Optimisation du procédé de création de voix en synthèse par sélection

Didier Cadic

N° d'ordre: 2011PA112076

Thèse de Doctorat

- Spécialité Physique -

Ecole Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

présentée par

Didier Cadic

Optimisation du procédé de création de voix en synthèse par sélection

Soutenue le 10 juin 2011 devant les membres du Jury :

M.	Olivier Boëffard	(Rapporteur, président du jury)
M.	Yoshinori Sagisaka	(Rapporteur)
M.	Christophe d'Alessandro	(Directeur de thèse)
M.	Thierry Moudenc	
M.	Yannis Stylianou	

Thèse préparée au sein des **Orange Labs**
Laboratoire Advertising Solutions, Audience and Profiling
Unité de recherche et développement Voice
2 avenue Pierre Marzin
22 307 Lannion CEDEX

Résumé

Cette thèse s'inscrit dans le cadre de la synthèse de parole à partir du texte. Elle traite plus précisément du procédé de création de voix en synthèse par sélection d'unités. L'état de l'art repose pour cela sur l'enregistrement d'un locuteur pendant une à deux semaines, suivant un script de lecture de plusieurs dizaines de milliers de mots. Les 5 à 10 heures de parole collectées sont généralement révisées par des opérateurs humains, pour en vérifier la segmentation phonétique et ainsi améliorer la qualité finale de la voix de synthèse.

La lourdeur générale de ce procédé freine considérablement la diversification des voix de synthèse ; aussi en proposons-nous ici une rationalisation. Nous introduisons une nouvelle unité, appelée « sandwich vocalique », pour l'optimisation de la couverture des scripts de lecture. Sur le plan phonétique, cette unité offre une meilleure prise en compte des limites segmentales de la synthèse par sélection que les unités traditionnelles (diphones, triphones, syllabes, mots, etc.). Sur le plan linguistique, un nouvel enrichissement contextuel nous permet de mieux focaliser la couverture, sans négliger les aspects prosodiques. Nous proposons des moyens d'accroître le contrôle sur les phrases du script lecture, tant dans leur longueur que dans leur pertinence phonétique et prosodique, afin de mieux anticiper le contenu du corpus de parole final et de rendre automatisable la tâche de segmentation. Nous introduisons également une alternative à la stratégie classique de condensation de corpus en mettant au point un algorithme semi-automatique de création de phrases, grâce auquel nous accroissons de 30 à 40% la densité linguistique du script de lecture.

Ces nouveaux outils nous permettent d'établir un procédé très efficace de création de voix de synthèse, procédé que nous validons à travers la création et l'évaluation subjective de nombreuses voix. Des scores perceptifs comparables à l'approche traditionnelle sont ainsi atteints dès 40 minutes de parole (une demi-journée d'enregistrement) et sans post-traitement manuel. Enfin, nous mettons à profit ce résultat pour enrichir nos voix de synthèse de diverses composantes expressives, multi-expressives et paralinguistiques.

Mots-clefs : Synthèse vocale, sélection d'unités, script de lecture, sandwich vocalique, création de phrases, évaluation, voix, expressivité.

Optimised voice creation for unit-selection synthesis

Abstract

This work falls within the scope of text-to-speech (TTS) technology. More precisely, focus is on the voice creation process for unit-selection synthesis. In a standard approach, a textual script of several thousands of words is read by a speaker in order to generate approximately 5 to 10 hours of useable speech. The recording time is spread out over one or two weeks and is followed by the considerable task of manually revising the phonetic segmentation for all of the speech.

Such a costly and time-consuming process presents a major obstacle to diversifying synthesized voices. In order to increase efficiency in this process, we introduce a new unit, called a "vocalic sandwich", to optimize coverage of the recording texts. Phonetically, this unit better addresses the segmental limitations of unit-selection TTS than state-of-the-art units (diphones, triphones, syllables, words...). Linguistically, a new set of contextual symbols focuses the coverage, allowing for more control and consideration of prosody. Practically, in order to automate the segmentation process, better anticipation of the phonetic and prosodic content desired in the final database is required. This is achieved here by increasing the readability and consistency of each sentence included in the script. As a side, these properties also help to facilitate the reading stage. Furthermore, as an alternative to the classic corpus condensation, a semi-automatic sentence building algorithm is developed in this work wherein sentences are built rather than selected from a reference corpus. Ultimately, the sentence building provides access to much denser scripts, specifically allowing for increases in density of between 30 and 40%.

In incorporating these new approaches and tools, the voice creation process is made very efficient, as is validated in this work through the preparation and evaluation of numerous synthesized voices. Perceptive scores that are comparable to the traditional process are achieved with 40 minutes of speech (half-day recording) and without any manual post-processing. Finally, we take advantage of these results in order to enhance our synthesized voices with various expressive, multi-expressive and paralinguistic features.

Keywords : Text-to-speech, unit selection, recording script, vocalic sandwich, sentence construction, evaluation, voice, expressiveness.

Remerciements

Tout d'abord un grand merci à Thierry pour sa confiance et pour m'avoir permis de mener cette thèse dans de très bonnes conditions. Je remercie également toute l'équipe des Orange Labs pour son soutien, tant sur le plan humain que sur le plan technique. Je pense en particulier à Cédric, avec qui la collaboration sur les problèmes de création de phrases a été fructueuse. Merci également à Christophe d'avoir apporté un regard avisé tout au long du déroulement de cette thèse.

Enfin j'ai une pensée particulière pour toute ma famille. Je remercie chaleureusement Sandrine qui m'a apporté un soutien indéfectible malgré les nombreux sacrifices, Alban qui a su distraire son papa pendant la rédaction de ce mémoire et Bastien qui en a précipité le point final.

Table des matières

Glossaire

ACR Absolute Category Rating

API Alphabet Phonétique International

CART Classification and Regression Tree

CCR Comparison Category Rating

CMOS Comparison Mean Opinion Score

DCR Degradation Category Rating

DMOS Degradation Mean Opinion Score

FSM Finite State Machine

FST Finite State Transducer

GMM Gaussian Mixture Model

GSM Groupe Spécial Mobile, devenu le Global System for Mobile communications

GSSSD Generic-Single-Source-Shortest-Distance (algorithme de recherche de meilleur chemin dans un automate de type WFST)

HMM Hidden Markov Model

HTS « H Triple S », pour HMM-based Speech Synthesis System

LPC Linear Predictive Coding

LSF Line Spectrum Frequencies

LSP Line Spectrum Pairs

MCA Multiple Centroïd Analysis

MFCC Mel Frequency Cepstral Coefficients

MOS Mean Opinion Score

NLG Natural Language Generation

PLP Perceptual Linear Predictive

PSOLA Pitch-Synchronous Overlap Adding

RMSE Root Mean Square Error

RSS Really Simple Syndication (on parle généralement de flux RSS ou fil RSS)

SLA Sclérose Latérale Amyotrophique (ou maladie de Charcot)

SMS Short Message Service

SSML Speech Synthesis Markup Language

TTS Text-to-Speech

VSCR Vocalic Sandwiches Coverage Rate

WFST Weighted Finite State Transducer

W3C World Wide Web Consortium

Introduction

Contexte

Elle est loin, la machine parlante du baron von Kempelen qui permettait, dès la fin du XVIII^{ème} siècle, de produire artificiellement des bribes de parole humaine ! Cet ingénieux système de soufflets, tubes et autres résonateurs avait déjà pris un coup de vieux avec l'apparition en 1939 du système électrique développé aux Bell Labs' par Homer Dudley, le Voder. Mais c'est l'avènement du numérique depuis les années 1970 qui a véritablement changé la donne et ouvert à la synthèse vocale de nouveaux horizons [Schroeder 93][Klatt 87]. L'automatisation du procédé, l'amélioration constante de l'intelligibilité et du naturel de la parole synthétique, ont suscité un intérêt croissant pour ce type de technologie.

Les opérateurs de télécommunications en ont été les premiers vrais utilisateurs. Permettant de convertir automatiquement n'importe quel texte en parole humaine, la synthèse vocale a en effet contribué, par un couplage avec les technologies également émergentes de reconnaissance vocale et de dialogue homme-machine, à l'apparition au tout début des années 90 d'interfaces vocales entièrement automatisées sur des serveurs téléphoniques destinés au grand public [Sorin 92]. Cependant la parole synthétique, encore jugée robotique à l'époque, a été fréquemment écartée de ce type de serveur vocal interactif. Leurs concepteurs lui préféraient la diffusion de messages pré-enregistrés, éventuellement juxtaposés de manière dynamique pour offrir un minimum de souplesse.

Il faudra attendre la fin des années 90 et la généralisation de la technique de synthèse par sélection d'unités (ou synthèse par corpus) pour que l'on atteigne un niveau de qualité susceptible de générer l'adhésion du grand public. Son utilisation sur les serveurs téléphoniques est alors devenue beaucoup plus systématique, au point de devenir quasiment incontournable.

Cantonnée pendant longtemps aux secteurs des télécommunications et du handicap, la synthèse vocale fait désormais son apparition sur des terrains inattendus. En effet, le naturel de la parole restituée a fait naître de nouveaux besoins : lecture de livres, synthèse de cartes de voeux ludiques, création de voix personnalisées, doublage de films ou documentaires, etc. Nous assistons aujourd'hui à une véritable explosion de la demande en voix expressives, voix typées ou encore voix célèbres...

Limites de la synthèse par sélection d'unités

Malheureusement la technologie ne permet pas, dans son état actuel, de répondre à toutes ces attentes. Elle est bridée par au moins deux contraintes majeures : d'une part la lourdeur des procédés de création d'une nouvelle voix, d'autre part la limitation à de la parole de style « lecture neutre ».

En synthèse par sélection d'unités, le signal de synthèse est obtenu par juxtaposition de segments de parole originale, ces derniers étant sélectionnés au sein d'un vaste corpus de parole naturelle [Sagisaka 88] [Hunt 96]. Selon l'état de l'art le corpus d'origine doit contenir plusieurs heures de parole d'un même locuteur, ce qui nécessite une à deux semaines d'enregistrements suivant un script de lecture de plusieurs milliers de phrases. Le script est généralement construit en sélectionnant des phrases parmi un vaste corpus textuel de référence [Van Santen 97b]. On parle alors de « condensation de corpus ». Cette condensation est traditionnellement opérée suivant un critère de couverture en diphones ou triphones [Gauvain 90]. La lecture du script est soumise à des contraintes très fortes : suivi méticuleux, constance de la voix, qualité sonore...

De l'homogénéité de ces enregistrements dépend la compatibilité des segments juxtaposés lors de la synthèse. Le timbre et le ton de la voix doivent donc rester neutres et uniformes tout au long de la lecture du script. Mais quelle que soit la rigueur du locuteur, le corpus final comporte inéluctablement des imprécisions. Pour cette raison on a souvent recours à une coûteuse phase de traitement manuel, qui consiste à réviser la segmentation phonétique de chaque phrase enregistrée.

La lourdeur générale de ce procédé de création de voix freine considérablement la recherche. Elle tend à verrouiller le catalogue de voix et à limiter la prise en compte de composantes expressives. Comme nous le verrons dans cette étude, de nombreux travaux se sont attaqués à ces verrous.

Une partie d'entre eux a porté sur l'amélioration de l'algorithme de sélection des unités. Les critères numériques de sélection ont souvent été remis en cause, que ce soit pour l'anticipation des distorsions acoustiques [Vepa 04] ou pour la mesure de l'adéquation des unités à leur contexte [Taylor 06]. Mais les critères symboliques ont également fait l'objet de nombreuses attentions. Plusieurs études ont proposé d'ajouter des critères d'expressivité dans l'algorithme de sélection, en se reposant sur de nouveaux symboles pour l'annotation du corpus de parole. Ces symboles ont par exemple porté sur les zones d'emphase [Black 03], les actes de dialogues [Syrdal 08] ou encore les émotions [Iida 02]. Un tel enrichissement présente toutefois l'inconvénient majeur de compliquer l'annotation du corpus et de décupler ses besoins de couverture, ce qui accroît la lourdeur du procédé de création de voix.

D'une manière générale lorsqu'on souhaite traiter de composantes expressives en synthèse par sélection, les besoins de matière sonore deviennent tels qu'on ne peut plus envisager de tout enregistrer [Black 03]. Des techniques de modélisation du signal sont requises pour extrapoler le contenu du corpus et compenser les lacunes de sa couverture. Elles peuvent être utilisées à une échelle très locale, par exemple pour lisser les transitions entre unités successives [Stylianou 01a] [Wouters 01] [Pfitzinger 04], ou bien jouer un rôle de plus haut niveau comme l'augmentation de l'expressivité [Raux 03] [Beller 07]. Mais ces techniques sont imparfaites et nuisent au naturel du signal de parole.

Que ce soit pour la réduction des besoins de matière sonore ou pour la prise en compte de nouvelles composantes expressives, la tentation est grande de recourir à des techniques de synthèse paramétrique. Ces techniques, qui reposent sur des modèles d'apprentissage et de génération de la parole, présentent au moins deux avantages : d'une part elles sont assez robustes aux imperfections ou insuffisances de la parole source, d'autre part elles offrent un contrôle accru sur les restitutions synthétiques. C'est en tout cas l'ambition nourrie par les systèmes actuels de synthèse par HMM [Tokuda 95], qui se sont hissés au fil des dernières années parmi les systèmes de référence. Ils peuvent fonctionner avec des corpus d'apprentissage réduits [Tamura 01] voire bruités [Yamagishi 08], et ont montré une certaine aptitude dans la restitution de composantes émotionnelles [Yamagishi 03]. Mais comme tous les modèles paramétriques, ceux de la synthèse par HMM tendent à gommer les aspérités de la parole qui participent pourtant pleinement à son naturel. Il en résulte un timbre de voix bourdonnant et aux consonances artificielles. Ce timbre est à ce jour moins naturel qu'avec la technologie de synthèse par sélection d'unités, du moins lorsque la matière acoustique est adaptée, ce qui est justement l'objet de notre étude.

Organisation du document

Dans la première partie de ce document nous décrivons en détail la synthèse par sélection d'unités, cadre général de nos travaux. Nous détaillons en particulier les procédés traditionnels de création de voix, de la préparation du script de lecture à la compilation des enregistrements.

Nous nous attachons à mieux cerner les tenants et aboutissants de ces procédés, en mettant en lumière les difficultés auxquelles ils font face. Nous approfondissons également la problématique d'évaluation des voix de synthèse, qui occupe une place centrale dans nos travaux à travers de nombreux tests perceptifs.

Dans la seconde partie nous traitons du critère d'optimisation des scripts de lecture. Nous introduisons une nouvelle unité, le sandwich vocalique [Cadic 09], pour remplacer les unités traditionnelles dans les considérations de couverture linguistique. Cette unité offre une prise en compte inédite des principaux facteurs segmentaux et prosodiques qui influencent la qualité de la synthèse par sélection.

Nous nous intéressons en troisième partie à l'algorithme d'optimisation lui-même. Nous en traçons les contours généraux par des considérations quantitatives et qualitatives, puis comparons deux procédés d'optimisation [Cadic 10a]. Le premier, conforme à l'état de l'art, consiste à condenser un corpus textuel de référence. Le second, plus expérimental, consiste à fabriquer des phrases qui maximisent la densité linguistique du script de lecture.

En quatrième partie nous décrivons une vaste expérience perceptive, portant sur 49 voix de synthèses créées suivant des procédés différents. Nous verrons que le soin apporté à la constitution de nos scripts de lecture, et notamment l'optimisation de la couverture en sandwichs vocaliques, permet une réduction drastique des enregistrements nécessaires à la création de voix de synthèse de haute qualité [Cadic 10b]. Ces résultats très favorables ouvrent de nouveaux horizons, notamment pour l'expressivité des voix de synthèse. Nous exposons à ce sujet quelques expériences prometteuses.

Première partie

La synthèse par sélection d'unités

Sommaire

1 Généralités sur la synthèse de parole à partir du texte

1.1 Composants essentiels

La conversion automatique d'un texte en signal de parole suit traditionnellement deux grandes étapes fonctionnelles, comme indiqué sur la figure 1 : les **traitements linguistiques** (ou hauts-niveaux) et les **traitements acoustiques** (ou bas-niveaux). Les premiers visent à convertir un texte d'entrée en une séquence phonétique enrichie de descripteurs linguistiques et prosodiques. Les seconds produisent à partir de cette séquence un signal de parole, correspondant à la vocalisation du texte initial.

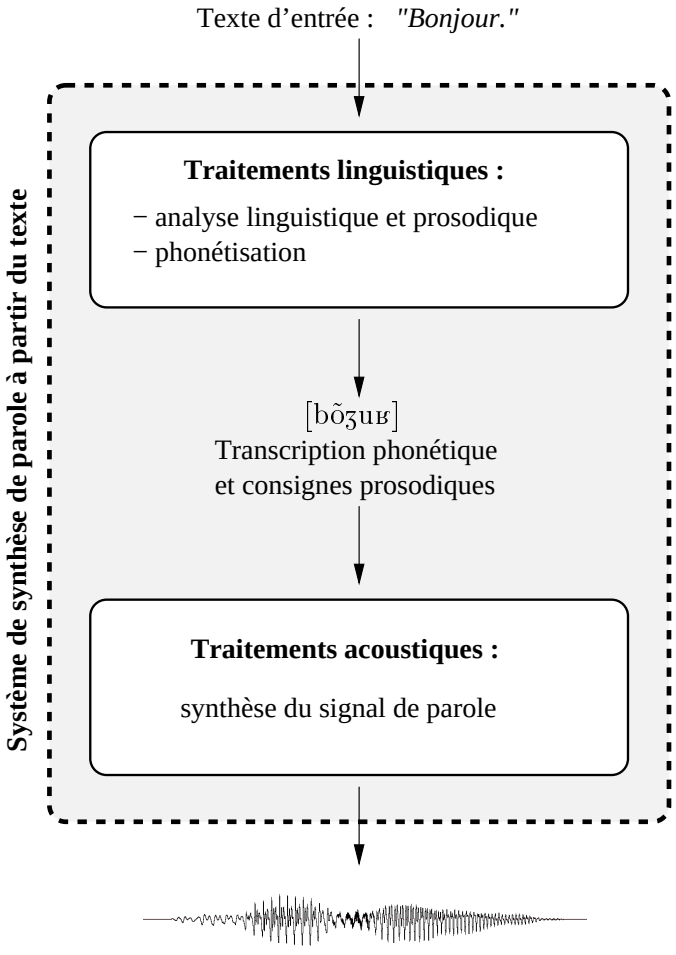


FIGURE 1 – Architecture générale d'un système de synthèse vocale à partir du texte

Dans la suite de la section 1, nous formulons quelques hypothèses sur l'entrée et la sortie du système de synthèse, puis décrivons succinctement les traitements linguistiques utilisés pour le français. Les traitements acoustiques, sur lesquels repose l'essentiel de notre travail, seront détaillés dans la section suivante.

1.2 Typologie de l'entrée

1.2.1 Choix du format textuel

Le travail exposé ici porte exclusivement sur la synthèse de parole à partir du texte, ou Text-To-Speech en anglais (TTS). C'est le cas de figure le plus courant même si, à proprement parler, la « synthèse de parole » peut renvoyer à des domaines très distincts utilisant d'autres types d'entrées : synthèse de parole à partir de murmure [Toda 05] (nam-to-speech où « nam » signifie « Not Audible Murmur »), synthèse de parole à partir d'enregistrements articulatoires [Hueber 07], conversion de voix, et plus généralement tout décodage visant à convertir des paramètres en parole.

Dans certains cas l'entrée textuelle peut être enrichie d'informations contextuelles à travers un langage de balises. Ces informations sont généralement d'ordre paralinguistique, c'est-à-dire qu'elles complètent l'information linguistique contenue dans le texte lui-même, en précisant par exemple des choix prosodiques (points d'emphase, rythme de lecture) ou en spécifiant le contexte dialogique dans lequel s'inscrit le message (informations sémantiques, discursives, pragmatiques). Le groupe W3C (World Wide Web Consortium) a proposé en 2004 un langage de balisage normalisé, SSML 1.0 [1][Burnett 04]. Les balises introduites restent assez limitées et permettent simplement de commander les fonctions classiques des systèmes de synthèse vocale : sélection de la voix, modification de hauteur/tessiture/rythme, exceptions de prononciation, etc. On note tout de même la présence de quelques balises de plus haut niveau, comme l'emphase d'une partie de texte.

Plus généralement, l'enrichissement paralinguistique des données textuelles s'inscrit dans le cadre de ce qu'on appelle la synthèse à partir de concepts, ou Concept-to-Speech en anglais. Un concept est une structure abstraite comportant des informations plus générales et plus riches qu'une simple donnée textuelle. Il peut être produit par des applications de génération automatique d'énoncés textuels (ou NLG [2]), comme le dialogue naturel ou la traduction, qui ont souvent à leur disposition des informations de haut niveau sur le contexte, la sémantique, etc. La notion de synthèse vocale à partir de concepts a été introduite dès 1979 dans [Young 79]. Les auteurs proposaient alors de convertir directement les données conceptuelles générées par un système d'information de distribution d'eau en un signal de parole. Même si l'une des étapes du traitement consistait à générer un contenu textuel, les données sémantiques et syntaxiques fournies par le système d'information restaient disponibles à tout moment, sans qu'elles doivent être inférées de manière imparfaite à partir du texte. Le Concept-to-Speech a été généralisé en 1997 à l'occasion d'un atelier de conférence dédié [Alter 97]. De nombreux schémas d'annotation de concepts, plus ou moins généralistes, ont été proposés pour l'intégration des briques NLG et TTS, sans réelle convergence [Shimei 97] [Hitzeman 99] [Mertens 01] [Xydas 04]. Dans la pratique le concept-to-speech tend à rester un modèle de pensée théorique, certes séduisant mais qui ne parvient pas à s'imposer dans les systèmes réels. Deux verrous technologiques essentiels subsistent : d'une part la génération automatique et pertinente de concepts de hauts-niveaux, d'autre part leur traduction acoustique réaliste.

Une partie de notre travail consistera à s'attaquer au deuxième verrou. En effet en section 14 nous proposerons une méthode d'introduction réaliste d'éléments paralinguistiques (rires, hésitations...) dans le flux de parole synthétique. Cette méthode supposera la présence dans le texte d'un enrichissement paralinguistique, consistant en des balises de commande des éléments à restituer. Mais **tout le reste de notre travail sera cantonné à une entrée textuelle simple, dépourvue d'enrichissement de ce type**.

1. la version 1.1 ayant été officialisée en septembre 2010
2. Natural Language Generation

1.2.2 Hypothèses simplificatrices sur l'entrée textuelle

Hypothèse 1 : un texte en français conventionnel

Chaque application d'un système TTS est associée à un « univers linguistique » spécifique : SMS[3], messagerie instantanée, tweets, courriers électroniques, flux RSS[4], recettes de cuisine, messages d'accueil, livres, pièces de théâtres... Les textes issus de ces univers se distinguent par des usages orthographiques, lexicaux, syntaxiques et structurels qui peuvent être très différents.

[Torzec 01] propose une étude détaillée des difficultés rencontrées par le système CVOX, ancêtre du système Baratinoo sur lequel nous travaillons, pour la vocalisation automatique de courriers électroniques : néologismes, abréviations, fautes d'orthographe, etc., mettent considérablement en défaut les mécanismes de lecture. Ces écueils sont encore plus nombreux lorsqu'il s'agit de vocaliser des SMS. A titre d'illustration nous en rapportons ci-dessous quelques exemples, extraits de [Guimier de Neef 07] et accompagnés d'une proposition de réécriture plus conventionnelle :

`g ht du kfé a+`	→	j'ai acheté du café à plus
`slt k f tu`	→	salut que fais-tu ?
`ssuuuppperr ! hhhhuuuuuummmm !`	→	super ! hum !
`g esayé 2tapelé pl1 2foi`	→	j'ai essayé de t'appeler plein de fois
`tu pe tokup du cha 2m1 ?`	→	tu peux t'occuper du chat demain ?
`je c pa ki c`	→	je sais pas qui c'est
`TU PE VENIR`	→	tu peux venir

Si la syntaxe du français reste à peu près respectée dans les SMS, l'orthographe est quant à elle totalement revisitée. [Bove 05] propose quelques évolutions simples du système de synthèse vocale pour en limiter les défaillances sur ce type d'entrée textuelle. Mais cela reste insuffisant, tant les mécanismes d'écriture des SMS peuvent être complexes [Anis 02] :

- graphies phonétisantes (« `kfé` pour « *café* », « `vi1` » pour « *viens* »)
- étirements graphiques (« `ssuuuppperr` »)
- squelettes consonantiques (« `slt` » pour « *salut* »)
- agglutinations voire absence de séparateur (« `moijaVpEr2pareusiradormir` » pour « *moi j'avais peur de pas réussir à dormir* »)

Pour une réécriture automatique acceptable des SMS, des techniques plus élaborées doivent être mises en place. Nous citerons simplement les travaux intéressants de [Kobus 08] : se reposant sur la proximité des SMS aux mécanismes du langage parlé, les auteurs proposent avec succès d'utiliser des techniques inspirées de la reconnaissance vocale.

Nous supposerons dans toute la suite que les entrées textuelles ne présentent pas d'écueil particulier pour le système de synthèse en français. Il s'agira donc de textes conformes aux règles d'écriture du français standard, ou au moins correctement interprétés par le système.

Hypothèse 2 : des phrases isolées

Pour finir, nous devons faire une dernière hypothèse simplificatrice sur l'entrée textuelle. Observons pour cela l'extrait de *L'Avare* de Molière rapporté en figure 2. Bien que les règles

3. Le « Short Message Service », composant de la norme GSM, permet de transmettre de courts messages textuels depuis un téléphone mobile.

4. « Really Simple Syndication » : format utilisé pour la syndication de contenus web, c'est-à-dire la mise à disposition d'une partie d'un site Web dans d'autres sites. De nombreux flux (ou fils RSS) sont accessibles sur le web.

Scène VI : Harpagon, Elise, Mariane, Frosine.

MARIANE
Je m'acquitte bien tard, madame, d'une telle visite.

ÉLISE
Vous avez fait, madame, ce que je devais faire, et c'était à moi de vous prévenir.

HARPAGON
Vous voyez qu'elle est grande ; mais mauvaise herbe croît toujours.

MARIANE, *bas, à Frosine.*
Oh ! l'homme déplaisant !

HARPAGON, *bas, à Frosine.*
Que dit la belle ?

FROSINE
Qu'elle vous trouve admirable.

HARPAGON
C'est trop d'honneur que vous me faites, adorable mignonne.

MARIANE, *à part*
Quel animal !

HARPAGON
Je vous suis trop obligé de ces sentiments.

MARIANE, *à part.*
Je n'y puis plus tenir.

HARPAGON
Voici mon fils aussi, qui vous vient faire la révérence.

MARIANE, *bas, à Frosine.*
Ah ! Frosine, quelle rencontre ! C'est justement celui dont je t'ai parlé.

FROSINE, *à Mariane.*
L'aventure est merveilleuse.

HARPAGON
Je vois que vous vous étonnez de me voir de si grands enfants ; mais je serai bientôt défait et de l'un et de l'autre.

FIGURE 2 – *L'Avare*, Molière, Acte III, scène VI.

du français aient évolué depuis le XVII^{ème} siècle, cet extrait ne présente aucune difficulté de déchiffrage. La mise en page, la structure dialogique, les indications en italique, et plus particulièrement l'enchaînement des phrases, donnent au texte une cohérence d'ensemble qui permet au lecteur de préciser son interprétation. Idéalement les systèmes de synthèse de parole à partir du texte devraient être capables de décliner ces notions globales en des choix prosodiques adaptés. Mais cette tâche est très ardue car elle requiert un niveau élevé de compréhension du texte. Si quelques travaux de recherche s'y attellent [Suciu 06], il n'en reste pas moins que **la plupart des systèmes de synthèse, dont le nôtre, négligent ces aspects narratifs et fragmentent l'entrée textuelle en une succession de phrases considérées isolément**.

1.3 Typologie de la parole produite

Les travaux présentés dans ce document ne s'appliquent pas à tous les styles vocaux. Nous allons donc à présent préciser le type de parole visé par notre système de synthèse vocale.

En matière de communication orale, la notion d'**expressivité** est souvent invoquée pour caractériser un message vocal ou plus généralement une voix. Cette notion ambiguë mérite toutefois quelques précisions. Voici pour commencer les définitions du terme « expressif » relevées

dans deux dictionnaires courants :

- `Expressif (`*`Petit Larousse`*`) :`

  `Qui exprime avec force une pensée, un sentiment, une émotion.`
- `Expressif (`*`Petit Robert`*`) :`
  1. `Qui exprime bien ce qu'on veut exprimer, faire entendre.`
  2. `Qui a beaucoup d'expression, de vivacité.`

Selon ces définitions, un message vocal peut être qualifié d'expressif dès l'instant où il transmet au récepteur, outre le contenu strictement verbal, des informations sur l'état cognitif de l'émetteur : sentiments, émotions, intentions... Ces informations peuvent être transmises de manière volontaire ou involontaire, et dans le premier cas elles peuvent être naturelles ou simulées. Elles sont essentiellement véhiculées par les traits acoustiques suivants :

- la **prosodie**. Elle rassemble usuellement trois caractéristiques acoustiques du signal de parole : la mélodie (variations de hauteurs), le rythme (variations de durée) et l'intensité (variations d'énergie). Plus précisément Vaissière donne la définition suivante [Vaissiere 80] : « La prosodie recouvre les phénomènes de variations dans l'actualisation des phonèmes (qui sont des unités « idéales ») ; ces variations peuvent être décrites sur le plan acoustique (description de l'évolution de la courbe de fondamental, des durées des segments successifs et de l'intensité comparée des phonèmes), sur le plan perceptif (perception du rythme des phrases, de leur mélodie, d'accent, d'intonation, etc.) et sur le plan fonctionnel (fonction linguistique et paralinguistique de ces variations). ».
- la **qualité vocale** (ou timbre de voix). Elle est l'empreinte du conduit vocal, abstraction faite des variations articulatoires et prosodiques. Elle permet de déceler certains aspects de l'état interne du locuteur, comme par exemple le niveau de stress (voix plus ou moins tendue), l'état de réveil (voix plus ou moins rauque), ou encore un rhume (nez bouché). Campbell propose cette composante comme quatrième dimension prosodique [Campbell 03].
- le **degré de réduction**, proposé par Pfitzinger comme cinquième dimension prosodique [Pfitzinger 06]. Il mesure en quelques sortes l'écart entre les suites phonétiques théorique et réalisée. La situation du locuteur et son état interne peuvent avoir un impact important sur les élisions, assimilations, etc., ainsi que sur le « triangle vocalique » de son articulation [Beller 08].
- les **éléments paralinguistiques**, comme les rires, hésitations, bâillements... Bien qu'ils apparaissent parfois isolément, ils présentent souvent des interactions avec le flux verbal environnant. Ces interactions se traduisent par une modification de la qualité vocale, de la prosodie et de l'articulation [Campbell 07].

Fónagy parle à ce titre d'un double-encodage de la parole [Fónagy 83]. Un premier encodage linguistique, par lequel l'émetteur transpose son message linguistique en une suite de phonèmes, est complété par un second encodage, de nature paralinguistique, qui réalise acoustiquement le premier. On parle également de composante suprasegmentale de la parole [Hockett 42], par opposition à la composante segmentale qui désigne la succession de phonèmes.

Au vu de ces éléments il ne peut pas exister de parole totalement dépourvue d'expressivité. Un engagement minimal du locuteur dans son message vocal est inéluctable. Même une lecture naïve, sans interprétation du contenu sémantique, implique un effort de déchiffrage et d'analyse syntaxique qui, par les choix prosodiques qui en découlent, trahissent en partie l'état cognitif du locuteur. Ainsi un enfant qui adopterait une prosodie trébuchante sur un texte délicat livrerait malgré lui des indications sur son état de stress interne.

Toutefois un lecteur expérimenté a la possibilité, pour la lecture d'un texte sans écueil, de se reposer sur des quasi-automatismes, sans réel engagement de sa part. En ce sens le

style « **lecture neutre** » apparaît chez un lecteur expérimenté comme le mode de vocalisation vraisemblablement le moins expressif, à défaut d'être totalement inexpressif.

L'essentiel de nos travaux portera sur des données vocales de type « lecture neutre », type auquel la quasi-totalité des systèmes actuels de synthèse vocale se limitent. Comme nous le verrons dans la partie IV, cette restriction laisse tout de même une marge de manoeuvre intéressante. En effet les « lectures neutres » diffèrent d'une personne à l'autre, en matière de timbre évidemment, mais aussi en matière de prosodie : lecture plus ou moins lente et articulée, mise en relief plus ou moins forte de la syntaxe, etc.

Nous verrons par ailleurs que les modèles introduits pour la lecture neutre peuvent parfois être utilisés pour de la parole plus « colorée ». Toute composante paralinguistique s'accordant avec les modèles peut être prise en compte simplement par effet de bord, sans traitement spécifique : voix basse, voix sensuelle, voix triste, etc. Dans les sections 13 et 14, **nous aborderons spécifiquement le problème d'une synthèse de parole plus expressive**, point chaud de la recherche à l'heure actuelle.

1.4 Traitements linguistiques

Pré-traitements

Les niveaux linguistiques, aussi appelés hauts-niveaux, effectuent en premier lieu un **pré-traitement du texte** consistant en la réécriture de certaines abréviations, acronymes ou nombres. Pour certaines applications spécifiques, cette phase peut être complétée d'algorithmes élaborés de réécriture de textes mal formés comme les SMS (cf. page 22).

Analyse linguistique

Une analyse morpho-lexicale permet ensuite d'établir une pré-catégorisation grammaticale des mots apparaissant dans le texte d'entrée. Elle se repose sur l'utilisation de lexiques ainsi que de règles de décomposition des mots inconnus en morphèmes (préfixes, racines, suffixes) [Larreur 89]. A ce stade de nombreuses ambiguïtés grammaticales ne peuvent être levées, comme par exemple le caractère verbal ou nominal du mot « *actions* » (pluriel de « *action* », ou conjugaison du verbe « *acter* »).

L'analyse syntaxique lève ces ambiguïtés en appliquant généralement un ensemble de règles grammaticales. Chaque mot se voit ainsi attribuer une unique étiquette morpho-lexicale (ou étiquette POS de l'anglais part-of-speech), comme l'illustre l'exemple de la table 1, extrait de [Krul 08].

Elle	PRONOM personnel, singulier, féminin
aime	VERBE, indicatif, présent, 3ème personne, singulier
beaucoup	ADVERBE
le	DÉTERMINANT, article défini, singulier, masculin
chocolat	NOM commun, singulier, masculin

TABLE 1 – Exemple d'étiquetage morpho-syntaxique de la phrase « *Elle aime beaucoup le chocolat.* »

Puis, dans le cadre d'une « lecture neutre » (donc en l'absence de considérations sémantiques ou paralinguistiques) l'ensemble de ces éléments lexicaux, morphologiques et syntaxiques, permettent d'établir une description symbolique de la structure prosodique du texte d'entrée. Dans

notre système cette étape se fait également suivant un ensemble de règles expertes. Les phénomènes prosodiques sont décrits sur deux niveaux hiérarchiques : le groupe accentuel, qui est un regroupement de mots par unités de sens (on parle aussi de mot prosodique), et le groupe intonatif, constitué de groupes accentuels appartenant à une même phrase prosodique [Rossi 81]. Les groupes intonatifs sont délimités par des ruptures prosodiques comme des pauses, allongements ou contours mélodiques majeurs. Dans la pratique notre module d'analyse prosodique place entre les groupes intonatifs des « pauses mineures » ou des « pauses majeures », mais qui sont en fait traitées de la même manière par les modules acoustiques. Dans toute la suite les groupes intonatifs seront donc assimilés aux groupes de souffles, simplement délimités par deux pauses.

Transcription phonétique

L'étape de transcription phonétique du texte de départ, aussi appelée conversion graphème-phonème, présente de nombreux écueils dont la résolution est facilitée par l'un ou l'autre des niveaux d'analyse précédents [Larreur 89].

Par exemple, la prononciation du graphème « x » varie en fonction du contexte lexical : /s/ dans « *six* », /ks/ dans « *axe* », /gz/ dans « *examen* » et muet dans « *noix* » (notre alphabet phonétique est précisé en annexe). Ce choix relève des analyses lexicale et morpho-lexicale précédentes.

Toutefois ces deux premiers niveaux d'analyse peuvent se révéler insuffisants. En particulier il existe des mots qui s'écrivent de la même manière mais se prononcent différemment (homographes hétérophones) : « *Les poules du **couvent couvent**.* », « *Le vent **est** à l'**est**.* », « *Tu **as** trois **as** dans ton jeu de carte.* ». Dans ces cas précis, ce sont les considérations syntaxiques qui permettent de statuer sur leur prononciation.

Mais ce n'est pas toujours le cas. Par exemple la phrase « *La couturière a perdu ses fils.* » ne peut être lue sans une compréhension du sens de la phrase et de son contexte. S'agit-il de ses bobines de fils (/fil/) ou de ses enfants (/fis/) ? Faute d'analyse sémantique ce cas ne peut pas être résolu dans la plupart des systèmes de synthèse.

Enfin, la conversion graphème-phonème fait face à d'autres difficultés, comme la prononciation des noms propres, des mots d'origine étrangère, etc. [Béchet 00] [Schmidt 93]

La prise en compte de contraintes articulatoires vient ensuite altérer cette prononciation théorique : ajouts de e-muets pour faciliter la prononciation de séquences consonantiques, liaisons (obligatoires, interdites ou facultatives), ou encore effets de coarticulation. Certains de ces phénomènes relèvent plus de considérations allophoniques que phonologiques. En d'autres termes la séquence des unités phonétiques effectivement prononcées (ou allophones, notés entre crochets) présente une certaine variété non reflétée dans la séquence de phonèmes (notés entre barres obliques). La transcription phonétique issue des traitements linguistiques du système TTS est bien de nature allophonique, puisque cette variété y est partiellement modélisée. Le terme « phone » désigne quant à lui l'unité acoustique, c'est-à-dire la réalisation sonore de l'allophone. Pour résumer, plusieurs personnes peuvent utiliser des allophones différents pour prononcer un même phonème : par exemple le [ʀ] grasseyé d'Edith Piaf, le [r] roulé et le [ʁ] « parisien » sont trois variantes du même phonème /ʁ/, puisque ces variations ne conduisent pas à une compréhension différente des mots prononcés. Par ailleurs une même personne qui prononce deux fois le même allophone produit deux phones distincts. Dans un souci de simplicité ces trois notions sont fréquemment assimilées dans la littérature. Nous utiliserons régulièrement le terme « phonème » pour désigner en fait un allophone issu des

hauts-niveaux, ou encore le terme « diphone » pour remplacer les expressions moins courantes « diphonème » et « allodiphone ».

Dans notre système la transcription phonétique est calculée par un ensemble de règles de réécriture contextuelle des graphèmes en phonèmes, utilisant tous les niveaux d'analyse disponibles et complétées par des lexiques de transcription. Bien que certaines règles puissent être inférées automatiquement à partir des lexiques [Bagshaw 98], ce procédé requiert un important travail d'expertise et se transpose donc difficilement d'une langue à l'autre. De nombreux algorithmes d'apprentissage automatique ont été proposés, à base de chaînes de Markov cachées [Taylor 05], de réseaux de neurones [Jensen 00], ou encore de prononciation par analogie [Dedina 91] [Yvon 96].

Prédiction prosodique

A ce stade, chaque unité phonétique est associée à une description symbolique de son contexte syntaxico-prosodique. Reprenons l'exemple de la table 1 :

Phrase textuelle : « *Elle aime beaucoup le chocolat.* »
Transcription phonétique : [ɛlɛmbokuløʃokola]

Dans cet exemple, une description du contexte syntaxico-prosodique de la voyelle finale [a] pourrait être : « voyelle d'une syllabe de type CV[5] en position finale d'un nom commun, lui-même situé en fin de groupe intonatif et dans un contexte affirmatif ».

Dans la plupart des systèmes de synthèse, cette représentation symbolique ne suffit pas à restituer une parole naturelle. Un module de prédiction prosodique fournit en complément une courbe intonative et un schéma rythmique, inférés à partir des données symboliques. De nombreux modèles et méthodes peuvent être utilisés pour calculer l'évolution temporelle de ces paramètres acoustiques.

Emerard propose de répertorier directement l'intonation de chaque mot dans une table en fonction de l'étiquette linguistique et du nombre de syllabes [Emerard 92]. Le contenu de cette table est établi de manière experte en se reposant sur l'observation de corpus de parole spécifiques. Une technique similaire peut être utilisée pour prédire l'allongement de chaque phonème par rapport à sa durée moyenne [Bartkova 87].

De nombreux algorithmes d'apprentissage automatique ont également été proposés. Le plus courant d'entre eux est sans doute celui de Donovan qui propose, pour chaque paramètre prosodique, d'apprendre un arbre de classification sur la base d'un corpus de parole étiqueté linguistiquement (en général le même que celui utilisé pour les traitements acoustiques), puis de modéliser chaque feuille avec des chaînes de Markov cachées [Donovan 96]. Les valeurs prédites correspondent alors aux moyennes observées dans les classes de l'arbre issues du même contexte. [Boidin 08] va plus loin en regroupant, au sein de chaque feuille de l'arbre, les réalisations prosodiques qui suivent des motifs différents. Ces regroupements sont matérialisés par des états cachés d'une chaîne de Markov, dont les transitions permettent de prendre en compte, de syllabe en syllabe, des phénomènes intonatifs globaux.

Parfois une étape intermédiaire de modélisation prosodique est ajoutée. Elle consiste à calculer dans un premier temps les paramètres d'un modèle prosodique, avant d'en déduire, dans un deuxième temps, des valeurs de fréquence fondamentale ou durée. On peut utiliser pour

5. les structures syllabiques sont notées V, CV, VC, CCV, etc., où C et V signifient respectivement « Consonne » et « Voyelle ». Ainsi le mot « *Bonjour* » se décompose en deux syllabes, [bõ] et [ʒuʁ], de structures respectives CV et CVC.

cela les paramètres numériques du modèle de Fujisaki, qui décrit l'intonation comme une superposition de deux phénomènes intonatifs élémentaires : les formes de groupe et les formes d'accent [Fujisaki 04]. On peut également utiliser la paramétrisation symbolique du modèle phonologique ToBI [Silverman 92], basé sur les travaux de Pierrehumbert [Pierrehumbert 83]. ToBI décrit la prosodie comme une succession d'évènements de trois types : accents intonatifs (« pitch accent »), tons (« tones »), et indices de coupures (« break indices »). Une autre paramétrisation courante est celle du modèle Tilt [Taylor 00], qui propose deux symboles pour désigner respectivement les accents et frontières prosodiques, eux-mêmes décrits temporellement par des jeux de paramètres acoustiques.

2 L'avènement de la synthèse par sélection d'unités

Nous nous intéressons à présent aux modules de traitements acoustiques, qui ont pour objet la génération d'un signal de parole à partir des données phonétiques et prosodiques issues des hauts-niveaux. Après un tour d'horizon des techniques existantes, nous détaillerons celle qui a motivé notre travail et qui est aussi la plus utilisée à l'heure actuelle : la synthèse par sélection d'unités, ou synthèse par corpus.

2.1 Les anciennes méthodes

2.1.1 La synthèse articulatoire

Les techniques de synthèse articulatoire ont pour ambition de reproduire numériquement le fonctionnement du conduit vocal humain [Rubin 81]. Elles sont donc les lointains successeurs de la machine parlante du baron von Kemperlen, qui avait une approche similaire mais avec des procédés mécaniques. Une modélisation complète du conduit vocal est requise, du larynx aux lèvres en passant par la cavité nasale.

Ces techniques présentent de grandes vertus explicatives sur les mécanismes de production de la parole, ainsi qu'un contrôle fin de la position des différents articulateurs : lèvres, mâchoire inférieure, joues, cordes vocales, etc. De tels modèles articulatoires sont détaillés dans [Mermelstein 73] et [Maeda 79]. Toutefois la construction d'un signal de parole à partir de ces paramètres articulatoires est une étape délicate et souvent coûteuse en temps de calcul. Au prix généralement de nombreuses approximations, la résolution d'équations différentielles complexes de mécanique des fluides permet d'estimer l'onde de pression acoustique rayonnée au niveau des lèvres, et d'en déduire ainsi un signal de parole synthétique [Richard 95].

D'une manière générale, la lourdeur du procédé et la faible qualité de la parole restituée font que les systèmes de synthèse articulatoire sont restés jusqu'à ce jour cantonnés à des travaux de recherches. Avec l'augmentation des puissances de calcul, ils pourraient cependant revenir un jour sur le devant de la scène. Le lecteur trouvera d'ailleurs à l'adresse suivante quelques démonstrations prometteuses : `http://sal.shs.arizona.edu/~bstory/`.

2.1.2 La synthèse par règles

La technique de synthèse par règles se veut moins ambitieuse et plus pragmatique que la synthèse articulatoire. Elle repose d'une part sur un modèle de décomposition paramétrique du signal de parole, d'autre part sur un ensemble de règles qui régissent l'évolution temporelle de ces paramètres, sans entrer dans le détail du processus phonatoire.

Généralement la modélisation acoustique utilise une décomposition source-filtre, où la source représente un flot d'air turbulent ou un train d'impulsions glottiques, et où le filtre modélise l'action du conduit vocal. Ce dernier est souvent réduit à ses paramètres « formantiques », d'où l'appellation fréquente de « synthèse par formants ». Les formants sont les fréquences de résonance du conduit vocal. Ils sont étroitement liés à sa forme et donc à la nature du son prononcé. On les caractérise généralement par trois paramètres : fréquence, amplitude et bande passante. La génération de parole à partir de paramètres formantiques (et non à partir de texte) a été initiée par le Voder de Dudley en 1939 et est restée en vogue jusqu'au début des années 1990. Parmi les synthétiseurs par formants les plus célèbres on peut citer PAT [6] de Walter Lawrence [Lawrence 53], OVE [7] de Gunnar Fant [Fant 53] et, un peu plus récemment, le synthétiseur de Klatt [Klatt 80].

Le calcul des trajectoires formantiques à partir des séquences phonétique et prosodique est également une étape délicate ; ces trajectoires doivent en effet refléter une articulation naturelle du conduit vocal. Un jeu de règles expertes permet d'établir l'évolution temporelle des paramètres formantiques et de la source. Parmi les systèmes complets de synthèse par règles, on peut citer le système anglais MITalk [Allen 87] basé sur le synthétiseur de Klatt, ainsi que le système multilingue Infovox [Carlson 82] basé sur une évolution de OVE [Liljencrants 68].

Longtemps fer de lance de la synthèse vocale, les systèmes par règles et/ou par formants sont aujourd'hui peu utilisés du fait de leur timbre de voix robotique, moins naturel que les solutions présentées ci-dessous.

2.1.3 La synthèse par concaténation de diphones

La synthèse par concaténation de diphones, ou plus simplement synthèse par diphones, est née du double constat suivant :

- d'une part, les transitions entre phones ont un impact très important sur l'intelligibilité et le naturel perçus d'un signal de parole [Harris 53],
- d'autre part les modèles acoustiques existants ne permettent pas suffisamment de rendre compte de la complexité d'un signal de parole naturel, en particulier dans les zones transitoires.

De ce dernier point découle l'idée d'utiliser des unités acoustiques pré-enregistrées par un locuteur et stockées dans une mémoire informatique. La génération du signal consiste alors à récupérer les unités adéquates et à les juxtaposer (ou concaténer). Afin de préserver les zones transitoires entre phones, le choix de l'unité élémentaire s'est logiquement porté sur le diphone, c'est-à-dire l'unité acoustique qui s'étend du milieu d'un phone au milieu du phone suivant (figure 3). Ses frontières appartiennent donc à des zones acoustiquement stables, ce qui facilite les concaténations. Une telle unité a été introduite spécifiquement pour la synthèse dans [Peterson 58], à l'époque sous le nom de « dyad ». L'appellation « diphone » est apparue plus tard, dans [Dixon 68].

Pour un alphabet (usuel en français) de 35 phonèmes le nombre théorique de diphones est de $35 \times 35 = 1225$, mais dans la pratique une centaine de diphones inutilisés peuvent être exclus. Une occurrence de chaque diphone est enregistrée puis stockée dans le dictionnaire acoustique du système de synthèse. Pour pouvoir être enchaînés, ils doivent tous être enregistrés avec la même voix, à la même hauteur, à la même vitesse, etc. Un contexte phonétique « neutralisant » de type logatome [8] est généralement utilisé lors de la lecture par le locuteur.

6. Parametric Artificial Talker
7. Orator Verbis Electris
8. Courte séquence phonétique dépourvue de sens

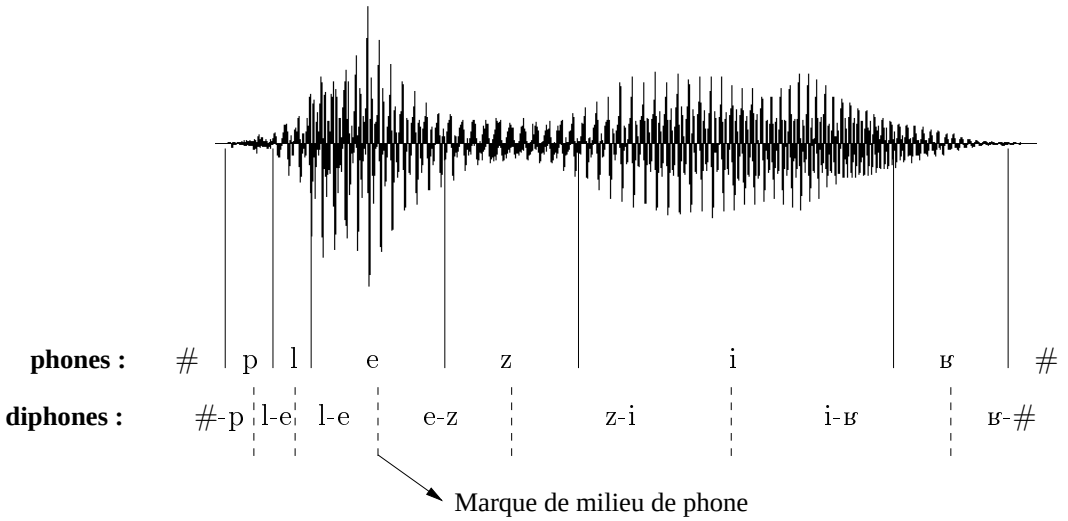


FIGURE 3 – Illustration d'un découpage en diphones, ici sur le mot « *plaisir* ». Le symbole # est une indication de pause.

Naturellement la parole obtenue par simple concaténation de ces diphones est dépourvue de prosodie : mélodie plate et rythme constant. Des algorithmes de traitement du signal sont alors utilisés pour plaquer une prosodie calculée par les hauts-niveaux. La prédiction de trajectoires prosodiques naturelles constitue alors une première difficulté des systèmes de synthèse par concaténation de diphones, les techniques rapportées en section 1.4 pouvant se révéler insuffisantes. Puis le plaquage des ces trajectoires sur le signal de parole représente un écueil supplémentaire. La méthode LPC [9] était initialement la plus utilisée, bien qu'elle dégradât significativement le signal. La mise au point de la technique PSOLA [10] [Moulines 90], plus souple et performante, a changé la donne et démocratisé la synthèse par diphones.

La figure 4 résume le fonctionnement d'un tel système de synthèse. Si la restitution d'un signal de parole par concaténation de diphones a été expérimentée dès la fin des années 60, il a fallu attendre les années 80 et l'arrivée de l'informatique pour observer les premiers sytèmes TTS complets (hauts-niveaux et bas-niveaux) fonctionnant sur ce principe [Olive 85].

Avec cette technique le signal de parole reste fortement pénalisé par plusieurs facteurs :

- Chaque phone fait l'objet d'une concaténation qui s'accompagne d'une discontinuité acoustique plus ou moins importante.
- La prosodie est artificielle, puisque calculée par les hauts-niveaux.
- Les techniques de placage qui permettent de restituer cette prosodie nuisent au naturel du signal (même avec PSOLA).
- Les phénomènes de coarticulation à long terme, qui s'étendent sur plusieurs phonèmes, sont mal pris en compte ; Pols rapporte par exemple des problèmes sur certains clusters consonantiques [Pols 87].

2.2 La synthèse par HMM

La synthèse par HMM (Hidden Markov Model) est une technique récente [Yoshimura 00], basée sur les travaux de Tokuda [Tokuda 95], dont le fer de lance est actuellement le système

9. Linear Predictive Coding
10. Pitch-Synchronous Overlap Adding

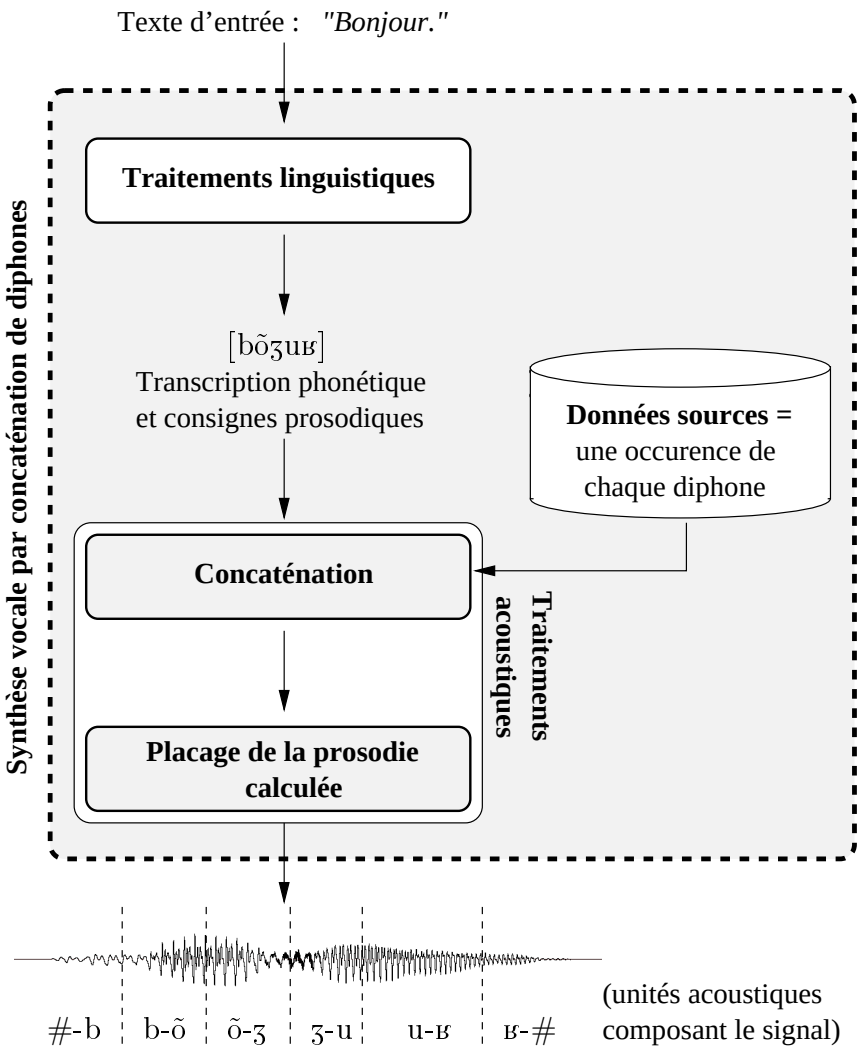


FIGURE 4 – Schéma de principe d'un système de synthèse par concaténation de diphones

HTS [11] développé par le groupe du même nom [12]. Cette technique de synthèse, rendue possible par l'augmentation de la puissance de calcul des ordinateurs, peut être considérée comme une évolution lointaine des synthétiseurs par formants.

D'une part, la représentation source-filtre du signal acoustique va bien au-delà des simples paramètres formantiques. Les coefficients MFCC [13] [Davis 90] sont généralement utilisés, enrichis de leurs dérivées temporelles ainsi que d'autres paramètres suivant les versions : coefficients d'apériodicité [Zen 05], paramètres de source [Cabral 08], filtres de mise en forme de la source [Maia 07], etc. Cette représentation plus complète permet de générer un signal de parole de meilleure qualité qu'avec les anciens synthétiseurs par formants.

D'autre part, les règles expertes de prédiction des paramètres acoustiques ont laissé place à de puissants algorithmes d'apprentissage. A partir d'un corpus de parole annoté, des modèles HMM sont appris conjointement à des arbres de décision. Ces arbres visent à regrouper les états HMM (entre 3 et 7 états par phonème) en fonction de leurs caractéristiques acoustiques et suivant des critères contextuels (phonétique, syntaxe...). Lors de la synthèse d'une phrase

11. « H Triple S », pour HMM-based Speech Synthesis System
12. Groupe d'experts coordonné par le Department of Computer Science du Nagoya Institute of Technology (NITECH) http://hts.sp.nitech.ac.jp/?Home
13. Mel Frequency Cepstral Coefficients

donnée, les consignes issues des hauts-niveaux permettent, en suivant les arbres de décision, d'obtenir des densités de probabilités dépendantes du contexte pour chaque paramètre acoustique. Une étape de décodage acoustique permet ensuite de générer les trajectoires des différents paramètres sur l'ensemble de la phrase.

Pour affiner la prédiction, un apprentissage multi-locuteurs peut être opéré au préalable sur un vaste ensemble de corpus acoustiques (typiquement 6 bases mono-locuteurs contenant chacune une dizaine d'heures de parole). Dans ce cas une deuxième phase d'adaptation des modèles permet de spécialiser la prédiction suivant le corpus d'un locuteur donné [Tamura 01]. Cette approche en deux temps améliore la robustesse des modèles, en particulier lorsque le corpus du locuteur cible est réduit (moins d'une heure de parole).

La figure 5 reprend le principe de fonctionnement d'un système de synthèse vocale par HMM.

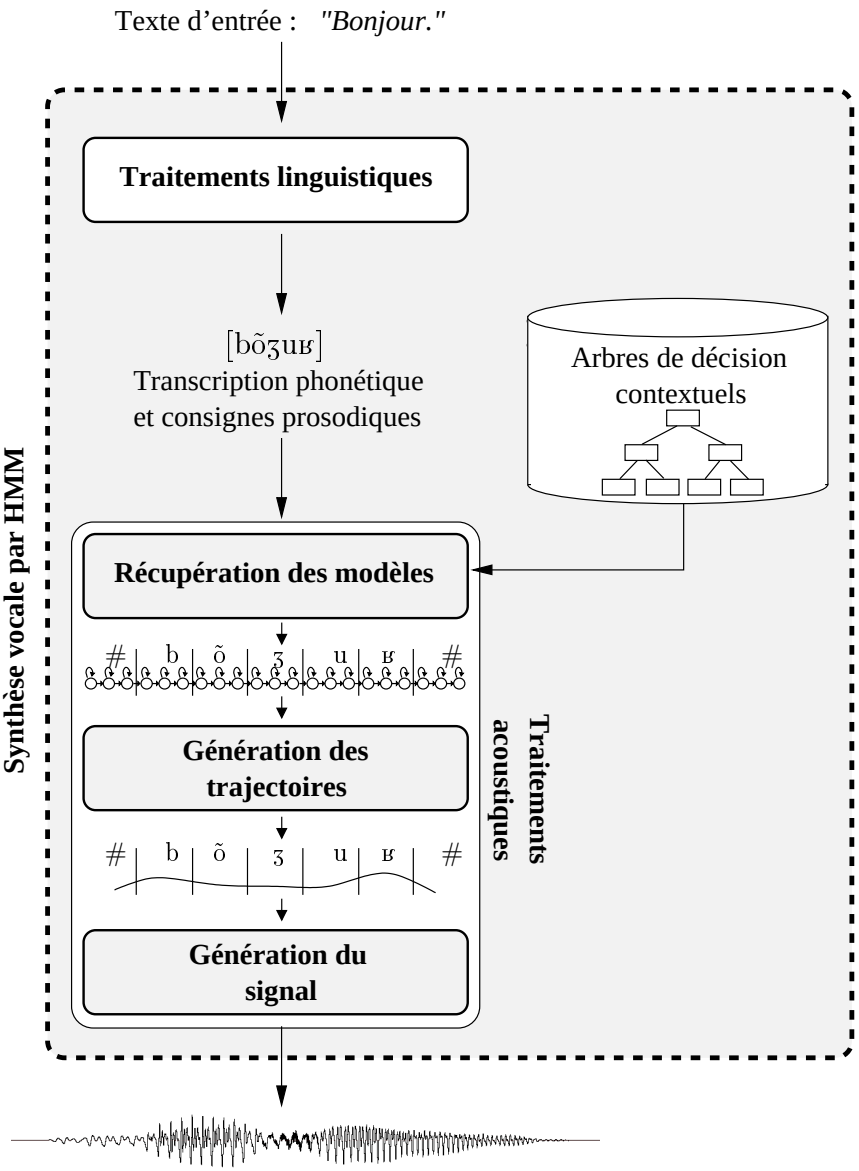


Figure 5 – Schéma de principe d'un système de synthèse par HMM

D'une manière générale ces sytèmes se sont hissés au fil des dernières années parmi les systèmes de référence. La synthèse par HMM fait preuve d'une grande flexibilité : synthèse à

partir de corpus réduits, à partir d'enregistrements bruités [Yamagishi 08], avec une intelligibilité quasi garantie en toutes circonstances. Toutefois le timbre de voix restitué souffre toujours d'un léger effet de « bourdonnement » qui rappelle son caractère artificiel. La technologie de synthèse par sélection d'unités présentée ci-dessous permet à ce jour de véhiculer une identité vocale plus fidèle à l'original, du moins lorsque la matière acoustique disponible est adaptée et en quantité suffisante, ce qui est l'objet de notre travail.

2.3 La synthèse par sélection d'unités

Nous allons maintenant détailler la technique de synthèse par sélection, cadre général de notre étude. Après un exposé des fondements de cette technique (paragraphe 2.3.1), nous présenterons l'algorithme de sélection (paragraphe 2.3.2) puis l'étape de traitement acoustique (paragraphe 2.3.3).

2.3.1 Fondements

Comme expliqué plus haut, la synthèse par concaténation de diphones souffre d'un manque de cohérence à long-terme, entre autres à cause de concaténations trop fréquentes (sur chaque milieu de phone). Le besoin d'utiliser des unités plus longues s'est donc vite fait ressentir. Les triphones, quadriphones, syllabes [Matoušek 05], ou encore les mots [Stöber 99] [Vosnidis 01] peuvent être envisagés mais un problème combinatoire se pose alors : la masse d'enregistrements nécessaire pour couvrir l'ensemble de ces unités devient rédhibitoire. A titre d'exemple il existe en théorie $35^3 = 42875$ triphones et $35^4 = 1500625$ quadriphones en français, sans tenir compte de la multiplicité des contextes ! Certains triphones ou quadriphones sont cependant beaucoup plus rares que d'autres, voire totalement inutilisés ; il n'est donc probablement pas nécessaire de tous les avoir à disposition dans la base de données.

Sagisaka a introduit pour la première fois en 1988 la notion de synthèse vocale par concaténation d'unités de longueur variable [Sagisaka 88], satisfaisant ainsi à la fois le besoin de recourir à des unités longues et la contrainte sur la taille des enregistrements. Cet article n'exposait toutefois qu'un schéma général de synthèse vocale, sans préciser le mode opératoire complet.

Cette lacune a été comblée en 1996 par Hunt et Black [Hunt 96] dans le système CHATR [Black 94]. Leurs expériences ont porté sur des bases de parole d'origines diverses (par exemple des annonces radios), contenant de 10 minutes à 1 heure de parole utile, pour les deux sexes, en anglais et en japonais [Black 95]. Dans chacune de ces bases de nombreuses unités, qu'il s'agisse de phones, diphones, ou polyphones, sont multi-représentées et dans des contextes linguistiques et prosodiques variés. Un algorithme de sélection d'unités doit donc être mis en place, afin de sélectionner, pour chaque phrase d'entrée, les unités acoustiques à concaténer. Cet algorithme doit bien sûr rechercher les unités qui présentent des caractéristiques proches de celles attendues dans la phrase de synthèse (en termes de contexte phonétique, hauteur, durée...), mais aussi qui minimisent les distorsions liées aux concaténations. L'apport principal de Hunt et Black a été d'établir une analogie avec la problématique de reconnaissance de la parole, et de proposer ainsi une approche d'optimisation par programmation dynamique pour la sélection d'unités non uniformes dans une large base de parole naturelle. Cette approche sera détaillée un peu plus bas.

Avec cette technique de (re-)construction du signal vocal, les modules de prédiction des paramètres prosodiques n'ont pas le même rôle qu'avec les autres techniques. Ils ne sont plus utilisés pour la génération du signal, mais pour le guidage de l'algorithme de sélection. Dans ce cadre, la prédiction de trajectoires acoustiques précises n'est plus un pré-requis. Certains

systèmes opèrent même leur sélection uniquement suivant des marqueurs contextuels symboliques. Si les paramètres numériques sont parfois utilisés pour la correction des unités retenues (placage prosodique, lissage des concaténations), l'altération du signal de parole reste dans tous les cas limitée. C'est le principal intérêt de la synthèse par sélection.

Bien entendu cette technique a pu voir le jour grâce à la disponibilité de ressources informatiques suffisantes. D'une part il faut pouvoir stocker de manière efficace de grandes quantités de signal, d'autre part l'algorithme de sélection requiert des capacités de calcul importantes. Avec l'augmentation des mémoires informatiques et des puissances des processeurs, la taille des corpus de parole utilisés a constamment augmenté et dépasse aujourd'hui régulièrement les 10 heures. La qualité de la parole synthétique est en effet directement liée à la richesse du corpus. Plus celui-ci est vaste, plus les segments de parole sélectionnés peuvent être longs et adaptés au contexte, ce qui minimise les distorsions. Pour cette raison la synthèse par sélection d'unités est également appelée « synthèse par corpus ». Aujourd'hui la taille du corpus vocal est davantage limitée par son coût d'acquisition que par les capacités matérielles. A titre d'exemple 10 heures de parole utile nécessitent environ 40 heures d'enregistrements, dans des conditions très contraignantes, et ces enregistrements sont souvent complétés d'une étape fastidieuse (et donc coûteuse) de post-traitement manuel des données.

La parole produite peut être de très haute qualité, voire indiscernable du naturel (au moins sur des textes courts d'une dizaine de mots). Le signal de sortie peut même être entièrement naturel, dans le cas extrême où la phrase à synthétiser est déjà présente dans le corpus. Toutefois, et c'est probablement son défaut majeur, la synthèse par corpus n'apporte aucune garantie sur le niveau de qualité en sortie. En fonction de la phrase d'entrée et du contenu du corpus, les segments de parole concaténés peuvent être plus ou moins longs et plus ou moins adaptés au contexte, ce qui conduit à des artefacts sonores plus ou moins nombreux et gênants. D'une phrase à l'autre le résultat peut ainsi se montrer très variable.

La figure 6 reprend les grandes étapes d'un système de synthèse par corpus. L'étape de correction du signal est souvent allégée, voire supprimée. Dans notre système, seul un lissage des unités aux endroits de concaténation est opéré. En l'absence de placage prosodique, la prosodie restituée est dite « intrinsèque ».

A ce jour, les systèmes de synthèse par corpus les plus connus sur le plan international sont ceux de Nuance, Acapela, Loquendo, ainsi que le logiciel libre Festival [14].

2.3.2 L'étape de sélection

L'étape de sélection des unités peut être décomposée en trois traitements successifs : la définition d'une séquence-cible (qui implique le choix d'une unité élémentaire), la présélection, et enfin l'algorithme de sélection proprement dit. Ces trois phases sont détaillées ci-dessous.

Définition de la séquence-cible

L'algorithme de sélection d'unités repose avant tout sur le choix d'une **unité élémentaire**. Celle-ci correspond au plus petit segment de parole qui puisse être extrait de la base de données en vue d'une concaténation avec d'autres segments. Comme expliqué plus haut, les diphones constituent assurément l'unité de prédilection pour la synthèse par concaténation. En préservant les zones transitoires, ils forcent la concaténation sur des zones spectralement

14. Développé au CSTR d'Edimbourg, le logiciel Festival implémente plusieurs autres techniques de synthèse vocale.

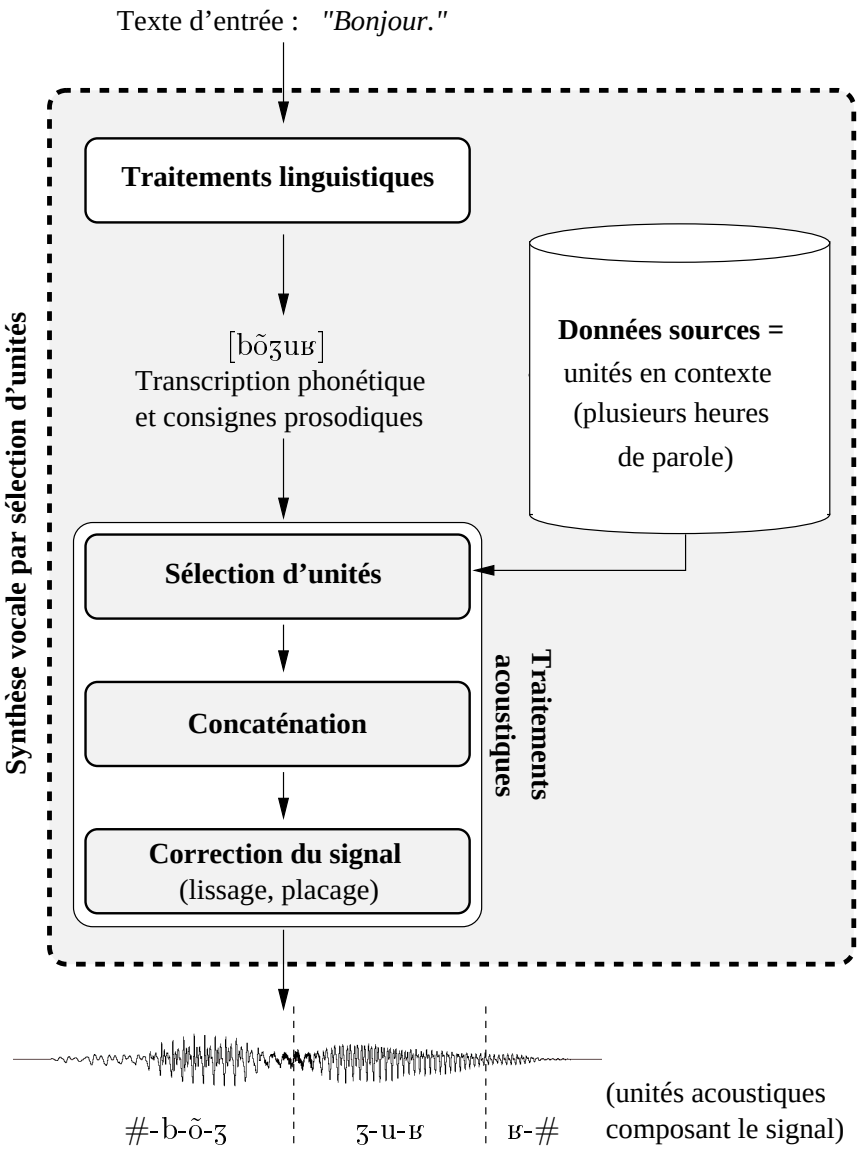


FIGURE 6 – Schéma de principe d'un système de synthèse par sélection d'unités, ou « synthèse par corpus ».

stables (milieux de phonèmes ou parties sourdes des occlusives), ce qui minimise les risques de discontinuité entre unités successives.

Des unités plus petites peuvent être utilisées, comme par exemple les demi-phones [Conkie 99], ou encore les fénèmes qui désignent chacun des 2 ou 3 états du modèle HMM associé à chaque phonème [Donovan 98]. L'intérêt de ces unités sub-phonémiques est d'offrir au module de sélection plus de liberté pour le placement des concaténations, ce qui est particulièrement appréciable dans certains cas de carence en diphones. Mais ce gain de souplesse se fait au prix d'une augmentation de la charge de calcul lors de la sélection. En effet d'une part le nombre d'unités constituantes d'une phrase est multiplié par 2 ou 3, d'autre part le nombre de candidats pour chaque unité augmente considérablement. Il y a par exemple beaucoup plus de « moitiés droites de [a] » et de « moitiés gauches de [b] » que de diphones [a-b]. Cela impose le recours à des techniques de présélection drastiques qui visent à éliminer une grande partie des unités candidates avant même de procéder à l'optimisation par programmation dynamique, ce qui tend fortement à restreindre la continuité du signal synthétique.

De manière générale, les systèmes reposant sur des unités sub-phonémiques opèrent la plupart des concaténations sur les zones stables et simulent donc indirectement un fonctionnement par diphones. Et bien qu'un peu plus contraignants, les diphones laissent la plupart du temps une latitude suffisante pour la construction d'enchaînements phonétiques complets. **L'ensemble de notre travail porte sur un système de synthèse par sélection d'unités à base de diphones**. Cependant les outils développés dans cette thèse pourront facilement être adaptés à des unités plus petites.

Les consignes phonétiques et prosodiques établies par les hauts-niveaux à partir de la phrase d'entrée doivent être projetées sur ce support élémentaire qui est désormais le diphone. On obtient ainsi une séquence-cible de diphonèmes[15] en contextes. Comme expliqué en section 1.4, l'enrichissement contextuel peut inclure des composantes symboliques et numériques très variées. Pour chacun des 2 demi-phonèmes composant chaque diphonème, le vecteur contextuel peut par exemple contenir les informations symboliques suivantes :

- type de groupe prosodique
- position dans le groupe prosodique
- type de mot prosodique
- position dans le mot prosodique
- étiquetage morpho-syntaxique du mot englobant
- structure de la syllabe englobante
- traits phonétiques environnants (par exemple « nasalité du phonème droit »)

ainsi que les traits acoustiques suivants :

- hauteur, voire courbe fréquentielle, pour les parties voisées
- durée
- énergie du signal
- caractérisation spectrale (LPC, MFCC)
- qualité vocale

Présélection

Pour chaque diphonème-cible, il peut exister de nombreux candidats dans la base de données vocale. Chacun de ces candidats correspond au bon diphonème, mais dans un contexte et avec des caractéristiques généralement différents. Le but de l'étape de présélection est de ne conserver que les candidats « crédibles », c'est-à-dire ceux qui présentent les caractéristiques les plus proches de la cible. Cela permet essentiellement d'alléger la charge calculatoire de l'algorithme de sélection. La réduction du nombre de candidats peut être plus ou moins sévère, suivant les stratégies et contraintes matérielles. Une technique classique de présélection consiste à classifier au préalable toutes les unités de la base de données dans un arbre de décision contextuel, et à ne retenir lors de la synthèse que les unités appartenant à la feuille la plus proche du contexte cible [Black 97].

Dans le cas où l'unité élémentaire est le diphone, la charge calculatoire est plus réduite qu'avec des unités sub-phonémiques et on peut se restreindre, en guise de présélection, à quelques règles de filtrage triviales (comme par exemple l'exclusion des unités de fin de groupe intonatif lorsque la cible est en milieu de phrase). Les autres techniques complexes de présélection peuvent alors être vues comme un cas particulier du coût-cible utilisé dans l'algorithme de Viterbi (voir plus bas). **Pour ces raisons nous négligerons dans toute la suite l'étape de présélection.**

15. Le « diphonème » renvoie à une unité phonétique abstraite tandis que le « diphone » désigne une réalisation acoustique d'un diphonème (voir page 26).

Algorithme de Viterbi

Il s'agit maintenant de détecter, parmi tous les diphones disponibles dans la base de données vocale du locuteur (et après présélection éventuelle), la séquence de diphones qui permet de reconstruire, par concaténation, le signal de parole le plus naturel possible.

Deux critères essentiels guident cette sélection :
- l'adéquation des unités aux caractéristiques phonétiques, syntaxiques et prosodiques issues des hauts-niveaux
- la minimisation des distorsions liées aux concaténations, tant dans leur nombre que dans leur amplitude

Dans l'approche par programmation dynamique proposée par Hunt, Black et Campbell, ces deux critères sont matérialisés par des contraintes locales agissant sur un treillis d'unités candidates. La figure 7 illustre un tel treillis de diphones ; dans la réalité il peut y avoir plusieurs milliers de candidats pour chaque diphonème-cible. Dans ce treillis les contraintes prennent la forme d'un **coût-cible** et d'un **coût de concaténation**. Le coût-cible pénalise chaque candidat en fonction de l'écart entre ses caractéristiques et celles de la cible correspondante. Le coût de concaténation porte quant à lui sur les transitions entre candidats successifs (symbolisées par des flèches sur la figure) : chaque transition se voit attribuer un coût quantifiant la distorsion acoustique qui résulterait d'une concaténation entre les deux diphones. Bien entendu ce dernier coût est nul pour les diphones qui sont déjà voisins dans la base vocale, ce qui permet au passage de privilégier la sélection de segments de parole longs (on parle dans ce cas d'unités « contiguës »). La teneur de ces fonctions de coût sera abordée à la section 2.4.

Plus précisément, si $(\hat{d_1}, ..., \hat{d_N})$ désigne la séquence des N diphonèmes-cibles en contexte et si, pour tout $n \in [\![1; N]\!]$, $\{d_n^1, ..., d_n^{K_n}\}$ désigne l'ensemble des K_n candidats pour l'unité $\hat{d_n}$, alors on recherche la séquence d'unités optimale D^* qui minimise la somme C des coûts-cibles $\mathrm{C_{cible}}$ et des coûts de concaténation $\mathrm{C_{concat}}$:

$$D^* = \mathrm{argmin}_{\alpha_1,...,\alpha_N} \mathrm{C}(\mathrm{d}_1^{\alpha_1}, ..., \mathrm{d}_\mathrm{N}^{\alpha_N}) \tag{1}$$

$$= \mathrm{argmin}_{\alpha_1,...,\alpha_N} \left(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{C_{cible}}(\mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) + \sum_{\mathrm{n}=2}^{\mathrm{N}} \mathrm{C_{concat}}(\mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}, \mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) \right) \tag{2}$$

Pour résoudre ce problème nous introduisons, pour tout diphone candidat $d_n^{\alpha_n}$, la grandeur $\mathrm{C_{min}}(\mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}})$ désignant le coût minimal des séquences partielles qui aboutissent à $d_n^{\alpha_n}$. On peut écrire, pour tout $n \in [\![1; N]\!]$ et pour tout $\alpha_n \in [\![1; K_n]\!]$:

$$\mathrm{C_{min}}(\mathrm{d}_1^{\alpha_1}) = \mathrm{C_{cible}}(\mathrm{d}_1^{\alpha_1}) \tag{3}$$

$$\mathrm{C_{min}}(\mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) = \mathrm{min}_{\alpha_1,...,\alpha_{\mathrm{n}-1}} \mathrm{C}(\mathrm{d}_1^{\alpha_1}, ..., \mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}, \mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) \tag{4}$$

$$= \mathrm{min}_{\alpha_1,...,\alpha_{\mathrm{n}-1}} \left\{ \mathrm{C}(\mathrm{d}_1^{\alpha_1}, ..., \mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}) + \mathrm{C_{concat}}(\mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}, \mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) + \mathrm{C_{cible}}(\mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) \right\} \tag{5}$$

$$= \mathrm{min}_{\alpha_{\mathrm{n}-1}} \left\{ \mathrm{C_{min}}(\mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}) + \mathrm{C_{concat}}(\mathrm{d}_{\mathrm{n}-1}^{\alpha_{\mathrm{n}-1}}, \mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) \right\} + \mathrm{C_{cible}}(\mathrm{d}_\mathrm{n}^{\alpha_\mathrm{n}}) \tag{6}$$

Les équations 3 et 6 donnent une formulation récursive du coût minimum partiel $\mathrm{C_{min}}$, ce qui permet une résolution très efficace avec un algorithme de type Viterbi [Viterbi 67]. La recherche du chemin optimal D^* est effectuée de manière conjointe : l'argument du minimum de l'équation 6 désigne en effet le prédécesseur de $d_n^{\alpha_n}$ sur le chemin optimal qui aboutit à $d_n^{\alpha_n}$. On obtient ainsi aisément le chemin optimal global, représenté sur la figure 7 par des traits plus épais. La complexité de cet algorithme est en $\mathcal{O}(N \times K^2)$, où $K = \mathrm{max}_{1 \leq \mathrm{n} \leq \mathrm{N}} \mathrm{K_n}$ représente le nombre maximum d'unités candidates pour un diphonème-cible.

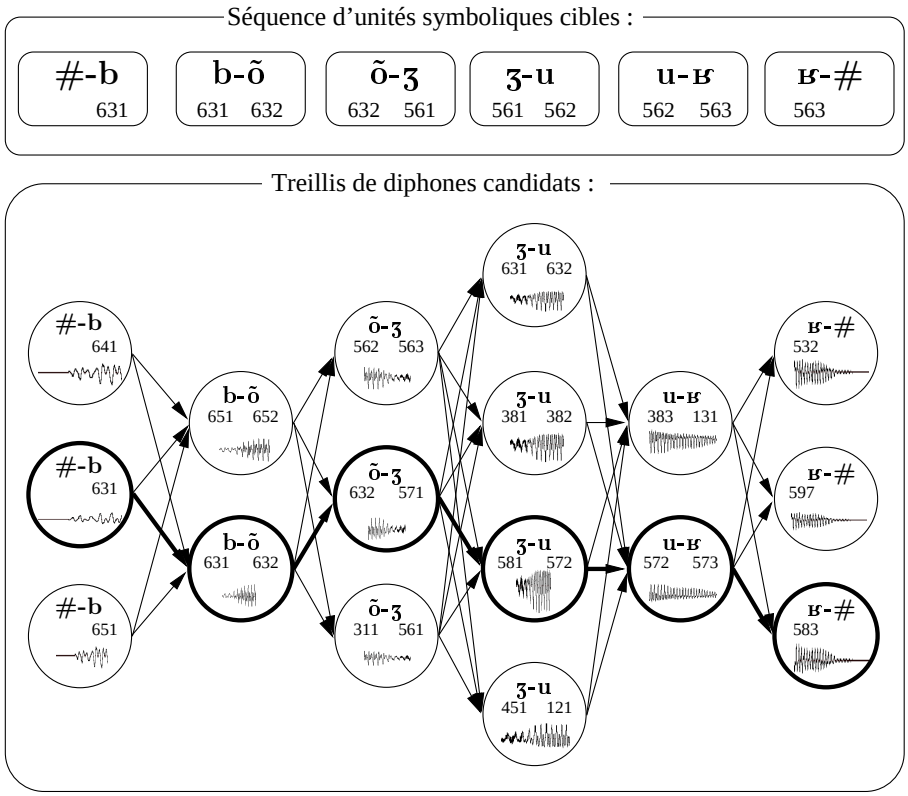


FIGURE 7 – Illustration d'un treillis de diphones candidats pour la séquence-cible correspondant à la phrase « *bonjour* ». Dans la pratique il peut y avoir plusieurs milliers de candidats pour chaque diphonème-cible. Les vecteurs contextuels sont ici représentés par des nombres, dans un but simplement illustratif. La séquence d'unités présentant un coût minimal est signalée par des traits épais.

Une telle approche par programmation dynamique est possible uniquement parce que les critères de sélection s'expriment localement (le coût-cible est d'ordre 1 et le coût de concaténation est d'ordre 2). Dans [Popescu 06] nous avons augmenté le treillis de recherche de contraintes globales portant sur des séquences de trois unités ou plus, comme par exemple la continuité de F0 autour de clusters consonantiques sourds, ou encore la recherche d'un contour général de F0 sans en imposer la hauteur moyenne. Comme ces contraintes globales rendent impossible la recherche exhaustive de l'optimum, nous avons proposé un algorithme de recuit simulé pour construire un chemin aussi optimal que possible. Une solution approchée peut également être dressée avec une succession de passes de Viterbi, comme suggéré dans [Hirai 02]. Cependant l'utilité de telles contraintes reste à établir.

A l'heure actuelle la totalité des solutions commercialisées de synthèse par corpus reposent sur un algorithme de Viterbi. Notons que, dans l'hypothèse où la taille des bases de données ne permettrait plus une optimisation complète par Viterbi en un temps raisonnable, Kumar a proposé un algorithme génétique pour parvenir plus rapidement à une solution acceptable [Kumar 04]. Black et Taylor proposent quant à eux une technique de sélection hiérarchique [Taylor 99], intitulée *Phonological Structure Matching* (PSM). Dans cette technique, les unités cibles et les unités de la base sont organisées en arbres phonologiques, qui rassemblent les informations syntaxiques, lexicales, syllabiques, phonétiques et accentuelles. Ces arbres permettent une sélection rapide des unités en privilégiant des séquences de haut-niveau (par exemple des syllabes ou mots entiers), ce qui garantit une certaine cohérence phonologique. En particulier des phénomènes complexes de réduction ou de coarticulation peuvent être pris en compte

automatiquement, sans traitement particulier. Toutefois l'espace des combinaisons explorées par l'algorithme PSM est nettement moins exhaustif qu'avec le Viterbi, ce qui laisse entendre que la solution trouvée est potentiellement moins optimale. En ce sens, l'approche PSM est à mi-chemin entre les étapes de présélection et de sélection, ce qui lui vaut d'être parfois utilisée exclusivement en guise de présélection [Schweitzer 03].

2.3.3 Les traitements acoustiques

Malgré la prise en compte d'un coût de concaténation lors de la sélection, l'aboutement brutal des unités sélectionnées risquerait fort de faire apparaître des ruptures acoustiques au niveau des frontières. Un lissage du signal est donc fréquemment opéré afin de rendre les transitions plus continues. Bien entendu ce traitement est inutile pour deux unités successives qui sont déjà voisines dans la base de départ.

Une approche courante, directement inspirée du procédé PSOLA dans le domaine temporel (ou Time Domain-PSOLA), consiste à interpoler les formes d'onde des unités gauche et droite sur une période de pitch. Si ce lissage reste simple à mettre en oeuvre et relativement efficace, il n'offre toutefois aucun contrôle sur les trajectoires formantiques et on assiste régulièrement à des évanouissements ou apparitions de formants qui ne reflètent pas les mécanismes phonatoires naturels [Dutoit 94].

Pour pallier cela des techniques plus complexes ont été proposées. Dans [Stylianou 01a], l'auteur propose d'utiliser, pour le lissage des concaténations entre unités successives, le modèle « Harmoniques + Bruit », ou Harmonic plus Noise Model (HNM), qu'il a lui-même introduit quelques années auparavant [Stylianou 96]. Ce modèle décompose le signal de parole en deux composantes :

- une composante quasi périodique, qui est donc composée d'harmoniques.
- une composante bruitée, qui représente les phénomènes non-périodiques : bruits de friction, bruit glottique, etc.

L'interpolation des paramètres harmoniques entre un point de départ dans l'unité gauche et un point d'arrivée dans l'unité droite permet alors de concaténer les deux unités tout en assurant une continuité de hauteur et de spectre. Il en résulte une qualité perceptive supérieure au PSOLA.

On peut également citer les travaux de Wouters et Macon [Wouters 01] qui proposent, pour chaque concaténation, de copier des mouvements spectraux naturels observés sur un segment de la base (fusion unit) apparaissant dans des contextes phonétique et prosodique semblables. Les paramètres utilisés pour ce « décalquage » sont les paramètres LSF [16], qui permettent un lissage assez respectueux des trajectoires formantiques.

Pour affiner la localisation et le suivi des formants, Pfitzinger propose dans [Pfitzinger 04] d'aligner les spectres de départ et d'arrivée par un Dynamic Frequency Warping (DFW), puis d'en déduire une suite de coefficients LPC réalisant une transformation progressive du spectre de départ vers le spectre d'arrivée.

Néanmoins aucune méthode ne permet d'annuler totalement les distorsions acoustiques qui apparaissent aux frontières. Les écarts de hauteur ou de spectre entre unités successives résultent souvent d'une incompatibilité plus globale, qui ne peut donc pas être résolue localement. Le lissage se traduit alors en partie par un étalement de l'artefact, ce qui atténue les ruptures mais dégrade le signal environnant. Le meilleur moyen de limiter les distorsions

16. Line Spectrum Frequencies [Wakita 81]

reste donc d'optimiser le processus de sélection, ce qui suppose avant tout une optimisation du contenu de la base de données.

Pour des raisons similaires l'étape de placage prosodique, qui permet d'imprimer une prosodie artificielle sur le signal de synthèse, est de plus en plus délaissée. Plusieurs techniques de modification prosodique existent toutefois. Parmi les plus connues on retrouve PSOLA et HNM, mais également Vocoder [Laroche 99] et STRAIGHT [Kawahara 99]. Chacune de ces techniques a ses propres avantages et inconvénients, mais aucune ne permet de réaliser des modifications prosodiques parfaitement naturelles. De nos jours, peu de systèmes de synthèse par corpus utilisent un placage prosodique ; on lui préfère un affinage du coût-cible conjointement à une diversification des motifs prosodiques présents dans la base.

2.4 Facteurs de qualité en synthèse par sélection d'unités

La qualité perceptive d'un message vocalisé par un système de synthèse par corpus est essentiellement influencée par trois facteurs : la pertinence des cibles issues des hauts-niveaux, l'adéquation du signal à ces cibles et la qualité des concaténations. Le choix des coût-cible et coût de concaténation, seuls guides du processus de sélection, découle entièrement de l'analyse de ces facteurs perceptifs. Nous détaillons ci-dessous chacun des trois facteurs, ainsi que leur impact éventuel sur le choix des coût-cible et coût de concaténation.

2.4.1 La pertinence des cibles issues des hauts-niveaux

Ce facteur tient à la qualité des traitements linguistiques. L'état de l'art offre un niveau de précision qui peut être considéré comme suffisant pour le style « lecture neutre » défini page 23. [Torzec 01] rapporte pour le système CVOX des taux d'erreurs très faibles dans les transcriptions phonétiques (moins de 1% de mots erronés) et dans le placement des frontières prosodiques (moins de 5%). Les systèmes qui génèrent des cibles prosodiques numériques sont également acceptables si on se tient à ce style neutre.

La diversité des marqueurs symboliques et numériques permet, dans un cadre de lecture neutre, une description assez exhaustive du contexte de chaque unité-cible. Toutefois elle conduit fréquemment à une sur-détermination des cibles, les traits contextuels étant trop nombreux au regard des réalisations acoustiques possibles. Il n'est pas rare que l'ensemble des combinaisons de marqueurs symboliques ou numériques destinés à guider la sélection atteigne plusieurs milliers pour une seule unité-cible ! Nul doute que ces combinaisons manquent de pouvoir discriminant. Théoriquement cela ne devrait pas pénaliser le système. Mais dans la pratique le module de sélection est excessivement contraint par ces multiples consignes qui sont souvent mal hiérarchisées. Il en résulte un accroissement non justifié de la fréquence des concaténations.

Dans le cas particulier où un placage prosodique est effectué, on peut également relever l'insuffisance des paramètres de modélisation de la parole. Les outils de modification prosodique ne prennent donc pas en compte toute la complexité du signal de parole, qui se trouve inéluctablement altéré.

Lorsqu'il s'agit de traiter de la parole plus expressive, le problème est de nature très différente. Les modules linguistiques sont souvent mis en défaut par leur incapacité à prendre en compte des critères plus complexes souvent liés à la sémantique, comme des traits accentuels marqués ou des variations de la qualité de voix.

D'une manière générale, les hauts-niveaux et bas-niveaux font l'objet de travaux assez cloisonnés et le module de sélection ne remet pas en question les consignes issues des hauts-niveaux. On trouve toutefois quelques expériences intéressantes sur une collaboration entre les deux niveaux. Ainsi dans les travaux de Bulyko les consignes phonétiques et prosodiques prennent la forme d'un automate qui rassemble plusieurs formulations possibles de la phrase d'entrée [Bulyko 02]. Cet automate est créé automatiquement par un système de génération de langage naturel, conformément à une approche de type « concept-to-speech » (voire page 21). La recherche du chemin optimal étant effectuée conjointement à la sélection d'unités, il en résulte que les séquences phonétiques et prosodiques sont conditionnées par le contenu de la base de données et par les critères de sélection. Dans [Revelin 05] nous proposons de prendre en compte des alternatives phonétiques plus localisées, qui concernent essentiellement des phénomènes de coarticulation : assimilation consonantique, diérèse ou synérèse, ouverture ou fermeture d'une voyelle, etc. Comme chez Bulyko, la fonction de coût du module de sélection est utilisée pour effectuer un choix entre ces alternatives phonétiques. Bien que les considérations coutumières et sociétales qui guident normalement les choix de prononciation soient occultées, les résultats sont encourageants.

2.4.2 L'adéquation du signal aux cibles spécifiées par les hauts-niveaux

En synthèse par corpus, et plus particulièrement en l'absence de placage prosodique comme c'est le cas dans notre système, le suivi de la séquence-cible issue des hauts-niveaux n'est pas une tâche triviale. Il faut trouver dans la base de données vocale les unités qui respectent au mieux les traits symboliques et numériques de la séquence-cible. C'est le rôle du module de sélection et plus particulièrement du coût-cible. Ce dernier est généralement défini comme une fonction de distance entre les attributs d'une unité candidate et les attributs de l'unité-cible correspondante. Afin d'alléger le treillis de sélection, les composantes les plus importantes du coût-cible sont fréquemment déplacées dans l'étape de présélection [Black 97], au risque de réduire la marge de manoeuvre de l'algorithme de sélection.

Pour les attributs numériques comme les paramètres prosodiques, de simples distances peuvent être utilisées. Des écarts de hauteur, durée, énergie, voire MFCC sont d'usage courant [Donovan 98]. Pour les attributs symboliques, des matrices de distance doivent être mises en place. La « distance triviale » est fréquemment utilisée. Elle correspond à une matrice composée de 0 sur la diagonale et de 1 partout ailleurs. Dans [Breen 98], les auteurs utilisent quant à eux une distance basée sur une organisation arborescente des contextes phonologiques. Une étude générale des différentes composantes du coût-cible est présentée dans [Taylor 06].

Un jeu de pondérations permet d'arbitrer entre ces composantes et ainsi de privilégier celles qui sont perceptivement les plus importantes. Si ces pondérations peuvent être choisies de manière experte, elles peuvent également être apprises automatiquement. On peut utiliser pour cela une approche non-supervisée, par exemple en mesurant dans la base vocale le niveau de différenciation acoustique induit par chaque attribut contextuel [Hunt 96]. Une approche supervisée est également possible, comme dans [Chu 01a] et [Toda 04] où les auteurs cherchent à reproduire au mieux, à partir de la fonction de coût, une notation perceptive établie par un panel d'auditeurs. Mais cette approche est lourde et difficilement généralisable à d'autres bases acoustiques éventuellement très différentes tant quantitativement que qualitativement.

Malgré ces nombreux travaux, la sélection d'unités conformes aux cibles reste perturbée par deux éléments essentiels. Le premier tient à la qualité de couverture de la base locuteur. En effet l'absence éventuelle d'une unité ciblée par les hauts-niveaux impose aux bas-niveaux la sélection d'une unité alternative, pas totalement adaptée au contexte. Le second élément tient à la précision de l'annotation de la base de données. Si celle-ci est imprécise, du fait

d'un manque d'assiduité du comédien dans le suivi des consignes de lectures, d'une annotation automatique incertaine, ou encore d'une mesure incorrecte de paramètres acoustiques, le contrôle des restitutions est nécessairement dégradé. Cette imprécision peut également être la conséquence d'une typologie d'annotation trop restrictive qui ne permettrait pas de discriminer toutes les nuances acoustiques observées dans la base.

2.4.3 La qualité des concaténations

Les discontinuités acoustiques qui apparaissent aux endroits de concaténation sont en grande partie responsables des consonances artificielles de la synthèse par corpus, souvent qualifiée de « métallique ». Plus grave encore, le caractère chevrotant qui en résulte peut dans certains cas nuire à l'intelligibilité du message.

Le coût de concaténation a pour fonction principale de limiter l'apparition de tels artefacts sonores, en contribuant à sélectionner des unités qui « se concatènent bien », c'est-à-dire qui présentent un écart acoustique minimal aux endroits de concaténation. Cette fonction de coût repose donc sur la définition d'une distance acoustique entre unités successives. Par définition cette distance prend des valeurs positives et s'annule théoriquement pour deux unités déjà contiguës dans la base de données. Plusieurs distances et paramètres acoustiques peuvent être utilisés. On rencontre entre autres les distances euclidienne, absolue, de Kullback-Leibler, de Mahanalobis ; on trouve des paramétrisations spectrales également très variées comme LPC, MFCC, LSP [17], PLP [18], MCA [19] ou plus simplement FFT, ainsi que quelques paramétrisations acoustiques non spectrales (F0, énergie). Une présentation complète des travaux sur le coût de concaténation peut être trouvée dans [Vepa 04]. Nous rapportons à présent ceux qui nous semblent les plus significatifs.

Dès 1998 et avant même d'adopter la technologie de synthèse par corpus, Klabbers et Veldhuis étudiaient la corrélation de plusieurs distances acoustiques à la notation perceptive de concaténations [Klabbers 98]. Cette étude, qui portait sur cinq voyelles de l'alphabet phonétique néerlandais, visait principalement à étendre une base de données de synthèse par diphones, en ajoutant pour certains diphones les contextes les plus discriminants sur le plan perceptif. La distance de Kullback-Leibler [Kullback 51] appliquée aux coefficients LPC semblait alors donner les meilleurs résultats. Dans une expérience assez similaire portant sur quatre voyelles américaines, Wouters et Macon avançaient quant à eux la distance euclidienne entre paramètres de type LPC calculés sur une échelle mel [Wouters 98].

L'équipe d'AT&T a appliqué ce type d'expérience au coût de concaténation utilisé en synthèse par corpus. Ainsi dans [Stylianou 01b], elle met de nouveau en avant la distance de KullBack-Leibler, mais sur des paramètres différents (spectre de puissance basé sur une transformée de Fourier discrète). Elle obtient avec cette distance une détection automatique de 37% des discontinuités perçues par l'oreille humaine. Par ailleurs dans [Syrdal 01] elle constate que les discontinuités sont plus audibles sur une voix féminine que sur une voix masculine, et surtout qu'elles sont fortement impactées par le type de phonème et le contexte phonétique. L'équipe est ainsi amenée à conduire plus tard [Syrdal 05] une étude détaillée en fonction des types de phonèmes, et intégrant par ailleurs des paramètres acoustiques non spectraux (hauteur, énergie, cross-corrélation sur le signal temporel). Parmi les conclusions, on note que les milieux de consonnes supportent mieux les concaténations que les milieux de voyelles et que les distances acoustiques doivent être adaptées au phonème.

17. Line Spectrum Pairs, souvent assimilées aux Line Spectrum Frequencies (LSF)
18. Perceptual Linear Predictive [Hermansky 90]
19. Multiple Centroïd Analysis [Crowe 87]

Les distances acoustiques introduites par Donovan et Bellegarda se conforment à cette dernière conclusion. Dans [Donovan 01], le contexte phonétique est utilisé pour la descente d'un arbre de décision, chaque feuille menant à une pondération spécifique des écarts spectraux. Dans [Bellegarda 04], pour chaque phonème de l'alphabet, les périodes de pitch proches du centre d'un tel phonème dans la base sont inventoriées, puis décomposées en valeurs singulières. On obtient ainsi une base de signaux mono-période représentatifs et discriminants, sur laquelle toutes les périodes limitrophes peuvent être décomposées. La distance qui en est déduite semble donner de très bons résultats. Dans [Pantazis 05], les auteurs reviennent à une approche indépendante du phonème. Ils étendent les travaux de [Stylianou 01b] et [Klabbers 01b] avec des modèles de décomposition non linéaires du signal de parole. Le taux de détection des discontinuités perçues par l'oreille humaine atteint 56%.

Le coût de concaténation revêt donc des aspects très variés en fonction des systèmes. Dans cette thèse nous nous intéressons aux principaux invariants phonétiques qui caractérisent les concaténations. On peut remarquer que les distorsions sont statistiquement plus importantes sur les phonèmes qui présentent :

- **une forte variabilité acoustique inter-occurrences**, comme par exemple les liquides, semi-voyelles ou voyelles. Les réalisations de ces phonèmes sont fortement influencées par les choix de prononciation du locuteur et par des effets de coarticulation liés au contexte phonétique [Lindblom 63].
- **une grande stabilité acoustique intra-occurrence** (par ex. les voyelles), et qui supportent donc mal les ruptures spectrales causées par les concaténations. Dans [Bulyko 02], le coût de concaténation est agrémenté d'un « splicing cost » qui tient compte de ce critère.
- **une énergie acoustique élevée**, pour des raisons perceptives évidentes (audibilité des artefacts). D'où la proposition de Donovan de corriger les mesures spectrales par des considérations énergétiques déduites de la perception [Donovan 01].
- **un caractère voisé**. Le voisement augmente, par la périodicité de l'excitation glottique, la cohérence temporelle du signal, qui n'est pas forcément respectée par les concaténations (rupture de périodicité).
- **une grande ouverture du conduit vocal** (et plus précisément de la bouche). Outre l'énergie du signal, l'augmentation du volume de résonance tend à accroître l'acuité des formants, ce qui constitue généralement un facteur aggravant pour les concaténations (rupture des trajectoires formantiques). Encore une fois les voyelles sont les principaux phonèmes visés.

On peut donc établir grossièrement la hiérarchie suivante entre les phonèmes, en commençant par ceux qui supportent le mieux les concaténations :

1. les occlusives sourdes [p], [t] et [k]
2. les autres consonnes sourdes [f], [s] et [ʃ]
3. les occlusives voisées [b], [d] et [g]
4. les consonnes nasales et fricatives voisées [m], [n], [v], [z] et [ʒ]
5. les liquides [l] et [ʁ]
6. les semi-voyelles et le schwa [j], [w], [ɥ] et [ə]
7. les voyelles [a], [ɔ], [o], [ɛ], [e], [ø], [œ], [ã], [õ], [ɛ̃], [œ̃], [i], [y] et [u]

Une approche simple mais relativement robuste consiste à injecter dans le coût de concaténation un jeu croissant de pénalités calqué sur ce classement, ce qui permet de privilégier le placement des concaténations sur les phonèmes qui les supportent le mieux. Les prémices d'une telle approche sont apparues dans [Yi 98], et ont depuis été largement réutilisées.

Une partie de notre travail vise à appréhender la problématique des concaténations uniquement avec des considérations linguistiques ; dans la suite nous nous reposerons donc en grande partie sur cette hiérarchie phonétique.

3 La préparation d'une base de données de synthèse par sélection

La base de données vocale constitue la matière première d'un système de synthèse par sélection. La richesse de son contenu, la qualité de l'annotation, mais aussi l'homogénéité de la voix et des conditions d'enregistrements, sont autant de paramètres qui conditionnent la qualité des restitutions.

La création d'une telle base de données peut être décomposée en trois grandes étapes :

1. Constitution du script de lecture
2. Lecture et enregistrement du script
3. Post-traitement des données

Chacune de ces étapes est détaillée ci-dessous. Nous présenterons ensuite au paragraphe 3.4 le cas particulier de la constitution de bases par collecte de rushes, c'est-à-dire à partir d'enregistrements non dédiés à la synthèse vocale, comme par exemple des bandes sonores de DVD, des enregistrements de discours politiques, etc.

3.1 Constitution du script de lecture

Le script de lecture, ou corpus textuel, désigne l'ensemble des phrases qui vont être lues par le locuteur ou la locutrice durant la phase d'enregistrement. Il peut s'agir d'un texte littéraire (comme par exemple un roman), d'une succession de textes indépendants (comme des articles du *Monde*), ou encore d'un ensemble de phrases isolées sans lien logique. Ce dernier cas est le plus fréquent dans le cadre d'une synthèse de style « lecture neutre », dans laquelle la cohérence prosodique entre phrases successives est négligée.

Le script de lecture a un impact direct sur la qualité finale de la voix de synthèse. De lui dépend la variété des unités disponibles dans la base. Plus le système aura à sa disposition des segments longs et adaptés au contexte, meilleures pourront être les phrases synthétiques.

3.1.1 Quantité vs. qualité

Le contenu du script peut être caractérisé sur un plan quantitatif et sur un plan qualitatif.

L'importance de l'aspect quantitatif n'est plus à établir. L'accroissement des tailles de corpus a été, au cours des années précédentes, un important facteur d'amélioration des voix de synthèse. Les premiers systèmes par sélection exploitaient environ une heure de parole utile [Hunt 96]. A titre indicatif cela correspond grossièrement à un script de 50 000 à 80 000 caractères et nécessite, du fait des pauses entre phrases et des reprises, 3 à 4 heures d'enregistrements. Bien entendu ces valeurs dépendent de nombreux facteurs : rapidité de la voix, aisance du locuteur, niveau d'exigence des superviseurs, etc. Les systèmes actuels utilisent le plus souvent entre 5 et 10 heures de parole utile. Parmi les corpus les plus vastes on peut citer ceux utilisés par ATR dans le système Ximera [Kawai 04], avec des durées de 20, 60 et même 110 heures de parole pour un seul locuteur ! L'équipe d'ATR insiste toutefois sur les problèmes d'homogénéité (qualité de voix, conditions d'acquisition...) posés par l'enregistrement de tels corpus, qui sont nécessairement étalés sur plusieurs mois. Elle constate par ailleurs que, au-delà de 30 heures de parole, on touche aux limites des critères de sélection et que la qualité de synthèse ne progresse plus.

L'aspect qualitatif a en revanche fait l'objet, par le passé, d'une attention moindre. Dans certains systèmes, les phrases du script de lecture sont même sélectionnées aléatoirement parmi un vaste ensemble de textes. Ainsi l'équipe d'AT&T a composé l'essentiel de ses scripts d'enregistrement en assemblant des articles journalistiques du Wall Street Journal et des prompts de services vocaux sans critère de choix précis [Beutnagel 98]. Seule une partie minoritaire du script a fait l'objet d'un travail spécifique, uniquement dans le but d'assurer une couverture complète des diphones de l'anglais. Outre la captation de phénomènes prosodiques d'ordre discursif, l'objectif de cette approche faiblement supervisée est de reproduire une distribution des évènements linguistiques qui soit similaire à celle du domaine lexical envisagé, c'est-à-dire, la plupart du temps, aussi général que possible.

D'autres travaux semblent confirmer la pertinence de cette approche dictée par l'aléa. Ainsi, dans [Kawai 04], les auteurs montrent qu'un ensemble de phrases optimisé ne réduit que très peu le coût de sélection moyen par rapport à un ensemble de phrases aléatoire. Pire, les résultats présentés dans [Lambert 07] indiquent une régression perceptive du système de synthèse lorsque le script a fait l'objet d'une optimisation. Il serait toutefois imprudent de généraliser ces conclusions tant les systèmes de synthèse et les processus d'optimisation du script d'enregistrement peuvent être variés.

L'optimisation du script suppose le double choix d'un **critère** et d'un **algorithme**. Le critère guide la constitution du script tandis que l'algorithme explore l'ensemble (potentiellement gigantesque) des scripts possibles pour satisfaire au mieux ce critère. L'essentiel de notre travail porte sur ces deux choix, qui feront respectivement l'objet des parties II et III. Nous décrivons ci-dessous l'état de l'art sur chacun des deux aspects.

3.1.2 Critère d'optimisation de script

La loi de Zipf

Il est naturel de s'intéresser en premier lieu à la couverture des unités de base dans le script, toute absence pouvant empêcher la reconstruction de certaines séquences phonétiques. Dans notre cas, qui est aussi le plus courant, il s'agit des diphones. Le français en utilise environ 1200 et l'anglais 1700, ces nombres pouvant varier selon les alphabets phonétiques retenus. Il a été proposé en 1988 un corpus de 89 phrases couvrant tous les diphones de la langue française [Charpentier 88]. Ce corpus, établi manuellement, nécessite 4849 diphones pour couvrir les 1290 diphones-cibles, soit un taux de redondance record de 73% seulement. Mais la « densité diphonique » de ces phrases rend leur lecture particulièrement difficile et ambiguë. En voici deux exemples :

```
Nous nous huilons les biceps comme des athlètes pour jouer oisivement
dans l'herbe printanière du parc zoologique.

Taïwan devient une arche de Noé, où on trouve quelques zorilles
yankees, quelques yacks, quelques wapitis wallons, des hibous
ougandais, des papillons yougoslaves, des oies hongroises, des grives
gabonaises et des escargots européens.
```

Dans le cadre d'un enregistrement de locuteur, il est souhaitable de recourir à des phrases plus courtes et plus simples. Ce choix s'accompagne inéluctablement d'une augmentation de la redondance et d'un appauvrissement en diphones, la couverture complète devenant alors difficile à atteindre. Ainsi le corpus de 4 millions de mots (constitué d'archives du Monde)

exploré par les auteurs de [Gauvain 90] n'apporte pas, malgré sa taille, une couverture totale. 65 diphones n'y sont pas représentés.

D'une manière générale la distribution des diphones suit une distribution logarithmique, fréquemment observée en linguistique statistique du fait de la loi de Zipf [Zipf 32]. Initialement cette loi empirique énonce que, si l'on ordonne les mots d'un texte par fréquence décroissante et si on leur attribue un rang, alors le rang trouvé est à peu près inversement proportionnel à la fréquence. En d'autres termes :

$$f(r) \simeq \frac{K}{r} \tag{7}$$

où $f(r)$ représente la fréquence relative du mot de rang r et K est une constante dépendante du corpus. La loi de Zipf est plus ou moins valide pour la plupart des unités phonétiques et linguistiques : syllabes, diphones, triphones, etc. La colonne gauche de la figure 8 montre en échelle bi-logarithmique la fréquence des mots, syllabes, quadriphones et diphones en fonction de leur rang. Les statistiques sont tirées d'un corpus textuel français de 2 500 000 mots, présenté en détail page 86. Si la distribution des mots présente, conformément à la loi de Zipf, une tendance linéaire de pente -1, on constate pour les autres unités une inflexion des distributions pour les rangs petits et une augmentation de la pente pour les rangs élevés. Ceci est dû à une dispersion relativement faible des unités observées, du fait notamment de leur taille inférieure à celle du mot, conduisant à une moindre « température informationnelle » suivant la théorie de Shannon. Cette courbure est parfaitement modélisée par la formule de Mandelbrot [Apostel 57], qui généralise la loi de Zipf en s'appuyant sur la théorie de l'information :

$$f(r) \simeq \frac{K}{(\varphi + r)^{\beta}} \tag{8}$$

où K, φ et β sont des constantes dépendantes du corpus et que l'on peut estimer graphiquement. En échelle bi-logarithmique, φ explique l'inflexion sur les unités fréquentes, tandis que β contrôle la pente sur les unités rares. β est généralement supérieure à 1 (sauf pour certains textes particuliers) et une faible dispersion des unités implique une valeur de β élevée. Lorsque l'unité observée est le mot, la valeur de β, proche de 1, est un indicateur de la richesse du vocabulaire présent dans le texte.

Mais ce qui nous intéresse avant tout, c'est la fonction de répartition des unités, c'est-à-dire la proportion $F(r)$ du texte qui est couverte de manière cumulée par les unités de rang 1 à r :

$$F(r) = \sum_{i=1}^{r} f(i) \tag{9}$$

Les fonctions de répartition des mots, syllabes, quadriphones et diphones sont présentées sur la colonne droite de la figure 8. On observe dans tous les cas une courbe d'allure logarithmique, fortement incurvée, qui découle des lois de Zipf-Mandelbrot. Ainsi, sur 1162 diphones rencontrés, 274 couvrent à eux seuls 80 % des occurrences de diphones relevées dans le corpus. De même les 10% de quadriphones les plus fréquents offrent une couverture de 80%. En contrepartie le nombre d'unités rares est très élevé et représente un poids statistique non négligeable (*i.e.* la probabilité de rencontrer une unité rare au sein d'une phrase est élevée). On utilise souvent l'acronyme LNRE (*Large Number of Rare Events*) pour décrire cette distribution des unités rares [Khmaladze 88], qui rend presque impossible la couverture dans le script de la totalité des formes existantes pour une unité donnée.

Dans un système à base de diphones, l'absence dans le corpus de certains diphones rares peut occasionner des « trous » dans les phrases synthétiques qui les requièrent. Des techniques palliatives peuvent alors être mises en place : remplacement par un diphone phonétiquement

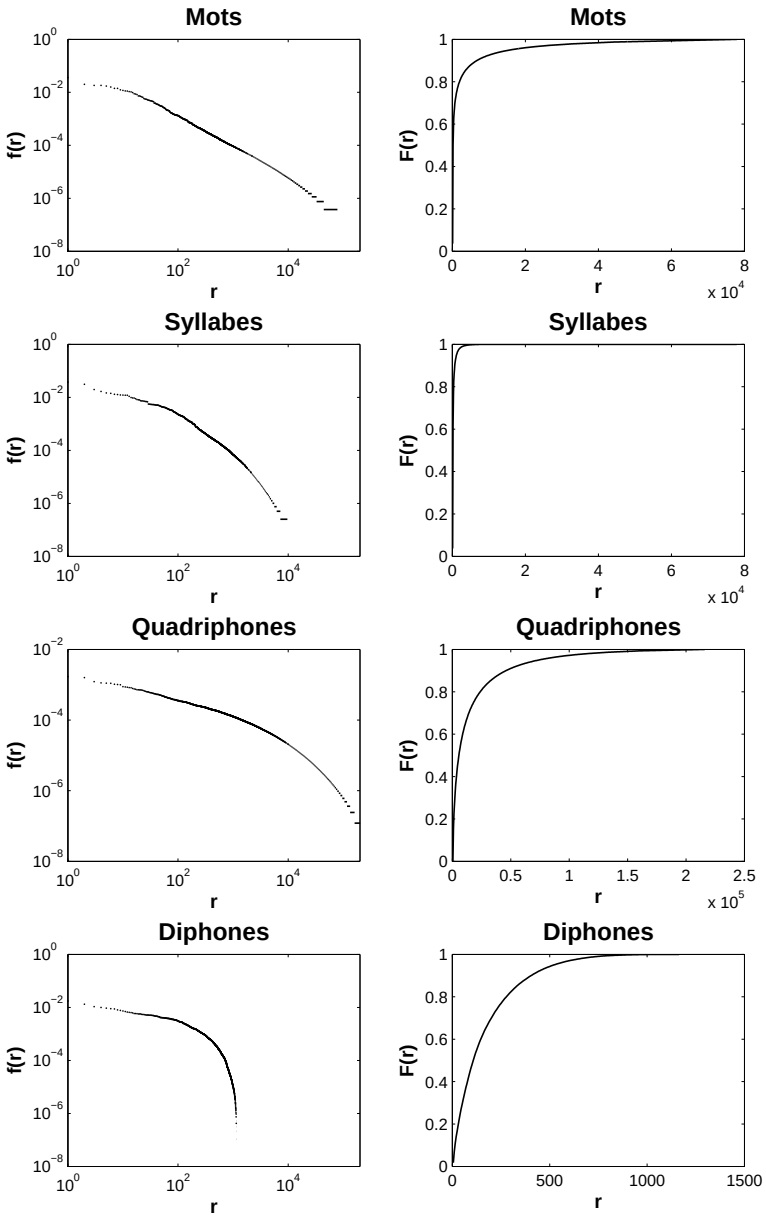


FIGURE 8 – Distribution de Zipf-Mandelbrot de différentes unités linguistiques. La colonne de gauche donne la fréquence $f(r)$ de chaque unité en fonction de son rang r (obtenu après classement par ordre de fréquence). La colonne de droite présente les fonctions de répartition correspondantes, c'est-à-dire la proportion $F(r)$ du texte de référence qui est couverte de manière cumulée par les unités de rang 1 à r.

proche (par exemple /ɛ̃-ã/ remplacé par /œ̃-ã/), reconstruction par concaténation d'unités plus petites, reconstruction paramétrique, utilisation d'unités provenant d'une autre base ou d'un autre locuteur, etc. [Möbius 03] propose à ce sujet une analyse intéressante des problèmes posés par les évènements rares en synthèse de parole et recense les techniques utilisées pour y remédier.

Le choix des unités à couvrir

Si les diphones occupent incontestablement une place centrale dans les modules de traitement acoustique, leur rôle dans la construction du script de lecture est plus nuancé. D'une

part leur petite taille ne permet pas d'anticiper les contraintes de continuité segmentale qui seront rencontrées lors de la synthèse. D'autre part leur faible nombre les rend inadaptés à l'optimisation de scripts de plusieurs milliers de phrases, qui offrent une latitude suffisante pour considérer des unités plus longues. Une partie de l'état de l'art porte donc sur la couverture d'autres unités dans le script de lecture. Les triphones sont d'usage courant, souvent en complément des diphones pour assurer une couverture minimale [François 01] [Bozkurt 03] [Lambert 04] [Ni 06]. Des unités plus longues comme les quadriphones sont parfois utilisées, ou encore des unités linguistiques comme les mots [Ni 06]. On relève également l'emploi fréquent d'unités phonétiques reposant sur des considérations acoustiques : demi-syllabes [Black 01] et syllabes [Isogai 05], surtout dans les langues tonales comme le Mandarin [Chu 01b] [Kuo 02]. A l'inverse certains travaux reposent en grande partie sur de simples considérations phonémiques [Kawai 00] [Black 01]. Une observation statistique de toutes ces unités est rapportée dans [Gauvain 90] pour le français. Parmi elles, seuls les diphones ont été conçus spécifiquement pour la synthèse par concaténation, toutes les autres étant d'usage courant en linguistique et dans les technologies vocales. Mais aucune n'est réellement dédiée à l'étape de constitution du script pour la synthèse par sélection. De plus le lien entre l'unité choisie pour le critère de couverture et la qualité finale de la voix de synthèse n'a jamais été analysé.

Outre les aspects phonétiques, on injecte fréquemment dans le critère d'optimisation du script des considérations de plus haut niveau. En effet pour la synthèse d'une phrase avec une qualité acceptable on ne peut pas, la plupart du temps, se contenter de la simple présence dans la base de données des unités phonétiques requises. Il est souhaitable que chacune des unités-cibles soit disponible dans un contexte similaire à celui qu'elle aura dans la phrase synthétique. C'est d'ailleurs en cela que la synthèse par corpus se distingue de la synthèse par diphones mono-représentés.

Pour cette raison on privilégie généralement, lors de la constitution du script, la couverture d'unités en contexte. Dans ce cas, une unité est caractérisée non seulement par son contenu phonétique, comme par exemple l'identité des phonèmes gauche et droit dans le cas des diphones, mais également par des traits contextuels qui précisent son environnement linguistique, prosodique, ou encore phonétique. Dans [Kawai 00], les auteurs proposent même d'enrichir les unités à couvrir avec des paramètres prosodiques numériques. Naturellement de tels paramètres, prédits à partir du texte, ne correspondent pas forcément à ce qui sera réalisé par le locuteur lors de la lecture du script. Les auteurs mesurent toutefois un taux de recoupement de 30% entre la prosodie prédite et celle effectivement produite, ce qui leur permet d'accroître légèrement leur couverture prosodique.

D'une manière générale les traits contextuels utilisés pour le choix du script (et pour l'annotation de la base) doivent être de la même nature que ceux recherchés par le système de sélection. Leur variété a un impact direct et combinatoire sur l'abondance des unités à couvrir ; aussi plusieurs travaux ont visé à réduire le nombre de contextes possibles pour chaque unité. Cette réduction peut être de nature experte, comme dans [Lambert 04] où seule l'information d'accentuation est retenue pour l'anglais : accent primaire, accent secondaire, ou unité non accentuée. Mais elle peut également être automatique. Ainsi dans [Black 01], les contextes sont regroupés au sein d'un arbre de classification calculé automatiquement à partir d'une base de parole pré-existante. Cette approche permet donc d'établir une hiérarchie entre les contextes, que les auteurs exploitent lors de la constitution du script pour couvrir en priorité les contextes les plus significatifs.

Précisons enfin que le critère phonétique et le niveau de précision contextuelle requis dépendent de plusieurs facteurs, comme le type d'expressivité envisagé et la taille de script visée, et qu'il peut évoluer tout au long de l'étape de création du script de lecture. A ce sujet, Kuo et Huang proposent pour le Mandarin un ensemble de traits contextuels qui est évolutif, tant dans

sa diversité (nombre de traits) que dans sa précision (nombre de symboles par trait) [Kuo 02].

Couverture en largeur ou couverture en fréquence ?

La quasi-totalité des travaux portant sur l'optimisation du script de lecture se basent sur des critères de couverture d'unités. Dans ce cadre, on distingue deux familles d'approches.

La première famille rassemble ce que nous appellerons les approches « en largeur ». Partant du constat qu'une couverture complète est hors de portée du fait de la loi de Zipf-Mandelbrot, ces approches visent à maximiser dans le script de lecture le nombre de formes différentes pour le type d'unité recherché. Il peut s'agir par exemple de maximiser le nombre de diphones différents, puis en second lieu le nombre de triphones. Cela peut être vu comme un problème de recouvrement d'ensemble [François 01] : un ensemble plus ou moins vaste d'unités à couvrir est défini, puis le script est construit de manière à assurer une couverture complète, ou pour le moins maximale [Lambert 04], de cet ensemble. La notion de « couverture en largeur » peut également être abordée de manière plus diffuse. Ainsi dans [Bozkurt 03], les contextes à couvrir pour chaque unité ne sont pas choisis à l'avance ; une fonction de coût, proche du coût-cible utilisé dans l'algorithme de sélection, permet d'éviter la couverture d'unités trop semblables en privilégiant des contextes éloignés. [Tian 05] propose une démarche plus générale n'utilisant aucune notion phonétique ni degré de couverture ; les auteurs utilisent en effet la distance de Levenshtein [20] pour mesurer le degré de nouveauté apporté par une phrase et ainsi maximiser la variété textuelle au sein d'un texte. Dans [Gauvain 90], on propose de mesurer l'entropie des phrases suivant des modèles markoviens. Les phrases d'entropie faible sont alors composées d'évènements phonétiques et linguistiques courants, tandis que les phrases d'entropie élevée contiennent des évènements plus rares. La sélection de ces dernières dans le script de lecture devrait permettre d'accroître la couverture en largeur ; mais les auteurs constatent qu'elles apportent principalement des passages en langue étrangère, ce qui n'est pas forcément souhaitable. C'est d'ailleurs l'une des difficultés majeures rencontrées par les approches en largeur : en se focalisant sur la couverture d'évènements rares elles apportent inéluctablement des textes complexes voire incorrects.

La deuxième famille concerne les approches « en fréquence ». Celles-ci visent à inclure dans le script de lecture en priorité les évènements fréquents, sous prétexte qu'ils seront plus utiles que les autres, quitte à réduire la diversité globale. Plus précisément, chaque unité-cible se voit attribuer, au moment de l'optimisation du script, un niveau d'importance qui est fonction de sa fréquence d'apparition dans un corpus de référence [Black 01] [Ni 06]. Bien entendu le choix de ce corpus de référence est primordial ; il doit refléter au mieux l'usage qui sera fait de la voix de synthèse finale. La transcription phonétique et l'annotation prosodique de ce corpus est effectuée simplement à l'aide des haut-niveaux du système de synthèse. Pour un script final de plusieurs milliers de mots, l'expérience montre qu'un corpus de référence de plusieurs millions de mots est nécessaire à l'établissement de statistiques pertinentes.

La difficulté porte alors sur la disponibilité de telles ressources textuelles avec une diversité lexicale et un niveau de correction (orthographe, formatage) suffisants. Les archives journalistiques sont pour cela d'une grande aide. Pour le français les archives du journal *Le Monde* sont facilement accessibles et largement utilisées pour de nombreux travaux en linguistique. Par exemple le corpus BREF créé par le LIMSI se base en grande partie sur cette source [Gauvain 90]. Les sources journalistiques ont également joué un rôle majeur dans la constitution de corpus de référence pour l'anglais [Beutnagel 98] [Ni 07]. Dans [Chu 01b] ce sont cinq

20. La distance de Levenshtein permet de mesurer la similarité entre deux chaînes de caractères en dénombrant les insertions, suppressions et substitutions qui sont nécessaires pour passer d'une chaîne à l'autre.

années d'un quotidien mandarin qui servent de base aux calculs statistiques. Toutefois, le style journalistique n'est généralement pas représentatif de l'usage qui sera fait de la synthèse vocale. La vocalisation de courriers électroniques ou de messages instantanés fera par exemple intervenir des champs lexicaux et structures syntaxiques très différents. D'autres sources abondantes, libres de droits, et probablement plus variées, peuvent être exploitées. Ainsi le projet Gutenberg [Lebert 08], qui vise à rassembler sur la toile tous les livres libres de droits, a servi à la constitution de plusieurs corpus dont ARCTIC [Kominek 03] qui est utilisé dans de nombreux travaux en synthèse de la parole. Il est par ailleurs fréquent de mêler plusieurs sources afin d'élargir le champ lexical. Pour le corpus ATRECSS [Ni 07], la source BTEC [21] constituée de 4 millions de mots a été utilisée en complément d'une source journalistique de 22 millions de mots pour améliorer la prise en compte du langage conversationnel anglais. Le corpus français IRISA [François 01] fait quant à lui intervenir 7 sources différentes, pour un total de 5 millions de mots environ : interviews d'auteurs de bandes dessinées, sous-titres de séries TV, un livre de Diderot, communications médicales, journal *Le Monde*, messages de forums et transcriptions de sessions parlementaires.

Toutefois la possibilité même de constituer un corpus de référence suffisamment diversifié pour être représentatif de la langue est contestée par Van Santen dans [Van Santen 97a]. En comparant plusieurs sources textuelles de grande taille, il a en effet constaté des différences significatives dans les distributions d'unités et en a déduit que la quête d'un corpus « universel » était vaine. Cette position est partagée par Bozkurt et al dans [Bozkurt 03] : en observant des statistiques de triphones entre corpus très différents, les auteurs constatent des recoupements assez faibles. Suivant ces conclusions il convient de prendre quelques précautions lors de la constitution du corpus de référence. Tout d'abord il faut bien délimiter les domaines d'utilisation de la synthèse vocale et choisir des sources textuelles appropriées. Ensuite, et c'est là le point le plus important, il ne faut pas surexploiter les statistiques extraites d'un tel corpus : si les mesures de fréquence des évènements courants peuvent être considérées comme fiables, celles des évènements rares restent fortement dépendantes des sources injectées dans le corpus et doivent être considérées comme telles. Notons que cette précaution s'accorde aisément avec une approche « en fréquence ».

Les approches « en largeur » et « en fréquence » ne sont pas forcément antagonistes et peuvent tout à fait cohabiter au sein d'un algorithme de création de script. Ainsi dans [Isogai 05], des critères fréquentiels (approche en fréquence) sont utilisés pour départager des phrases qui apportent un même nombre de nouveaux triphones (approche en largeur). Dans [Kuo 02], les deux approches sont pondérées et mêlées au sein d'un unique critère d'optimisation. Enfin, la couverture d'une unité dans le corpus peut être considérée de manière plus riche qu'avec une simple distinction binaire « couvert » / « non-couvert ». Dans plusieurs travaux, l'objectif est de disposer dans le script d'un nombre minimal d'occurences pour chaque unité-cible [François 01]. La redondance introduite de cette manière dans la base de données est censée permettre, pendant l'étape de sélection, de réduire le coût de concaténation en choisissant des candidats qui s'enchaînent mieux. Cet ajout *a priori* de redondance permet également de se prémunir d'éventuels écarts phonétiques ou prosodiques entre les réalisations du locuteur et la chaîne symbolique attendue, évitant ainsi l'apparition *a posteriori* de trous dans la couverture. [Chu 01b] assouplit cette contrainte en imposant un nombre minimal de 10 contextes différents pour chaque syllabe mandarine : on est alors à mi-chemin entre la recherche de redondance et l'élargissement de la couverture. L'objectif de redondance peut également être fonction de la fréquence de l'unité. Krul introduit à ce sujet une technique de construction de script qui permet de tendre vers une distribution pré-établie des unités [Krul 08]. Cette distribution-cible peut par exemple être uniforme, ou encore être extraite d'un corpus spécifique.

21. Basic Travel Expression Corpus

D'une manière générale, la présence dans le script d'unités répétées est absolument inéluctable du fait de la loi de Zipf-Mandelbrot, au moins pour les unités fréquentes. Mais l'injection volontaire d'une redondance additionnelle est une stratégie contestable. Son intérêt pour la qualité de synthèse finale n'a jamais été établi, malgré un impact sévère sur la taille du script et une forte pénalisation du niveau de couverture général.

3.1.3 Algorithme d'optimisation de script

Dans les paragraphes précédents nous avons discuté des critères d'optimisation possibles pour la constitution du script de lecture. Comme nous l'avons vu, les critères les plus courants s'expriment sous la forme d'un taux de couverture d'un ou de plusieurs type(s) d'unité(s), ce taux étant mesuré en largeur (nombre d'unités distinctes) ou en fréquence (pourcentage d'occurrences couvertes dans un corpus de référence). Nous discutons ci-dessous des algorithmes qui permettent d'explorer l'univers des scripts possibles et de construire celui qui satisfait au mieux le critère choisi.

Citons pour commencer les travaux intéressants de [Boeffard 97], dans lesquels les phrases d'un corpus textuel sont générées de manière optimale à l'aide d'un algorithme génétique. Cette stratégie est intéressante car elle permet de créer de toutes pièces un ensemble de phrases extrêmement dense. Dans cette étude, le corpus de 100 phrases ainsi créé a pour vocation d'être lu et enregistré, mais dans un but très différent du nôtre. Il sert à l'apprentissage par réseaux de neurones des modèles prosodiques[22] d'un système de synthèse par diphones monoreprésentés, dans un contexte applicatif très contraint. Les phrases considérées sont toutes de la forme « *Jean Dupont, poste 12 00* », où les champs variables sont le prénom, le nom et le numéro de poste. L'approche par génération de phrases est donc grandement facilitée par la combinatoire et la syntaxe très contrôlées, et serait donc difficilement réutilisable pour notre problème plus générique de la création du script de lecture en synthèse par corpus.

Une approche universelle : la condensation de corpus

Dans notre cadre, l'état de l'art repose intégralement sur des approches de type « condensation de corpus », dans lesquelles les phrases du script de lecture sont sélectionnées parmi un ensemble très vaste de phrases candidates. Appelons « corpus de pioche » cet ensemble initial. En théorie le corpus de pioche peut être différent du corpus de référence mentionné plus haut, qui sert à observer les distributions statistiques des différentes unités et à préciser le critère d'optimisation. Mais dans la pratique il y a peu de raisons de les distinguer et tous les travaux s'accordent à utiliser le même corpus pour les deux usages.

Bien entendu plus le corpus de pioche est grand et varié, plus la marge de manoeuvre pour l'optimisation du script est importante. Mais la complexité de la recherche d'un sous-ensemble optimal augmente également, et de manière combinatoire, avec la taille de départ. Il s'agit d'un problème algorithmique NP-difficile [Garey 79] ; la quête de l'optimum global est hors de portée et le recours à des stratégies sous-optimales doit être envisagé.

Les stratégies itératives

L'algorithme glouton (*greedy* en anglais) est un algorithme très utilisé pour les problèmes de couverture d'ensemble. Il a été introduit en synthèse vocale par Van Santen et Buchsbaum

22. c'est-à-dire de la fonction de prédiction des paramètres de placage prosodique à partir des données symboliques déduites du texte, voir page 27.

en 1997 [Van Santen 97b]. Cet algorithme extrait une à une les phrases les plus intéressantes du corpus de pioche et les ajoute au script final. Le choix de la Nième phrase à ajouter au script dépend bien entendu des $N-1$ phrases qui ont déjà été choisies. Dans une approche en largeur, la Nième phrase pourra par exemple être celle qui, parmi toutes les phrases restantes dans le corpus de pioche, apporte le plus de triphones nouveaux (c'est-à-dire non contenus dans les $N-1$ premières phrases).

La plupart des travaux mentionnés dans les pages précédentes ont recours à un algorithme glouton. Bien que sous-optimal il offre de nombreux avantages :

- Tout d'abord il n'implique qu'une charge de calcul très réduite, ce qui facilite l'expérimentation et permet l'utilisation de gros corpus de pioche.
- Ensuite il garantit, par construction même, une forme d'optimalité sur le début du script. En effet le choix des phrases 1 à N ne dépend pas des phrases suivantes et peut donc être exploité isolément. Cette propriété est intéressante pour deux raisons. D'une part chaque phrase peut être soumise à une validation manuelle, par exemple pour exclure celles qui sont incorrectes (faute d'orthographe, langue étrangère, vulgarité), sans remettre en cause le choix des phrases précédentes. D'autre part elle permet une exploitation partielle du script. En effet la quantité d'enregistrements visée pour la création d'une voix de synthèse dépend de nombreux facteurs : budget disponible, caractéristiques de la voix, rapidité du locuteur, attentes qualitatives, etc. Un tel script, permettant un enregistrement partiel avec une perte de qualité minimale, est donc attrayant.
- Enfin, ses performances sont très proches de l'optimum, si l'on en croit les conclusions de [Chevelu 07]. Pour un critère d'optimisation particulier, ce travail propose de constituer le script de lecture avec une technique de relaxation lagrangienne [Caprara 00]. Cette technique fournit avantageusement une borne inférieure au problème de condensation de corpus, ce qui permet à Chevelu et al de conclure que, dans ce contexte et pour un niveau de couverture donné, la sous-optimalité du glouton n'est responsable que d'une augmentation de 10% de la taille du script final par rapport à l'optimum. L'utilisation du glouton semble donc peu pénalisante au regard de son efficacité calculatoire.

Si la relaxation lagrangienne se montre, dans le contexte expérimental de [Chevelu 07], légèrement plus performante que le glouton, elle présente d'autres inconvénients. Tout d'abord elle se révèle 10 fois plus coûteuse sur le plan calculatoire, pour un apport somme toute assez marginal. Ensuite elle n'offre aucune « progressivité » du script, contrairement au glouton (voir le deuxième point ci-dessus). Enfin elle n'a été appliquée qu'à des problèmes figés de couverture d'ensemble (ensemble d'unités et nombre d'occurrences imposés) ; on peut douter de sa pertinence pour des problèmes moins contraints comme des approches en fréquence.

[Francois 02] compare plusieurs stratégies itératives de condensation : l'algorithme glouton, l'algorithme cracheur et l'algorithme d'échange par paire. Le cracheur est simplement l'inverse du glouton, c'est-à-dire qu'il exclut une à une les phrases les moins intéressantes du corpus de pioche. L'algorithme d'échange par paire initialise quant à lui le script de manière aléatoire à la taille finale souhaitée, et sélectionne itérativement les paires de phrases (l'une faisant partie du script, l'autre appartenant au complémentaire dans le corpus de pioche) dont l'échange est le plus bénéfique. La conclusion de [Francois 02] est que le glouton fournit les meilleurs résultats, mais que l'application *a posteriori* d'une courte passe de « crachage » les améliore légèrement.

Les traitements manuels

Des traitements manuels peuvent accompagner chacune des phases de la création du script de lecture.

Tout d'abord la relecture du corpus de référence permet d'éliminer des phrases aberrantes (ex. : phrases en langue étrangère), de corriger certains passages (ex. : mots mal orthographiés, abréviations), voire de réécrire entièrement des textes mal formés (ex. : SMS). Ces corrections textuelles ont pour objectif principal de faciliter l'interprétation du corpus de référence par les haut-niveaux du système de synthèse, ce qui permet d'obtenir une transcription phonétique et une annotation prosodique du corpus qui sont conformes à une lecture intuitive. La qualité du script final en dépend : pertinence de la couverture linguistique, lecture aisée et conforme à la séquence phonétique attendue. Si une relecture complète du corpus de référence est hors de portée car trop coûteuse, on peut bien entendu se cantonner à ses parties critiques. Par exemple des traces de messagerie instantanée bénéficieront bien plus d'une intervention manuelle que des sources journalistiques ou littéraires.

Tout au long de la construction du script de lecture, la supervision des phrases retenues peut s'avérer très bénéfique, par exemple pour rejeter des phrases trop complexes ou mal phonétisées. Cet élagage peut être effectué au fur et à mesure de la construction du script ou bien *a posteriori*. Dans [Kominek 03], la sélection par glouton de 1297 phrases est suivie de plusieurs passes de relecture qui aboutissent à la suppression de 165 phrases complexes, incorrectes ou encore injurieuses. Il est également possible de corriger la transcription phonétique de certaines phrases du script, pour éviter les écarts entre la séquence attendue et celle effectivement réalisée par le(s) futur(s) locuteur(s). Ce type de modification ou suppression *a posteriori* peut toutefois causer des trous dans la couverture linguistique ; un traitement « en temps réel » (c'est-à-dire au fil de la construction du script) lui est donc préférable, car ainsi la disparition d'unités occasionnée par une suppression de phrase ou une correction phonétique peut être automatiquement compensée dans les phrases suivantes.

La surveillance de certains indicateurs, comme les taux de couvertures de plusieurs types d'unités ou la longueur des phrases, offre également la possibilité d'affiner voire corriger le critère d'optimisation en temps réel.

Enfin le script final peut faire l'objet d'annotations ou de réécritures pour en faciliter la lecture et favoriser la conformité entre la séquence phonétique attendue et celle effectivement produite : simplification de l'écriture des mots complexes, francisation de mots étrangers pour lever les ambiguïtés de prononciation, indications de liaison, etc.

3.1.4 Spécialisation sur un domaine applicatif

La plupart des systèmes et voix de synthèse se veulent généralistes, c'est-à-dire capables de vocaliser avec une qualité acceptable et si possible uniforme n'importe quel contenu textuel. On rencontre cependant fréquemment des contextes applicatifs dans lesquels le champ lexical, les structures syntaxiques, etc., évoluent dans un **domaine restreint**.

Parfois même, l'ensemble des textes possibles en entrée du synthétiseur est fini et connu à l'avance ; on parle dans ce cas de **domaine limité**. Par exemple les prompts vocalisés sur une horloge parlante respectent une nomenclature prédéfinie, avec des champs variables faisant intervenir un vocabulaire bien identifié : « *Il est exactement 11h21 et nous sommes vendredi 15 octobre 2010* ». Si, pour certaines applications, l'enregistrement complet de toutes les phrases possibles du domaine limité est envisageable, le recours à un système de synthèse est souvent préférable car il permet de rationaliser la masse d'enregistrements requis.

Le domaine d'utilisation de la voix de synthèse peut également être restreint sans être limité. Ainsi la vocalisation de recettes de cuisine fait intervenir de nombreux mots et expressions récurrents (« *Mélangez les oeufs et la farine* », « *Faîtes fondre le beurre* », « *Préchauffez le four thermostat 7* », etc.), tout en présentant une variété qui peut être considérée comme illimitée.

Plusieurs travaux ont traité de l'adaptation de systèmes de synthèse à des domaines restreints [Krul 08]. Yi et Glass y voient même un point de passage obligé lorsque la qualité de synthèse est placée au coeur des préoccupations [Yi 98]. Plus précisément ils suggèrent d'améliorer les systèmes de synthèse en élargissant progressivement les domaines d'applications plutôt qu'en augmentant la quantité de données sur des domaines généralistes. Selon cette approche on ne construit que des systèmes qui produisent de la parole hautement naturelle, quel que soit l'investissement financier, mais en contrepartie ces systèmes ne sont opérationnels que sur des domaines textuels restreints. Black et Lenzo ont ainsi expérimenté la création de scripts de lecture spécialisés, pour plusieurs applications en domaines restreints (limités ou non) [Black 00]. Ils ont rassemblé pour cela de nombreuses phrases issues des domaine cibles, sans critère de sélection particulier. Les scripts constitués ont été enregistrés puis exploités conformément à l'approche de synthèse par corpus, et de très bonnes qualités de synthèse ont pu être obtenues sur chacun des domaines cibles. Toutefois en présence de mots hors-domaine, les auteurs proposent une bascule (*back-off*) sur une ancienne technique de synthèse par diphones. La qualité est donc fortement dégradée dès que l'on sort du domaine spécifique.

Une approche plus courante consiste à partir d'un système généraliste, capable de vocaliser à peu près n'importe quelle phrase avec une qualité acceptable, puis d'en optimiser les performances sur un domaine restreint. On trouve dans [Fischer 04] une évaluation comparative de plusieurs techniques d'adaptation de voix généralistes à un domaine restreint. Il ressort en particulier de ce travail que l'ajout dans la base de données acoustique de phrases spécifiques du domaine a un impact très positif sur la qualité de synthèse dans ce domaine. C'est incontestablement la technique d'adaptation la plus performante et la plus utilisée. Par exemple si la future voix de synthèse est destinée à vocaliser des informations de Football, il est recommandé d'ajouter au script de lecture un ensemble dédié de phrases comportant des mots, expressions et noms propres typiques de ce sport. Statistiquement, la base de données finale pourra ainsi fournir au système de sélection des segments acoustiques plus longs et mieux adaptés au contexte. [Chu 02] propose une méthode de construction d'un tel script complémentaire, dans laquelle on commence par identifier les chaînes textuelles spécifiques et récurrentes du domaine. Parmi ces chaînes on extrait celles qui amélioreraient le plus le système de synthèse si elles étaient ajoutées au script de lecture. Enfin, un algorithme glouton de condensation de corpus permet de sélectionner les phrases du domaine qui couvrent au mieux ces chaînes différenciantes. Dans d'autres travaux, l'élargissement de la base est complété d'adaptations portant sur le système de sélection ou de présélection [Taylor 99] [Schweitzer 03].

Le présent document traite de la constitution de scripts généraux inspirés de nombreux domaines classiques d'exploitation des voix de synthèse (voir en section 7.1 page 86). Néanmoins les outils développés restent aisément applicables à des domaines applicatifs plus restreints. Nous suggérons de créer dans un premier temps un script de lecture généraliste, réutilisable pour la création de nombreuses voix, puis de constituer d'éventuels scripts complémentaires spécifiques au domaine applicatif dans lequel chaque voix sera utilisée. Une telle prise en compte du domaine applicatif n'est pas toujours possible car les voix de synthèse font l'objet d'applications variées et souvent imprévisibles au moment de leur création.

3.2 Lecture et enregistrement du script

3.2.1 Casting

L'étape de casting a pour objet la sélection d'une voix parmi un panel de locuteurs. La plupart du temps, la création d'une voix de synthèse s'inscrit dans un projet de service précis. Plusieurs intervenants participent alors à la sélection de la voix. La maîtrise d'oeuvre impose

certains critères de sélection : aptitude du locuteur à suivre des consignes de lecture parfois complexes, aptitude à contrôler son organe vocal, fluidité de la lecture, motivation, etc. D'autres critères peuvent être imposés par la maîtrise d'ouvrage ou définis conjointement : sexe, timbre de voix, style de lecture, accent régional, niveau d'articulation, etc.

Le casting est l'occasion d'un échange enrichissant entre la communauté scientifique de synthèse vocale et la communauté artistique. Les locuteurs qui sont démarchés ou qui se portent candidats sont généralement des professionnels de la voix : comédiens, voix-off, chanteurs...

La sélection peut être effectuée sur la base de simples enregistrements existants des différentes voix, ou bien porter sur des enregistrements dédiés effectués en conditions réelles, ce qui est préférable. Pour cela on peut utiliser un extrait du script de lecture. Une approche judicieuse mais un peu lourde consiste à retenir un sous-ensemble de phrases qui apporte a priori (sur des critères essentiellement phonétiques) les unités nécessaires à la synthèse d'une dizaine de phrases de test. De cette manière on peut simuler le comportement en synthèse vocale de chaque voix. Les phrases ainsi synthétisées font l'objet d'une évaluation perceptive et un classement entre les voix candidates peut être établi. Mais la plupart du temps on se contente d'effectuer un simple choix expert sur la base des enregistrements de casting.

Notons enfin que dans certains cas l'étape de casting n'est pas nécessaire. La voix cible peut être identifiée à l'avance, comme dans le cas de la synthèse personnalisée d'une célébrité ou d'un amateur.

3.2.2 Enregistrement

La lecture et l'enregistrement du script par le locuteur constituent une étape décisive et probablement la plus délicate du processus de création de voix.

La chaîne d'acquisition

Tout d'abord, la chaîne d'acquisition sonore doit être constante tout au long des enregistrements. Cette contrainte est inhérente aux mécanismes de la synthèse par concaténation : les unités concaténées doivent présenter des caractéristiques acoustiques les plus proches possibles. En particulier la pièce, le mobilier, le matériel d'acquisition ainsi que la position du locuteur par rapport au microphone doivent rester inchangés.

Ensuite la réverbération doit être extrêmement faible car elle nuit aux aspects segmentaux de la parole : les unités qui « bavent » les unes sur les autres ne peuvent pas être segmentées ni concaténées proprement.

Enfin le bruit ambiant doit être aussi « blanc » que possible, c'est-à-dire qu'il ne doit pas présenter de cohérence temporelle (échantillons successifs non corrélés). Des évènements sonores non ponctuels comme une discussion d'arrière-plan, des grincements de portes, passages de véhicules, sonneries de téléphones, même faibles, sont à proscrire. Le modèle du bruit blanc est théorique ; dans la pratique il suffit que le bruit ambiant présente des caractéristiques constantes et que les évènements qui le composent aient une durée d'évolution significativement plus courte que les phonèmes. Sous cette condition il n'est pas incompatible avec la synthèse par corpus. C'est le cas par exemple d'une ventilation d'ordinateur (sauf modulation liée à la charge CPU). On retrouvera toutefois ce bruit à l'identique dans la synthèse finale, ce qui peut nuire tant à l'intelligibilité qu'à l'image du service utilisant la synthèse vocale. On préfère donc le minimiser autant que possible.

L'utilisation de chambres (partiellement) anéchoïques, ou au moins de pièces traitées acoustiquement, permet de réduire la réverbération et le bruit ambiant à des niveaux tout à fait acceptables.

Pour une acquisition sonore de qualité, l'utilisation d'un matériel haut de gamme et finement réglé est recommandée. Un superviseur veillera à empêcher l'apparition de saturations ainsi que l'enregistrement de souffles ou bruits de bouches. Le recours à un électro-glottographe est également fréquent. A l'aide de deux électrodes fixées sur le cou du locuteur, au niveau du larynx, il permet un enregistrement de l'activité glottique, ce qui facilite grandement la tâche de pitch-marquage discutée plus loin. Cependant cette connectique supplémentaire peut incommoder certains locuteurs et nuire au naturel des enregistrements.

Des algorithmes de traitement acoustique peuvent être utilisés pour corriger le signal a posteriori. Mais ce type de correction doit rester occasionnel. La recommandation ITU-T P.56 propose à ce sujet un algorithme de mesure du niveau de parole actif, qui peut être appliqué afin de normaliser le volume des phrases enregistrées [ITU-T 93]. Une égalisation du canal d'acquisition doit par ailleurs être envisagée lorsque les enregistrements proviennent de sources sonores variées. Stylianou propose pour cela de modéliser l'espace acoustique des différentes sources à l'aide de GMM [23], puis de compenser les variations avec des filtres autorégressifs [Stylianou 99]. On peut également faire appel à des outils de débruitage afin de réduire le bruit ambiant. La plupart d'entre eux reposent sur une approche de soustraction spectrale par filtrage de Wiener, la densité spectrale du bruit ambiant étant apprise sur des périodes de silence [Boll 79]. Notons enfin qu'il n'existe pas à ce jour de technique satisfaisante de réduction de la réverbération, hormis dans des contextes très contraints comme celui d'une acquisition multi-capteurs.

Le rôle du locuteur

La tâche du locuteur est assez ardue. Malgré la durée (habituellement plusieurs jours) et la pénibilité de l'exercice il doit faire preuve d'une grande constance dans le timbre de sa voix qui, si l'on n'y veille pas, peut dériver naturellement au fil des phrases. Avec l'envie d'en finir il devient souvent plus tendu. La prosodie doit également rester constante tout au long des enregistrements, tant dans ses valeurs moyennes que dans ses variations. Conformément au style « lecture neutre », la prosodie doit être portée par la syntaxe et éventuellement un peu par le lexique, mais en aucun cas par la sémantique. Enfin, le locuteur doit respecter certaines consignes de prononciation (liaisons, e-muets...) ou d'intonation (notamment sur les fins de phrase).

Pour favoriser la constance et faciliter le respect des consignes, un retour casque doit être mis en place. Il permet au locuteur de s'entendre, en temps réel comme en réécoute, et de communiquer avec le(s) superviseur(s). Le support de lecture peut être imprimé sur papier, ou bien affiché sur un écran via une interface dédiée.

La supervision

Les nombreuses directives exposées plus haut rendent indispensable la mission de supervision. Menée par une ou plusieurs personnes, elle consiste à s'assurer continuellement :

- de la qualité sonore des enregistrements

23. Gaussian Mixture Model

- de la constance du canal d'acquisition, en vérifiant notamment le positionnement du locuteur
- du suivi par le locuteur des consignes de lecture
- de la constance du locuteur en matières de timbre et de prosodie

Ce dernier aspect est probablement le plus subtil. Une lente dérive de la voix du locuteur est quasi inéluctable et difficile à percevoir. Une comparaison régulière des dernières phrases enregistrées à deux ou trois références fixes permet à une oreille exercée de détecter les dérives. Quelques mesures automatiques de paramètres acoustiques peuvent aider : tracé de la hauteur médiane, de l'énergie moyenne, de la durée phonémique moyenne, de l'écart-type de hauteur par phrase, etc.

Lorsque l'interface d'acquisition le permet, l'équipe de supervision réalise en temps réel la découpe en phrases des enregistrements. Ceci implique notamment d'éliminer les phrases erronées, d'effectuer un choix parmi plusieurs reprises d'une même phrase et d'écarter les enregistrements parasites (bruits divers, discussions entre le locuteur et les superviseurs, etc.).

Enfin, l'équipe de supervision a une mission d'accompagnement du locuteur : expliquer clairement l'objectif de sa prestation, le soutenir inlassablement pendant cet exercice d'endurance, formuler en douceur les demandes de correction, calmer d'éventuelles sautes d'humeur, etc. Bref il s'agit, tout au long des sessions d'enregistrement, d'assurer l'interface entre des univers scientifique et artistique, ce qui requiert parfois diplomatie, pédagogie, et dans tous les cas une attention particulière.

3.3 Post-traitement des données

3.3.1 Segmentation phonétique

Chaque phrase du script de lecture est associée à une transcription phonétique, qui a généralement été établie dès le départ afin de guider la constitution du script. Elle est déduite du texte par les hauts-niveaux du système de synthèse. La segmentation des enregistrements en unités élémentaires consiste, pour chaque phrase enregistrée, à aligner cette transcription phonétique sur le signal correspondant (figure 9).

Des outils de segmentation standards peuvent être utilisés à cet effet. Le plus populaire est probablement la bibliothèque HTK (Hidden Markov model ToolKit), issue des laboratoires de l'Université de Cambridge [Young 05]. Conçue pour la modélisation HMM de n'importe quel processus temporel, elle est principalement utilisée pour la reconnaissance de parole.

HTK comporte en particulier des outils d'apprentissage qui permettent d'estimer automatiquement les paramètres de modèles HMM à partir de phrases de référence déjà segmentées. Nous utilisons ces outils pour créer de premiers modèles « multi-locuteurs », à partir d'un ensemble de bases existantes comme par exemple les bases d'apprentissage utilisées en reconnaissance de parole. Ces bases comportent un assortiment très large d'évènements acoustiques et phonétiques prononcés par de nombreux locuteurs. Les modèles HMM qui en sont extraits se révèlent donc assez robustes et permettent la segmentation de n'importe quelle voix. A l'aide de ces modèles et des outils d'alignement également fournis dans HTK, nous obtenons une première segmentation phonétique de nos enregistrements.

Mais la grande variété acoustique prise en compte dans les modèles multi-locuteurs nuit à leur précision lorsqu'il s'agit de segmenter une unique voix bien identifiée. C'est pourquoi cette première base, segmentée approximativement avec les modèles multi-locuteurs, est utilisée comme point de départ pour l'apprentissage de modèles « mono-locuteurs », toujours avec HTK.

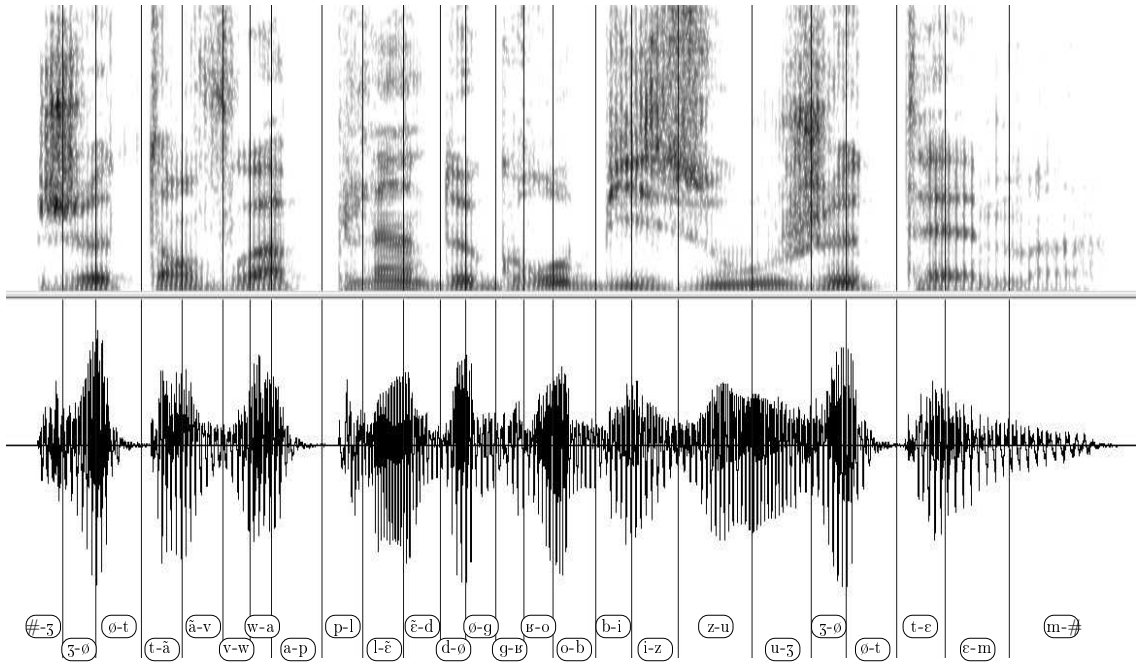


FIGURE 9 – Segmentation en diphones de la phrase « *Je t'envoie plein de gros bisous je t'aime* ». La partie inférieure correspond à la forme d'onde et la partie supérieure au spectrogramme. On peut y distinguer les parties stables du signal sur lesquelles sont placées les frontières de diphones.

On obtient ainsi des modèles plus performants car spécialisés sur la voix en question. Les outils d'alignement sont alors appliqués pour obtenir une nouvelle segmentation, plus précise, de l'ensemble de la base. Et ainsi de suite... on peut réitérer ce processus d'adaptation à la voix du locuteur jusqu'à ce qu'un critère d'arrêt soit vérifié, comme par exemple un déplacement non significatif des marques de segmentation. L'efficacité de l'adaptation est étroitement liée à la durée de parole disponible. Si la base contient moins de 10 minutes de parole utile, la segmentation ne sera probablement pas améliorée, voire sera dégradée. Au-delà de 30 minutes de parole, l'amélioration de la segmentation peut être très significative. Le gain dépend également de la voix et de la diction du locuteur. S'il a un timbre très particulier, ou si certains phonèmes sont régulièrement prononcés de manière singulière, les modèles mono-locuteurs ont de fortes chances de surpasser les modèles multi-locuteurs.

La stratégie à adopter dépend donc du contenu de la base. Dans tous les cas, des erreurs d'alignement subsistent. Il peut s'agir soit d'imprécisions dans le placement des marques de segmentation, comme fréquemment sur les liquides ou semi-voyelles, soit de différences entre la suite phonétique prévue et celle effectivement prononcée par le locuteur : assimilations, ajouts ou retraits de pauses, introductions de e-muets, ou encore prononciation inattendue de noms propres. Les écarts phonétiques peuvent être lourds de conséquences. Ils s'accompagnent souvent d'imprécisions sur la segmentation des phonèmes environnants voire, en cas d'erreurs groupées, d'un déraillement complet du processus d'alignement sur l'ensemble de la phrase. Pour pallier ce type d'erreur, des variantes peuvent être introduites dans la séquence phonétique, comme illustré en figure 10. Le chemin phonétique le plus proche de la réalisation acoustique est alors choisi de manière automatique par les outils d'alignement. Quoique cette approche par variantes phonétiques soit séduisante, nous expliquerons en partie IV pourquoi nous avons préféré nous en tenir à l'alignement forcé d'une unique séquence phonétique.

La segmentation phonétique est d'autant plus précise que la supervision des enregistrements a été rigoureuse : respect de la séquence phonétique, constance de la voix, qualité sonore, etc.

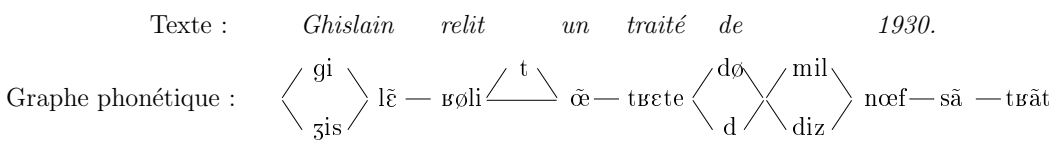


FIGURE 10 – Graphe des variantes phonétiques pour la phrase « *Ghislain relit un traité de 1930* ». Le prénom et l'année peuvent tous les deux être prononcés de deux manières différentes, la liaison après « *relit* » est facultative et enfin le [ø] final de « *de* » peut être élidé.

Mais pour obtenir une segmentation irréprochable, la tentation est grande de procéder à une ultime vérification manuelle... du moins lorsque les moyens financiers le permettent, car cette tâche est extrêmement coûteuse : environ un mois de travail est nécessaire à un expert pour revoir, avec des outils adaptés, la segmentation phonétique d'une heure de parole. Le résultat est-il fiable pour autant ? Pas forcément, si l'on en croit les conclusions de [Kawai 00]. Dans ce travail cinq opérateurs ont procédé, chacun de leur côté, à une correction manuelle de la segmentation phonétique de 50 phrases japonaises. Les auteurs ont ensuite mesuré les écarts de segmentation entre opérateurs et les ont comparés aux imprécisions d'une segmentation automatique : 90% des étiquettes manuelles sont à moins de 3.5 ms de la segmentation de référence (définie comme la moyenne des segmentations manuelles), contre 4.6 ms pour les étiquettes automatiques. On en déduit que, dans la configuration de cette expérience, l'étiquetage manuel n'offre pas une fiabilité absolue et qu'il surpasse seulement de très peu le système automatique. Toutefois, dans le même travail, des tests perceptifs montrent une légère préférence des auditeurs pour la synthèse vocale résultant de la segmentation manuelle. Cette dernière conclusion n'est cependant pas unanime, puisque dans [Makashay 00] l'évaluation perceptive conclut plutôt en faveur d'une segmentation automatique. Nous présenterons en partie IV nos propres résultats dans ce domaine.

Qu'elle soit manuelle ou automatique, la segmentation de la base porte généralement sur des phonèmes (dans un souci de simplicité et de performance de l'alignement). Dans le cas des diphones, il faut donc convertir les frontières de phonèmes obtenues en frontières de diphones. Par définition il s'agit alors de localiser les zones acoustiquement stables. On se contente généralement de choisir les milieux de phonèmes. Pour les occlusives, une recherche de la zone de plus faible énergie peut être envisagée. Le placement précis des marques milieux peut également être réalisé à l'exécution de la synthèse : chez [Conkie 97] le lieu de concaténation optimal est ainsi choisi dynamiquement, de façon à minimiser la distorsion acoustique entre diphones consécutifs pendant l'étape de concaténation.

3.3.2 Annotation des données segmentées

Pour pouvoir être utilisées dans le système de synthèse, les unités élémentaires constituant la base de données doivent préalablement être annotées. Ces annotations apportent des précisions sur le contexte et les caractéristiques acoustiques de chaque unité. Elles permettent d'affiner le travail du système de sélection, tant pour le calcul des distorsions entre unités successives (coût de concaténation) que pour la comparaison des candidats à la séquence-cible (coût-cible). Ce dernier aspect implique que les annotations comportent au moins les composantes du vecteur contextuel produit par les hauts-niveaux pour la caractérisation des unités-cibles. Une liste non exhaustive de ces composantes est proposée page 36. Mais d'autres annotations peuvent être ajoutées, comme par exemple les mesures acoustiques nécessaires au calcul du coût de concaténation.

Les marqueurs symboliques, liés aux contextes phonétique et linguistique, sont calculés di-

rectement par les hauts-niveaux à partir des phrases du script d'enregistrement. Pour chaque phrase, une étape d'alignement et d'adaptation des marqueurs peut être nécessaire si la séquence phonétique a fait l'objet d'une correction manuelle suivant la prononciation du locuteur.

Quant aux marqueurs numériques, ils sont traditionnellement issus de mesures acoustiques et dépendent donc entièrement des données enregistrées par le locuteur. Outre le guidage de l'algorithme de sélection, ces marqueurs peuvent servir aux éventuels traitements de lissage et de placage prosodique. Parmi les paramètres acoustiques fréquemment utilisés pour enrichir les enregistrements, on trouve le pitch-marquage, l'énergie et les mesures spectrales.

Le pitch-marquage consiste à localiser les instants de fermeture glottique, ou Glottal Closure Instants (GCI), dans le signal de parole. Cette opération, qui va de pair avec l'estimation de la fréquence fondamentale, est aussi indispensable que délicate. De nombreux algorithmes de traitement du signal vocal travaillent en effet de manière pitch-synchrone, ce qui suppose la disponibilité d'un pitch-marquage fiable. Son calcul est certes aisé lorsqu'un signal électroglottographique mesurant l'activité glottique est disponible. Mais en l'absence de tels enregistrements, il faut recourir à des techniques d'analyse du signal de parole. Ces dernières exploitent le plus souvent la périodicité des signaux voisés : les pics de corrélation sont alors recherchés au sein du signal, en se focalisant sur des extremums locaux [Laprie 98]. Cependant la périodicité du signal est imparfaite et ces approches conduisent fréquemment à des multiplications ou divisions de périodes. Même si certaines de ces erreurs peuvent être corrigées avec des mécanismes de suivi temporel, les imprécisions restent fréquentes, notamment pour la détection des parties voisées. Vincent et Rosec ont proposé un algorithme performant basé sur un modèle de décomposition source-filtre de type ARX-LF [Vincent 06]. Bien que cet algorithme puisse être mis en défaut sur certaines voix aiguës ou pathologiques, il présente une grande robustesse aux problèmes de doublement, et surtout un niveau de précision sans précédent. En contrepartie il nécessite d'importantes ressources CPU : environ 1 minute de calcul pour 1 seconde de parole sur un processeur courant. On est donc bien loin du temps réel, ce qui interdit l'utilisation de cette technique pour la création de voix de synthèse « à la volée », c'est-à-dire au fur et à mesure des enregistrements (par exemple dans le cas d'une interface entièrement automatisée de création de voix).

Les mesures utilisées pour la caractérisation de l'énergie et du spectre découlent directement des fonctions de coût décrites en 2.4. Notons enfin que certains marqueurs symboliques peuvent être déduits de mesures acoustiques par l'intermédiaire de modèles prosodiques, comme évoqué page 27 : ToBI, Tilt... Ils jouent dans ce cas un rôle intermédiaire entre les marqueurs numériques et les marqueurs purement symboliques.

3.3.3 Compilation du dictionnaire de synthèse

La base de données numérique des enregistrements segmentés et annotés du locuteur se présente le plus souvent sous la forme de nombreux fichiers, plus ou moins redondants et plus ou moins lisibles. Cette organisation se révèle sous-optimale lorsqu'il s'agit d'exécuter la synthèse vocale dans un environnement contraint : téléphone mobile, serveur téléphonique recevant de nombreux appels simultanés, lecteur de livre électronique, etc. Pour minimiser le besoin en ressources matérielles (mémoire vive, espace disque et CPU), un réaménagement « binaire » des données doit être envisagé. Nous appelons cette nouvelle organisation le « dictionnaire de synthèse ». Il permet une meilleure utilisation de l'espace mémoire et une accélération du processus de sélection, notamment grâce à un regroupement des accès mémoire. Ce réaménagement est également l'occasion de compresser les données acoustiques avec l'un des nombreux codecs existants. Pour diminuer encore plus l'empreinte mémoire de la voix de synthèse, il est

également possible de réduire l'inventaire des unités, par exemple en ne conservant que celles qui sont statistiquement les plus utilisées [Rutten 02].

Ce dictionnaire inclut également le résultat de tous les algorithmes d'apprentissage exécutés sur la base de données : arbres de classification des unités, modèles de prédiction prosodique, pré-calcul de certaines composantes du coût de concaténation, etc. Enfin il peut faire intervenir des techniques de chiffrage permettant de protéger les données (par exemple en vue d'un licensing).

3.4 Une approche différente : la collecte de rushes

Nous présentons maintenant une technique radicalement différente pour la préparation d'une voix de synthèse par corpus. Plutôt que de procéder à des enregistrements spécifiques, elle consiste à recueillir des données vocales à partir de sources non dédiées à la synthèse : films sur DVD, discours politiques mis en ligne sur internet, chroniques radio, livres audio... Nous appelons cela la « synthèse par rushes ». Au premier abord cette approche semble séduisante car elle permet de collecter des données à moindre coût, de synthétiser la voix de personnes décédées ou encore de personnes qui ont perdu l'usage de la parole. Pour certaines de ces sources, la disponibilité d'une transcription textuelle permet même d'envisager un processus entièrement automatique de création de voix : transcription phonétique à partir du texte, segmentation, annotation, pitch-marquage, etc.

En pratique la « synthèse par rushes » fait face à de nombreuses difficultés. La première porte sur la collecte de données vocales propres, homogènes, et en quantité suffisante. Par exemple dans les bandes sonores des productions audiovisuelles, une grande partie des dialogues sont inutilisables en synthèse par corpus pour cause de chevauchement entre les répliques des différents personnages, ajout de musique ou présence d'un bruit de fond. Ainsi pour Homer, personnage principal de la série *les Simpsons*, nous n'avons pu récolter qu'une heure et demie de parole utile sur trois saisons complètes de la série (74 épisodes). En outre, le canal d'acquisition peut varier significativement selon que les scènes se jouent en intérieur ou en extérieur (essentiellement pour les films ou séries en version originale). Dans le cas de discours politiques, les données vocales peuvent certes se montrer beaucoup plus abondantes, mais des problèmes d'homogénéité subsistent. Ainsi les 11 heures de parole que nous avons pu récolter pour l'ancien président Jacques Chirac présentent des inégalités gênantes dans le cadre d'une synthèse par corpus : distance au microphone, niveau sonore ambiant, amplitude et durée de la réverbération, et surtout une évolution de la voix tout au long des deux mandats (devenue plus grave et plus rocailleuse au fil des années). A l'inverse les livres audio constituent une source de matière acoustique de qualité supérieure.

Outre les problèmes de qualité et de quantité sonore, on rencontre avec les rushes de fréquentes approximations dans la transcription textuelle (lorsqu'elle est disponible). Par exemple les sous-titrages de productions audiovisuelles sont beaucoup plus synthétiques que les dialogues d'origine et comportent couramment des erreurs. Les personnalités publiques ou chroniqueurs prennent également de nombreuses libertés dans leurs discours par rapport à la préparation écrite. Encore une fois, seuls les livres audio se démarquent en faisant preuve d'une grande fidélité au texte d'origine. D'une manière générale les erreurs de transcription sont responsables d'incohérences souvent grossières entre la chaîne phonétique prédite et celle effectivement réalisée. Il peut en résulter de grosses inconsistances dans la base de données, avec un impact très négatif sur la qualité de synthèse. Pour cette raison on réalise souvent une vérification de la transcription textuelle, voire une saisie complète lorsqu'aucune transcription n'est disponible.

On peut utiliser pour cela l'outil Transcriber [24], développé par la DGA [25], dont l'ergonomie a été spécialement pensée pour ce type de travail. Il permet de saisir ou corriger la transcription textuelle tout en contrôlant efficacement sa synchronisation phrase par phrase avec le signal de parole. La figure 11 présente une capture d'écran de cette application libre et gratuite.

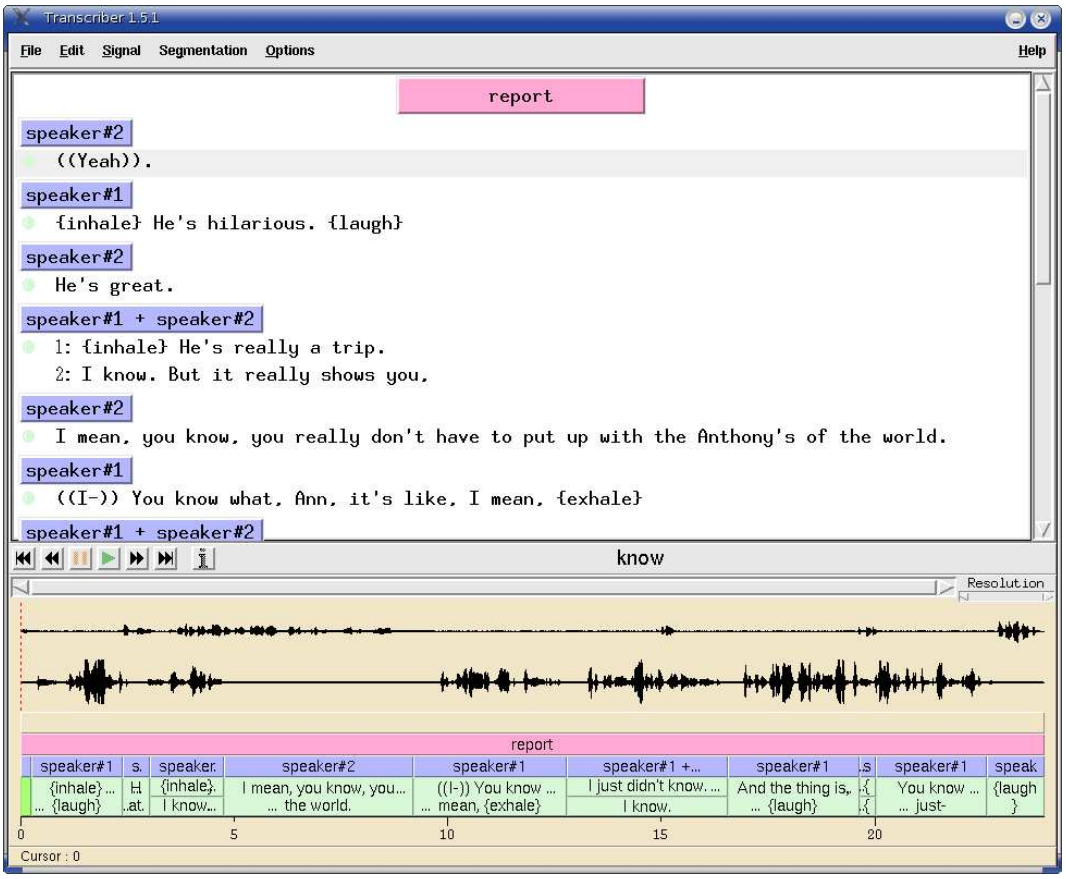


Figure 11 – Capture d'écran de l'application Transcriber, développée par la DGA, permettant l'annotation manuelle d'un signal de parole : transcription, découpage en phrases, repérage des tours de parole, etc. Cette image est extraite du site de Transcriber.

Une autre difficulté liée aux rushes concerne leur richesse prosodique. Cette matière sonore a généralement été enregistrée dans un cadre moins contraint que celui de la création d'une voix de synthèse. Il en résulte un niveau d'expressivité qui dépasse largement les possibilités de nos modèles prosodiques et de nos systèmes d'annotation automatique des données, conçus pour la lecture neutre. On peut ainsi rencontrer d'importantes modulations de qualité de voix, d'intensité, de hauteur, de rythme, voire même de timbre. En cela les livres audio sont très pénalisés, surtout lorsqu'il s'agit d'un narrateur professionnel : changement du timbre dans les dialogues en fonction du personnage, mise en relief de la narration et de la sémantique par la prosodie, imprégnation de la qualité de voix et de la diction par la tension de la scène, etc. Toute cette variété constitue indéniablement une grande richesse qui doit pouvoir être modélisée et exploitée. Néanmoins elle met considérablement en défaut notre approximation « lecture neutre » et, incontrôlée, elle se répercute de manière anarchique dans le signal de synthèse : sauts permanents de F0, rythme chaotique, timbre instable, etc.

Enfin la dernière difficulté de la synthèse par rushes, qui n'est pas la moindre, provient du fait que nous n'avons aucun contrôle sur la couverture linguistique de la base. Naturellement il

24. http://trans.sourceforge.net/en/presentation.php
25. Direction Générale de l'Armement

n'est pas question ici d'optimiser le script de lecture. Les transcriptions textuelles dépendent entièrement des données collectées et se montrent souvent très redondantes : répétitions de noms de personnages, de noms de lieux, de mots spécifiques liés au contexte de l'action ou du discours... On trouve par exemple dans les discours de Chirac une sur-représentation des vocabulaires politique, diplomatique et économique : noms de présidents, noms de pays en guerre, indicateurs économiques... Tous les rushes proviennent inéluctablement d'un domaine qui les contextualise de manière parfois très restrictive. On peut bien entendu envisager de les utiliser uniquement pour la synthèse vocale de ce même domaine restreint ; mais même dans ce cadre, la couverture linguistique reste fortement pénalisée par le niveau élevé de redondance. Nous présenterons en partie IV quelques résultats quantitatifs à ce sujet et comparerons, en termes de qualité de synthèse, les performances des rushes à celles d'enregistrements dédiés.

Notons pour finir que ces deux approches ne sont pas totalement incompatibles ; elles peuvent tout à fait être réunies pour la création d'une unique base composite. Dans ce cas les rushes apportent la richesse prosodique, tandis que les enregistrements dédiés complètent la couverture linguistique. La mise au point d'un système de sélection performant est cependant requise pour dépasser le verrou technologique du contrôle prosodique des rushes.

4 La problématique de l'évaluation

Le présent document propose de nouvelles techniques de création de voix de synthèse. Pour en apprécier la pertinence nous devrons mettre en oeuvre des méthodologies d'évaluation qualitative de la synthèse de parole à partir du texte. Il s'agit là d'un sujet épineux car il touche à la perception humaine de la parole, qui revêt de nombreux caractères difficiles à appréhender. Les mécanismes de perception peuvent considérablement varier suivant le sujet, ses usages et le domaine applicatif : langues tonales *vs.* non tonales, vocalisation par téléphone d'informations boursières *vs.* lecture de livres numériques...

L'étape d'évaluation est cependant essentielle car elle permet de mesurer les progrès effectués et de guider la recherche. Elle peut porter sur la qualité globale du système de synthèse, ou bien être focalisée sur un module précis [van Santen 93] : découpage syntaxique, conversion graphème-phonème, prédiction prosodique, sélection d'unités, lissage des concaténations ou encore placage prosodique. Nos travaux, qui portent sur la préparation de bases de données en synthèse par corpus, influencent essentiellement les bas-niveaux du système. Dans la suite nous ne traiterons donc pas de l'évaluation des hauts-niveaux ; nous nous concentrerons sur l'évaluation des scripts de lecture et de la qualité de synthèse globale.

4.1 Critères objectifs

4.1.1 Vers une évaluation globale et automatique de la synthèse

De nombreux travaux ont pour ambition de modéliser, au moins partiellement, la perception humaine d'un signal de parole, afin d'éviter le recrutement long et coûteux d'auditeurs pour les campagnes d'évaluation perceptive. Aucune fonction de mesure objective ne permet à ce jour de remplacer totalement l'oreille humaine. Il existe néanmoins des outils qui remplissent très partiellement cet objectif et qui peuvent s'avérer utiles dans des cadres expérimentaux spécifiques.

Le domaine du codage de la parole a vu de nombreuses avancées dans ce domaine. La méthode intitulée « Perceptual Evaluation of Speech Quality » [Rix 01], abrégée PESQ, offre un

moyen d'appréciation automatique de la dégradation subie par un signal de parole à travers un réseau téléphonique complet. Normalisée dans la recommandation P.862 de l'ITU-T [ITU-T 01], elle utilise différentes techniques d'alignement temporel et de représentation psychoacoustique. La corrélation de cette mesure objective à une vraie notation subjective a été estimée à 0.935, ce qui semble très bon.

Cernak et Rusko ont appliqué cette méthode à la synthèse de parole à partir du texte [Cernak 05] et ont obtenu d'excellentes corrélations aux notations subjectives. Toutefois leurs expériences n'ont porté que sur la synthèse de mots isolés et avec un système de synthèse par diphones, l'objectif étant de comparer différentes techniques de placage prosodique (LP, RELP et PSOLA). Le contexte est donc finalement assez proche des problématiques de codage de la parole. En synthèse par corpus les mécanismes de dégradation du signal de parole sont différents et tendent à survenir de manière très localisée : ruptures prosodiques et spectrales, coarticulation inadaptée au contexte, courbe mélodique inconsistante, etc. Un peu plus tôt, Chen et Campbell avaient proposé une approche similaire dans le cadre spécifique de la synthèse par corpus (sans placage prosodique) [Chen 99]. Les meilleures corrélations obtenues avec des notations subjectives étaient de l'ordre de 0.75. La marge d'erreur associée reste non négligeable, d'autant qu'elle vient s'ajouter à celle de la notation subjective elle-même (que nous détaillerons plus bas). Apparemment ces techniques ne permettent pas, en synthèse par corpus, de comparer finement des voix de qualité relativement proche.

D'une manière générale, les composantes du coût de sélection prennent déjà en compte des aspects segmentaux, prosodiques et articulatoires, dans le but de refléter au mieux la perception humaine du signal reconstitué. Il est donc naturel de mesurer les corrélations entre ces différentes composantes (ou leur somme) et des notations subjectives. Aux paragraphes 2.4.2 et 2.4.3 (voir page 41), nous avons rapporté de nombreux travaux qui s'intéressent à cette corrélation pour améliorer la définition des coût-cible et coût de concaténation. Dans [Chu 01a], une vaste campagne d'évaluation subjective est menée et une corrélation avec la fonction de coût supérieure à 0.8 est observée. Les erreurs dans la prédiction de notation sont alors de l'ordre de 0.4 sur une échelle de 5, ce qui semble prometteur. Toutefois ces expériences concernent un système (celui de Microsoft Research China -MSRCN-) et une langue (le Mandarin) particuliers ; ils ne sont pas forcément reproductibles dans d'autres configurations. Les mesures rapportées dans ce document seront en tout cas très différentes. Quoi qu'il en soit, l'observation des coûts de sélection, si elle ne permet probablement pas de se passer des mesures subjectives, est porteuse d'informations et nous y aurons recours par la suite.

4.1.2 Évaluation des scripts de lecture

La méthode idéale pour apprécier la qualité d'un script de lecture consiste à l'utiliser pour créer puis évaluer une ou plusieurs voix de synthèse. On peut ainsi comparer de manière très fiable plusieurs scripts de lecture en les soumettant à un panel identique de locuteurs, en construisant de la même manière les voix de synthèse, puis en procédant à une évaluation perceptive et comparative équitable. Cette méthode n'a été que très rarement utilisée par le passé car, pour les tailles de script courantes (plusieurs jours d'enregistrement), elle se révèle extrêmement onéreuse. Nous l'utiliserons largement en partie IV.

Nous ferons également appel par la suite à de nombreux indicateurs objectifs, qui permettent d'anticiper partiellement la qualité de synthèse finale à partir du script. Naturellement le premier indicateur est la taille du script. Plus il est grand et meilleure pourra être la synthèse finale, mais en contrepartie les coûts d'enregistrement et de post-traitement sont plus importants. Par ailleurs, comme nous l'avons vu dans la section 3.1, les taux de couverture de différentes unités linguistiques sont fréquemment utilisés : diphones, triphones, syllabes, mots,

etc., toutes ces unités pouvant être inventoriées avec ou sans leur contexte linguistique ou prosodique. Ces taux de couverture permettent d'anticiper, simplement à partir du script, la marge de manoeuvre dont disposera le système de synthèse pour sélectionner les unités nécessaires lors de l'exécution. Nous proposerons dans ce document une analyse inédite du lien entre ces mesures objectives, le coût de synthèse et la qualité perceptive finale.

4.2 Critères subjectifs

Dès l'enregistrement par un locuteur, les scripts de lecture peuvent faire l'objet d'une première évaluation subjective. On peut en effet recueillir les impressions du locuteur, par exemple sur la lisibilité ou la monotonie des phrases. On peut également se positionner à mi-chemin entre l'objectif et le subjectif en relevant des indicateurs représentatifs de la facilité de lecture : temps de lecture moyen par mot, nombre de reprises, etc. Mais, encore une fois, le meilleur moyen d'apprécier la qualité du script reste l'évaluation subjective de la (ou des) voix de synthèse finale(s). Une étude complète des techniques d'évaluation subjective en synthèse de la parole est proposée dans [Boëffard 02] et [D'Alessandro 04]. Cette évaluation peut porter sur plusieurs aspects. La littérature en retient généralement deux, que nous exposons ci-après : l'intelligibilité et le naturel.

4.2.1 Intelligibilité

En première approximation les systèmes TTS servent à véhiculer une information linguistique sur un support vocal. Il convient donc avant tout de vérifier dans quelle mesure cette fonction est bien réalisée. La difficulté essentielle provient du fait que l'être humain perçoit et comprend un texte de manière contextuelle. Ainsi, à l'oral, de nombreux mots peuvent être déduits de leur contexte par des considérations sémantiques et syntaxiques, même s'ils n'ont pas été compris isolément. Cela peut biaiser l'estimation de l'intelligibilité en synthèse de parole. Une synthèse de mauvaise qualité peut ainsi se révéler compréhensible lorsqu'il s'agit de transmettre un contenu linguistique fortement prédictible, comme par exemple une histoire pour enfants. Ce mécanisme de compréhension globale est d'ailleurs partiellement reproduit dans les systèmes de dictée vocale, où des « modèles de langage » apportent des informations statistiques sur les séquences de mots afin de pallier les déficiences éventuelles du décodage acoustico-phonétique.

Des tests d'intelligibilité permettant de dépasser cette difficulté ont été mis au point. Ils consistent principalement à synthétiser des stimuli hors-contexte et non porteurs de sémantique, puis à les soumettre à un panel d'auditeurs.

Parmi les approches courantes, on trouve le **test de rime**, ou *Rhyme test*. Introduit par Fairbanks en 1958 [Fairbanks 58], il vise à évaluer l'intelligibilité de mots monosyllabiques isolés. Il existe plusieurs déclinaisons de ce test, toutes focalisées sur la compréhension de consonnes attenantes à un noyau vocalique constant (d'où la notion de « rime »). Le *Fairbanks Rhyme Test* (FRT) porte sur des stimuli monosyllabiques de type VC. Les sujets écoutent des séries de stimuli et retranscrivent ce qu'ils pensent avoir entendu. Le *Modified Rhyme Test* (MRT) [House 65] utilise des stimuli de type CVC, qui diffèrent soit par la consonne initiale, soit par la consonne finale, comme par exemple « *FUN, SUN, BUN, NUN, RUN, GUN* » en anglais. Les sujets cochent leur réponse parmi des choix multiples. Dans le *Diagnostic Rhyme Test* (DRT), des paires de mots CVC sont constituées, chacune servant à tester un trait distinctif bien précis de la consonne initiale : voisement, nasalité, etc. Ce test permet donc un diagnostic plus précis des confusions phonétiques.

Une autre approche courante, issue du projet européen ESPRIT-SAM [26], consiste à évaluer l'intelligibilité sur des phrases sémantiquement imprédictibles. Plus précisément le *Semantically Unpredictable Sentences Test*, ou **SUS test**, repose sur une technique de création automatique de telles phrases, à partir de cinq structures syntaxiques de base et à partir de lexiques de mots monosyllabiques [Benoît 96]. L'approche a été transposée dans 6 langues européennes dont le Français. Voici deux exemples de SUS en anglais :

```
The unsure steaks overcame the zippy rudder.
The dank geniuses woke the humane emptiness.
```

Ces phrases imprédictibles sont vocalisées avec le système de synthèse à évaluer, puis les prompts audio sont soumis à des auditeurs qui doivent les retranscrire. Le pourcentage de phrases correctement retranscrites, abstraction faite des homophones (*mère*/*mer*), donne alors, pour un panel représentatif d'auditeurs, le niveau moyen d'intelligibilité d'un système donné. Le test SUS reproduit mieux les conditions moyennes d'utilisation de la synthèse vocale que les méthodes basées sur des mots isolés. En ce sens il est probablement plus proche de nos préoccupations.

4.2.2 Naturel

L'évaluation du naturel ou de l'agrément global d'une voix de synthèse est très répandue. D'une manière générale, elle consiste à faire écouter à un panel d'auditeurs natifs et naïfs (c'est-à-dire ignorants de la technologie et du contexte expérimental) un ensemble représentatif de prompts synthétiques auxquels ils attribuent une note perceptive. Le critère de notation, les catégories de réponses, l'ordonnancement des prompts, etc., peuvent évoluer en fonction des protocoles.

La recommandation ITU-T P.800, qui est une mise à jour de la recommandation ITU-T P.80, normalise un ensemble de protocoles pour l'évaluation subjective des systèmes de transmission vocale [ITU-T 96]. La recommandation distingue les *essais d'opinion de conversation* des *essais d'opinion d'écoute*.

Les *essais d'opinion de conversation* visent à reproduire les conditions réelles d'un service téléphonique telles qu'elles sont perçues par les usagers : deux sujets sont alors invités à converser librement à travers le système de transmission vocale évalué. Bien entendu ce test est avant tout conçu pour l'évaluation des réseaux téléphoniques et/ou techniques de codage. Mais de la même manière on peut tout à fait envisager, dans le cas des systèmes TTS, de mettre des utilisateurs dans les conditions réelles d'utilisation de la synthèse vocale, par exemple en reproduisant ou simulant un service vocal téléphonique complet. Si cette approche est parfois suivie, elle présente cependant deux inconvénients. Tout d'abord elle peut apparaître trop spécialisée, la plupart des voix de synthèse étant destinées à des usages et services multiples. Ensuite sa mise en place peut se révéler complexe voire coûteuse.

Pour ces raisons les *essais d'opinion d'écoute* sont généralement préférés. Il en existe essentiellement trois types.

Le « Absolute Category Rating » (ACR) est le plus répandu en synthèse vocale. Il est d'ailleurs repris dans la recommandation ITU-T P.85, qui l'applique au cas spécifique de l'évaluation de la qualité de la parole synthétique sur un serveur vocal [ITU-T 94]. L'ACR consiste à demander aux sujets de noter des prompts synthétiques de manière absolue. La table 2 rapporte trois exemples de critères absolus, ainsi que les échelles de notation correspondantes.

26. Multilingual Speech Input/Output : Assessment, Methodology and Standardisation

Contrairement à l'ACR, le « Degradation Category Rating » (DCR) compare le système-cible avec une référence fixe de haute qualité (par exemple de la parole naturelle). Des paires A-B sont ainsi présentées aux auditeurs, où A est un échantillon de référence et B sa version dégradée par le système. Des paires répétées A-B-A-B peuvent aussi être utilisées pour affiner le jugement de l'auditeur. La table 3 précise l'échelle de notation utilisée pour les DCR.

Enfin, le « comparison category rating » (CCR) permet d'estimer les niveaux de qualité relatifs de deux systèmes A et B. Il se distingue du DCR par deux aspects : d'une part l'ordre de présentation des deux systèmes comparés varie aléatoirement d'un prompt à l'autre (A-B ou B-A), d'autre part l'échelle de notation est symétrique. La table 4 rapporte l'échelle la plus simple qui puisse être envisagée dans le cadre d'un CCR ; elle invite simplement les auditeurs à exprimer leur préférence ou non pour l'un des deux systèmes. La table 5 rapporte une autre échelle, plus riche, qui réunit en fait deux jugements de la part de l'auditeur : « Quel est l'échantillon de meilleure qualité ? » et « Quelle est la différence de qualité entre les deux échantillons ? ». Pour ces deux échelles, un post-traitement doit être opéré pour harmoniser la polarité de chaque note en fonction de l'ordre de présentation A-B ou B-A.

Impression générale

Comment jugez-vous la qualité sonore de ce que vous venez d'entendre ?	*Note*
Excellente	5
Bonne	4
Passable	3
Médiocre	2
Mauvaise	1

Effort d'écoute

Comment qualifiez-vous l'effort d'écoute nécessaire pour comprendre le message ?	*Note*
Détente absolue ; aucun effort	5
Attention nécessaire, pas d'effort appréciable	4
Effort modéré	3
Effort considérable	2
Incompréhensible en dépit de tous les efforts possibles	1

Débit

Le débit moyen de l'énoncé était :	*Note*
Beaucoup trop rapide	5
Un peu trop rapide	4
Satisfaisant	3
Un peu trop lent	2
Beaucoup trop lent	1

TABLE 2 – Exemples de critères d'évaluation absolus, avec les échelles de notation correspondantes, tirés de [ITU-T 94].

Quelle que soit la stratégie retenue, l'analyse des résultats se fait souvent par le calcul de la moyenne arithmétique des notes attribuées par l'ensemble des auditeurs et sur l'ensemble des

Échelle des catégories de dégradation

Catégorie de dégradation	*Note*
Dégradation inaudible	5
Dégradation audible mais pas gênante	4
Dégradation un peu gênante	3
Dégradation gênante	2
Dégradation très gênante	1

Table 3 – Échelle des catégories de dégradation [ITU-T 96].

Échelle des catégories de comparaison

Catégorie de comparaison	*Note*
Meilleure	1
Equivalente	0
Moins bonne	-1

Table 4 – Échelle simplifiée des catégories de comparaison.

prompts. On parle alors de ***Mean Opinion Score***, ou **MOS**. Les symboles DMOS ou CMOS sont également utilisés lorsqu'il s'agit respectivement d'un DCR ou d'un CCR.

La présence d'échantillons de contrôle au sein du test, comme de la parole naturelle en lieu et place de prompts synthétiques, ou bien des paires de prompts identiques dans le cas d'un CCR, permet de valider certains résultats, de contrôler la pertinence de certains auditeurs, ou encore de déterminer la borne supérieure accessible pour le score moyen.

L'une des difficultés de ces évaluations, surtout lorsqu'il s'agit de critères absolus, porte sur le référentiel de notation. En effet chaque auditeur distribue les notes d'une manière qui lui est propre, en fonction de ses habitudes et de son expérience. Certains utiliseront toute l'échelle, d'autres se concentreront sur les notes hautes, d'autres encore sur les notes centrales ou basses. Pour un utilisateur donné, cette distribution peut même évoluer au fil de l'évaluation, en fonction des échantillons qu'il a l'occasion d'entendre et de noter. Ainsi des échantillons de haute qualité ont tendance à influencer à la baisse la suite de la notation, et réciproquement. Ce dernier effet, appelé « effet d'ordre », est facilement contrecarré en modifiant aléatoirement l'ordre de présentation des prompts d'un auditeur à l'autre. La question du référentiel global est plus délicate. On la résout partiellement en précédant le test d'un ensemble représentatif d'échantillons de qualités variées (dont des échantillons de voix naturelle). L'auditeur découvre ainsi, avant de commencer, l'éventail des qualités qu'il pourra rencontrer, ce qui fixe plus ou moins son référentiel de notation. Néanmoins les résultats MOS restent dépendants des conditions dans lesquelles il ont été mesurés. C'est pourquoi on évite généralement de comparer des MOS issus d'expériences différentes ; on préfère la confrontation de résultats obtenus de manière parfaitement équitable, par exemple au cours du même test.

Échelle des catégories de comparaison

Catégorie de comparaison	*Note*
Bien meilleure	3
Meilleure	2
Légèrement meilleure	1
A peu près équivalente	0
Un peu moins bonne	-1
Moins bonne	-2
Beaucoup moins bonne	-3

TABLE 5 – Échelle complète des catégories de comparaison [ITU-T 96].

4.3 Illustration sur un cas concret : le Blizzard Challenge

Le Blizzard Challenge a été introduit par Black et Tokuda en 2005 [Black 05]. Il s'agit d'une compétition internationale annuelle entre les systèmes de synthèse vocale, dans un cadre bien précis : la création (en un temps limité) d'une voix de synthèse à partir d'un corpus de parole commun fourni par les organisateurs. Les sous-catégories de ce challenge se sont étoffées au fil des années. Il porte désormais sur deux langues, l'anglais britannique et le Mandarin, et plusieurs corpus et sous-corpus. Les critères d'évaluation des voix de synthèse sont variés : naturel, intelligibilité, degré de ressemblance au locuteur d'origine, ainsi que l'acceptabilité dans le cadre d'applications spécifiques [King 09].

Les campagnes d'évaluation se déroulent sur internet, de sorte que des auditeurs d'origine et d'expérience très variées peuvent participer : spécialistes de la parole, étudiants natifs rémunérés, ou encore volontaires novices. Pour la partie anglaise de l'édition 2009, plus de 400 sujets ont ainsi pu être réunis pour des tests d'écoute de plusieurs types :

- Les critères de naturel, d'acceptabilité et de ressemblance au locuteur d'origine ont chacun été évalués sous la forme de MOS, suivant des échelles classiques à 5 points. Le naturel a été mesuré sur deux types d'entrées textuelles : des nouvelles journalistiques et des scripts conversationnels. Pour l'acceptabilité, un contexte de service vocal a été simulé autour des prompts synthétiques.
- Pour l'intelligibilité, les systèmes ont été soumis au test SUS décrit plus haut, en distinguant les auditeurs anglophones natifs des non-natifs.
- Enfin, un test de comparaison par paires a été utilisé, dans lequel chaque paire correspond à deux prompts synthétiques issus de deux systèmes différents. Pour chaque paire l'auditeur a dû décider si, indépendamment du contenu linguistique, les deux prompts ont une qualité sonore équivalente ou différente. Ce test est destiné à une analyse multidimensionnelle, dans le but de mieux comprendre les éléments acoustiques qui influencent la perception humaine de la parole synthétique [Mayo 05]. Quelques éléments de réponse sont rapportés dans [Clark 07] ; en l'état ce type d'analyse apporte un éclairage mitigé.

La pertinence statistique des résultats MOS doit recevoir une attention particulière : il n'est pas possible de tirer des conclusions sans tenir compte des marges d'erreurs. Plusieurs facteurs d'aléas entrent en ligne de compte : le choix des phrases de test (la qualité de synthèse peut, pour un système donné, varier en fonction de la phrase), les varations inter-auditeurs (différences de perception d'un sujet à l'autre), les variations intra-auditeurs (un auditeur peut noter différemment un même prompt à quelques minutes d'intervalle), etc. Différentes techniques d'analyse statistique permettent de préciser le contour des conclusions « autorisées ».

Dans nos travaux nous nous contenterons d'estimer l'intervalle de confiance (par exemple à 95%) de chaque résultat, à partir de la variance empirique des notes collectées et en supposant qu'elles suivent une loi de distribution gaussienne.

C'est d'ailleurs en partie pour éviter les conclusions hâtives que les organisateurs du Blizzard Challenge passent sous silence les résultats MOS. Ils préfèrent présenter les distributions de notes obtenues pour chaque système, à travers les valeurs de quartiles[27] (dont la valeur médiane). Les MOS ne sont utilisés que pour l'ordonnancement des systèmes dans les graphiques. L'autre raison pour ce choix de présentation est que les barèmes sur 5 points utilisés dans les évaluations MOS sont avant tout ordinaux, c'est-à-dire qu'ils indiquent un ordre mais ne quantifient pas vraiment l'écart perceptuel entre deux points consécutifs, sans même garantir que cet écart est constant [Clark 07]. Par conséquent l'extraction d'une moyenne arithmétique a une valeur statistique litigieuse, tandis que la comparaison de valeurs médianes et de quartiles est parfaitement justifiée.

27. Les quartiles sont des valeurs qui divisent une distribution statistique en 4 parties d'effectifs égaux.

Deuxième partie

Un appétit naturel pour les sandwichs vocaliques

Sommaire

Dans cette partie nous nous intéressons au critère d'optimisation utilisé pour la constitution de scripts de lecture. Nous introduisons pour cela une nouvelle unité phonétique, associée à un jeu simplifié de traits contextuels linguistiques et prosodiques. Cette unité tient mieux compte des spécificités de la synthèse par corpus que les unités traditionnelles et son taux de couverture est un indicateur fiable de la qualité de synthèse permise par un script de lecture.

5 Rationalisation des traits contextuels

La plupart des systèmes de synthèse vocale utilisent de nombreuses informations contextuelles pour préciser les cibles déduites du texte. Ces informations sont issues de considérations lexicales, syntaxiques, phonétiques, prosodiques et parfois sémantiques, et peuvent prendre des valeurs symboliques ou numériques. Chacun de ces plans étant associé à un ensemble de valeurs très varié, il existe pour chaque unité-cible un nombre très grand de combinaisons possibles, parfois même infini. La multiplication des contextes, bien que bénéfique pour de nombreuses applications, s'avère néfaste lorsqu'il s'agit de créer un script d'enregistrement. Nous avons déjà évoqué page 48 l'impact combinatoire des traits contextuels (attachés aux unités à couvrir) sur la dispersion de la couverture linguistique ; or une telle capacité de description excède largement le nombre des réalisations acoustiques possibles par un locuteur donné. En dispersant de manière excessive les objectifs de couverture d'unités, cette surdétermination contextuelle a pour conséquence de créer un fossé entre la couverture théorique et la couverture acoustique effectivement obtenue après enregistrement.

Pour constituer des scripts de lecture denses et efficaces, il est important de ne considérer que les traits contextuels qui ont un réel pouvoir discriminant sur les réalisations acoustiques. Dans notre cas cette rationalisation porte sur les deux aspects suivants : l'abandon des traits contextuels attachés aux consonnes et la réduction de leur nombre pour les voyelles.

5.1 L'abandon des traits contextuels sur les consonnes

Les syllabes constituent assurément, dans un flux verbal, le support premier des composantes prosodiques : rythme, accentuation et intonation se manifestent avant tout par les caractéristiques intrinsèques et relatives des syllabes successives. Par ailleurs les noyaux vocaliques, qui sont les éléments constitutifs centraux des syllabes, portent l'essentiel de la sonorité (voisement, énergie vocale). Il est donc usuel de considérer les voyelles [28] comme les porteurs principaux de la prosodie.

Bien entendu les consonnes sont aussi, mais dans une moindre mesure, influencées par leur contexte. Par exemple l'occlusion d'un [p] peut être plus ou moins longue selon le niveau d'accentuation des voyelles environnantes. Mais d'une manière générale les consonnes tendent plus à accompagner la prosodie qu'à la conduire. Dans un cadre de lecture neutre, ce dernier rôle est dévolu aux voyelles.

Partant de ce constat, on peut légitimement s'interroger sur la nécessité, dans le système de sélection, d'utiliser un coût-cible pour la sélection des unités consonantiques : les traits contextuels linguistiques et prosodiques attachés aux consonnes sont-ils vraiment significatifs du point de vue des réalisations acoustiques ? Les contraintes de continuité imposées lors de la sélection d'unités ne suffisent-elles pas à préciser, par effet de bord, la réalisation acoustique des consonnes, et notamment les phénomènes de coarticulation ?

28. Naturellement nous parlons ici des unités phonétiques, et non des graphèmes. Notre alphabet phonétique contient 14 voyelles (voir annexe) : [a], [ɔ], [o], [ɛ], [e], [ø], [œ], [ã], [õ], [ɛ̃], [œ̃], [i], [y], [u].

Pour répondre à ces questions nous avons conduit une évaluation comparative de type CCR (voir page 67) sur deux systèmes de synthèse : d'une part le système de référence des Orange Labs (système A), d'autre part une version modifiée dans laquelle nous avons supprimé le coût-cible sur les consonnes (système B). La voix Agnès, reposant sur une base standard de 6 heures de parole utile, et la voix Chirac, créée à partir de 10 heures de rushes (voir section 3.4), ont fait l'objet de deux tests séparés. Six auditeurs (trois naïfs et trois experts, tous ignorant l'objet précis du test) ont comparé, pour chacune des deux voix, 34 paires de phrases synthétisées avec les deux systèmes concurrents. Ces phrases ont été piochées aléatoirement pour chaque sujet dans un corpus textuel de 120 phrases issues de domaines variés. L'échelle de notation simplifiée de la table 4 page 68 a été utilisée. Les 4 premières phrases, dédiées à la mise en condition des sujets, ont été écartées pour l'analyse des résultats.

La figure 12 présente les histogrammes de préférence relevés sur les voix Agnès et Chirac. Les MOS sont très proches de la neutralité : 0.03 pour Agnès et 0.006 pour Chirac, sur une échelle de -1 à 1. Nous observons toutefois des distributions de notes très inégales. Pour Agnès, les auditeurs ont rarement tranché entre les deux systèmes : 86% des paires de phrases ont été jugées équivalentes. Pour Chirac les notes sont beaucoup plus dispersées : seules 27% des phrases ont reçu la note centrale. Ceci s'explique par le caractère aléatoire et instable de la voix de synthèse basée sur des rushes. Les unités candidates sont très nombreuses, redondantes par certains aspects contextuels mais hautement variables dans leurs réalisations acoustiques ; il en résulte un faible contrôle sur les restitutions et la moindre perturbation du système de sélection peut conduire à des séquences acoustiques très différentes.

Les résultats obtenus sur ces deux voix radicalement différentes semblent donc indiquer une **équivalence perceptive des deux systèmes**. La fiabilité statistique de cette conclusion repose sur le calcul du risque de deuxième ordre, c'est-à-dire le risque d'acceptation à tort de l'hypothèse nulle « les deux systèmes sont équivalents ». Ce risque est évalué à travers la *puissance observée* du test qui donne, en fonction de plusieurs paramètres dont le nombre de notes et leur écart-type, la probabilité de détecter une différence s'il y en a une. Nous estimons cette puissance grâce à la fonction `power.t.test` du logiciel libre `R`[29]. Ainsi dans la configuration de notre test sur la voix d'Agnès, nous avons, malgré le faible nombre d'auditeurs, une probabilité de 97% (puissance observée) de détecter une tendance de plus de 10% en faveur de l'un ou l'autre système, lorsqu'un seuil de significativité de 90% est utilisé. On peut donc conclure de manière assez fiable à l'équivalence entre les deux systèmes. Cette conclusion est en revanche moins tranchée pour la voix de Chirac, la puissance observée du test étant seulement de 46% (en raison d'une dispersion des notes plus grande).

A la lumière de ce résultat, la prise en compte du contexte des consonnes paraît superflue. Nous décidons donc de l'occulter pour l'optimisation de la couverture du script de lecture. Ce choix va nous permettre de limiter la dispersion des unités-cibles et ainsi de faciliter la couverture de phénomènes plus significatifs.

5.2 La simplification des traits contextuels sur les voyelles

5.2.1 Approche par arbre de classification et de régression

La position d'un noyau vocalique dans le mot et dans la phrase, ses contextes phonétique gauche et droit, sa structure syllabique, ses contextes syntaxique et prosodique [Larreur 94], sont autant de marqueurs symboliques utilisés par le système des Orange Labs pour spécifier la séquence-cible et guider la sélection. Le nombre théorique de combinaisons possibles pour

29. http://www.r-project.org/

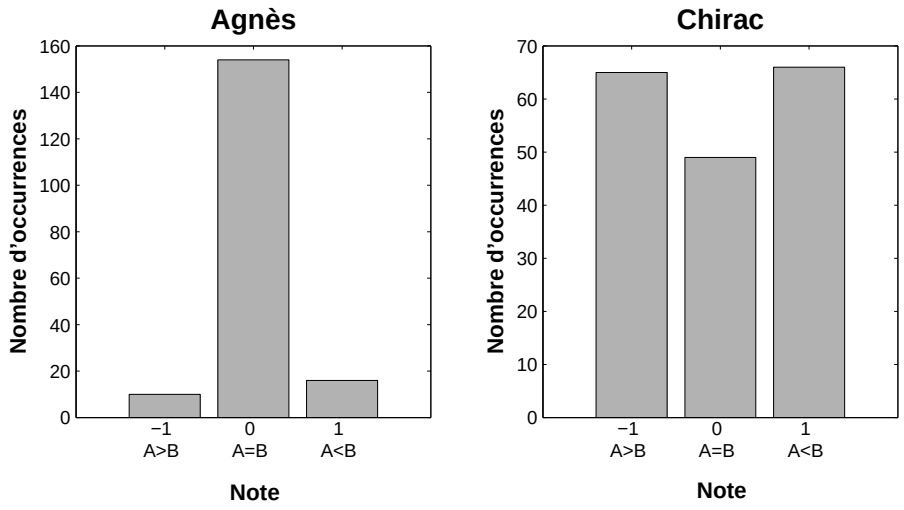


FIGURE 12 – Histogrammes des préférences entre le système A (original) et le système B (sans coût-cible sur les consonnes), pour les voix Agnès et Chirac.

décrire le contexte d'une voyelle est alors de plusieurs millions ; dans la pratique on peut en rencontrer plusieurs dizaines de milliers.

Naturellement toutes ces informations ne sont pas d'une importance capitale. Dans cette section nous souhaitons d'une part **évaluer leurs vertus descriptives**, d'autre part **établir des groupes de contextes qui permettent une réelle discrimination entre les réalisations acoustiques**. Pour cela nous suggérons de construire un arbre de régression et de classification (CART [30]).

Pour l'apprentissage de cet arbre, nous utilisons un corpus d'apprentissage de 83 252 noyaux vocaliques issus de la base féminine d'Agnès. Appelons ce corpus `agnes_appr`. Pour chacune de ces voyelles, nous disposons bien sûr de marqueurs contextuels (variables catégorielles), mais aussi de descripteurs acoustiques (variables numériques). Il s'agit donc de regrouper automatiquement les marqueurs contextuels selon une structure arborescente, de manière à prédire au mieux les descripteurs acoustiques.

Nous utilisons plus précisément **quatre variables numériques** : durée en millisecondes de la voyelle, hauteurs en demi-tons des début, milieu et fin de voyelle. Pour équilibrer ces paramètres acoustiques entre eux, nous les centrons (soustraction de leurs moyennes), les réduisons (division par l'écart-type), et les décorrélons grâce à une analyse en composantes principales (ACP). Cette dernière opération permet entre autres d'éviter que la hauteur moyenne, qui influence fortement les mesures en début/milieu/fin de phonème, ait une importance presque trois fois plus importante que la durée dans la classification.

Concernant les variables catégorielles, nous avons dû nous restreindre pour des raisons calculatoires. Certains traits contextuels ont laissé place à des versions simplifiées, déjà exploitées par ailleurs dans notre système de synthèse. La position dans le mot et la position dans la phrase ont ainsi été combinées de manière experte en un seul marqueur simplifié, tout comme les contextes syntaxique et prosodique [de Tournemire 04]. Notre apprentissage repose au final sur **trois variables catégorielles**, pour un total d'environ 300 combinaisons réalisables : structure syllabique, position syllabique et contexte syntaxico-prosodique.

Les critères phonétiques (identité et contexte) ont été volontairement écartés, tout comme

30. Classification and Regression Tree

les descripteurs spectraux (MFCC). Nous souhaitons en effet obtenir des **classes contextuelles génériques utilisables pour tous les noyaux vocaliques**, soit 14 unités phonétiques. Cette approximation va nous permettre de simplifier les considérations contextuelles dans la suite de notre travail, sans réelle perte de précision pour au moins deux raisons :

- Si des facteurs phonétiques peuvent effectivement avoir une influence marginale sur la prosodie, la structure de l'arbre a peu de raisons d'être impactée par l'identité phonétique. Autrement dit, nous faisons l'hypothèse que l'influence des contextes linguistique et prosodique va dans le même sens pour tous les noyaux vocaliques, indépendamment de leur identité.
- Les phénomènes de coarticulation, qui dépendent quant à eux étroitement de la séquence phonétique, seront bien pris en compte par les « sandwichs vocaliques » que nous introduirons dans la prochaine section. En effet ces unités englobent les noyaux vocaliques et intègrent naturellement la plupart de ces phénomènes.

C'est une différence notable avec les nombreux arbres de classification que l'on peut trouver sur ce thème dans l'état de l'art. Dans [Black 97], des techniques de CART sont utilisées pour regrouper les unités candidates dans un but de présélection, ce qui nécessite évidemment la prise en compte du contexte et de l'identité phonétiques. Ces derniers critères jouent aussi un rôle important dans [Donovan 98], où des arbres de décision sont utilisés à la fois pour la présélection et pour la prédiction d'une séquence prosodique cible. On peut également citer la synthèse par HMM, qui fait grand usage d'arbres de décision (avec critères phonétiques) pour la prédiction des paramètres acoustiques [Yoshimura 00]. Toutes ces applications sont très différentes de la nôtre : nous souhaitons seulement, par ce travail préalable, **simplifier notre description symbolique des noyaux vocaliques** en déterminant les contextes les plus discriminants sur le plan prosodique. Pour cela les considérations phonétiques peuvent être écartées.

Nous utilisons pour la construction de l'arbre le paquetage `rpart` du logiciel `R`. Ce paquetage est initialement prévu pour la prédiction d'une unique variable numérique ; nous utilisons donc une version modifiée qui l'étend à une prédiction multidimensionnelle [Boidin 09a], ce qui nous permet d'intégrer nos quatre descripteurs acoustiques. L'algorithme de classification mis en oeuvre par la fonction `rpart` divise récursivement les observations en deux sous-ensembles, présentant chacun une certaine homogénéité acoustique (matérialisée par la variance). Cette division (ou split) est opérée suivant l'un des marqueurs symboliques. Pour tous les marqueurs et toutes les partitions en deux groupes de symboles, l'algorithme mesure le gain d'homogénéité acoustique à l'intérieur des deux sous-ensembles ; le split qui offre le meilleur gain est alors retenu. Et ainsi de suite, jusqu'à ce que l'amélioration devienne inférieure à un seuil ou que le nombre d'observations devienne insuffisant. Dans `rpart`, le seuil minimum d'amélioration nécessaire pour accepter un split est contrôlé par un paramètre de complexité *cp*. Le nombre de feuilles (ou classes) de l'arbre final dépend directement de ce paramètre.

5.2.2 Résultats de la régression

La courbe inférieure de la figure 13 présente la diminution de l'erreur de prédiction RMSE [31] en fonction de la taille de l'arbre, matérialisée par son nombre de feuilles (ou classes). La RMSE initiale de la courbe vaut 1 simplement parce qu'avec une seule classe, toutes les observations sont regroupées et la RMSE n'est autre que l'écart-type acoustique (moyenne sur les 4 dimensions), qui vaut 1 du fait de la réduction évoquée plus haut.

On constate que des arbres petits offrent une description symbolique pertinente sur le plan prosodique, avec une réduction très appréciable de la RMSE dès 4 classes. La

31. Root Mean Square Error, ou racine carrée de l'écart quadratique moyen entre une valeur observée et le centroïde de la classe correspondante

figure 14 rapporte à titre informatif la structure de l'arbre correspondant à un découpage en 4 classes contextuelles. Cette structure souligne l'importance des mouvements prosodiques en fin de groupe intonatif (contours montants vs. contours descendants). Pour le reste, on constate un regroupement plutôt inattendu entre les syllabes de mots grammaticaux (par ex. les articles) et les syllabes non finales de mots finaux : il semblerait qu'il s'agisse, dans les deux cas, de noyaux en moyenne peu accentués. L'arbre à 4 classes permet donc de prédire, à partir du contexte symbolique, les 4 grands motifs prosodiques suivants : syllabe de fin de phrase montante, syllabe de fin de phrase descendante, syllabe accentuée et syllabe non-accentuée.

Pour les tailles d'arbre supérieures la RMSE ne diminue que très peu. Ceci confirme donc notre intuition que les contextes symboliques initiaux sont trop nombreux et trop précis au regard des réalisations prosodiques. Il faut souligner que, même avec un nombre élevé de classes, 32% seulement de la variance prosodique parvient à être modélisée. La part incontrôlée des réalisations est en fait régie par de nombreux facteurs inaccessibles : notions sémantiques, contexte dialogique éventuel, environnement textuel, état psychologique du locuteur, etc.

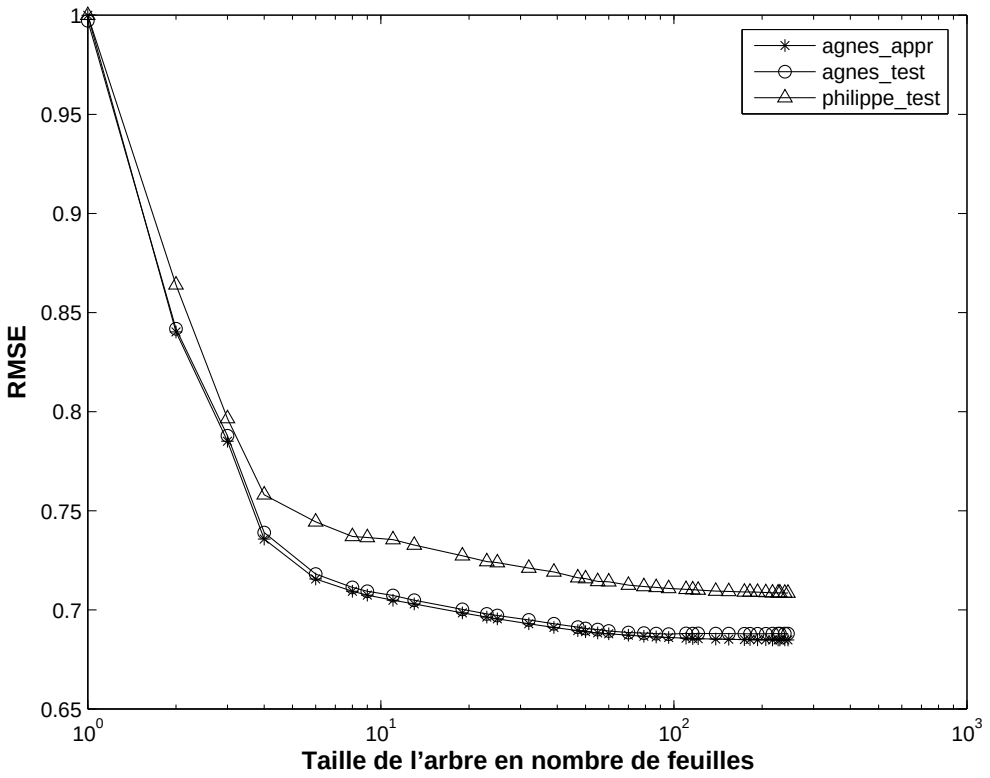


FIGURE 13 – Evolution de l'erreur de prédiction en fonction du nombre de classes contextuelles (feuilles de l'arbre de prédiction).

Nous avons ensuite effectué une validation croisée des résultats sur deux corpus de test :

- `agnes_test` : 20 811 voyelles d'Agnès, non incluses dans `agnes_appr`. Le centrage, la réduction et la décorrélation par ACP des variables numériques ont été effectués avec les paramètres mesurés sur `agnes_appr`.
- `philippe_test` : 104 479 voyelles issues de la base masculine de Philippe. Le centrage et la réduction et la décorrélation par ACP ont été effectués sur ces données, afin de compenser les écarts de tessiture et de rythme entre Agnès et Philippe.

Les CART de différentes tailles appris sur `agnes_appr` ont été appliqués sur ces corpus de

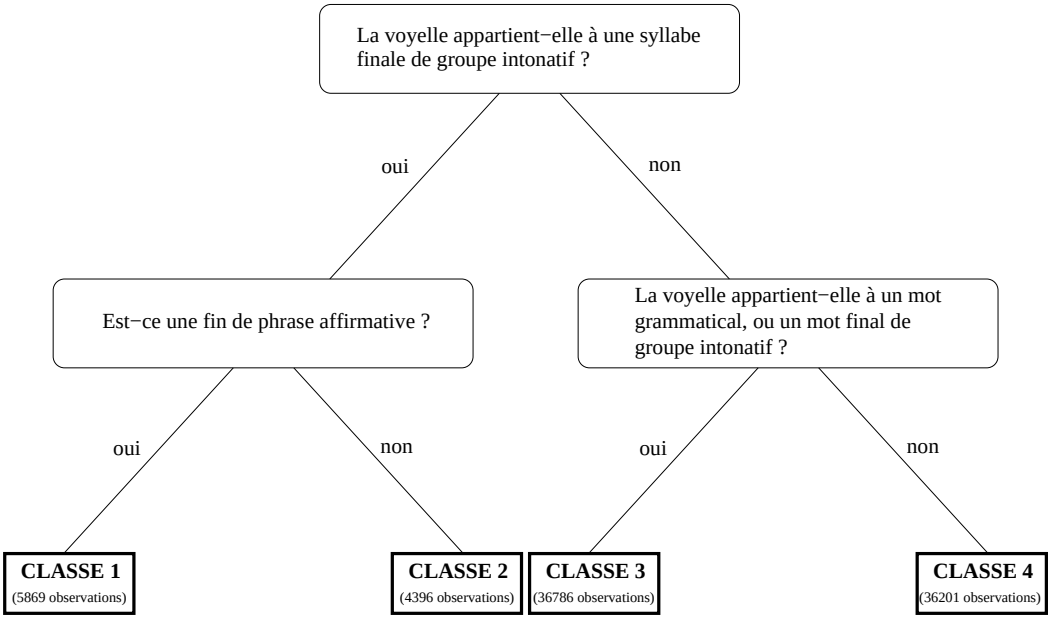


FIGURE 14 – Structure d'un arbre de classification et régression obtenu sur `agnes_appr`. Cet arbre définit ici 4 classes contextuelles.

test. Les courbes supérieures de la figure 13 rapportent l'évolution de la RMSE en fonction de la taille de l'arbre pour `agnes_test` et `philippe_test`. Les résultats sur `agnes_test` sont très proches de ceux sur `agnes_appr`, ce qui est naturel du fait que les deux corpus, bien qu'ils soient disjoints, sont issus en réalité de la même base acoustique et présentent donc des caractéristiques très voisines. La courbe mesurée sur `philippe_test` est certes moins bonne, mais présente une physionomie très comparable : l'essentiel de la réduction de RMSE est concentré sur les premières classes contextuelles.

Le nombre de classes est bien sûr majoré par le nombre de contextes possibles[32] ; il semblerait que dans notre expérience cette limite intervienne avant l'apparition du sur-apprentissage. En effet, pour les arbres de grandes tailles, la courbe de RMSE ne remonte pas sur `philippe_test` et que très peu sur `agnes_test`. Cela signifie que toutes les valeurs de nos marqueurs simplifiés apportent, au moins de manière très marginale, un certain pouvoir descriptif indépendant de la base. Nous aurions probablement observé un phénomène de sur-apprentissage si nous avions utilisé l'intégralité des contextes symboliques disponibles au départ.

En conclusion, ce travail préliminaire nous permet de regrouper l'ensemble des informations contextuelles en un nombre restreint de classes présentant des caractéristiques prosodiques bien distinctes. Ce nombre peut être dimensionné à souhait ; nos résultats montrent cependant qu'il est possible de modéliser l'essentiel des mouvements prosodiques du style « lecture neutre » simplement avec une dizaine de classes symboliques. La hiérarchie entre les classes peut bien sûr être exploitée ; mais dans notre cas elle sera « mise à plat », ce qui se justifie par le fait que nous n'utiliserons que quelques classes bien distinctes sur un plan acoustique, sans recoupement majeur.

32. En réalité il est aussi limité par la quantification numérique de nos descripteurs acoustiques, qui peut occasionner l'égalité parfaite entre plusieurs vecteurs acoustiques, et ainsi empêcher la division de certains noeuds quel que soit le seuil de complexité utilisé.

6 La notion de sandwich vocalique

6.1 Motivation

Comme nous l'avons expliqué page 47, le diphone n'est pas adapté à l'optimisation des scripts de lecture. Pour cette tâche on considère traditionnellement des unités plus longues comme les triphones, quadriphones, syllabes ou mots. Mais ces unités ne sont pas dédiées à la synthèse par sélection ; elles sont d'usage général en linguistique et dans les technologies vocales. Leur lien avec la qualité de synthèse finale n'a jamais été analysé et pourrait être, selon nous, pénalisé par le fait qu'elles n'intègrent pas les spécificités de la synthèse par sélection.

Nous nous interrogeons plus particulièrement sur la relation entre d'une part la couverture des unités traditionnelles dans le script de lecture initial, d'autre part les composantes du coût de sélection dans la synthèse finale (après enregistrement du script). Le coût-cible est certes favorisé par la prise en compte de traits contextuels lors de la constitution du script ; aussi avons-nous vu dans la section précédente comment rationaliser l'utilisation de ces traits. Mais qu'en est-il du coût de concaténation ? La couverture d'unités longues dans le script contribue sans doute à un accroissement de la longueur des segments sélectionnés lors de la synthèse. Néanmoins cette implication est à la fois imprécise et insuffisante : nous ne contrôlons pas vraiment l'impact du niveau de couverture sur le coût de concaténation et donc sur la qualité de synthèse.

Pour progresser sur cet aspect et mieux apprécier la qualité de couverture d'un script, nous proposons de nous inspirer du fonctionnement du système de sélection. Il est en particulier une caractéristique essentielle de la synthèse par corpus, commune à tous les systèmes : **à travers les considérations acoustiques du coût de concaténation, certaines classes phonétiques « sensibles » tendent à être naturellement préservées des concaténations.** Cette tendance générale des systèmes de sélection est entièrement justifiée par les invariants acoustiques qui caractérisent les différentes classes de phonèmes. Nous avons ainsi établi au paragraphe 2.4.3 la hiérarchie suivante, en commençant par les classes phonétiques qui supportent le mieux les concaténations :

1. occlusives sourdes [p], [t] et [k]
2. autres consonnes sourdes [f], [s] et [ʃ]
3. occlusives voisées [b], [d] et [g]
4. consonnes nasales et fricatives voisées [m], [n], [v], [z] et [ʒ]
5. liquides [l] et [ʁ]
6. semi-voyelles et schwa [j], [w], [ɥ] et [ə]
7. voyelles [a], [ɔ], [o], [ɛ], [e], [ø], [œ], [ã], [õ], [ɛ̃], [œ̃], [i], [y] et [u]

On peut, pour simplifier, distinguer deux groupes de phonèmes : d'une part les **phonèmes « robustes »** (ou « sécables ») qui supportent en général bien les concaténations, d'autre part les **phonèmes « fragiles »**, dont l'intégrité doit être protégée autant que possible. Le premier groupe comporte typiquement les consonnes non liquides, tandis que le second rassemble les voyelles, semi-voyelles, schwa et liquides. Avec notre système de synthèse et la voix d'Agnès, on observe effectivement une plus grande fréquence des concaténations sur les phonèmes robustes : 58% d'entre eux font l'objet d'une concaténation contre seulement 9% des voyelles [33].

Pour assurer une bonne protection des phonèmes fragiles, le système de sélection doit bien sûr disposer de séquences phonétiques adéquates dans la base de données. La figure 15 illustre

33. Nous avons effectué cette mesure sur un corpus de 147 phrases issues de domaines variés

ce mécanisme sur la phrase suivante : « *Et le plâtrier est revenu lundi matin.* » Les différentes unités de la transcription phonétique ont été plus ou moins foncées en fonction de leur tolérance ***a priori*** aux concaténations. Cette tolérance tient compte du contexte phonétique ; par exemple les liquides sont plus fragiles dans l'environnement direct d'un phonème sourd à cause de phénomènes importants de coarticulation, et notamment d'un risque de dévoisement. Pour la phrase donnée en exemple, il est probable que, du fait des critères de sélection, le système de sélection va essayer de sélectionner des segments de la base qui épargnent les phonèmes fragiles (en clair sur la figure). Cela signifie que les séquences [#eløp], [plat], [tʁijeɛʁøv], [vøn], [nulɛ̃d], [dim], [mat] et [tɛ̃#] doivent être présentes dans la base, et dans un contexte adéquat, sans quoi certaines concaténations apparaîtront inéluctablement sur des phonèmes fragiles.

Et le plâtrier est revenu lundi matin.

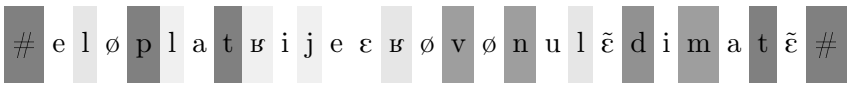


FIGURE 15 – Exemple d'énoncé textuel en entrée du système de synthèse vocale, avec sa transcription phonétique. Le niveau de gris de chaque phonème indique sa tolérance *a priori* aux concaténations, les phonèmes foncés étant les plus robustes.

Les séquences phonétiques qui s'étendent d'un phonème robuste au suivant, protégeant en leur sein une séquence plus ou moins longue de phonèmes fragiles, semblent donc d'intérêt pour l'étude de la couverture de la base. Elles permettent de prévenir les concaténations qui pourraient être gênantes lors de la synthèse. Elles permettent également de prendre en compte la plupart des phénomènes de coarticulation, en intégrant dans leur environnement phonétique les phonèmes qui sont le plus sujets à ce type de variations : voyelles, semi-voyelles et liquides.

6.2 Définition des sandwichs vocaliques

6.2.1 Caractérisation phonétique

Suivant les observations précédentes nous introduisons une nouvelle unité phonétique, appelée « **sandwich vocalique** », pour désigner **toute séquence de phonèmes fragiles entourée par deux phonèmes robustes**. Plus rigoureusement nous définissons le sandwich vocalique à l'aide de l'expression régulière suivante :

$$\boxed{C\ (\ W^*\ V\ W^*\ )^+\ C} \tag{10}$$

où :

C désigne l'ensemble des phonèmes robustes, c'est-à-dire ceux qui constituent un lieu privilégié de concaténation, ainsi que l'unité « silence »,
W désigne l'ensemble des phonèmes fragiles hormis les voyelles,
V désigne l'ensemble des voyelles,
le quantifieur * définit un groupe présent zéro, une ou plusieurs fois,
le quantifieur $^+$ définit un groupe présent une ou plusieurs fois.

Cette expression régulière offre aux voyelles un statut particulier, différent des autres phonèmes fragiles : tout sandwich vocalique contient au moins une voyelle. Cela permet de distinguer les sandwichs vocaliques des clusters consonantiques, ce qui n'aurait pas été le cas avec la

définition suivante, plus intuitive mais moins précise :

$$\boxed{C\ \overline{C}^{+}\ C} \tag{11}$$

où $\overline{C}$ est le complémentaire de C, c'est-à-dire l'ensemble des phonèmes fragiles (donc $W \cup V$). C'est néanmoins cette dernière définition que, dans un souci de simplicité, nous avons présentée dans notre première publication sur les sandwichs vocaliques [Cadic 09].

Naturellement le choix de l'ensemble C des phonèmes robustes conditionne beaucoup la composition des sandwichs vocaliques. Si les consonnes occlusives, fricatives et nasales, ainsi que l'unité silence, peuvent être rattachées à cet ensemble sans hésitation, le sort des liquides [l] et [ʁ] est moins tranché. Dans la suite nous les classerons de manière variable, en fonction du contexte phonétique et du niveau de protection envisagé (cf. section 6.3). La figure 16 montre le découpage en sandwichs vocaliques de l'exemple précédent, dans le cas où les liquides sont rattachées à l'ensemble W des consonnes fragiles.

On constate notamment sur cette figure que **les sandwichs vocaliques peuvent contenir plusieurs voyelles et même traverser les frontières de mots**, comme c'est le cas des sandwichs [#eløp] (« *Et le p...* »), [tʁijeɛʁøv] (« *... trier est rev...* ») et [nulɛ̃d] (« *... nu lund...* »).

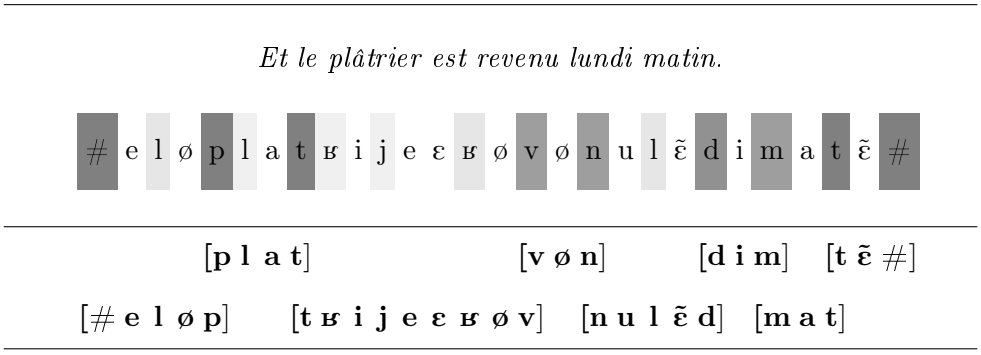


FIGURE 16 – Exemple de découpage en sandwichs vocaliques, dans le cas où les phonèmes liquides [l] et [ʁ] sont considérés comme fragiles et donc rattachés à W.

6.2.2 La notion de sandwich vocalique en contexte

Reprenons la phrase d'exemple précédente. Pour la synthèse vocale de cette phrase, la présence des sandwichs complets dans la base de données semble être, au premier abord, un gage de qualité. Elle devrait en effet permettre au système de sélection de diphones de placer les concaténations exclusivement sur des phonèmes considérés comme robustes. Mais en réalité cela n'est possible que si les sandwichs présents dans la base sont porteurs d'une prosodie adaptée au contexte. A titre d'exemple, voici trois contextes d'apparition très différents d'un même sandwich [plat] :

« *Et le plâtrier est revenu lundi matin.* » [#elø**plat**ʁijeɛʁøvønulɛ̃dimatɛ̃#]
« *La terre est plate.* » [#latɛʁe**plat**#]
« *Ce plat te plaît-il ?* » [#sø**plat**øplɛtil#]

Les sandwichs [plat] apparaissant dans ces trois contextes ne sont aucunement interchangeables. Et en cas d'inadéquation entre le sandwich-cible et le(s) candidat(s) présent(s) dans la base, le coût-cible risque de l'emporter sur le coût de concaténation et ainsi de forcer des concaténations sur les phonèmes fragiles.

Pour éviter cela, les contextes linguistique et prosodique des sandwichs doivent être pris en compte dès le départ, lors de la constitution du script de lecture. On parle alors de « **sandwich**

vocalique en contexte ». Conformément aux conclusions de la section 5.1, seules les informations contextuelles des noyaux vocaliques sont utilisées pour cet enrichissement. La figure 17 reprend notre exemple en ajoutant au-dessus de chaque voyelle son marqueur contextuel (illustré ici par une simple étiquette numérique). Grâce à notre travail de rationalisation des traits contextuels (section 5.2), plusieurs niveaux de précision peuvent être retenus pour le jeu de marqueurs. Nous verrons plus bas les différents jeux de marqueurs utilisés dans nos travaux.

Dans la suite nous utiliserons presque toujours des sandwichs vocaliques en contexte (sauf mention contraire), même si pour simplifier nous parlerons souvent de « sandwich » ou « sandwich vocalique ».

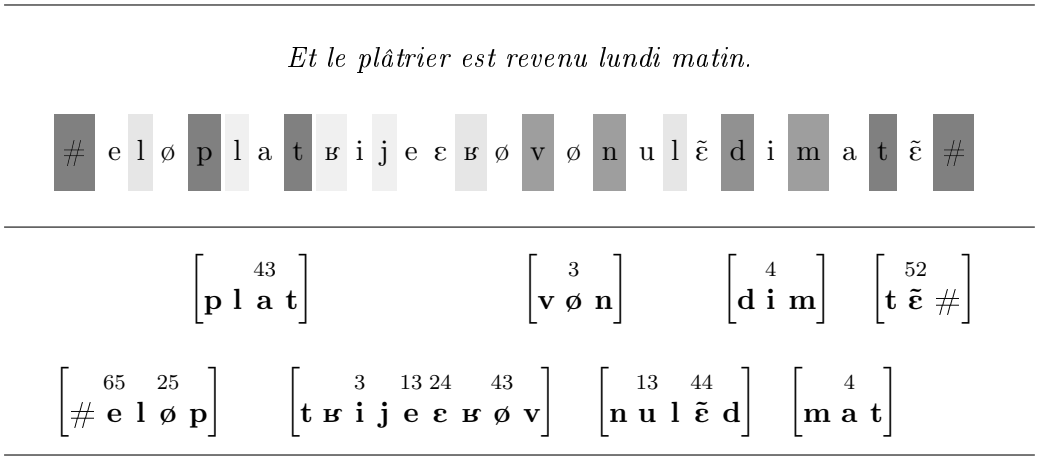


FIGURE 17 – Sandwichs vocaliques enrichis d'informations contextuelles, d'ordre linguistique et prosodique. Les marqueurs contextuels ne portent que sur les noyaux vocaliques.

6.2.3 Position par rapport à l'état de l'art

L'idée d'introduire des unités phonétiques plus longues que les diphones, tout et tenant compte des spécificités de la synthèse par concaténation, n'est pas nouvelle.

Dès 1985, Laferrière et al ont utilisé pour leur système de synthèse des segments plus larges que les diphones, appelés « polysons » [Laferrière 85]. Ces unités sont des demi-syllabes [34] du français, choisies pour préserver certaines consonnes vocaliques des problèmes de segmentation ou de concaténation. Plus précisément les phonèmes protégés sont des semi-voyelles ou liquides, dans des contextes particulièrement critiques (par exemple au contact d'une autre consonne). L'inventaire des diphones a ainsi été complété par 2 630 « polysons » enregistrés dans un contexte phonologique neutre.

Une approche comparable a été proposée pour l'allemand [Breuer 04]. Dans ce travail, on traite comme blocs unitaires certaines séquences phonétiques présentant des dynamiques articulatoires complexes. Ceci conduit à l'introduction d'unités appelées « phoxsy » (PHOne eXtensions for SYnthesis), dont la plupart sont de type consonne-voyelle. Elles sont utilisées dans le système de synthèse BOSS [Klabbers 01a]. Celui-ci reposant sur les phones comme unités de base (ou les mots lorsqu'ils sont couverts), les phoxsy sont nécessaires pour améliorer la prise en compte des phénomènes transitoires.

De par leurs définitions, les unités polyson et phoxsy présentent des différences majeures avec les sandwichs vocaliques. Elles sont plus courtes et généralement délimitées par des zones

34. Partie de syllabe composée soit de l'attaque et du noyau, soit du noyau et de la coda.

de stabilité spectrale comme les voyelles. Les frontières de sandwichs n'interviennent quant à elles jamais sur les voyelles ; c'est d'ailleurs l'objet principal des sandwichs que de protéger ces phonèmes qui, malgré leur grande stabilité spectrale intra-occurence (voir 2.4.3), supportent très mal les concaténations. A ce titre le support d'un sandwich peut tout à fait rassembler plusieurs voyelles et traverser des frontières de mots.

D'autre part les polysons et phoxsy jouent un rôle principalement acoustique : ces unités sont enregistrées puis stockées dans la base de données en vue d'une utilisation unitaire dans l'étape de concaténation. Ce rôle est très différent de celui réservé aux sandwichs vocaliques. Nous les destinons en effet aux études amonts, comme la constitution d'un script de lecture ou plus généralement la préparation d'une base de données, où ils permettent d'anticiper la qualité de synthèse finale. Mais notre système de synthèse (hauts-niveaux comme bas-niveaux) conserve intégralement un fonctionnement à base de diphones.

Une autre unité acoustique, baptisée en anglais « syllable-like unit », présente des similitudes notables avec nos sandwichs vocaliques. Introduite à l'occasion de travaux en reconnaissance de parole [Hu 96], elle a été utilisée dans [Thomas 06] pour la synthèse vocale du Tamil, une langue régionale indienne. La définition d'une « syllable-like unit » repose exclusivement sur des considérations acoustiques : il s'agit d'une portion de signal délimitée par deux régions d'énergie minimale. La segmentation du signal en « syllable-like units » est effectuée à l'aide d'un algorithme basé sur les délais de groupe [Nagarajan 03]. On peut noter une certaine analogie entre ces régions faiblement énergétiques et nos phonèmes robustes ; mais la correspondance n'est pas systématique et la définition des « syllable-like units » ne repose sur aucun critère phonétique ni notion linguistique. Il s'agit de segments purement acoustiques qui, contrairement aux sandwichs vocaliques, ne peuvent pas évoluer dans un contexte symbolique. Les « syllable-like units » sont utilisées dans les bas-niveaux pour contraindre le processus de sélection (comme nous pourrions probablement le faire en modifiant le coût de concaténation), mais ne sont pas adaptées à des études linguistiques plus générales comme la constitution d'un script de lecture.

6.3 Variantes

La définition des sandwichs vocaliques est assez souple : elle ne précise ni le contenu de l'ensemble C des phonèmes robustes, ni le jeu de paramètres contextuels. De manière avantageuse, ces ensembles seront choisis en fonction du contexte applicatif. Ainsi, pour la création d'un petit script de lecture de quelques milliers de mots[35], on envisagera une couverture simplifiée des sandwichs vocaliques : avec un ensemble C maximal et un jeu réduit de traits contextuels, la dispersion des unités est limitée et leur couverture facilitée. A l'inverse des choix plus ambitieux pourront être effectués pour la création d'un script de lecture plus long.

Plus précisément, nous utiliserons dans la suite deux classifications différentes pour le groupe des liquides ([l] et [ʁ]), ce qui engendre deux variantes phonétiques des sandwichs vocaliques. **Dans le premier cas nous considérerons les liquides comme des phonèmes fragiles** ; elles seront donc exclues de C pour être rattachées à W, aux côtés des semi-voyelles et du schwa. C'est la solution de luxe, qui permet notamment de prévenir les phénomènes de coarticulation qui impactent la production des liquides, voire les traversent. **Dans le second cas nous utiliserons une classification dépendante du contexte phonétique, dans laquelle les liquides seront qualifiées de « semi-robustes »** : elles seront considérées comme robustes, sauf dans l'environnement direct d'un phonème sourd. Ce contexte suscite, entre autres effets de coarticulation, un risque incontrôlé de dévoisement des liquides et pourrait, si ces dernières n'étaient pas protégées, provoquer des segmentations et concaténations

35. soit moins d'une heure d'enregistrement, pour quelques minutes de parole utile

hasardeuses. Notons que cette dépendance de C au contexte phonétique constitue une légère entorse à notre définition (10). La figure 18 illustre l'impact des deux classifications sur le découpage en sandwichs vocaliques.

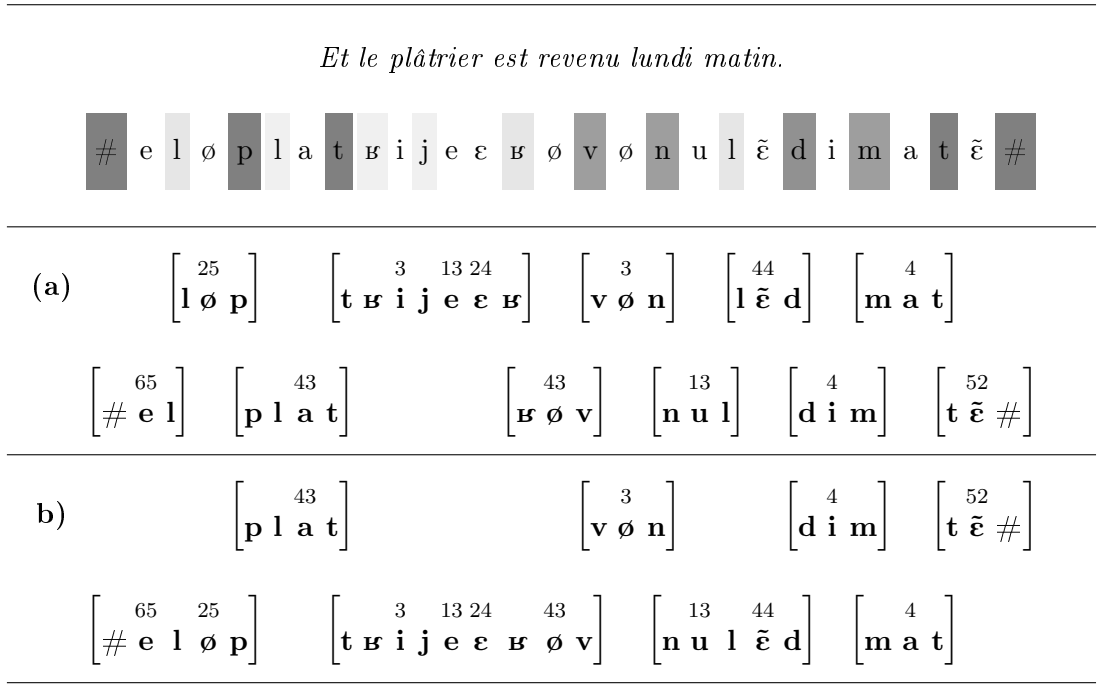


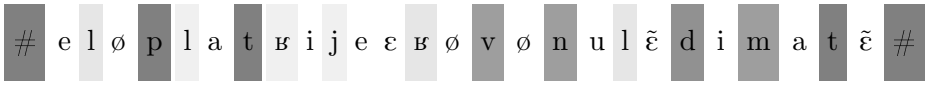
FIGURE 18 – Deux découpages en sandwichs vocaliques, suivant que les liquides sont classées (a) parmi les phonèmes robustes sauf dans l'environnement direct d'un phonème sourd (liquides semi-robustes), ou bien (b) parmi les phonèmes fragiles.

L'autre facteur permettant de moduler la complexité (et donc la dispersion) des sandwichs vocaliques est le choix du jeu de paramètres utilisé pour caractériser le contexte des voyelles. En retenant les noeuds les plus discriminants des classifications arborescentes obtenues en section 5.2, le niveau de précision contextuel peut être adapté aux attentes qualitatives et à la taille envisagée pour le script de lecture. Nous utiliserons dans la suite **trois niveaux de précision contextuelle** :

- un premier niveau sans aucune information contextuelle (unités hors contexte).
- un second niveau qui vise à décrire uniquement les situations contextuelles les plus significatives sur le plan acoustique, au moyen de 4 classes seulement (voir figure 14 page 77)
- un troisième niveau qui couvre, avec 13 classes, la plupart des contextes pertinents pour le style « lecture neutre ».

Enfin il existe certaines configurations où une couverture linguistique de très haute qualité est envisageable et où, même en retenant 13 contextes et en considérant [l] et [ʁ] comme des phonèmes fragiles, les sandwichs vocaliques se révèlent trop simples et insuffisamment dispersés. C'est par exemple le cas lorsqu'il s'agit de préparer un script conséquent (plusieurs dizaines de milliers de mots) pour l'enregistrement d'une voix en domaine restreint. On peut alors s'intéresser aux **bigrammes de sandwichs**, c'est-à-dire aux séquences de deux sandwichs consécutifs. De telles unités permettent la prise en compte de phénomènes linguistiques et prosodiques plus globaux, tout en garantissant la protection des phonèmes fragiles et donc une bonne qualité segmentale. La figure 19 présente le découpage d'une phrase en bigrammes. Pour des raisons pratiques les début et fin de phrase doivent être matérialisés par des sandwichs virtuels, que nous notons respectivement <*deb*> et <*fin*>.

Et le plâtrier est revenu lundi matin.



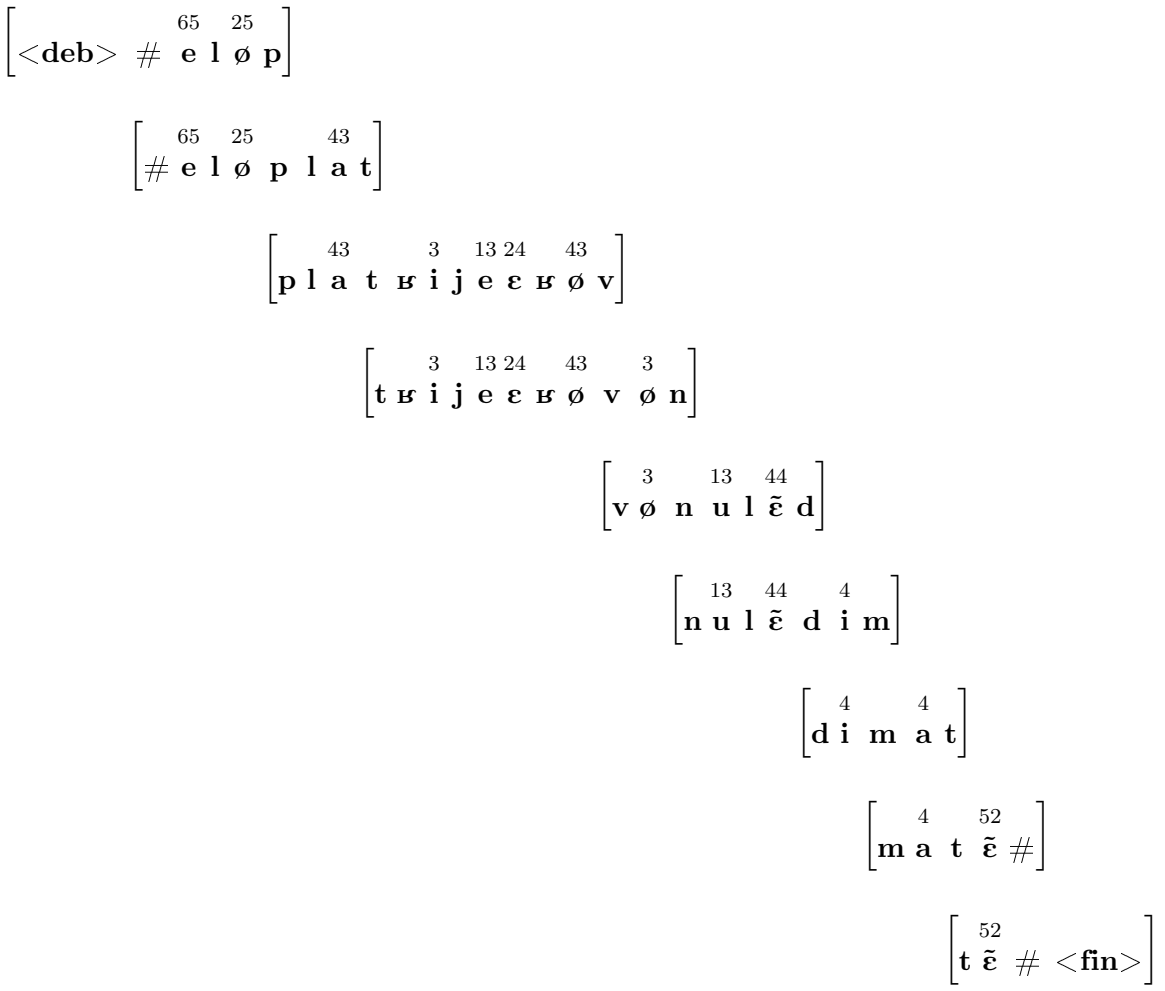


Figure 19 – Découpage en bigrammes de sandwichs vocaliques. Chaque séquence regroupe deux sandwichs successifs, ainsi qu'un éventuel cluster consonantique central.

6.4 Le traitement des clusters consonantiques

L'introduction des sandwichs vocaliques est liée au choix du diphone comme unité de base pour la sélection, car nous supposons que les concaténations sont opérées sur les milieux de phonèmes. Cela implique entre autres que les phonèmes limitrophes d'un sandwich vocalique n'appartiennent que pour moitié à ce sandwich. Ainsi une chaîne phonétique quelconque bordée par deux phonèmes silences peut toujours être décomposée en une alternance de sandwichs vocaliques et de séquences de la forme CW^*C. Ces dernières séquences sont en fait des **clusters consonantiques « robustes »**, c'est-à-dire bordés par des phonèmes robustes. Les phonèmes extrémaux d'un tel cluster consonantique empiètent pour moitié sur les sandwichs voisins. La figure 20 illustre le découpage d'une phrase complète comportant ce type de clusters consonantiques. On y reporte également les bigrammes de sandwichs, dont on constate qu'ils absorbent les clusters consonantiques situés entre leurs deux sandwichs constitutifs.

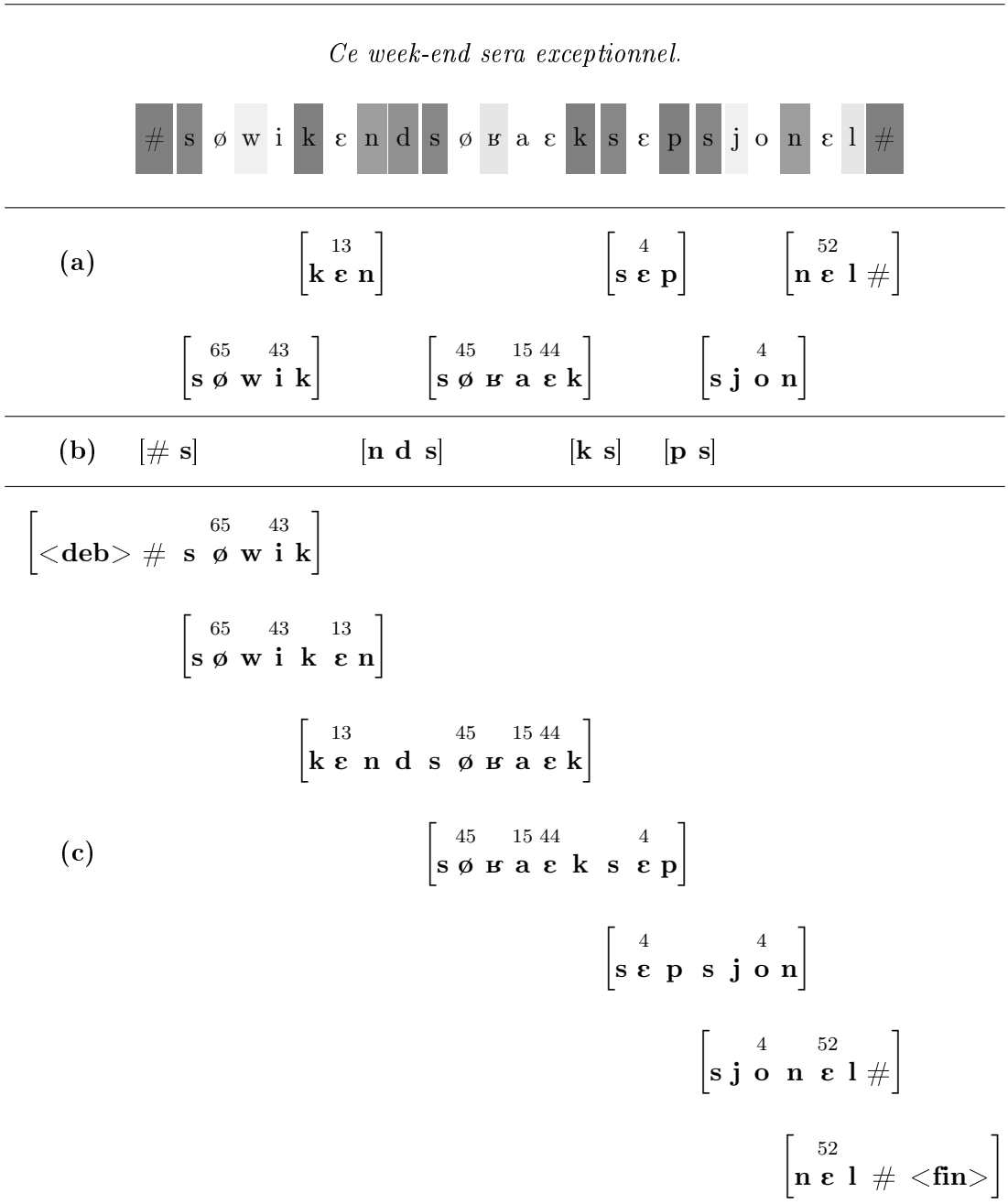


FIGURE 20 – Découpage d'une phrase comportant des clusters consonantiques robustes : (a) sandwichs vocaliques, (b) clusters consonantiques robustes, (c) bigrammes de sandwichs.

Dans notre étude, ces clusters consonantiques robustes feront l'objet d'une attention moindre pour les trois raisons suivantes :

- Ils sont peu nombreux dans la langue française et donc faciles à couvrir : en considérant [l] et [ʁ] comme des phonèmes fragiles, 90% des occurrences[36] sont couvertes avec seulement 100 clusters. Tout script raisonnablement diversifié apporte une bonne couverture de ces clusters sans qu'il soit nécessaire de s'en préoccuper.
- Ne comportant pas de voyelle, ils ne sont pas ou peu porteurs de prosodie.
- Ils peuvent aisément être reconstruits à partir d'unités plus petites. En effet, selon notre hiérarchie phonétique établie plus haut, ils supportent généralement bien les concaténa-

36. relevées dans le corpus présenté en section 7.1

tions et ne nécessitent pas de protection particulière.

7 Evaluation objective

7.1 Constitution d'un corpus de référence

Nous avons vu dans la partie précédente (page 49) que la constitution d'un corpus textuel de référence est une étape importante du processus de création d'un script de lecture. Le corpus de référence permet d'observer les statistiques de différents évènements de la langue et ainsi d'établir les priorités de couverture, par exemple à travers une pondération fréquentielle du critère d'optimisation. Il peut également faire office de corpus de pioche pour la sélection des phrases qui sont ajoutées au script de lecture (voir page 51). Dans tous les cas, le niveau de correction de ce corpus a un impact direct sur la qualité du script final (voir page 52).

Si la constitution d'un corpus « universel » représentatif de la langue est hors de portée [Van Santen 97a] (voir page 50), notre objectif est de **couvrir de manière significative et équilibrée la plupart des applications envisagées pour nos voix de synthèse**.

Nous avons ainsi constitué un corpus de 2 500 000 mots (166 000 lignes, 250 000 phrases, 540 000 groupes de souffle) avec un niveau élevé de vérification. Pour cela plusieurs sources ont été assemblées. Nous les détaillons ici par ordre décroissant de taille :

1. **SMS Louvain (692 000 mots)**
   Cet ensemble de 30 000 SMS a été collecté par l'Université de Louvain-la-Neuve dans le cadre de l'opération « *Faites don de vos SMS à la science* » puis retranscrit en français standard [Fairon 06]. Nous avons, de notre côté, effectué un travail de relecture afin de s'assurer de la bonne interprétation du texte par les hauts-niveaux de notre système ; en particulier certaines erreurs récurrentes de transcription phonétique nous ont amenés à corriger le texte ou les hauts-niveaux.
2. **Le Monde (680 000 mots)**
   Extrait de l'année 2002.
3. **Livres et pièces de théâtre (240 000 mots)**
   Il s'agit de documents anciens ou contemporains, libres de droit, récoltés en grande partie sur internet. Le travail de supervision a essentiellement porté sur le formatage : retrait des indications typographiques (par ex. des tirets pour séparer deux parties), retrait de passages superflus ou très redondants (par ex. « *Acte III, scène 1* »), etc.
4. **Phrases de service (210 000 mots)**
   Nous avons rassemblé de nombreux scripts de services vocaux téléphoniques, dans des domaines variés comme la bourse, l'automobile, ou encore le dépannage internet. L'univers des télécommunications y est fortement représenté. Nous avons également inclus des modes d'emploi de services multimédia.
5. **SMS Aix (150 000 mots)**
   10 000 SMS, compilés et retranscrits par le laboratoire DELIC de l'Université de Provence [Hocq 06]. Nous l'avons révisé de la même manière que « SMS Louvain ».
6. **Flux RSS (134 000 mots)**
   Plusieurs flux RSS captés sur internet en 2007-2008. Ils sont principalement composés de nouvelles politiques, économiques et sportives.
7. **Sous-titres de séries TV (128 000 mots)**
   Un traitement automatique d'erreurs récurrentes ainsi qu'une relecture sommaire ont été effectués.

8. **Phrases équilibrées (97 000 mots)**
   Il s'agit de plusieurs groupes de phrases, utilisés par le passé pour différents travaux de recherche en synthèse vocale, comme par exemple l'apprentissage de modèles prosodiques.
9. **Recettes de cuisine (63 000 mots)**
10. **Tchat 18-25 ans (50 000 mots)**
    Extrait du « corpus de français tchaté » constitué par l'Université Grenoble I [Falaise 05], que nous avons entièrement retranscrit.
11. **Chroniques radio (50 000 mots)**
    Transcriptions textuelles de chroniques radio portant sur les nouvelles technologies.
12. **Traces de tchat (18 000 mots)**
    Traces du système de messagerie instantanée Microsoft MSN, collectées auprès d'un panel de collaborateurs et connaissances d'âges variés. Nous avons entièrement retranscrit ces traces en français standard.

Nous pouvons constater qu'une large part de notre corpus de référence est consacrée au langage parlé : SMS, sous-titres de séries TV et traces de tchat représentent plus de 40% des données textuelles collectées. Ceci reflète notre volonté de nous écarter des usages traditionnels de la synthèse vocale pour aborder de nouveaux champs applicatifs, plus proches de nos modes de communication réels. Nous pensons par exemple à des applications grand public basées sur l'échange de messages ludiques, vocalisés avec des voix célèbres et/ou caricaturales. Malheureusement ces sources textuelles sont les plus rares et les plus onéreuses : d'une part elles posent des problèmes de droits (d'où les difficultés pour s'en procurer), d'autre part leur niveau de correction orthographique est généralement très bas et implique le recours à de coûteux traitements correctifs.

7.2 Distributions

En s'appuyant sur notre corpus de référence, nous pouvons désormais observer les propriétés statistiques des sandwichs vocaliques : longueur moyenne, nombre d'unités distinctes, etc. Une telle étude est d'autant plus importante que **nous allons, pour l'optimisation des scripts de lecture, adopter une « approche en fréquence »** (voir pages 49 à 50). Cette approche va consister à couvrir en priorité les sandwichs vocaliques les plus fréquents, sous prétexte qu'ils seront statistiquement plus utiles pour la synthèse finale. Nous nous intéressons donc au taux de couverture du corpus de référence, ce qui revient à considérer un taux de couverture pondéré par les fréquences d'apparition dans le corpus de référence.

Van Santen conteste cette approche fréquentielle :

> *« There is a risk, however, in relying too much on frequencies observed in a particular corpus, because of the instability of frequency distributions across text corpora. (...) Tying system construction too closely to a particular corpus runs the risk of neglecting units that may prove unexpectedly frequent in new corpora. »* [Van Santen 97b]

Dans notre cas le problème se pose un peu différemment. D'une part, en nous focalisant sur les unités vraiment fréquentes, nous n'exploiterons pas les statistiques des unités rares, qui sont le plus soumises à l'aléa du choix du corpus. D'autre part, comme nous l'avons expliqué un peu plus haut, notre corpus de référence ne se veut pas universel, mais représentatif des domaines spécifiques dans lesquels nous envisageons d'exploiter les voix de synthèse ; le risque de se focaliser sur des unités superflues ou de manquer des unités très importantes est donc

faible. Quoiqu'il en soit nous n'avons pas vraiment le choix : comme nous souhaitons créer des scripts courts et denses, seule une approche en fréquence peut nous offrir l'efficacité escomptée.

La table 6 rapporte quelques grandeurs relatives aux distributions de plusieurs variantes de sandwichs dans notre corpus de référence ; ces valeurs peuvent être comparées aux unités traditionnelles grâce à la table 7. Pour chaque unité, les grandeurs mesurées sont :

- **Longueur moyenne** : nombre moyen de phonèmes contenus dans une unité
- **Unités distinctes** : nombre total d'unités distinctes rencontrées dans le corpus de référence
- **Couverture à 80%** : nombre minimum d'unités nécessaires pour couvrir 80% des occurences relevées dans le corpus
- **Densité** [37] : rapport entre le nombre total d'unités (pas forcément distinctes) et le nombre total de phonèmes dans le corpus
- **Taux de chevauchement**[37] : chevauchement phonétique moyen entre deux unités successives, en proportion de la première unité. Par exemple dans la séquence [#bõvwaja3#] (« *Bon voyage* »), les sandwichs [bõv] et [vwaja3] ont un phonème en commun, donc un taux de chevauchement de 1/3. De même deux diphones successifs ont un taux de chevauchement de 1/2. Même si ces unités sont acoustiquement disjointes (car les frontières interviennent sur les milieux de phonèmes), le taux de chevauchement phonétique est un indicateur de la corrélation qui existe entre les supports phonétiques des unités successives.
- **Entropie** [38] : mesure de la dispersion des unités, en bits/unité :

$$E = -\sum_{i=1}^{N} f_i \cdot log_2 f_i$$

où f_i désigne la fréquence de la i^{ème} unité et N le nombre total d'unités distinctes. L'entropie E est maximale pour une distribution uniforme.

Du fait de la loi de Zipf-Mandelbrot (voir page 45), les indicateurs « unités distinctes », « couverture à 80% » et « entropie » nous livrent des informations assez complètes sur la distribution des différentes unités. En particulier, de petites valeurs de ces indicateurs font référence à une distribution compacte et donc une unité facile à couvrir. A l'inverse, ces indicateurs augmentent avec la complexité de l'unité, en rapport avec sa longueur moyenne et la précision de son information contextuelle (limitée ici aux noyaux vocaliques pour toutes les unités).

On constate sur la table 6 de gros écarts statistiques entre la version la plus simple des sandwichs vocaliques (hors contexte, [l] et [ʁ] semi-robustes) et la version la plus complexe (13 contextes possibles pour les voyelles, [l] et [ʁ] fragiles). La première variante est plutôt destinée à la définition de critères minimalistes pour la création et l'analyse de scripts très courts, tandis que la seconde englobe des considérations de haute qualité dédiées à des scripts plus longs. C'est d'ailleurs pour pouvoir s'adapter à différents niveaux de besoins que nous avons introduit plusieurs variantes de sandwichs. A titre indicatif, les fonctions de répartition de ces deux variantes sont présentées sur la figure 21. Indirectement, les fonctions de répartition décrivent la couverture maximale que l'on peut attendre d'un script de lecture de taille donnée, ou encore la taille minimale du script de lecture nécessaire pour atteindre un certain taux de couverture. Par

37. Suivant le rôle accordé aux silences, la mesure de cet indicateur peut s'écarter sensiblement des valeurs intuitives.

38. Dans les tables 6 et 7, l'entropie donne des indications sur les statistiques de 1er ordre de différentes unités (n-grammes de phones ou de sandwichs). Elle augmente avec la complexité des unités et en particulier avec n. Cette mesure diffère de celle rencontrée dans d'autres contextes applicatifs, comme par exemple en reconnaissance vocale, où l'on mentionne souvent l'entropie de modèles d'ordre n pour quantifier la (non-)prédictibilité d'une unité à partir des $n-1$ unités précédentes. Dans ce cadre très différent, l'entropie diminue avec l'ordre n des modèles.

Unité			Longueur moyenne	Unités distinctes	Couverture à 80%	Densité	Taux de chevauchement	Entropie
Type	**Classement des liquides**	**Nombre de symboles contextuels**						
sandwichs	semi-robustes	0 contexte	3.18	40453	1371	0.38	0.17	11.17
sandwichs	semi-robustes	4 contextes	3.18	54786	2332	0.38	0.17	11.87
sandwichs	semi-robustes	13 contextes	3.18	93173	5427	0.38	0.17	13.01
sandwichs	fragiles	0 contexte	3.66	118748	3038	0.33	0.18	12.27
sandwichs	fragiles	4 contextes	3.66	144326	5098	0.33	0.18	12.88
sandwichs	fragiles	13 contextes	3.66	203307	10463	0.33	0.18	13.82
bigrammes	semi-robustes	0 contexte	5.16	559313	71754	0.44	0.55	15.84
bigrammes	semi-robustes	4 contextes	5.16	662568	107342	0.44	0.55	16.37
bigrammes	semi-robustes	13 contextes	5.16	859427	183283	0.44	0.55	16.90
bigrammes	fragiles	0 contexte	5.80	794552	160285	0.39	0.56	16.49
bigrammes	fragiles	4 contextes	5.80	889847	220220	0.39	0.56	16.94
bigrammes	fragiles	13 contextes	5.80	1032196	301461	0.39	0.56	17.32

TABLE 6 – Quelques grandeurs relatives aux distributions de différentes variantes de sandwichs et de leurs bigrammes.

	Unité	Longueur moyenne	Unités distinctes	Couverture à 80%	Densité	Taux de chevauchement	Entropie
Unités en contexte (13 symboles)	phones, 1-grammes	1.00	215	46	1.00	0.00	6.02
	phones, 2-grammes	2.00	10953	1247	1.00	0.50	10.79
	phones, 3-grammes	3.00	163367	14887	0.93	0.64	14.32
	phones, 4-grammes	4.00	756268	105516	0.87	0.71	16.84
	phones, 5-grammes	5.00	1703332	445042	0.81	0.75	18.59
	phones, 6-grammes	6.00	2613222	1173168	0.76	0.78	19.77
	phones, 7-grammes	7.00	3309110	1970141	0.70	0.80	20.55
	phones, 8-grammes	8.00	3745395	2502026	0.65	0.81	21.05
	syllabes, 1-grammes	2.21	31315	1324	0.45	0.00	10.80
	syllabes, 2-grammes	4.38	467062	70545	0.39	0.42	16.05
	mots, 1-grammes	3.31	181276	7920	0.30	0.00	12.23
	mots, 2-grammes	6.42	723715	300345	0.24	0.35	17.24
	sandwichs, 1-grammes	3.18	93173	5427	0.38	0.17	13.01
	sandwichs, 2-grammes	5.16	859427	183283	0.44	0.55	16.90
Unités hors contexte	phones, 1-grammes	1.00	35	17	1.00	0.00	4.76
	phones, 2-grammes	2.00	1162	273	1.00	0.50	8.74
	phones, 3-grammes	3.00	24149	2950	0.93	0.64	12.14
	phones, 4-grammes	4.00	215682	23363	0.87	0.71	14.96
	phones, 5-grammes	5.00	835385	134720	0.81	0.75	17.14
	phones, 6-grammes	6.00	1746807	514494	0.76	0.78	18.72
	phones, 7-grammes	7.00	2583790	1244821	0.70	0.80	19.80
	phones, 8-grammes	8.00	3182459	1939090	0.65	0.81	20.52
	syllabes, 1-grammes	2.21	9666	255	0.45	0.00	8.68
	syllabes, 2-grammes	4.38	227777	18380	0.39	0.42	14.40
	mots, 1-grammes	3.31	77893	1864	0.30	0.00	10.51
	mots, 2-grammes	6.42	496641	120053	0.24	0.35	16.02
	sandwichs, 1-grammes	3.18	40453	1371	0.38	0.17	11.17
	sandwichs, 2-grammes	5.16	559313	71754	0.44	0.55	15.84

TABLE 7 – Quelques grandeurs relatives aux distributions de différentes unités. Pour les sandwichs, les liquides sont ici considérées comme semi-robustes. Pour les unités en contexte, les symboles ne portent que sur les noyaux vocaliques.

exemple dans le cas `LRsemirobustes_0contexte` (respectivement `LRfragiles_13contextes`), au moins 1 371 sandwichs (respectivement 10 463) sont nécessaires pour offrir une couverture de 80% des occurrences observées dans le corpus, comme indiqué dans la table 6 et illustré sur la figure 21 par les lignes en pointillés. Ceci correspondrait à des scripts de lecture parfaitement denses, contenant exclusivement les sandwichs les plus fréquents et sans aucune redondance. Dans la réalité une certaine redondance est inéluctable, tout comme la présence de sandwichs plus rares. **Les fonctions de répartition constituent donc un optimum de couverture**, inaccessible en pratique mais que nous essayerons d'approcher au mieux.

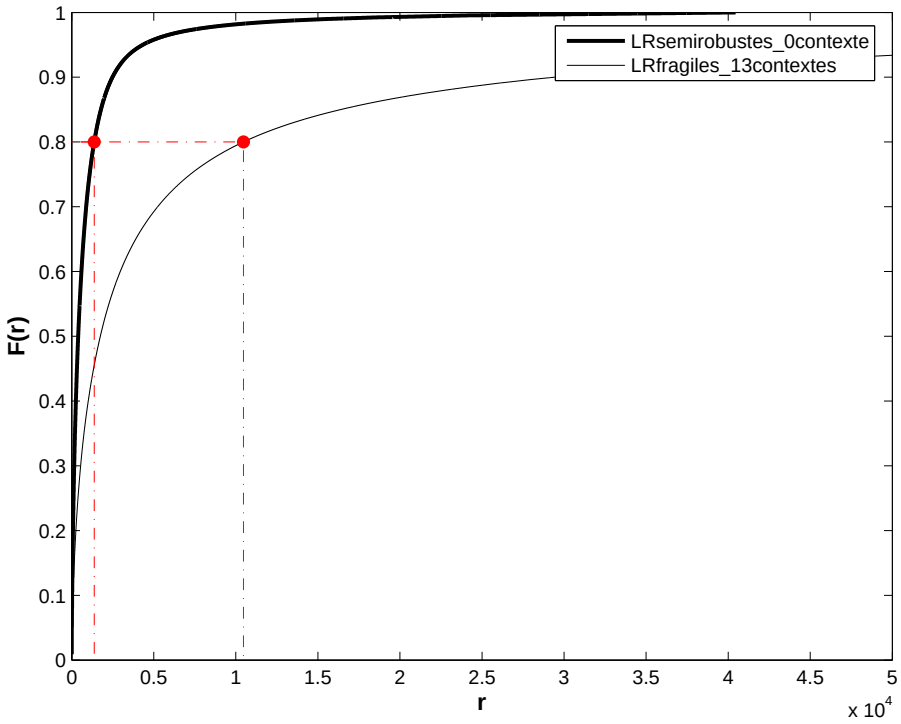


FIGURE 21 – Fonction de répartition de deux variantes de sandwichs. $F(r)$ est la proportion du corpus de référence qui est couverte de manière cumulée par les sandwichs de rang 1 à r (classés par ordre inverse de fréquence).

Lorsque les liquides sont considérées comme semi-robustes, les 1-grammes et 2-grammes de sandwichs ont une longueur moyenne de 3.18 et 5.16 phonèmes respectivement. La table 7 indique que, d'après les indicateurs « couverture à 80% » et « entropie », leurs distributions sont plus compactes que les n-grammes de phones[39] de longueur équivalente. Ils semblent donc relativement faciles à couvrir, tout en intégrant avantageusement des phénomènes vocaux d'assez long terme. L'indicateur « densité » joue cependant en leur défaveur : à longueur à peu près équivalente, ils occupent un espace phonétique plus important que les n-grammes de phones, donc avec un impact plus élevé sur la taille du script. Les n-grammes de phones se caractérisent en effet par des densités proches de 1, chaque phonème du corpus initiant un nouveau n-gramme[40]. En contrepartie les n-grammes de phones présentent des taux de chevauchement importants et chacun d'entre eux impose une contrainte phonétique sur ses

39. En théorie nous devrions plutôt parler de « n-grammes de phonèmes » ou de « n-grammes d'allophones » car il ne s'agit pas, à ce stade, de réalisations acoustiques. Mais dans un souci d'homogénéité avec les diphones et triphones, pour lesquels nous avons choisi d'unifier les notations (voir page 26), nous parlerons de « n-grammes de phones ».

40. Sauf aux extrémités des phrases où des effets de bord interviennent. Ceux-ci sont plus importants pour les n-grammes longs, ce qui explique que leur densité diminue avec n.

$n-1$ successeurs. Ce n'est pas le cas des sandwichs qui, avec un taux de chevauchement de seulement 0.17, offrent une certaine souplesse pour l'enchaînement des unités. La sélection ou la construction de phrases denses, c'est-à-dire apportant un maximum d'unités à couvrir avec un minimum de redondance, est donc facilitée.

Pour finir, il est important de rappeler que notre définition des « taux de couverture » ne tient pas compte de la longueur des sandwichs. Par conséquent, les sandwichs [kʁwajɛk] et [pɛʒ], qui apparaissent tous les deux 50 fois dans le corpus de référence (hors considérations contextuelles), ont le même impact sur le taux de couverture. Pourtant le premier offre la protection de 5 phonèmes fragiles, contre un seul pour le second. Une manière de compenser cela pourrait être de pondérer chaque sandwich par le nombre de phonèmes fragiles qu'il englobe. On obtient alors un nouveau taux de couverture, ou plutôt un « taux de protection des phonèmes fragiles ». Celui-ci n'est autre que la proportion des phonèmes fragiles du corpus de référence qui appartiennent à un sandwich couvert. La figure 22 montre, dans le cas `LRsemirobustes_0contexte`, l'impact de cette pondération sur les taux mesurés. Dans la suite nous n'utiliserons pas cette version pondérée, par commodité mais aussi parce qu'au fil des expériences nous avons constaté qu'elle influait très peu sur la création des scripts.

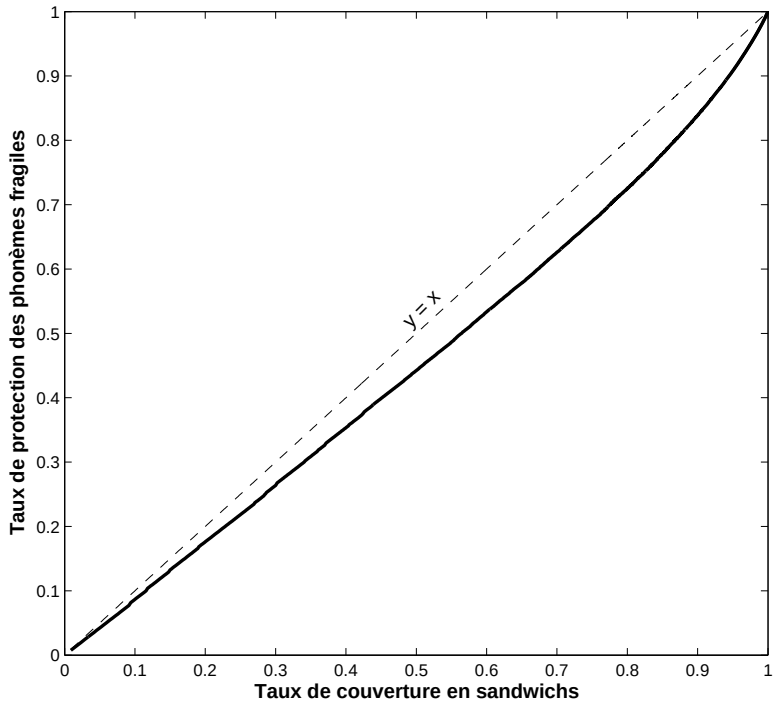


FIGURE 22 – Différence entre le taux de couverture en sandwichs et le taux de protection des phonèmes fragiles, ici pour le cas `LRsemirobustes_0contexte`. Les courbes ont été obtenues en suivant la fonction de répartition (c'est-à-dire en ajoutant les sandwichs par ordre inverse de fréquence). La droite « $y = x$ » est tracée en pointillés.

7.3 Corrélation au coût de sélection

Les observations précédentes nous ont offert une vision globale des distributions des différentes unités. Nous avons ainsi pu constater la relative accessibilité d'une couverture en sandwichs vocaliques. Avant d'exploiter réellement ces unités pour la création de scripts de lec-

ture, nous en présentons ici une évaluation objective dans le contexte spécifique de la synthèse par sélection. Cette évaluation préliminaire sera complétée dans la partie IV d'une évaluation subjective menée *a posteriori*, c'est-à-dire après avoir constitué plusieurs scripts de lecture et enregistré de nombreux locuteurs.

Pour mieux prendre en compte la perception humaine du signal de synthèse, les sandwichs vocaliques ont été introduits suivant le modèle du coût de sélection. La présente expérience consiste à vérifier que les sandwichs vocaliques reflètent effectivement mieux le coût de sélection que les autres unités. Pour cela nous proposons de mesurer la **corrélation entre le coût de sélection et le taux de couverture des différentes unités**. Plus précisément nous calculons, pour chaque type d'unité et pour chacune des 250 000 phrases du corpus de référence :

(a) la **proportion des unités de la phrase qui sont disponibles dans la base** :

$$\frac{card(P \cap B)}{card(P)}$$

où P désigne la liste des unités constituant la phrase et B l'ensemble des unités couvertes dans la base de données de la voix de synthèse (mais qui ne sont pas forcément sélectionnées par le système de synthèse)

(b) le **coût de sélection** obtenu pour la synthèse de cette phrase (somme des coût-cible et coût de concaténation).

Puis, pour chaque type d'unité, la corrélation entre (a) et (b) est mesurée sur l'ensemble du corpus de référence.

L'expérience a été répétée sur trois voix de synthèse. La première est celle d'Agnès, voix féminine dont la base comporte 258 519 diphones. La seconde est un simple sous-ensemble d'Agnès, approximativement le premier quart, pour un total de 66 532 diphones. Il est important de noter que, lors de sa création, la base d'Agnès avait fait l'objet d'une optimisation suivant des considérations de couverture en diphones et triphones. Même si, dans le cadre de cette évaluation objective, l'hypothèse d'un biais consécutif à cette phase d'optimisation est largement contestable, elle ne peut être totalement écartée. Pour cette raison nous avons utilisé une troisième base, constituée de rushes de la voix de Chirac (438 603 diphones). L'intérêt majeur de cette base est que son contenu, collecté sur internet (voir section 3.4), n'a fait l'objet d'aucune optimisation. On peut donc affirmer qu'elle n'introduit aucun biais en faveur de l'une ou l'autre unité.

Les résultats sont présentés en table 8. Pour chaque unité, nous avons reporté la corrélation observée sur chacune des trois bases, ainsi que la moyenne. Nous avons par ailleurs indiqué le classement des unités en fonction de cette corrélation moyenne.

La table ne contient logiquement que des valeurs négatives : le coût de sélection augmente avec la difficulté du système à trouver des unités adéquates, qui est elle-même inversement reliée aux taux de couverture. **Les 1-grammes de sandwichs vocaliques accaparent le haut du classement, ce qui confirme la bonne prise en compte des composantes du coût de sélection par ces unités**. Des corrélations de l'ordre de 0.8 ont ainsi pu être relevées. Les résultats sont tout à fait comparables d'une base à l'autre, ce qui infirme l'hypothèse d'un biais lié à l'optimisation de la base de données pour la couverture de l'une ou l'autre unité. Du fait de la grande taille du corpus de référence, les mesures de corrélation peuvent être jugées comme fiables : toutes les marges de confiance à 95% sont inférieures à 0.005 [Jolion 06].

Naturellement ces résultats pourraient être mis en défaut dans des configurations extrémales. Par exemple sur de très petites bases (quelques milliers de diphones) la plupart des taux de couverture pourraient devenir exagérément bas et des unités plus petites, comme par exemple

Unité	Corrélation sur la voix féminine Agnès (base complète, 258 519 diphones)	Corrélation sur la voix féminine Agnès (base réduite, 66 532 diphones)	Corrélation sur la voix masculine Chirac (rushes, 438 603 diphones)	Corrélation moyenne	Rang
phones, 1-grammes, 13 contextes	-0.02	-0.02	-0.09	**-0.04**	**48**
phones, 1-grammes, 4 contextes	N.A.	N.A.	N.A.	**N.A.**	
phones, 1-grammes, 0 contexte	N.A.	N.A.	N.A.	**N.A.**	
phones, 2-grammes, 13 contextes	-0.52	-0.56	-0.45	**-0.51**	**22**
phones, 2-grammes, 4 contextes	-0.44	-0.53	-0.52	**-0.50**	**25**
phones, 2-grammes, 0 contexte	-0.10	-0.32	-0.37	**-0.26**	**42**
phones, 3-grammes, 13 contextes	-0.68	-0.69	-0.73	**-0.70**	**6**
phones, 3-grammes, 4 contextes	-0.64	-0.68	-0.74	**-0.69**	**9**
phones, 3-grammes, 0 contexte	-0.66	-0.71	-0.73	**-0.70**	**7**
phones, 4-grammes, 13 contextes	-0.62	-0.58	-0.65	**-0.62**	**13**
phones, 4-grammes, 4 contextes	-0.62	-0.61	-0.67	**-0.63**	**11**
phones, 4-grammes, 0 contexte	-0.70	-0.66	-0.73	**-0.70**	**8**
phones, 5-grammes, 13 contextes	-0.53	-0.47	-0.54	**-0.51**	**21**
phones, 5-grammes, 4 contextes	-0.56	-0.50	-0.57	**-0.54**	**20**
phones, 5-grammes, 0 contexte	-0.61	-0.55	-0.62	**-0.59**	**16**
phones, 6-grammes, 13 contextes	-0.48	-0.40	-0.47	**-0.45**	**28**
phones, 6-grammes, 4 contextes	-0.50	-0.42	-0.49	**-0.47**	**27**
phones, 6-grammes, 0 contexte	-0.53	-0.46	-0.53	**-0.51**	**23**
phones, 7-grammes, 13 contextes	-0.44	-0.34	-0.40	**-0.39**	**33**
phones, 7-grammes, 4 contextes	-0.45	-0.36	-0.43	**-0.41**	**30**
phones, 7-grammes, 0 contexte	-0.48	-0.40	-0.46	**-0.45**	**29**
phones, 8-grammes, 13 contextes	-0.40	-0.29	-0.34	**-0.35**	**38**
phones, 8-grammes, 4 contextes	-0.42	-0.31	-0.36	**-0.36**	**37**
phones, 8-grammes, 0 contexte	-0.44	-0.33	-0.40	**-0.39**	**34**
syllabes, 1-grammes, 13 contextes	-0.48	-0.48	-0.54	**-0.50**	**24**
syllabes, 1-grammes, 4 contextes	-0.48	-0.48	-0.51	**-0.49**	**26**
syllabes, 1-grammes, 0 contexte	-0.41	-0.38	-0.44	**-0.41**	**31**
syllabes, 2-grammes, 13 contextes	-0.40	-0.36	-0.43	**-0.40**	**32**
syllabes, 2-grammes, 4 contextes	-0.36	-0.34	-0.41	**-0.37**	**36**
syllabes, 2-grammes, 0 contexte	-0.38	-0.35	-0.43	**-0.39**	**35**
mots, 1-grammes, 13 contextes	-0.28	-0.24	-0.33	**-0.28**	**41**
mots, 1-grammes, 4 contextes	-0.30	-0.25	-0.34	**-0.30**	**40**
mots, 1-grammes, 0 contexte	-0.34	-0.27	-0.36	**-0.32**	**39**
mots, 2-grammes, 13 contextes	-0.25	-0.19	-0.21	**-0.22**	**45**
mots, 2-grammes, 4 contextes	-0.26	-0.20	-0.24	**-0.23**	**44**
mots, 2-grammes, 0 contexte	-0.27	-0.20	-0.27	**-0.25**	**43**
sandwichs, 1-grammes, liquides semi-robustes, 13 contextes	-0.77	-0.74	-0.78	**-0.76**	**2**
sandwichs, 1-grammes, liquides semi-robustes, 4 contextes	-0.78	-0.76	-0.80	**-0.78**	**1**
sandwichs, 1-grammes, liquides semi-robustes, 0 contexte	-0.74	-0.71	-0.74	**-0.73**	**3**
sandwichs, 1-grammes, liquides fragiles, 13 contextes	-0.73	-0.69	-0.72	**-0.71**	**5**
sandwichs, 1-grammes, liquides fragiles, 4 contextes	-0.74	-0.69	-0.72	**-0.72**	**4**
sandwichs, 1-grammes, liquides fragiles, 0 contexte	-0.70	-0.64	-0.67	**-0.67**	**10**
sandwichs, 2-grammes, liquides semi-robustes, 13 contextes	-0.61	-0.55	-0.62	**-0.59**	**15**
sandwichs, 2-grammes, liquides semi-robustes, 4 contextes	-0.64	-0.59	-0.65	**-0.63**	**12**
sandwichs, 2-grammes, liquides semi-robustes, 0 contexte	-0.63	-0.55	-0.62	**-0.60**	**14**
sandwichs, 2-grammes, liquides fragiles, 13 contextes	-0.58	-0.51	-0.57	**-0.56**	**18**
sandwichs, 2-grammes, liquides fragiles, 4 contextes	-0.60	-0.54	-0.59	**-0.58**	**17**
sandwichs, 2-grammes, liquides fragiles, 0 contexte	-0.60	-0.50	-0.56	**-0.55**	**19**
clusters consonantiques robustes, liquides semi-robustes	-0.06	-0.09	-0.14	**-0.10**	**46**
clusters consonantiques robustes, liquides fragiles	-0.05	-0.06	-0.11	**-0.08**	**47**

TABLE 8 – Corrélation entre le coût de sélection et les taux de couverture de différentes unités, pour trois voix de synthèse. Les marges de confiance à 95% sont inférieures à 0.005 pour toutes les valeurs de corrélation. Pour les lignes 2 et 3, les taux de couverture saturés à 100% n'ont pas permis d'estimer une corrélation.

les diphones, pourraient être propulsées en tête du classement. Les deux versions d'Agnès pré-

sentent une légère tendance dans ce sens : en ce qui concerne les diphones et triphones, les corrélations mesurées sur la version réduite sont plus élevées que sur la version complète, tandis que les unités plus longues montrent l'évolution inverse. Les trois bases que nous avons choisies pour cette expérience présentent toutefois des configurations assez standard, tant par leur taille que par leurs taux de couverture. A titre indicatif la table 9 rapporte différents taux de couverture mesurés sur les trois voix de synthèse. L'éventail des valeurs est comparable à celui que nous rencontrerons par la suite dans la plupart de nos scripts de lecture ; en première approximation, les niveaux de corrélation et classements d'unités rapportés en table 8 peuvent donc être considérés comme valables dans tout le cadre de notre étude.

			Taux de couverture			
Nom de la base	**Nombre de phrases dans la base**	**Nombre de diphones dans la base**	**Sandwichs (liquides semi-robustes, 4 symboles contextuels)**	**Sandwichs (liquides fragiles, 13 symboles contextuels)**	**Diphones (4 symboles contextuels)**	**Triphones (13 symboles contextuels)**
Agnes, base complète	7855	258519	89%	74%	100%	87%
Agnes, base réduite	1960	66532	82%	59%	99%	73%
Chirac	27618	438603	86%	68%	100%	82%

TABLE 9 – Quelques taux de couvertures mesurés sur les trois bases.

8 Discussion

8.1 Sur la redondance dans les scripts d'enregistrement

Dans notre travail nous avons une vision binaire de la couverture dans les bases ou scripts de lecture : une unité est soit absente soit couverte. Nous avons vu page 50 que d'autres approches sont possibles, par exemple en visant pour chaque unité un nombre minimum d'occurrences, ce que nous appellerons « facteur de redondance ». Cet ajout volontaire de redondance a deux objectifs principaux :

- se prémunir d'éventuels écarts phonétiques ou prosodiques entre les réalisations du locuteur et la chaîne symbolique attendue, évitant ainsi l'apparition *a posteriori* de trous dans la couverture.
- accroître le nombre d'unités candidates dans le treillis de sélection et ainsi faciliter la minimisation de la fonction de coût.

Un accompagnement assidu du locuteur tout au long de la lecture du script permet de prévenir la plupart des écarts entre les prévisions symboliques et les réalisations acoustiques. Reste à savoir si le deuxième point justifie à lui seul d'introduire de la redondance dans le script de lecture.

L'inconvénient majeur de la redondance est son impact sur la taille du script de lecture. En effet une redondance de facteur K multiplie d'autant la taille de script requise pour atteindre un niveau de couverture donné. Réciproquement pour une taille fixée elle pénalise le taux de couverture, comme illustré sur la figure 23 pour le cas `LRfragiles_13contextes`. Par exemple

pour une taille de script permettant théoriquement d'atteindre 70% de couverture en sandwichs, nous ne pourrons atteindre au mieux que 58% si un facteur 2 de redondance est imposé, 50% avec un facteur 3 et 45% avec un facteur 4. Exprimé autrement, le pourcentage de sandwichs non couverts augmente respectivement de 41%, 67% et 84%, ce qui peut se traduire dans la synthèse finale par une augmentation très significative du nombre de concaténations sur les phonèmes fragiles.

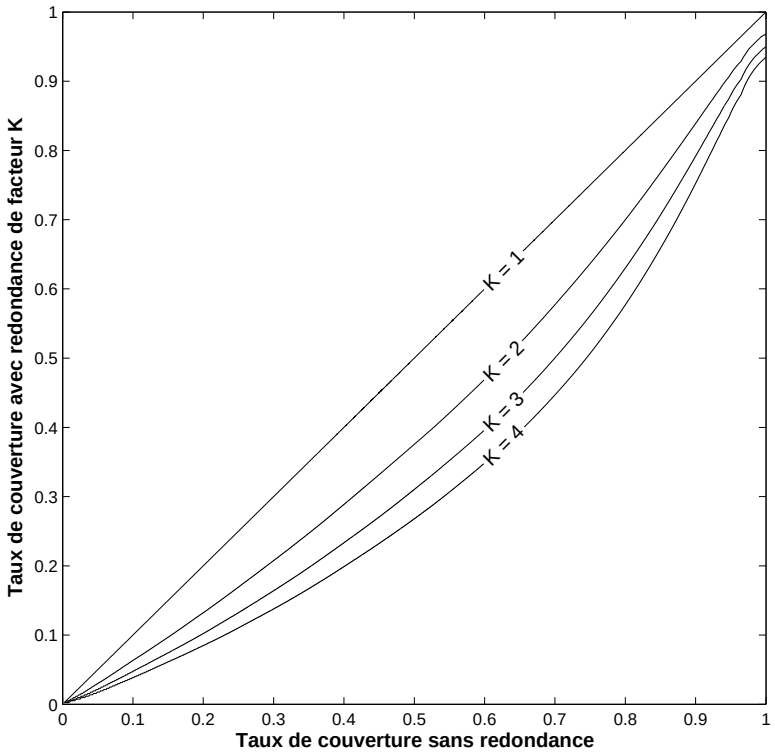


FIGURE 23 – Impact du facteur de redondance sur le taux de couverture, dans le cas `LRfragiles_13contextes`. Pour une taille de script donnée, on reporte le taux de couverture maximal avec une redondance de facteur K en fonction du taux de couverture maximal sans redondance. Pour $K = 1$ (pas de redondance), on suit donc la droite « $y = x$ ».

En nuisant à la couverture en sandwichs, la redondance diminue la protection des voyelles dans l'étape de sélection des diphones. Mais en contrepartie elle offre plus de souplesse pour la reconstruction des sandwichs qui sont couverts. Considérons deux sandwichs successifs apparaissant dans la séquence de diphones-cibles et supposons qu'ils sont couverts chacun K fois dans la base. L'algorithme de sélection dispose alors de K^2 combinaisons possibles pour reconstruire la séquence phonétique, tout en garantissant la protection des phonèmes fragiles. Ceci permet en particulier une réduction de la distorsion acoustique apparaissant sur le(s) phonème(s) robuste(s) entre les sandwichs. En imaginant que la frontière de chaque segment acoustique puisse être modélisée par un point dans un espace acoustique, le problème revient à sélectionner le couple de distance minimale parmi deux ensembles de K points (un ensemble « gauche » vs. un ensemble « droite »). Pour donner un ordre de grandeur, si notre espace est composé de 10 dimensions acoustiques indépendantes, que les réalisations acoustiques suivent une distribution gaussienne centrée réduite, et que nous utilisons une distance euclidienne, le taux moyen de réduction de la distance minimale est de 17% pour $K = 2$, 25% pour $K = 3$ et 31% pour $K = 4$.

La réalité des productions acoustiques et de la perception humaine est bien entendue plus complexe et nous n'entrerons pas dans les détails. Nous nous reposerons simplement sur notre intuition pour formuler la conclusion suivante : si la redondance des sandwichs permet, lors de la synthèse, une légère amélioration des concaténations sur les phonèmes robustes, cela nous semble superflu au regard de son impact très négatif sur la protection des phonèmes fragiles. Une certaine redondance des sandwichs fréquents sera de toute manière inévitable dans le script de lecture ; selon nous il serait contre-productif de chercher à l'accroître... tout au moins pour les scripts de lecture courts et efficaces que nous cherchons à créer.

8.2 Vers une généralisation de l'approche

Nous avons introduit les sandwichs vocaliques suivant le modèle du coût de sélection, qui est lui-même le reflet de la perception humaine d'un signal de synthèse. Pour cette raison le taux de couverture en sandwichs vocaliques dans un script de lecture ou une base est un bon indicateur du niveau de qualité prévisionnel de la synthèse finale. Mais ce taux de couverture n'anticipe que partiellement les mécanismes de sélection :

- la priorité entre symboles contextuels est omise (voir en fin de section 5.2.2) ;
- la tolérance des différents phonèmes aux concaténations est seulement traitée de manière binaire (phonèmes fragiles vs. phonèmes robustes).

Nous avons déjà évoqué les mécanismes proposés par plusieurs auteurs pour dépasser, sur d'autres unités, la première limitation. Dans [Black 01] un traitement hiérarchique des symboles contextuels est utilisé afin de couvrir en priorité, et avec un maximum de finesse, les symboles les plus fréquents du corpus de référence. Dans [Bozkurt 03] une fonction de coût, proche du coût-cible de l'algorithme de sélection, permet d'éviter la couverture d'unités trop semblables en privilégiant des contextes éloignés. Nous proposons dans cette section une méthode plus globale qui consiste à guider la création des scripts avec une version approchée du coût de sélection, intégrant à la fois le coût-cible et le coût de concaténation.

Au stade de la constitution du script de lecture, les composantes acoustiques nécessaires au calcul du coût de sélection ne sont naturellement pas encore disponibles. Nous proposons donc d'utiliser une **approximation symbolique $\tilde{C}$ du coût de sélection** C. Pour la projection symbolique du coût-cible, il suffit de se reposer sur les symboles contextuels. Le coût d'échange entre ces symboles est la plupart du temps déjà défini dans le système de synthèse ; il consiste à pénaliser un écart entre les marqueurs de la cible et les marqueurs d'une unité candidate, par exemple en s'inspirant d'une classification arborescente des contextes. Pour la projection symbolique du coût de concaténation, qui fait quant à lui grand usage de considérations acoustiques, nous proposons de nous reposer sur la hiérarchie de classes phonétiques présentée en 2.4.3. La pénalité associée à une concaténation est dans ce cas « forfaitaire », en fonction du type de phonème. Nous ne détaillerons pas ici les valeurs précises de pénalités, car elles dépendent entièrement du système : allure globale de la fonction de coût (logarithmique, linéaire...), pondération du coût-cible par rapport au coût de concaténation, propriétés des algorithmes de traitement du signal utilisés pour le lissage, etc.

Avec $\tilde{C}$, on définit en quelques sortes la **synthèse symbolique** d'un contenu textuel. Comme $\tilde{C}$ est indépendant des réalisations acoustiques, cette synthèse symbolique ne nécessite que la connaissance du script de lecture. Il devient alors possible d'estimer la qualité globale d'un script de lecture de manière plus complète qu'avec des taux de couverture. Étant donné un script de lecture $script\_courant$ nous mesurons $\tilde{C}(corpus\_ref|script\_courant)$, c'est-à-dire le **coût total obtenu pour la synthèse symbolique du corpus de référence**. Ceci nous fournit une estimation assez précise de la qualité de synthèse qu'offrira ce script de lecture après l'enregistrement de voix. Il est en particulier possible de comparer la qualité de plusieurs

scripts et donc de guider l'algorithme de construction du script : pour une taille donnée, le script optimal est celui qui minimise le coût de synthèse symbolique sur le corpus de référence.

Par ce procédé nous faisons implicitement l'hypothèse que les enregistrements ultérieurs respecterons les chaînes symboliques prévues dans le script de lecture. Ce n'est pas forcément le cas : les futurs locuteurs introduiront peut-être une certaine variabilité, comme par exemple des prononciations inattendues. Mais avec une supervision assidue des enregistrements et en l'absence de révision manuelle ces différences restent marginales voire inexistantes.

Une telle approche, reposant sur un « coût global de synthèse symbolique », nous semble prometteuse. Elle est toutefois très gourmande en calculs, surtout lorsque l'algorithme de constitution du script requiert de nombreuses itérations (ce qui est à peu près toujours le cas). A chaque itération et pour toutes les modifications possibles du script de lecture, la synthèse symbolique des 250 000 phrases du corpus de référence doit être recalculée. Même s'il est possible de réduire la complexité algorithmique en factorisant certains calculs, **nous n'avons pas trouvé à ce jour d'implémentation suffisamment efficace pour être utilisable sur les ordinateurs actuels**. Dans ce contexte, l'utilisation de taux de couverture en sandwichs vocaliques semble être l'approximation la plus réaliste et la plus satisfaisante de notre problème.

Troisième partie

Quand les gloutons font la course aux sandwichs

Sommaire

Dans la partie précédente nous avons proposé un nouveau critère pour guider la constitution du script de lecture : le taux de couverture en sandwichs vocaliques, ou VSCR [41]. Nous nous intéressons à présent à l'optimisation du script suivant ce critère. Il s'agit donc d'explorer l'ensemble incommensurable des scripts possibles afin d'en extraire celui qui, pour une longueur prédéfinie, maximise le taux de couverture en sandwichs. Après avoir détaillé quelques choix généraux, nous présenterons et comparerons deux procédés d'optimisation [Cadic 10a]. Le premier, assez traditionnel, consiste en la condensation d'un vaste corpus textuel. Le second, expérimental, consiste à fabriquer de manière semi-automatique des phrases les plus denses possibles.

9 Stratégie globale d'optimisation du script

9.1 Le groupe de souffle comme élément de base

Tous les travaux de l'état de l'art utilisent des phrases complètes comme élément constitutif de base des scripts de lecture. Nous avons préféré fonctionner avec des **groupes de souffle**. Ces parties de phrases délimitées par des pauses correspondent, dans notre système, aux groupes intonatifs (voir page 26). De ce fait ils sont **prosodiquement autonomes**, du moins en ce qui concerne les plans syntaxique et lexical qui animent le style « lecture neutre ». Les plans supérieurs comme la sémantique peuvent introduire des interactions entre groupes intonatifs, mais dont les systèmes de synthèse actuels ne savent pas tenir compte. La restriction, dans nos travaux, à des groupes intonatifs isolés ne semble donc en aucun cas pénalisante.

Un intérêt majeur des groupes de souffle porte sur leur longueur. Tandis que les phrases peuvent se prolonger indéfiniment avec des propositions ou appositions, les groupes de souffle restent contraints par les nécessités respiratoires. Les hauts-niveaux de notre système de synthèse tiennent compte de ces contraintes pour proposer automatiquement un découpage assez réaliste des phrases en groupes de souffle. Nous avons appliqué ce découpage sur notre corpus de référence et avons reporté en figure 24 les distributions de longueurs obtenues pour les groupes de souffle et les phrases. 90% des groupes de souffle ont une taille inférieure à 31 phonèmes, contre seulement 63% des phrases. En moyenne, les phrases comportent 35.1 phonèmes et sont composées de 2.1 groupes de souffles (de longueur moyenne 16.4 phonèmes).

L'utilisation des groupes de souffles **facilite la lecture du script final**. D'une part leur taille inférieure à celle de la phrase permet une diminution du nombre d'hésitations et de la durée des reprises. D'autre part ils permettent de clarifier les consignes de lecture : aucune pause ne doit intervenir à l'intérieur d'un groupe de souffle. Ceci a également pour effet de réduire la variabilité des enregistrements, ce qui améliore les segmentation et annotation automatiques qui sont très sensibles à la présence de pauses imprévues. Enfin, nous verrons plus bas que leur petite taille offre une **flexibilité accrue pour l'optimisation de la couverture**.

Pour toutes ces raisons, nous pensons que le choix du groupe de souffle comme élément de base est pertinent. Dans la suite, nous continuerons à parler de « phrases » dans un souci de simplicité, même si par ce terme nous désignerons en réalité des groupes de souffle isolés.

9.2 La question de la longueur des phrases

La gestion de la longueur des phrases constitue une difficulté courante dans les travaux de création de textes. Ici, l'intérêt de chaque phrase est évalué en fonction de sa contribution à la

41. Vocalic Sandwiches Coverage Rate

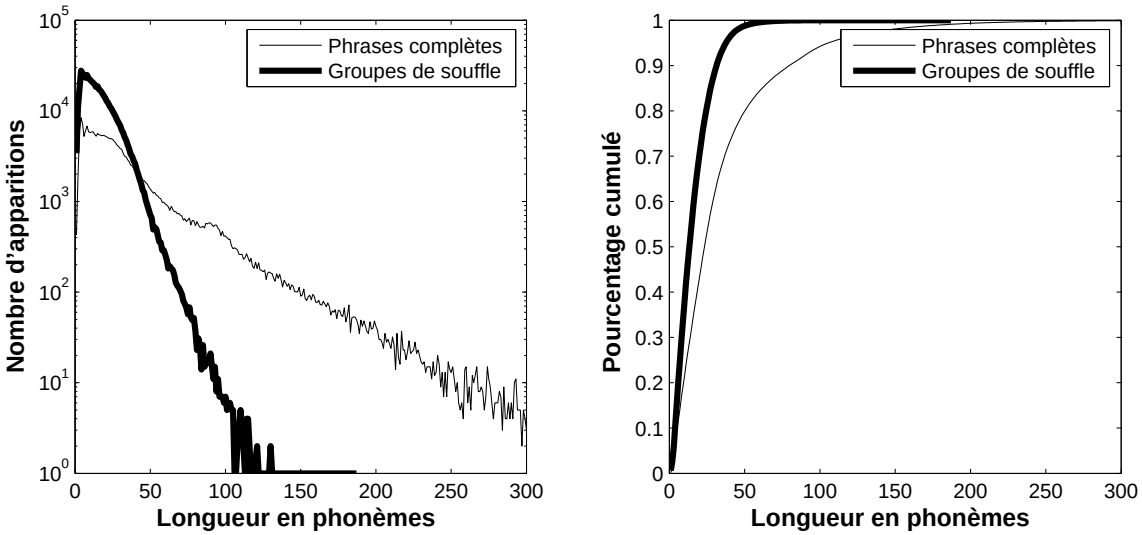


FIGURE 24 – Distributions des longueurs des groupes de souffle et des phrases dans notre corpus de référence. Sur le graphique de gauche sont rapportés les nombres d'apparitions de chaque longueur, en échelle logarithmique. Le graphique de droite présente les fonctions de répartition, c'est-à-dire le pourcentage de réalisations inférieures à une longueur donnée.

couverture linguistique globale. Par conséquent, lorsqu'elle est considérée de manière unitaire, une phrase longue a statistiquement plus d'intérêt qu'une phrase courte. Dans la pratique ceci n'est évidemment pas souhaitable, pour au moins deux raisons. Tout d'abord, pour anticiper correctement la durée d'enregistrement d'un script de lecture, sa longueur ne doit pas être mesurée en nombre de phrases mais plutôt en nombre de caractères, de mots, ou mieux, de phonèmes. Les phrases longues sont donc plus pénalisantes que les phrases courtes. Ensuite, les phrases trop longues suscitent lors de la lecture un nombre accru d'hésitations et augmentent la durée nécessaire pour chaque reprise ; le taux de parole utile récoltée peut ainsi régresser significativement.

Une approche simple et naturelle consiste alors à normaliser l'apport de chaque phrase par sa longueur. Mais en faisant cela on tombe rapidement dans l'excès inverse ! En effet, ce sont dès lors les phrases très courtes qui présentent le plus d'intérêt pour la maximisation de la couverture. Elles offrent une souplesse accrue à l'algorithme d'optimisation, qui peut ainsi mieux cibler les unités cibles et éviter la redondance. A l'inverse les phrases longues deviennent contraignantes, car l'apport d'unités intéressantes est souvent conditionné à l'acceptation d'autres unités moins intéressantes, voire redondantes, qui font baisser la moyenne.

La présence d'un trop grand nombre de phrases courtes dans le script de lecture n'est pas non plus souhaitable, car la proportion élevée de pauses entre les phrases nuit à l'efficacité des enregistrements. La pertinence de notre modèle prosodique de « lecture neutre », matérialisé par un jeu réduit de symboles contextuels, est également menacée. Les phrases très courtes, souvent porteuses d'une emphase exagérée, risquent en effet de le mettre en défaut. Des symboles plus complexes, tenant compte d'une position plus précise dans la phrase, pourraient pallier cela. En augmentant les contraintes de couverture ils forceraient par la même occasion le choix de phrases plus longues. Mais nous ne souhaitons pas augmenter la complexité des symboles contextuels. Il nous faut donc contraindre différemment la longueur des phrases.

Après plusieurs essais, il nous a semblé qu'un bon compromis entre le taux de couverture et la lisibilité du script pouvait être obtenu simplement en pénalisant les phrases trop longues, sans normaliser leur apport par leur longueur. Cette pénalité est introduite de manière croissante au-delà d'un seuil de longueur et est dimensionnée de sorte que toute unité supplémentaire

devienne pénalisante, même si elle est très fréquente. Logiquement le seuil de longueur devrait être exprimé en nombre de phonèmes, car c'est une mesure simple et fiable de la durée d'enregistrement. Le fonctionnement intrinsèque de nos outils, et plus particulièrement des outils de construction exposés en section 11, nous a toutefois incités à l'exprimer en nombre de sandwichs. Ceci nous offre un contrôle moins rigoureux des longueurs de phrases, puisque les sandwichs ont une longueur variable. La différence est toutefois marginale et ce choix nous permet de rester cohérents avec nos autres choix : utilisation d'un taux de couverture en sandwichs plutôt qu'un taux de protection des phonèmes fragiles (voir page 91), présentation des taux de couverture en fonction du nombre de sandwichs et non en fonction du nombre de phonèmes.

Le seuil a évolué au fil de nos expériences pour finalement se stabiliser à 15 sandwichs. En théorie notre choix de pénalité n'est pas rédhibitoire : le seuil de longueur peut être dépassé si cela est vraiment indispensable pour couvrir certains enchaînements d'unités importantes. Ce pourrait être le cas, par exemple, de phrases qui apparaissent fréquemment dans leur intégralité : « *Pour accéder au menu principal appuyez sur la touche étoile.* ». Mais en pratique ce cas est extrêmement rare et toutes les phrases sont composées de sous-séquences de sandwichs qui peuvent être couvertes avec des phrases plus courtes : « *Pour accéder au menu principal,* », « *C'est le principal appui du gouvernement.* », « *Appuyez sur la touche étoile.* ». Notre système de pénalité revient donc exactement à **interdire les phrases trop longues**. Une étude quantitative de cette stratégie sera proposée plus bas, dans les cas particuliers de la condensation de corpus et de la construction de phrases.

9.3 Le glouton « fréquents d'abord »

Nous avons vu en section 3.1.3 que l'**algorithme glouton** est fréquemment utilisé pour l'optimisation des scripts de lecture. Il s'agit d'une approche « grossissante » : le script final est construit en y ajoutant une à une des phrases qui maximisent un certain critère. Nous avons également évoqué page 52 les principales vertus de l'algorithme glouton : rapidité, optimalité des scripts partiels, performances proches de l'optimum. En outre, la légère sous-optimalité du glouton est en grande partie due à une redondance des unités fréquentes, dont nous avons vu au paragraphe 8.1 qu'elle participait également à la qualité de la synthèse finale. Pour toutes ces raisons, nous n'utiliserons dans la suite de nos travaux que des processus gloutons.

Nous avons par ailleurs justifié en section 7.2 notre choix de suivre une approche fréquentielle pour la constitution de nos scripts de lecture : nous souhaitons couvrir avant tout les unités fréquentes. Pour cela, deux critères de sélection de phrases sont possibles. Le glouton « fréquents d'abord », qui est aussi le plus courant, consiste à inclure en priorité les phrases qui apportent le plus d'unités fréquentes non couvertes. Autrement dit il maximise l'augmentation à court terme du taux de couverture, ce qui garantit au passage l'optimalité des scripts partiels mentionnée plus haut. Le glouton « rares d'abord » s'attache quant à lui à optimiser la couverture du script dans son ensemble, en délaissant le court terme au profit du long terme. Partant du constat que la couverture d'unités rares apporte aussi des unités fréquentes, ce procédé consiste à inclure en priorité les phrases qui apportent le plus d'unités rares non couvertes. Il implique une croissance assez lente du taux de couverture en début de processus, mais par la suite les performances peuvent être améliorées. L'optimalité des scripts partiels n'est en revanche pas garantie.

François et Boëffard ont comparé ces deux procédés dans un contexte précis [Francois 02] et les résultats se sont montrés plutôt favorables au **glouton « fréquents d'abord »**. C'est l'approche que nous avons retenue, également parce qu'elle garantit l'optimalité des scripts partiels. Ce choix sera conforté en section 10.2.2 par une étude quantitative des deux approches dans notre contexte spécifique et dans le cas particulier d'une condensation de corpus.

10 L'approche par condensation de corpus

10.1 Principe et évaluation

Dans la section précédente nous avons justifié quelques choix généraux pour l'optimisation de nos scripts de lecture : nous utiliserons des processus gloutons basés sur les groupes de souffle et suivant un critère de sélection « fréquents d'abord ». Nos scripts seront donc constitués en ajoutant une à une des phrases qui maximisent l'augmentation du VSCR. Mais la question de l'origine de ces phrases reste ouverte.

Dans cette section, nous expérimentons la « condensation de corpus », méthode classique dans laquelle les phrases sont sélectionnées au sein d'un vaste corpus de pioche. Comme suggéré page 51, notre corpus de référence fera office de corpus de pioche. Cependant, du fait de notre contrainte sur la longueur des phrases, une partie de ce corpus ne sera pas exploitable pour la constitution de scripts : environ 5% des groupes de souffle dépassent le seuil de longueur de 15 sandwichs que nous avons fixé et ne pourront donc pas être sélectionnés.

Nous avons observé les performances de la condensation de corpus dans le cas `LRsemirobustes_13contextes`. La figure 25 rapporte l'évolution du VSCR tout au long du processus glouton de sélection des phrases. Nous y avons également reporté une borne haute et une borne basse. La borne haute n'est autre que la fonction de répartition des sandwichs vocaliques ; elle correspond donc à un processus idéal qui permettrait la création de scripts parfaitement denses, contenant uniquement les sandwichs les plus fréquents et sans aucune redondance. La borne basse présente quant à elle l'évolution du VSCR dans le cas où les phrases sont tirées aléatoirement du corpus de pioche, ce qui correspond à un processus d'optimisation totalement inefficace. Sur la figure les tailles de scripts sont indiquées en nombre de sandwichs ; à titre indicatif un sandwich « pèse » en moyenne 2.63 phonèmes [42] et 4.05 caractères textuels.

Les courbes montrent que **le processus glouton de condensation donne accès à des scripts 2 à 3 fois plus denses que la sélection aléatoire**. Par exemple un script de 10 400 sandwichs permet d'atteindre un taux de couverture de 80%, alors que 25 000 sandwichs sont nécessaires avec une sélection aléatoire. **Malgré tout les performances restent très en deçà de l'optimum**, puisqu'une couverture de 80% est théoriquement accessible avec seulement 5 400 sandwichs (les plus fréquents).

Cette sous-optimalité s'explique en partie par l'utilisation du glouton : en n'explorant qu'une infime partie de l'ensemble des possibles, le glouton est responsable d'une inflation des scripts par rapport à une condensation optimale. Dans un contexte un peu différent, [Chevelu 07] évalue cet impact à environ 10% de la taille du script. L'autre cause de sous-optimalité est liée à la condensation elle-même. Cette approche reste en effet contrainte par le nombre limité de séquences contenues dans le corpus de pioche. Les enchaînements de sandwichs y sont souvent redondants et ne permettent pas d'envisager une couverture très dense. Cette restriction est d'ailleurs aggravée par notre volonté de favoriser, pour des raisons de lisibilité, les phrases de longueur moyenne (voir en 9.2) : ceci tend d'une part à réduire le choix de phrases, d'autre part à écarter les phrases courtes qui offrent le plus de souplesse pour la condensation.

42. Cela ne correspond pas exactement à la longueur moyenne de 3.18 phonèmes indiquée en table 6 page 89, car il faut soustraire les chevauchements phonétiques. La valeur de 2.63 correspond en fait à l'inverse de la densité (qui vaut 0.38).

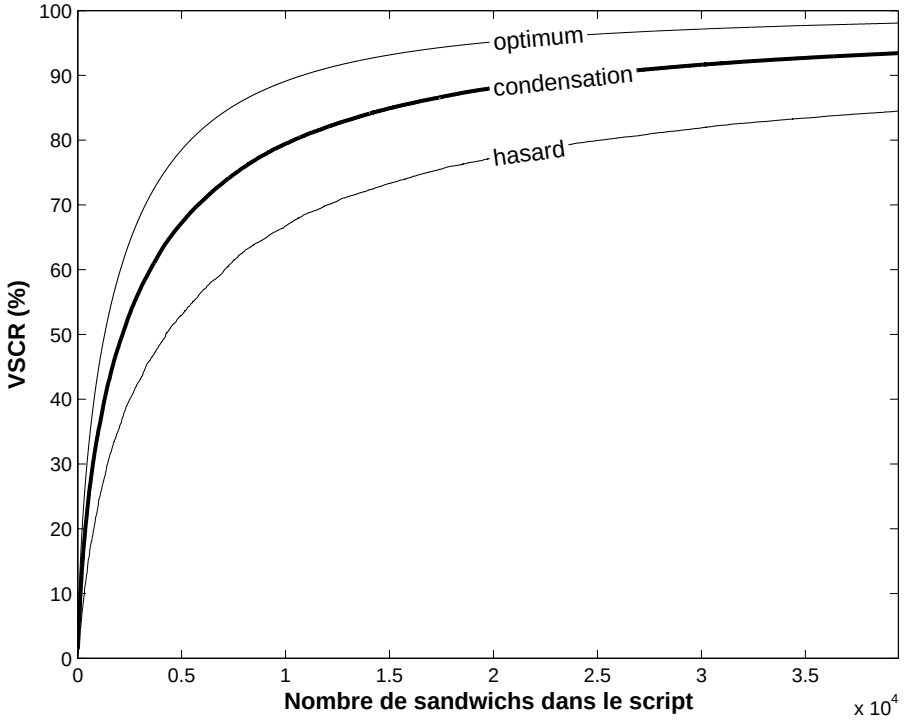


FIGURE 25 – Evolution du VSCR au fil du processus glouton de sélection de phrases (*condensation*), dans le cas `LRsemirobustes_13contextes`. La borne haute (*optimum*) correspond à la fonction de répartition des sandwichs et la borne basse (*hasard*) à une sélection de phrases aléatoire.

10.2 Validation du critère de sélection des phrases

10.2.1 Impact de la contrainte de longueur

Pour favoriser les phrases de longueur moyenne, nous avons fait le choix de ne pas normaliser l'apport en VSCR par la longueur de la phrase, mais plutôt de pénaliser les phrases longues. Nous souhaitons à présent mesurer, dans le cas d'une condensation, l'impact de cette stratégie sur la densité du script final. Pour cela nous la comparons à deux autres stratégies : l'une avec normalisation et sans pénalité, l'autre sans normalisation ni pénalité.

La figure 26 présente l'évolution du VSCR pour les trois stratégies, en fonction des tailles de scripts. Nous avons également relevé, pour chacune des trois stratégies, la distribution des longueurs de phrases dans le script présentant un VSCR de 80%. Ces distributions sont rapportées en figure 27, en nombre de sandwichs (partie gauche) et en nombre de phonèmes (partie droite).

On constate sans surprise la supériorité du critère normalisé en ce qui concerne le compromis entre le taux de couverture et le nombre de sandwichs dans le script. Il s'agit, par définition, du critère le plus naturel si on se focalise exclusivement sur la densité. Néanmoins, conformément à notre analyse du paragraphe 9.2, les phrases sélectionnées sont très courtes, avec une longueur moyenne de 3.8 sandwichs (9.4 phonèmes) ; aussi avons-nous vu les problèmes que cela pouvait poser, tant pour l'efficacité des enregistrements que pour le respect de notre modèle prosodique de « lecture neutre ».

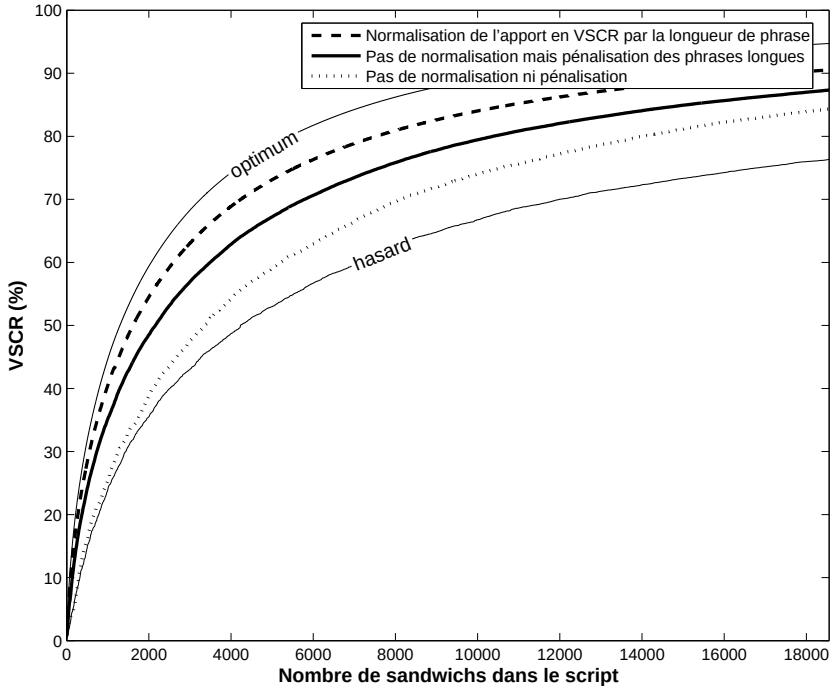


FIGURE 26 – Evolution du VSCR pour les trois stratégies de gestion des longueurs de phrases.

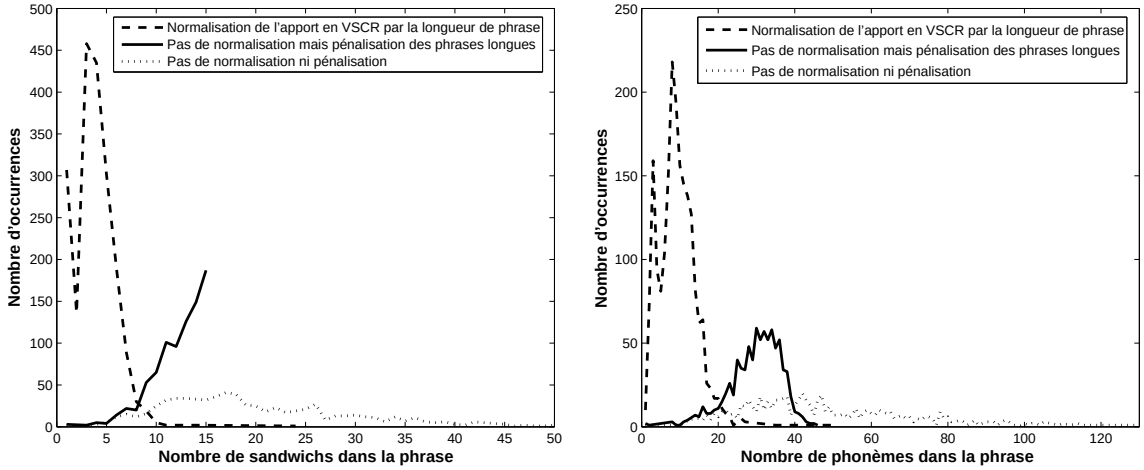


FIGURE 27 – Distribution des longueurs de phrases, en nombre de sandwichs et en nombre de phonèmes, pour les trois stratégies.

Notre stratégie (avec pénalité sur les phrases longues et sans normalisation) offre quant à elle une distribution plus satisfaisante des longueurs de phrases. La longueur moyenne des phrases sélectionnées est de 12.2 sandwichs (29.8 phonèmes). Les relevés confirment que le niveau de notre pénalité est tel que le seuil de longueur de 15 sandwichs n'est jamais franchi, ce qui revient à interdire les phrases trop longues. Concernant la densité, notre critère se révèle significativement moins efficace que le critère normalisé : à VSCR fixé, il induit une augmentation de 30 à 40% sur la taille du script. Ce surcoût est assez élevé ; il semblerait donc que, dans le cadre d'une condensation, nous payions assez cher notre aversion pour les phrases trop courtes.

La troisième stratégie, qui exploite les phrases de manière unitaire et sans aucune gestion de leur longueur, se révèle inadaptée sur tous les plans. Avec une longueur moyenne de 49.3 phonèmes, les phrases sélectionnées sont particulièrement longues, au point de s'étendre jusqu'à 187 phonèmes ! La densité de la couverture est également faible : à VSCR fixé, les scripts constitués sont environ deux fois plus longs qu'avec le critère normalisé.

Au vu de ces résultats notre stratégie apparaît comme un compromis raisonnable entre la longueur des phrases et la densité de la couverture en sandwichs vocaliques.

10.2.2 « Fréquents d'abord » vs. « rares d'abord »

Notre approche « fréquents d'abord » consiste à ajouter en priorité les phrases qui maximisent l'augmentation du VSCR. Nous comparons ici ses performances, toujours dans le cadre d'une condensation de corpus, à celles d'une approche « rares d'abord ».

Pour l'implémentation de cette dernière, nous réutilisons l'algorithme « fréquents d'abord » mais en modifiant, dans le critère de sélection, l'ordre des sandwichs de la fonction de répartition. Plus précisément nous introduisons un seuil d'inversion $\alpha \in [0;1]$ et désignons par N_α le rang du sandwich qui permet à la fonction de répartition de franchir le seuil α :

$$\begin{cases} F(N_\alpha - 1) & = \sum_{r=1}^{N_\alpha - 1} f(r) & < \alpha \\ F(N_\alpha) & = \sum_{r=1}^{N_\alpha} f(r) & \geq \alpha \end{cases} \tag{12}$$

Puis nous définissons les fréquences partiellement inversées $f_{inv}(r)$ à partir des fréquences d'origine $f(r)$, en renversant l'ordre des sandwichs de rang fréquentiel inférieur à N_α :

$$\begin{cases} \forall r \leq N_\alpha, & f_{inv}(r) = f(N_\alpha + 1 - r) \\ \forall r > N_\alpha, & f_{inv}(r) = f(r) \end{cases} \tag{13}$$

La figure 28 illustre cette opération. L'intérêt est d'adapter l'ensemble des sandwichs à couvrir au taux de couverture global envisagé. Les « sandwichs rares » ne sont en effet pas les mêmes suivant qu'on vise un VSCR de 50% ou un VSCR de 90%. Notre implémentation du « rares d'abord » cible en priorité les sandwichs les plus rares, mais dans la limite d'un certain taux de couverture.

La figure 29 présente l'évolution du VSCR pour les stratégies « fréquents d'abord » et « rares d'abord » avec, pour cette dernière, deux valeurs du seuil d'inversion α : 80% et 100%. Notons au passage que $\alpha = 100\%$ implique un renversement complet des fréquences de sandwichs, ce qui revient simplement à désactiver le mécanisme d'inversion partielle.

La figure montre que, dans notre contexte expérimental, ce mécanisme est indispensable. Sans lui (c'est-à-dire avec $\alpha = 100\%$) l'approche « rares d'abord » présente des performances très inférieures à l'approche « fréquents d'abord ».

Avec un seuil $\alpha = 80\%$ la tendance des résultats dépend de la taille des scripts. Pour les petits scripts, de VSCR inférieur à 80%, l'approche « fréquents d'abord » l'emporte sur

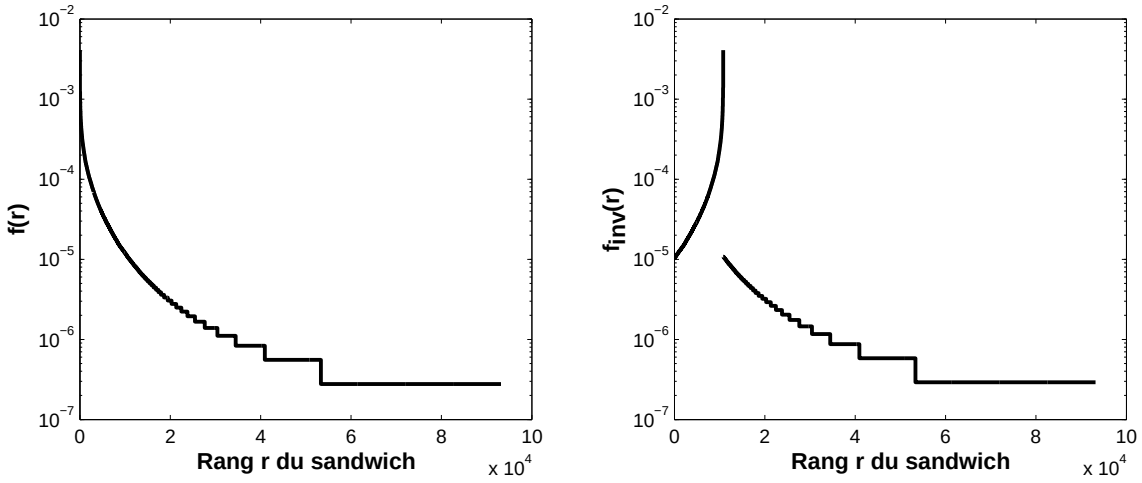


FIGURE 28 – Illustration du renversement partiel des sandwichs opéré pour la stratégie « rares d'abord ». Ici $\alpha = 0.9$, ce qui conduit dans le cas `LRsemirobustes_13contextes` à $N_\alpha = 10800$.

l'approche « rares d'abord, seuil de 80% ». Le VSCR de 80% constitue un pivot, qui est atteint pour la même taille de script dans les deux cas (10370 sandwichs). Ensuite les résultats sont très légèrement favorables à l'approche « rares d'abord » sur quelques dizaines de phrases, puis sont à peu près confondus pour les scripts de taille supérieure. Notons que ce comportement est identique quelle que soit la valeur de α (dans le cas particulier où $\alpha = 100\%$, nous n'observons en fait que la première phase).

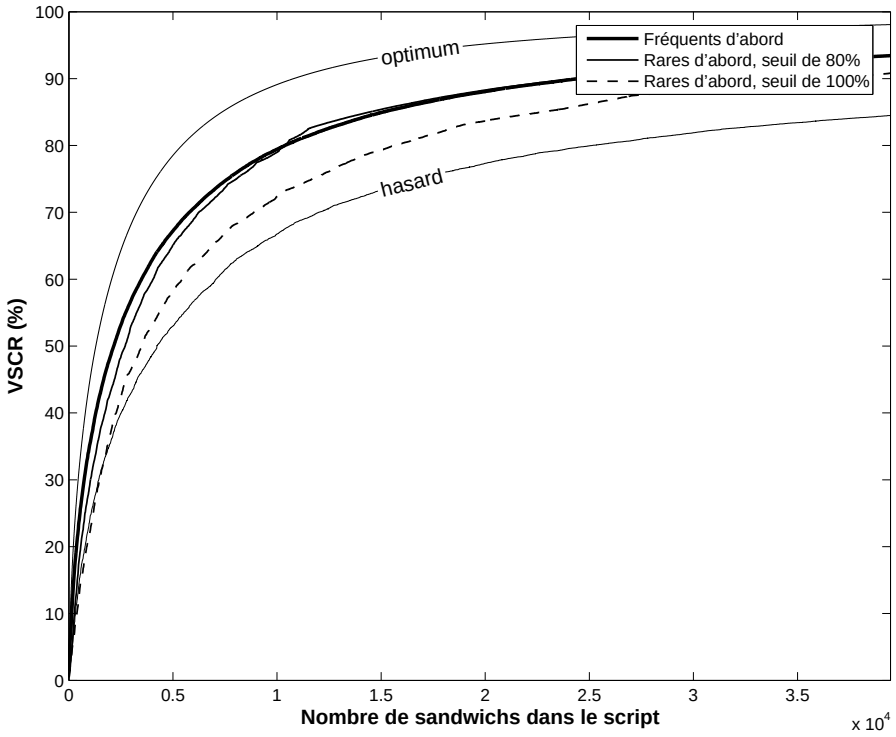


FIGURE 29 – Evolution du VSCR pour les stratégies « fréquents d'abord » et « rares d'abord ».

L'approche « rares d'abord » permet donc, lorsque son seuil d'inversion est bien dosé, d'accroître très légèrement la densité globale du script de lecture. Néanmoins cet apport reste

marginal et surtout il se fait au détriment de l'optimalité des scripts partiels. Cette expérience, menée dans le cas particulier d'une condensation, **conforte donc notre choix général de suivre une approche « fréquents d'abord »**.

10.3 La condensation en pratique

Le guidage d'un processus de condensation en situation réelle et à des fins d'exploitation nécessite certaines précisions. Ce sera l'objet du paragraphe 10.3.1. Dans le paragraphe 10.3.2 nous ferons l'inventaire des scripts constitués suivant ce procédé, chacun correspondant à une situation et à des choix différents.

10.3.1 Guidage de la condensation

Modulations du critère de sélection

Nous avons proposé au paragraphe 6.3 plusieurs variantes pour la définition des sandwichs vocaliques, avec notamment deux classifications possibles pour les phonèmes liquides et plusieurs niveaux de précision contextuelle pour les voyelles. Nous avons également suggéré l'utilisation de bigrammes de sandwichs pour intégrer des phénomènes linguistiques et prosodiques plus globaux.

Toutes ces variantes présentent des distributions d'allure logarithmique (voir en 7.2). La couverture des plus complexes d'entre elles peut être très coûteuse en taille de script, à moins de se contenter d'une couverture « plancher ». Prenons l'exemple des unités les plus complexes que nous ayons introduites : les bigrammes de sandwichs dans le cas `LRfragiles_13contextes`. Évidemment leur dispersion est très grande et, d'après la table 6 page 89, un script d'au moins 301 461 sandwichs est nécessaire pour atteindre une couverture de 80% de ces bigrammes ! Toutefois une couverture beaucoup plus modeste de 30% reste largement accessible puisqu'elle ne nécessite théoriquement que 3 317 sandwichs. Cette couverture minimale peut constituer un « socle » très avantageux pour la future voix de synthèse. Elle garantit en effet que près de la moitié des phonèmes [43] utilisés lors de la synthèse pourront être accompagnés d'un large extrait de leur environnement original (de l'ordre de 6 phonèmes) et avec une caractérisation contextuelle précise (13 symboles). De plus, les 3 317 sandwichs qui permettent cette couverture de 30% des bigrammes contribuent par la même occasion à la couverture des autres variantes : 52% des 1-grammes de type `LRfragiles_13contextes`, 67% des 1-grammes de type `LRsemirobustes_4contextes`, etc.

Pour la constitution d'un script de lecture, il peut donc être avantageux d'assurer une couverture « plancher » de variantes complexes avant de se focaliser sur des variantes plus abordables. Nous suggérons par exemple, pour une taille de script de l'ordre de 20 000 à 30 000 sandwichs, d'utiliser successivement ces quatre variantes, de la plus contraignante à la moins contraignante :

- **2-grammes de sandwichs, liquides fragiles, 13 contextes**
  Ces unités riches ont une longueur moyenne de 5.8 phonèmes. Leur couverture, difficilement accessible, garantit une haute qualité de synthèse. Un taux de couverture de 30% est envisageable.

43. Du fait des chevauchements entre bigrammes, la proportion des phonèmes appartenant à un bigramme couvert est significativement supérieure à la proportion de bigrammes couverts.

- **1-grammes de sandwichs, liquides fragiles, 13 contextes**
  La longueur moyenne de ces unités est de 3.7 phonèmes et une couverture cible de 80% semble raisonnable pour un script de 20 000-30 000 sandwichs.

- **1-grammes de sandwichs, liquides semi-robustes, 13 contextes**
  La longueur moyenne de ces unités est de 3.2 phonèmes et un taux de couverture de 90% peut être envisagé.

- **1-grammes de sandwichs, liquides semi-robustes, 4 contextes**
  Il s'agit d'une variante très simple, pour laquelle on visera typiquement un taux de couverture de 95%.

La variante `LRsemirobustes_0contexte`, encore plus simpliste, se révèle inadaptée pour un script de cette taille. On l'utilisera exclusivement pour la création de scripts très courts.

Enfin, d'autres unités que les sandwichs vocaliques peuvent être prises en compte dans le critère de sélection des phrases. On peut ainsi se préoccuper de la couverture des diphones hors contexte, qui sont l'élément de base du système pour la sélection et la concaténation. L'absence de l'un d'entre eux peut en effet occasionner un « trou » difficile à combler. De la même manière les clusters consonantiques robustes (voir 6.4) peuvent faire l'objet d'une attention particulière. Cependant il s'agit dans les deux cas de couvertures faciles à obtenir, qui viennent assez naturellement sans qu'il soit vraiment nécessaire de s'en préoccuper.

Critères d'arrêt

La plupart du temps, la constitution d'un script de lecture fait suite au choix d'une durée d'enregistrement cible, qui est elle-même fonction d'attentes qualitatives et de contraintes budgétaires. Dans ce cas l'approche la plus naturelle est d'interrompre le processus de condensation dès que le nombre de phonèmes cible est dépassé. Mais un autre critère d'arrêt possible peut être le franchissement d'un certain niveau de couverture.

Nous avons vu que notre processus est découpé en sous-phases consécutives, portant chacune sur la couverture d'une variante précise de sandwichs vocaliques. Pour ces sous-phases, l'utilisation d'un critère d'arrêt de type « taux de couverture » est parfois plus adaptée. Les seuils sont alors définis en tenant compte de plusieurs éléments :

- **la taille de script envisagée** ;
- **le taux optimal qui peut être atteint**, d'après la distribution de la variante dans le corpus de référence et étant donné la taille de script envisagée ;
- **la densité permise par le processus de condensation**, qui est environ deux fois moindre que l'optimum (voir 10.1) ;
- **Le nombre d'occurrences des « unités marginales »** dans le corpus de référence.

Nous entendons par unités marginales les unités les plus fréquentes qui ne sont pas encore couvertes. Ainsi dans le cas des sandwichs `LRsemirobustes_13contextes`, un taux de couverture supérieur à 95% nous amènerait à considérer des sandwichs qui apparaissent moins de 10 fois dans le corpus de référence. Il en résulterait une certaine dépendance aux sources textuelles utilisées pour la construction de ce corpus [Van Santen 97b]. Pour limiter ce phénomène nous préférons nous focaliser sur les unités qui apparaissent au moins 25 fois dans le corpus de référence. Les taux limites que nous évoquions au paragraphe précédent satisfont cette contrainte.

Supervision

Nous avons évoqué page 52 les bénéfices des traitements manuels dans ce type d'application. Nous y avons eu largement recours pour la constitution de nos scripts de lecture. Tout au long du processus de condensation, les phrases sélectionnées ont fait l'objet d'une validation manuelle. Les phrases incorrectes ou trop complexes ont ainsi été rejetées et des annotations textuelles ont permis de clarifier le script en vue des futurs enregistrements.

10.3.2 Inventaire des scripts obtenus par condensation

La table 10 donne la liste des scripts obtenus par condensation qui ont été exploités dans le cadre de cette thèse. Les noms des scripts sont composés de la **durée approximative d'enregistrement (qui n'est pas la durée de parole utile)** et du type d'unité principalement utilisé dans le critère de sélection des phrases. Les durées d'enregistrement ont été évaluées empiriquement au fil des sessions de lecture (voir partie IV) ; bien qu'elles varient suivant les locuteurs, elles dépendent essentiellement de la taille du script et de la complexité des phrases. Pour chaque script nous donnons les taux de couverture de quelques unités. Ces taux peuvent difficilement être comparés à l'état de l'art pour plusieurs raisons :

- ils sont spécifiques à la langue française ;
- peu de travaux de l'état de l'art (surtout en français) s'intéressent à des taux de couverture fréquentiels et sans redondance ;
- les taux sont liés au corpus de référence sur lequel ils sont mesurés ;
- notre prise en compte du contexte symbolique est très spécifique et influe beaucoup sur la dispersion des unités et donc sur la mesure des taux.

Le mode opératoire utilisé pour la création de chaque script est détaillé ci-dessous. Les scripts « 8days_nphones » et « 45min_nphones » n'ont pas été créés par nos soins ; ils sont issus d'autres travaux de l'équipe des Orange Labs. Ils figurent tout de même dans la table car nous les avons largement réutilisés, notamment pour l'évaluation comparative de voix de synthèse rapportée en partie IV.

« 5h_sandwiches » :

C'est le dernier script en date. Nous l'avons conçu pour des enregistrements planifiés sur une seule journée (pauses comprises). Il **suit à la lettre les choix techniques détaillés dans cette section**. Son optimisation a porté sur des 1-grammes et 2-grammes de sandwichs.

« 3h_nphones » :

Il s'agit d'un script plus ancien, **représentatif de l'état de l'art**.

Il a été créé pour l'enregistrement de patients atteints de SLA, ou Sclérose Latérale Amyotrophique. Également appelée maladie de Charcot, cette affection grave s'accompagne d'une dégénérescence progressive des neurones moteurs. Il en résulte une perte de l'usage des muscles, qui débute souvent par l'appareil phonatoire. Des solutions de communication alternative à base de synthèse vocale sont alors proposées aux patients. A l'époque notre projet visait à créer les voix de synthèse personnalisées de plusieurs patients, en anticipant leur perte de parole éventuelle [Cadic 06].

Une expérience similaire avait été menée quelques années auparavant en japonais [Iida 01]. Seule une petite partie du script était obtenue par condensation, le reste étant composé de

Nom du script	**45min_nphones**	**3h_nphones**	**5h_nphones**	**5h_sandwiches**	**8days_nphones**
Nombre de phrases	547	610	1 532	2 860	7 919
Nombre de diphones	5 760	18 632	40 995	57 768	261 275
Nombre de sandwichs (cas `LRsemirobustes`)	2 973	6 949	15 351	22 406	99 374
Longueur moyenne des phrases (en diphones)	10.5	30.2	26.8	20.2	33
Durée prévisionnelle d'enregistrement	45 min	3 h	5 h	5 h	8 jours
Durée prévisionnelle de parole utile	8 min	25 min	55 min	1 h 15 min	6 h
Type de phrases	groupes nominaux	phrases complètes	phrases complètes	groupes de souffle	phrases complètes
Critère d'optimisation principal	diphones	diphones	diphones et triphones	sandwichs	diphones et triphones
Nombre indicatif de symboles contextuels	3	4	300	13	300
Taux de couverture 2-grammes de sandwichs, liquides fragiles, 13 contextes	5%	11%	17%	27%	33%
1-grammes de sandwichs, liquides fragiles, 13 contextes	14%	38%	55%	75%	75%
1-grammes de sandwichs, liquides semi-robustes, 13 contextes	18%	48%	69%	86%	85%
1-grammes de sandwichs, liquides semi-robustes, 4 contextes	31%	67%	80%	90%	90%
Diphones (4 symboles contextuels)	96%	98%	100%	100%	100%
Triphones (13 symboles contextuels)	20%	47%	69%	82%	88%

TABLE 10 – Inventaire des scripts obtenus par condensation.

phrases spécifiques à l'environnement personnel du locuteur. La composition de « 3h_nphones » accorde plus d'importance au procédé de condensation :

- 80 phrases couvrant les besoins quotidiens de patients SLA ainsi que les situations courantes.
- 36 phrases médicales contenant les termes-clés pour une bonne interaction entre le patient et les équipes médicales.
- 24 phrases apportant des listes de lettres, chiffres, années, etc., pour faciliter diverses épellations.
- 470 phrases obtenues par condensation d'un corpus de pioche. Elles ont pour but de compléter la couverture phonétique. Le critère de sélection, de type « fréquents d'abord », a porté dans une première phase sur les diphones en contexte (4 contextes [44]), puis sur les diphones hors contexte.
- Optionnel : 10 à 20 phrases ou mots choisis par le patient parmi ses expressions les plus courantes. Il s'agit surtout de noms propres personnalisés qui n'apparaissent pas dans le

44. Bien que concentrés également sur les mouvements prosodiques de fins de groupes intonatifs, les 4 symboles contextuels utilisés ici diffèrent légèrement de ceux présentés en 5.2.

reste du script : lieux spécifiques, noms de proches...

Le corpus de pioche utilisé pour cette condensation est moins fourni que celui présenté en 7.1. Il comporte 62 000 lignes (700 000 mots) issues principalement de comédies et sous-titres de séries TV. La sélection a été opérée sur des phrases complètes et non sur des groupes de souffle. Le script final comporte de nombreuses phrases longues qui nuisent à l'efficacité des enregistrements.

« 8days_nphones » :

C'est le script « historique » de la synthèse par corpus à France Télécom R&D (devenu Orange Labs) ; il est **représentatif de l'état de l'art**. Les voix Agnès et Philippe se basent dessus. Il a été obtenu par condensation, suivant une approche « en largeur » portant sur les diphones et triphones nouveaux, le corpus de pioche étant essentiellement composé d'archives du Monde (2 millions de mots environ). Les informations contextuelles ont porté sur tous les phonèmes et avec l'ensemble des marqueurs symboliques (plus de 300 classes au total).

« 5h_nphones » :

Nous avons construit ce script par condensation, à partir d'un corpus de pioche constitué du script de lecture « 8days_nphones ». « 5h_nphones » est donc un sous-script de « 8days_nphones » et s'inscrit dans la continuité de l'approche historique.

Nous avons logiquement utilisé pour cette condensation une stratégie similaire à celle suivie pour la création de « 8days_nphones ». A chaque itération, la phrase sélectionnée est celle apportant le plus de nouveaux diphones (approche en largeur). Les informations contextuelles sont conservées pour tous les phonèmes, voyelles ou consonnes, et de manière non regroupée (environ 300 combinaisons possibles).

Ceci nous permet donc de disposer d'un script **représentatif de l'état de l'art**, pour lequel plusieurs voix sont disponibles (en l'occurrence Agnès et Philippe), et de **durée d'enregistrement comparable à notre script « 5h_sandwiches »**.

« 45min_nphones » :

Ce script, créé en 2008 **conformément à l'état de l'art**, visait à couvrir les diphones de manière minimaliste, pour une durée d'enregistrement très réduite. Les phrases du Monde (23 millions de mots), unique corpus de pioche alors disponible, avaient été jugées trop longues et complexes pour cet usage. Les groupes nominaux et groupes prépositionnels en avaient alors été extraits de manière automatique, en se reposant sur l'analyse linguistique des hauts-niveaux. La souplesse de l'algorithme de condensation était ainsi accrue, tout en offrant une certaine cohérence à la lecture. Pour la sélection de ces groupes, deux approches avaient cohabité au sein d'une unique critère pondéré : une approche « fréquents d'abord » pour la couverture des diphones de fin de phrase (2 contextes, contour montant ou descendant) et une approche « rares d'abord » pour la couverture des diphones en contexte neutre (c'est-à-dire hors fins de phrases).

11 L'approche par construction de phrases

Dans la section précédente nous avons étudié la condensation de corpus. Cette technique, de mise en oeuvre rapide et aisée, donne accès à des scripts relativement denses, mais qui restent très en deçà de l'optimum. Sa sous-optimalité s'explique par le contenu limité du corpus de pioche, qui contraint fortement l'ensemble des possibles, d'autant plus que nous imposons des restrictions sur les longueurs de phrases. Aussi explorons-nous dans la présente section une nouvelle piste : elle consiste à élargir les possibilités de l'algorithme d'optimisation en mettant à sa disposition des phrases créées de toutes pièces, par recombinaison de portions de phrases issues du corpus de départ [Cadic 10a].

11.1 Fonctionnement

11.1.1 Principe

L'algorithme de création de script que nous détaillons ci-dessous se conforme aux choix généraux de la section 9. Il fonctionne donc par groupes de souffle, en interdisant ceux comportant plus de 15 sandwichs, et suit une approche gloutonne de type « fréquents d'abord ».

Son principe est de **composer artificiellement des enchaînements de sandwichs qui maximisent, à chaque itération, l'augmentation du VSCR**. Naturellement la pertinence des séquences construites relève de considérations lexicales, syntaxiques et sémantiques qui peuvent être complexes. Faut-il par exemple autoriser les phrases « *Le chat relit le canapé de la poussette.* » ou « *J'écris à mon rapport.* » ? La première présente une incohérence sémantique et la seconde une incohérence syntaxique qui risquent toutes deux de piéger le locuteur lors des enregistrements. Si le locuteur ou le superviseur relèvent l'erreur et reprennent la lecture de cette phrase, ceci se traduit simplement par une perte d'efficacité. S'ils ne relèvent pas l'erreur, des problèmes d'alignement risquent d'apparaître dans le processus automatique de création de voix et d'induire une perte de qualité de la synthèse finale. Il est donc important que les phrases construites respectent des critères minimaux de lisibilité.

Notre algorithme ne modélise que très partiellement ces critères langagiers, en exploitant des considérations statistiques sur les enchaînements de sandwichs. Une intervention humaine complémentaire permet d'orienter le processus de création de phrases vers des résultats acceptables. Il s'agit donc d'un **processus semi-automatique**.

11.1.2 L'automate de référence

D'une manière générale nous proposons de générer chaque séquence de sandwichs par une **recherche de chemin optimal dans une machine à états finie**, ou Finite State Machine (FSM) en anglais. Dans notre cas il s'agira plus exactement d'un automate transducteur pondéré, ou Weighted Finite State Transducer (WFST) [Allauzen 07].

Etant donné un semi-anneau [45] $(\mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1})$, un transducteur pondéré peut être défini comme un sextuplet [46] $(\Sigma, \Phi, Q, q^0, F, T)$ tel que :

45. Un semi-anneau est un ensemble $\mathbb{K}$ muni de deux lois d'opérations binaires $\oplus$ et $\otimes$, tels que $(\mathbb{K}, \oplus, \bar{0})$ est un monoïde commutatif, $(\mathbb{K}, \otimes, \bar{1})$ est un monoïde (pas forcément commutatif), $\otimes$ est distributive par rapport à $\oplus$ et $\forall x \in \mathbb{K}, \bar{0} \otimes x = x \otimes \bar{0} = \bar{0}$. En d'autres termes, un semi-anneau est un anneau auquel il peut manquer l'inversibilité de la somme.

46. Nous proposons ici une définition adaptée à notre usage. La définition générale autorise plusieurs états initiaux ainsi que des coûts spécifiques pour les états initiaux et finaux.

Σ est l'ensemble fini des symboles d'entrée (ou alphabet d'entrée) ;

Φ est l'ensemble fini des symboles de sortie (ou alphabet de sortie) ;

Q est l'ensemble fini des états (ou noeuds) ;

$q^0 \in Q$ est l'état initial ;

$F \subset Q$ est l'ensemble des états finaux ;

T est l'ensemble des transitions, $T \subset Q \times (\Sigma \cup \{\epsilon\}) \times (\Phi \cup \{\epsilon\}) \times \mathbb{K} \times Q$;

Il s'agit autrement dit d'un **graphe orienté pondéré**, pourvu d'un état initial q^0 et d'états finaux F, et dont chaque transition (ou arc) est munie d'un symbole d'entrée $\sigma \in \Sigma$, d'un symbole de sortie $\phi \in \Phi$ et d'un coût $k \in \mathbb{K}$. Le passage d'un état à l'autre est dicté par la séquence des symboles d'entrée. La machine produit alors une séquence de symboles de sortie, ou séquence traduite, d'où le nom de « transducteur ». Le symbole ϵ est en quelques sortes un « symbole vide » ; il permet de définir des ϵ-transitions, c'est-à-dire des transitions que la machine peut suivre sans consommer de symbole d'entrée. La figure 30 présente un exemple de transducteur pondéré. L'état initial est signalé par un entourage gras et les états finaux par un entourage double. Les étiquettes des arcs (comme « `a:v/0.7` ») indiquent le symbole d'entrée (« `a` ») auquel est associée la transition, le symbole de sortie (« `v` ») et le coût de la transition (0.7). Le transducteur illustré ici est non-déterministe, c'est-à-dire qu'un même symbole d'entrée peut être associé à plusieurs transitions partant du même état (en l'occurence le symbole « `b` » ouvre une alternative à partir de l'état 0).

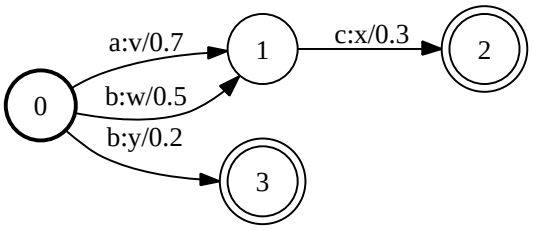


Figure 30 – Exemple de transducteur pondéré.

Conformément à une interprétation probabiliste, la loi $\otimes$ sert à calculer le coût d'un chemin dans le graphe (c'est-à-dire le coût d'une suite de transitions), tandis que la loi $\oplus$ sert à combiner les coûts de chemins « parallèles ». En particulier, $\oplus$ est nécessaire pour calculer le coût d'une séquence de symboles d'entrée dans le cas non-déterministe (plusieurs chemins possibles). Pour la définition des coûts nous utilisons un **semi-anneau $\mathbb{K}$ de type tropical**. Suivant ce choix les coûts sont des éléments de $\overline{\mathbb{R}}$ (nombres réels ou infinis), la loi $\otimes$ est la loi d'addition traditionnelle et la loi $\oplus$ est la fonction *minimum*. L'élément neutre $\bar{0}$ de $\oplus$ est $+\infty$ et l'élément neutre $\bar{1}$ de $\otimes$ est 0.

Plusieurs tels automates interviennent dans notre algorithme. Nous introduisons ci-dessous le premier d'entre eux, `WFST_REF`, **dont la topologie est dictée par les observations de séquences de sandwichs issues de notre corpus de référence**. Un état de l'automate `WFST_REF` correspond à un sandwich observé dans le corpus ; un arc entre états correspond à un bigramme de sandwichs également observé dans le corpus. La transition d'un état vers un autre est déclenchée par un symbole d'entrée, qui est ici le nom du sandwich final, et s'accompagne d'un symbole de sortie qui est le nom du bigramme. Le symbole de sortie permet entre autres de préciser le cluster consonantique qui sépare éventuellement les deux sandwichs successifs. On peut noter à ce sujet que `WFST_REF` est un automate non-déterministe, puisque deux bigrammes composés des mêmes sandwichs (donc mêmes noeuds de départ et d'arrivée) peuvent différer par leur cluster consonantique (donc arcs distincts), comme les bigrammes finaux [kat ʁ# <fin>] et [kat # <fin>] des exemples suivants :

Cent vingt-quatre. [#sɑ̃vɛ̃nkatʁ#]
Elle est délicate. [#ɛlɛdelikat#]

Le coût associé à un arc de `WFST_REF` est fixé à $(-\delta_{VSCR})$, où $\delta_{VSCR} \in [0;1]$ désigne **l'incrément de VSCR** offert soit par le sandwich final de l'arc, soit par le bigramme de l'arc, selon que l'on considère une couverture de sandwichs ou une couverture de bigrammes. On peut également utiliser une pondération des deux grandeurs pour un critère d'optimisation mixte « sandwichs+bigrammes ». Dans tous les cas le coût associé à un arc dépend de l'état du script de lecture en cours de création et l'automate évolue après chaque ajout de phrase dans ce script : les coûts des arcs dont le sandwich final (ou le bigramme) vient d'être couvert sont modifiés pour prendre la valeur maximale 0. Ceci traduit le fait qu'un nouvel ajout de ces sandwichs (ou bigrammes) dans les phrases suivantes du script n'améliore pas la couverture globale.

La figure 31 présente l'automate `WFST_REF` [47] correspondant à un corpus de référence simpliste de seulement trois phrases :

C'est un essai. [#sɛtœ̃nesɛ#]
Juste un exemple. [#ʒystœ̃negzãpl#]
Une esquisse. [#yneskis#]

On constate en particulier que l'automate autorise, outre les phrases du corpus de référence, de nouvelles phrases tout à fait valables : [#sɛtœ̃negzãpl#] (« *C'est un exemple.* ») et [#ʒystœ̃nesɛ#] (« *Juste un essai.* »). Mais il autorise également des séquences inacceptables : [#ynesɛ#] (« *Une essai.* »), [#sɛtœ̃neskis#] (« *C'est un esquisse.* »), ou encore [#ʒystœ̃neskis#] (« *Juste un esquisse.* »).

Dans la pratique `WFST_REF` est une structure bien plus vaste, possédant autant de noeuds qu'il y a de sandwichs dans le corpus de référence et autant d'arcs qu'il y a de bigrammes, soit 93 173 noeuds et 859 427 arcs dans le cas `LRsemirobustes_13contextes`.

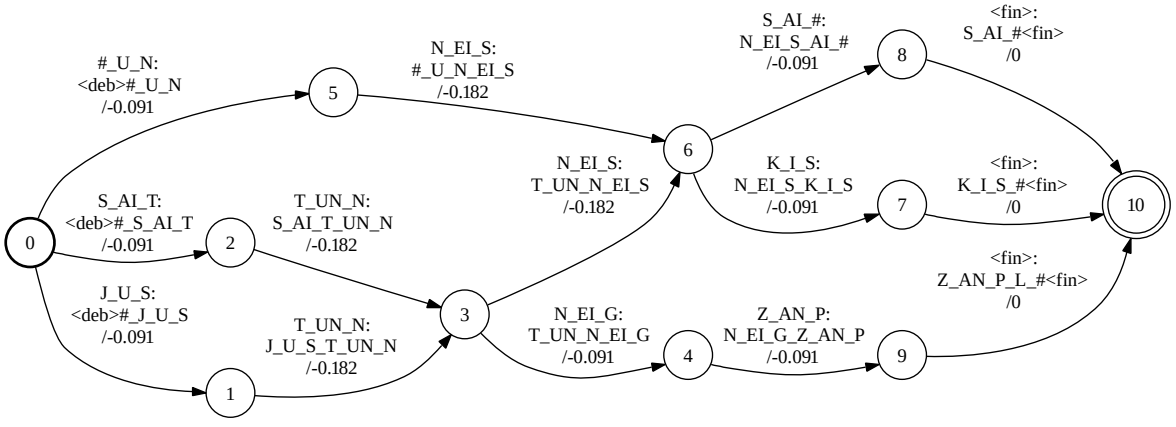


FIGURE 31 – Exemple d'automate `WFST_REF` construit à partir d'un corpus de référence comportant uniquement trois phrases. Pour simplifier la figure nous avons volontairement omis les symboles contextuels attachés aux différents sandwichs.

De par notre définition des coûts de transition, la recherche de séquences de sandwichs qui maximisent l'incrément de VSCR est étroitement liée à la recherche de chemins de coût minimum au sein de `WFST_REF`. Toutefois la structure de `WFST_REF` ne nous permet pas d'envisager directement ce type d'optimisation. En effet, conformément à notre analyse générale de la section 9.2, la négativité des coûts fait que ce sont les chemins les plus longs qui minimisent

47. Par commodité nous utilisons dans les représentations d'automates un alphabet phonétique interne, plus lisible que l'API.

le coût global. Mais contrairement à l'approche par condensation, la longueur des chemins de notre automate n'est pas bornée : un tronçon de phrase peut toujours être prolongé par un autre tronçon de phrase en « rebondissant » sur un sandwich commun (même si ce n'est pas le cas dans notre exemple simplifié de la figure 31). La recherche du meilleur chemin ne peut donc pas aboutir.

Il est possible de forcer la convergence en ajoutant un même offset $\Delta \geq 1$ à toutes les transitions. Les coûts devenant tous positifs, la convergence est assurée et la solution optimale sera d'autant plus courte que Δ sera élevé. Mais le lien entre la valeur de Δ, la longueur de la séquence optimale et l'incrément de VSCR est mal maîtrisé.

Un algorithme de minimisation du coût moyen peut également être appliqué, comme par exemple celui proposé dans [Rozenknop 01]. Cet algorithme itératif consiste à rechercher le meilleur chemin dans une succession d'automates. Les automates sont déduits de `WFST_REF` par soustraction d'un offset qui évolue au fil des itérations. Les chemins optimaux successifs (du moins à partir du deuxième) ont la propriété de décroître strictement en longueur, jusqu'à converger vers le chemin optimal au sens du coût moyen. Cet algorithme est très efficace : en général quelques itérations suffisent.

Cependant des cas de divergence peuvent à nouveau apparaître, du fait de la présence de cycles dans notre automate. Rien n'interdit en effet à un chemin de passer plusieurs fois par le même sandwich (ou bigramme) et de former ainsi une boucle. Si par hasard une telle boucle devient prioritaire pour l'augmentation du VSCR, alors l'algorithme de minimisation du coût moyen a de fortes chances de la répéter à l'infini et donc de diverger. Concrètement cela se traduit par l'apparition d'un cycle de coût négatif dans l'automate après la première itération de l'algorithme. Nous avons rencontré ce cas de figure plusieurs fois au fil de nos expérimentations. Un exemple typique porte sur la phrase « *Bisous bisous bisous.* » qui est présente de nombreuses fois dans le corpus de référence. Par construction, `WFST_REF` contient alors un cycle de coût très faible qui autorise des chemins du type « *Bisous bisous bisous bisous bisous bisous bisous...* ». La figure 32 illustre cet exemple. On peut noter au passage que les noeuds 1 et 3 sont associés au même sandwich [biz], mais qu'ils ne sont pas assimilés dans l'automate car leurs symboles contextuels (non reportés sur la figure) diffèrent.

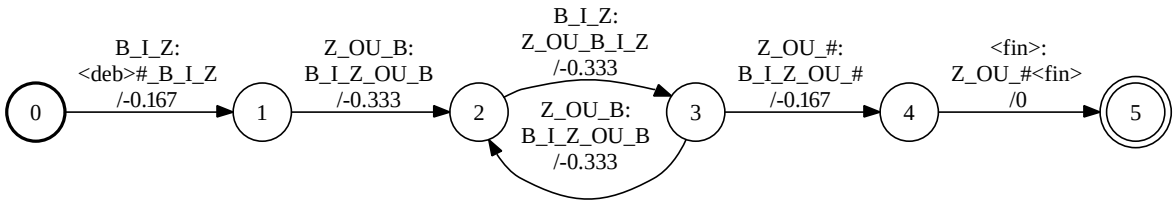


Figure 32 – Exemple de cycle, causé par la phrase « *Bisous bisous bisous.* » du corpus de référence.

Le rôle de tels cycles dans la minimisation du coût moyen montre une faiblesse de notre automate concernant la modélisation des phénomènes de couverture. Si la répétition d'un sandwich (ou bigramme) au sein d'une même séquence n'est pas préjudiciable en soi, le VSCR ne devrait théoriquement bénéficier que de la première occurrence. Malheureusement les coûts des transitions de notre automate ne peuvent être mis à jour qu'entre deux phrases du script en cours de création ; les évolutions de la couverture ne sont donc pas prises en compte au sein d'un même chemin.

Cette limitation structurelle de `WFST_REF` rend délicate toute tentative d'extraction de che-

min optimal (et pas uniquement avec le critère de coût moyen). Pour garantir la convergence de notre procédé et offrir un contrôle fin sur la longueur des phrases, nous devons introduire des contraintes topologiques supplémentaires.

11.1.3 Automates avec contrainte de longueur

Les observations précédentes nous amènent à définir de nouveaux automates, obtenus par composition de `WFST_REF` avec des automates de contrainte. La composition est une opération non symétrique qui peut être appliquée à deux transducteurs pondérés A_1 et A_2, lorsqu'ils reposent sur le même semi-anneau $\mathbb{K}$ commutatif et que l'alphabet de sortie du premier coïncide avec l'alphabet d'entrée du second. Si $A_1 = (\Sigma_1, \Theta, Q_1, q_1^0, F_1, T_1)$ et $A_2 = (\Theta, \Phi_2, Q_2, q_2^0, F_2, T_2)$, alors le transducteur composé $A_2 \circ A_1$ est défini comme suit, pour toute séquence d'entrée $x \in \Sigma_1^*$ et séquence de sortie $y \in \Phi_2^*$:

$$|A_2 \circ A_1|(x, y) = \bigoplus_{z \in \Theta^*} |A_1|(x, z) \otimes |A_2|(z, y) \tag{14}$$

où $|A|(a, b)$ désigne le coût de l'ensemble des chemins de A allant de l'état initial à un état final en suivant la séquence d'entrée a et en produisant la séquence de sortie b. La définition 14 suffit à déterminer la structure de $A_2 \circ A_1 = (\Sigma_{2\circ1}, \Phi_{2\circ1}, Q_{2\circ1}, q_{2\circ1}^0, F_{2\circ1}, T_{2\circ1})$, avec notamment :

- $\Sigma_{2\circ1} \subset \Sigma_1$
- $\Phi_{2\circ1} \subset \Phi_2$
- $Q_{2\circ1} \subset Q_1 \times Q_2$
- $q_{2\circ1}^0 = (q_1^0, q_2^0)$
- $F_{2\circ1} \subset F_1 \times F_2$
- $T_{2\circ1} \subset (Q_1 \times Q_2) \times (\Sigma_1 \cup \{\epsilon\}) \times (\Phi_2 \cup \{\epsilon\}) \times \mathbb{K} \times (Q_1 \times Q_2)$

La figure 33 donne un exemple d'automates A_1, A_2 et $A_2 \circ A_1$. Cet exemple est inspiré de [Mohri 02b].

Cette loi de composition nous permet d'introduire des **automates avec contrainte de longueur**. Étant donné la longueur maximale de 15 sandwichs choisie en 9.2, nous définissons :

$$\left\{\begin{array}{lcll}
\texttt{WFST_REF_1} & = & \texttt{LONGUEUR_1} & \circ \ \texttt{WFST_REF} \\
\texttt{WFST_REF_2} & = & \texttt{LONGUEUR_2} & \circ \ \texttt{WFST_REF} \\
\texttt{WFST_REF_3} & = & \texttt{LONGUEUR_3} & \circ \ \texttt{WFST_REF} \\
\ldots & & \ldots & \ldots \\
\texttt{WFST_REF_13} & = & \texttt{LONGUEUR_13} & \circ \ \texttt{WFST_REF} \\
\texttt{WFST_REF_14} & = & \texttt{LONGUEUR_14} & \circ \ \texttt{WFST_REF} \\
\texttt{WFST_REF_15} & = & \texttt{LONGUEUR_15} & \circ \ \texttt{WFST_REF}
\end{array}\right. \tag{15}$$

où les automates `LONGUEUR_L` servent à contraindre la longueur totale des chemins à L sandwichs (sans compter les sandwichs virtuels <*deb*> et <*fin*>). Il s'agit simplement d'**accepteurs non pondérés**, c'est-à-dire de transducteurs dont les symboles de sortie sont égaux aux symboles d'entrée et dont les coûts de transition sont tous nuls. La figure 34 montre la structure de `LONGUEUR_L`. Il est constitué de $L + 2$ états, qui représentent en quelque sorte les points de passage obligé de toute séquence. Chaque maillon accepte la totalité des N sandwichs existants[48], ce qui se traduit par N transitions possibles entre deux noeuds successifs. Sur la figure les N sandwichs distincts relevés dans le corpus de référence sont désignés par $Sand(n)$

48. $N = 93173$ dans le cas `LRsemirobustes_13contextes`

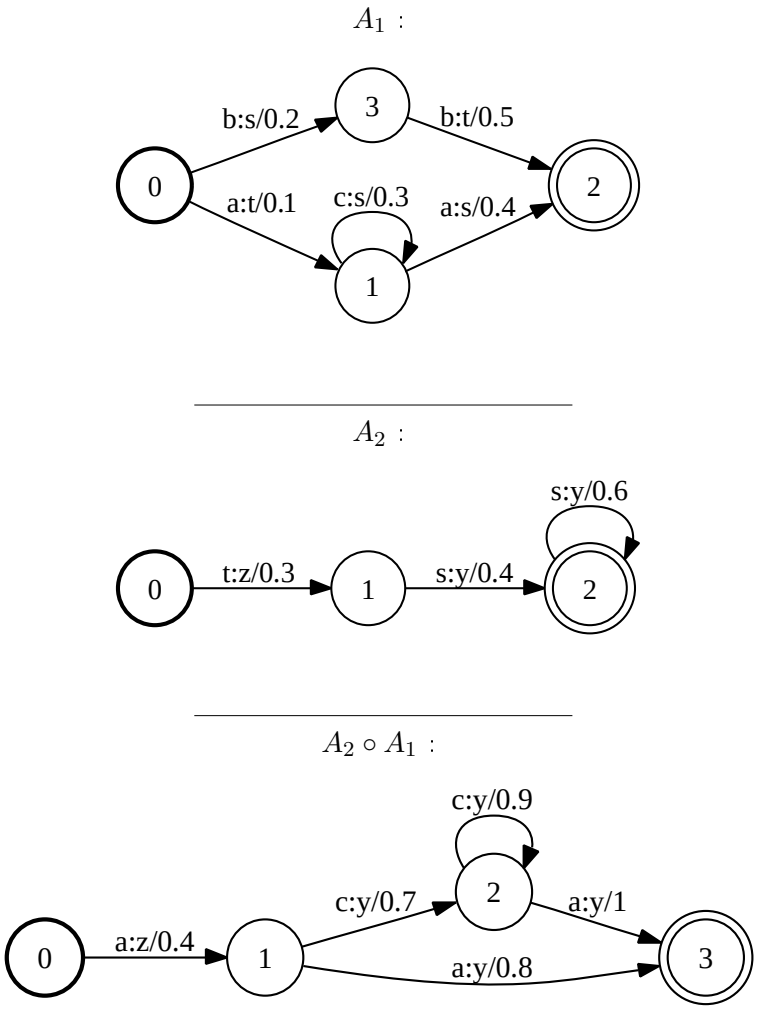


FIGURE 33 – Exemple de composition d'automates.

pour $n \in [\![1; N]\!]$. La transition entre les deux derniers états a toutefois un statut particulier, puisqu'elle est associée au sandwich virtuel <*fin*> qui clôture obligatoirement toutes les séquences [49].

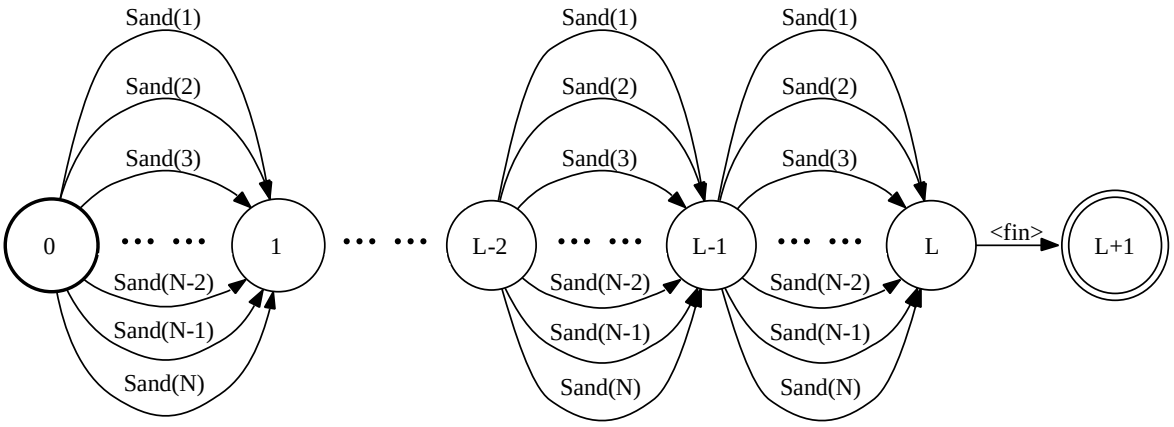


FIGURE 34 – Structure de l'accepteur `LONGUEUR_L`.

49. Nous aurions pu conserver les N sandwichs en parallèle de cette dernière transition, tout comme nous aurions pu ajouter le sandwich virtuel <*fin*> aux maillons précédents du graphe ; ceci n'aurait pas modifié l'automate composé, les chemins supplémentaires autorisés ne trouvant aucune correspondance dans l'automate de référence.

Si nous reprenons l'exemple simplifié de la figure 31, nous constatons que les seuls chemins possibles sont de longueur 3 ou 4 (hors sandwichs virtuels). En particulier il n'y a pas de cycle autorisant la prolongation des chemins à l'infini. Par conséquent seuls les automates contraints `WFST_REF_3` et `WFST_REF_4` sont non-vides. Nous les avons représentés en figure 35.

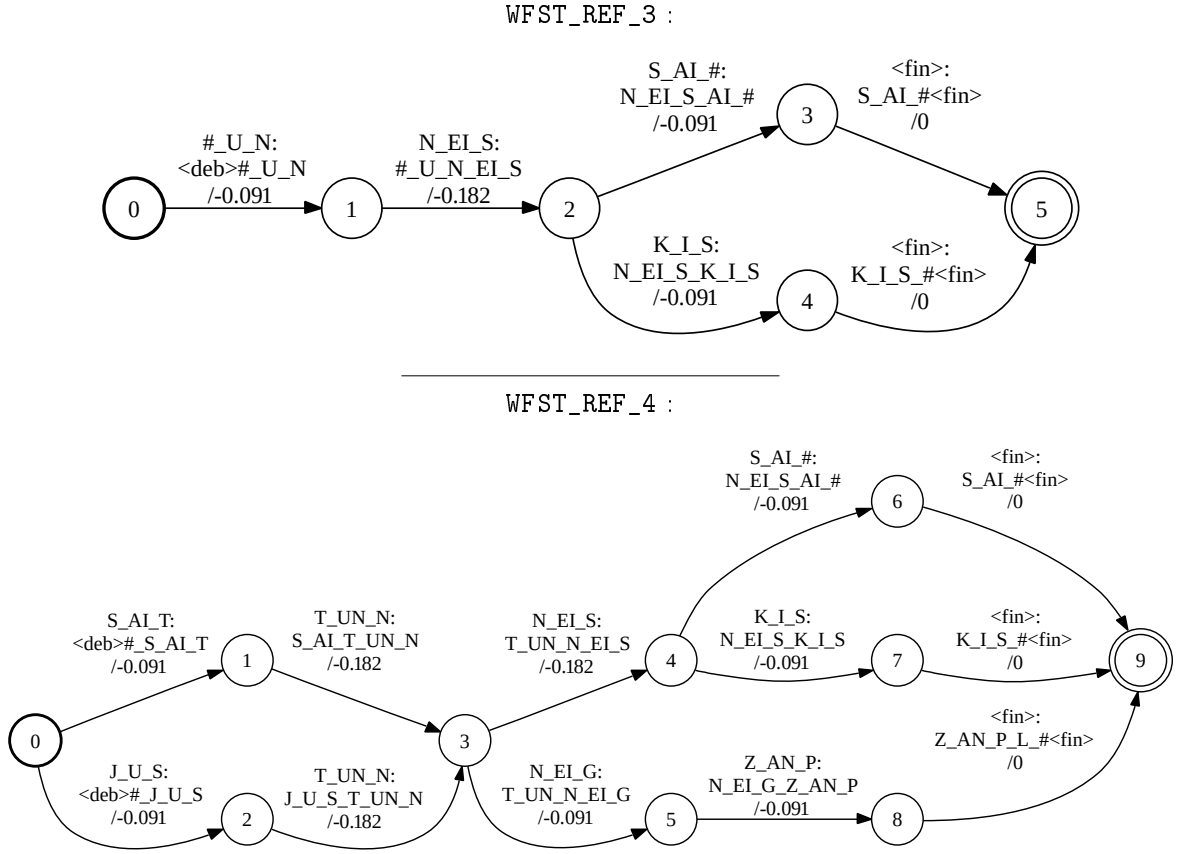


FIGURE 35 – Exemples d'automates de référence à longueur contrainte (L=3 et 4).

La recherche des meilleurs chemins dans `WFST_REF_3` et `WFST_REF_4` nous donne respectivement les séquences de sandwichs de longueurs 3 et 4 qui maximisent le VSCR. En l'occurence `WFST_REF_3` autorise deux chemins de coûts identiques :

[#ynesɛ#]	(« *Une essai.* »)	→ coût de -0.364
[#yneskis#]	(« *Une esquisse.* »)	→ coût de -0.364

On constate au passage qu'un seul de ces deux chemins est grammaticalement correct. L'automate `WFST_REF_4` autorise quand à lui six chemins :

[#sɛtœ̃nesɛ#]	(« *C'est un essai.* »)	→ coût de -0.546
[#sɛtœ̃neskis#]	(« *C'est un esquisse.* »)	→ coût de -0.546
[#ʒystœ̃nesɛ#]	(« *Juste un essai.* »)	→ coût de -0.546
[#ʒystœ̃neskis#]	(« *Juste un esquisse.* »)	→ coût de -0.546
[#sɛtœ̃negzãpl#]	(« *C'est un exemple.* »)	→ coût de -0.455
[#ʒystœ̃negzãpl#]	(« *Juste un exemple.* »)	→ coût de -0.455

Parmi ces six chemins, le coût minimal de -0.546 est atteint par les quatre premiers, dont deux sont grammaticalement corrects.

Nous reposant sur l'analyse de la section 9.2, nous retenons le chemin qui offre le meilleur incrément de VSCR dans la limite de 15 sandwichs. Comme $-0.546 < -0.364$, il s'agit dans notre cas de n'importe lequel des quatre chemins optimaux de longueur 4. Notons que le résultat aurait été le même au sens du coût moyen, puisque $\frac{-0.546}{4} < \frac{-0.364}{3}$.

D'une manière générale, imaginons qu'un script de N phrases ait déjà été constitué et que le corpus de référence complet du paragraphe 7.1 soit utilisé. Grâce à une mise à jour systématique des coûts de transition après chaque création de phrase, les 15 automates `WFST_REF_L` tiennent compte des N phrases existantes. Il s'agit désormais de créer la $(N+1)^{\text{ème}}$ phrase. Les recherches de meilleurs chemins dans chacun des automates aboutissent à 15 séquences optimales de sandwichs [50], pour chacune des 15 longueurs permises. Parmi ces 15 séquences, on retient celle qui offre le coût minimal (ou bien le coût minimal par sandwich si cette stratégie est retenue malgré les recommandations de la section 9.2).

Les automates de référence à longueur contrainte nous offrent une grande souplesse de mise en oeuvre. Tout d'abord, leur caractère acyclique garantit l'existence d'un chemin optimal. Ensuite ils permettent une gestion libre et précise des effets de longueurs, grâce à laquelle nous pouvons trivialement nous conformer aux conclusions de la section 9.2, mais aussi expérimenter la minimisation du coût moyen (voir page 130). Enfin ces automates peuvent facilement être distribués sur des mémoires et processeurs distincts afin de paralléliser les calculs et ainsi accélérer l'optimisation.

11.1.4 L'intervention manuelle

La simple recherche de chemins optimaux dans les automates `WFST_REF_L` permet de construire des séquences de sandwichs extrêmement denses sur le plan de la couverture. Bien entendu, l'optimalité des séquences est relative aux contraintes imposées, à savoir l'utilisation de sandwichs et bigrammes existants et fréquents (suivant la définition retenue pour les coûts de transition). Ces contraintes ne garantissent pas la lisibilité des phrases issues du processus automatique ; aussi la plupart d'entre elles ne peuvent pas être ajoutées telles quelles au script de lecture. Nous avons à ce sujet relevé un peu plus haut quelques incohérences grammaticales sur les chemins optimaux extraits des `WFST_REF_L`. Mais il s'agit plus souvent de séquences phonétiques qui ne peuvent même pas être transcrites textuellement. C'est le cas par exemple de la séquence [dãlaʒuʁdɥi#] : faut-il la transcrire « *dans la jourd'hui* » ? Cette séquence est lexicalement incorrecte, bien que composée de sandwichs et bigrammes très fréquents. En effet les trois premiers sandwichs ([dãl], [laʒ] et [ʒuʁ]) correspondent au début de « *dans la journée* », tandis que les troisième et dernier sandwichs ([ʒuʁ] et [dɥi#]) correspondent à la fin de « *aujourd'hui* ».

Dans notre approche, nous avons recours à une intervention experte pour améliorer la lisibilité des phrases construites. Tout au long du processus de construction d'un script, un opérateur humain intervient itérativement sur les séquences optimales en ajoutant des contraintes dans le système d'automates `WFST_REF_L`. Tout en restant guidé vers les sandwichs non-couverts les plus fréquents, l'opérateur parvient ainsi à produire des séquences lisibles, c'est-à-dire qui peuvent être transcrites textuellement sans présenter de piège pour les locuteurs futurs. En particulier les néologismes, erreurs grammaticales ou noms propres complexes peuvent être exclus. Une cohérence sémantique absolue n'est toutefois pas obligatoire. On tolérera par exemple la phrase « *La petite souris mange le gros chat.* », tandis que la phrase « *Le gros chat mange petite souris.* » comporte un piège inacceptable qui risque de tromper les locuteurs (absence de l'article « *la* »). Les critères de validité d'une phrase restent à la discrétion de l'opérateur. Nous avons tenté de les objectiver en donnant la consigne générale suivante : « il doit exister un contexte réaliste dans lequel la phrase aurait un sens ». Mais l'expérience montre que la frontière reste floue et les décisions souvent subjectives.

50. Avec le corpus de référence complet, aucun des automates `WFST_REF_L` n'est vide et nous disposons bien à chaque itération de 15 chemins optimaux, contrairement à l'exemple simpliste précédent.

Pour décrire les moyens d'action offerts à l'opérateur, considérons la figure 36. La ligne 1 donne un exemple de séquence produite automatiquement, par recherche du meilleur chemin dans les 15 automates `WFST_REF_L`. L'outil que nous avons développé (et dont les détails techniques de l'implémentation seront donnés plus bas) présente cette séquence initiale à l'opérateur. Ce dernier doit l'interpréter à partir des seules données phonétiques et contextuelles, ce qui nécessite évidemment une bonne maîtrise des deux notations. Une fonction de vocalisation est disponible pour accélérer le déchiffrage, mais en pratique quelques heures d'entraînement suffisent à être à l'aise.

Dans l'exemple, la phrase initiale ne satisfait pas les critères de lisibilité textuelle : « *Je ne la semaine des six,* ». L'opérateur engage donc un processus de modification. Pour cela il doit **identifier un tronçon de la séquence initiale qui lui semble « prometteur »**. Il s'agit d'une sous-séquence la plus large possible, porteuse d'un embryon de sens et qui pourrait s'intégrer aisément dans une phrase valide. Une telle sous-séquence existe toujours, du fait des contraintes de sandwichs, bigrammes et symboles contextuels, qui imposent au moins une cohérence à moyen-terme dans la séquence initiale. Sur la ligne 1 de la figure 36, le tronçon choisi par l'opérateur est encadré. Nous en donnons à titre indicatif l'équivalent textuel, qui n'est pas fourni par nos outils : « *Je ne la s...* ». Les 15 automates `WFST_REF_L` sont alors modifiés comme suit :

- l'état initial est positionné à la fin de la sous-séquence choisie par l'opérateur ;
- la transition avec le sandwich suivant est interdite (plus précisément tous les arcs dont le symbole de sortie correspond au bigramme $\left[\begin{smallmatrix}25 & 43\\ \mathbf{l\,a} & \mathbf{s\,\emptyset\,m}\end{smallmatrix}\right]$).

Certains des automates se retrouvent naturellement hors jeu car ils ne permettent pas de suivre la sous-séquence choisie. Dans notre exemple c'est au moins le cas de `WFST_REF_1` et `WFST_REF_2`, puisque la sous-séquence retenue comporte 3 sandwichs. Mais les contraintes topologiques liées aux bigrammes et symboles contextuels font que d'autres automates plus longs peuvent être concernés. Dans tous les cas il reste au moins l'automate qui est à l'origine de la séquence initiale (`WFST_REF_7` dans l'exemple).

Une nouvelle séquence est ensuite calculée, suivant le même procédé que la première. Elle est **presque optimale**, dans la mesure où les corrections apportées à l'automate sont a priori peu pénalisantes. Il s'agit en quelque sorte de la **deuxième meilleure séquence** incluant la sous-séquence choisie par l'opérateur. Seule la fin a changé, comme illustré sur la ligne 2 de la figure 36.

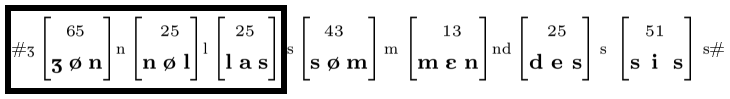
Et ainsi de suite, jusqu'à ce que la séquence satisfasse les critères de lisibilité (étapes 3 à 6 de la figure). Pour valider la séquence finale, l'opérateur doit en saisir une transcription textuelle. Le système analyse automatiquement cette transcription pour vérifier la concordance avec la séquence attendue de phonèmes et de symboles contextuels. Si une différence est constatée, cela peut signifier deux choses : soit l'opérateur a mal interprété la séquence, soit le système a mal analysé la transcription textuelle proposée par l'opérateur. Ce dernier doit donc réviser sa transcription textuelle ou bien reprendre le processus de modification. Des fonctions de remise à zéro et d'annulation des dernières actions sont disponibles.

Dès que l'opérateur est parvenu à valider la phrase, celle-ci est ajoutée au script de lecture et les automates évoluent de la manière suivante :

- l'état initial de chaque automate est repositionné sur le noeud de départ ;
- les transitions momentanément interdites sont réhabilitées ;
- les coûts de transition sont mis à jour en fonction de la couverture apportée par la nouvelle phrase.

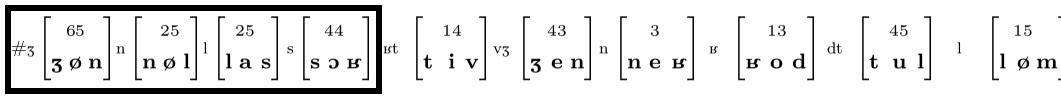
Nous avons illustré ci-dessus le cas où l'opérateur choisit de conserver le début de la séquence. Mais **l'opération inverse est également possible, à savoir le blocage de la fin**

1.



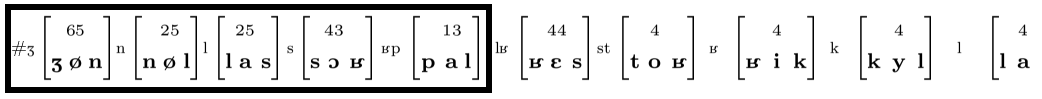
Je ne la s...

2.



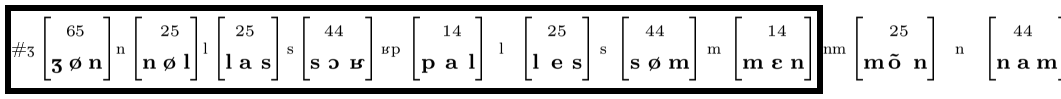
Je ne la sors...

3.



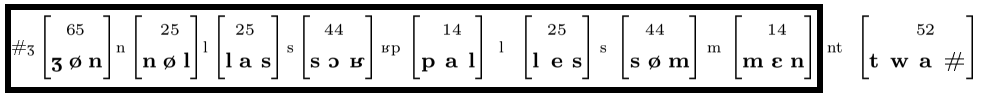
Je ne la sors pas l...

4.



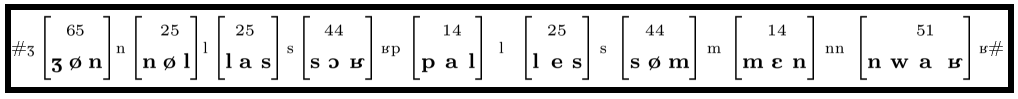
Je ne la sors pas les semaines...

5.



Je ne la sors pas les semaines...

6.



Je ne la sors pas les semaines noires,

FIGURE 36 – Exemple d'application du processus semi-automatique de création de phrase. A chaque étape, l'opérateur retient une portion de phrase « prometteuse » et demande la génération automatique d'un nouvel environnement quasi-optimal.

et la génération automatique d'un nouveau début. L'alternance entre les deux stratégies permet alors de conserver n'importe quelle sous-séquence. Pour cela on introduit de nouveaux automates `WFST_REF_L`$^{-1}$, obtenus par simple **renversement** des automates `WFST_REF_L`. Cette transformation est triviale du fait que les automates `WFST_REF_L` comportent un unique état initial et un unique état final[51] : il suffit d'échanger, dans chaque automate, les états de départ et d'arrivée de tous les arcs. L'utilisation des `WFST_REF_L`$^{-1}$ est ensuite équivalente à celle des `WFST_REF_L`. Lorsque l'état initial est inchangé on obtient la même séquence optimale (juste renversée), et de manière parfaitement symétrique le déplacement de l'état initial et l'interdiction de certaines transitions permet de bloquer une fin de séquence et de régénérer le début. Notons que les transitions interdites par l'opérateur doivent être impactées à l'identique dans les deux classes d'automates, quel que soit le sens choisi pour les modifications. Les automates `WFST_REF_L` et `WFST_REF_L`$^{-1}$ évoluent donc en parallèle.

À l'instar des lignes 2 à 4 de la figure 36, la plupart des chemins générés par le système d'automates s'étendent sur la longueur maximale, soit 15 sandwichs. C'est une conséquence logique de l'utilisation d'un critère de couverture non pondéré par la longueur de phrase. Cela peut devenir problématique lorsqu'une courte sous-séquence est presque satisfaisante du point de vue de l'opérateur, mais que toutes les régénérations lui accolent un environnement trop large. L'opérateur ne parvient pas à la clore. Pour ces situations nous avons mis en place un mécanisme de « **forçage d'une terminaison de phrase** ». Sur la demande de l'opérateur, le système force ainsi la génération d'un court début ou d'une courte fin de phrase, selon que la régénération se fait respectivement vers la gauche ou vers la droite. Cette option ne doit pas devenir une solution de facilité pour l'opérateur, qui n'est censé l'utiliser que lorsqu'il se trouve dans une réelle impasse liée à la longueur des séquences générées. Le forçage est obtenu en pénalisant simplement les séquences optimales issues des différents automates, de manière proportionnelle à leur longueur :

$$\forall L \in [\![1;15]\!], coût_{pond}(\widehat{C_L}) = coût(\widehat{C_L}) + \alpha * L \tag{16}$$

où $\widehat{C_L}$ désigne le meilleur chemin de `WFST_REF_L` (ou `WFST_REF_L`$^{-1}$ dans le cas d'une régénération vers la gauche), α un coefficient constant que nous prenons égal à 0.2, $coût$ la fonction de coût de l'automate et $coût_{pond}$ la fonction de coût pondérée. Dans le cas d'un forçage de terminaison de phrase, nous utilisons $coût_{pond}$ au lieu de $coût$ pour élire le chemin optimal parmi les 15 meilleurs chemins.

Rappelons pour finir que, bien que ces automates soient tous acycliques, la répétition d'un sandwich au sein d'une même séquence reste possible. On n'est donc pas à l'abri d'un **bouclage momentané sur une séquence de sandwichs** particulièrement intéressante pour l'augmentation du VSCR. Certes, la longueur de la séquence globale est contrôlée par l'automate, donc un tel bouclage est toujours borné et il n'y a aucun risque de divergence. Néanmoins cela peut avoir un impact négatif sur le VSCR puisque, comme nous l'avons déjà expliqué, l'évolution de la couverture n'est pas prise en compte au sein d'un même chemin. **L'intervention manuelle que nous venons d'exposer constitue un rempart naturel très efficace contre cette faiblesse algorithmique**. L'opérateur a en effet une tendance naturelle à éliminer les redondances de la phrase... surtout si les consignes l'y invitent.

51. Sans cette condition des ϵ-transitions apparaîtraient dans les automates renversés, sans gravité toutefois pour le fonctionnement de notre algorithme.

11.1.5 Réalisation technique

Architecture logicielle

Pour permettre la création rapide de scripts de lecture notre outil est capable de répartir le travail entre plusieurs opérateurs, suivant une approche de type clients-serveur. Par ailleurs la charge de calculs et l'occupation mémoire liées aux (gros) automates ont été distribuées sur plusieurs machines, afin d'offrir une réactivité suffisante à chacun des opérateurs. De ces choix découle l'architecture logicielle résumée en figure 37.

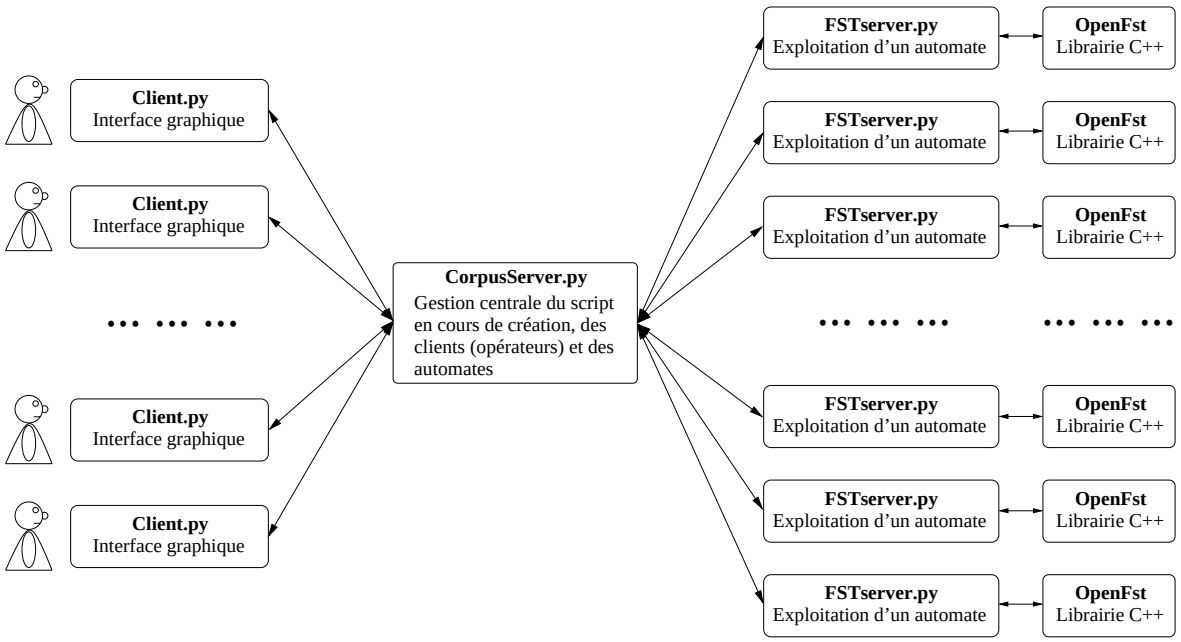


FIGURE 37 – Architecture logicielle de notre outil de construction de phrases.

Les différents composants seront détaillés dans les pages suivantes. La plupart d'entre eux sont implémentés en Python[52], d'où les extensions *py*. De manière très synthétique :

- chacun des automates impliqués dans le processus est géré par une instance de `FSTserver.py`, qui est notamment chargée d'appeler les fonctions C++ de la librairie OpenFst ;
- `Client.py` est l'outil graphique mis à la disposition des opérateurs pour intervenir sur les phrases construites ;
- `CorpusServer.py` centralise tous les échanges entre les automates et les opérateurs.

Chaque phrase du script final est créée par un et un seul opérateur. L'attribution automatique des numéros de phrases est gérée par `CorpusServer.py`.

Création et exploitation des automates avec OpenFst et `FSTserver.py`

Pour la gestion des automates, nous utilisons la librairie OpenFst [Allauzen 07]. Dédiée aux transducteurs pondérés à états finis, elle offre de nombreuses fonctionnalités disponibles sous forme de fonctions C++ ou de commandes shell. En particulier la commande `fstcompile` permet de convertir une description textuelle d'automate en une structure binaire optimisée et utilisable par les autres fonctions. Les commandes `fstcompose` et `fstreverse` implémentent

52. http://www.python.org/

respectivement les opérations de composition[53] et d'inversion. La création de nos automates `WFST_REF_L` et `WFST_REF_L`$^{-1}$, pour $L \in [\![1; 15]\!]$, est donc particulièrement aisée. Nous obtenons 30 fichiers, dont les tailles s'étendent de 160 Ko pour les plus petits ($L = 1$) à 120 Mo pour les plus grands ($L = 15$).

Les commandes shell nécessitant un rechargement systématique des automates en mémoire, elles se révèlent inadaptées à une mise en oeuvre en temps réel. Pour l'exploitation des automates nous préférons appeler directement les fonctions C++ depuis notre script Python `FSTserver.py`, en nous reposant sur la Python/C API [Foundation 10]. Ceci évite de répéter inutilement les chargements en mémoire et accès disques.

La recherche des meilleurs chemins est effectuée avec la fonction `ShortestPath` de la librairie. Basée sur l'algorithme Generic-Single-Source-Shortest-Distance (GSSSD) [Mohri 02a], cette fonction se montre particulièrement souple et performante. L'algorithme GSSSD fait des hypothèses moins fortes sur le contenu des automates que les algorithmes d'optimisation classiques. Il propose une version unifiée des différentes heuristiques de parcours de graphe (ou *queue disciplines*), la plupart des algorithmes classiques pouvant alors être vus comme des choix d'heuristiques spécifiques. Par exemple l'algorithme de Dijkstra, qui est adapté aux graphes de coûts positifs, correspond à une instanciation de l'algorithme GSSSD avec l'heuristique « plus court d'abord » (*shortest-first order*). Pour nos automates `WFST_REF_L` et `WFST_REF_L`$^{-1}$, qui ne sont pas positifs mais ont la propriété d'être acycliques, un parcours fondé sur l'**ordre topologique** est plus pertinent. La commande `fsttopsort` permet de renuméroter les états de chaque automate suivant l'ordre topologique, c'est-à-dire de façon à ce que toutes les transitions suivent des numéros d'états croissants. Il s'agit d'un pré-calcul peu complexe, réalisé off-line, qui permet d'accélérer significativement les recherches ultérieures de chemins optimaux : lorsque l'automate est trié topologiquement, la fonction `ShortestPath` s'exécute en un temps linéaire, contre un temps exponentiel dans le cas général !

Pour permettre la mise à jour des automates au fil du processus de création de phrases, nous avons dû ajouter à la librairie OpenFst nos propres fonctions C++ :

- `setstartfrompath(path)` suit le chemin `path` à partir de l'état initial et marque le noeud d'arrivée comme le nouvel état initial de l'automate.
- `setstartfromstate(state)` marque le noeud `state` comme le nouvel état initial de l'automate. Cette fonction sert essentiellement à rétablir le noeud initial après une application de `setstartfrompath`.
- `addilabelweights(isymbol, delta)` ajoute la valeur `delta` aux coûts de toutes les transitions associées au symbole d'entrée `isymbol`. Elle permet notamment de mettre à jour les automates en tenant compte des sandwichs nouvellement couverts.
- `addolabelweights(osymbol, delta)` ajoute la valeur `delta` aux coûts de toutes les transitions associées au symbole de sortie `osymbol`. Elle permet non seulement la mise à jour des automates en fonction des bigrammes nouvellement couverts (lorsque le critère d'optimisation tient compte de cette couverture), mais également l'interdiction du bigramme pointé par l'opérateur lors du choix d'une sous-séquence. Cette interdiction temporaire est simplement obtenue en ajoutant une pénalité rédhibitoire (+100) à toutes les transitions ayant ce bigramme comme symbole de sortie ; leur réhabilitation se fait ensuite par soustraction du même montant.

L'interface graphique `client.py`

Le script `client.py` a été développé avec le module Tkinter de Python, qui fournit une adaptation de la bibliothèque graphique Tk. La figure 38 montre une capture d'écran de cet

53. Les arcs des automates à composer doivent préalablement être ordonnés avec `fstarcsort`.

outil, dans la situation initiale de la figure 36. L'interface affiche à la fois la séquence phonétique et la séquence de sandwichs ; l'alphabet phonétique et les symboles contextuels suivent des notations internes, choisies en concertation avec les opérateurs. Le bouton `PLAY` vocalise la séquence courante. Le bouton `ANNULER` annule la dernière action. Le bouton `RESET` provoque la remise à zéro de la séquence : toutes les modifications sont annulées et on revient à la séquence initiale. Le bouton `DUMP` copie sur le disque toutes les informations courantes ainsi que la version actuelle des 30 automates, à des fins de déboguage. La ligne de saisie permet à l'opérateur de proposer un texte correspondant à la séquence symbolique, lorsque celle-ci lui convient. L'action `VALIDER` convertit ce texte en une séquence phonético-prosodique à l'aide des hauts-niveaux, puis la compare à celle attendue. Si elles concordent, la phrase est ajoutée au script de lecture et les automates sont mis à jour. Sinon, un message d'erreur affiche les deux séquences symboliques en surlignant les différences.

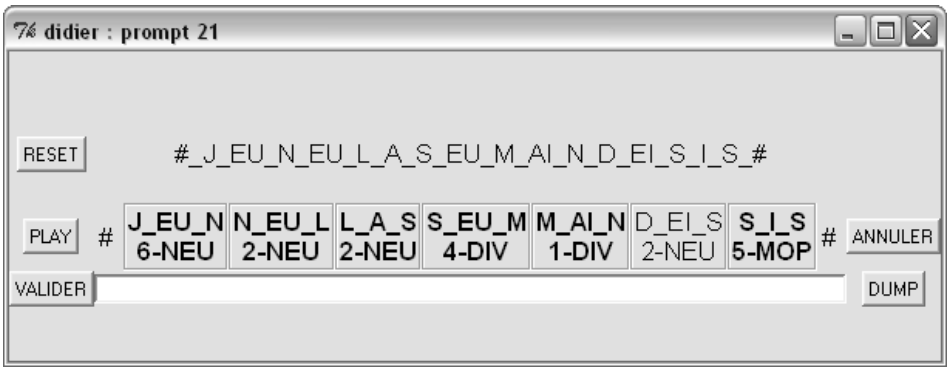


FIGURE 38 – Capture de l'interface graphique de construction de phrases.

Les modifications successives des séquences et automates sont commandées directement avec la souris. Par un « *Clic gauche* » sur un sandwich, l'opérateur interdit le bigramme menant à ce sandwich et une nouvelle fin est automatiquement proposée. L'opération symétrique, soit la régénération d'un début, est obtenue avec un « *Control-Clic gauche* ». De la même manière le forçage d'une fin courte se fait avec « *Clic droit* » et le forçage d'un début court avec « *Control-Clic droit* ».

L'interface peut également indiquer le niveau d'importance de chaque unité (sandwich ou bigramme) pour l'amélioration de la couverture. L'objectif est d'encourager l'opérateur à conserver dans la séquence finale les unités les plus fréquentes et non déjà couvertes. Pour éviter de surcharger l'interface nous nous sommes contentés d'une simple distinction binaire, en affichant en gras les sandwichs nouveaux. Sur l'exemple de la figure 38 l'avant-dernier sandwich a déjà été couvert dans une phrase précédente et ne représente donc aucun intérêt pour la couverture en sandwichs. Il nous a semblé au fil des expériences que cette information était parfois utile, même si elle n'est pas toujours prise en compte par les opérateurs.

L'entité centrale `CorpusServer.py`

Les composants `Client.py` et `FSTserver.py` ne peuvent pas dialoguer directement ; c'est le script `CorpusServer.py` qui joue le rôle d'intelligence centrale. Il suit l'évolution du script de lecture, répartit le travail entre les opérateurs, recueille les meilleurs chemins de chaque automate, commande les mises à jour d'automates, *etc.* À la façon d'un serveur http, il traite toutes les requêtes qui lui parviennent sur son interface réseau en provenance des différentes instances de `Client.py` (il y en a une par opérateur). La communication avec les instances de `FSTserver.py` se fait quant à elle par de simples sockets TCP ; on ouvre une socket pour chaque `FSTserver.py` et donc pour chaque automate.

Ainsi tous les composants logiciels peuvent aisément être distribués sur différentes machines au sein d'un réseau local. Si la parallélisation des automates ne soulève aucune difficulté technique, la coordination multi-utilisateurs est en revanche plus problématique. En particulier lorsqu'une séquence est proposée à un opérateur, les autres opérateurs ne doivent pas recevoir la même séquence, même si elle n'a pas encore été validée par le premier opérateur. Sinon les phrases construites en parallèle risquent de se montrer très redondantes. Par ailleurs, les bigrammes interdits par un opérateur doivent impacter uniquement sa propre séquence, de façon à ce que ces mêmes bigrammes puissent être proposés en parallèle aux autres opérateurs. Enfin les aller-retours dans l'historique via les fonctions « annuler » et « reset » doivent garantir, pour des raisons ergonomiques, une certaine conformité aux séquences du passé, quelle que soit l'évolution de la couverture provoquée entre temps par les autres opérateurs.

Pour assurer un tel fonctionnement, il est indispensable que chaque opérateur soit **identifié et interagisse avec son propre jeu d'automates** `WFST_REF_L(user)` **et** `WFST_REF_L`$^{-1}$`(user)`. Ces automates personnalisés sont déclinés à partir des automates de référence `WFST_REF_L` et `WFST_REF_L`$^{-1}$ par simple recopie ; ils sont initialisés lorsque l'opérateur entame la construction d'une nouvelle phrase et persistent jusqu'à la validation de cette phrase.

Les automates de référence n'évoluent qu'avec la couverture effective du script, donc à chaque fois qu'une phrase est validée par un opérateur. À l'inverse, les automates personnalisés évoluent à chaque action de l'opérateur : régénération partielle, forçage d'une terminaison, annulation d'action ou remise à zéro. `CorpusServer.py` maintient également une liste des sandwichs et bigrammes qui ont été proposés à un opérateur tout au long du processus de construction de la phrase courante. Ces unités sont temporairement considérées comme couvertes pour les autres opérateurs, avec une mise à jour systématique des coûts de transition de leurs automates[54]. Ainsi les unités qui ont été proposées à un opérateur deviennent temporairement sans intérêt pour les autres opérateurs, même si elles n'ont pas encore été validées dans une séquence finale.

Ce mécanisme de réservation temporaire des unités évite toute redondance entre les phrases construites en parallèle. Il peut théoriquement nuire à l'optimalité des séquences générées, mais en pratique les unités sont libérées très rapidement (à chaque validation de phrase) et l'impact sur la croissance du VSCR est totalement négligeable.

Une telle implémentation de `CorpusServer.py` offre une efficacité appréciable. D'une part les opérateurs s'inscrivent et se désinscrivent au gré de leur disponibilité (nous avons expérimenté jusqu'à cinq opérateurs simultanés avec succès). D'autre part la répartition multi-processeurs des automates donne accès à des **temps de calcul tout à fait acceptables sur le plan ergonomique**. Ainsi la recherche du chemin optimal se fait approximativement en 0.2 seconde ; à chaque changement de phrase, la recopie des automates de référence pour créer les automates personnalisés prend environ 1 seconde ; la mise à jour des coûts dans un jeu de 30 automates nécessite quant à elle 0.1 seconde environ.

11.2 Performances de la construction de phrases

Malgré cette efficacité logicielle, **la procédure de construction d'une phrase reste très coûteuse en temps humain** : un opérateur met en moyenne 3 minutes pour créer une phrase acceptable, soit une cinquantaine d'étapes (régénérations partielles ou annulations). Il s'agit là d'un inconvénient majeur de notre procédé. Rappelons néanmoins qu'un même

54. En réalité, pour chaque opérateur tiers, cette mise à jour peut être reportée jusqu'à la prochaine demande de régénération partielle.

script de lecture peut être utilisé pour l'enregistrement de nombreux locuteurs. La création d'un script dense est donc un investissement durable qui peut être amorti sur le long terme. L'utilisation de modèles génératifs plus contraints, tenant compte de critères lexicaux et/ou de règles grammaticales, pourrait probablement réduire les délais de construction, voire éviter le recours à une intervention humaine. Mais la perte de souplesse dans la génération des séquences de sandwichs s'accompagnerait inéluctablement d'une perte de densité.

Dans la suite de cette section nous évaluons l'apport de notre procédé de création de phrases sur la densité globale des scripts de lecture. Nous mesurons également l'impact de notre gestion des longueurs de phrases (maximisation du VSCR) en la comparant au critère normalisé (maximisation du VSCR par sandwich) ; nous verrons que le résultat n'est pas le même que dans le cas d'une condensation.

11.2.1 Amélioration de la densité globale

Comme nous le verrons en section 11.3, les scripts que nous avons effectivement constitués suivant le procédé de construction ne permettent pas une confrontation directe au procédé de condensation. Comme il serait trop coûteux de créer un nouveau script spécifiquement pour cette évaluation, nous nous contentons d'estimer les performances générales de notre algorithme.

La situation est identique à celle de la section 10.1, où nous avons évalué le procédé de condensation. D'après la topologie de nos automates et la définition des coûts de transitions, il s'agit toujours d'un glouton « fréquent d'abord », basé sur les groupes de souffles de moins de 15 sandwichs. Comme en 10.1, le critère d'optimisation repose ici sur la couverture des sandwichs de type `LRsemirobustes_13contextes`. En particulier les couvertures de bigrammes ne sont pas prises en compte.

Une **borne haute** de notre algorithme de construction peut facilement être estimée, en validant automatiquement toutes les séquences optimales, sans intervention humaine. La plupart des phrases construites suivant ce **procédé tout-automatique** sont donc dépourvues de sens, mais l'absence de contrainte « d'origine humaine » dans les automates aboutit nécessairement à une couverture plus dense. Pour affiner cette borne nous avons tout de même ajouté un mécanisme de correction des redondances intra-phrases. En temps normal cette correction, qui a un impact positif sur la croissance du VSCR, est effectuée naturellement par les opérateurs (voir en 11.1.4 page 123). Aussi pouvons-nous la simuler en parcourant chaque séquence optimale de gauche à droite et en commandant automatiquement une nouvelle fin chaque fois qu'un sandwich est répété au sein de la séquence. Il s'agit probablement d'une vision réductrice du comportement d'un opérateur face à ce type de redondance intra-phrase ; notre borne haute est donc plutôt pessimiste.

La détermination d'une **borne basse** est plus délicate, car elle est intimement liée à l'action de l'opérateur humain sur les séquences générées. L'impact de cette action sur la croissance du VSCR est d'ailleurs susceptible d'évoluer au fil de la construction du script, en fonction du niveau de couverture global. Pour mesurer cela nous avons appliqué le procédé semi-automatique sur de petits groupes de 10 à 30 phrases, en alternance avec le procédé tout-automatique. Plus précisément nous avons décomposé la création d'un script de 1500 phrases en 20 sessions comme suit :

- 30 phrases suivant le procédé semi-automatique (donc avec intervention humaine) ;
- 20 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 40 phrases en mode tout-automatique ;

- 12 phrases suivant le procédé semi-automatique ;
- 38 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 40 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 90 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 90 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 90 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 190 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 290 phrases en mode tout-automatique ;
- 10 phrases suivant le procédé semi-automatique ;
- 490 phrases en mode tout-automatique.

Nous obtenons ainsi un échantillonnage représentatif du procédé semi-automatique, réparti sur toute l'échelle de VSCR (de 1 à 93%). Du fait de l'alternance, les niveaux de VSCR atteints au terme de chaque session ne sont pas directement exploitables. En revanche, la dérivée du VSCR par rapport au nombre de sandwichs est riche d'information : observée sur les sessions semi-automatiques, elle nous indique ce que pourrait être l'évolution du VSCR sur un script complet si le procédé semi-automatique était appliqué intégralement. La figure 39 montre, entre autres, l'échantillonnage obtenu pour ces mesures de dérivée au sein des sessions semi-automatiques. Les mesures sont lissées sur 10 phrases successives et reportées en fonction du VSCR. La figure présente également les dérivées de la borne haute, du procédé de condensation et de la distribution de référence.

La plupart des fonctions de répartition de type $VSCR = F(Nombre\ de\ sandwichs)$ que nous avons observées jusqu'à présent se déduisent assez bien les unes des autres par de simples homothéties suivant l'axe des abscisses. Il est donc raisonnable de penser que l'échantillonnage précédent peut être interpolé de manière satisfaisante par une homothétie de la borne haute :

$$VSCR = F_{bh}(\alpha \cdot NbSand) \tag{17}$$

où F_{bh} est la fonction de répartition correspondant à la borne haute, $NbSand$ la variable portant sur le nombre de sandwichs du script, et α le coefficient d'homothétie. On constate sur la figure qu'une bonne approximation des échantillons peut effectivement être obtenue avec le coefficient $\alpha = 0.87$.

Du fait de l'alternance entre les sessions semi-automatiques et tout-automatiques, cette approximation nous donne une estimation basse des performances de notre algorithme. En effet, les sessions tout-automatiques ont tendance à construire des phrases plus denses, qui ont pour conséquence de réduire la marge de manoeuvre des sessions semi-automatiques suivantes. Donc les dérivées de VSCR enregistrées sur les sessions semi-automatiques sont a priori plus faibles que ce qu'elles pourraient être si l'intégralité du script était construite sur ce mode. D'où la notion de « borne basse ».

Nous disposons à présent d'une fourchette fiable pour les performances de notre algorithme. La figure 40 présente les fonctions de répartition correspondantes. On constate que notre algorithme de construction, dont la zone de fonctionnement est grisée sur la figure, améliore significativement la densité du script par rapport à une approche par condensation. On peut en effet espérer une **réduction des tailles de script de 30 à 40% pour un VSCR donné**, ce qui est évidemment très avantageux.

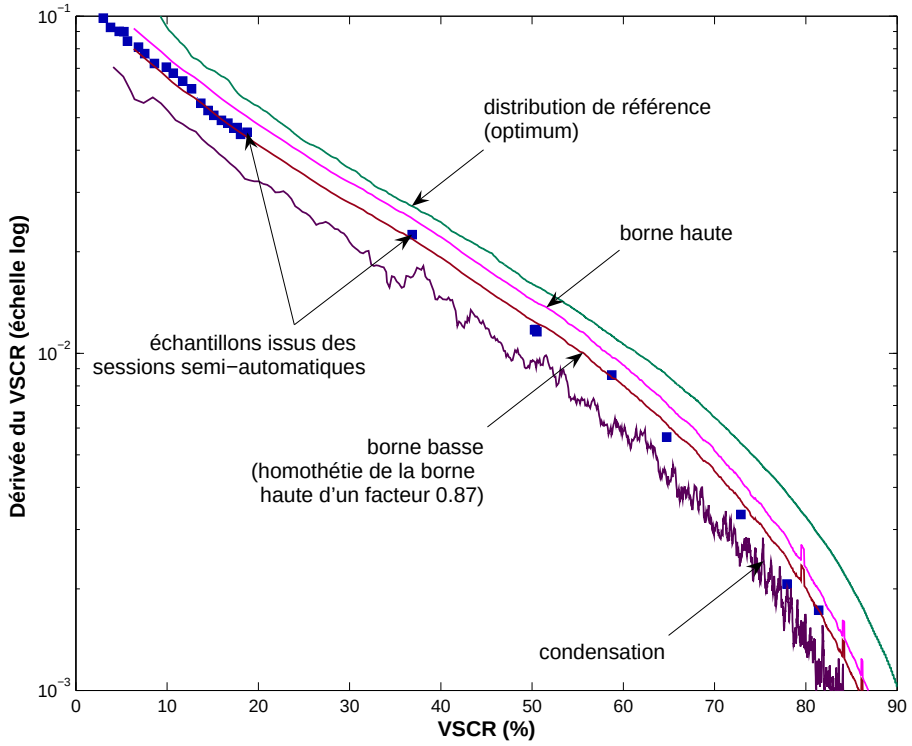


FIGURE 39 – Évolution de la dérivée du VSCR suivant différentes approches, en fonction du VSCR. L'évolution du VSCR sur les sessions semi-automatiques est bien décrite par une homothétie de la borne haute.

11.2.2 Impact modéré de la contrainte de longueur

Nous avons vu au paragraphe 10.2.1 qu'en choisissant d'exclure les phrases trop longues plutôt que de normaliser leur apport par leur longueur, nous provoquons, dans le cas d'une condensation, une inflation de 30 à 40% sur la taille des scripts. Même si cette inflation se justifie par une meilleure distribution des longueurs de phrases, elle reflète le manque de souplesse du procédé de condensation, lié au contenu limité de son corpus de pioche. On peut légitimement penser que notre procédé de construction de phrases, en élargissant considérablement l'ensemble des possibles, est à même de réduire cet écart.

Pour vérifier cela nous utilisons la version tout-automatique de notre procédé, décrite plus haut. Le comportement de l'algorithme complet, c'est-à-dire en mode semi-automatique, pourrait conduire à des résultats sensiblement différents ; il ne s'agit donc là que d'une expérience indicative.

Pour chacune des deux stratégies, un script est créé en mode tout-automatique. Plus précisément pour la première stratégie on retient, à chaque itération, la séquence offrant le coût minimal (*i.e.* celle qui maximise le VSCR) parmi les 15 séquences optimales correspondant aux 15 longueurs permises. C'est l'approche préconisée dans ce document ; aussi retrouvons-nous précisément la borne haute définie à la section précédente. Pour la seconde stratégie on retient la séquence qui minimise le coût moyen, autrement dit celle qui maximise l'accroissement du VSCR par sandwich. Notons que, dans les deux cas, une majoration de la longueur de phrase à 15 sandwichs est imposée par notre jeu d'automates. C'est une légère différence avec l'évaluation du 10.2.1, dans laquelle le critère normalisé n'était pas combiné à une majoration

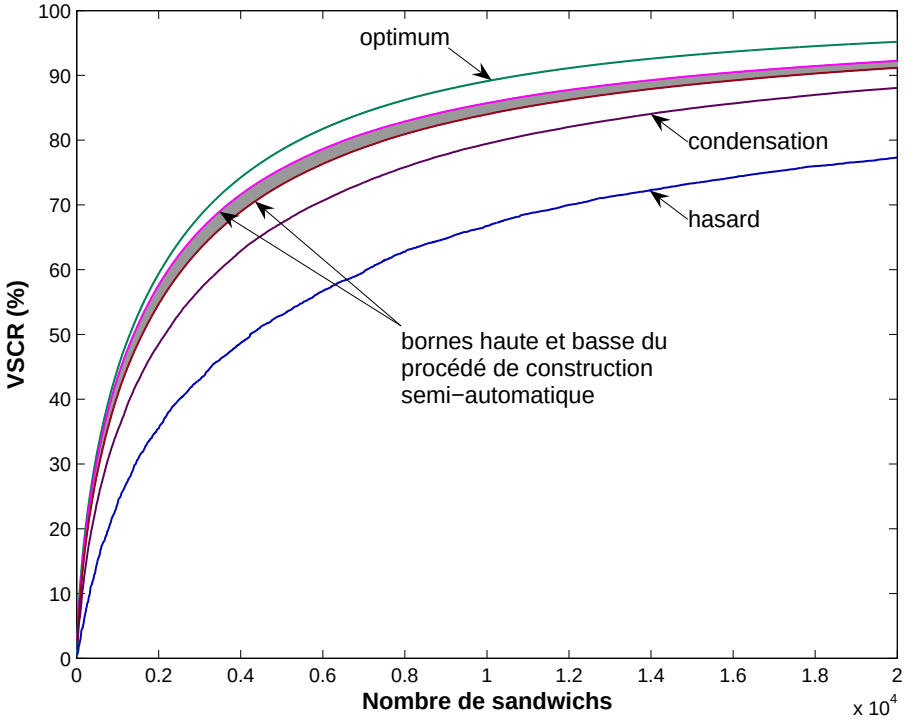


FIGURE 40 – Estimation de l'évolution du VSCR au fil des créations de phrases (zone grise), en comparaison de l'optimum (distribution de référence), de la condensation et de la sélection aléatoire.

des longueurs de phrases. Mais cette différence est sans importance, puisque les longueurs de phrases qui avaient alors été relevées étaient de toute façon bien inférieures à 15 sandwichs.

La figure 41 rapporte l'évolution de VSCR observée sur chacun des deux scripts. La figure 42 présente quant à elle les distributions de longueurs de phrases, en sandwichs et en phonèmes, pour un VSCR de 80%.

On constate que l'écart entre les deux stratégies est beaucoup plus faible que dans le cas de la condensation. La différence de taille de script est par exemple de 5% pour un VSCR de 50% et de 17% pour un VSCR de 80%, alors qu'elle était respectivement de 34% et 38% avec la condensation. La dépendance de cet écart au VSCR est également une nouveauté : dans le cas de la condensation il était à peu près constant. Ceci s'explique par le fait que les phrases construites suivant le critère non normalisé comportent presque toutes 15 sandwichs. Le taux de redondance, qui peut être très faible initialement grâce au mécanisme de construction, augmente nécessairement au fur et à mesure que la couverture s'étend. Le manque de souplesse de la condensation conduisait à l'inverse à des phrases plus courtes, avec une forte redondance dès le début.

D'une manière générale, le procédé de construction offre des longueurs de phrases supérieures au procédé de condensation, du moins en mode tout-automatique. Les phrases obtenues suivant le critère normalisé comportent en moyenne 5.8 sandwichs (14.3 phonèmes), alors qu'elles comportaient 3.8 sandwichs (9.4 phonèmes) avec la condensation. Les phrases obtenues suivant le critère non normalisé sont presque toutes saturées à 15 sandwichs (36.5 phonèmes), contre 12.2 sandwichs (29.8 phonèmes) dans le cas de la condensation.

Ces résultats démontrent l'**aptitude de notre algorithme à maximiser la densité**

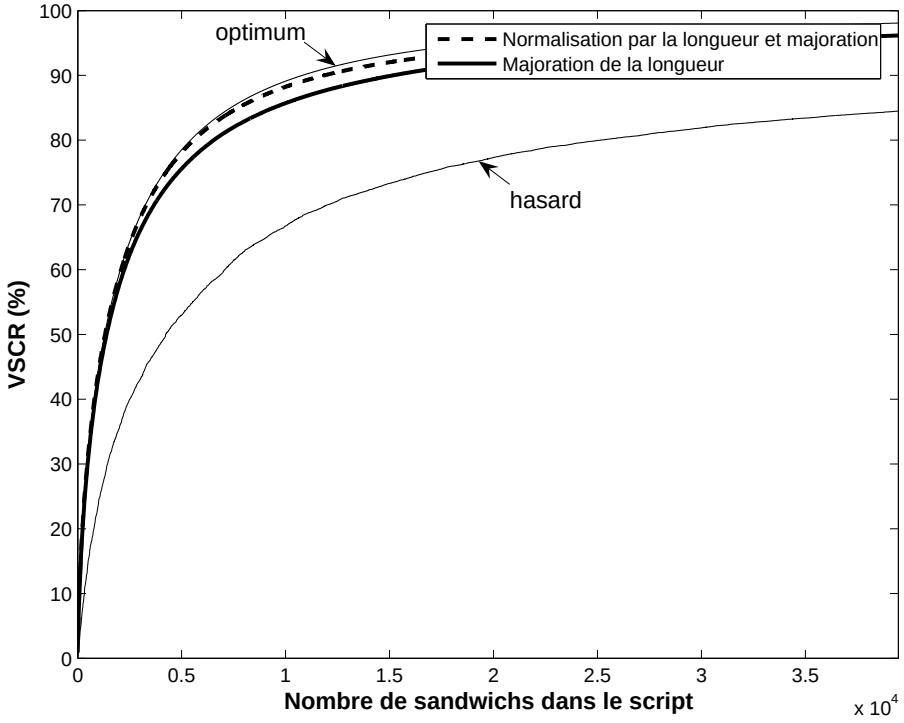


FIGURE 41 – Evolution du VSCR avec le procédé tout-automatique et pour les deux stratégies de gestion des longueurs de phrases.

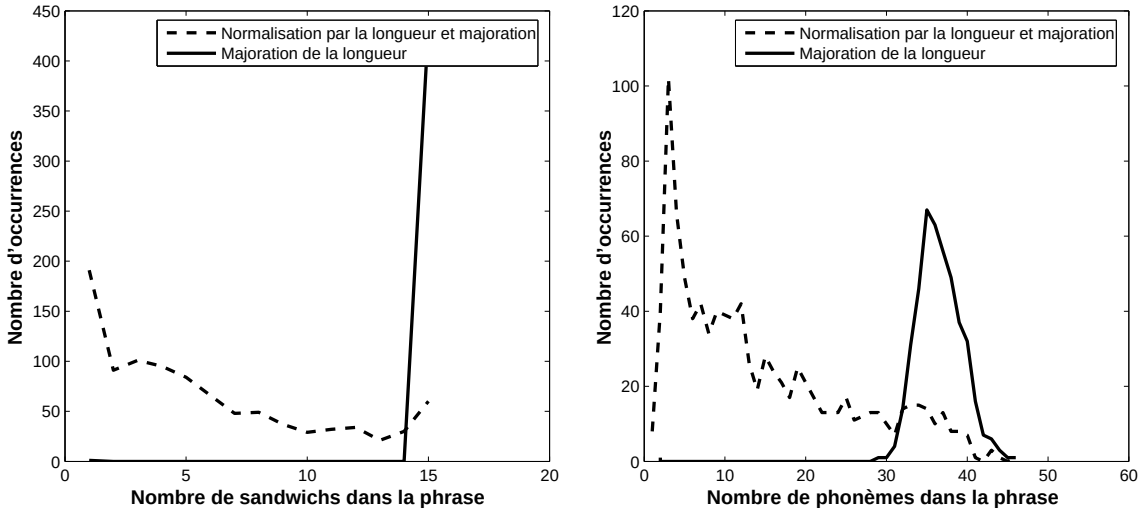


FIGURE 42 – Distribution des longueurs de phrases, en nombre de sandwichs et en nombre de phonèmes, pour les deux stratégies.

sans recourir à des phrases excessivement courtes. En particulier, il semble faiblement impacté par notre stratégie visant à favoriser les phrases de longueur moyenne.

11.3 La construction de phrases en pratique

Pour l'utilisation de notre procédé semi-automatique de construction de phrases en situation réelle, il convient de définir plusieurs phases d'optimisation successives, exactement comme dans

le cas de la condensation (voir en 10.3.1). On gagne en effet à moduler le critère d'optimisation au fil des phrases, du plus contraignant au moins contraignant, afin d'assurer une couverture minimale de phénomènes linguistiques complexes. A chacune de ces phases doit correspondre un critère d'arrêt spécifique.

La table 11 donne la liste des scripts de lecture créés par construction de phrases, ainsi que quelques ordres de grandeurs pour chacun d'entre eux. Comme pour la condensation, nous avons nommé les scripts en fonction de leur durée approximative d'enregistrement (et non de parole utile) et du type d'unité principalement utilisé dans le critère d'optimisation.

	Nom du script	**25min_sandwiches**	**1h_sandwiches**	**3h_sandwiches**	**4h30_sandwiches**
	Nombre de phrases	214	600	1 613	2 919
	Nombre de diphones	4 429	10 224	28 686	47 719
	Nombre de sandwichs (cas `LRsemirobustes`)	1 783	3 939	10 783	18 004
	Longueur moyenne des phrases (en diphones)	20.7	17.0	17.8	16.3
	Durée prévisionnelle d'enregistrement	25 min	1 h	3 h	4 h 30 min
	Durée prévisionnelle de parole utile	6 min	14 min	38 min	1 h
	Type de phrases	groupes de souffle	groupes de souffle	groupes de souffle	groupes de souffle
	Critère d'optimisation principal	sandwichs	sandwichs	sandwichs	sandwichs
	Nombre indicatif de symboles contextuels	4	4	13	13
Taux de couverture	2-grammes de sandwichs, liquides fragiles, 13 contextes	7%	10%	16%	26%
	1-grammes de sandwichs, liquides fragiles, 13 contextes	32%	36%	59%	71%
	1-grammes de sandwichs, liquides semi-robustes, 13 contextes	44%	48%	75%	81%
	1-grammes de sandwichs, liquides semi-robustes, 4 contextes	56%	75%	87%	88%
	Diphones (4 symboles contextuels)	90%	95%	99%	99%
	Triphones (13 symboles contextuels)	37%	41%	66%	77%

TABLE 11 – Caractéristiques principales des quatre scripts obtenus par construction de phrases.

Les scripts « **25min_sandwiches** », « **1h_sandwiches** » et « **3h_sandwiches** » couvrent les besoins de **création de voix express, dans la mesure où ils requièrent moins d'une demi-journée d'enregistrement**. Pour les deux premiers, le critère d'optimisation a porté exclusivement sur les sandwichs de type `LRsemirobustes_4contextes`. Le premier script correspond au début du second. Pour le troisième nous avons utilisé une série de critères plus complexes, sans toutefois aller jusqu'aux bigrammes de sandwichs. Ces trois scripts ont été réalisés avec une version intermédiaire de notre algorithme de construction, moins aboutie que celle présentée dans ce document. D'une part les séquences étaient obtenues

sans utilisation de la librairie OpenFst, par une approximation locale du problème de recherche du meilleur chemin ; d'autre part notre gestion des longueurs était assez confuse ; enfin le corpus de référence n'était que partiellement constitué. La densité de ces scripts ne peut donc pas être confrontée aux prévisions de la section précédente.

Les versions finales de notre algorithme et du corpus de référence sont apparues au fil de la construction de « **4h30_sandwiches** ». De nombreux ajustements ont été effectués tout au long de ce script. En particulier nous avons progressivement délaissé le critère d'optimisation normalisé au profit du critère d'optimisation absolu et plusieurs valeurs du seuil de longueur ont été expérimentées. Là encore, la complexité des évolutions rend impossible l'analyse objective des résultats en matière de densité.

Par conséquent aucun des quatre scripts constitués ne permet, malheureusement, de confirmer l'estimation des performances proposée à la section précédente. D'autres enseignements plus qualitatifs peuvent toutefois en être tirés.

Tout d'abord le **coût de l'intervention humaine** est confirmé, avec une moyenne d'environ 3 minutes pour la création de chaque phrase par un opérateur. Le processus est plus rapide en début de script, mais s'alourdit au fur et à mesure que la couverture croît. L'augmentation des coûts de transition portant sur des unités fréquentes s'accompagne en effet d'une complexification des séquences générées, qui rend la création de phrase de plus en plus délicate. Cette contrainte conduit par ailleurs les opérateurs à valider des phrases de longueur légèrement décroissante.

Ensuite nous avons constaté que la plupart des phrases construites présentent une **cohérence sémantique plus faible** que les phrases issues du corpus de référence. C'est une conséquence logique de notre volonté de contrer la distribution naturelle des unités, qui se traduit normalement par une redondance des unités fréquentes et une apparition fréquente d'unités rares. Le gain de densité se fait au détriment de la sémantique, d'autant plus qu'on avance dans le script. Les phrases suivantes, extraites de la fin du script « 4h30_sandwiches », en attestent :

```
À l'église Saint-Thomas t'as la soirée privée avant 11 heures.
C'est le big stress pour accoucher avec Fred.
Par crainte de ressource suivez les arbres,
La ligne officielle spécialiste de l'immeuble,
Mais le diagnostic des rapports entre les provinces,
Car il n'y a que 300 000 touristes fréquemment meilleurs.
Il répond aux news par fatigue,
Souhaitons qu'il puisse résolument imposer les employés.
```

Cette limitation sémantique peut avoir des répercussions sur la phase d'enregistrement, avec un nombre accru d'hésitations et de reprises, des productions peu naturelles de la part du locuteur, voire un certain agacement. Dans certains cas ces désagréments peuvent contrebalancer les bénéfices offerts par la densification du script, ce que nous discuterons dans la partie suivante.

Quatrième partie

Création et évaluation de voix de synthèse

Sommaire

Nous disposons à présent de scripts de lecture spécialement conçus pour la synthèse par corpus. Ils laissent entrevoir une meilleure efficacité dans le processus de création de voix, pour deux raisons. La première est la prise en compte inédite, à travers l'optimisation de la couverture en sandwichs vocaliques, des principaux facteurs segmentaux et prosodiques qui influencent la qualité de la synthèse par corpus. La seconde tient aux outils et stratégies utilisés, qui confèrent à ces scripts une grande densité.

Dans cette partie nous exploitons ces scripts pour la création de nombreuses voix de synthèse, à partir d'enregistrements allant de quelques minutes à une journée. Nous verrons que ces procédés « express » nous permettent d'ouvrir de nouvelles possibilités pour l'expressivité des voix de synthèse.

12 Évaluation des procédés de création de voix

12.1 Acquisition d'enregistrements dédiés

12.1.1 Les scripts de lecture

Suite aux travaux exposés précédemment, nous disposons de plusieurs scripts de lecture optimisés sur des critères à base de sandwichs :

- « 25min_sandwiches » ;
- « 1h_sandwiches » ;
- « 3h_sandwiches » ;
- « 4h30_sandwiches » ;
- « 5h_sandwiches ».

Qu'ils soient obtenus par construction ou par condensation, ces scripts offrent, à taille égale, des niveaux de VSCR bien supérieurs à l'état de l'art. En atteste la table 12, qui reprend les caractéristiques principales de tous nos scripts de lecture, incluant ceux conformes à l'état de l'art :

- « 45min_nphones » ;
- « 3h_nphones » ;
- « 5h_nphones » ;
- « 8days_nphones ».

Parmi ces derniers, « 8days_nphones » constitue une référence de qualité quasi-universelle : la plupart des systèmes commerciaux de synthèse par corpus reposent sur des scripts similaires pour la création de voix de haute qualité. Son enregistrement sur 8 jours implique des coûts considérablement plus élevés que nos autres scripts, qui requièrent tous moins d'une journée.

Pour valider nos travaux nous avons mené une expérience à grande échelle, consistant à créer et évaluer de nombreuses voix de synthèse avec chacun des scripts [Cadic 10b]. C'est cette expérience que nous nous attachons à décrire maintenant.

12.1.2 Les enregistrements

Nous avons développé une interface graphique facilitant le suivi des enregistrements, tant pour le locuteur que pour l'équipe de supervision. La figure 43 en présente une capture d'écran. Dans cette interface, chaque phrase peut être validée ou refusée par le(s) superviseur(s) par un simple appui sur une touche, ce qui permet au passage un découpage en temps réel des enregistrements. Chaque phrase peut instantanément être réécoutée, par exemple pour vérifier

Nom du script	Nombre de phrases dans la base	Nombre de diphones dans la base	Durée prévisionnelle de parole utile	Critère et algorithme d'optimisation	Taux de couverture : Sandwichs (liquides semi-robustes, 4 symboles contextuels)	Taux de couverture : Sandwichs (liquides fragiles, 13 symboles contextuels)	Taux de couverture : Diphones (4 symboles contextuels)	Taux de couverture : Triphones (13 symboles contextuels)
25min_sandwiches	214	4 429	6 min	sandwichs, construction	56%	32%	90%	37%
45min_nphones	547	5 760	8 min	diphones, condensation	31%	14%	96%	20%
1h_sandwiches	600	10 224	14 min	sandwichs, construction	75%	36%	95%	41%
3h_nphones	617	18 632	25 min	diphones, condensation	67%	38%	98%	47%
3h_sandwiches	1 613	28 686	38 min	sandwichs, construction	87%	59%	99%	66%
4h30_sandwiches	2 919	47 719	1 h	sandwichs, construction	88%	71%	99%	77%
5h_nphones	1 532	40 995	55 min	di/triphones, condensation	80%	55%	100%	69%
5h_sandwiches	2 860	57 768	1 h 15	sandwichs, condensation	90%	75%	100%	82%
8days_nphones	7 919	261 275	6 h	di/triphones, condensation	90%	75%	100%	88%

TABLE 12 – Caractéristiques principales des différents scripts utilisés dans nos travaux.

sa concordance avec la séquence phonétique attendue. L'équipe de supervision reste totalement maître du déroulement des opérations ; le locuteur, qui observe la même interface par un système de double écran, se contente de lire les phrases surlignées.

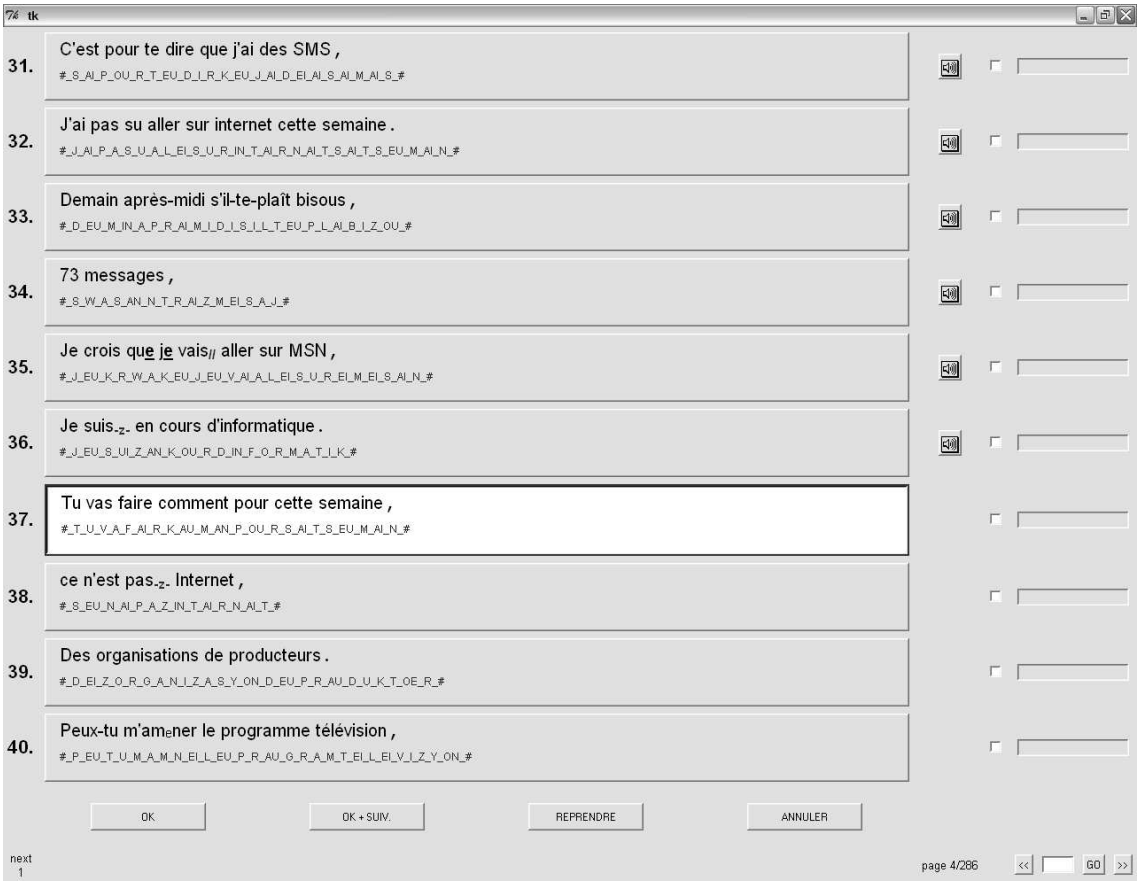


FIGURE 43 – Capture d'écran de l'interface d'acquisition utilisée pour la plupart de nos enregistrements de voix. Dans cet exemple, le locuteur doit lire la phrase numéro 37.

Les consignes de lecture portent sur certains choix phonétiques, sur le respect de la ponctua-

tion et plus généralement sur la conformité au style « lecture neutre ». L'équipe de supervision surveille en outre la constance vocale du locuteur et la qualité sonore des acquisitions (voir page 56).

La durée de parole utile correspond approximativement au quart de la durée des sessions d'enregistrement. Cette proportion dépend en partie du contenu du script de lecture ; en particulier les phrases de longueur moyenne, soit entre 20 et 40 phonèmes, offrent la meilleure efficacité. Notre gestion des longueurs de phrases (voir en 9.2) est particulièrement bénéfique dans ce domaine. La restriction à des groupes de souffles isolés, ainsi que la validation systématique, dès la constitution des scripts, des séquences phonétiques, sont d'autres éléments avantageux de notre stratégie : elles rendent la lecture fluide et agréable.

Nous avons constaté que **les scripts obtenus par construction étaient responsables d'une perte d'efficacité chez certains locuteurs**. Le manque de cohérence sémantique mis en avant page 134 a en effet suscité un nombre accru d'hésitations. Le ralentissement global de la lecture a pu atteindre environ 30% dans le pire des cas, facteur qui dépend grandement du niveau de professionnalisme du locuteur. Les plus agiles d'entre eux n'ont pour ainsi dire pas été gênés. Sachant que le bénéfice de notre algorithme de construction est de 30 à 40% sur la taille du script, la prise en compte de ce facteur est capitale. **La lecture des scripts obtenus par construction doit être réservée aux locuteurs expérimentés. Pour les novices, l'utilisation de scripts condensés est préférable**. C'est d'ailleurs pour cette raison que nous avons construit « 5h_sandwiches », dernier script en date, par condensation.

12.1.3 Le choix de voix

Pour une comparaison rigoureuse des performances des différents scripts de lecture, l'idéal aurait été d'enregistrer avec chacun d'entre eux un panel de locuteurs identique. Mais ici les enregistrements de voix ont été collectés tout au long de nos travaux [55] à l'occasion de projets variés, impliquant à chaque fois des locuteurs différents. Si certaines voix sont effectivement disponibles sur plusieurs scripts, l'intersection globale est nulle et les scripts sont évalués sur des ensembles de voix assez différents. Ceci peut théoriquement introduire un biais ; mais la diversité des voix disponibles nous permet tout de même d'envisager sereinement une simple comparaison « par nuages ».

Toutes nos voix de synthèse n'ont pas pu être retenues pour cette évaluation. D'abord parce que notre test perceptif aurait été trop coûteux, ensuite parce que certaines d'entre elles n'ont pas été enregistrées dans des conditions équitables. Nous avons par exemple exclu quatre voix correspondant au script « 4h30_sandwiches » : d'une part les conditions de supervision n'étaient pas réunies donc ces voix ont beaucoup fluctué, d'autre part les enregistrements ont été interrompus prématurément (environ au milieu du script, qui n'a pas été conçu pour ça puisque le début est optimisé sur des bigrammes), enfin les voix ne respectent pas le style « lecture neutre » : fortes variations rythmiques et mélodiques, imprégnation de la prosodie par la sémantique, etc.

Au total, parmi les différents dictionnaires de synthèse créés avec des enregistrements dédiés, nous en avons retenu 33 pour l'évaluation, dont 22 issus d'un script à base de sandwichs. Les voix ont été produites par 17 locuteurs différents (5 femmes, 12 hommes), dont 10 professionnels. Nous entendons par professionnels des personnes dont l'expérience laisse penser qu'ils ont une maîtrise avancée de leur organe. Il s'agit de comédiens, chanteurs, chroniqueurs ou encore orthophonistes. Parfois un même locuteur a procédé à plusieurs enregistrements avec des voix

55. Seules les voix enregistrées suivant le script « 8days_nphones » sont antérieures à l'étude présentée dans ce document.

différentes, en jouant essentiellement sur le timbre, mais aussi sur la prosodie et le niveau d'articulation.

12.2 Récupération de rushes

Pour notre évaluation nous avons également retenu 5 dictionnaires de synthèse construits à partir de rushes (voir section 3.4). Leur contenu n'est pas dédié à la synthèse vocale et n'a donc fait l'objet d'aucune optimisation linguistique. C'est une différence majeure avec les dictionnaires issus d'enregistrements dédiés ; les rushes constituent en cela des éléments de comparaison intéressants pour notre évaluation.

Ils présentent un niveau d'expressivité très variable, allant de la lecture quasiment neutre au jeu d'acteur expressif. La table 13 présente les caractéristiques principales des 5 bases de rushes retenues. Elle peut être comparée à la table 12 page 137, qui synthétise les caractéristiques des différents scripts de lecture. On constate logiquement que, à taille comparable, les rushes offrent des niveaux de couverture inférieurs aux scripts dédiés.

Les bases « Marge_rushes » et « Homer_rushes » ont été collectées à partir de DVD de la série *les Simpsons*, tandis que « PPDA_rushes » et « Chirac_rushes » ont été capturées respectivement sur les sites internet de TF1 et de l'Élysée. Toutes ces bases étaient à l'origine accompagnées de transcriptions textuelles. Mais le manque de fidélité de ces transcriptions nous a contraints à les ressaisir entièrement, par le biais de prestations externes. La base « Jonathan_rushes » est issue du projet de recherche collaboratif VIVOS, subventionné par l'ANR[56] et le CNC[57] dans le cadre du réseau RIAM[58].

					Taux de couverture			
Nom du script	**Nombre de phrases dans la base**	**Nombre de diphones dans la base**	**Durée de parole utile**	**Style vocal et origine**	**Sandwichs (liquides semi-robustes, 4 symboles contextuels)**	**Sandwichs (liquides fragiles, 13 symboles contextuels)**	**Diphones (4 symboles contextuels)**	**Triphones (13 symboles contextuels)**
Marge_rushes	1 157	26 328	36 min	Très expressif, dessin animé	72%	47%	97%	58%
Homer_rushes	2 593	50 469	1 h 10	Très expressif, dessin animé	78%	54%	99%	67%
PPDA_rushes	2 570	141 709	2 h 52	Neutre, journal TV	83%	62%	99%	77%
Jonathan_rushes	1 633	181 271	3 h 35	Expressif, narration	84%	62%	99%	76%
Chirac_rushes	27 618	438 603	10 h 25	Expressif, discours politique	86%	68%	100%	82%

TABLE 13 – Caractéristiques principales des cinq bases de rushes utilisées dans nos travaux.

12.3 Traitements et analyse acoustique

Qu'elles aient été enregistrées spécifiquement ou collectées par rushes, nous avons traité les données acoustiques conformément à l'exposé de la section 3.3. Le volume des phrases a été **normalisé** suivant la recommandation ITU-T P.56 [ITU-T 93] et nous avons utilisé pour le **pitch-marquage** les outils ARX-LF de Vincent et Rosec [Vincent 06]. Les bases « PPDA_rushes »

56. Agence Nationale de la Recherche
57. Centre national du cinéma et de l'image animée
58. Recherche et Innovation en Audiovisuel et Multimédia

et « Chirac_rushes », souffrant respectivement d'un bruit de codage et d'un bruit ambiant importants, ont fait l'objet d'un **débruitage** avec des outils existants basés sur un filtrage de Wiener [Boll 79] et développés par les Orange Labs.

Pour la **segmentation phonétique**, nous avons utilisé la bibliothèque HTK avec des modèles multi-locuteurs, complétés d'un apprentissage mono-locuteur au-delà de 15 minutes de parole disponibles. Pour cette étape nous n'avons pas souhaité utiliser des variantes phonétiques (voir page 58). De telles variantes ont en effet pour conséquence de disperser la couverture linguistique, en éloignant la séquence phonétique finale de celle attendue. Ceci est particulièrement gênant dans le cas d'enregistrements dédiés, le script de lecture ayant fait l'objet d'une optimisation phonético-linguistique préalable. Mais c'est également vrai pour les rushes, car les séquences phonétiques générées automatiquement à partir de la transcription présentent au moins l'intérêt d'être conformes à celles qui seront ensuite rencontrées sur des textes semblables. Pour plus de rigueur, il faudrait introduire des variantes identiques à la fois dans l'étape *offline* de segmentation phonétique et dans l'étape *online* de sélection d'unités [Bulyko 02][Revelin 05], l'équilibre global étant assurément difficile à trouver. Notre supervision rigoureuse nous permet cependant de supposer que les enregistrements sont fidèles aux séquences phonétiques attendues. Les imprécisions résiduelles relèvent essentiellement de spécificités articulatoires (par exemple [r] au lieu de [ʁ]) et ne justifient pas le recours à des mécanismes de variantes phonétiques.

Sur les 38 dictionnaires de synthèse retenus pour l'évaluation (enregistrements dédiés et rushes confondus), 8 ont fait l'objet d'une révision manuelle complète de la segmentation phonétique. Pour chacun d'entre eux plusieurs dictionnaires peuvent alors être déclinés :

auto/auto : dictionnaire créé en mode tout-automatique, c'est-à-dire en conservant la séquence phonétique attendue et la segmentation automatique.

manuel/auto : dictionnaire créé avec les séquences phonétiques révisées et une segmentation automatique tenant compte de ces séquences révisées.

manuel/manuel : dictionnaire créé avec la segmentation entièrement révisée.

Ces niveaux de révision graduels nous permettront dans la section suivante d'évaluer séparément l'intérêt de réviser manuellement les séquences phonétiques et leurs alignements temporels.

12.4 Matériel de test

Notre évaluation a porté au total sur **49 dictionnaires de synthèse** : 33 dictionnaires créés à partir d'enregistrements dédiés, 5 dictionnaires créés à partir de rushes et 11 dictionnaires déclinés des précédents en faisant varier le niveau de révision. 25 voix différentes interviennent dans cet échantillonnage, avec des timbres, sexes, styles d'élocution et niveaux de professionnalisme variés. Certaines voix sont multi-représentées, chaque dictionnaire correspondant alors à un procédé différent : rushes ou enregistrements dédiés, scripts de taille et d'origine diverses, segmentation manuelle ou automatique. Une évaluation détaillée des principaux aspects du processus de création de voix est donc possible. Le détail des dictionnaires et des voix sera présenté dans les tables de résultats ; des exemples sonores originaux et synthétiques sont également fournis sur le CD d'accompagnement.

Pour évaluer la qualité de la synthèse vocale offerte par chacun de ces 49 dictionnaires, nous avons effectué des **essais d'opinion d'écoute de type *Degradation Category Rating*** (voir en 4.2.2). Dans notre situation le protocole DCR est préférable au protocole ACR (*Absolute Category Rating*), car il ne dépend pas des préférences absolues entre les voix naturelles. Notre test comporte en effet plusieurs timbres de voix célèbres ou ludiques susceptibles de biaiser le résultat dans le cas d'un ACR. Le critère évalué ici est le « niveau de dégradation de la voix de

synthèse par rapport à la voix naturelle ». Conformément à la recommandation ITU-T P.800 [ITU-T 96], l'échelle de notation est composée de cinq catégories de dégradation, allant de « 1 = Dégradation très gênante » à « 5 = Dégradation inaudible » (voir la table 3 page 68).

51 auditeurs, tous français natifs et extérieurs au domaine de la synthèse vocale, ont participé à l'évaluation perceptive. Avec 26 femmes et 25 hommes, la représentation des sexes est équilibrée. La figure 44 présente l'histogramme des âges, la moyenne étant de 35 ans et l'écart-type de 16 ans. Ce panel d'auditeurs a été recruté pour partie dans l'entourage proche de l'équipe et pour partie dans le corps étudiant de l'ENSSAT, école d'ingénieur lannionnaise (ce qui explique la surreprésentation de la tranche 21-30 ans). La majorité des participants ont reçu une petite compensation financière.

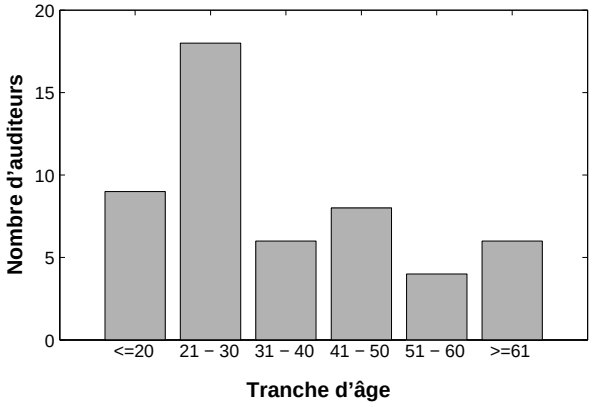


FIGURE 44 – Répartition des âges au sein du panel d'auditeurs.

Nous avons développé une interface logicielle dédiée à ce test. Une capture d'écran en est rapportée figure 45. Cette interface est organisée en pages, portant chacune sur une voix de synthèse donnée. Chaque page propose une phrase naturelle et trois phrases synthétiques de la voix concernée. Il s'agit donc d'une version légèrement revisitée du protocole DCR standard, qui oppose normalement une référence naturelle à chaque phrase synthétique. L'unique but de cette adaptation est de réduire la durée du test. Dans ce type d'évaluation, le premier avis est souvent le meilleur ; afin d'éviter les tergiversations, les phrases synthétiques ne peuvent être jouées qu'une seule fois et doivent être notées immédiatement après l'écoute. Les boutons correspondants sont activés puis désactivés automatiquement, de manière séquentielle. Sur l'exemple de la figure 45, les deux premières phrases synthétiques sont déjà notées ; l'auditeur peut soit modifier sa notation de la deuxième phrase, soit passer à l'écoute de la troisième phrase. A l'inverse, la phrase naturelle peut être écoutée pendant toute la durée d'affichage de la page, autant de fois que l'auditeur le souhaite.

Équipé d'un casque sonore de qualité, chaque auditeur a noté les 49 voix de synthèse, ainsi que **3 voix de contrôle**. Ces dernières, dissimulées parmi les voix de synthèse, sont en fait des enregistrements naturels. D'une part elles contribuent à calibrer l'échelle de notation des auditeurs, d'autre part elles nous permettent d'estimer la borne haute de la notation, borne qu'aucune voix de synthèse ne peut a priori dépasser. L'ordre de présentation des 52 pages est aléatoire. Pour affiner encore le calibrage du test, l'ensemble est précédé de **3 pages d'apprentissage**, portant sur des voix choisies de manière experte comme étant représentatives de l'échelle de qualité : une voix de très mauvaise qualité, une voix de très bonne qualité et une voix intermédiaire. Les auditeurs sont prévenus qu'il s'agit d'une phase d'apprentissage et qu'aucune des notes attribuées sur ces pages n'est prise en compte.

L'interface comporte donc 52 pages cibles, sollicitant chacune 3 notes, soit un total de 156

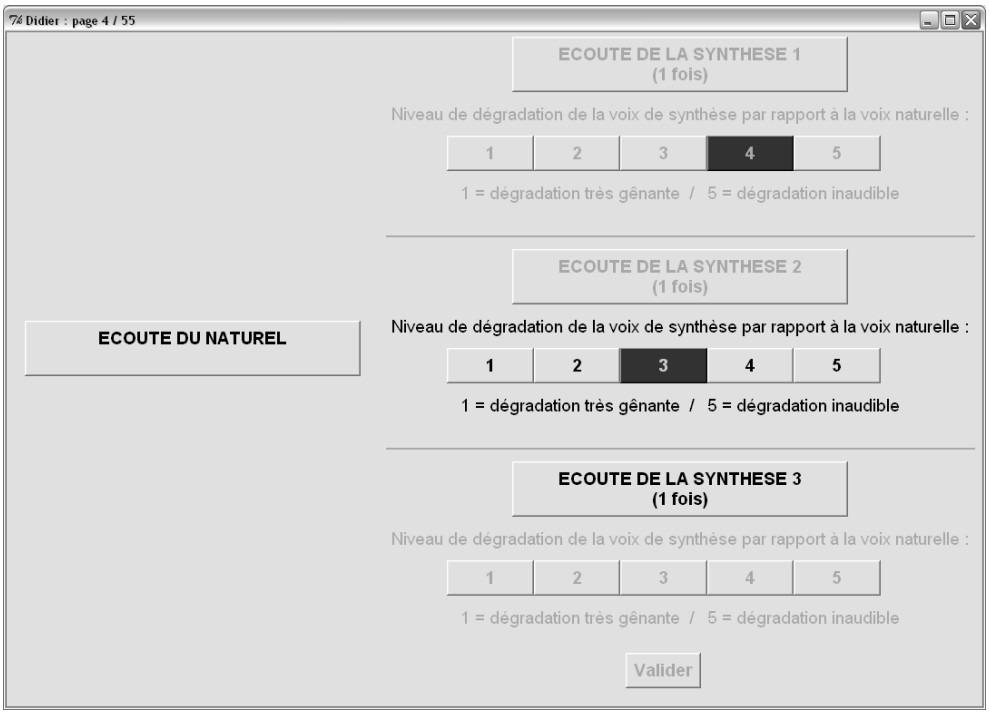


FIGURE 45 – Capture d'écran de l'interface de notation utilisée pour notre évaluation subjective globale.

notes par auditeur. **Le déroulement complet du test nécessite entre 20 et 40 minutes** selon les personnes.

Afin de réduire la marge d'erreur liée au choix des phrases synthétiques, nous les avons fait varier d'un auditeur à l'autre. Plus précisément nous avons constitué au préalable un corpus textuel de $49 \times 3 = 147$ phrases textuelles, issues de domaines variés représentatifs des usages de la synthèse vocale. Ce corpus n'interfère pas avec le contenu des dictionnaires de synthèse ; en particulier il est totalement indépendant du corpus de référence présenté dans ce document. La bonne interprétation des 147 phrases par les hauts-niveaux du système de synthèse a été rigoureusement contrôlée, de sorte qu'aucune erreur de transcription phonétique ne vienne perturber la notation. Les phrases ont ensuite été réparties entre les auditeurs et les voix de manière à minimiser les répétitions. L'interface d'évaluation n'affiche volontairement pas le texte sous-jacent ; les problèmes occasionnels d'intelligibilité de certaines voix de synthèse sont ainsi mieux pris en compte dans la notation.

Au final, chaque voix de synthèse a reçu 153 notes, attribuées par 51 auditeurs sur 30 phrases différentes, chaque phrase étant notée par 5 à 6 auditeurs.

12.5 Résultats de l'évaluation perceptive

12.5.1 Cas d'une segmentation phonétique automatique

Considérons tout d'abord les 38 dictionnaires de synthèse obtenus par segmentation automatique. La figure 46 rapporte leurs DMOS en fonction de leur taille. La taille est indiquée en nombre de diphones et suit une **échelle logarithmique**. Pour les enregistrements dédiés, le nom du script de lecture est indiqué verticalement. A chaque type de script est associée une forme de marque, les marques correspondant à la même voix étant reliées par des arcs. Les

voix naturelles de contrôle ont obtenu des DMOS respectifs de 4.7, 4.8 et 4.3, soit un niveau moyen de 4.6. Ce niveau est repris sur la figure 46 par une ligne en pointillés, symbolisant la borne haute de l'évaluation.

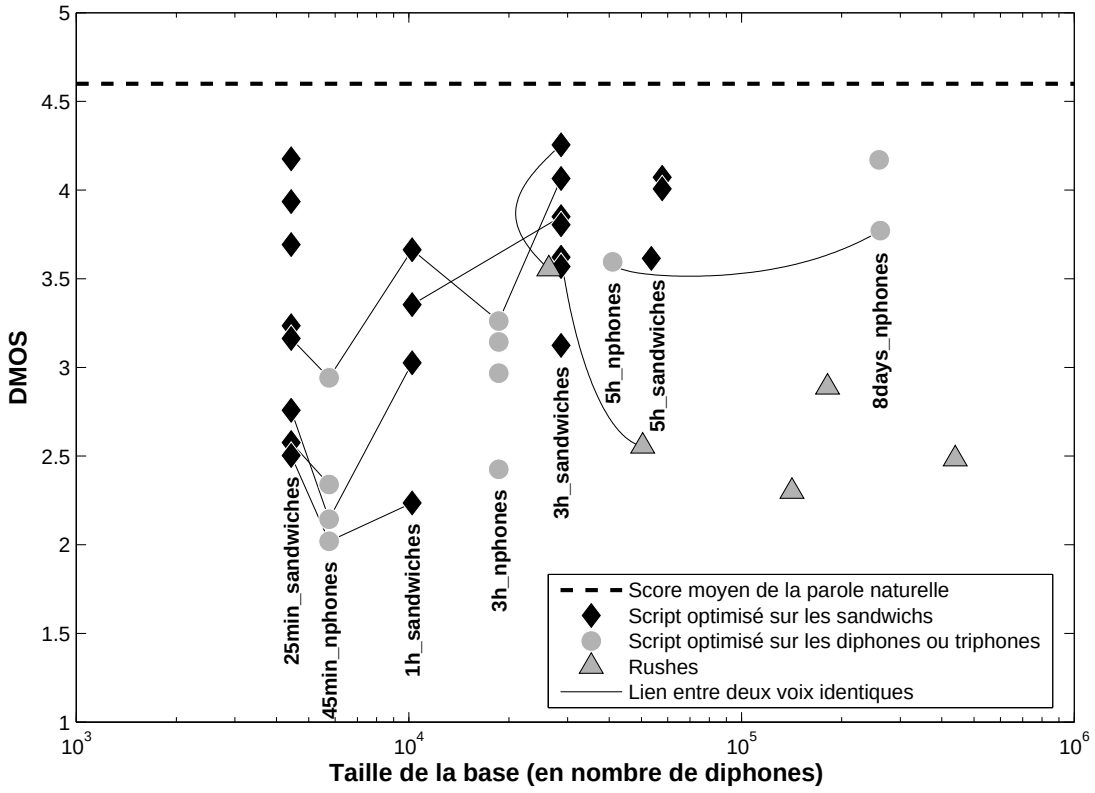


FIGURE 46 – Scores d'opinion moyens mesurés sur les 38 dictionnaires construits par segmentation automatique.

La table 14 détaille les résultats en ajoutant des informations complémentaires sur les voix et les bases de données. Elle rapporte également les intervalles de confiance des DMOS à 95%, qui sont simplement déduits de l'écart-type des notes suivant une hypothèse gaussienne. On constate que ces intervalles sont tous inférieurs à ± 0.19, ce qui nous autorise à dégager des tendances de manière fiable.

Les résultats nous apportent de nombreux enseignements. Tout d'abord ils confirment que, pour une stratégie d'optimisation donnée, **les voix de synthèse tendent à progresser lorsque la taille du script croît**. Le niveau de qualité reste tout de même **fortement dépendant de la voix enregistrée**. Par exemple le script « 25min_sandwiches » conduit à des DMOS très variables, allant de 2.5 à 4.2. Les scores élevés des voix Fouras, Chut et Darkside, construites suivant ce script, sont cependant le fruit de particularités acoustiques. En effet Fouras et Chut sont majoritairement non-voisées ; de tels signaux bruités supportent très bien les concaténations, du fait d'une cohérence temporelle réduite. Darkside présente quant à lui une tessiture très étroite (1.4 dt), ce qui est également avantageux dans le cadre d'une synthèse par corpus. D'une manière générale on remarque que, pour des voix présentant un voisement normal, la qualité de synthèse tend à décroître avec l'amplitude des variations mélodiques. Mais d'autres facteurs prosodiques, non pris en compte dans la mesure de l'écart-type de hauteur, restent probablement prépondérants : régularité de la prosodie entre les phrases de la base, respect du style « lecture neutre », etc.

Avec un DMOS moyen de 2.8 **les rushes procurent des synthèses vocales de qualité très médiocre**, et cela malgré des corpus de grande taille (jusqu'à 11 heures de parole). Les

Voix	Sexe	Nom du script	Nombre de diphones	Durée de parole utile	Hauteur moyenne	Ecart-type de hauteur (en demi-tons)	DMOS	Intervalle de confiance à 95%
Pierre-Yves	male	**25min_sandwiches**	4 429	6 min	113 Hz	2.1 dt	**2.50**	±0.17
Virginia	female			6 min	215 Hz	2.6 dt	**2.54**	±0.17
Marc	male			5 min	125 Hz	2.1 dt	**2.76**	±0.18
Didier	male			6 min	121 Hz	1.8 dt	**3.16**	±0.18
G.I.	male			8 min	116 Hz	1.8 dt	**3.24**	±0.19
Fouras	male			9 min	160 Hz *	2.0 dt *	**3.69**	±0.17
Chut	male			6 min	136 Hz *	2.6 dt *	**3.93**	±0.15
Darkside	male			12 min	97 Hz	1.4 dt	**4.18**	±0.15
Pierre-Yves	male	**45min_nphones**	5 760	9 min	113 Hz	2.1 dt	**2.02**	±0.14
Marc	male			8 min	125 Hz	2.1 dt	**2.14**	±0.15
Virginia	female			10 min	215 Hz	2.6 dt	**2.34**	±0.15
Didier	male			9 min	123 Hz	1.7 dt	**2.94**	±0.17
Pierre-Yves	male	**1h_sandwiches**	10 224	18min	116 Hz	1.8 dt	**2.24**	±0.18
Marc	male			17 min	128 Hz	1.8 dt	**3.03**	±0.18
Sidoo	male			14 min	160 Hz	2.6 dt	**3.35**	±0.17
Didier	male			14 min	120 Hz	1.6 dt	**3.65**	±0.16
Eric	male	**3h_nphones**	18 632	25 min	133 Hz	3.0 dt	**2.42**	±0.16
Annie	female			27 min	226 Hz	3.5 dt	**2.97**	±0.17
Thierry	male			26 min	87 Hz	2.9 dt	**3.14**	±0.17
Didier	male			26 min	126 Hz	2.5 dt	**3.26**	±0.17
Marge	female	**Marge_rushes**	26 328	37 min	297 Hz *	6.3 dt *	**3.56**	±0.18
Ghislain	male	**3h_sandwiches**	28 686	32 min	124 Hz	2.1 dt	**3.12**	±0.18
Chat-Potté	male			40 min	110 Hz	4.2 dt	**3.57**	±0.18
Homer	male			45 min	165 Hz	3.8 dt	**3.60**	±0.16
Loic	male			40 min	113 Hz	4.0 dt	**3.80**	±0.18
Sidoo	male			40 min	178 Hz	2.7 dt	**3.81**	±0.17
Didier	male			36 min	130 Hz	1.8 dt	**4.07**	±0.15
Marge	female			44 min	256 Hz *	7.0 dt *	**4.25**	±0.15
Philippe	male	**5h_nphones**	40 995	1 h 03 min	118 Hz	4.0 dt	**3.59**	±0.15
Homer	male	**Homer_rushes**	50 469	1 h 10 min	185 Hz	4.8 dt	**2.56**	±0.19
Matteo	boy	**5h_sandwiches****	53 571	1 h 14 min	279 Hz	2.1 dt	**3.61**	±0.18
Electra	female		57 768	2 h 01 min	154 Hz	2.9 dt	**4.01**	±0.16
Guy	male		57 768	1 h13 min	116 Hz	3.8 dt	**4.07**	±0.15
PPDA	male	**PPDA_rushes**	141 709	2 h 55 min	88 Hz	3.1 dt	**2.30**	±0.18
Jonathan	male	**Jonathan_rushes**	181 271	3 h 40 min	140 Hz	5.4 dt	**2.89**	±0.19
Philippe	male	**8days_nphones****	261 432	6h 21 min	118 Hz	4.0 dt	**3.77**	±0.17
Agnes	female		259 134	5 h 58 min	177 Hz	3.3 dt	**4.17**	±0.14
Chirac	male	**Chirac_rushes**	438 603	10 h 42 min	105 Hz	5.3 dt	**2.48**	±0.17

* Mesure indicative : Fouras et Chut sont majoritairement non-voisés et Marge présente un voisement très erratique.
** Le script n'a pas été enregistré intégralement par tous les locuteurs, d'où de légères variations sur les nombres de diphones.

TABLE 14 – Caractéristiques principales et scores DMOS des 38 dictionnaires de synthèse obtenus par segmentation automatique.

obstacles inhérents aux rushes ont été détaillés en section 3.4 : erreurs d'annotation, expressivité incontrôlée et couverture linguistique limitée. Leur impact sur la qualité de synthèse finale est décisif, du moins avec le système des Orange Labs.

Enfin, on constate que **les scripts optimisés sur des critères à base de sandwichs offrent d'excellents compromis entre la taille de la base et la qualité de synthèse**. Sur la figure 46 les losanges noirs sont en effet particulièrement haut placés, au-dessus des rushes et des scripts à base de n-phones (du moins à taille comparable). Cette tendance globale est confortée par une comparaison par paire détaillée des voix disponibles dans plusieurs conditions (voir les arcs sur la figure, concernant les voix Didier, Marc, Pierre-Yves, Virginia, Sidoo, Homer, Marge et Philippe). Bien qu'il s'agisse d'une segmentation automatique, les scripts à base de

sandwichs donnent accès à une synthèse de haute qualité avec des enregistrements très réduits : des valeurs de DMOS de l'ordre de 4.0 peuvent être atteintes dès 40 minutes de parole, soit une demi-journée d'enregistrement.

En conclusion, **nos procédés permettent de réduire environ d'un facteur 10 la durée des enregistrements nécessaires à la création d'une voix de synthèse de haute qualité**. L'essentiel de ce gain s'explique par la meilleure densité des scripts en matière de couverture des sandwichs vocaliques. Mais, comme expliqué plus haut, on relève également une efficacité accrue de la phase d'enregistrement [59], liée à un contrôle plus fin des longueurs de phrase et de leur pertinence phonétique.

12.5.2 Cas d'une segmentation phonétique révisée

La révision manuelle de la segmentation phonétique est une tâche très coûteuse en temps humain, mais souvent considérée comme bénéfique pour la synthèse finale. Plusieurs études invitent cependant à nuancer ce propos [Kawai 00][Makashay 00]. La présence dans notre test de dictionnaires avec des niveaux de révision graduels (voir page 140) nous permet d'établir nos propres conclusions.

La table 15 donne le détail des dictionnaires évalués suivant plusieurs niveaux de révision, avec les scores DMOS correspondants. Pour les séquences phonétiques comme pour les marques d'alignement, on indique le taux de modification moyen relativement au mode tout-automatique. **Le niveau de révision manuel/manuel apporte une amélioration par rapport à auto/auto**, les DMOS moyens étant respectivement de 3.51 et 3.31. D'après un test de Student à deux échantillons, cette amélioration est statistiquement significative pour les dictionnaires concernés ($p < 10^{-4}$). La légère dégradation de DMOS observée sur les voix Eric et Loic n'est pas significative (avec $p = 0.1$). Le niveau de révision manuel/auto apporte une légère amélioration par rapport à auto/auto, mais qui n'est pas non plus significative (avec $p = 0.1$). Ceci nous amène à la conclusion que **les imprécisions d'alignement semblent avoir plus d'impact sur la qualité de synthèse que les erreurs dans la séquence phonétique elle-même**.

Ces résultats sont difficilement généralisables, tant ils dépendent du système de synthèse, de la langue, de la voix, de la stratégie experte de segmentation, de l'outil de segmentation automatique et de son implémentation. Un enseignement commun avec [Kawai 00] et [Makashay 00] se dégage tout de même : **la révision manuelle de la segmentation ne joue pas un rôle déterminant dans la création de voix de synthèse de haute qualité**. Dans notre expérience, l'amélioration moyenne du DMOS n'est que de 0.15 pour les dictionnaires construits à partir d'enregistrements dédiés. Seuls les rushes semblent vraiment profiter d'une révision manuelle ; en atteste le score de la voix PPDA, qui passe de 2.30 à 2.84. C'est une conséquence logique du faible contrôle sur le contenu des rushes, qui met considérablement en défaut le processus de segmentation automatique.

12.5.3 Mesures de corrélation

L'évaluation perceptive a permis au total de collecter 7 497 notes, portant sur 1 470 phrases synthétiques différentes. Dans ce paragraphe, les 1 470 scores moyens obtenus par chacune de ces phrases sont confrontés à différents indicateurs objectifs.

59. matérialisée par le ratio $\frac{duree\ de\ parole\ utile}{duree\ des\ enregistrements}$

Voix	Nom du script	Niveau de révision de la séquence phonétique	Niveau de révision de l'alignement	% phonèmes impactés (suppressions, insertions, substitutions)	% marques d'alignement déplacées ($\geq 20ms$)	DMOS	Intervalle de confiance à 95%
Eric	3h_nphones	auto	auto			**2.42**	±0.16
Eric	3h_nphones	**manuel**	**manuel**	7.60%	29.29%	**2.35**	±0.17
Thierry	3h_nphones	auto	auto			**3.14**	±0.17
Thierry	3h_nphones	**manuel**	**manuel**	6.80%	29.67%	**3.33**	±0.17
Didier	3h_nphones	auto	auto			**3.26**	±0.17
Didier	3h_nphones	**manuel**	**manuel**	2.58%	20.06%	**3.69**	±0.15
Homer	3h_sandwiches	auto	auto			**3.60**	±0.16
Homer	3h_sandwiches	**manuel**	**manuel**	0.53%	20.36%	**3.74**	±0.18
Loic	3h_sandwiches	auto	auto			**3.80**	±0.18
Loic	3h_sandwiches	**manuel**	**manuel**	0.95%	27.29%	**3.78**	±0.17
PPDA	PPDA_rushes	auto	auto			**2.30**	±0.18
PPDA	PPDA_rushes	**manuel**	auto	3.44%	3.57%	**2.35**	±0.18
PPDA	PPDA_rushes	**manuel**	**manuel**	3.44%	23.13%	**2.84**	±0.18
Philippe	8days_nphones	auto	auto			**3.77**	±0.17
Philippe	8days_nphones	**manuel**	auto	2.52%	1.59%	**3.97**	±0.15
Philippe	8days_nphones	**manuel**	**manuel**	2.52%	27.35%	**4.07**	±0.15
Agnes	8days_nphones	auto	auto			**4.17**	±0.14
Agnes	8days_nphones	**manuel**	auto	2.56%	1.55%	**4.13**	±0.14
Agnes	8days_nphones	**manuel**	**manuel**	2.56%	18.03%	**4.27**	±0.13

TABLE 15 – Scores DMOS des dictionnaires disponibles suivant plusieurs niveaux de révision phonétique.

Par un procédé analogue à celui de la section 7.3, nous mesurons la corrélation entre le DMOS et les taux de couverture de différentes unités. Les unités présentant les meilleurs taux de corrélation sont rapportées en table 16. **Les sandwichs arrivent en tête du panel d'unités, avec un niveau de corrélation atteignant 0.41** dans le cas `LRsemirobustes_13contextes`. Au premier abord ce niveau peut paraître assez faible. Mais il est en réalité très proche de celui obtenu par le coût de sélection, qui représente pourtant la meilleure projection connue de la perception sur un espace objectif. Nous mesurons en effet une corrélation de 0.43 entre le coût de sélection et le DMOS. **La perception de la synthèse vocale est donc presque aussi bien décrite par le taux de couverture en sandwichs vocaliques que par le coût de sélection**, malgré de nombreuses contraintes liées à l'utilisation d'un « taux de couverture linguistique » : non prise en compte des facteurs acoustiques, approximation binaire des écarts contextuels, approximation binaire de la fragilité des phonèmes, approximation binaire de la redondance, etc. Ce résultat rejoint celui du 7.3, où nous avions relevé de très bonnes corrélations entre le coût de sélection et les taux de couverture en sandwichs vocaliques.

La corrélation de 0.43 entre le DMOS et le coût de sélection est significativement plus faible que celles mesurées chez [Chu 01a] et [Toda 04], qui avoisinent 0.8. Il y a deux raisons principales à cela. D'une part les conditions expérimentales sont très différentes, en particulier la langue et le système TTS, et justifient des écarts de fonctionnement importants. D'autre part notre expérience a porté sur un matériel de test bien plus diversifié : nous avons utilisé de nombreuses voix de synthèse, de qualité et de niveau de segmentation variés, ce qui tend à

Type d'unité		**Corrélation au DMOS**	
sandwichs, 1-grammes,	liquides semi-robustes, 13 contextes	**0.411**	±0.039
sandwichs, 1-grammes,	liquides fragiles, 13 contextes	**0.384**	±0.040
phones, 3-grammes,	13 contextes	**0.376**	±0.041
sandwichs, 1-grammes,	liquides semi-robustes, 4 contextes	**0.371**	±0.041
phones, 3-grammes,	4 contextes	**0.361**	±0.041
sandwichs, 2-grammes,	liquides semi-robustes, 0 contexte	**0.358**	±0.041
sandwichs, 1-grammes,	liquides fragiles, 4 contextes	**0.357**	±0.041
sandwichs, 2-grammes,	liquides semi-robustes, 13 contextes	**0.355**	±0.041
phones, 4-grammes,	0 contexte	**0.353**	±0.041
sandwichs, 2-grammes,	liquides semi-robustes, 4 contextes	**0.352**	±0.041
sandwichs, 2-grammes,	liquides fragiles, 0 contexte	**0.349**	±0.042
sandwichs, 1-grammes,	liquides semi-robustes, 0 contexte	**0.347**	±0.042
phones, 4-grammes,	4 contextes	**0.347**	±0.042
syllabes, 1-grammes,	13 contextes	**0.344**	±0.042
phones, 4-grammes,	13 contextes	**0.343**	±0.042
phones, 2-grammes,	13 contextes	**0.342**	±0.042
phones, 5-grammes,	0 contexte	**0.340**	±0.042

TABLE 16 – Niveaux de corrélation entre le DMOS et les taux de couverture de différentes unités, avec les intervalles de confiance à 95% [Jolion 06]. Seules les unités présentant les meilleures corrélations sont reportées.

accroître la variabilité des résultats et donc à réduire les niveaux de corrélations.

Cette diversité procure à nos mesures une pertinence statistique appréciable, dont nous ne disposions pas encore lors des premières publications mentionnant les sandwichs vocaliques et leur corrélation à la perception humaine [Cadic 09] [Boidin 09b]. A noter tout de même que les sandwichs y étaient déjà classés au premier rang des unités linguistiques.

12.6 Discussion

D'après l'étude précédente, **les scripts de lecture issus de nos travaux permettent la création de voix de synthèse de haute qualité en moins d'une journée**, incluant les enregistrements et le post-traitement des données avec une segmentation automatique. Des scores perceptifs de l'ordre de 4.0, équivalents à ceux des voix de synthèse créées suivant le procédé traditionnel sur 8 jours, ont en effet été atteints de cette manière.

Comme expliqué en fin de partie précédente, les scripts constitués suivant notre algorithme de construction de phrases sont trop expérimentaux pour se démarquer, en matière de densité, de ceux obtenus par condensation. Il est donc logique que, à critère comparable, les scripts construits n'offrent pas une meilleure synthèse que les scripts condensés. Mais il est important de signaler que nous n'observons pas non plus de dégradation, malgré les difficultés de lecture rencontrées par certains locuteurs. **Le manque de cohérence sémantique des phrases construites, responsable parfois d'un certain ralentissement des enregistrements, semble sans conséquence sur la qualité des voix de synthèse.** En particulier il ne met pas en défaut le style « lecture neutre », qui est par définition indépendant de la sémantique.

La couverture en sandwichs vocaliques a joué un rôle important dans la densification du procédé de création de voix. L'un de ses objectifs principaux est de préserver des concaténations un ensemble de phonèmes dits fragiles. Il convient de vérifier que cette protection est effective. La figure 47 indique, pour chaque voix et sur les 147 phrases du corpus de test, la proportion des

phonèmes fragiles [60] qui ont fait l'objet d'une concaténation. A taille de script donnée, ce taux de fragmentation est plus faible pour les scripts optimisés sur des critères à base de sandwichs, ce qui confirme la pertinence de ces unités pour la préservation des phonèmes fragiles. On observe au passage que, pour un script donné, le taux de fragmentation dépend peu de la voix ; cela signifie que, avec le coût de sélection utilisé dans notre système, les composantes acoustiques ont un faible impact sur le positionnement des concaténations.

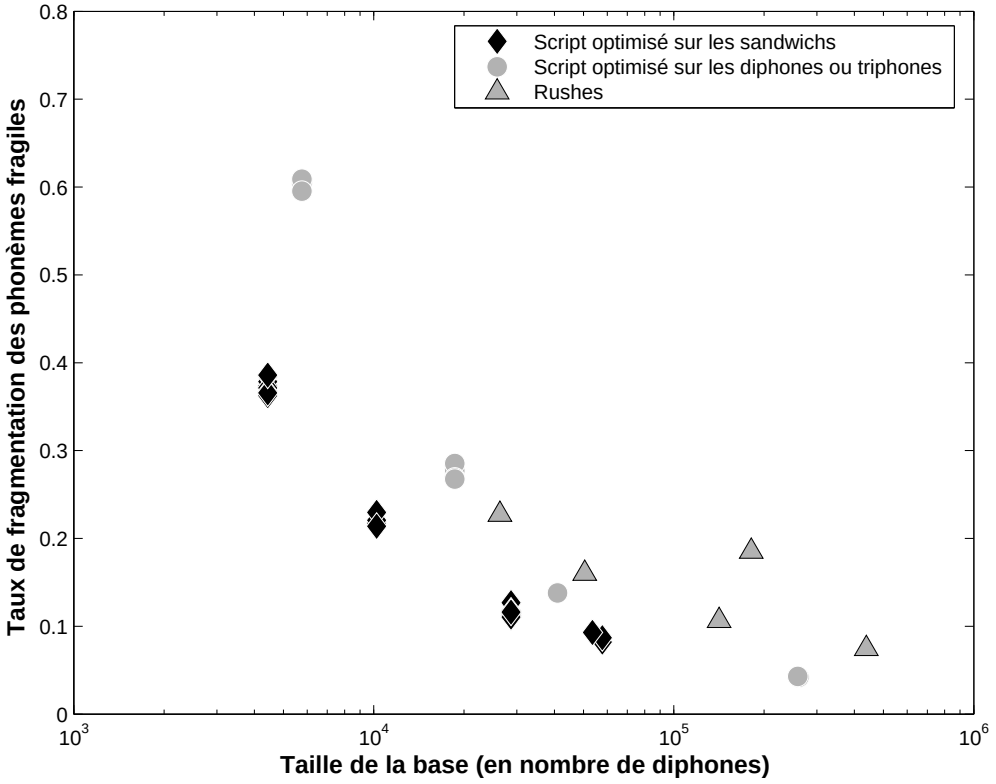


FIGURE 47 – Taux moyen de fragmentation des phonèmes fragiles, avec chacun des 38 dictionnaires segmentés automatiquement et sur les 147 phrases du corpus de test.

Au fil des expériences, il nous a semblé que la **couverture en sandwichs vocaliques** n'était **pas complètement adaptée aux scripts de lecture très courts**, c'est-à-dire en dessous de 10 000 diphones, soit 15 minutes de parole. On constate par exemple qu'il manque aux dictionnaires basés sur « 25min_sandwichs » et « 1h_sandwichs » respectivement 4.4% et 1.3% des diphones hors contexte observés dans le corpus de référence [61]. Dans un système de synthèse par diphones comme le nôtre, ces lacunes sont problématiques. Même si certaines d'entre elles peuvent être comblées par l'utilisation de diphones proches, comme /œ̃-ã/ à la place de /ɛ̃-ã/, il subsiste de nombreux « trous » acoustiques gênants. Une simple prise en compte de la couverture des diphones hors contexte pourrait certainement améliorer les résultats sur ces petits scripts, avec un impact très modéré sur leur longueur.

A l'inverse, nous pouvons nous interroger sur les **perspectives de nos procédés pour la création de grosses bases de données**, reposant sur plusieurs jours d'enregistrements. Suffit-il de prolonger nos scripts de lecture, suivant les mêmes procédés de construction ou de condensation, pour produire des voix de synthèse de DMOS supérieurs, qui seraient donc indiscernables de la parole neutre naturelle ? Faute d'en avoir fait l'expérience nous ne pourrons pas répondre de manière tranchée. Mais il est à peu près certain que nous toucherions

60. pour cette mesure nous avons considéré les liquides comme semi-robustes
61. donc pondérés par les fréquences d'apparition

très vite à plusieurs limites. La première est liée à la taille du corpus de référence. Nous avons vu au paragraphe 10.3.1 que, lors de la constitution d'un script de lecture, l'un des critères d'arrêt principaux était le nombre d'occurrences des unités marginales. Sur cet aspect, le script « 5h_sandwiches » est déjà de taille maximale ; son extension supposerait donc avant tout d'élargir le corpus de référence. Mais cela poserait le problème plus fondamental de l'universalité du corpus [Van Santen 97a]. La seconde limite tient au glouton lui-même, dont les performances par rapport à l'optimum décroissent lorsqu'on atteint des niveaux de couverture très élevés. L'algorithme de construction de phrase, également basé sur une approche gloutonne, ne permettrait pas d'empêcher ce déclin. Enfin, la troisième et dernière limite porte sur les sandwichs vocaliques eux-mêmes. Leur définition, qui exclut la caractérisation contextuelle des consonnes et y libéralise les concaténations, deviendrait probablement insuffisante pour des taux de couverture très élevés. Même en utilisant davantage les critères à base de bigrammes, les clusters consonnantiques resteraient négligés au profit des voyelles, ce qui interdirait peut-être le franchissement d'un pallier de qualité résiduel. Pour toutes ces raisons nous n'envisageons pas, à l'heure actuelle, d'augmenter les durées d'enregistrement en prolongeant nos scripts de lecture.

Pour finir, il est important de rappeler que l'évaluation précédente a porté principalement sur le naturel des voix de synthèse. Mais qu'en est-il de l'intelligibilité ? En occultant la transcription textuelle des phrases évaluées, notre protocole de test tient partiellement compte des problèmes d'intelligibilité. Ces problèmes apparaissent essentiellement sur les voix à base de rushes, mais parfois aussi sur d'autres voix de faible qualité. Ce point pourrait être approfondi. **Pour affiner nos résultats, en particulier ceux portant sur la comparaison des niveaux de révision, il serait intéressant de les compléter d'une mesure de scores d'intelligibilité**. Le protocole SUS, exposé en 4.2.1, pourrait être utilisé à cet effet.

13 Application à la multi-expressivité

13.1 De la lecture neutre à la lecture expressive

Nous nous sommes cantonnés jusqu'à présent au style « lecture neutre » défini en section 1.3. Ce style fait référence à une lecture désengagée, reposant uniquement sur des automatismes de déchiffrage lexical ou syntaxique. Dans ce cadre, nous avons proposé en section 5.2 un pavage des contextes qui permet d'expliquer l'essentiel des mouvements prosodiques avec une dizaine de classes seulement. Pour garantir la pertinence et la régularité de la parole synthétique finale, les enregistrements de voix doivent alors se conformer à ce découpage : ils ne doivent pas comporter d'excursion prosodique non expliquée par ce modèle.

Cette contrainte de modélisation n'interdit pas l'introduction d'une dose d'expressivité dans les voix enregistrées. En effet, le découpage contextuel peut supporter certaines composantes paralinguistiques sans remettre en cause la cohérence de la synthèse, en particulier si ces composantes sont relativement uniformes ou indépendantes du contexte : voix basse, voix sensuelle, voix triste, etc. Elles sont prises en compte automatiquement dans la synthèse par corpus, par simple effet de bord. Il ne s'agit donc plus à proprement parler d'une lecture neutre puisque la parole véhicule, au-delà du contenu strictement linguistique, des informations sur son émetteur (tristesse, sensualité, etc.).

Nous avons expérimenté l'ajout de telles composantes dans de nombreuses voix de synthèse. L'évaluation perceptive de la section précédente comportait ainsi, outre les voix à base de rushes, quelques voix légèrement colorées sur le plan expressif : Electra (lecture sensuelle), Homer (lecture niaise), G.I (lecture bestiale), Chut (lecture à voix basse)... Les bons résultats

obtenus par ces voix confirment la compatibilité de certaines composantes paralinguistiques avec la technologie de synthèse par corpus.

Toutefois **le modèle contextuel sous-jacent ne peut supporter qu'un unique patron prosodique**. Autrement dit, même s'il ne s'agit pas d'une lecture neutre au sens strict, la correspondance entre les paramètres acoustiques et les facteurs contextuels doit rester univoque. Or les patrons prosodiques régissant la parole humaine sont susceptibles d'évoluer régulièrement à travers un flux verbal, en fonction de facteurs environnementaux variés : contexte dialogique, contexte narratif, état émotif... Cette variété, que nous appelons « **multi-expressivité** » ne peut pas être prise en compte par notre procédé de création de voix. La création d'un dictionnaire à partir d'enregistrements comportant divers styles expressifs aboutit à une voix de synthèse erratique, dont la richesse prosodique est incontrôlée. C'est par exemple le cas de la voix Jonathan dans l'évaluation précédente (DMOS=2.9).

13.2 Un bref état de l'art de la multi-expressivité

Au cours des dernières années, de nombreux travaux ont porté sur la diversification expressive des voix de synthèse. Leur objectif est d'offrir des alternatives au style prosodique par défaut, par la prise en compte de composantes expressives variées. Ces composantes peuvent être assez globales, comme l'expression d'une émotion, ou plus locales, comme l'emphase sur un groupe de mot. Pour cela le texte d'entrée doit être enrichi de considérations paralinguistiques, ce qui suppose la disponibilité d'une intelligence de haut-niveau. Cette intelligence peut découler d'une intervention humaine (placement manuel de balises textuelles) ou bien être fournie par un système de type concept-to-speech (voir page 21).

Pour la restitution de composantes multi-expressives en synthèse vocale, il existe deux grandes familles d'approches : les approches numériques et les approches symboliques.

Les **approches numériques** ont pour ambition de modéliser les réalisations acoustiques qui accompagnent les différentes composantes expressives. Elles reposent généralement sur la disponibilité de corpus de parole illustratifs et sur des techniques d'apprentissage automatique. Il en ressort des fonctions de déformation prosodique qui servent à corriger le signal synthétique. Par exemple dans [Raux 03] des modèles intonatifs de l'emphase sont appris sur un corpus acoustique comportant 968 mots accentués. Le placage sur le signal synthétique des courbes intonatives prédites est ensuite effectué avec une technique LPC. Beller traite quant à lui de composantes émotives, avec plusieurs niveaux d'intensité [Beller 07] : joie douce, joie explosive, tristesse contenue, tristesse larmoyante, dégoût, etc. De manière experte, l'auteur initie un premier jeu de règles contextuelles permettant la transformation de la parole neutre en parole expressive. Ces règles sont ensuite affinées automatiquement à l'aide d'un réseau bayésien opérant sur un corpus d'apprentissage multi-locuteurs. Les paramètres acoustiques pris en compte sont la hauteur, la durée de syllabe, l'intensité et le degré de réduction. Le placage est effectué avec une technique de Vocoder de phase. Dans [Yamagishi 03], les auteurs appliquent les outils de la synthèse par HMM à trois corpus de 500 phrases, correspondant chacun à une émotion précise. Pour accroître la cohérence à long-terme de la composante émotionnelle dans la synthèse HMM, Hirose introduit de nouveaux paramètres dans la modélisation HMM [Hirose 06].

Si les techniques numériques ont permis des progrès significatifs dans la prise en compte de composantes expressives variées, aucune ne parvient véritablement à produire des signaux synthétiques réalistes. Elles sont limitées à la fois par le manque de précision des modèles prosodiques et par les dégradations dues aux modifications acoustiques.

Les **approches symboliques** se veulent plus pragmatiques. Conformément au paradigme de la synthèse par corpus, elles reposent uniquement sur l'annotation d'unités porteuses de composantes expressives dans le corpus de parole. [Black 03] et [Syrdal 08] rapportent des résultats prometteurs à partir d'annotations d'emphase, de style prosodique, ou encore d'actes de paroles (impératif, interrogatif, assertif...). Les unités acoustiques ainsi annotées viennent simplement enrichir le corpus de parole ; l'étape de sélection, guidée par les nouveaux marqueurs symboliques, traite ces unités comme des boîtes noires. On peut dès lors envisager de créer une voix de synthèse pour chaque patron prosodique que l'on souhaite couvrir. [Iida 02] expérimente ainsi, pour deux locuteurs de sexes différents, la création de 4 voix de synthèse correspondant à des émotions différentes : neutre, colère, joie, tristesse. [Black 03] insiste toutefois sur les limites combinatoires de l'approche symbolique, liées à la lourdeur du procédé de création de voix en synthèse par corpus. L'auteur conclut à la nécessité de recourir à des techniques numériques complémentaires, pour diviser le problème sur un plan acoustique et ainsi casser cette combinatoire. Il envisage donc une approche mixte, utilisant à la fois des concepts numériques et symboliques. On peut à ce sujet citer [Eide 04], qui rapporte une expérience connue sous le nom de « Good news, Bad news ». Dans ces travaux, un petit ensemble de phrases est enregistré suivant plusieurs styles : styles associés à la communication d'une bonne nouvelle ou d'une mauvaise nouvelle, style interrogatif et emphase contrastive. Ces mini-corpus sont utilisés à la fois pour apprendre des modèles prosodiques dédiés et pour enrichir l'ensemble des unités candidates à la sélection.

En conclusion, les approches symboliques décrites dans l'état de l'art ne sont que d'un secours limité par rapport aux approches numériques. Elles donnent certes accès à des signaux de haute qualité, mais en contrepartie elle font preuve d'une grande rigidité. Le recours obligatoire à une abondante matière sonore les rend lourdes et onéreuses. Ce facteur empêche d'atteindre la diversité de styles requise par la multi-expressivité, trois ou quatre composantes expressives ne suffisant pas à rendre compte de la richesse de la parole humaine.

13.3 Expériences

13.3.1 Acquisition de données multi-expressives

Les procédés de création de voix exposés dans ce document ouvrent de nouvelles perspectives à l'approche symbolique de la multi-expressivité. En réduisant considérablement la durée d'enregistrement nécessaire à la création d'une voix de haute qualité, ils permettent d'envisager une réelle diversification des styles expressifs. Suivant le script « 5h_sandwiches » nous avons ainsi pu créer, avec un même comédien professionnel, près de vingt dictionnaires de synthèse correspondant à des styles de lecture différents ! Chaque dictionnaire a nécessité en moyenne 5 heures d'enregistrements, soit une journée en studio, et a fait l'objet d'une segmentation phonétique automatique.

La table 17 présente les caractéristiques principales de **16 styles expressifs enregistrés avec la voix de Guy**[62]. Nous avons exclu certains autres styles enregistrés par ce même comédien, soit parce qu'ils correspondent à des timbres de voix ludiques sans rapport direct avec la multi-expressivité, soit parce que le style mal défini et les consignes floues ont conduit à des enregistrements trop hétérogènes pour la synthèse par corpus. Le choix de la palette expressive a été entièrement guidé par l'expertise. Cette palette couvre quelques composantes émotives (Guy_triste, Guy_enjoue, Guy_fort) mais aussi et surtout de nombreux styles prosodiques qui relèvent de la narration. Nous avons pour cela tenu compte du niveau de tension, du niveau

62. La voix de Guy utilisée dans l'évaluation perceptive de la section précédente est en réalité Guy_narratif_calme

de didactisme, des passages dialogiques, mais aussi des schémas prosodiques standards. Par exemple Guy_virgule et Guy_point ont vocation à se succéder d'un groupe de souffle à l'autre, de manière à reproduire l'alternance naturelle entre phrases prosodiques montantes et descendantes. D'autres styles ont un rôle plus expérimental, comme Guy_dictee et Guy_insiste. Le premier correspond à une énonciation lente et sur-articulée, à la façon d'une dictée scolaire. Le second consiste en une emphase quasi-systématique des mots non grammaticaux ; il est destiné à remplacer localement le style neutre dans la synthèse, de manière à restituer des emphases ponctuelles. Le CD d'accompagnement illustre la diversité des enregistrements en rapportant un extrait de chaque style.

N°	Nom de la voix	Hauteur moyenne et écart-type	Durée moyenne des voyelles et écart-type	Commentaire
1	**Guy_grave**	82 ±13 Hz	77 ±40 ms	Grave et monotone
2	**Guy_triste**	92 ±13 Hz	81 ±39 ms	Empreint de tristesse
3	**Guy_narratif_calme**	124 ±26 Hz	77 ±37 ms	Lecture neutre
4	**Guy_narratif_tendu**	133 ±21 Hz	75 ±35 ms	Narration avec tension
5	**Guy_dialogue**	151 ±21 Hz	79 ±45 ms	Voix tendue, comme en conversation
6	**Guy_enjoue**	153 ±39 Hz	84 ±35 ms	Voix enjouée et souriante
7	**Guy_fort**	197 ±31 Hz	87 ±42 ms	Empreint de colère
8	**Guy_dictee**	124 ±39 Hz	142 ±89 ms	Style lent et sur-articulé
9	**Guy_didactique**	139 ±42 Hz	102 ±40 ms	À la façon d'un discours politique
10	**Guy_insiste**	145 ±31 Hz	110 ±51 ms	Emphase sur tous les mots
11	**Guy_contraste_bas**	99 ±15 Hz	78 ±35 ms	Ton offusqué
12	**Guy_enumeration**	136 ±28 Hz	87 ±47 ms	A la façon d'une énumération
13	**Guy_suspensif**	109 ±17 Hz	73 ±32 ms	Neutre, fins de phrases interrompues
14	**Guy_point**	113 ±19 Hz	80 ±36 ms	Affirmatif, terminaisons en points
15	**Guy_virgule**	108 ±21 Hz	75 ±33 ms	Registre bas, terminaisons montantes
16	**Guy_voix_basse**	85 ±12 Hz*	76 ±45 ms	Voix basse

* Mesure indicative, Guy_voix_basse étant majoritairement non-voisée.

TABLE 17 – Liste des 16 principaux styles expressifs enregistrés avec la voix de Guy.

D'une manière générale, les enregistrements se sont révélés plus difficiles que prévu. Le principal problème a été le maintien d'un même style tout au long d'une journée d'enregistrement. La distinction entre certains styles repose en effet sur des nuances prosodiques subtiles, délicates à appréhender en dehors d'un contexte narratif adéquat. Il en résulte que **certains styles « extrémaux » jouent le rôle d'attracteurs prosodiques**. C'est par exemple le cas de Guy_narratif_calme, Guy_fort, ou encore Guy_voix_basse, dont les caractéristiques sont particulièrement faciles à cerner. Les enregistrements des autres styles ont tendance à dériver au fil des phrases vers un style extrémal prosodiquement proche. Une supervision très attentive et des ré-enregistrements partiels ont permis de limiter ces fluctuations au sein d'une même base.

De par son professionnalisme, le comédien n'a pas particulièrement souffert de l'effort vocal. La seule exception a porté sur Guy_fort, style très intense par définition, que nous avons dû fractionner en deux journées d'enregistrement pour ménager la voix du comédien.

13.3.2 Observation de la distribution acoustique des styles

Les valeurs moyennes reportées dans la table 17 offrent une vision globale de la dispersion des 16 styles. Mais **pour une différenciation plus rigoureuse des patrons prosodiques, il faut entrer dans le détail des contextes syntaxico-prosodiques**. Pour cela nous nous sommes focalisés sur l'un des symboles contextuels issus de la classification proposée en 5.2. Ce symbole $\mathcal{C}$ fait partie du jeu de 13 symboles utilisé à plusieurs reprises dans notre étude. Il correspond à la description contextuelle suivante : « voyelle finale de mot non grammatical et non situé en fin de phrase ». Nous l'avons choisi car il fournit des résultats visuellement intéressants ; mais d'autres contextes permettent des observations tout à fait similaires.

Nous avons rassemblé, sur l'ensemble des 16 bases, toutes les voyelles[63] apparaissant dans ce contexte $\mathcal{C}$. Pour chacune des voyelles, nous disposons de quatre mesures acoustiques : hauteur de début, hauteur de milieu, hauteur de fin et durée de phonème. Des expressions logarithmiques de ces quatre indicateurs ont été utilisées afin d'équilibrer les distributions et de s'approcher de la perception humaine. Les données ont été centrées. Par une analyse factorielle discriminante, nous avons extrait le plan acoustique qui sépare au mieux les 16 styles. Ce plan est caractérisé par deux axes factoriels présentant un pouvoir discriminant maximal. Ils sont obtenus par combinaisons linéaires des quatre indicateurs acoustiques de départ, leurs coefficients respectifs étant $(1.67, -0.15, 0.44, 0.04)$ et $(-1.21, 0.77, 0.49, 1.13)$[64]. Le premier axe factoriel (ou « dimension acoustique n°1 ») correspond donc plus ou moins à la hauteur de début, tandis que le second axe factoriel (ou « dimension acoustique n°2 ») est plus composite.

La figure 48 présente, dans ce nouveau système de représentation, les distributions acoustiques des 16 bases. Chaque distribution est représentée par son ellipse de Mahalanobis, dimensionnée à une fois l'écart-type suivant ses deux directions principales, ce qui englobe environ 40% des réalisations dans l'hypothèse multi-gaussienne. Les ellipses des différentes bases sont numérotées conformément à la table 17. Dans un but de comparaison la figure rapporte également la distribution de la base Jonathan, en se limitant toujours aux voyelles qui sont dans le contexte $\mathcal{C}$. Cette base de rushes (voir page 139) consiste en des lectures professionnelles et expressives de livres variés. Il s'agit donc d'une référence en matière de narration. Comme pour Guy ses données ont été centrées, ce qui permet au passage d'annuler l'écart acoustique moyen entre les deux voix.

On peut tirer de la figure 48 quelques enseignements importants. Tout d'abord il semblerait que **certains styles enregistrés par Guy relèvent de la caricature**. En effet plusieurs ellipses s'écartent significativement du spectre prosodique couvert par Jonathan. Ce constat n'est pas surprenant pour des styles comme Guy_dictee (n°8) ou Guy_voix_basse (n°16), qui ne sont censés jouer qu'un rôle très marginal dans la narration ; il l'est beaucoup plus pour Guy_triste (n°2), Guy_fort (n°7) ou encore Guy_enjoue (n°6), styles que nous imaginions pourtant utiles. Si ces derniers styles peuvent tout de même être adaptés à des situations narratives particulières, il semble que l'essentiel d'une narration expressive naturelle repose sur des nuances stylistiques plus subtiles que ces simples caricatures. Ceci rejoint le problème des « attracteurs prosodiques » évoqués un peu plus haut : la caricature est malheureusement un défaut naturel dans ce type d'exercice.

D'un autre côté, on constate que l'espace prosodique de la voix Jonathan est bien couvert par un dizaine de styles de la voix Guy. Mais cette **couverture est très imprécise**. Les ellipses de Guy se chevauchent largement, au point que certaines d'entre elles sont presque

63. Cette restriction aux voyelles découle des choix de la section 5.1.

64. Ces vecteurs ne sont pas orthogonaux au sens de la base de départ mais au sens des données, c'est-à-dire que les projections des données sur ces deux vecteurs ne sont pas corrélées.

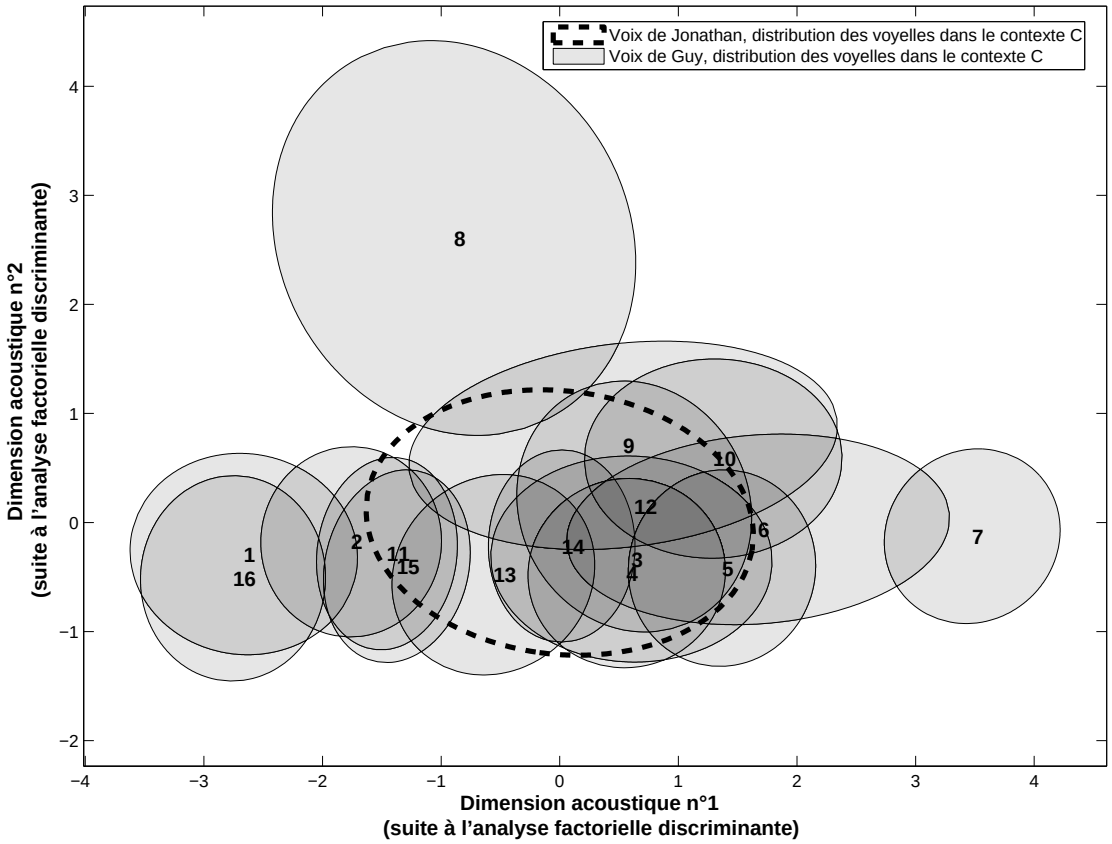


FIGURE 48 – Distribution acoustique des 16 styles de Guy, dans un contexte syntaxico-prosodique donné. La distribution de la voix très expressive de Jonathan dans ce même contexte est reportée en pointillés.

incluses dans d'autres (voir par exemple Guy_narratif_tendu -n°4- et Guy_narratif_calme -n°3-). Ceci remet en cause non seulement la définition des styles, mais aussi la précision de leurs enregistrements. En effet chaque style expressif a inévitablement fluctué au fil des enregistrements. Les contours définis pour chacun se sont par conséquent estompés, au point de les rendre indiscernables. Ainsi les distributions, bien que plus étroites que Jonathan, restent relativement étendues. Dans le contexte $\mathcal{C}$, on relève des écart-types minimaux de $1.5dt$ en hauteur et $23ms$ en durée. La table 18 rapporte plus généralement l'erreur de prédiction RMSE de chaque style suivant notre découpage contextuel en 13 symboles (voir section 5.2). Une RMSE de 1 signifie que la prosodie du locuteur est indépendante du découpage, tandis qu'une RMSE de 0 indique une corrélation parfaite. On trouve pour nos 16 styles des valeurs comprises entre 0.771 et 0.896, ce qui est significativement supérieur aux mesures de la figure 13 (avec 13 classes). Cela signifie que les styles enregistrés ne s'accordent pas très bien avec notre découpage contextuel, même s'il y a clairement des progrès par rapport à la voix de Jonathan (RMSE=0.947). Pour un contrôle fin des voix de synthèse, les symboles contextuels devraient permettre une description quasi-complète des mouvements prosodiques enregistrés ; dans ce cas idéal, les ellipses seraient réduites à des points.

N°	Nom de la voix	RMSE
1	**Guy_grave**	0.881
2	**Guy_triste**	0.836
3	**Guy_narratif_calme**	0.821
4	**Guy_narratif_tendu**	0.840
5	**Guy_dialogue**	0.837
6	**Guy_enjoue**	0.874
7	**Guy_fort**	0.849
8	**Guy_dictee**	0.891
9	**Guy_didactique**	0.881
10	**Guy_insiste**	0.848
11	**Guy_contraste_bas**	0.807
12	**Guy_enumeration**	0.866
13	**Guy_suspensif**	0.866
14	**Guy_point**	0.773
15	**Guy_virgule**	0.771
16	**Guy_voix_basse**	0.896*

* Mesure indicative, Guy_voix_basse étant majoritairement non-voisée.

TABLE 18 – Erreur de prédiction RMSE de notre découpage contextuel pour chacun des styles enregistrés par Guy.

13.3.3 Rendu final

Nous avons longuement expérimenté les dictionnaires de synthèse issus des enregistrements de Guy, en les appliquant à la synthèse de textes narratifs complets. Quelques exemples sonores de telles synthèses sont disponibles sur le CD d'accompagnement. Les styles ont été attribués de manière experte aux différents groupes de souffle, en s'attachant à optimiser le rendu global. Cette restriction à un unique style par groupe de souffle nous est imposée par le fonctionnement actuel du système de synthèse. Dans la pratique elle ne nous éloigne pas trop du découpage stylistique effectué naturellement par un narrateur. A terme, une légère adaptation du module de sélection permettrait de dépasser cette limite.

D'une manière générale, la diversité des styles disponibles a grandement facilité la tâche. Mais nous avons rencontré plusieurs difficultés qui nous ont empêchés d'atteindre l'expressivité d'une narration naturelle. Ces difficultés sont la conséquence logique des défauts relevés dans l'analyse objective précédente.

Tout d'abord, **certains styles ne peuvent pas être introduits de manière douce dans la synthèse**. Il s'agit des styles caricaturaux observés sur la figure 48. Leur utilisation suscite des « surprises prosodiques » qui trouvent difficilement une justification narrative. La quasi-totalité de nos essais de narrations a donc été construite avec seulement 6 ou 7 styles centraux.

Ensuite, **l'imprécision de la couverture prosodique s'est révélée particulièrement gênante**. Le flou des distributions constaté dans le contexte $\mathcal{C}$ est en effet décuplé lorsqu'on considère une séquence complète de phonèmes et symboles contextuels. Il en résulte une incertitude prosodique à l'échelle du groupe de souffle, qui a deux conséquences. D'une part elle crée une confusion entre les styles, ce qui complique leur attribution, d'autre part elle provoque de fréquentes incohérences prosodiques dans le rendu d'un style.

Ces aléas nous ont contraints à recourir à des outils de « synthèse assistée par opérateur ». Ces outils permettent à un opérateur de retraiter le signal synthétique de différentes manières, par exemple en intervenant sur la séquence d'unités sélectionnées[65]. Grâce à eux nous avons pu

65. http://baratinoo.elibel.tm.fr/spo (codes d'accès requis)

rectifier en quelques clics les motifs prosodiques inconsistants constatés sur certains groupes de souffle. Ces outils nous ont également permis de corriger d'autres défauts de la synthèse vocale, comme des distorsions aux concaténations ou encore des erreurs de transcription phonétique. D'une manière générale, **l'objectif de multi-expressivité est apparu incompatible avec les petites imperfections de la synthèse par corpus**, imperfections qui passent pourtant inaperçues dans un contexte de lecture neutre. Même à l'intérieur d'un style la barre qualitative est rehaussée, ce qui dans notre cas rend ce type de retouche incontournable.

Grâce à ce retraitement de la synthèse, **le rendu final nous semble de bonne qualité. la narration est nettement plus crédible qu'avec une voix de synthèse monolithique**, c'est-à-dire limitée au traditionnel style « lecture neutre ». Néanmoins des artefacts demeurent, tant sur le plan prosodique que sur le plan acoustique, et nuisent au naturel global du flux synthétique. Par ailleurs la restriction, dans un souci d'homogénéité, à quelques styles centraux tend à appauvrir la narration : le spectre prosodique de notre voix multi-expressive reste plus réduit que celui d'un narrateur professionnel.

13.3.4 Perspectives

La multi-expressivité constitue selon nous la prochaine rupture technologique dans le domaine de la synthèse vocale. En couvrant plusieurs variantes expressives d'une même voix de synthèse elle devrait permettre, dans un futur proche, une vocalisation crédible de textes narratifs. Si les résultats des premières expériences nous confortent dans cette opinion, ils mettent également en avant plusieurs difficultés. Pour les contourner nous envisageons de travailler sur les thématiques suivantes.

Tout d'abord il faut **revoir le pavage de l'espace multi-expressif**, en se concentrant sur la détermination de styles centraux, non caricaturaux. Une plus grande finesse est requise pour reproduire les nuances expressives de la parole humaine. Ceci pose évidemment le problème de la précision des enregistrements, puisqu'il n'est pas envisageable de demander à un comédien de respecter, tout au long d'une journée d'enregistrement, des nuances prosodiques aussi fines que celles requises dans la palette de styles. Il faut donc poursuivre les efforts de recherche, en quête d'un meilleur compromis entre les contraintes d'enregistrement d'une part et la finesse de la couverture multi-expressive d'autre part. Des techniques d'acquisition plus efficaces devront être imaginées. On pourra par exemple accroître la supervision, mettre en place une vérification temps-réel de la prosodie, faire écouter au comédien le patron idéal de chaque phrase, etc.

Comme mentionné plus haut, le **module de sélection devra également être adapté**, pour offrir à la synthèse multi-expressive toute la flexibilité qu'elle requiert. L'objet principal des évolutions portera sur la possibilité de spécifier un style à l'échelle du mot, et non simplement à l'échelle du groupe de souffle. Ceci sera notamment utile pour le style Guy_insiste, destiné à la restitution d'emphases sur des mots ou groupes de mots.

D'autres aspects plus exploratoires pourront également être abordés, comme par exemple l'**attribution automatique de styles aux différents segments d'une narration**. Il n'est certes pas envisageable, du moins à moyen terme, d'appréhender les contextes dialogique et sémantique dans leur intégralité. Mais la mise en place de règles simples, basées sur des considérations lexicales ou typographiques (détection de superlatifs, de dialogues, etc.), pourrait offrir un niveau appréciable d'automatisation dans le choix des styles.

Enfin, la conduite d'un **test perceptif** permettra de valider de manière rigoureuse les progrès effectués et de mieux orienter les efforts de recherche.

14 Au-delà du texte

Parmi les traits essentiels d'expressivité attendus par certaines applications, figurent les éléments paralinguistiques [Campbell 07]. Par définition, ces éléments représentent les bruits et mimiques qui accompagnent le flux linguistique d'un locuteur et influencent, volontairement ou non, les informations véhiculées : hésitations, éclaircissements de voix, rires, éternuements, mais aussi expressions faciales, mouvements des mains... Bien entendu nous nous intéressons ici uniquement aux **éléments paralinguistiques vocaux**, c'est-à-dire ceux qui sont produits par l'appareil phonatoire. On peut citer à ce sujet les travaux intéressants d'Esposito, qui démontrent que le canal vocal véhicule plus d'informations émotionnelles que le canal visuel [Esposito 07].

Les éléments paralinguistiques vocaux s'accompagnent généralement de déformations spécifiques sur les phonèmes environnants. Ainsi, un rire survenant au milieu d'une phrase peut être annoncé par une altération du timbre et une élévation de la fréquence fondamentale. Ces déformations tendent ensuite à persister dans le reste de la phrase. La problématique est alors la suivante : comment intégrer « sans couture » un élément paralinguistique vocal dans un flux de parole synthétique ?

Nous proposons dans la suite une technique d'introduction réaliste de ces éléments dans la synthèse, que nous appliquons aux rires et hésitations. Comme pour la multi-expressivité, cette technique suppose l'existence d'une intelligence de haut-niveau pour spécifier la position et la nature des éléments à restituer. Nous ne traitons pas ici ces aspects ; aussi nous contentons-nous de placer manuellement les balises adéquates dans l'entrée textuelle.

La présente section a fait l'objet d'une publication partielle dans [Cadic 08] et [Ségalen 08]. Ces travaux constituent en réalité le point de départ de toute notre étude. Plusieurs idées ont en effet découlé des expériences décrites ci-dessous : protection des phonèmes fragiles par des phonèmes robustes (sandwichs vocaliques), constitution artificielle de scripts denses (algorithme de création de phrases) et choix statistiques motivés par un corpus de référence équilibré. La section peut toutefois être vue comme une extension de nos procédés aux contenus non textuels, d'où sa position finale dans ce document.

14.1 Approche pour l'introduction d'éléments paralinguistiques

Comme illustré sur la figure 49, nous pouvons distinguer plusieurs phases dans un flux de parole contenant un élément paralinguistique. La phase d'**annonce** correspond au tronçon de parole précédant l'élément et présentant des signes précurseurs de celui-ci, comme par exemple une augmentation du volume ou de la fréquence fondamentale. Nous définissons de la même manière la phase de **reprise de la parole**, qui fait parfois l'objet d'une prolongation des modifications vocales. La nature de ces zones transitoires est étroitement liée au type et à l'intensité de l'élément paralinguistique lui-même. Ainsi un rire intense s'accompagnera de modifications plus importantes sur les phonèmes environnants qu'un rire contenu.

Peu de travaux ont traité de l'introduction d'éléments paralinguistiques dans la synthèse vocale. Aussi se contentent-ils tous d'insérer des éléments isolés au sein d'une portion silencieuse du signal de parole, c'est-à-dire entre deux groupes de souffle successifs. Les phases d'annonce et de reprise ne sont donc pas prises en compte. Dans [Trouvain 04] et [Lasarcyk 07] cette technique d'insertion est utilisée pour l'ajout de rires, d'origine humaine ou synthétique. Les auteurs concluent logiquement que seuls des rires très contenus peuvent être restitués de cette manière.

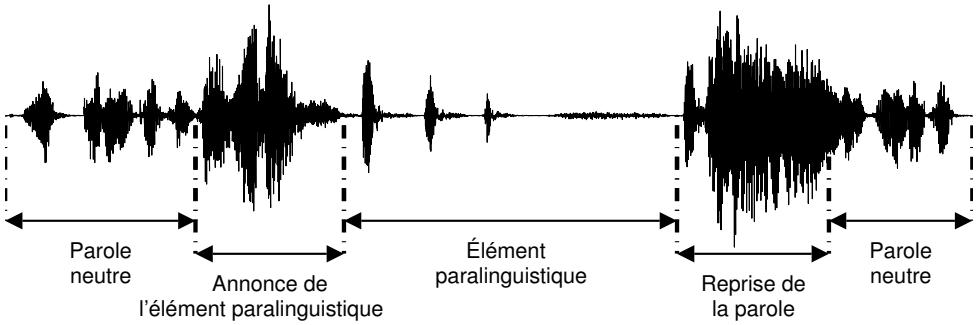


FIGURE 49 – Phases successives d'un flux de parole contenant un élément paralinguistique (ici un rire).

En réalité, le corps d'un élément paralinguistique est difficilement détachable de son annonce ou de la phase de reprise. Partant de ce constat, nous proposons d'adopter les fondements suivants pour la restitution d'éléments paralinguistiques en synthèse vocale :

1. enrichir notre base de synthèse par corpus avec des réalisations variées d'éléments paralinguistiques en contexte ;
2. considérer les séquences « annonce + élément + reprise » comme un continuum acoustique, à préserver autant que possible des fragmentations de la synthèse par corpus.

Il y a une analogie directe avec les sandwichs vocaliques. L'élément paralinguistique constitue un noyau acoustique à protéger, au même titre que les voyelles des sandwichs vocaliques. Les phases d'annonce et de reprise correspondent quant à elles à des suites de phonèmes fragiles, à protéger également des concaténations. On obtient ainsi un **sandwich paralinguistique**, délimité par les phonèmes robustes environnants. Ce sandwich paralinguistique englobe une phase d'annonce, l'élément lui-même et une phase de reprise. La longueur prise en compte pour les phases d'annonce et de reprise dépend donc du contexte phonétique de l'élément paralinguistique. Dans le cas extrême où l'élément paralinguistique est entouré directement de deux consonnes, l'annonce et la reprise sont réduites à un unique phonème. La figure 50 donne un exemple de sandwich paralinguistique englobant un éternuement. Le niveau de gris des phonèmes suit les mêmes conventions graphiques que dans la partie II : un phonème est d'autant plus foncé qu'il tolère *a priori* les concaténations. Dans toute la suite de cette section, nous nous plaçons dans le cas `LR_semirobustes`, c'est-à-dire que les liquides sont considérées comme des phonèmes robustes sauf dans l'environnement direct d'un phonème sourd. Pour la spécification d'un élément paralinguistique dans l'entrée textuelle, nous utilisons un format de balisage inspiré du SSML.

(a) *Je crois que je vais éternuer <paraling type="sneeze"/> aah ça fait du bien.*

(b) # ʒ ø k ʁ w a k ø ʒ ø v ɛ z e t ɛ ʁ n ɥ e ★ a s a f ɛ d y b j ɛ̃ #

(c) **[n ɥ e ★ a s]**

FIGURE 50 – (a) Exemple d'énoncé textuel comportant un élément paralinguistique, symbolisé par ★, avec (b) la transcription phonétique et (c) le sandwich paralinguistique correspondant.

Notre approche repose sur l'acquisition d'un **corpus représentatif de sandwichs paralinguistiques**. Pour cela nous suggérons de procéder à des enregistrements spécifiques, en demandant à un locuteur de simuler les différents éléments. Le coeur de notre travail porte sur la définition d'un script de lecture dédié, qui couvre de façon dense les sandwichs paralinguistiques les plus fréquents. Ceux-ci pourront ainsi, lors de la synthèse, être préservés des concaténations. Comme pour les sandwichs vocaliques, la notion de fréquence est attachée à un corpus textuel de référence ; mais ici, des hypothèses complémentaires doivent être formulées sur les lieux possibles d'apparition des éléments paralinguistiques.

14.2 Constitution d'un script de lecture paralinguistique

Le script de lecture doit être suffisamment riche pour apporter dans la plupart des cas un continuum satisfaisant entre l'élément paralinguistique et la phrase synthétique environnante. Mais il doit à l'inverse rester le plus court possible pour des raisons de coût évidentes.

Afin de rationaliser les besoins de couverture, **nous faisons l'hypothèse que les éléments paralinguistiques ne peuvent pas survenir à l'intérieur d'un mot**. Ainsi la phase d'annonce porte nécessairement sur la fin du mot précédent et la phase de reprise sur le début du mot suivant. Cette restriction nous permet de réduire considérablement la dispersion des contextes possibles pour chaque élément paralinguistique, sans trop nous éloigner de la réalité phonatoire.

Dans notre expérience, nous supposons en outre que les éléments paralinguistiques sont toujours suivis d'une pause et qu'ils n'interagissent pas avec la phrase synthétique suivante. Autrement dit **nous ne traitons pas la phase de reprise** ; seule l'annonce est intégrée au continuum que nous cherchons à préserver des concaténations. Le traitement de la reprise est réservé à des travaux futurs.

Dans ces conditions, la distribution des sandwichs paralinguistiques est largement simplifiée. Leur contenu ne dépend que de la fin du mot précédent, comme illustré à travers quelques exemples dans la table 19.

Mot précédent		Sandwich paralinguistique
cinéma	[sinema]	[ma ★ #]
remercier	[ʁømɛʁsje]	[sje ★ #]
déplacent	[deplas]	[s ★ #]
peuple	[pœplə]	[plə ★ #]
plâtrier	[platʁije]	[tʁije ★ #]

TABLE 19 – Quelques exemples de sandwichs paralinguistiques, suite aux hypothèses simplificatrices. Le symbole ★ désigne tout type d'élément paralinguistique.

Nous avons conduit une étude statistique de ces sandwichs paralinguistiques dans la langue française. En ajoutant l'hypothèse qu'un élément paralinguistique, quel qu'il soit, peut survenir de manière équiprobable à la fin de n'importe quel mot, les contextes phonétiques des sandwichs paralinguistiques peuvent facilement être dénombrés. Le corpus textuel utilisé pour cela est un état intermédiaire du corpus de référence présenté au 7.1. Il comportait à l'époque 130 000 groupes de souffle issus de sous-titres de films, de pièces de théâtre contemporaines, de recettes de cuisine et d'articles du Monde. 453 548 sandwichs paralinguistiques y ont été recensés, dont seulement 3 340 distincts. La figure 51 présente la fonction de répartition de ces sandwichs, classés par ordre décroissant de fréquence d'apparition. On retrouve l'allure logarithmique induite par la loi de Zipf-Mandelbrot, qui régit la plupart des distributions de ce type. La courbe suggère ici que l'**on peut obtenir une couverture de 92% des occurrences avec seulement 200 sandwichs distincts**, ce qui est très satisfaisant pour notre application.

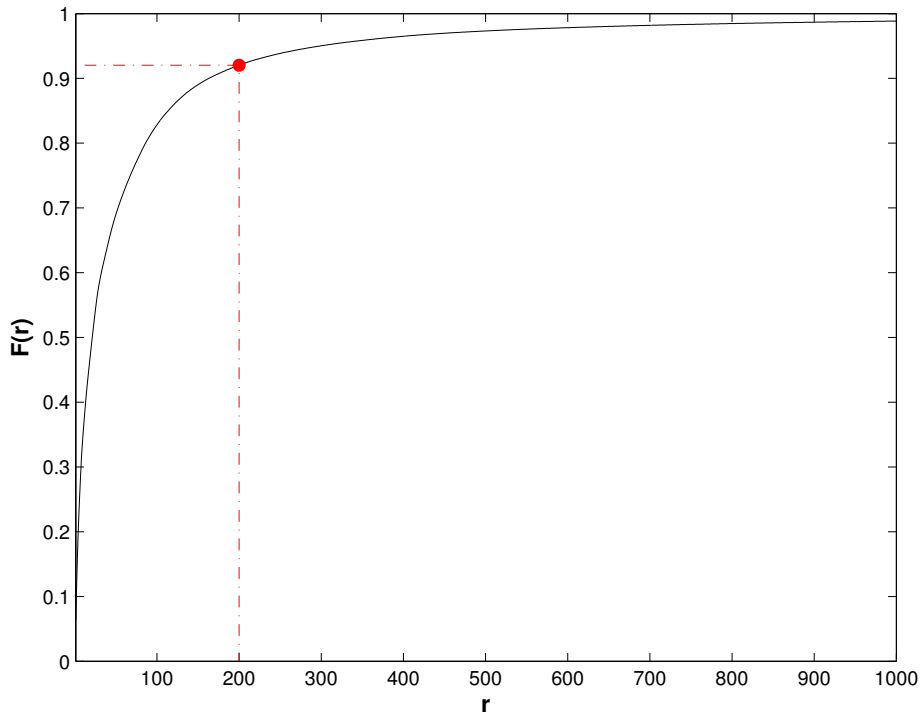


FIGURE 51 – Fonction de répartition des sandwichs paralinguistiques.

Les 200 sandwichs paralinguistiques les plus fréquents ont donné lieu à un script de lecture de 200 prompts. Ces prompts prennent l'aspect de courtes séquences phonétiques. Dans chacun d'entre eux, le sandwich est précédé d'un logatome visant à neutraliser son contexte phonétique. Par exemple le prompt [#tœtʁije★#] sert à couvrir le sandwich [tʁije★#]. Il est censé capturer une réalisation d'élément paralinguistique conjointement à une phase d'annonce portée par les phonèmes [tʁije].

Le script de 200 prompts phonétiques est ici commun à tous les types d'éléments paralinguistiques : rire, hésitation, éternuement, bâillement, toux, hoquet, éclaircissement de voix... Un enregistrement complet du script est à prévoir pour chaque type d'élément que l'on souhaite couvrir.

14.3 Enregistrements

14.3.1 Éléments paralinguistiques traités

Le rire

Le rire est l'un des éléments paralinguistiques les plus importants, du fait qu'on l'associe souvent à un idéal d'expressivité. Nous l'avons donc inclus dans cette étude, bien qu'il constitue un véritable écueil pour la synthèse vocale. Il a en effet de grandes répercussions sur la prosodie (hauteur, énergie...) et le timbre des mots voire des phrases environnantes. En outre il est souvent associé à de la parole souriante, dont il a été montré que les caractéristiques acoustiques sont perceptibles [Émond 06]. La définition du rire est rendue complexe par le fait qu'il ne possède pas de caractère stéréotypé mais plutôt de nombreuses variantes [Trouvain 03].

L'hésitation

Nous avons souhaité traiter l'hésitation car elle est fréquente à l'oral. Bien qu'elle puisse prendre des formes très différentes en fonction du locuteur et du contexte [Boula de Mareüil 05], nous nous sommes restreints au type de réalisation le plus courant en français : le classique *euh*, plus ou moins long et plus ou moins laryngalisé, qui s'intercale généralement entre 2 mots. On parle également de pause remplie (ou filled pause), par opposition aux pauses silencieuses qui ne sont pas des marques d'hésitation [Campione 05]. L'observation des pauses remplies indique qu'elles peuvent aussi bien être précédées d'une courte pause silencieuse que prolonger sans interruption le mot qui les précède. La première manière peut aisément être restituée en synthèse vocale (simple insertion au milieu d'une pause d'un *euh* enregistré isolément) mais est inappropriée lorsqu'une certaine rapidité et fluidité sont requises. Dans cette étude, nous nous sommes donc intéressés à ce deuxième cas, plus délicat à traiter car il nécessite d'assurer un continuum entre la phrase de synthèse et l'élément paralinguistique.

14.3.2 Déroulement des enregistrements

Pour le rire comme pour l'hésitation, nous avons enregistré le script de lecture complet, soit 200 prompts. L'auteur a prêté sa voix pour cet exercice, en utilisant une interface d'acquisition similaire à celle de la page 137. L'une des consignes de lecture a été de **conserver une voix neutre jusqu'à la consonne initiale du sandwich paralinguistique**, ce dernier étant souligné dans l'interface. Cette contrainte est censée garantir la bonne intégration du sandwich dans les futures phrases de synthèse. L'autre consigne a été de **produire des éléments paralinguistiques les plus crédibles possibles**, avec une transition naturelle de la voix neutre vers l'élément. Pour la simulation de cette phase d'annonce les phonèmes de début du sandwich ont pu être exploité. Ceci relève d'un jeu d'acteur pour lequel l'auteur ne montre pas de prédisposition particulière ; aussi la tâche s'est-elle révélée plutôt ardue en ce qui concerne la simulation des 200 rires.

L'enregistrement du script complet a nécessité un peu moins de 30 minutes, aussi bien pour les rires que pour les hésitations. L'investissement semble donc tout à fait raisonnable au regard du temps passé à enregistrer une base complète de synthèse par corpus.

14.4 Intégration dans le moteur de synthèse

Les 400 prompts ainsi collectés ont été ajoutés à l'unique dictionnaire de synthèse dont nous disposions à l'époque pour la voix Didier. Il est basé sur le script de lecture « 3h_nphones », antérieur aux sandwichs vocaliques. Rappelons que ce dictionnaire contient 26 minutes de parole utile et qu'il a obtenu un DMOS de 3.26 dans l'évaluation de la section 12 (voir table 14 page 144). La qualité moyenne de synthèse peut sembler en décalage avec le réalisme des éléments paralinguistiques que nous cherchons à restituer, de la même manière que la synthèse multi-expressive s'est montrée incompatible avec le moindre artefact. Nous expliquerons un peu plus loin comment nous avons compensé cette faiblesse pour l'évaluation de notre procédé d'insertion d'éléments paralinguistiques.

Les rires et hésitations ont été segmentés manuellement et annotés de manière spécifique dans la base. Pour cela nous avons simplement ajouté deux phonèmes respectifs dans l'alphabet phonétique du système. Bien entendu, les éléments paralinguistiques dont il est question ne sont pas vraiment des phonèmes en tant que sons élémentaires de la langue ; néanmoins cette astuce nous a permis de prendre en compte très facilement les nouveaux éléments dans le

moteur de sélection de diphones. Une pénalité élevée de concaténation leur a été attribuée de façon à préserver autant que possible le continuum paralinguistique. Les balises *<paraling type="laugh"/>* et *<paraling type="euh"/>* ont par ailleurs été ajoutées au module de parsing, afin d'insérer automatiquement les phonèmes respectifs dans la séquence phonétique. Nous ne détaillerons pas plus ces aspects techniques.

Lors de la synthèse environ 92% des sandwichs paralinguistiques peuvent être introduits de manière intacte. Ce résultat est rendu possible grâce au script de lecture dédié, mais aussi grâce aux règles générales de sélection et à la pénalité élevée de concaténation. L'élément paralinguistique et son annonce sont dans ce cas introduits par une concaténation sur le milieu de la dernière consonne du mot précédent. Les 8% de sandwichs paralinguistiques restants doivent être reconstitués à partir d'une concaténation à l'intérieur de la phase d'annonce, sur une voyelle ou semi-voyelle précédant l'élément. Ces cas s'accompagnent évidemment d'un risque important de distorsion acoustique.

14.5 Évaluation

Une évaluation subjective de cette technique de restitution d'éléments paralinguistiques a été conduite. Elle a consisté en des essais d'opinion d'écoute de type *Absolute Category Rating* (voir en 4.2.2).

14.5.1 Matériel de test

Pour cette évaluation un ensemble de phrases textuelles, contenant chacune une spécification d'élément paralinguistique, a été vocalisé de trois manières :

- **Référence** : lecture naturelle de la phrase par le locuteur, en simulant au mieux l'élément paralinguistique.
- **Synthèse contextuelle** (notre système) : l'élément paralinguistique et sa phase d'annonce sont sélectionnés dans une base acoustique dédiée, qui couvre les sandwichs paralinguistiques les plus fréquents.
- **Synthèse basique** (état de l'art) : simple insertion, au milieu d'un silence, d'une réalisation autonome de l'élément paralinguistique.

Chaque phrase restituée par notre système est ainsi mise en regard de deux solutions limites : une version naturelle et une version synthétique de base. Cette dernière peut être vue comme la solution de l'état de l'art.

Pour les hésitations comme pour les rires, nous avons rédigé 10 phrases de test spécifiques. Dans chacune d'elles, l'élément paralinguistique est entouré de deux groupes de mots, l'ensemble formant un tout **sémantiquement cohérent**. Pour les versions synthétiques de ces phrases (basique et contextuelle), nous avons souhaité focaliser l'attention des auditeurs autour de l'élément paralinguistique et non sur la synthèse médiocre des phrases environnantes. Afin d'éviter l'apparition d'artefacts qui parasiteraient les zones annexes, nous avons **simulé une synthèse de très haute qualité** en ajoutant simplement une version neutre des 20 phrases, sans élément paralinguistique, au dictionnaire de synthèse. Ainsi le système n'effectue que 2 concaténations, de part et d'autre de l'élément. La première concaténation intervient au début de la phase d'annonce si le sandwich paralinguistique est couvert dans la base, ou à l'intérieur de la phase d'annonce si ce n'est pas le cas. La seconde concaténation est placée au milieu du silence qui suit l'élément paralinguistique.

L'appréciation des rires étant extrêmement subjective, la notation des différentes solutions risquait d'être influencée par la crédibilité de l'élément correspondant. Dans un souci d'équité,

nous avons tâché de réutiliser pour la version synthétique de base les rires obtenus avec la version contextuelle. Cela a été possible chaque fois que le rire pouvait être séparé de son annonce, en exploitant une interruption entre éclats successifs. Lorsqu'une telle extraction était malaisée, ainsi que pour les hésitations, nous avons inséré dans la version basique l'élément paralinguistique qui nous semblait le plus approprié parmi un ensemble de réalisations produites isolément par le locuteur.

Les **60 prompts sonores** (10 phrases suivant 3 méthodes et pour 2 types d'éléments) sont disponibles sur le CD d'accompagnement. Ils ont été écoutés et notés dans un ordre aléatoire par **10 auditeurs naïfs d'origine française**. Afin de décorréler l'impact perceptif des trois phases du flux paralinguistique, nous avons retenu **trois critères de notation** :

- le *naturel de la phase d'annonce* (transition entre la parole et l'élément paralinguistique) ;
- le *naturel de l'élément paralinguistique* lui-même ;
- le *naturel de la reprise de vocalisation* (transition entre l'élément et la parole suivante).

Les auditeurs ont noté chaque prompt suivant chacun des 3 critères, sur une échelle de 1 à 5 conforme aux recommandations de l'UIT [ITU-T 94] : 1 = mauvais, 2 = médiocre, 3 = passable, 4 = bon, 5 = excellent.

14.5.2 Résultats

La figure 52 rapporte les scores perceptifs moyens. Pour chacune des trois méthodes, on distingue le MOS obtenu pour les trois critères de notation, en indiquant les intervalles de confiance à 95%.

Concernant la **phase d'annonce**, les résultats indiquent un **apport très net de notre système par rapport au procédé basique d'insertion d'un élément paralinguistique**. Pour les hésitations, nous obtenons même une note très proche du naturel. Pour les rires les résultats restent inférieurs à la référence, malgré une amélioration notable par rapport à l'état de l'art. Avec une longueur de seulement quelques phonèmes, notre phase d'annonce est vraisemblablement trop courte pour introduire les rires de façon naturelle. Dans 25% des cas, l'annonce est réduite à un seul phonème : la progressivité des rires en est pénalisée, même si la continuité du signal est bel et bien garantie (au moins pour les 92% de sandwichs couverts).

Concernant le **naturel des éléments eux-mêmes**, on observe pour les hésitations un échelonnage similaire à celui de la phase d'annonce. Ceci traduit la difficulté des auditeurs à différencier les critères de notation : l'hésitation et son annonce tendent à former un tout indissociable, dont la perception ne peut être que globale. Il en résulte une corrélation entre les deux premiers critères. La notation des rires est quant à elle plus équilibrée. Du fait probablement de leur caractère intermittent, ils s'accordent davantage avec le modèle en 3 phases et les auditeurs parviennent mieux à séparer les trois notes.

Enfin la **reprise de vocalisation** est très mal notée avec les deux types de synthèse, que ce soit pour les rires ou pour les hésitations. Ce résultat est logique, dans la mesure où nous n'assurons aucune continuité entre l'élément paralinguistique et la parole qui suit. Une simple pause matérialise la transition, comme dans le procédé basique.

14.6 Perspectives

En conclusion, ces résultats démontrent la pertinence de notre méthode pour la restitution de transitions naturelles d'une phrase synthétique vers un élément paralinguistique. Mais ils invitent également à **améliorer le traitement de la reprise** de parole. Suite aux remarques

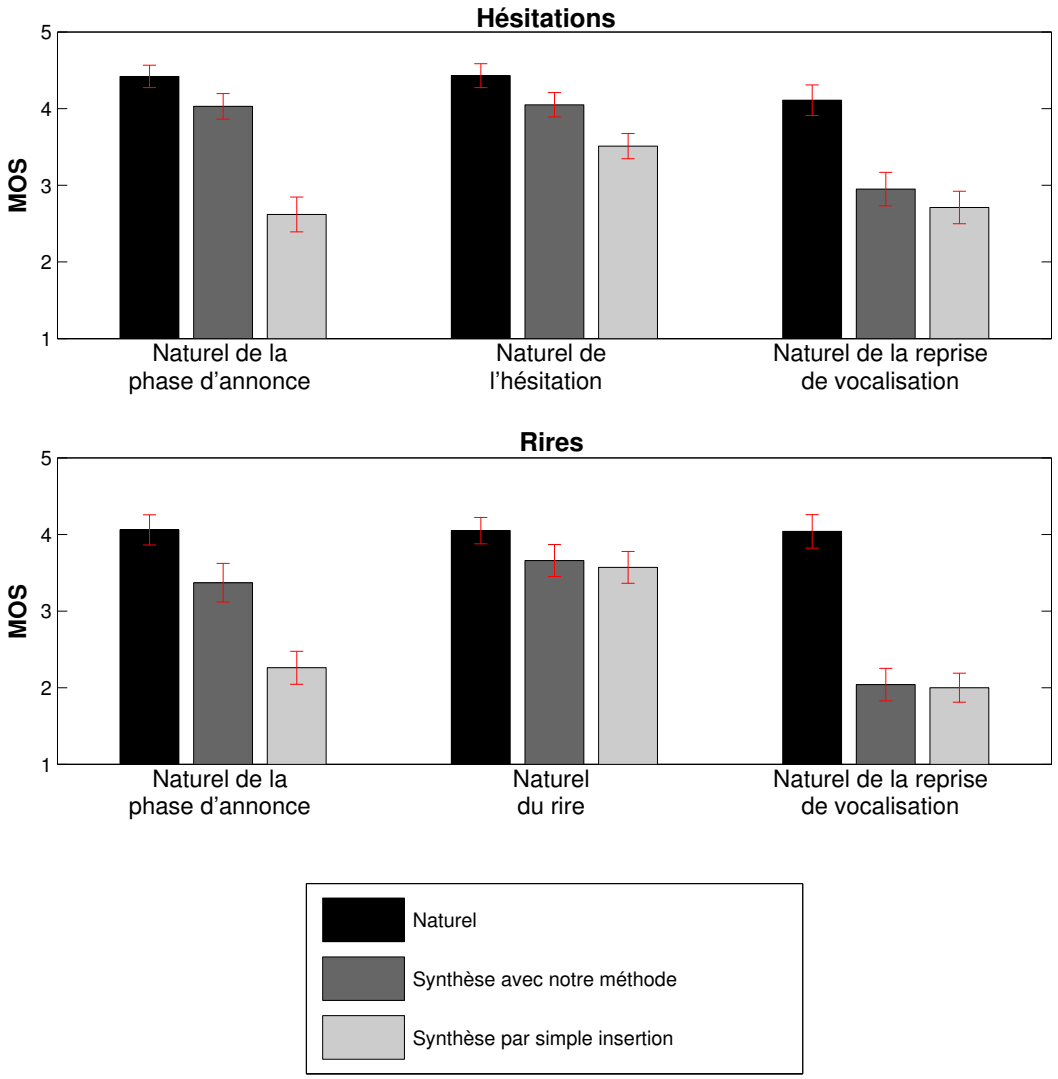


FIGURE 52 – Scores moyens relevés, dans le cas des hésitations et dans le cas des rires.

des différents sujets, il semblerait que l'insertion d'une inspiration juste avant la phrase de synthèse suivant l'élément puisse à elle seule améliorer sensiblement l'appréciation. Mais on peut également construire une approche similaire à la phase d'annonce. Pour éviter le problème combinatoire posé par la prise en compte de sandwichs paralinguistiques complets (c'est-à-dire intégrant à la fois l'annonce et la reprise), il convient cependant de disjoindre le traitement des deux phases transitoires. Dans ce but on peut envisager, pour chaque type d'élément paralinguistique, d'imposer un état central commun à toutes les réalisations. Dans le cas des rires, il pourrait s'agir d'une interruption prédéfinie, associée à un état émotif constant. Dans le cas des pauses remplies (hésitations), cet état central pourrait être défini par une hauteur et une cible phonétique précises. Tous les sandwichs paralinguistiques pourraient alors être recomposés par simple concaténation de deux moitiés de sandwichs, chacune des moitiés étant couverte de manière optimale dans un script de lecture dédié.

Outre les travaux sur la phase de reprise, il serait intéressant de chercher à accroître le contrôle sur les caractéristiques de l'élément paralinguistique, notamment sur son intensité. La réalisation de certains éléments nécessitant de réels talents d'acteurs, il semble important de recourir à un locuteur professionnel pour la suite de ces expériences.

Conclusion

Nos travaux ont porté sur la création de voix dans le cadre de la synthèse par corpus. Nous avons tout d'abord décrit les tenants et aboutissants des méthodes traditionnelles, qui nécessitent l'enregistrement d'un locuteur en studio pendant plusieurs jours. Elles reposent sur un script de lecture de plusieurs milliers de phrases, obtenu par condensation d'un vaste corpus textuel. Le critère de condensation porte généralement sur la maximisation de la couverture en diphones ou triphones. Les enregistrements effectués suivant ce type de script sont souvent complétés d'une coûteuse phase de traitement manuel, consistant à vérifier intégralement la segmentation phonétique. La lourdeur générale de cet état de l'art freine la diversification du catalogue de voix et contribue à limiter les restitutions à un style de « lecture neutre ».

Contributions principales

Partant de ce constat, nous avons proposé de nouveaux critères pour l'optimisation du script de lecture. Ces critères reposent sur la couverture d'unités nouvelles, les « sandwichs vocaliques », qui intègrent mieux les spécificités de la synthèse par corpus que les unités traditionnelles. La nouveauté réside à la fois dans leur définition phonétique et dans leur enrichissement linguistique.

Sur le plan phonétique, les sandwichs vocaliques permettent une prise en compte inédite des limites segmentales de la synthèse par corpus. Pour cela les « phonèmes robustes », qui tolèrent généralement bien les concaténations de la synthèse par corpus, sont distingués des « phonèmes fragiles », dont il faut préserver l'intégrité autant que possible. Le premier groupe comporte typiquement les consonnes non liquides et le silence, tandis que le second rassemble les voyelles, semi-voyelles, schwa et liquides. Les sandwichs vocaliques désignent alors toute séquence phonétique qui s'étend d'un phonème robuste au suivant, protégeant en son sein une séquence plus ou moins longue de phonèmes fragiles (dont au moins une voyelle). Les sandwichs sont donc des unités de longueur variable, pouvant comporter plusieurs voyelles et franchir des frontières de mots.

Sur le plan linguistique, les sandwichs sont enrichis de traits contextuels largement simplifiés. Nous avons en effet démontré qu'une dizaine de symboles linguistiques suffisent à expliquer les principaux mouvements prosodiques de la lecture neutre. Un test perceptif nous a en outre permis de restreindre l'enrichissement contextuel aux seules voyelles, en tant que porteuses principales de la prosodie.

La définition des sandwichs vocaliques ne fige ni l'ensemble des phonèmes robustes ni le jeu de symboles contextuels. Aussi en avons-nous proposé plusieurs variantes, ce qui permet de moduler leur complexité en fonction de contraintes variées (attentes qualitatives, taille visée pour le script de lecture...).

Pour le suivi de la couverture en sandwichs vocaliques, nous avons justifié une approche en fréquence. Celle-ci consiste à pondérer les unités cibles en fonction de leur fréquence d'apparition. La constitution d'un corpus textuel de référence est alors requise, ce qui a fait l'objet d'une attention particulière dans ce travail. Avec un niveau élevé de vérification, le corpus de référence que nous avons constitué couvre de manière significative et équilibrée la plupart des applications envisagées pour nos voix de synthèse. Une place volontairement importante a été accordée au langage parlé ; elle reflète notre volonté de nous écarter des usages traditionnels de la synthèse vocale pour aborder de nouveaux champs applicatifs, comme par exemple l'échange de messages ludiques entre internautes.

Plusieurs mesures objectives ont confirmé la pertinence des sandwichs vocaliques pour l'étude qualitative des scripts de lecture. D'une manière générale, l'optimisation de leur couverture dans un script est un gage de qualité pour les futures voix de synthèse, car elle garantit une bonne prise en compte des principaux facteurs segmentaux et prosodiques qui influencent la qualité de la synthèse par corpus.

Étant donné ces critères d'optimisation du script de lecture, basés sur une couverture fréquentielle des différentes variantes de sandwichs vocaliques, nous nous sommes intéressés à l'algorithme d'optimisation lui-même. Des considérations quantitatives et qualitatives ont d'abord permis de tracer les contours généraux de cet algorithme. Nous avons entre autres retenu une stratégie gloutonne de type « fréquents d'abord », consistant à ajouter une à une les phrases qui apportent le plus d'unités fréquentes non couvertes. Nous avons aussi proposé une gestion inédite des longueurs de phrases, reposant sur un seuil maximal de longueur et sur l'utilisation du groupe de souffle, plutôt que la phrase, comme élément constitutif des scripts de lecture. Ces choix offrent un bon compromis entre la densité des scripts et leur lisibilité.

L'origine des phrases (en réalité des groupes de souffle) ajoutées à chaque itération du glouton a fait l'objet d'une étude approfondie. Nous avons expérimenté le procédé traditionnel de condensation de corpus, suivant lequel les phrases sont simplement sélectionnées au sein du corpus de référence. La densité des scripts ainsi constitués est deux à trois fois supérieure à celle que l'on obtient en sélectionnant les phrases aléatoirement. Elle est néanmoins deux fois moindre que l'optimum, qui fait référence à un script virtuel apportant uniquement les unités les plus fréquentes et sans redondance. La condensation de corpus laisse donc une marge de manoeuvre importante pour la densification des scripts. Nous l'avons utilisée pour la constitution de plusieurs scripts de lecture, dont « 5h_sandwiches » ; comme son nom l'indique, il est optimisé sur des critères à base de sandwichs vocaliques et requiert en moyenne cinq heures d'enregistrement (pour une durée de parole utile d'environ 75 minutes).

Nous avons également proposé un algorithme d'optimisation inédit, concurrent de la condensation. Il consiste à élargir les possibilités du glouton en mettant à sa disposition des phrases créées de toutes pièces. Pour cela nous avons conçu un ensemble d'automates transducteurs pondérés, qui dépendent des statistiques du corpus de référence ainsi que de l'état courant du script cible. À chaque étape du glouton, une simple recherche de meilleur chemin dans ces automates nous permet de composer, en une fraction de seconde, un enchaînement de sandwichs qui maximise l'augmentation du taux de couverture. La cohérence globale de cet enchaînement n'est pas garantie ; aussi devons-nous recourir à une intervention humaine pour améliorer sa lisibilité. Un outil semi-automatique de création de phrases a été développé dans ce but. Il offre à un opérateur humain la possibilité de contraindre la génération des séquences automatiques, de façon à en assurer la lisibilité sans trop nuire à leur optimalité. Nous avons estimé, de manière fiable, que cet algorithme permet d'accroître la densité des scripts de 30 à 40% par rapport au procédé de condensation. Il est toutefois très coûteux en temps humain, puisqu'en moyenne trois minutes de travail sont nécessaires à un opérateur pour créer chaque phrase.

Nous avons utilisé la construction de phrases pour la création de 4 scripts de lecture. Malheureusement la variété des conditions expérimentales dans lesquelles ces scripts ont été réalisés empêche de vérifier les estimations de performances précédentes. Nous avons tout de même pu tirer de cette mise en application quelques observations qualitatives, la principale étant le manque de cohérence sémantique des phrases construites. C'est là une conséquence logique de notre volonté de contrer la distribution naturelle des sandwichs : le gain de densité se fait au détriment de la sémantique.

Les différents scripts de lecture ont été enregistrés par de nombreux locuteurs d'âge, de sexe et de niveau de professionnalisme variés. Nous avons à cette occasion obtenu la confirmation

que la restriction à des groupes de souffle de longueur contrôlée accroît l'efficacité des enregistrements. Les scripts obtenus par construction de phrase sont en revanche responsables, du fait de leur manque de cohérence sémantique, d'un ralentissement global du rythme de lecture chez certains locuteurs. Pour que les bénéfices de la densification ne soient pas contrebalancés par un tel ralentissement, ces scripts doivent être réservés aux locuteurs expérimentés.

La diversité des bases constituées nous a ensuite permis de conduire une évaluation perceptive à grande échelle. 49 dictionnaires de synthèse ont ainsi été soumis à un large panel d'auditeurs, afin d'évaluer l'impact sur la qualité de synthèse de deux facteurs : le script de lecture d'une part, le niveau de révision de la segmentation phonétique d'autre part. Sur le premier aspect les résultats sont sans appel. À taille comparable, les scripts optimisés sur des critères à base de sandwichs offrent une meilleure qualité de synthèse que les scripts de l'état de l'art. Ce constat est valable aussi bien pour les scripts obtenus par condensation que pour ceux obtenus par construction de phrases. Des scores perceptifs très proches du naturel peuvent être atteints dès 40 minutes de parole, soit une demi-journée d'enregistrement. Cela correspond approximativement à une réduction d'un facteur 10 de la durée des enregistrements nécessaires à la création d'une voix de synthèse de haute qualité. Le gain s'explique à la fois par l'optimisation de la couverture en sandwichs vocaliques et par l'efficacité accrue des sessions d'enregistrement. Concernant la révision manuelle des données segmentales, notre évaluation a permis de conclure qu'elle ne joue pas un rôle déterminant dans la création de voix de synthèse de haute qualité, ce qui est également une conclusion rassurante.

Il devient donc possible de créer de très bonnes voix de synthèse avec moins d'une journée d'enregistrement et en utilisant uniquement des post-traitements automatiques. Nous avons mis à profit ce résultat pour créer près de vingt dictionnaires de synthèse d'un même locuteur, chaque dictionnaire étant dédié à un unique style vocal. L'objectif de cette « multi-expressivité » est de pouvoir adapter le patron prosodique ou la qualité vocale de la voix de synthèse en fonction des contextes dialogique et sémantique, simplement en passant d'un dictionnaire de synthèse à l'autre. Cette expérience a donné des résultats très encourageants. Néanmoins une observation de la dispersion acoustique des différents styles a mis en évidence des failles importantes dans notre pavage de l'espace multi-expressif.

Enfin, nous avons proposé une extension de nos procédés aux contenus non textuels que sont les éléments paralinguistiques vocaux (rires, hésitations, toux, *etc.*). La notion de sandwich paralinguistique a été introduite pour désigner le continuum acoustique qui entoure l'élément paralinguistique et participe pleinement à son réalisme. À l'instar des sandwichs vocaliques, ce continuum doit être préservé autant que possible des concaténations de la synthèse par corpus. Nous avons donc construit un script de lecture dédié, couvrant de manière dense les sandwichs paralinguistiques les plus fréquents. De l'enregistrement de ce script résulte une petite base acoustique qui permet de compléter un dictionnaire de synthèse neutre. Il est alors possible, grâce à ce complément, d'insérer sans couture des éléments paralinguistiques dans la synthèse vocale. Le procédé a été appliqué aux rires et hésitations, puis soumis à une expérience perceptive dont les résultats sont très favorables.

Perspectives

De manière générale, les travaux présentés dans cette thèse ont permis une réelle avancée en matière de création de voix de synthèse. Le design soigné des scripts de lecture et la densification de leur couverture en sandwichs vocaliques se sont révélés être des facteurs décisifs dans l'accélération du processus de création de voix.

Sur ces aspects, des efforts de recherche complémentaires sont envisageables. On peut par

exemple chercher à accroître la cohérence sémantique des scripts issus de notre algorithme de construction de phrases, pour pouvoir en généraliser l'utilisation à tous les locuteurs, quel que soit leur niveau de professionnalisme. Il s'agirait en d'autres termes de mieux contrôler le compromis entre la densité et la lisibilité des phrases générées. Pour cela des modèles génératifs plus contraints pourraient être imaginés, en tenant compte par exemple de critères lexicaux ou de règles grammaticales simples. On allégerait par la même occasion le coût humain de création des phrases.

Il serait par ailleurs intéressant de supplanter les taux de couverture par un unique critère d'optimisation global, fondé sur un « coût de synthèse symbolique ». Comme expliqué en 8.2 cette approche repose sur une approximation symbolique du coût de sélection, qui permet de simuler de manière assez précise l'impact d'un script de lecture sur la synthèse finale. Il en découle un critère très pertinent pour l'optimisation des scripts de lecture ; malheureusement nous ne sommes pas encore parvenus à en surmonter la complexité calculatoire. L'approche est néanmoins prometteuse et un travail de recherche algorithmique pourrait, à moyen terme, la rendre au moins partiellement exploitable.

Les travaux sur l'optimisation des scripts ont donc encore de belles perspectives. Mais les scripts issus de nos travaux, en particulier « 5h_sandwichs », offrent déjà de très bons résultats qu'il convient de valoriser. Dans les sections précédentes nous avons ouvert plusieurs pistes à ce sujet. La multi-expressivité est probablement l'axe de valorisation le plus important ; elle constitue selon nous la prochaine rupture technologique en synthèse vocale. À la lumière de nos expériences dans ce domaine, les travaux futurs devront porter sur une amélioration du pavage de l'espace multi-expressif. Des techniques plus contraignantes d'acquisition de parole devront être mises au point afin de couvrir des styles vocaux plus précis et en meilleure adéquation avec notre découpage contextuel. Il en résultera un contrôle accru sur la prosodie des restitutions synthétiques, ce qui facilitera l'attribution manuelle des styles aux différentes phases de la narration. Sur ce dernier aspect, des procédés d'automatisation sont également à explorer. En effet des heuristiques de sélection des styles doivent pouvoir être composées à partir d'éléments lexicaux ou typographiques simples de l'entrée textuelle. De manière générale, ces différents axes de progression devraient permettre de limiter le recours à la « synthèse assistée par opérateur » dans le cadre de la multi-expressivité.

Pour finir, l'internationalisation de nos procédés est incontournable et constitue sans aucun doute la prochaine étape de nos travaux. Pour chaque nouvelle langue, les principales tâches identifiées sont la constitution d'un corpus de référence, la simplification des symboles contextuels, la définition des sandwichs vocaliques et la transposition de nos outils de constitution de scripts. Pas ou peu d'évolutions du système de synthèse sont à prévoir : pour les langues déjà couvertes, nos outils de création de voix peuvent reposer sur les hauts-niveaux et les règles de sélection existants. L'incertitude majeure réside dans la dispersion des sandwichs vocaliques. On peut en effet craindre que certaines langues ne soient plus « voisées » que le français, avec un impact combinatoire sur la distribution des sandwichs vocaliques, ce qui rendrait les objectifs de couverture difficiles à atteindre. Les langues européennes majeures ne devraient cependant pas poser ce type de problème. Des mesures préliminaires indiquent en effet que les fréquences d'apparition des phonèmes robustes en anglais (52%) et en espagnol (48%) sont supérieures à celle du français (42%).

Annexe

Alphabet phonétique

	symbole phonétique	exemple
consonnes	[p]	**p**ère
	[t]	**t**erre
	[k]	**c**ou
	[b]	**b**on
	[d]	**d**ans
	[g]	**g**are
	[z]	**z**éro
	[ʒ]	**j**e
	[v]	**v**ous
	[s]	**s**ale
	[ʃ]	**ch**at
	[f]	**f**eu
	[m]	**m**ain
	[n]	**n**ous
	[l]	**l**ent
	[ʁ]	**r**ue
semi-voyelles	[j]	**y**eux
	[w]	**ou**i
	[ɥ]	l**ui**
voyelles	[a]	pl**a**t
	[ɔ]	m**o**rt
	[o]	m**o**t
	[ɛ]	l**ai**t
	[e]	bl**é**
	[ø]	p**eu**
	[œ]	p**eu**r
	[ɑ̃]	s**an**s
	[ɔ̃]	b**on**
	[ɛ̃]	pl**ein**
	[œ̃]	br**un**
	[i]	**i**l
	[y]	r**u**e
	[u]	gen**ou**
	[ə]	montagn**e**

TABLE 20 – Liste des unités phonétiques utilisées dans le système de synthèse vocale des Orange Labs, suivant les notations de l'Alphabet Phonétique International (API).

La table 20 présente la liste des unités phonétiques (ou allophones) utilisées dans notre système de synthèse. On notera l'absence de plusieurs unités de l'alphabet phonétique français :

- Le [ɑ] de "p**â**tes", dont la distinction avec le [a] de "pl**a**t" tend à disparaître.
- Le [ɲ] de "a**gn**eau", que nous assimilons à la séquence [nj].

- Les variantes [ʀ] et [r], utilisées dans certains contextes et accents régionaux, que nous assimilons à la liquide principale [ʁ].

Certaines unités phonétiques d'origine étrangère ont également été écartées :
- Le "h aspiré" [h] de "**h**op"
- Le [ŋ], utilisé essentiellement dans des anglicismes comme "campi**ng**" et qui peut être remplacé de manière acceptable par la séquence [ng]"
- La consonne [x] de "mo**j**ito" (origine hispanique) ou "**kh**amsin" (origine arabe)

Contrairement à ces unités phonétiques les **phonèmes**, qui relèvent de la phonologie et non de la phonétique, désignent les différents sons d'une langue en les distinguant par leurs traits pertinents **sans entrer dans le détail de leur prononciation**. Ainsi les unités phonétiques [ʀ], [r] et [ʁ] correspondent à plusieurs modes de production d'un même phonème que nous notons /ʁ/ (entre barres obliques). Notre alphabet **phonémique** est ainsi calqué sur l'alphabet **phonétique** listé en table 20. La seule nuance porte sur le [ə]. Dans notre système il désigne la voyelle introduite par épenthèse, c'est-à-dire pour faciliter la prononciation des consonnes environnantes, ce qui correspond généralement à la production d'un schwa (voyelle neutre centrale) très court. Dans les cas non épenthétiques, c'est-à-dire lorsqu'il a une place entière dans la suite phonétique, nous l'approchons par un [ø] ou un [œ]. Par conséquent, notre utilisation du [ə] relève de considérations allophoniques et ce n'est donc pas un phonème à part entière.

Références

[Allauzen 07] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut & M. Mohri. *OpenFst : A general and efficient weighted finite-state transducer library.* In 12th international conference on Implementation and application of automata, pages 11–23, 2007. (Cité pages 113 and 124)

[Allen 87] J. Allen, M.S. Hunnicutt, D.H. Klatt, R.C. Armstrong & D.B. Pisoni. *From text to speech : the MITalk system.* Cambridge Studies In Speech Science And Communication, page 216, 1987. (Cité page 29)

[Alter 97] Kai Alter, Hannes Pirker & Wolfgang Finkler, editeurs. Concept to speech generation systems, 1997. Workshop in conjunction with 35th Annual Meeting of the Association for Computational Linguistics. (Cité page 21)

[Anis 02] J. Anis. *Communication électronique scripturale et formes langagières : chat et SMS.* In 4èmes rencontres reseaux humains/Reseaux technologiques, 2002. http://rhrt.edel.univ-poitiers.fr/document.php?id=547. (Cité page 22)

[Apostel 57] L. Apostel, B. Mandelbrot & A. Morf. Logique, langage et théorie de l'information. Presses Universitaires de France, 1957. (Cité page 46)

[Bagshaw 98] P.C. Bagshaw. *Phonemic transcription by analogy in text-to-speech synthesis : Novel word pronunciation and lexicon compression.* Computer Speech & Language, vol. 12, no. 2, pages 119–142, 1998. (Cité page 27)

[Bartkova 87] K. Bartkova & C. Sorin. *A model of segmental duration for speech synthesis in French.* Speech Communication, vol. 6, no. 3, pages 245–260, 1987. (Cité page 27)

[Béchet 00] F. Béchet & F. Yvon. *Les noms propres en traitement automatique de la parole.* Traitement automatique des langues, vol. 41, no. 3, pages 671–707, 2000. (Cité page 26)

[Bellegarda 04] J.R. Bellegarda. *A novel discontinuity metric for unit selection text-to-speech synthesis.* In 5th ISCA Workshop on Speech Synthesis (SSW), 2004. (Cité page 43)

[Beller 07] G. Beller. *Context dependent transformation of expressivity in speech using a bayesian network.* Paraling'07, 2007. (Cité pages 16 and 150)

[Beller 08] G. Beller, N. Obin & X. Rodet. *Articulation degree as a prosodic dimension of expressive speech.* Speech Prosody, pages 681–684, 2008. (Cité page 24)

[Benoît 96] C. Benoît, M. Grice & V. Hazan. *The SUS test : A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences.* Speech Communication, vol. 18, no. 4, pages 381–392, 1996. (Cité page 66)

[Beutnagel 98] M. Beutnagel, A. Conkie & A.K. Syrdal. *Diphone synthesis using unit selection.* In Proc. 3th ESCA/COCOSDA workshop on speech synthesis, Jenolan Caves, 1998. (Cité pages 45 and 49)

[Black 94] A.W. Black & P. Taylor. *CHATR : a generic speech synthesis system.* In 15th International conference on Computational linguistics, volume 2, page 986. Association for Computational Linguistics, 1994. (Cité page 33)

[Black 95] A.W. Black & N. Campbell. *Optimising selection of units from speech databases for concatenative synthesis.* In Proc. of Eurospeech, pages 581–584. International Speech Communication Association, 1995. (Cité page 33)

[Black 97] A.W. Black & P. Taylor. *Automatically clustering similar units for unit selection in speech synthesis.* In 5th European Conference on Speech Communication and Technology, 1997. (Cité pages 36, 41, and 75)

[Black 00] A.W. Black & K.A. Lenzo. *Limited Domain Synthesis.* In 6th International Conference on Spoken Language Processing. ISCA, 2000. (Cité page 54)

[Black 01] A.W. Black & K.A. Lenzo. *Optimal data selection for unit selection synthesis.* In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001. (Cité pages 48, 49, and 96)

[Black 03] A.W. Black. *Unit selection and emotional speech.* In 8th European Conference on Speech Communication and Technology, 2003. (Cité pages 16 and 151)

[Black 05] A.W. Black & K. Tokuda. *The blizzard challenge - 2005 : evaluating corpus-based speech synthesis on common datasets.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 69)

[Boeffard 97] O. Boeffard & F. Emerard. *Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm.* In 5th European Conference on Speech Communication and Technology. ISCA, 1997. (Cité page 51)

[Boëffard 02] O. Boëffard & C. d'Alessandro. Traitement Automatique du Langage Parlé, directed by Joseph Mariani, volume 1, chapitre Synthèse de la parole, pages 115–154. Hermès, 2002. (Cité page 65)

[Boidin 08] C. Boidin & O. Boeffard. *Modeling Intonation Variability with HMM for Speech Synthesis.* In Speech Prosody, 2008. (Cité page 27)

[Boidin 09a] B. Boidin. *Modélisation statistique de l'intonation de la parole expressive.* PhD thesis, Université de Rennes 1, 2009. (Cité page 75)

[Boidin 09b] C. Boidin, V. Rieser, L. van der Plas, O. Lemon & J. Chevelu. *Predicting how it sounds : Re-ranking dialogue prompts based on TTS quality for adaptive Spoken Dialogue Systems.* Interspeech Special Session : Machine Learning for Adaptivity in Spoken Dialogue, pages 2487–2490, 2009. (Cité page 147)

[Boll 79] S. Boll. *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pages 113–120, 1979. (Cité pages 56 and 140)

[Boula de Mareüil 05] P. Boula de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda & P. Paroubek. *A quantitative study of disfluencies in French broadcast interviews.* In Disfluency In Spontaneous Speech (DiSS) Workshop, pages 27–32, 2005. (Cité page 161)

[Bove 05] R. Bove. *Etude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS.* RECITAL, pages 625–634, 2005. (Cité page 22)

[Bozkurt 03] B. Bozkurt, O. Ozturk & T. Dutoit. *Text design for TTS speech corpus building using a modified greedy selection.* In 8th European Conference

on Speech Communication and Technology, 2003. (Cité pages 48, 49, 50, and 96)

[Breen 98] A.P. Breen & P. Jackson. *Non-uniform unit selection and the similarity metric within BT's Laureate TTS system.* In The 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, 1998. (Cité page 41)

[Breuer 04] S. Breuer & J. Abresch. *Phoxsy : Multi-phone segments for unit selection speech synthesis.* In 8th International Conference on Spoken Language Processing. ISCA, 2004. (Cité page 81)

[Bulyko 02] I. Bulyko. *Flexible speech synthesis using Weighted Finite State Transducers.* PhD thesis, University of Washington, 2002. (Cité pages 41, 43, and 140)

[Burnett 04] Daniel C. Burnett, Mark R. Walker & Andrew Hunt. *Speech Synthesis Markup Language (SSML) Version 1.0.* Rapport technique, W3C - World Wide Web Consortium, 2004. http ://www.w3.org/TR/speech-synthesis/. (Cité page 21)

[Cabral 08] J. Cabral, S. Renals, K. Richmond & J. Yamagishi. *Glottal spectral separation for parametric speech synthesis.* In 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 1829–1832, 2008. (Cité page 31)

[Cadic 06] D. Cadic, A. Le Forestier, E. Gougis, T. Moudenc, A. Furby & O. Boeffard. *Etude préliminaire d'une nouvelle synthèse vocale pour les patients atteints de sclérose latérale amyotrophique.* In Journées de Neurologie de Langue Française, avril 2006. (Cité page 110)

[Cadic 08] D. Cadic & L. Segalen. *Paralinguistic elements in speech synthesis.* In 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2008. (Cité page 157)

[Cadic 09] D. Cadic, C. Boidin & C. d'Alessandro. *Vocalic sandwich, a unit designed for unit selection TTS.* In 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 2079–2082, 2009. (Cité pages 17, 80, and 147)

[Cadic 10a] D. Cadic, C. Boidin & C. d'Alessandro. *Towards Optimal TTS Corpora.* In International conference on Language Resources and Evaluation (LREC), 2010. (Cité pages 17, 100, and 113)

[Cadic 10b] D. Cadic & C. d'Alessandro. *High quality TTS voices within one day.* In 7th ISCA Speech Synthesis Workshop (SSW7), 2010. (Cité pages 17 and 136)

[Campbell 03] N. Campbell & P. Mokhtari. *Voice quality : the 4th prosodic dimension.* In 15th. International Congress of Phonetic Sciences, pages 2417–2420, 2003. (Cité page 24)

[Campbell 07] N. Campbell. *On the use of nonverbal speech sounds in human communication.* Verbal and Nonverbal Communication Behaviours, pages 117–128, 2007. (Cité pages 24 and 157)

[Campione 05] E. Campione & J. Véronis. *Pauses and hesitations in French spontaneous speech.* In Disfluency in Spontaneous Speech. ISCA, 2005. (Cité page 161)

[Caprara 00] A. Caprara, P. Toth & M. Fischetti. *Algorithms for the set covering problem.* Annals of Operations Research, vol. 98, no. 1, pages 353–371, 2000. (Cité page 52)

[Carlson 82] R. Carlson, B. Granstrom & S. Hunnicutt. *A multi-language text-to-speech module.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 1604–1607, 1982. (Cité page 29)

[Cernak 05] M. Cernak & M. Rusko. *An evaluation of synthetic speech using the PESQ measure.* In European Congress on Acoustics, pages 2725–2728, 2005. (Cité page 64)

[Charpentier 88] F. Charpentier & F. Emerard. *Corpus de 89 phrases couvrant tous les diphones du français.* Rapport technique, Centre National d'Etudes des Télécommunications, Division Traitement du Signal de Parole et Services, 1988. Référence 502/LAA/TSS/RCP. (Cité page 45)

[Chen 99] J.D. Chen & N. Campbell. *Objective distance measures for assessing concatenative speech synthesis.* In 6th European Conference on Speech Communication and Technology, 1999. (Cité page 64)

[Chevelu 07] J. Chevelu, N. Barbot, O. Boeffard & A. Delhay. *Lagrangian relaxation for optimal corpus design.* In 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6), pages 211–216, 2007. (Cité pages 52 and 103)

[Chu 01a] M. Chu & H. Peng. *An objective measure for estimating MOS of synthesized speech.* In 7th European Conference on Speech Communication and Technology, 2001. (Cité pages 41, 64, and 146)

[Chu 01b] M. Chu, H. Peng, H. Yang & E. Chang. *Selecting non-uniform units from a very large corpus for concatenative speech synthesizer.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 785–788, 2001. (Cité pages 48, 49, and 50)

[Chu 02] M. Chu, C. Li, H. Peng & E. Chang. *Domain adaptation for TTS systems.* In International Conference on Acoustics, Speech and Signal Processing, volume 1. IEEE, 2002. (Cité page 54)

[Clark 07] R.A.J. Clark, M. Podsiadlo, M. Fraser, C. Mayo & S. King. *Statistical analysis of the Blizzard Challenge 2007 listening test results.* Blizzard Challenge 3, in Proceedings of SSW6, 2007. (Cité pages 69 and 70)

[Conkie 97] A. Conkie & S. Isard. *Optimal coupling of diphones.* Progress in speech synthesis, pages 293–304, 1997. (Cité page 59)

[Conkie 99] A. Conkie. *A robust unit selection system for speech synthesis.* In 137th meeting of the Acoustical Society of America, page 978, 1999. (Cité page 35)

[Crowe 87] A. Crowe & MA Jack. *Globally optimising formant tracker using generalised centroids.* Electronics Letters, vol. 23, no. 19, pages 1019–1020, 1987. (Cité page 42)

[D'Alessandro 04] C. D'Alessandro. L'évaluation des systèmes de traitement de l'information, chapitre L'évaluation des systèmes de synthèse de la parole, pages 215–239. Hermès, Lavoisier, Paris, 2004. (Cité page 65)

[Davis 90] S.B. Davis & P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* Readings in speech recognition, pages 65–74, 1990. (Cité page 31)

[de Tournemire 04] S. de Tournemire. *Document de travail VMI-71, Description des marqueurs utilisés dans CNETVOX et observation de leur réalisation sur le corpus.* Rapport technique, France Télécom R&D, 2004. (Cité page 74)

[Dedina 91] M.J. Dedina & H.C. Nusbaum. *PRONOUNCE : a program for pronunciation by analogy.* Computer Speech & Language, vol. 5, no. 1, pages 55–64, 1991. (Cité page 27)

[Dixon 68] N. Dixon & H. Maxey. *Terminal analog synthesis of continuous speech using the diphone method of segment assembly.* IEEE Transactions on Audio and Electroacoustics, vol. 16, no. 1, pages 40–50, 1968. (Cité page 29)

[Donovan 96] R. Donovan. *Trainable speech synthesis.* PhD thesis, Cambridge University Engineering Department, 1996. (Cité page 27)

[Donovan 98] R.E. Donovan & E.M. Eide. *The IBM trainable speech synthesis system.* In 5th International Conference on Spoken Language Processing, 1998. (Cité pages 35, 41, and 75)

[Donovan 01] R.E. Donovan. *A new distance measure for costing spectral discontinuities in concatenative speech synthesizers.* In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001. (Cité page 43)

[Dutoit 94] T. Dutoit & H. Leich. *On the ability of various speech models to smooth segment discontinuities in the context of text-to-speech synthesis by concatenation.* In EUSIPCO, volume 1, pages 8–12, 1994. (Cité page 39)

[Eide 04] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny & J. Pitrelli. *A corpus-based approach to< ahem/> expressive speech synthesis.* In 5th ISCA Speech Synthesis Workshop, 2004. (Cité page 151)

[Emerard 92] F. Emerard, L. Mortamet & A. Cozannet. *Prosodic processing in a text-to speech synthesis system using a database and learning procesures.* Talking Machines : Theories, Models, and Designs, pages 225–254, 1992. (Cité page 27)

[Émond 06] C. Émond. *Une analyse prosodique de la parole souriante : une étude préliminaire.* In XXVIe Journées d'étude sur la parole (JEP), pages 147–150, 2006. (Cité page 160)

[Esposito 07] A. Esposito. *The amount of information on emotional states conveyed by the verbal and nonverbal channels : some perceptual data.* Progress in nonlinear speech processing, pages 249–268, 2007. (Cité page 157)

[Fairbanks 58] G. Fairbanks. *Test of phonemic differentiation : The rhyme test.* The Journal of the Acoustical Society of America, vol. 30, page 596, 1958. (Cité page 65)

[Fairon 06] C. Fairon & S. Paumier. *A translated corpus of 30,000 French SMS.* In International conference on Language Resources and Evaluation (LREC), 2006. (Cité page 86)

[Falaise 05] A. Falaise. *Constitution d'un corpus de français tchaté.* RECITAL, 2005. (Cité page 87)

[Fant 53] G. Fant. *Speech communication research.* IVA Royal Swedish Acad. Eng. Sci., vol. 24, pages 331–337, 1953. (Cité page 29)

[Fischer 04] V. Fischer, J.B. Ordinas & S. Kunzmann. *Domain Adaptation Methods in the IBM trainable Text-to-speech System.* In 8th International Conference on Spoken Language Processing. ISCA, 2004. (Cité page 54)

[Fónagy 83] I. Fónagy & R. Jakobson. La vive voix : essais de psycho-phonétique. Bibliothèque scientifique Payot, 1983. (Cité page 24)

[Foundation 10] Python Software Foundation. *Python/C API Reference Manual*, 2010. http://docs.python.org/c-api/index.html. (Cité page 125)

[Francois 02] H. Francois & O. Boeffard. *The greedy algorithm and its application to the construction of a continuous speech database.* In International conference on Language Resources and Evaluation (LREC), volume 5, pages 1420–1426, 2002. (Cité pages 52 and 102)

[François 01] H. François & O. Boëffard. *Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem.* In 7th European Conference on Speech Communication and Technology, 2001. (Cité pages 48, 49, and 50)

[Fujisaki 04] H. Fujisaki. *Information, prosody, and modeling-with emphasis on tonal features of speech.* In Speech Prosody. ISCA, 2004. (Cité page 28)

[Garey 79] M.R. Garey & D.S. Johnson. Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences. WH Freeman and Company, San Francisco, California, 1979. (Cité page 51)

[Gauvain 90] J.L. Gauvain, L.F. Lamel & M. Eskénazi. *Design Considerations and Text Selection for BREF, a large French read-speech corpus.* In First International Conference on Spoken Language Processing, 1990. (Cité pages 15, 46, 48, and 49)

[Guimier de Neef 07] E. Guimier de Neef, A. Debeurme & J. Park. *TiLT correcteur de SMS : évaluation et bilan quantitatif.* In Conférence sur le Traitement Automatique des Langues (TALN), pages 123–32, 2007. (Cité page 22)

[Harris 53] C.M. Harris. *A study of the building blocks in speech.* The Journal of the Acoustical Society of America, vol. 25, pages 962–969, 1953. (Cité page 29)

[Hermansky 90] H. Hermansky. *Perceptual linear predictive (PLP) analysis of speech.* Journal of the Acoustical Society of America, vol. 87, no. 4, pages 1738–1752, 1990. (Cité page 42)

[Hirai 02] T. Hirai, S. Tenpaku & K. Shikano. *Speech unit selection based on target values driven by speech data in concatenative speech synthesis.* In IEEE Workshop on Speech Synthesis, pages 43–46, 2002. (Cité page 38)

[Hirose 06] K. Hirose, Y. Asano & N. Minematsu. *Corpus-Based Generation of Fundamental Frequency Contours Using Generation Process Model and Considering Emotional Focuses.* In 9th International Conference on Spoken Language Processing, 2006. (Cité page 150)

[Hitzeman 99] J. Hitzeman, A. Black, C. Mellish, J. Oberlander, M. Poesio & P. Taylor. *An annotation scheme for concept-to-speech synthesis.* In European Workshop on Natural Language Generation, pages 59–66, 1999. (Cité page 21)

[Hockett 42] C.F. Hockett. *A system of descriptive phonology.* Language, pages 3–21, 1942. (Cité page 24)

[Hocq 06] S. Hocq. Etude des sms en français : constitution et exploitation d'un corpus aligné sms-langue standard. Master's thesis, Rapport de Master II, Industries des Langues, Université de Provence, 2006. (Cité page 86)

[House 65] A.S. House, C.E. Williams, M.H.L. Hecker & KD Kryter. *Articulation-Testing Methods : Consonantal Differentiation with a Closed-Response Set.* The Journal of the Acoustical Society of America, vol. 37, page 158, 1965. (Cité page 65)

[Hu 96] Z. Hu, J. Schalkwyk, E. Barnard & R. Cole. *Speech recognition using syllable-like units.* In 4th International Conference on Spoken Language Processing (ICSLP), volume 2, pages 1117–1120, 1996. (Cité page 82)

[Hueber 07] T. Hueber, G. Chollet, B. Denby, M. Stone & L. Zouari. *Ouisper : corpus based synthesis driven by articulatory data.* In 16th International Congress of Phonetic Sciences, pages 2193–2196, 2007. (Cité page 21)

[Hunt 96] A. Hunt & A.W. Black. *Unit selection in a concatenative speech synthesis system using a large speech database.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 373–376, 1996. (Cité pages 15, 33, 41, and 44)

[Iida 01] A. Iida & N. Campbell. *A database design for a concatenative speech synthesis system for the disabled.* In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001. (Cité page 110)

[Iida 02] A. Iida, N. Campbell, F. Higuchi & M. Yasumura. *A corpus-based speech synthesis system with emotion.* Speech Communication, vol. 40, no. 1-2, pages 161–187, 2002. (Cité pages 16 and 151)

[Isogai 05] M. Isogai, H. Mizuno & K. Mano. *Recording script design for corpus-based TTS system based on coverage of various phonetic elements.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 301–304, 2005. (Cité pages 48 and 50)

[ITU-T 93] ITU-T. *P. 56 recommandation - Objective Measurement of Active Speech Level.* Telephone transmission quality, telephone installations, local line networks, 1993. (Cité pages 56 and 139)

[ITU-T 94] ITU-T. *P. 85 recommandation - A method for subjective performance assessment of the quality of speech voice output devices* . Telephone transmission quality, telephone installations, local line networks, 1994. (Cité pages 66, 67, and 163)

[ITU-T 96] ITU-T. *P. 800 recommandation - Methods for subjective determination of transmission quality.* Telephone transmission quality, telephone installations, local line networks, 1996. (Cité pages 66, 68, 69, and 141)

[ITU-T 01] ITU-T. *P. 862 recommandation - Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.* Telephone transmission quality, telephone installations, local line networks, 2001. (Cité page 64)

[Jensen 00] K.J. Jensen & S. Riis. *Self-organizing letter code-book for text-to-phoneme neural network model.* In 6th International Conference on Spoken Language Processing. ISCA, 2000. (Cité page 27)

[Jolion 06] J.-M. Jolion. *Probabilités et Statistique, Intervalle de confiance sur le coefficient de corrélation.* Rapport technique, INSA-Lyon, Département Génie Industriel, 2006. http://rfv.insa-lyon.fr/~jolion/STAT/node87.html. (Cité pages 92 and 147)

[Kawahara 99] H. Kawahara, I. Masuda-Katsuse & A. de Cheveigné. *Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0extraction : Possible role of a repetitive structure in sounds.* Speech Communication, vol. 27, no. 3, pages 187–208, 1999. (Cité page 40)

[Kawai 00] H. Kawai, S. Yamamoto, N. Higuchi & T. Shimizu. *A design method of speech corpus for text-to-speech synthesis taking account of prosody.* In

6th International Conference on Spoken Language Processing, volume 3, 2000. (Cité pages 48, 59, and 145)

[Kawai 04] H. Kawai, T. Toda, J. Ni, M. Tsuzaki & K. Tokuda. *XIMERA : A new TTS from ATR based on corpus-based technologies.* In 5th ISCA Workshop on Speech Synthesis (SSW), 2004. (Cité pages 44 and 45)

[Khmaladze 88] EV Khmaladze. *The statistical analysis of a large number of rare events.* Rapport technique MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, 1988. (Cité page 46)

[King 09] S. King & V. Karaiskosb. *The blizzard challenge 2009*, 2009. (Cité page 69)

[Klabbers 98] E. Klabbers & R. Veldhuis. *On the reduction of concatenation artefacts in diphone synthesis.* In 5th International Conference on Spoken Language Processing, 1998. (Cité page 42)

[Klabbers 01a] E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner & S. Breuer. *Speech synthesis development made easy : The Bonn Open Synthesis System.* In 7th European Conference on Speech Communication and Technology, 2001. (Cité page 81)

[Klabbers 01b] E. Klabbers & R. Veldhuis. *Reducing audible spectral discontinuities.* IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, pages 39–51, 2001. (Cité page 43)

[Klatt 80] D.H. Klatt. *Software for a cascade/parallel formant synthesizer.* Journal of the Acoustical Society of America, vol. 67, no. 3, pages 971–995, 1980. (Cité page 29)

[Klatt 87] D.H. Klatt. *Review of text-to-speech conversion for English.* The Journal of the Acoustical Society of America, vol. 82, pages 737–793, 1987. (Cité page 15)

[Kobus 08] C. Kobus, F. Yvon & G. Damnati. *Normalizing SMS : are two metaphors better than one ?* In 22nd International Conference on Computational Linguistics, volume 1, pages 441–448. Association for Computational Linguistics, 2008. (Cité page 22)

[Kominek 03] J. Kominek & A. Black. *The CMU ARCTIC speech databases for speech synthesis.* Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177, http://festvox.org/cmu_arctic/cmu_arctic_report.pdf, 2003. (Cité pages 50 and 53)

[Krul 08] A. Krul. *Construction et réduction de la base de parole adaptées à une application spécifique de la synthèse par corpus.* PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2008. (Cité pages 25, 50, and 54)

[Kullback 51] S. Kullback & R.A. Leibler. *On information and sufficiency.* The Annals of Mathematical Statistics, vol. 22, no. 1, pages 79–86, 1951. (Cité page 42)

[Kumar 04] R. Kumar. *A genetic algorithm for unit selection based speech synthesis.* In 8th International Conference on Spoken Language Processing, 2004. (Cité page 38)

[Kuo 02] C.C. Kuo & J.Y. Huang. *Efficient and scalable methods for text script generation in corpus-based TTS design.* In 7th International Conference on Spoken Language Processing (ICSLP). ISCA, 2002. (Cité pages 48, 49, and 50)

[Laferrière 85] P. Laferrière, G. Chollet, L. Miclet & J.P. Tubach. *Segmentation d'une base de données de "polysons", application à la synthèse de la parole.* In XIVèmes Journées d'Etude sur la Parole, 1985. (Cité page 81)

[Lambert 04] T. Lambert & A. Breen. *A database design for a TTS synthesis system using lexical diphones.* In 8th International Conference on Spoken Language Processing. ISCA, 2004. (Cité pages 48 and 49)

[Lambert 07] T. Lambert, N. Braunschweiler & S. Buchholz. *How (Not) to Select Your Voice Corpus : Random Selection vs. Phonologically Balanced.* In 6th ISCA Speech Synthesis Workshop (SSW6), 2007. (Cité page 45)

[Laprie 98] Y. Laprie & V. Colotte. *Automatic pitch marking for speech transformations via TD-PSOLA.* In Eusipco : European signal processing conference, pages 1133–1136, 1998. (Cité page 60)

[Laroche 99] J. Laroche & M. Dolson. *Improved phase vocoder time-scale modification of audio.* IEEE Transactions on Speech and Audio processing, vol. 7, no. 3, page 323, 1999. (Cité page 40)

[Larreur 89] D. Larreur, F. Emerard & F. Marty. *Linguistic and prosodic processing for a text-to-speech synthesis system.* In First European Conference on Speech Communication and Technology, 1989. (Cité pages 25 and 26)

[Larreur 94] D. Larreur. *Document de travail DT/461/LAA/TSS/RCP, Marqueurs utilisés dans CNETVOX94.* Rapport technique, Centre National d'Etudes des Télécommunications, Département Recherche en Communication par la Parole, 1994. (Cité page 73)

[Lasarcyk 07] E. Lasarcyk & J. Trouvain. *Imitating conversational laughter with an articulatory speech synthesizer.* In Interdisciplinary Workshop on The Phonetics of Laughter, pages 43–48, 2007. (Cité page 157)

[Lawrence 53] W. Lawrence. *The synthesis of speech from signals which have a low information rate.* In Symposium on applications of communication theory, Institution of Electrical Engineers, London, pages 460–469. Butterworths Scientific Publications, 1953. (Cité page 29)

[Lebert 08] M. Lebert. *Le Projet Gutenberg (1971-2008).* Projet Gutenberg, 2008. http://www.gutenberg.org/ebooks/27045. (Cité page 50)

[Liljencrants 68] JC Liljencrants. *The OVE III speech synthesizer.* IEEE Transactions on Audio and Electroacoustics, vol. 16, no. 1, pages 137–140, 1968. (Cité page 29)

[Lindblom 63] B. Lindblom. *Spectrographic study of vowel reduction.* The journal of the Acoustical society of America, vol. 35, page 783, 1963. (Cité page 43)

[Maeda 79] S. Maeda. *An articulatory model of the tongue based on a statistical analysis.* The Journal of the Acoustical Society of America, vol. 65, page S22, 1979. (Cité page 28)

[Maia 07] R. Maia, T. Toda, H. Zen, Y. Nankaku & K. Tokuda. *A trainable excitation model for HMM-based speech synthesis.* In 8th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 1909–1912, 2007. (Cité page 31)

[Makashay 00] M.J. Makashay, C.W. Wightman, A.K. Syrdal & A. Conkie. *Perceptual evaluation of automatic segmentation in text-to-speech synthesis.* In 6th International Conference on Spoken Language Processing (ICSLP), volume 2, pages 431–434, 2000. (Cité pages 59 and 145)

[Matoušek 05] J. Matoušek, Z. Hanzlıcek & D. Tihelka. *Hybrid Syllable/Triphone Speech Synthesis.* 6th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 2529–2532, 2005. (Cité page 33)

[Mayo 05] C. Mayo, R.A.J. Clark & S. King. *Multidimensional scaling of listener responses to synthetic speech.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 69)

[Möbius 03] B. Möbius. *Rare events and closed domains : Two delicate concepts in speech synthesis.* International Journal of Speech Technology, vol. 6, no. 1, pages 57–71, 2003. (Cité page 47)

[Mermelstein 73] P. Mermelstein. *Articulatory model for the study of speech production.* Journal of the Acoustical Society of America, vol. 53, no. 4, pages 1070–1082, 1973. (Cité page 28)

[Mertens 01] P. Mertens, A. Auchlin, J.P. Goldman & A. Grobet. *L'intonation du discours : une implémentation par balises ; motifs et premiers résultats.* Journées Prosodie, 2001. (Cité page 21)

[Mohri 02a] M. Mohri. *Semiring frameworks and algorithms for shortest-distance problems.* Journal of Automata, Languages and Combinatorics, vol. 7, no. 3, pages 321 – 350, 2002. (Cité page 125)

[Mohri 02b] M. Mohri, F. Pereira & M. Riley. *Weighted finite-state transducers in speech recognition.* Computer Speech & Language, vol. 16, no. 1, pages 69 – 88, 2002. (Cité page 117)

[Moulines 90] E. Moulines & F. Charpentier. *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones.* Speech Communication, vol. 9, no. 5-6, pages 453–467, 1990. (Cité page 30)

[Nagarajan 03] T. Nagarajan, H.A. Murthy & R.M. Hegde. *Segmentation of speech into syllable-like units.* In Eurospeech, 2003. (Cité page 82)

[Ni 06] J. Ni, T. Hirai & H. Kawai. *Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for English speech synthesis.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, 2006. (Cité pages 48 and 49)

[Ni 07] J. Ni, T. Hirai, H. Kawai, T. Toda, K. Tokuda, M. Tsuzaki, S. Sakai, R. Maia & S. Nakamura. *ATRECSS-ATR english speech corpus for speech synthesis.* In Speech Synthesis Workshop 6, Blizzard Challenge, 2007. (Cité pages 49 and 50)

[Olive 85] J.P. Olive & MY Liberman. *Text to speech - an overview.* The Journal of the Acoustical Society of America, vol. 78, page S6, 1985. (Cité page 30)

[Pantazis 05] Y. Pantazis, Y. Stylianou & E. Klabbers. *Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 43)

[Peterson 58] G.E. Peterson, S. William, Y. Wang & E. Sivertsen. *Segmentation techniques in speech synthesis.* The Journal of the Acoustical Society of America, vol. 30, pages 739–742, 1958. (Cité page 29)

[Pfitzinger 04] H.R. Pfitzinger. *DFW-based spectral smoothing for concatenative speech synthesis.* In 8th International Conference on Spoken Language Processing, 2004. (Cité pages 16 and 39)

[Pfitzinger 06] H.R. Pfitzinger. *Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction.* In Speech Prosody, pages 6–9, 2006. (Cité page 24)

[Pierrehumbert 83] JB Pierrehumbert. *Linguistic units for FO synthesis.* In Abstracts of the 10th International Congress of Phonetic Sciences, pages 137–144, 1983. (Cité page 28)

[Pols 87] L.C.W. Pols, J.P. Lefevre, G. Boxelaar & N. Son. *Word intelligibility of a rule synthesis system for French.* In European Conference on Speech Technology. ISCA, 1987. (Cité page 30)

[Popescu 06] A. Popescu, C. Boidin & D. Cadic. *Contraintes globales pour la sélection des unités en synthèse vocale.* In XVIèmes Journées d'Etude sur la Parole, 2006. (Cité page 38)

[Raux 03] A. Raux & A.W. Black. *A unit selection approach to f0 modeling and its application to emphasis.* In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 700–705. IEEE, 2003. (Cité pages 16 and 150)

[Revelin 05] S. Revelin, D. Cadic & C. Waast-Richard. *Optimization of Text-to-Speech Phonetic Transcriptions using A-Posteriori Signal Comparison.* In 9th European Conference on Speech Communication and Technology. ISCA, 2005. (Cité pages 41 and 140)

[Richard 95] G. Richard, M. Liu, D. Snider, H. Duncan, Q. Lin, J.L. Flanagan, S. Levinson, D. Davis & S. Slimon. *Numerical simulations of fluid flow in the vocal tract.* In 4th European Conference on Speech Communication and Technology. ISCA, 1995. (Cité page 28)

[Rix 01] A.W. Rix, J.G. Beerends, M.P. Hollier & A.P. Hekstra. *Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 2, pages 749–752, 2001. (Cité page 63)

[Rossi 81] M. Rossi, A. Di Cristo, D. Hirst, P. Martin & Y. Nishinuma. L'intonation : de l'acoustique à la sémantique. Klincksieck, 1981. (Cité page 26)

[Rozenknop 01] A. Rozenknop & M.C. Silaghi. *Algorithme de décodage de treillis selon le critère du coût moyen pour la reconnaissance de la parole.* Conférence sur le Traitement Automatique des Langues (TALN), 2001. (Cité page 116)

[Rubin 81] P. Rubin, T. Baer & P. Mermelstein. *An articulatory synthesizer for perceptual research.* Journal of the Acoustical Society of America, vol. 70, no. 2, pages 321–328, 1981. (Cité page 28)

[Rutten 02] P. Rutten, M.P. Aylett, J. Fackrell & P. Taylor. *A statistically motivated database pruning technique for unit selection synthesis.* In 7th International Conference on Spoken Language Processing (ICSLP). ISCA, 2002. (Cité page 61)

[Sagisaka 88] Y. Sagisaka. *Speech synthesis by rule using optimal selection of non-uniform synthesis units.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 679–682, 1988. (Cité pages 15 and 33)

[Schmidt 93] M. Schmidt, S. Fitt, C. Scott & M.A. Jack. *Phonetic transcription standards for European names (ONOMASTICA).* In 3rd European Conference on Speech Communication and Technology, 1993. (Cité page 26)

[Schroeder 93] M.R. Schroeder. *A brief history of synthetic speech.* Speech Communication, vol. 13, no. 1-2, pages 231–237, 1993. (Cité page 15)

[Schweitzer 03] A. Schweitzer, N. Braunschweiler, T. Klankert, B. M "obius & B. Sauberlich. *Restricted unlimited domain synthesis.* In 8th European Conference on Speech Communication and Technology, 2003. (Cité pages 39 and 54)

[Ségalen 08] L. Ségalen & D. Cadic. *Introduction d'éléments paralinguistiques en synthèse vocale.* In XVIIèmes Journées d'Etude sur la Parole, 2008. (Cité page 157)

[Shimei 97] Pan Shimei & McKeown Kathleen R. *Integrating language generation with speech synthesis in a concept to speech system.* In ACL workshop on Concept to Speech Generation Systems, pages 23–28, 1997. (Cité page 21)

[Silverman 92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert & J. Hirschberg. *ToBI : A standard for labeling English prosody.* In 2nd International Conference on Spoken Language Processing, volume 2, pages 867–870, 1992. (Cité page 28)

[Sorin 92] C. Sorin, D. Jouvet, M. Toularhoat, D. Dubois, B. Cherbonnel, D. Bigorgne & C. Cagnoulet. *CNET speech recognition and text-to-speech in telecommunications applications.* In First IEEE Workshop on Interactive Voice Technology for Telecommunications applications, 1992. (Cité page 15)

[Stöber 99] K. Stöber, T. Portele, P. Wagner & W. Hess. *Synthesis by word concatenation.* In 6th European Conference on Speech Communication and Technology, 1999. (Cité page 33)

[Stylianou 96] Y. Stylianou. *Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modification.* PhD thesis, École nationale supérieure des télécommunications, 1996. (Cité page 39)

[Stylianou 99] Y. Stylianou. *Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 377–380, 1999. (Cité page 56)

[Stylianou 01a] Y. Stylianou. *Applying the harmonic plus noise model in concatenative speech synthesis.* IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, pages 21–29, 2001. (Cité pages 16 and 39)

[Stylianou 01b] Y. Stylianou & A.K. Syrdal. *Perceptual and objective detection of discontinuities in concatenative speech synthesis.* In International Conference on Acoustics, Speech and Signal Processing, volume 2. IEEE, 2001. (Cité pages 42 and 43)

[Suciu 06] I. Suciu, I. Kanellos & T. Moudenc. *What about the text ? Modeling global expressiveness in speech synthesis.* IEEE International Conference on Information and Communication Technologies : from Theory to Applications (ICTTA), vol. 1, 2006. (Cité page 23)

[Syrdal 01] A.K. Syrdal. *Phonetic effects on listener detection of vowel concatenation.* In 7th European Conference on Speech Communication and Technology, 2001. (Cité page 42)

[Syrdal 05] A.K. Syrdal & A.D. Conkie. *Perceptually-based data-driven join costs : comparing join types.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 42)

[Syrdal 08] A.K. Syrdal & Y.J. Kim. *Dialog speech acts and prosody : Considerations for TTS.* In Speech Prosody, pages 661–665, 2008. (Cité pages 16 and 151)

[Tamura 01] M. Tamura, T. Masuko, K. Tokuda & T. Kobayashi. *Text-to-speech synthesis with arbitrary speaker's voice from average voice.* In 7th European Conference on Speech Communication and Technology, 2001. (Cité pages 16 and 32)

[Taylor 99] P. Taylor & A.W. Black. *Speech synthesis by phonological structure matching.* In 6th European Conference on Speech Communication and Technology, 1999. (Cité pages 38 and 54)

[Taylor 00] P. Taylor. *Analysis and synthesis of intonation using the tilt model.* The Journal of the acoustical society of America, vol. 107, pages 1697–1714, 2000. (Cité page 28)

[Taylor 05] P. Taylor. *Hidden Markov Models for grapheme to phoneme conversion.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 27)

[Taylor 06] P. Taylor. *The Target Cost Formulation in Unit Selection Speech Synthesis.* In 9th International Conference on Spoken Language Processing. ISCA, 2006. (Cité pages 16 and 41)

[Thomas 06] S. Thomas, M.N. Rao, H.A. Murthy & C.S. Ramalingam. *Natural sounding TTS based on syllable-like units.* In EUSIPCO, 2006. (Cité page 82)

[Tian 05] J. Tian, J. Nurminen & I. Kiss. *Optimal subset selection from text databases.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 305–308, 2005. (Cité page 49)

[Toda 04] T. Toda, H. Kawai & M. Tsuzaki. *Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, 2004. (Cité pages 41 and 146)

[Toda 05] T. Toda & K. Shikano. *NAM-to-speech conversion with Gaussian mixture models.* In 6th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2005. (Cité page 21)

[Tokuda 95] K. Tokuda, T. Kobayashi & S. Imai. *Speech parameter generation from HMM using dynamic features.* In International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 660–663. IEEE, 1995. (Cité pages 16 and 30)

[Torzec 01] N. Torzec, T. Moudenc & F. Emerard. *Prétraitement et analyse linguistique dans le système de synthèse TTS CVOX : Application à la vocalisation automatique d'e-mails.* Traitement Automatique des Langues, vol. 42, no. 1, pages 17–46, 2001. (Cité pages 22 and 40)

[Trouvain 03] J. Trouvain. *Segmenting phonetic units in laughter.* In 15th International Congress of Phonetic Sciences, pages 2793–2796, 2003. (Cité page 160)

[Trouvain 04] J. Trouvain & M. Schröder. *How (not) to add laughter to synthetic speech.* Affective Dialogue Systems, pages 229–232, 2004. (Cité page 157)

[Vaissiere 80] J. Vaissiere. *La structuration acoustique de la phrase française*. Annali della scuola normale superiore di Pisa, vol. 3, no. 10, pages 529–560, 1980. (Cité page 24)

[van Santen 93] J.P.H. van Santen. *Perceptual experiments for diagnostic testing of text-to-speech systems*. Computer Speech & Language, vol. 7, no. 1, pages 49–100, 1993. (Cité page 63)

[Van Santen 97a] J.P.H. Van Santen. *Combinatorial issues in text-to-speech synthesis*. In 5th European Conference on Speech Communication and Technology. ISCA, 1997. (Cité pages 50, 86, and 149)

[Van Santen 97b] J.P.H. Van Santen & A.L. Buchsbaum. *Methods for optimal text selection*. In Eurospeech, volume 97, page 2, 1997. (Cité pages 15, 52, 87, and 109)

[Vepa 04] J. Vepa & S. King. *Join Cost for Unit Selection Speech Synthesis*. In Abeer Alwan & Shri Narayanan, editeurs, Speech Synthesis. Prentice Hall, 2004. (Cité pages 16 and 42)

[Vincent 06] D. Vincent, O. Rosec & T. Chonavel. *Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, 2006. (Cité pages 60 and 139)

[Viterbi 67] A. Viterbi*et al.* *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. IEEE transactions on Information Theory, vol. 13, no. 2, pages 260–269, 1967. (Cité page 37)

[Vosnidis 01] C. Vosnidis & V. Digalakis. *Use of Clustering Information for Coarticulation Compensation in Speech Synthesis by Word Concatenation*. In 7th European Conference on Speech Communication and Technology, 2001. (Cité page 33)

[Wakita 81] H. Wakita. *Linear prediction voice synthesizers : Line spectrum pairs (LSP) is the newest of several techniques*. Speech Technol, vol. 1, pages 17–22, 1981. (Cité page 39)

[Wouters 98] J. Wouters & M.W. Macon. *A perceptual evaluation of distance measures for concatenative speech synthesis*. In 5th International Conference on Spoken Language Processing, 1998. (Cité page 42)

[Wouters 01] J. Wouters & M.W. Macon. *Control of spectral dynamics in concatenative speech synthesis*. IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, 2001. (Cité pages 16 and 39)

[Xydas 04] G. Xydas, D. Spiliotopoulos & G. Kouroupetroglou. *Modeling prosodic structures in linguistically enriched environments*. Text, Speech and Dialogue, pages 521–528, 2004. (Cité page 21)

[Yamagishi 03] J. Yamagishi, K. Onishi, T. Masuko & T. Kobayashi. *Modeling of various speaking styles and emotions for HMM-based speech synthesis*. In 8th European Conference on Speech Communication and Technology. ISCA, 2003. (Cité pages 16 and 150)

[Yamagishi 08] J. Yamagishi, Z. Ling & S. King. *Robustness of HMM-based speech synthesis*. In 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2008. (Cité pages 16 and 33)

[Yi 98] J.R.W. Yi & J.R. Glass. *Natural-Sounding Speech Synthesis Using Variable-Length Units*. In 5th International Conference on Spoken Language Processing. ISCA, 1998. (Cité pages 43 and 54)

[Yoshimura 00] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi & T. Kitamura. *Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis.* IEICE Transactions on Information and Systems, vol. 83, no. 11, pages 2099–2107, 2000. (Cité pages 30 and 75)

[Young 79] SJ Young & F. Fallside. *Speech synthesis from concept : a method for speech output from information systems.* The Journal of the Acoustical Society of America, vol. 66, pages 685–695, 1979. (Cité page 21)

[Young 05] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev & P. Woodland. *The HTK book*, 2005. http://htk.eng.cam.ac.uk/docs/docs.shtml. (Cité page 57)

[Yvon 96] F. Yvon. *Prononcer par analogie : motivation, formalisation et évaluation.* PhD thesis, École nationale supérieure des télécommunications, 1996. (Cité page 27)

[Zen 05] H. Zen & T. Toda. *An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005.* In 9th European Conference on Speech Communication and Technology, 2005. (Cité page 31)

[Zipf 32] G. K. Zipf. Selective studies and the principle of relative frequency in language. Harvard University Press, Cambridge, MA, 1932. (Cité page 46)


diphones
sélection
séquences
unités
construction
création
groupes
travaux
fonction
phrases
texte
voyelles
vocale
segmentation
fragiles
paralinguistique
élément
séquence
script
acoustique
partie
robustes
parole
lecture
niveau
enregistrement
nombre
symboles
systèmes
contexte
mots
sans
bien
automates
longueur
couverture
automatique
sandwich
synthèse
qualité
concaténations
référence
critères
taille
sandwichs
locuteur
phonétique
taux
effet
données
choix
ensemble
exemple
enregistrements
coût
phonèmes
système
prosodique
évaluation
phones
signal
VSCR
vocaliques
voix
état
cible
condensation
liquides
approche
unité
base
optimisation
procédé
algorithme
contextuels
corpus
concaténation
résultats
critère
rushes