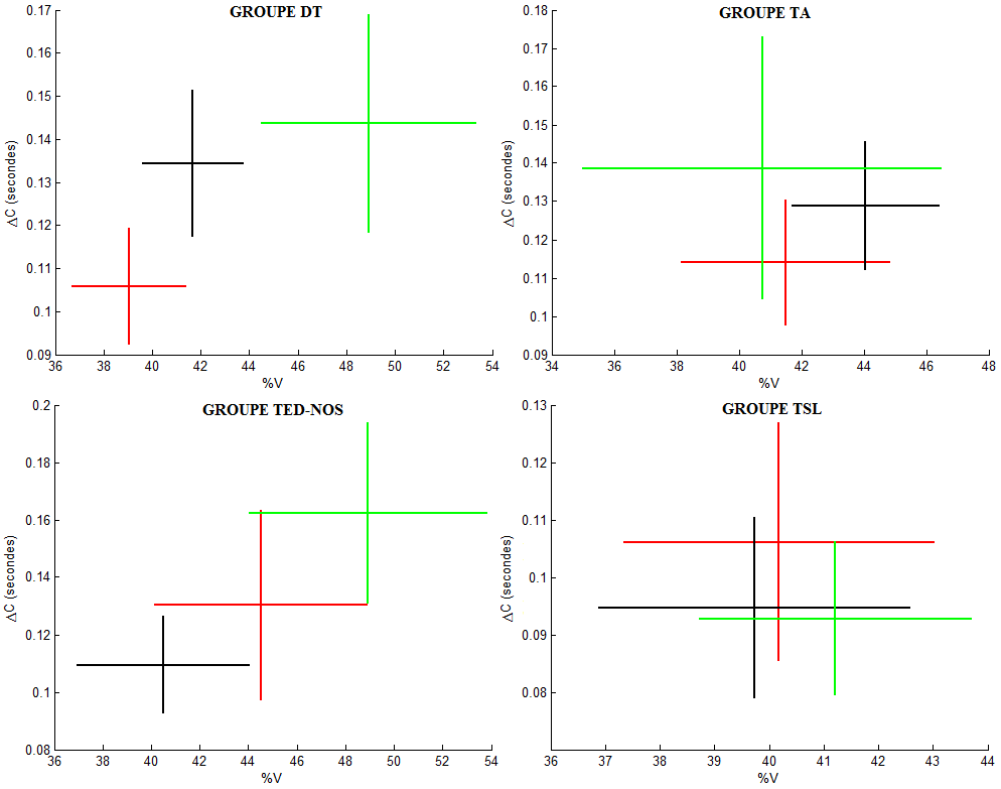


Fig. A.5.1 Comparaison des paramètres *conventionnels* du rythme selon la valence affective (vert : *Positive*, noir : *Neutre* et rouge : *Négatif*) et les groupes d'enfants.

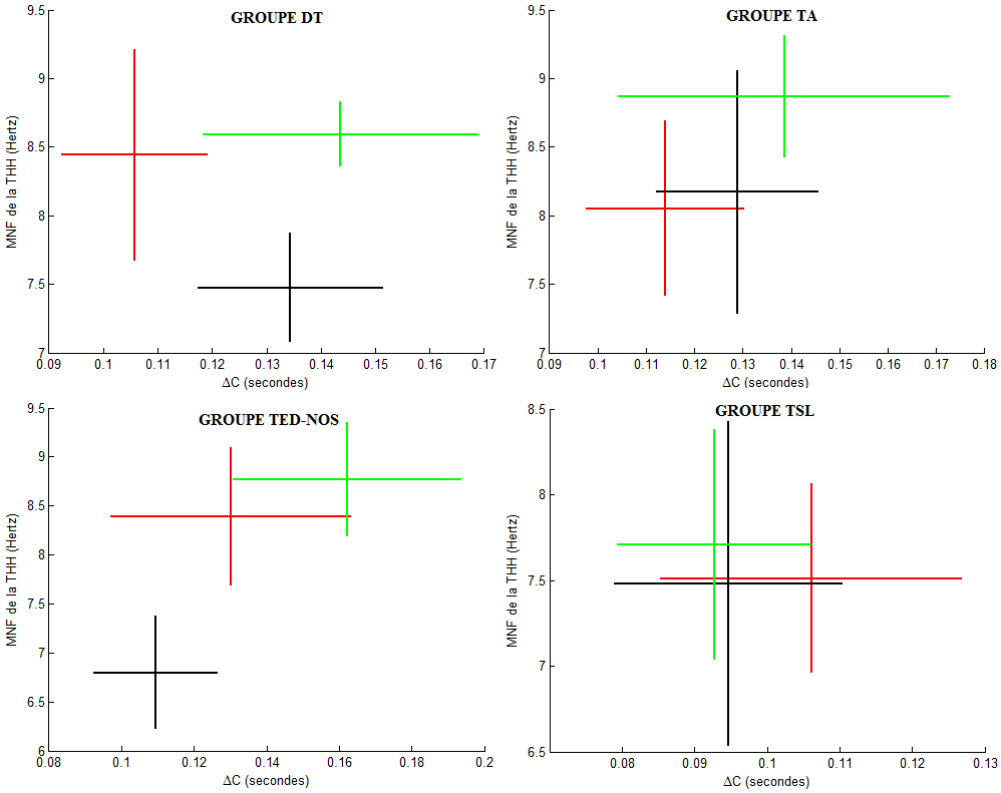


Fig. A.5.2 Comparaison des paramètres *conventionnels* et *non-conventionnels* du rythme selon la valence affective (vert : *Positive*, noir : *Neutre* et rouge : *Négatif*) et les groupes d'enfants.

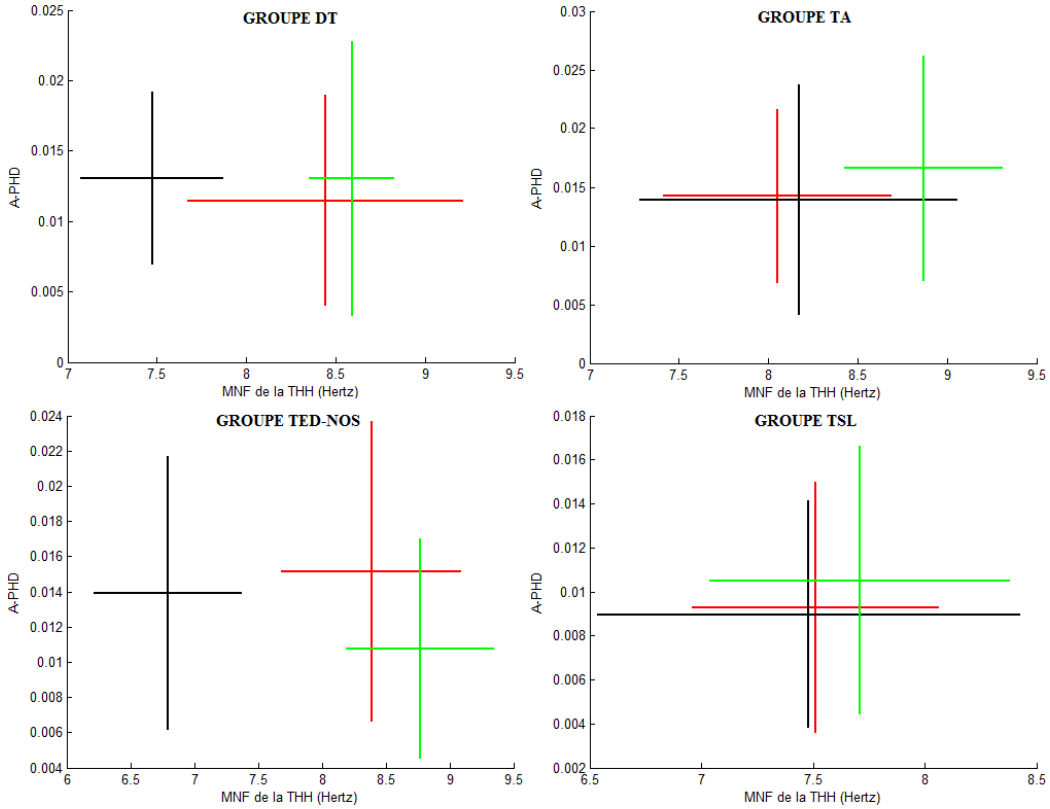


Fig. A.5.3 Comparaison des paramètres *non-conventionnels* du rythme selon la valence affective (vert : *Positive*, noir : *Neutre* et rouge : *Négatif*) et les groupes d'enfants.

BIBLIOGRAPHIE

[ABE67] D. Abercrombie, *Elements of general phonetics*, dans Edinburgh University Press, Mar. 1967.

[ABE09] A. Abel, A. Hussain, Q. D. Nguyen, F. Ringeval, M. Chetouani et M. Milgram, "Maximising audiovisual correlation with automatic lip tracking and vowel based segmentation", dans *LNCS*, J. Fierrez, J. Ortega, A. Esposito A. Drygajlo and M. Faundez-Zanuy [Eds], *joint COST 2101 and 2102 Inter. C. on BioID_MultiComm*, Madrid, Spain, Sep. 16-18 2009, vol. 5707, pp. 65–72.

[ABR08] J. C. Abric, *Psychologie de la communication : théories et méthodes*, dans Armand Colin, 3ème Ed., Fev. 2008.

[ALL77] J. B. Allen et L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", dans *proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[ALL92] D. A. Allen et I. Rapin, "Autistic children are also dysphasic", dans H. Naruse and E. M. Ornitz [Eds], *Neurobiology of infantile autism*, Amsterdam: Excerpta Medica, pp. 157–168, 1992.

[ALM09] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J. C. Martin, S. Ondas, C. Pelachaud, J. Urbain et M. Yilmaz, "Generating robot/agent backchannels during a storytelling experiment", dans proc. *IEEE Inter. C. on Rob. and Automation*, Kobe, Japan, May 12-17 2009, pp. 2477–2482.

[ANA08] S. Ananthakrishnan et S. Narayanan, "Fine-grained pitch accent and boundary tones labeling with parametric f0 features", dans proc. *ICASSP*, Las Vegas (NV), USA, Mar. 30-Apr. 4 2008, pp. 4545–4548.

[ANA09] S. Ananthakrishnan et S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition", dans *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 17, no. 1, pp. 138–149, Jan. 2009.

[ANG97] J. P. Angoujard, *Théorie de la syllabe, rythme et qualité*, dans CNRS [Eds], Paris, 1997.

[ANG02] J. Ang, R. Dhillon, E. Schriberg et A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in Human-computer dialog", dans proc. *Interspeech, 7th ICSLP*, Denver (CO), USA, Sep. 16-20 2002, pp. 2037–2040.

[ARA19] O. Aran et D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversation", dans proc. *ICPR*, Istanbul, Turkey, Aug. 23-26 2010, pp. 3687–3690.

[ARG67] M. Argyle, *The Psychology of Interpersonal Behaviour*, dans Penguin, 1967.

[ARI07] Y Arimoto, S. Ohno et H. Iida, "Acoustic features of anger utterances during natural dialog", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2217–2220.

[ARU01] S. Arunachalam, D. Gould, E. Andersen, D. Byrd et S. Narayanan, "Politeness and frustration language in child-computer interactions", dans proc. *Eurospeech*, Aalborg, Denmark, Sep. 3-7 2001, pp. 2675–2678.

[ARV10] A. Arvaniti et T. Ross, "Rhythm classes and speech perception", dans proc. *Speech Prosody*, Chicago (IL), USA, May 11-14 2010, paper id. 100173:1-4.

[ASP44] H. Asperger, "Autistic psychopathy in childhood", (Translation and annotation by U. Frith of the original paper), dans U. Frith [Eds], *Autism and Asperger Syndrome*, Cambridge University Press (1991, pp. 37–92), 1944.

[AUS05] A. Austermann, N. Esau, L. Kleinjohann et B. Kleinjohann, "Prosody based emotion recognition for MEXI", dans proc. *IROS*, Edmonton, Alberta, Canada, Aug. 2-6 2005, pp. 83–104.

[AVE75] J. R. Averill, "A semantic atlas of emotional concepts", dans *JSAS Catalog of Selected Documents in Psycho.*, vol. 5, pp. 330, 1975.

[BAG91] B. Baghemil, "Syllable structure in Bella Coola", dans *Ling. Inquiry*, vol. 22, no. 4, pp. 589–646, Aut. 1991.

[BAN72] T. de Banville, *Petit traité de poésie*, dans Chez l'écho de la Sorbonne, Paris, 1872.

[BAN01] T. Bänziger, D. Grandjean, P. J. Bernard, G. Klasmeyer et K. R. Scherer, "Prosodie de l'émotion : étude de l'encodage et du décodage", dans *Cahiers de Lingu. Française*, no. 23, pp. 11–37, 2001.

[BAN05] T. Bänziger, V. Tran, et K. R. Sherer, "The Geneva Emotion Wheel: A tool for the verbal report of emotion reactions", dans *ISRE*, Bari, Italy, Jul. 11-15, 2005, pp. 241–254.

[BAR05] P. A. Barbosa, P. Arantes, A. Meireles et J. M. Vieira, "Abstractness in speech-metronome synchronisation: p-centres as cyclic attractors", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 1441–1444.

[BAT99] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth et H. Niemann, "Prosodic feature evaluation: brute force or well designed?", dans proc. *14th ICPhS*, San Francisco (CA), USA, Aug. 1999, pp. 2315–2318.

[BEL09] G. Beller, *Analyse et modèle génératif de l'expressivité: Application à la musique et à l'interprétation musicale*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 2009.

[BER02] F. Beritelli, S. Casale, G. Rugeri, et S. Serrano, "Performance evaluation and

comparison of G.729/AMR/fuzzy voice activity detectors", dans *IEEE Signal Proc. Letters*, vol. 9, no. 3, pp. 85–88, 2002.

[BLA09] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee et S. Narayanan, "Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children", dans proc. *ICASSP*, Taipei, Taiwan, Apr. 19-24 2009, pp. 4861–4864.

[BOL68] D. Bolinger, *Aspects of language*, dans Harcourt, Brace and World, New York, 1968.

[BOT83] K. de Bot "Visual feedback of intonation: effectiveness and induced practice behavior", dans *Lang. and Speech*, vol. 26, no. 4, pp. 331–350, Oct.-Dec. 1983.

[BRA06] M. C. Brady et R. F. Port, "Quantifying vowel onset periodicity in Japanese", dans proc. *16th ICPhS*, Saarbrücken, Germany, Aug. 6-10 2006, pp. 337–342.

[BRE02] C. Breazeal et L. Aryananda, "Recognition of affective communicative intent in robot-directed speech", dans *Autonomous Robots*, Kluwer Academic Publishers, vol. 12, no. 1, pp. 83–104, Jan. 2002.

[BRO07] L. Bronakowski, K. Slot, J. Cichosz et J. Kim, "Application of Poincare map-based description of vowel pronunciation variability for emotion assessment in speech signal", dans *Int. Symp. on Inf. and Tech. Conv.*, Jeonju, Korea, Nov. 23-24 2007, pp. 175–178.

[BUL05] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. Lee, S. Lee, et S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 801–804.

[BUL07] M. Bulut, S. Lee et S. Narayanan, "Analysis of emotional speech prosody in terms of part of speech tags", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 626–629.

[BUR05] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier et B. Weiss, "A database of German emotional speech", dans proc. *Interspeech*, Lisbon, Portugal, Sep. 4-8 2005, pp. 1517–1520.
http://pascal.kgw.tu-berlin.de/emodb/download/download.zip.

[BUS04] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, S. Lee, U. Neumann, et S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information", dans proc. *6th ICMI*, State College (PA), USA, Oct. 13-15 2004, pp. 205–211.

[CAM92] N. Campbell, *Multi-level speech timing control*, PhD thesis, University Sussex, UK, 1992.

[CAM09] N. Campbell, "An audio-visual approach to measuring discourse synchrony in multimodal conversation data", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 2159–2162.

[CHA02] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", dans *J. of Artificial Intel. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[CHE04] M. Chetouani, *Codage neuro-prédictif pour l'extraction de caractéristiques de signaux de parole*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris, 2004.

[CHE09a] M. Chetouani, M. Faundez-Zanuy, B. Gas et J. L. Zarader, "Investigation on LP-residual representations for speaker identification", dans *Pattern Recogn.*, vol. 42, no. 3, pp. 487–494, 2009.

[CHE09b] M. Chetouani, A. Mahdhaoui et F. Ringeval, "Time-scale feature extractions for emotional speech characterization", dans *Cognitive Comp.*, Springer Verlag, vol. 1, no. 2, pp. 194–201, 2009.

[CHW88] K. Chwalisz, E. Diener et D. Gallagher, "Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured", dans *J. of Personality and Social Psycho.*, vol. 54, pp. 820–828, 1988.

[CLA07] C. Clavel, *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales*, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, TELECOM Paris, 2007.

[COR96] R. R. Cornelius, *The science of emotion. Research and tradition in the psychology of emotion*, dans Prentice Hall, Upper Saddle River, NJ, 1996.

[COR00] R. R. Cornelius, "Theoretical approaches to emotion", dans proc. *ISCA Tut. and Res. W. on Speech and Emotion*, Newcastle, Northern Ireland, Sep. 5-7 2000, pp. 3–10.

[COW01] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz et J. Taylor, "Emotion recognition in Human-computer interaction", dans *IEEE Signal Proc. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[COW03] R. Cowie et R. Cornelius, "Describing the emotional states that are expressed in speech", dans *Speech Comm.*, vol. 40, pp. 2–32, 2003.

[CRY82] D. Crystal, *Profiling Linguistic Disability*, Edward Arnold, London, 1982.

[CUM98] F. Cummins et R. Port, "Rhythmic constraints on stress timing in English", dans *J. of Phonetics*, vol. 26, pp. 145–171, 1998.

[CUM02] F. Cummins, "Speech rhythm and rhythmic taxonomy", dans proc. *Speech Prosody*, Aix-en-Provence, France, Apr. 11-13 2002, pp. 121–126.

[CUM09] F. Cummins, "Rhythm as entrainment: The case of synchronous speech", dans *J. of Phonetics*, vol. 37, no. 1, pp. 16–28, Jan. 2009.

[DAR72] C. Darwin, *The Expression of Emotion in Man and Animals*, dans John Murray, London, 1872 (P. Ekman [Eds], Oxford University Press, 3th Ed., 1998).

[DAU83] R. M., Dauer, "Stress-timing and syllable-timing reanalysed", dans *J. of Phonetics*, vol. 11, pp. 51–62, 1983.

[DAV52] K. H. Davis, R. Biddulph et S. Balashek, "Automatic recognition of spoken digits", dans *The J. of the Acoust. Soc. of Amer.*, vol. 24, pp. 637–642, 1952.

[DAV80] S. Davis et P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", dans *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[DEL85] F. Dell, et M. Elmedlaoui, "Syllabic consonants and syllabification in Imdlawn Tashlhyit Berber", dans *J. of African Lang. and Ling.*, vol. 7, pp. 105–130, 1985.

[DEL03] V. Dellwo et P. Wagner, "Relations between language rhythm and speech rate", dans proc. *15th ICPhS*, Barcelona, Spain, Aug. 3-9 2003, pp. 471–474.

[DEL06] V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC", dans proc. *Lang. and Lang. Proc., 38th Ling. Colloq.*, Piliscsaba, Hungary, Sep. 6-8 2006, pp. 231–241.

[DEL08] V. Dellwo, "The role of speech rate in perceiving speech rhythm", dans proc. *Speech Prosody*, Campinas, Brasil, May 6-9, 2008, pp. 375–378.

[DEM77] P. Dempster, N. M. Laird et D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", dans *J. of Acoust. Soc. of Amer.*, vol. 39, pp. 1–38, 1977.

[DEM09] J. Demouy, "*Caractéristiques prosodiques des enfants et adolescents autistes, dysharmoniques, dysphasiques et sans pathologie*", Mémoire de l'école d'orthophonie de la Pitié-Salpêtrière, Université Pierre et Marie Curie, Paris 6, 2009.

[DIC00] A. Di Cristo, "*Interpréter la prosodie*", dans proc. *23èmes JEP*, Aussois, France, Jun. 19-23 2000, pp. 13–29.

[DIC04] A. Di Cristo, "La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions", dans *Travaux interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, no. 23, pp. 67–211, 2004.

[DOU08] *L'Homme et ses rythmes*, séminaire dirigé par C. Doumet et A. W. Lasowski, Université Paris 8, Vincennes, Saint-Denis à l'ENS de la rue d'Ulm, France,

2006-2008.

[DRU94] R. Drullman, J. M. Festen et R. Plomp, "Effect of temporal envelope smearing on speech reception", dans *J. of the Acous. Soc. of Amer.*, vol. 95, pp. 1053–1064, 1994.

[DSM94] American Psychiatric Association, *Diagnostic and Statistical Manual of mental disorders*, 4th Ed., Washington, DC, 1994.

[DUD00] R. O. Duda, P. E. Hart et D. G. Stork, *Pattern classification*, 2nd Ed., New York: Wiley, 2000.

[DUF41] E. Duffy, "An explanation of 'emotional' phenomena without the use of the concept 'emotion'", dans *J. of General Pscyho.*, vol. 25, pp. 283–293, 1941.

[EKM69] P. Ekman et W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding", dans *Semiotica*, vol. 1, pp. 49–98, 1969.

[ELF02] H. A. Elfenbein et N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis", dans *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, Mar. 2002.

[ELE04] D. Elenius et M. Blomberg, "Comparing speech recognition for adults and children", dans proc. *FONETIK*, Stockholm, Sweden, May 26-28 2004, pp. 105–108.

[ENG96] I. S. Engbert et A. V. Hansen, "Documentation of the Danish Emotional Speech Database DES", dans *Tech. Rep. Center for PersonKommunikation*, Aalborg University Denmark, 1996, http://cpk.auc.dk~tb/speech/Emotions.

[ESP09] A. Esposito, M. Teresa Riviello et N. Bourbakis, "Cultural specific effects on the recognition of basic emotions: A study on Italian subjects", dans A. Holzinger and K. Miesenherger [Eds], *LNCS*, *USAB*, vol. 5889, pp. 135–148, 2009.

[EST05] P. A. Estevez, N. Becerra-Yoma, N. Boric, et J. A. Ramirez, "Genetic programming based voice activity detection", dans *IEEE Electronic Letters*, vol. 41, no. 20, pp. 1141–1142, 2005.

[EVA86] J. R. Evans, et M. Clynes, *Rhythm in psychological, linguistic, and musical processes*, dans C. Thomas [Eds], Springfield, 1986.

[FAK97] N. Fakotakis, K. Georgila et A. Tsopanoglou, "A continuous HMM text-independent speaker recognition system based on vowel spotting", dans proc. *5th Eur. Conf. on Speech Comm. and Tech.* (*Eurospeech*), Rhodes, Greece, Sep. 22-25 1997, vol. 5, pp. 2247–2250.

[FAN60] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton [Eds],

1960.

[FAR01] J. Farinas et F. Pellegrino, "Automatic rhythm modeling for language identification", dans proc. *Eurospeech*, Aalborg, Denmark, Sep. 3-7 2001, pp. 2539–2542.

[FEH84] B. Fehr et J. A. Russel, "Concept of emotion viewed from a prototype perspective", dans *J. of Experim. Psycho.*, vol. 113, pp. 464–486, 1984.

[FEN10] G. Fenk-Oczlon et A. Fenk, "Measuring basic tempo across languages and some implications for speech rhythm", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010, pp. 1537–1540.

[FOM92] E. Fombonne et C. du Mazaubrun, "Prevalence of infantile autism in 4 French regions", dans *Social Psychiatry and Psychiatric Epidemiology*, vol. 27, pp. 203–210, 1992.

[FOM97] E. Fombonne, C. du Mazaubrun, C. Cans et H. Grandjean, "Autism and associated medical disorders in a French epidemiological survey", dans *J. of the Amer. Acad. of Child and Adolesc. Psychiatry*, vol. 36, pp. 1561–1569, 1997.

[FOM03] E. Fombonne, "Epidemiological surveys of autism and other pervasive developmental disorders: an update", dans *J. of Autism and Develop. Disorders*, vol. 33, no. 4, pp. 365–382, 2003.

[FOS99] S. Fosnot et S. Jun, "Prosodic characteristics in children with stuttering or autism during reading and imitation", dans proc. *14th Annual Congress of Phonetic Sc.*, San Francisco (CL), Aug. 1-7 1999, pp. 103–115.

[FOW79] C. Fowler, "Perceptual centers", dans *Speech Production and Perception, Perception and Psychophysics*, vol. 25, pp. 375–388, 1979.

[FRA56] P. Fraisse (préface de A. Michotte), *Les structures rythmiques : étude psychologique*, dans Louvain [Eds], publications universitaires de Louvain, 1956.

[FRI86] N. H. Frijda, *The emotion*, dans Cambridge University Press, 1986.

[FRI90] W. Friedman, *About time. Inventing the fourth dimension*, dans Cambridge: MIT Press, 1990.

[FUJ04] H. Fujisaki, "Information, prosody, and modeling – with emphasis on tonal features of speech", dans *Speech Prosody*, Nara, Japan, Mar. 23-26 2004, invited paper.

[FUR05] S. Furui, "50 years of progress in speech and speaker recognition", dans proc. *SPECOM*, Patras, Greece, Oct. 17-19 2005, pp. 1–9.

[FUR09] S. Furui, "Selected topics from 40 years of research on speech and speaker recognition", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 1–8.

[GAL02] A. Galves Jesus, J. Garcia, D. Duarte et C. Galves, "Sonority as a basis for rhythm class discrimination", dans proc. *Speech Prosody*, Aix-en-Provence, France, Apr. 11-13 2002, pp. 11–13.

[GAR68] P. Garde, *L'accent*, dans Presses Universitaires de France [Eds], Paris, 1968.

[GAR93] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren et V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus", dans *Linguistic Data Consortium*, Philadelphia (PA), USA, 1993.

[GER56] G. Gerbner, "Toward a general model of communication", dans *Audio-Visual Comm. Review*, vol. 4, pp. 171–199, 1956.

[GER93] C. L. Gerard, *L'enfant dysphasique*, 1ère Ed., Paris, 1993.

[GIB01] D. Gibbon et U. Gut, "Measuring speech rhythm", dans proc. *Eurospeech*, Aalborg, Denmark, 3-7 Sep. 2001, pp. 95-98.

[GOB03] C. Gobl et A. Ní Chasaide, "The role of voice quality in communication emotion, mood and attitude", dans *Speech Comm.*, vol. 40, no. 1-2, pp. 189–212, Apr. 2003.

[GOR06] J. M. Górriz, J. Ramírez, C. G. Puntonet, et J. C. Segura, "Generalized LRT-based voice activity detector", dans *IEEE Signal Proc. Letters*, vol. 13, no. 10, pp. 636–639, Oct. 2006.

[GOU05] F.Gouyon, *A Computational Approach to Rhythm Description: Audio Features for the Computation of Rhythm Periodicity Functions and Their Use in Tempo Induction and Music Content Processing*, PhD thesis, University Pompeu Fabra, Barcelona, Spain, 2005.

[GOU10] M. Goudbeek et M. Broersma, "Language specific effects of emotion on phoneme duration", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010 (pas de pagination).

[GRA98] E. Grabe, F. Nolan et K. Farrar, "IViE—A comparative transcription system for intonational variation in English", dans proc. *ICSLP*, Sydney, Australia, Nov. 30-Dec. 4 1998, pp. 1259–1262.

[GRA00] E. Grabe, B. Post, F. Nolan et K. Farrar, "Pitch accent realization in four varieties of British English", dans *J. of Phonetic*, vol. 28, no. 2, pp. 161–185, Jun. 2000.

[GRA02a] E. Grabe et E. L. Low, "Durational variability in speech and the rhythm class hypothesis", dans C. Gussenhoven & N. Warner [Eds], *Papers in laboratory phonology VII*, The Hague: Mouton de Gruyter, vol. 7, pp. 515–546, 2002.

[GRA02b] H. P. Graf, E. Cosatto, V. Strom et F. J. Huang, "Visual prosody: Facial movements accompanying speech", dans proc. *5th AFGR*, Washington DC,

USA, May 21 2002, pp. 381–386.

[GRI00] S. Griffin, L. Wilson, L. Clark, et S. McLeod, "Speech pathology applications of automatic speech recognition technology", dans L. Wilson & S. Hewat [Eds], dans proc. *8th Australian Inter. Conf. on Speech Sci. and Tech.*, Melbourne, Australia, Dec. 5-7 2000, pp. 362–366.

[GUP07] P. Gupta et N. Rajput, "Two-stream emotion recognition for call-center monitoring", dans proc. *10th Eurospeech – Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2241–2244.

[GUS01] S. Gustafson-Capková, "*Emotions in speech: Tagset and acoustic correlates",* dans *Term paper in Speech Technology*, GSLT, Stockholm University, Depart. of Ling., Aut. 2001.

[HAR89] P. Hargrove et C. P. Sheran, "The use of stress by language impaired children", dans *J. of Comm. Disorders,* vol. 22, no. 5, pp. 361–373, Oct. 1989.

[HER90] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech", dans *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[HIR93] D. Hirst et R. Espesser, "Automatic modeling of fundamental frequency using a quadratic spline function", dans *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, vol. 15, pp. 75–85, 1993.

[HOH96] G. H. Hohmann, "Some effects of spinal cord lesions on experiences emotional feelings", dans *Psychophysiology,* vol. 3, pp. 143–156, 1996.

[HOW00] A. W. Howitt, "Vowel landmark detection", dans proc. *6th ICSLP*, Beijing, China, Oct. 16-20 2000, vol. 4, pp. 628–631.

[HUA98] N. Huang, Z. Shen, S. Long, *et al.*, "The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis", dans proc. *R. Soc. London*, ser. A, vol. 454, pp. 903–995, Mar. 1998.

[IZA94] C. E. Izard, "Innate and universal facial expressions: Evidence from developmental and cross-cultural research", dans *Psychological Bulletin*, vol. 115, no. 2, pp. 288–299, 1994.

[JAK60] R. Jakobson, *Closing statement: Linguistics and poetics,* dans T. Sebeok [Eds], *Style in Language,* 1960.

[JAM84] W. James, "What is an emotion?", dans *Mind,* vol. 19, pp. 188-205, 1884.

[JAM76] E. James, "The acquisition of prosodic features of speech using a speech visualizer", dans *Inter. R. of Applied Ling.*, vol. 14, pp. 227–243, 1976.

[JAN90] C. Jankowski, A. Kalyanswamy, S. Basson et J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwith speech database", dans

proc. *ICASSP*, Albuquerque (NM), USA, Apr. 3-6 1990, vol. 1, pp. 109–112.

[JEL75] F. Jelinek, L. Bahl et R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech", dans *IEEE Trans. on Information Theory*, vol. 21, no. 3, pp. 250–256, May 1975.

[JON87] M. R. Jones, "Dynamic pattern structure in music: recent theory and research", dans *Perception & psychophysics*, Psychonomic Society, Austin (TX), USA, vol. 41, no. 6, pp. 621–634, Jun. 1987.

[JUA92] B. H. Juang et S. Katagiri, "Discriminative learning for minimum error classification", dans *IEEE Trans. on Signal Proc.*, vol. 40, pp. 3043–3054, Dec. 1992.

[KAN43] L. Kanner, "Autistic disturbances of affective contact", dans *Nervous child*, vol. 2, pp. 217–250, 1943.

[KAP07] A. Kapoor, W. Burleson et R. W. Picard, "Automatic prediction of frustration", dans *Inter. J. of Human-Computer Studies*, vol. 65, pp. 724–736, 2007.

[KAY84] J. D. Kaye et J. Lowenstamm, "De la syllabicité", dans F. Dell, D. J. Hirst, et J. R. Vergnaud [Eds], *La forme sonore du langage*, Hermann, Paris, pp. 123–159, 1984.

[KEL00] B. Zellner Keller, et E. Keller, The chaotic nature of speech rhythm: hints for fluency in the language acquisition process, dans Delcloque, Ph., Holland, V.M. [Eds], *Integrating Speech Technology in Language Learning*, Swets & Zeitlinger, in press, 2000.

[KEL05] E. Keller, "The analysis of voice quality in speech processing", dans *LNCS*, Springer-Verlag, vol. 3445, pp. 54–73, 2005.

[KEN96] R. D. Kent, "Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders", dans *Amer. J. of Speech-Lang. Pathology*, vol. 5, no. 3, pp. 7–23, Aug. 1996.

[KHO01] A. Khomsi, *Evaluation du Langage Oral*, dans Paris: ECPA, 2001.

[KIM06] C. Kim, K. D. Seo et W. Sung, "A robust formant extraction algorithm combining spectral peak picking and root polishing", dans *EURASIP J. on Applied Signal Proc.*, vol. 2006, article id. 67960, 2006.

[KIM07] S. Kim, P. G. Georgiou, S. Lee, et S. Narayanan, "Real-time emotion detection system using speech: multi-modal fusion of different timescale features", dans proc. *9th W. on MMSP*, Chania, Crete, Greece, Oct. 1-3 2007, pp. 48–51.

[KIS10] G. Kiss et J. van Santen, "Automated vocal emotion recognition using phoneme class specific features", dans proc. *Interspeech*, Makuhari, Japan, Sep. 26-30 2010 (pas de pagination).

[KLE03] W. B. Kleijn, "Signal processing representations of speech", dans *IEICE Trans. on Information and Systems*, vol. E86-D, no. 3, pp. 359–376, Mar. 2003.

[KUH04] P. Kuhl, "Early language acquisition: cracking the speech code", dans *Nature Reviews, Neuroscience*, vol. 5, pp. 831–843, Nov. 2004.

[KUN04] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, dans Wiley & Sons Inc., Hoboken, New Jersey, Jul. 2004.

[LAA97] Gitta P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and read speaking style", dans *Speech Comm.*, vol. 22, pp. 43–65, Mar. 1997.

[LAB05] L. Labrune, "Autour de la syllabe : les constituants prosodiques mineurs en phonologie", dans N. Nguyen, S. Wauquiers et J. Durand [Eds], *Phonétique et phonologie, approches contemporaines*, Hermès, pp. 95–116, 2005.

[LAN81] D. van Lancker, D. Canter et D. Terbeek, "Disambiguation of ditropic sentences: Acoustic and phonetic cues", dans *J. of Speech and Hearing Res.* vol. 24, no. 3, pp. 330–335, Sep. 1981.

[LAZ94] R. S. Lazarus, "The stable and the unstable in emotion. Fundamental questions", dans P. Ekman and R. Davidson [Eds], *The Nature of Emotion: Fundamental Questions*, pp. 79–85, Oxford University Press, 1994.

[LEH77] I. Lehiste, "Isochrony reconsidered", dans *J. of Phonetics*, vol. 5, no. 3, pp. 253–263, Mar. 1977.

[LEE99a] D. D. Lee et H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", dans *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[LEE99b] T. van Leeuwen, *Speech, Music, Sound*, dans Palgrave Macmillan [Eds], Oct. 1999.

[LEE02] C. M. Lee, S. Narayanan et R. Pieraccini, "Combining acoustic and language information for emotion recognition", dans proc. *7th ICSLP*, Denver (CO), USA, Sep. 16-20 2002, pp. 873–876.

[LEE04] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee et S. Narayanan, "Emotion recognition based on phoneme classes", dans proc. *Interspeech*, Jeju Island, Korea, Oct. 4-8 2004, pp. 205–211.

[LEE05] C. M. Lee et S. Narayanan, "Toward detecting emotions in spoken dialogs", dans *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 2, pp. 293–303, Feb. 2005.

[LEI98] F. Leisch, L. C. Jain et K. Hornik, "Cross-validation with active pattern selection for neural network classifiers", dans *IEEE Trans. on Neural Net.*, vol. 9, no. 1, pp. 35–41, Jan. 1998.

[LEN08] M. T. Le Normand, S. Boushaba et A. Lacheret-Dujour, "Prosodic disturbances in autistic children speaking French", dans proc. *Speech Prosody*, Campinas, Brazil, May 6–9 2008, pp. 195–198.

[LER96] F. Lerdahl et R. Jackendoff, *A generative theory of tonal music*, dans MIT Press, Jun. 1996.

[LEV92] R. W. Levenson, "Autonomic nervous system differences among emotions", dans *Psycho. Sci.*, vol. 3, pp. 23–27, 1992.

[LIM10] F. Limousin et M. Blondel, "Prosodie et acquisition de la langue des signes française", dans *Lang., Interaction and Acquisition*, J. Benjamins Publishing Company, vol. 1, pp. 82–109, 2010.

[LIN80] Y. Linde, A. Buzo et R. M. Gray : "An algorithm for Vector Quantizer design", dans *IEEE Trans. on Comm.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[LOR94] C. Lord, M. Rutter et A. Le Couteur, "Autism Diagnostic Interview-Revised: A revision version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders", dans *J. of Autism and Develop. Disorders*, vol. 24, no. 5, pp. 659–685, Oct. 1994.

[LUE09] I. Luengo, E. Navas et I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 332–335.

[MAH08] A. Mahdhaoui, M. Chetouani et C. Zong, "Motherese detection based on segmental and supra-segmental features", dans proc. *19th ICPR*, Tampa (FL), Dec. 8–11 2008, pp. 1–4.

[MAH09] A. Mahdhaoui, F. Ringeval et M. Chetouani, "Emotional speech characterization based on multi-features fusion for face-to-face interaction", dans proc. *3rd Inter. C. on Signals, Circuits and Systems*, Djerba, Tunisia, Nov. 6-8 2009, pp. 1–6.

[MAH10] A. Mahdhaoui, *Analyse des signaux sociaux pour la modélisation de l'interaction face-à-face*, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 2010.

[MAI09] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster et E. Nöth, "PEAKS – A system for the automatic evaluation of voice and speech disorder", dans *Speech Comm.*, vol. 51, no. 5, pp. 425–437, May 2009.

[MAR81] S. Marcus, "Acoustic determinants of perceptual center (p-center) location", dans *Perception and Psychophysics*, vol. 30, no. 3, pp. 247–256, Sep. 1981.

[MAR02] M. Marzinzik et B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics", dans *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

[MAR08] P. Martínez-Castilla et S. Peppé, "Developing a test of prosodic ability for speakers of Iberian-Spanish", dans *Speech Comm.*, vol. 50, no. 11-12, pp. 900–915, Mar. 2008.

[MAR09] C. R. Marshall, S. Harcourt Brown, F. Ramus et H. J. K. Van der Lely, "The link between prosody and language skills in children with SLI et/or dyslexia", dans *Inter. J. of Lang. and Comm. Disorders,* vol. 44, no. 4, pp. 466–488, Jul. 2009.

[MAT05] Y. Y. Mathieu, "Annotation of emotions and feelings in texts", dans proc. *ACII*, Beijing, China, Oct. 22-24 2005, vol. 3784, pp. 350–357.

[MAY69] M. Mayer, *Frog where are you?,* dans New York: Dial Books for young readers, 1969.

[MCC03] J. McCann et S. Peppé, "Prosody in autism: a critical review", dans *Inter. J. of Lang. and Comm. Disorders,* vol. 38, no. 4, pp. 325–350, Oct.-Dec. 2003.

[MCC07] J. McCann, S. Peppé, F. Gibbon, A. O'hare, et M. Rutherford, "Prosody and its relationship to language in school-aged children with high-functioning autism", dans *Inter. J. of Lang. and Comm. Disorders,* vol. 42, no. 6, pp. 682–702, Nov.-Dec. 2007.

[MEH96] J. Mehler, E. Dupoux, T. Nazzi, et G. Dehaene-Lambertz, "Coping with linguistic diversity: the infant's viewpoint", dans J.L. Morgan and K. Demuth [Eds], *Signal to syntax: bootstrapping from speech to grammar in early acquisition*, Erlbaum, Mahwah (NJ), USA, 1996.

[MEH67] A. Mehrabian et S. R. Ferris, "Inference of attitude from nonverbal communication in two channels", dans *J. of Counseling Psycho.*, vol. 31, no. 3, pp. 248–252, Jun. 1967.

[MEI08] A. Meireles et P. A. Barbosa, "Speech rate effects on speech rhythm", dans proc. *Speech Prosody*, Campinas, Brasil, May 6-9, 2008, pp. 327–330.

[MER04] P. Mertens, "The Prosogram: 'Semi-automatic transcription of prosody based on a tonal perception model'", dans *Speech Prosody*, Nara, Japan, Mar. 23-26 2004, pp. 143–146.

[MEU97] S. van der Meulen et P. Janssen, "Prosodic abilities in children with Specific Language Impairment", dans *J. of Comm. Disorders*, vol. 30, pp. 155–170, May-Jun. 1997.

[MON09] E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy et J. Sole-Casals, "Maximum likelihood linear programming data fusion for speaker recognition", dans *Speech Comm.*, vol. 51, no. 9, pp. 820–830, Sep. 2009.

[MOR94] P. J. Moreno et R. M. Stern, "Sources of degradation of speech recognition in the telephone network", dans proc. *ICASSP*, Pittsburgh (PA), USA, Apr. 19-22

1994, vol. 1, pp. 109–112.

[MOR96] J. Morgan et K. Demuth, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, dans J. L. Morgan and K. Demuth [Eds], Mahwah, NJ: Erlbaum, 1996.

[MOU74] G. Mounin, *Dictionnaire de la Linguistique*, dans Paris: Presses Universitaires de France, 1974.

[MUR93] I. R. Murray et J. L. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion", dans *J. of Acous. Soc. of Amer.*, vol. 93, no. 2, pp. 1097–1198, 1993.

[NAD02] J. Nadel, "Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism", dans A. N. Meltzoff and W. Prinz [Eds], *The imitative mind: Development, evolution and brain bases*, pp. 2–14, Cambridge University Press, 2002.

[NEM01] E. Nemer, R. Goubran et S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain", dans *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.

[NEW53] T. M. Newcomb, "An approach to the study of communicative acts", dans *Psychological Review*, vol. 60, pp. 393–404, 1953.

[NOO99] S. Nooteboom, "The prosody of speech: melody and rhythm", dans W. J. Hardcastle and J. Laver [Eds], *The handbook of phonetic sciences*, pp. 640–673, Hardcastle, Blackwell, Oxford, 1999.

[NOT01] E. Nöth, A. Batliner, H. Niemann, G. Stemmer, F. Gallwitz et J. Spilker, "Language models beyond word strings", dans proc. *ASRU*, Trento, Italy, Dec. 9-13 2001 (pas de pagination).

[OBR88] R. André-Obrecht, "A new statistical approach for automatic speech segmentation", dans *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 36, no. 1, pp. 29–40, Jan. 1988.

[OBR93] R. André-Obrecht, *Segmentation et parole ?*, Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, 1993.

[OBR97] R. André-Obrecht et B. Jacob, "Direct identification vs. correlated models to process acoustic and articulatory information in automatic speech recognition", dans proc. *ICASSP*, Munich, Germany, Apr. 21-24 1997, pp. 989–992.

[OLL10] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel et S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development", dans *PNAS of the USA*, Jun. 2010, Appendix, pp. 16–21.

[ORT90] A. Ortony et T. J. Turner, “What's basic about basic emotions?”, dans *Psychological Review*, vol. 97, pp. 315–331, 1990.

[OUD03] P. Oudeyer, “The production and recognition of emotions in speech: features and algorithms”, dans *Inter. J. of Human-Computer Studies*, vol. 59, pp. 157–183, Jul. 2003.

[PAE00] A. Paeschke et W. Sendlmeier, “Prosodic characteristics of emotional speech: measurements of fundamental frequency movements”, dans proc. *ISCA ITRW on Speech and Emotion*, Belfast, United-Kingdom, Sep. 5-7 2000, pp. 75–80.

[PAN08] M. Pantic, A. Pentland, A. Nijholt et T. Huang, “Human-centred intelligent human-computer interaction (hci2): how far are we from attaining it ?”, dans *Inter. J. of Autonomous and Adaptive Comm. Systems*, vol. 1, no. 2, pp. 168–187, Aug. 2008.

[PAU05a] R. Paul, A. Augustyn, A. Klin et F. R. Volkmar, “Perception and production of prosody by speakers with autism spectrum disorders” dans *J. of Autism and Develop. Disorders*, vol. 35, no. 2, pp. 205–220, Apr. 2005.

[PAU05b] R. Paul, L. Shriberg, J. Mc Sweeny, D. Ciccheti, A. Klin et F. Volkmar, “Brief Report : relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders”, dans *J. of Autism and Develop. Disorders*, vol. 35, no. 6, pp. 861–869, Dec. 2005.

[PAU08] R. Paul, N. Bianchi, A. Agustyn, A. Klin et F. Volkmar, “Production of syllable stress in speakers with autism spectrum disorders”, dans *Res. in Autism Spectrum Disorders*, vol. 2, pp. 110–124, Jan.-Mar. 2008.

[PEL98] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques*, Thèse de Doctorat, Université Paul Sabatier, Toulouse, 1998.

[PEN07] A. Pentland, “Social Signal Processing”, dans *IEEE Signal Proc. Magazine*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[PFA98] T. Pfau et G. Ruske, “Estimating the speaking rate by vowel detection”, dans proc. *ICASSP*, Seattle (WA), USA, May 12-15 1998, vol. 2, pp. 945–948.

[PFI96] H. Pfitzinger, S. Burger et S. Heid, “Syllable detection in read and spontaneous speech”, dans proc. *4th ICSLP*, Philadelphia (PA), USA, Oct. 3-6 1996, vol. 2, pp. 1261–1264.

[PIC97] R. W. Picard, *Affective Computing*. Perceptual computing section, Technical report no. 321, Cambridge, Massachusetts, M.I.T. Media Lab., 1997.

[PIK45] K. L. Pike, *The intonation of American English*, dans University of Michigan Press, Ann Arbor, 1945.

[PLA02] M. Plaza, D. Chauvin, O. Lanthier, M. T. Rigoard, J. Roustit, M. P. Thibault, et M. Touzin, "Validation longitudinale d'un outil de dépistage des troubles du langage écrit. Etude d'une cohorte d'enfants dépistés en fin de CP et réévalués en fin de CE1", dans *Glossa*, vol. 81, pp. 22–33, 2002.

[PLU62] R. Plutchik, *The Emotions: Facts, Theories, and a New Model*, dans Random House, New York, 1962.

[PLU80] R. Plutchik, Emotion: *A Psychoevolutionary Synthesis*, dans Harper & Row, New York, 1980.

[PLU84] R. Plutchik, "Emotions: A general psychoevolutionary theory", dans K. R. Scherer and P. Ekman [Eds], *Approaches to Emotion*, Erlbaum, Hillsdale (NJ), 1984.

[POL00] T. S. Polzin et A. Waibel, "Emotion-sensitive Human-computer interfaces", dans proc. *ISCA Tutorial and Res. W. on Speech and Emotion*, Newcastle, Northern Ireland, Sep. 5-7 2000, pp. 201–206.

[POL09] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner et F. Metze, "Emotion classification in children's speech using fusion of acoustic and linguistic features", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 340–343.

[POT07] A. Potamianos et S. Narayanan, "A review of the acoustic and linguistic properties of children's speech", dans proc. *IEEE 9th W. on MMSP*, Chania, Greece, Oct. 23 2007, pp. 22–25.

[PRE05] L. Prevost, *Cours de Reconnaissance des Formes*, Master en Sciences de l'Ingénieur, Université Pierre et Marie Curie, Paris 6, 2005.

[QUA07a] V. M. Quang, L. Besacier et E. Castelli, "Automatic question detection: prosodic-lexical features and crosslingual experiments", dans proc. *Interspeech ICSLP*, Antwerp, Belgium, Aug. 27–31 2007, pp. 2257–2260.

[QUA07b] V. M. Quang, *Exploitation de la prosodie pour la segmentation et l'analyse automatique des signaux de parole*, Thèse de Doctorat, Institut National Polytechnique de Grenoble, 2007.

[RAB75] L. R. Rabiner et M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances", dans *The Bell System Technical J.*, vol. 54, no. 2, pp. 297–315, 1975.

[RAB78] L. R. Rabiner et R. W. Schafer, *Digital Processing of Speech Signals*, dans Prentice Hall, Upper Saddle River, NJ, 1978.

[RAM99] F. Ramus, M. Nespor, et J. Mehler, "Correlates of linguistic rhythm in the speech signal", dans *Cognition*, vol. 73, no. 3, pp. 265–292, Dec. 1999.

[RAM02] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives", dans proc. *Speech Prosody*, B. Bel and I. Marlin [Eds], Aix-en-Provence, France, Apr. 11-13 2002, pp. 115–120.

[RAM04] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, et Á. Rubio, "A new Kullback-Leibler VAD for robust speech recognition", dans *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 266–269, Feb. 2004.

[RAM06] J. Ramírez, P. Yélamos, J. M ; Górriz, et J. C. Segura, "SVM-based speech endpoint detection using contextual speech features", dans *IEEE Electronic Letters*, vol. 42, no. 7, pp. 426–428, Apr. 2006.

[REY03] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones et B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", dans proc. ICASSP, Hong Kong, China, pp. 784–787, 2003.

[RIL03] G. Riling, P. Flandrin et P. Gonçalvès, "On empirical mode decomposition and its algorithms", dans proc. *6th IEEE-EURASIP W.* on *NSIP*, Grado, Italy, Jun. 8-11, 2003 (pas de pagination).

[RIN08a] F. Ringeval et M. Chetouani, "Exploiting a vowel based approach for acted emotion recognition", dans *LNCS*, A. Esposito, N. G. Bourbakis, N. Avouris and I. Hatzilygeroudis [Eds], *Verbal and Nonverbal Features of Human-Human and Human-machine Interaction*, Springer Verlag, vol. 5042, pp. 243–254, 2008.

[RIN08b] F. Ringeval et M. Chetouani, "Une approche basée voyelle pour la reconnaissance automatique des émotions actées", dans proc. *JEP*, Avignon, France, Jun. 9-13 2008 (pas de pagination).

[RIN08c] F. Ringeval et M. Chetouani, "A vowel based approach for acted emotion recognition", dans proc. *Interspeech*, Brisbane, Australia, Sep. 22-26 2008, pp. 2763–2766.

[RIN08d] F. Ringeval, M. Chetouani, D. Sztahó et K. Vicsi, "Automatic prosodic disorders analysis for impaired communication children", dans proc. *1st W. on Child, Computer and Interaction*, Chania, Greece, Oct. 23 2008 (pas de pagination).

[RIN09] F. Ringeval et M. Chetouani, "Hilbert-Huang transform for non-linear characterization of speech rhythm", dans proc. *NOLISP*, Vic, Spain, Jun. 25-27 2009 (pas de pagination).

[RIN10] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, et M. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children", dans *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. PP, no. 99, Oct. 2010.

[ROA82] P. Roach, *On the distinction between 'stress-timed' and 'syllable-timed' languages*, dans D. Crystal [Eds], Linguistic Controversies, Arnold, London, 1982.

[ROB89] L. Robert, *Les horloges biologiques*, dans Paris: Flammarion, 1989.

[ROB03] M. Robnik et I. Konenko, "Theoretical and empirical analysis of ReliefF and RReliefF", dans *Mach. Learn. J.*, vol. 53, pp. 23–69, Oct.-Nov. 2003.

[ROD09] W. R. Rodríguez et E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool", dans *ISCA Inter. W. on Speech and Lang. Techn. in Educ.*, Wroxall Abbey Estate, UK, Sep. 3-5 2009 (pas de pagination).

[RON] E. Rondeau, L. Klein, A. Masse, N. Bodeau, D. Cohen, J. M. Guilé, "Is Pervasive Developmental Disorder Not Otherwise Specified less stable than Autistic Disorder?", dans *J. of Autism and Develop. Disorder* (en révision).

[ROS03] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio et B. De Carolis, "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent", dans *Inter. J. of Human-Computer Studies, Application of affective computing in Human–computer interaction*, vol. 59, no. 1-2, pp. 81–118, Jul. 2003.

[ROS09] A. Rosenberg et J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level", dans proc. *Human Lang. Tech.: The 2009 Ann. C. of the North Amer. Chapter of the Assoc. for Comp. Ling.*, Boulder, Colorado, USA, May 31 - Jun. 5 2009, pp. 81–84.

[ROT05] M. Rotaru, D. J. Litman et K. Forbes-Riley, "Interactions between speech recognition problems and user emotions", dans proc. *Interspeech*, Lilsbon, Portugal, Sep. 4-8 2005, pp. 1–4.

[ROU05] J. L. Rouas, *Caractérisation et identification automatique des langues*, Thèse de Doctorat, Université Paul Sabatier, Toulouse, 2005.

[RUS80] J. A. Russel, "A circumplex model of affect", dans *J. of Personality and Social Psycho.*, vol. 39, pp. 1161–1178, 1980.

[RUT02] M. D. Rutherford, S. Baron-Cohen et S. Wheelwright, "Reading the mind in the voice: a study with normal adults and adults with Asperger syndrome and high functioning autism", dans *J. of Autism and Develop. Disorders*, vol. 32, no. 3, pp. 189–194, Jun. 2002.

[SAL88] G. Salton et C. Buckley, "Term-weighting approaches in automatic text retrieval", dans *Inf. Proc. and Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[SAM03] C. Samuelsson, C. Scocco et U. Nettelbladt, “Towards assessment of prosodic abilities in Swedish children with language impairment”, dans *Logopedics*

Phoniatrics Vocology, vol. 28, no. 4, pp. 156–166, Oct. 2003.

[SAN09] J. P. H. van Santen, E. T. Prud'hommeaux et L. M. Black, "Automated assessment of prosody production", dans *Speech Comm.*, vol. 51, no. 11, pp. 1082–1097, Nov. 2009.

[SAR06] I. Saratxaga, E. Navas, I. Hernaez et I. Luengo, "Designing and recording an emotional speech database for corpus based speech synthesis in Basque", dans proc. *LREC*, Genoa, Italy, May 24-26 2006, pp. 2126–2129.

[SCH54] H. Schlosberg, "Three dimensions of emotion", dans *Psycho. R.*, vol. 61, pp. 81–88, 1954.

[SCH80] E. Schopler, R. Reichler, R. Devellis et K. Daly, "Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS)", dans *J. of Autism and Develop. Disorders*, vol. 10, no. 1, pp. 91–103, 1980.

[SCH84] K. R. Scherer, "Emotion as a multicomponent process: a model and some cross-cultural data", dans *R. of Personality and Social Pscyho.*, vol. 5, pp. 37–63, 1984.

[SCH85a] K. R. Scherer, S. Feldstein, R. N. Bond et R. Rosenthal, "Vocal cues to deception: A comparative channel approach", dans *J. of Psycholingu. Res.*, vol. 14, pp. 409–425, 1985.

[SCH85b] M. R. Schroeder et B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates", dans proc. *ICASSP*, Tampa (FL), Mar. 26-29 1985, pp. 937–940.

[SCH86] K. R. Scherer, "Vocal affect expression: a review and a model for future research", dans *Psychological Bulletin*, vol. 99, no. 2, pp. 143–165, Mar. 1986.

[SCH89] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance", dans H. L. Wagner and A. Manstead [Eds], *Handbook of Social Psychophysiology*, Wiley and Sons Ltd, London, chap. 7, pp. 165–197, 1989.

[SCH90] L. D. Schriberg, J. Kwiatkowski, et C. Rasmussen, *The Prosody-Voice Screening Profile*, dans Tuscon, AZ: Communication Skill Builders, 1990.

[SCH94] K. R. Scherer, "Towards a concept of modal emotions", dans P. Ekman & R. Davidson [Eds], *The Nature of Emotion: Fundamental Questions*, pp. 25–31, Oxford University Press, 1994.

[SCH99] K. R. Scherer, "Appraisal Theory", dans T. Dalgleish and M. Power [Eds], *Handbook of Cognition and Emotion*, pp. 637–663, John Wiley, New York, 1999.

[SCH00] K. R. Scherer, "Psychological models of emotion", dans J. C. Borod, [Eds], *The neuropsychology of emotion*, pp. 137–162, Oxford University Press, Ox-

ford, New York, 2000.

[SCH01a] K. R. Scherer, R. Banse et H. G. Wallbott, “Emotion inferences from vocal expression correlate across languages and cultures”, dans *J. of Cross-Cultural Psycho.*, vol. 32, pp. 76–92, 2001.

[SCH01b] M. Schröder, “Emotional speech synthesis: A review”, dans proc. *Interspeech*, Aalborg, Denmark, Sep. 3-7 2001, vol. 1, pp. 561–564.

[SCH03] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms”, dans *Speech Comm.*”, vol. 40, pp. 227–256, 2003.

[SCH04a] D. Schreuder et D. Gilbers, “The influence of speech rate on rhythm patterns”, dans Gilbers, D., Schreuder, M. and N. Knevel [Eds], *On the Boundary of Phonology and Phonetics*, pp. 183–202, 2004.

[SCH04b] B. Schuller, G. Rigoll et M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid Support Vector Machine-Belief Network architecture”, dans proc. *ICASSP*, Montreal, Canada, May 17-21 2004, pp. 577–580.

[SCH05] B. Schuller, R. Jiménez Villar, G. Rigoll et M. Lang, “Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition”, dans proc. *ICASSP*, Philadelphia (PA), USA, Apr. 6-10 2005, pp. 325–328.

[SCH06a] B. Schuller, J. Stadermann et G. Rigoll, “Affect-robust speech recognition by dynamic emotional adaptation”, dans *Speech Prosody*, Dresden, Germany, May 2-5 2006, paper 169.

[SCH06b] B. Schuller, D. Arsík, F.Wallhoff et G. Rigoll, “Emotion recognition in the noise applying large acoustic features sets”, dans proc. *Speech Prosody*, Dresden, Germany, May. 2-5 2006.

[SCH06c] B. Schuller et G. Rigoll, “Timing levels in segment-based speech emotion recognition”, dans proc. *Interspeech*, Pittsburgh, (PA), Sep. 17-21 2006, pp. 1818–1821.

[SCH07a] B. Schuller, D. Seppi, A. Batliner, A. Maier et S. Steidl, “Towards more reality in the recognition of emotional speech”, dans proc. *ICASSP*, Honolulu (HI), USA, Apr. 15-20 2007, pp. 941–944.

[SCH07b] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, A. Noam, L. Kessous, et V. Aharonson, “The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals”, dans *Interspeech*, Antwerp, Belgium, Aug. 27-31 2007, pp. 2253–2256.

[SCH07c] B. Schuller, *Mensh, Maschine, Emotion – Erkennung aus sprachlicher und manueller Interaktion*, dans VDM Verlag Dr. Müller, Saarbrücken, 2007.

[SCH08] B. Schuller, A. Batliner, S. Steidl et D. Seppi, "Does affect affect automatic recognition of children's speech?", dans proc. *1st W. on Child, Computer and Interaction*, Chania, Greece, Oct. 23 2008.

[SCH09a] B. Schuller, *Emotion Recognition in the Next Generation: an Overview and Recent Development*, Tutoriel à *Interspeech*, Brighton, UK, Sep. 6-10 2009.

[SCH09b] B. Schuller, S. Steidl, et A. Batliner, "The Interspeech 2009 emotion challenge", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009.

[SCH09c] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll et A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances", dans proc. *W. on ASRU*, Merano, Italy, Dec. 13-17 2009, pp. 552–557.

[SCO93] S. K. Scott, *P-centers in speech: an acoustic analysis*, Thèse de doctorat, University College London, 1993.

[SHA07] M. Shami et W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", dans *Speech Comm.*, vol. 49, no. 3, pp. 201–212, Mar. 2007.

[SHA48] C. Shannon et W. Warren, "A mathematical theory of communication", dans *The Bell Systems Techn. J.*, (reprinted with corrections), vol. 27, pp.379–423, (pp. 623–656), Jul. (Oct.) 1948.

[SHA87] P. Shaver, J. Schwartz, D. Kirson et C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach", dans *J. of Personality and Social Pscyho.*, vol. 52, pp. 1061–1086, 1987.

[SHR00] E. Shriberg, A. Stolcke, D. Hakkani-Tur et G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", dans *Speech Comm.*, vol. 32, no. 1-2, pp. 127–154, Sep. 2000.

[SEC83] B. Secrest et G. Doddington, "An integrated pitch tracking algorithm for speech systems", dans proc. *ICASSP*, Boston (MA), USA, Apr. 14-16 1983, pp. 1352–1355.

[SIL92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert et J. Hirschberg, "ToBI: A standard scheme for labeling prosody", dans proc. *2nd ICSLP*, Banff (AL), Canada, Oct. 13-16 1992, pp. 867–869.

[SJO00] K. Sjöleter et J. Beskow, "WaveSurfer - an open source speech tool", dans proc. *6th ICSLP*, vol. 4, Beijing, China, Oct. 16-20 2000, pp. 464–467. Disponible à l'adresse suivante : http://www.speech.kth.se/wavesurfer/.

[SKO04] M. D. Skowronski et J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition", dans *J. Acoust. Soc. Amer.*, vol. 116, no. 3, pp. 1774–1780, Sep. 2004.

[SMI00] L. M. Smith, *A Multiresolution Time-Frequency Analysis and Interpretation of Musical Rhythm*, PhD thesis, University of Western Australia, 2000.

[SNO31] W. B. Snow, "Audible frequency ranges of music, speech and noise", dans *J. of the Acous. Soc. Amer.*, Jul. 1931.

[SNO98a] D. Snow, "Children's imitations of intonation contours: are rising tones more difficult than falling tones?", dans *J. of Speech, Lang. and Hearing Research*, vol. 41, pp. 576–587, Jun. 1998.

[SNO98b] D. Snow, "Prosodic markers of syntactic boundaries in the speech of 4-year-old children with normal and disordered language development", dans *J. of Speech, Lang. and Hearing Research*, vol. 41, pp. 1158–1170, Oct. 1998.

[SOH99] J. Sohn, N. Soo Kim, et W. Sung, "A statistical model-based voice activity detection", dans *IEEE Signal Proc. Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[STE92] N. Stein et K. Oatley, "Basic emotions: Theory and measurement", dans *Cognition and Emotion*, vol. 6, pp. 161–168, 1992.

[STE09] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, PhD thesis, University Friedrich-Alexander, Erlangen-Nuremberg, Germany, 2009.

[STE10] S. Steidl, A. Batliner, D. Seppi et B. Schuller, "On the impact of children's emotional speech on acoustic and language models", dans *EURASIP J. on Audio, Speech, and Music Proc.*, vol. 2010, article ID 783954, 2010.

[STO87] C. Storm et T. Storm, "A taxonomic study of the vocabulary of emotion", dans *J. of Personality and Social Psycho.*, vol. 53, pp. 805–816, 1987.

[SZA09] G. Szaszák, D. Sztahó et K. Vicsi, "Automatic intonation classification for speech training systems", dans proc. *Interspeech*, Brighton, UK, Sep. 6-10 2009, pp. 1899–1902.

[SZC10] B. Szczepek Reed "Speech rhythm across turn transitions in cross-cultural talk-in-interaction", dans *J. of Pragmatics*, vol. 42, no. 4, pp. 1037–1059, Apr. 2010.

[TAY94] P. Taylor, "The Rise/Fall/Connection model of intonation", dans *Speech Comm*, vol. 15, no. 1-2, pp. 169-186, Oct. 1994.

[THU93] C. Thurber et H. Tager-Flusberg, "Pauses in the narratives produced by autistic, mentally retarded, and normal children as an index of cognitive demand", dans *J. of Autism and Develop. Disorders*, vol. 23, no. 2, pp. 309–322, 1993.

[TIL08] S. Tilsen et K. Johnson, "Low-frequency Fourier analysis of speech rhythm", dans *J. of Acoust. Soc. of Amer.*, Express Letters, vol. 124, no. 2, pp. 34–39, Aug. 2008.

[TIL09] S. Tilsen, "Multitimescale dynamical interactions between speech rhythm and gesture", dans *Cogn. Sci.*, vol. 33, pp. 839–879, Jul. 2009.

[TOM84] S. S. Tomkins, "Affect theory", dans K. R. Scherer and P. Ekman [Eds], *Approaches to Emotion*, Erlbaum, Hillsdale, NJ, pp. 163–196, 1984.

[TOT08] S. L. Tóth, D. Sztahó et K. Vicsi, "Speech emotion perception in human and machine", dans A. Esposito, N. G. Bourbakis, N. Avouris and I. Hatzilygeroudis [Eds], *Verbal and Non-verbal features of Human-Human and Human-machine interaction*, Springer-Verlag, *LNCS*, vol. 5024, pp. 213–224, 2008.

[TOU03] H. H. Touma, *La musique des Arabes*, Amadeus Press, 2003.

[TUC92] R. Tucker, "Voice activity detection using a periodicity measure", dans proc. *IEEE Comm., Speech, and Vision, Inst. Elect. Eng.*, vol. 139, no. 4, pp. 377–380, Aug. 1992.

[TVE69] A. Tversky, "Intransitivity of preferences" dans *Psycho. R.*, vol. 76, pp. 31–48, Jan. 1969.

[VER04] A. Veroloes et E. Excoffier, "Dysphasie : aspects génétiques", dans C. L. Gerard et V. Brun [Eds], *Les dysphasies Rencontres en rééducation*, Paris : MASSON, pp. 17–22, 2004.

[VER03] D. Ververidis et C. Kotropoulos, "A review of emotional speech databases", dans 9th *Panhellenic C. in Informatics*, Thessaloniki, Greece, Nov. 1-23, 2003, pp. 560–574.

[VER06] D. Ververidis et C. Kotropoulos, "Emotional speech recognition: Resources, features and methods", dans *Speech Comm.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[VIC01] K. Vicsi, *A multimedia multilingual teaching and training system for speech handicapped children*, Final Annual Report, SPECO-977126, University of Technology and Economics, Departements of Telecommunications and Telematics, 09.1998 - 08.2001, http://alpha.tmit.bme.hu/speech/speco/index.html.

[VIN09] A. Vinciarelli, M. Pantic et H. Bourlard, "Social signal processing: Survey of an emerging domain", dans *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[VLA07] B. Vlasenko, B. Schuller, A. Wendemuth et G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing", dans proc. 2nd *Inter. C. on Affective Comp. and Intel. Interaction*, Lisbon, Portugal, pp. 139–147, Sep. 12-14 2007.

[VOG05] T. Vogt et E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", dans proc. *ICME*, Amsterdam, The

Netherlands, Jul. 6-8 2005, pp. 474–477.

[VOG06] T. Vogt et E. André, “Improving automatic emotion recognition from speech via gender differentiation”, dans proc. *LREC*, Genoa, Italy, May 24-26 2006, pp. 1123–1126.

[VOL05] F. Volkmar, “Handbook of Autism and Pervasive Developmental Disorder”, dans Wiley and Sons, New jersey: Hoboken, 2005.

[WAR96] P. Warren, “Parsing and prosody: An introduction”, dans *Prosody and parcing*, East Sussex, UK: Psychology Press, pp. 1–16, 1996.

[WEL03] B. Wells et S. Peppé, “Intonation abilities of children with speech and language impairments”, dans *J. of Speech, Lang. and Hearing Research*, vol. 46, pp. 5–20, Feb. 2003.

[WHI89] C. Whissel, “The dictionary of affect in language”, dans R. Plutchik and H. Kellerman, [Eds], *Emotion: Theory, Research and Experience*, vol. 4, *The Measurement of Emotion*, pp. 113–131, Academic Press, New York, 1989.

[WIG02] C. W. Wightman, “ToBI or not ToBI ?”, dans proc. *Speech Prosody*, Aix-en-provence, France Apr. 11-13 2002, pp. 25–29.

[WIN79] L. Wing et J. Gould, “Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification”, dans *J. of Autism and Develop. Disorders*, vol. 9, no. 1, pp. 11–29, 1979.

[WIN88] E. Winner, *The point of words: Children’s understanding of metaphor and irony*, dans Cambridge, Harvard University Press, 1988.

[WOO00] K. Woo, T. Yang, K. Park et C. Lee, “Robust voice activity detection algorithm for estimating noise spectrum”, dans *Electronics Letters*, vol. 36, no. 2, pp. 180–181, Jan. 2000.

[WUN93] W. Wundt, *Grundzüge der Physiologischen Psycho.*, dans Leipzig: Verlag von Wilhelm Engelmann, 4th revised Ed. [first published 1873], 1893.

[XIE06] H. Xie et Z. Wang, “Mean frequency derived via hilbert-huang transform with application to fatigue emg signal analysis”, dans *Computer Methods and Programs in Biomedicine*, vol. 82, no. 2, pp. 114-120, May 2006.

[YIL04] S. Yildrim, M. Buhut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee et S. Narayanan, “An acoustic study of emotions expressed in speech”, dans proc. *Interspeech, 8th ICSLP*, Jeju Island, Korea, 4-8 Oct. 2004 (pas de pagination).

[YIL09] S. Yildirim et S. Narayanan, “Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information”, dans *IEEE Trans. on Audio Speech and Lang. Proc.*, vol. 17, no. 1, pp. 2–12, Jan. 2009.

[YIL11] S. Yildirim, S. Narayanan et A. Potamianos, “Detecting emotional state of a child in a conversational computer game”, dans *Computer Speech and Lang.*, vol. 25, no. 1, pp. 29–44, Jan. 2011.

[ZAM98] V. L. Zammuner, “Concepts of emotion: ‘emotionness’ and dimensional ratings of Italian words”, dans *Cogn. and Emotion*, vol. 12, pp. 243–272, 1998.

[ZON09] C. Zong et M. Chetouani, Hilbert-Huang transform based physiological signals analysis for emotion recognition, dans proc. *ISSPIT*, Ajman, UAE, Dec. 14-17 2009, pp. 334–339.

[ZWI90] E. Zwicker et H. Fastl, *Psychoacoustics: facts and models*, dans Springer-Verlag, Heidelberg, 1990.