Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française

Sébastien Le Maguer

▶ **To cite this version:**

Sébastien Le Maguer. Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française. Autre [cs.OH]. Université de Rennes, 2013. Français. NNT : 2013REN1S088 . tel-00913565v2

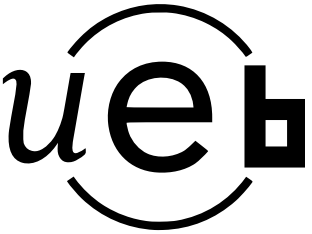
HAL Id: tel-00913565
https://theses.hal.science/tel-00913565v2

Submitted on 21 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANNÉE 2013



THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Sébastien LE MAGUER

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
IFSIC

Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française

Thèse soutenue à Lannion
le 2 Juillet 2013

devant le jury composé de :

Christophe D'ALESSANDRO
Directeur de Recherche CNRS / *Président, Rapporteur*
Yannick ESTÈVE
Professeur à l'université du Maine / *Rapporteur*
Vincent COLOTTE
Maître de conférence au LORIA / *Examinateur*
Olivier BOEFFARD
Professeur à l'Université de Rennes1 / *Directeur de thèse*
Nelly BARBOT
Maître de conférence à l'Université de Rennes1 / *Co-directrice de thèse*

Où on va ? J'en sais rien mais on y va !
Pierre Fournier

Remerciements

Je tiens à remercier le jury pour avoir jugé et analysé mes travaux. Un grand merci à Christophe d'Alessandro pour avoir présidé mon jury et pour avoir rapporté mes travaux. Merci à Yannick Estève pour avoir également rapporté mes travaux et à Vincent Colotte pour avoir participé à ce jury. Je remercie également le conseil général des Côtes d'Armor pour avoir financé mes travaux.

Je remercie chaleureusement Nelly Barbot et Olivier Boëffard pour m'avoir fait l'honneur d'encadrer mes travaux de thèse, de partager avec moi leurs nombreuses connaissances et de m'avoir guider durant ces cinq ans. Malgré ses difficultés, la thèse a été pour moi un enrichissement tant sur le plan des connaissances que sur le plan humain. Merci encore pour tout !

Je remercie mes collègues de l'ENSSAT qui m'ont si bien accueilli et fait sentir que je faisais parti de la « famille » ENSSAT. Un grand merci à Daniel et Jean Christophe pour m'avoir donné la chance d'enseigner à l'ENSSAT et de m'avoir fait confiance. Merci à Pierre, Hélène, Damien, Nelly, Laure et Vincent pour leur accueil et les échanges si amicaux. Merci à Laurent et Olivier pour les discussions culturelles autour des civilisations anciennes ou futures. Merci à Jonathan et Gwénolé pour les moments de détentes qui, parfois, redevenaient des discussions de travail.

Merci aux doctorants de l'époque, Nourédine, Katia, Amine, David, William, Larbi et Anouar, avec qui nous avons pu échanger et partager beaucoup plus que le simple couloir pendant ces années. Merci aux doctorants que j'ai rencontré lors des formations pour leurs échanges si amicaux : Laëtitia, François et Alice.

Un grand merci à mes deux familles : Papa, Maman, Belle-Maman, le « frangin », Hubert, Brigitte, Clément, Matthieu, Héloïse, Estelle, Stéphane et Juliette. Merci pour votre soutien sans faille. Sans vous je n'aurais jamais pu espérer arriver là où je suis.

Merci aux enseignants que j'ai eu au lycée qui m'ont appris le goût d'apprendre et le goût de faire apprendre : Nelly, Emmanuelle, Gilles, Jean-Yves. Merci également à mes enseignants de DUT devenus collègues et amis : Arnaud, Adib, Jean-Christophe, ainsi que tout le département informatique.

Je remercie également les étudiants à qui j'ai enseigné et qui m'ont appris beaucoup de chose.

Je tiens enfin à remercier les musiciens (en particulier Anneke, Rob, Johann et Tony) ayant composé les musiques qui ont jalonnées mon parcours de thèse et qui, dans les moments de doutes,

m'ont redonné l'énergie nécessaire pour arriver au bout de ces travaux.

Table des matières

Introduction

La production artificielle de la parole a vu le jour en 1791 lorsque le baron Von Kempelen a mis au point la première machine imitant la physiologie (connue à l'époque) de l'appareil phonatoire pour générer une vingtaine de sons mécaniquement. Le XX^e siècle a vu l'essor de la production automatique de la parole qui passa du stade purement mécanique à des systèmes électriques (le premier système étant le VODER présenté à la fin des années 1930 par Homer Dudley) puis informatiques.

Durant les années 1960, une nouvelle méthodologie de production automatique de la parole fut proposée : la synthèse à partir du texte ou synthèse TTS. Contrairement aux méthodes de synthèse précédentes, la synthèse à partir du texte ne consiste plus à utiliser un opérateur pour contrôler le système de production mais à extraire des informations du texte pour produire le signal audio correspondant. À cette période, le premier système de synthèse dit par concaténation d'unités a été publié. Cette méthodologie tranche avec l'état de l'art de la synthèse de l'époque, qui repose sur un ensemble de règles permettant de prédire le signal de parole à générer, car elle repose sur l'utilisation d'unités de parole naturelle réellement dictées par un locuteur humain. L'augmentation de la puissance de calcul et de la capacité de stockage des ordinateurs a permis l'évolution de la synthèse par concaténation : les systèmes de synthèse, dits par sélection d'unités, associent dorénavant plusieurs représentants pour une unité à produire. L'algorithme de synthèse consiste alors à déterminer quelles sont les meilleures unités à utiliser pour générer le signal de parole associé au texte à synthétiser. La dernière évolution majeure de la synthèse de la parole eut lieu au milieu des années 1990. Toujours grâce à l'évolution des machines de calcul, la synthèse paramétrique basée sur des modèles statistiques a fait son apparition. Il s'agit d'une méthode de synthèse qui dérive de la synthèse par règles. Des règles statistiques remplacent les règles proposées dans les années 1950 pour représenter la source et le filtre.

Depuis le milieu des années 2000, deux méthodes sont prédominantes dans le domaine de la synthèse de la parole : la sélection d'unités et la synthèse paramétrique basée sur des modèles de Markov cachés (ou HMM). À l'heure actuelle, le système de synthèse HTS (HMM-based Speech Synthesis System ou H Triple S), qui est le représentant le plus connu et le plus utilisé pour la synthèse par HMM, fait l'objet d'une attention particulière. En effet, ce système présente l'avantage de pouvoir produire une synthèse intelligible et fluide en utilisant moins de données qu'un système par sélection d'unités. En revanche,

contrairement à la synthèse par sélection, le signal de synthèse produit par HTS est de plus faible qualité.

Problématique

Pour effectuer une synthèse en utilisant le système HTS, il est nécessaire d'associer au signal une description basée sur un ensemble conséquent de descripteurs linguistiques et prosodiques. Nous pouvons distinguer deux limites importantes résultant de la combinatoire issue de cette description. Tout d'abord, définir un corpus permettant de couvrir l'ensemble des combinaisons possibles est irréalisable. Cela implique que, lors de la phase de synthèse, des combinaisons que l'on souhaite synthétiser peuvent ne pas avoir été vues lors de la phase d'apprentissage. De plus, le système HTS repose sur une modélisation statistique. Ainsi, le nombre d'occurrences associées à une combinaison de descripteurs, dans le corpus d'apprentissage, peut également être trop faible pour obtenir un modèle pertinent. Bien que le système HTS propose l'utilisation d'arbres de décision pour pallier ces problèmes, la question de la pertinence des descripteurs utilisés se pose et est d'autant plus critique que l'un des avantages du système HTS consiste à pouvoir effectuer une synthèse en utilisant un corpus d'apprentissage de taille relativement faible.

Les travaux de thèse présentés dans ce document ont pour objet l'évaluation de l'influence des descripteurs sur la modélisation effectuée par HTS et donc la qualité du signal de synthèse. De plus, nous avons restreint ces travaux à l'évaluation dans le cadre d'une synthèse en français. Pour effectuer cela, nous avons proposé un ensemble de 44 descripteurs. Le premier objectif consiste à déterminer s'il est possible d'obtenir un ensemble de descripteurs (nous appellerons cet ensemble un jeu de descripteurs) qui, bien que réduit, permet d'obtenir une qualité de synthèse équivalente à l'utilisation au jeu de descripteurs complet. Si un tel ensemble a pu être défini alors le second objectif consiste à déterminer quel est l'apport des descripteurs retenus sur la modélisation effectuée par HTS.

Organisation du document

Pour présenter les travaux de thèse qui ont été réalisés, ce document s'articule autour de trois parties. Tout d'abord, un état de l'art concernant la synthèse de la parole, et plus spécifiquement le système HTS, est réalisé. La seconde partie est réservée à la présentation des protocoles et données expérimentales proposés pour répondre à la problématique. Enfin, la dernière partie présente l'ensemble des résultats obtenus par les protocoles précédemment évoqués.

La première partie, intitulée « Le système HTS », présente le contexte scientifique dans lequel se situe HTS. Le premier chapitre (« Synthèse de la parole » page 9) débute

par une présentation générale du phénomène linguistique et acoustique qu'est la parole. Ce chapitre se poursuit par la description des méthodes de synthèse TTS dominantes que sont la sélection d'unités et la synthèse par HMM. Le second chapitre (« Système HTS - Présentation » page 27) permet de présenter en détail le système HTS que nous allons étudier. Pour cela, ce chapitre débute par l'introduction des concepts utilisés par ce système ainsi que les processus d'apprentissage et de synthèse. La dernière partie de ce chapitre est consacrée à la présentation de la configuration du système HTS utilisée pour réaliser les expériences. Enfin le troisième chapitre (« Système HTS - Jeux de descripteurs » page 53) est consacré à la définition du jeu de descripteurs proposé pour effectuer une synthèse en français. Ainsi, ce chapitre débute par la présentation des jeux de descripteurs connus pour effectuer une synthèse dans diverses langues. À l'heure actuelle, aucun jeu adapté au français n'ayant fait l'objet d'une publication, le chapitre se poursuit par la présentation du jeu de descripteurs que nous avons proposé pour effectuer notre étude. Enfin, la dernière section de ce chapitre porte sur l'analyse des études ayant pour objet l'influence des descripteurs sur la modélisation effectuée par HTS.

La seconde partie, intitulée « Évaluation HTS - Méthodologie et données expérimentales », présente les protocoles expérimentaux et les données nécessaires pour la mise en place de ces protocoles. Ainsi, le chapitre 4 (« Méthodologie d'évaluation », page 75) présente deux méthodes d'évaluation objective. Le premier protocole consiste à évaluer l'espace acoustique des coefficients générés par le système HTS en utilisant un méthodologie issue du domaine de l'identification du locuteur. Le second protocole repose sur le calcul de distances entre les coefficients générés par HTS et les coefficients extraits du signal naturel. En combinant les distances obtenues selon les caractéristiques linguistiques et prosodiques des segments auxquels appartiennent les coefficients, ce protocole permet d'évaluer plus finement l'influence de descripteurs. Le second chapitre 5 (« Données expérimentales » page 89) présente le corpus utilisé lors des évaluations puis les jeux de descripteurs évalués.

La dernière partie, intitulée « Évaluation HTS - Résultats pour le français », présente les résultats des expériences et leur analyse. Cette partie est composée de trois chapitres. Les chapitres 6 et 7 (« Évaluation objective - Évaluation par GMM » page 119 et « Évaluation objective - Évaluation non paramétrique » page 137) décrivent, respectivement, les résultats des premier et second protocoles d'évaluation objective. Le dernier chapitre de ce document (« Évaluation subjective » page 163) présente deux évaluations subjectives réalisées dans le but d'éprouver les résultats obtenus par les évaluations objectives.

Première partie

Le système HTS

Introduction à la première partie

La première partie de ce document présente un état de l'art concernant le système HTS. Cette partie se décompose en trois chapitres. Le premier chapitre (intitulé « Synthèse de la parole », page 9) permet de situer le système HTS dans le domaine de la synthèse de la parole à partir du texte. Ce chapitre débute par une introduction générale à la parole pour se focaliser sur l'architecture et les méthodes utilisées dans le cadre de la synthèse à partir du texte.

Le second chapitre (intitulé « Système HTS - Présentation », page 27) détaille les concepts utilisés par HTS pour obtenir des modèles qui seront utilisés pour générer le signal de parole. Ce chapitre débute par la présentation des concepts utilisés par HTK[Young1993, Young2005] (HMM ToolKit), qui est un système de reconnaissance de la parole, sur lequel se base le système HTS. Après avoir indiqué quels problèmes se posent, pour l'utilisation de HMM dans un cadre génératif, nous détaillerons les algorithmes et concepts utilisés pour apprendre les modèles utilisés pour générer les trajectoires lors de la phase de synthèse. Enfin, nous présenterons les processus d'apprentissage et de génération ainsi que la configuration du système HTS utilisée dans le cadre de ces travaux.

Le troisième et dernier chapitre de cette partie (intitulé « Système HTS - Jeux de descripteurs », page 53) présente un état de l'art concernant les jeux de descripteurs. Un jeu de descripteurs permet de caractériser symboliquement un signal de parole et est nécessaire lors de la phase de synthèse pour déterminer les modèles à utiliser. Ce chapitre débute par la présentation du jeu de descripteurs standard, proposé pour l'anglais, puis se poursuit par une comparaison des jeux proposés pour d'autres langues par rapport à ce jeu standard. Aucun jeu de descripteurs n'ayant été publié pour le français, les descripteurs proposés pour les modules de prédiction de durées et de prosodies des systèmes de synthèse de parole, pour le français, seront ensuite présentés. En se basant sur ce recensement, nous avons proposé un jeu de descripteurs spécifique au français pour effectuer une synthèse HTS qui sera présenté dans un troisième temps. La dernière section de ce chapitre expose les études existantes qui permettent d'évaluer l'influence des descripteurs sur la modélisation, et donc la synthèse, effectuée par HTS. Nos travaux ayant pour objectif de déterminer quelle est l'influence des descripteurs sur la modélisation effectuée par HTS, cette dernière section présente les études auxquelles nous devrons nous comparer.

Chapitre 1

Synthèse de la parole

Dans ce premier chapitre, nous allons présenter un état de l'art concernant la synthèse de la parole à partir du texte. Cette présentation va s'effectuer en deux temps. Tout d'abord, nous allons décrire globalement la parole et la modélisation de ce phénomène grâce à des concepts issus de la théorie du traitement du signal. Nous présenterons ensuite la synthèse de la parole à partir du texte (ou TTS pour Text-To-Speech) en décrivant l'architecture d'un système TTS et les méthodes de synthèse dominantes.

1.1 La parole

L'objectif de cette section est de présenter les notions concernant la parole et sa modélisation sur lesquelles reposeront les principes de synthèse de parole à partir du texte. Pour cela, nous allons décrire rapidement le dispositif physiologique utilisé par l'être humain pour produire une parole. Dans un second temps, nous décrirons la parole d'un point de vue acoustique afin d'expliciter la nature complexe de ce phénomène. Ceci nous permettra d'introduire la notion de modèle source/filtre sur laquelle se base le

système de synthèse objet de nos travaux. Le dernier point de cette section est consacré à la perception de la parole.

1.1.1 Physiologie

L'appareil physiologique de production du signal de parole, appelé appareil phonatoire et illustré par la figure 1.1, se décompose en trois parties : l'appareil respiratoire (composé des poumons et de la trachée) origine d'un flux d'air ; le larynx qui permet d'obtenir un flux de nature vibratoire par l'action des cordes vocales, ou signal glottique ; le conduit vocal, composé des articulateurs, qui permet de moduler le signal glottique pour obtenir le signal de parole rayonné aux lèvres.

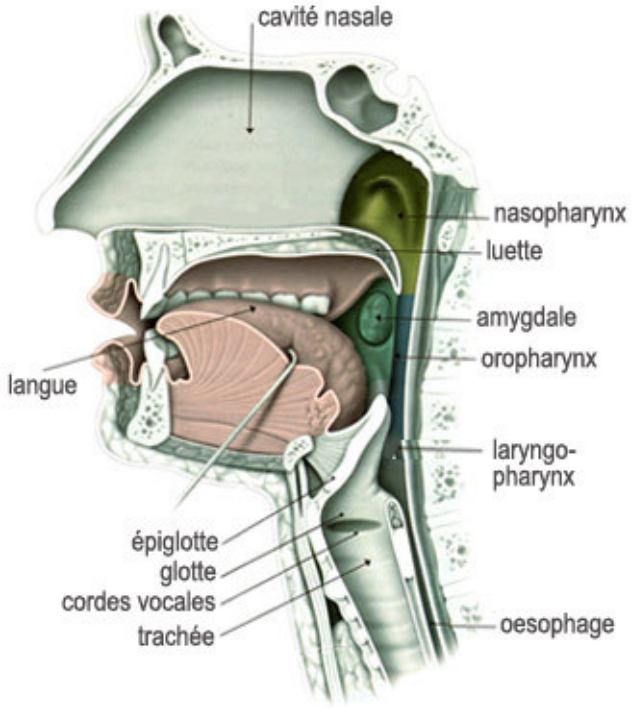


FIGURE 1.1 – Appareil physiologique pour la production de la parole [Kamina2006]

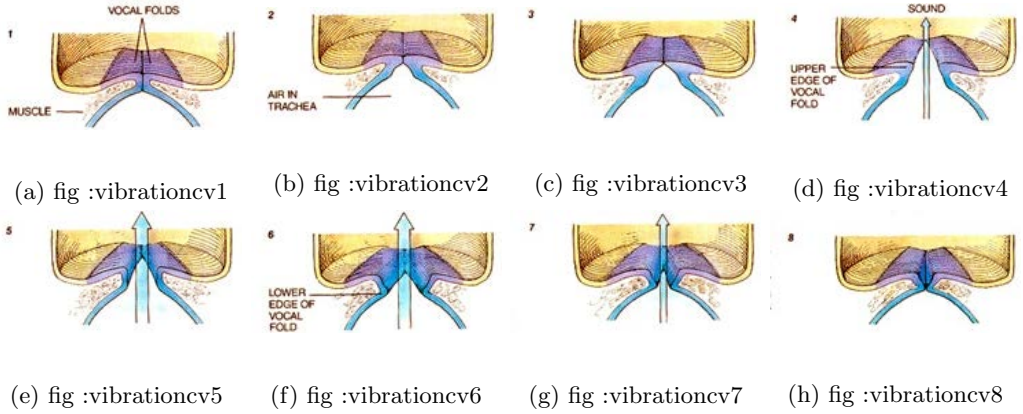
Une caractéristique importante du signal glottique est, pour certains sons, sa nature périodique. La fréquence fondamentale du signal de parole, également nommée F0, correspond à la hauteur du son. Pour obtenir un son de nature périodique le flux d'air va entretenir une vibration à la hauteur du larynx, selon un processus illustré figure 1.2 :

1. Au stade initial, les cordes vocales sont accolées l'une à l'autre bloquant ainsi l'expulsion de l'air provenant des poumons. L'accumulation d'air accentue la pression sub-glottique au contact des cordes vocales (figures 1.2 à 1.2b) ;
2. Lorsque la force de pression se fait trop importante, les cordes vocales s'écartent et l'air s'échappe par le conduit vocal conduisant à une chute de pression (figures 1.2c à 1.2e) ;

3. Cette dépression crée une force d'adduction qui conduit à l'accolement des cordes vocales (figures 1.2f à 1.2g) ;
4. Tant que la pression sub-glottique est suffisante, les étapes 1-3 sont itérées.

La fréquence fondamentale, ou F0, correspond au nombre de cycles d'ouverture/fermeture des cordes vocales effectués par unité de temps. Pour des raisons anatomiques, les plages de F0 qui peuvent être émises diffèrent selon les individus. Pour donner un ordre de grandeurs, [Calliope1989] indique les plages suivantes :

– entre 100 et 150Hz pour une voix masculine,
– entre 140 et 240Hz pour une voix féminine.



(a) fig :vibrationcv1 (b) fig :vibrationcv2 (c) fig :vibrationcv3 (d) fig :vibrationcv4

(e) fig :vibrationcv5 (f) fig :vibrationcv6 (g) fig :vibrationcv7 (h) fig :vibrationcv8

FIGURE 1.2 – Séquence d'illustration de la production d'un signal périodique via la vibration des cordes vocales. (extrait du site web du centre pour la parole de l'université de Pittsburgh, http://www.pitt.edu/~crosen/voice/normcords.html). En (a), les cordes vocales obstruent le passage de l'air vers le conduit vocal. En (b) et (c), la pression de l'air augmente ce qui aboutit (en (d), (e) et (f)), à l'écartement des cordes vocales. Cette ouverture provoque une diminution de la pression ce qui conduit au rapprochement des cordes vocales, en (g), jusqu'au retour à l'état initial (h). La fréquence fondamentale correspond au nombre de répétitions de ce schéma par seconde.

Dans certains cas, le flux d'air transite par le larynx sans faire vibrer les cordes vocales. Les sons obtenus sont alors qualifiés de *non-voisés* en opposition aux sons *voisés* qui ont la caractéristique d'être périodiques.

Le signal glottique situé à la hauteur du larynx est ensuite modulé par le conduit vocal. Comme cela est illustré dans la figure 1.1, le conduit vocal se compose de deux sous-conduits : le conduit nasal et le conduit oral. Ce dernier est le plus complexe et contient des articulateurs qui sont, par ordre d'importance, la langue (articulateur le plus mobile), les lèvres, le voile du palais et les mâchoires.

1.1.2 Le signal de parole

La parole est considérée comme un signal acoustique réel, périodique ou aléatoire, continu, d'énergie finie [Calliope1989, Boite2000]. D'après la théorie de Fourier,

Tout signal périodique, voire apériodique, d'énergie finie peut se décomposer sur la base d'une série de composantes sinusoïdales.

Pour décrire une sinusoïde, il est nécessaire de caractériser trois paramètres :

- la *fréquence*, en Hertz Hz, qui correspond au nombre de répétitions d'une période élémentaire,
- la *phase*, en radian, qui identifie le décalage de la sinusoïde à l'instant $t = 0$,
- l'*amplitude* qui peut être reliée à l'énergie apportée par le signal. L'énergie s'exprime le plus souvent sur une échelle logarithmique (décibels, dB).

Ainsi, en supposant un signal numérique $s(n)$ comportant N échantillons, la transformée de Fourier discrète permet de passer du domaine temporel au domaine fréquentiel par la relation suivante :

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} s(n) e^{-j \frac{2\pi}{N} kn} \tag{1.1}$$

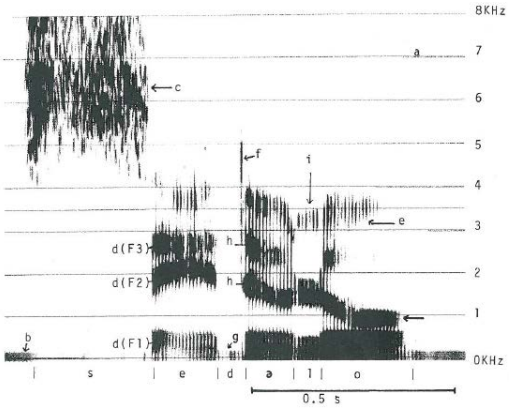
où $|X(k)|$, respectivement $arg(X(k))$, correspond à l'amplitude et la phase de la sinusoïde de fréquence $2\pi k/N$. Pour déterminer une DFT, l'algorithme le plus couramment utilisé est la transformée de Fourier rapide (ou FFT pour Fast Fourier Transform).

Un signal de parole est un signal quasi-stationnaire car ses propriétés fréquentielles évoluent au cours du temps. Toutefois, certaines propriétés peuvent être observées localement. Il est nécessaire de trouver un compromis entre résolution spectrale et résolution temporelle. La transformée de Fourier à court terme (STFT, Short Time Fourier Transform), décrite par la relation (1.2), est souvent utilisée en traitement de la parole pour déterminer l'évolution temporelle du spectre fréquentiel du signal acoustique :

$$X_m(k) = \sum_{n=0}^{N-1} s(n) w(n-m) e^{-\frac{knj2\pi}{N}} \tag{1.2}$$

où w correspond à la fonction de fenêtre. La fonction de fenêtre, communément appelé fenêtre, permet de pouvoir réduire les distorsions spectrales dues au découpage du signal en bloc. Plusieurs types de fenêtres existent et chacune possède des propriétés qui lui sont propres [harris1978]. En traitement automatique de la parole, les fenêtres de Hamming, de Hanning ou de Blackman sont généralement utilisées.

En se basant sur le spectre d'amplitude, il est possible de visualiser le signal de parole comme l'illustre la figure 1.3. Ainsi, pour représenter le spectre, en plus des axes temps/fréquences, un spectrogramme intègre une troisième dimension pour visualiser l'amplitude de chaque de sinusoïde. Cette dimension est représentée par le niveau de gris d'un point : plus le point est noir plus l'amplitude est élevée. De plus, [Calliope1989] indique que l'oreille est peu sensible à la phase. Ainsi dans la suite du document lorsque la notion de spectre sera évoquée, il s'agira implicitement du spectre d'amplitude.



a. ligne de calibration ;
b. *ligne de base* produite par l'appareil en l'absence de signal ;
c. bruit de friction du /s/ ;
d. formants : la mesure se fait au centre de la bande noire ;
e. modulation de l'énergie à la fréquence fondamentale ;
f. barre d'explosion de l'occlusive ;
g. barre de voisement ;
h. transitions formantiques ;
i. formant faible de la sonante /l/.

FIGURE 1.3 – Spectrogramme de la phrase "C'est de l'eau". Cette figure est extraite de [Calliope1989].

Le spectrogramme est un outil fondamental en traitement de la parole car il permet de visualiser les propriétés à court terme du signal de parole. Ainsi, comme l'illustre la figure 1.3, un son voisé, comme le segment étiqueté /o/ par exemple, possède une structure régulière contrairement à un son non voisé, comme le segment étiqueté /s/, dont la structure fréquentielle est plus aléatoire. Ces propriétés spectrales locales, sont généralement observées sur des longueurs de l'ordre de 20/30ms[Boite2000].

Jusqu'à présent, nous avons supposé que le signal de parole était numérisé et donc discrétisé. Toutefois, comme nous l'avons indiqué, un signal de parole est un signal continu. Pour obtenir la représentation numérique de ce signal, deux modes existent[Boite2000] : une représentation directe de la forme d'onde ou bien une représentation paramétrique. La représentation directe de la forme d'onde consiste à ne poser aucune hypothèse sur le signal de parole. Sous sa forme la plus simple, comme celle utilisé par le système MIC [1], le codage direct vise à représenter chaque échantillon du signal indépendemment de tous les autres. Par opposition, une représentation paramétrique suppose une modélisation définie *a priori* et permettant d'obtenir une description d'un signal de parole moins coûteuse que la représentation directe. La plupart des modèles paramétriques font l'hypothèse d'un modélisation source/filtre proposée par G. Fant [Fant70].

1.1.3 Modélisation source/filtre

Le modèle source/filtre, illustré figure 1.4, considère le signal de parole $s(n)$ comme le résultat de la convolution du signal glottique $e(n)$ (la source) par un filtre $h(n)$ qui représente le comportement fréquentiel du conduit vocal soit :

$$s(n) = e(n) * h(n) \tag{1.3}$$

1. Modulation par Impulsions Codées, plus connu sous le nom de PCM

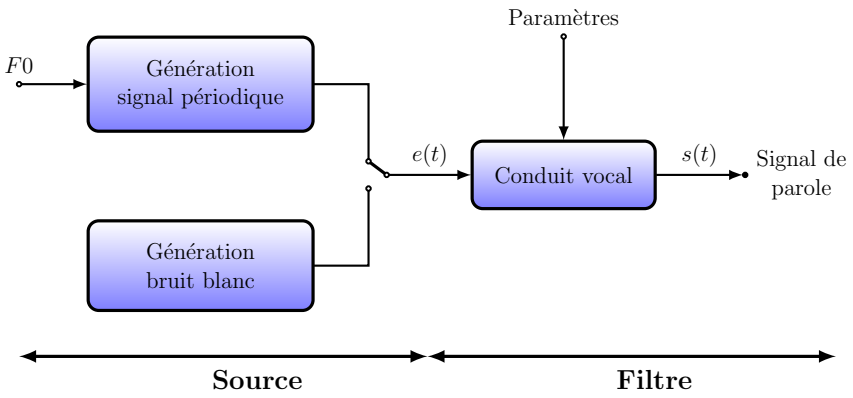


FIGURE 1.4 – Modélisation Source/Filtre de la parole telle que proposée par G. Fant [Fant70]

Dans sa représentation la plus simple, ce modèle repose sur deux contraintes fortes [Rothenberg2008] :

– le filtre est posé comme un système linéaire ;

– le filtre et la source sont indépendants.

Lors de la phase d'analyse du signal de parole, seul $s(n)$ est connu. Représenter un signal selon une modélisation source/filtre consiste donc à résoudre une équation à deux inconnues ce qui implique de faire des hypothèses simplificatrices sur la source ou sur le filtre. Généralement, ces hypothèses sont appliquées au modèle de la source qui est réduit à deux états possibles : la source correspond à un signal périodique si le son est voisé, ou la source correspond à un signal bruité si le son est non-voisé. Le filtre correspond ici à un spectre qui, par convolution, va permettre d'amplifier certaines fréquences du signal issues de la source.

Afin de simplifier l'opération de déconvolution et ainsi de déterminer plus aisément la contribution de la source et du filtre, cette opération est effectuée dans l'espace dit *cepstral* [Calliope1989, Boite2000]. Cet espace, appelé domaine quéfrentiel, est obtenu en effectuant une transformée de Fourier inverse sur le logarithme du spectre ce qui permet de substituer l'opérateur de convolution à un opérateur d'addition. La relation suivante est alors obtenue :

$$\tilde{c}(n) = \tilde{e}(n) + \tilde{h}(n) \tag{1.4}$$

où $\tilde{c}(n)$ correspond aux coefficients dits cepstraux et $\tilde{e}(n)$, respectivement $\tilde{h}(n)$, correspond à $e(n)$, respectivement $h(n)$, dans le domaine quéfrentiel.

En se basant sur ces hypothèses, plusieurs méthodes permettent de paramétrer le signal de parole. Parmi ces méthodes, les vocodeurs (ou vocoder pour VOice CODER) à canaux, dont le principe a été proposé par H. Dudley [Dudley1939] en 1939, repose sur l'utilisation de bancs de filtres associés à une plage de fréquences déterminées. En excitant chaque filtre par un signal source (périodique ou bruit blanc), et en combinant la sortie de chaque filtre, le signal de parole est reproduit. Bien que la méthodologie soit ancienne, de

nouveaux modèles, comme STRAIGHT, dérivant de ce principe ont récemment été mis au point et permettent d'obtenir un signal de bonne qualité.

Modèle STRAIGHT

Parmi les systèmes qui découlent de la modélisation source/filtre, le modèle STRAIGHT [2] [Kawahara1999] est devenu le modèle de référence pour les systèmes HTS.

En supposant que la fréquence fondamentale (F0) soit définie, le modèle STRAIGHT considère que le signal associé à un segment voisé est représenté comme la somme de K harmoniques [3], considérés comme les canaux du vocodeur, comme suit :

$$s(t) = \sum_{k=1}^{K} \alpha_k(t) sin\Big[\int_{t_0}^{t} k(\omega(\tau) + \omega_k(\tau))d\tau + \phi_k \Big] \tag{1.5}$$

où $t_0 = 1/F0$ et $\omega(\tau)$ correspond à une pulsation globale. ϕ_k, $\alpha_k(t)$, $\omega_k(\tau)$ correspondent respectivement à la phase, l'amplitude et la pulsation associée à la k-ème harmonique.

Pour obtenir le spectre fréquentiel, [Kawahara1999] propose une méthode utilisant la reconstruction de surfaces basée sur des informations partielles. Ces informations sont extraites du signal grâce à l'utilisation de fenêtres adaptatives, tant dans le domaine temporel que dans le domaine fréquentiel et des B-splines sont alors utilisées pour effectuer la reconstruction. L'objectif est d'obtenir une enveloppe spectrale [4] dépourvue d'information due à la périodicité.

Le signal de parole n'étant pas exclusivement périodique ou bruité, le modèle STRAIGHT introduit un troisième paramètre : l'apériodicité qui correspond à l'énergie associée aux fréquences non-harmoniques [Kawahara2001]. En supposant l'enveloppe spectrale supérieure $|S_U(\omega)|^2$ (représentant les composantes périodiques du signal) et $|S_L(\omega)|^2$ l'enveloppe spectrale inférieure (représentant les composantes de bruit), l'apériodicité est définie comme la normalisation de $|S_L(\omega)|^2$ par $|S_U(\omega)|^2$ [Kawahara2001].

Lors de la phase de synthèse, STRAIGHT utilise l'apériodicité pour pondérer l'apport du signal périodique et l'apport du signal bruité avant l'application du filtre. Cette méthode de synthèse, dont le principe est illustré figure 1.5, est appelée *Mixed-mode excitation* [Kawahara2001].

Pour générer un signal de parole en utilisant STRAIGHT, il est donc nécessaire de fournir trois données : la fréquence fondamentale, les coefficients d'apériodicité et les coefficients spectraux. La dimension des vecteurs de coefficients d'apériodicité et des coefficients spectraux étant élevée (généralement supérieure à 512), une opération supplémentaire

2. Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
3. Une harmonique correspond à une fréquence multiple de la fréquence fondamentale
4. D'après la définition donnée par [Robel05], une enveloppe spectrale est une fonction continue et *lisse* qui passe par les pics spectraux

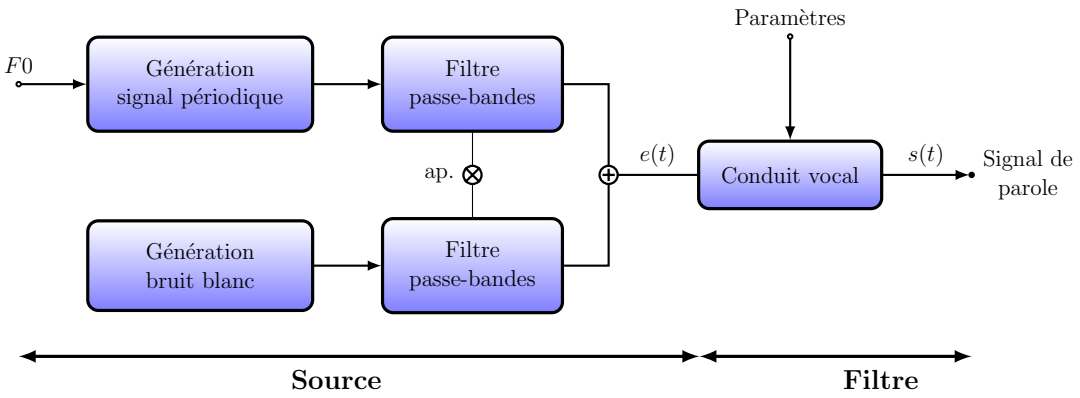


FIGURE 1.5 – Principe du vocodeur STRAIGHT. Le signal glottique est considéré comme une combinaison d'un signal périodique et d'un signal aléatoire. L'apériodicité permet de pondérer cette combinaison. Cette figure est basée sur celle présentée dans [Yoshimura2001]

est réalisée pour réduire cette dimension. Pour les coefficients d'apériodicités, l'espace des fréquences est généralement découpé en bandes et un coefficient moyen est affecté à chaque bande. Pour le spectre, un filtre MLSA (Mel-Log Spectrum Approximation) est généralement utilisé pour obtenir une approximation du spectre basée sur un ensemble réduit de coefficients : les coefficients MGC (Mel Generalized Cepstral Coefficient).

Filtre MLSA

Pour modéliser le spectre via les coefficients MGC, [Fukada1992] propose la relation suivante :

$$H(\omega) = \exp \sum_{m=0}^{M} c(m) e^{-j\tilde{\omega}m} \tag{1.6}$$

où $c(m)$ correspond aux coefficients MGC les coefficients mel-cepstraux d'ordre M. $\tilde{\omega}$ correspond à la phase de la fonction de transfert et est caractérisée en se basant sur un coefficient α dont la valeur est déterminée relativement à la fréquence d'échantillonage. Dans [Imai1983], ces valeurs ont été déterminées subjectivement et sont présentées dans le tableau 1.1. De plus, le coefficient $c(0)$ correspond au gain du filtre.

Fréq. d'éch. (kHz)	α
≤ 8	0.31
≤ 10	0.35
≤ 12	0.37
≤ 16	0.42
≤ 22.05	0.45

TABLE 1.1 – Valeur du coefficient α en fonction de la fréquence d'échantillonnage

Pour déterminer les coefficients $\tilde{c}(m)$, [Fukada1992] propose une méthode itérative basée sur un algorithme de gradient (plus précisément l'algorithme de Newton-Raphson)

et dont le critère à minimiser, proposé dans [Imai1988], est le suivant :

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{exp(R(\omega)) - R(\omega) - 1\} d\omega \tag{1.7}$$

où

$$R(\omega) = log(X(\omega)) - log(|H(\omega)|^2) \tag{1.8}$$

$R(\omega)$ identifie l'erreur entre le spectre modèle $H(\omega)$, obtenu à partir des coefficients MGC et décrit par l'équation (1.6), et le spectre d'amplitude $X(\omega)$, extrait du signal $x(n)$ sur une échelle logarithmique.

1.1.4 Perception de la parole

Après avoir exposé les mécanismes de production de la parole puis sa modélisation en utilisant le modèle source/filtre, nous allons maintenant présenter la perception de ce phénomène qui s'effectue sur deux niveaux : le niveau segmental et le niveau suprasegmental également appelé niveau prosodique.

Description segmentale de la parole

Tout d'abord, il est important de noter que l'oreille ne perçoit pas les sons sous une échelle linéaire mais logarithmique. Plusieurs échelles de perception ont été proposées pour se rapprocher de l'analyse effectuée par l'oreille. L'une des plus courantes est l'échelle Mel[Stevens1937]. Le coefficient α du modèle STRAIGHT, présenté section précédente, permet d'approcher cette échelle.

Comme nous l'avons vu section 1.1.2 de ce chapitre, la parole est perçue, au niveau segmentale, comme une succession de sons élémentaires, les phones. Ces phones peuvent être classés en trois catégories principales :

- les voyelles qui sont des sons vibrants. Elles peuvent être classées selon leur nasalité/oralité, leur degré d'ouverture du conduit vocal, la position de la constriction principale ainsi que l'aperture des lèvres[Calliope1989],
- les consonnes qui peuvent êtres des sons vibrants ou non. Elles peuvent être classées selon le voisement, le mode d'articulation ainsi que la position de la constriction principale (lieu d'articulation),
- les semi-voyelles qui correspondent à des sons intermédiaires.

Le tableau 1.2 présente les différents phones pour le français classés en fonction de leurs catégories principales et des différentes caractéristiques articulatoires liées à ces catégories.

Consonnes				**Voyelles**				
Lieux / Modes	*Labiales*	*Dentales*	*Vélo-pal.*		Ant.		Post.	
Occlusives				Orales	NA	Arr.		NA
- non-voisées	p	t	k	- Fermés	i	y	u	
- voisées	b	d	g		e	ø	o	
					ɛ	œ	ɔ	
				- Ouvertes	a			
Nasales	m	n	ɲ	Nasales	Ant.		Post.	
				- Fermés	ɛ̃		õ	
				- Ouvertes		ɑ̃		
Fricatives								
- non-voisées	f	s	ʃ					
- voisées	v	z	ʒ					
Glissantes	w	ɥ	j					
Liquides		l	ʀ					

TABLE 1.2 – Classification des phonèmes du français [Calliope1989]

La prosodie

Une définition de la prosodie est donnée par DiCristo [Dicristo2000] :

> **La prosodie** (ou la prosodologie) est une branche de la linguistique consacrée à la description (aspect phonétique) et à la représentation formelle (aspect phonologique) des éléments de l'expression orale tels que les accents, les tons, l'intonation et la quantité, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F0), de la durée et de l'intensité (paramètres prosodique physiques), ces variations étant perçues par l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation pragmatique dans le flux du discours.

En se basant sur cette définition, nous pouvons distinguer trois paramètres principaux de la prosodie corrélés avec une représentation acoustique du signal de parole : la fréquence fondamentale et la durée qui permettent de décrire la mélodie de l'énoncé ; l'intensité qui correspond à la perception de la force sonore de la voix. En se basant sur ces paramètres, l'étude de la prosodie consiste à analyser différents phénomènes tels l'accentuation, l'intonation ou le débit. Il en existe d'autres mais nous allons nous focaliser sur ces trois phénomènes car ils sont souvent utilisés en synthèse de la parole à partir du texte.

Tout d'abord, dans le cadre du français, l'unité de l'accentuation est la syllabe. Ainsi, [Mertens1992] définit une syllabe accentuée, ou syllabe proéminente, comme une syllabe qui ressort sur son entourage par sa force particulière, par un contraste d'intensité subjective. Le français est considéré comme une langue à accent fixe ce qui implique que l'accent

est appliqué à la syllabe en fonction de sa position dans le mot.

L'intonation correspond à la courbe mélodique de l'énoncé. Elle est donc basée sur deux paramètres prosodiques : la durée et le F0. Pour le français, l'unité du groupe intonatif est une séquence de syllabes ne comportant qu'un unique accent final [Mertens1993, Mertens2001]. P. Delattre [Delattre1966] a mis en évidence dix cas d'intonation de base et montre également qu'une substitution d'intonation, sur un même contenu, implique un changement de sens (comme par exemple la perception d'un énoncé comme une interrogation ou bien comme une affirmation). La figure 1.6 présente les courbes des dix intonations identifiées par P. Delattre.

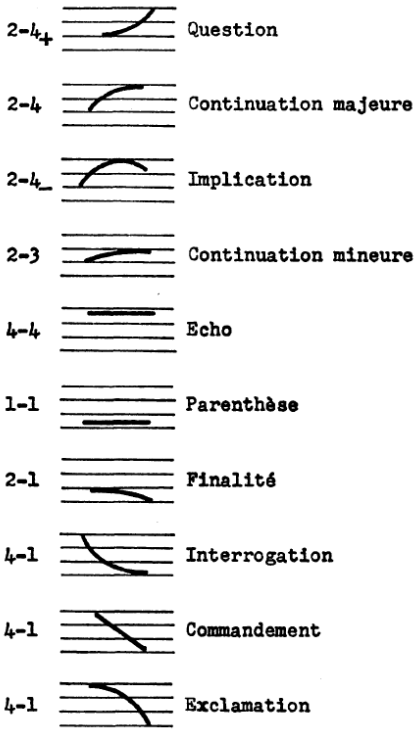


FIGURE 1.6 – Les dix intonations identifiées par Delattre [Delattre1966]

Enfin le débit qui correspond à la vitesse d'élocution et qui s'exprime en unité de parole par temps. Dans le cadre du français, le débit syllabique, nombre de syllabes par seconde, est généralement utilisé. Toutefois, B. Zellner [Zellner1998] a montré que cet élément est en réalité influencé par plusieurs facteurs : les pauses, l'allongement/raccourcissement des syllabes ou bien encore l'ajout de sons dans le cadre de l'hyper-articulation. Dans son étude B. Zellner [Zellner1998] note que le débit syllabique du français se situe entre 4 (débit lent) et 7 (débit rapide) syllabes par seconde.

Pour conclure concernant la prosodie, I. Fónagy [Fonagy2003] a effectué une synthèse des différentes fonctions de l'intonation. Parmi les fonctions principales, la prosodie permet la distinction entre homonyme (cas de l'anglais) ou bien de distinguer les frontières de mots (cas du français). La prosodie a donc une fonction de structuration de l'énoncé

importante. Ensuite, la prosodie a une fonction de focalisation qui permet de nuancer le sens d'un même énoncé en faisant ressortir l'entité importante. Au delà de ces fonctions linguistiques, la prosodie renseigne également sur l'état émotif du locuteur, son attitude, son niveau social ou culturel. Il s'agit donc d'une composante fondamentale en parole.

1.2 Synthèse de la parole à partir du texte

Les travaux présentés dans ce document se situent dans le cadre de la synthèse de la parole à partir du texte. À l'heure actuelle, deux méthodes dominent ce domaine : la synthèse par corpus, qui consiste à utiliser des segments [5] audios issus d'un corpus de parole enregistré par un locuteur, et la synthèse paramétrique, qui consiste à modéliser le signal par des coefficients acoustiques puis à générer la forme d'onde du signal de parole en utilisant ces modèles.

Cette section débute par la présentation de l'architecture commune à tout système de synthèse TTS. Les deux méthodes précédemment évoquées sont ensuite décrites. Cette présentation se base sur le livre *Text-To-Speech synthesis* de P. Taylor [Taylor2009].

1.2.1 Principe d'un système TTS

La synthèse de la parole à partir du texte a pour objectif de produire un signal de parole correspondant à un texte donné. Pour cela, il est nécessaire de distinguer deux étapes comme l'illustre la figure 1.7 :

1. Les traitements linguistiques et prosodiques qui permettent d'obtenir un ensemble de descripteurs qualifiant le texte à synthétiser ;
2. Les traitements acoustiques qui, à partir de la séquence des descripteurs obtenue à l'étape précédente, permet de générer le signal correspondant au texte. C'est lors de cette étape que les deux méthodes précédemment évoquées, synthèse par sélection et synthèse paramétrique, vont être utilisées.

L'étape des traitements linguistiques consiste à enrichir le texte de descripteurs visant à qualifier ce texte selon plusieurs échelles de description. Tout d'abord, en guise de prétraitement, il est nécessaire de développer les acronymes, abréviations et de réécrire les nombres afin d'unifier la représentation dite de surface.

Une fois ces éléments du texte explicités, une analyse morpho-lexicale est appliquée pour identifier les mots et leur associer une catégorie grammaticale (Part-Of-Speech). Cette étape permet, par exemple, de lever de nombreuses ambiguïtés telles que par exemple la distinction entre le verbe *couvent* issu de *couver* et le nom *couvent* de la phrase suivante : « *Les poules du couvent couvent* ». L'analyse syntaxique permet, en s'appuyant

5. Un segment est une partie du signal identifiée par un instant de début et un instant de fin

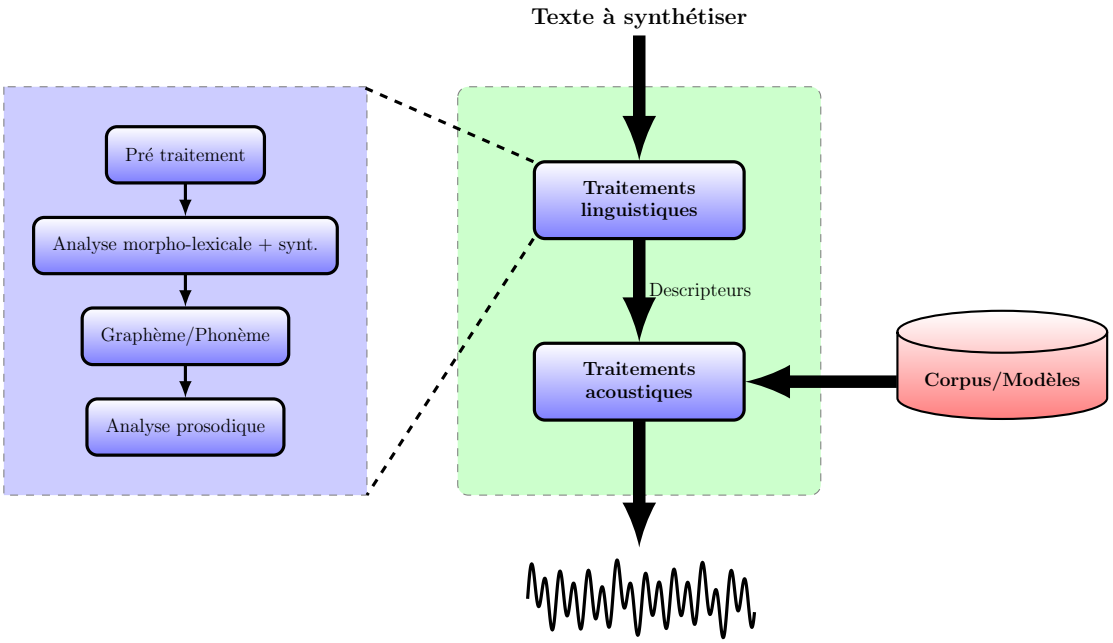


FIGURE 1.7 – Illustration du processus de synthèse d'un système TTS

sur les étiquettes précédemment obtenues, de structurer l'énoncé sous la forme d'un arbre où chaque noeud correspond à une subdivision de cet énoncé ou syntagme. Les feuilles correspondent alors aux mots.

L'étape de transcription phonétique, dénommée *Graphème/Phonème* sur la figure 1.7, se base sur les annotations obtenues aux étapes précédentes pour définir la séquence de phonèmes associée à l'énoncé que l'on souhaite synthétiser. Ainsi, le système de phonétisation doit pouvoir distinguer les deux mots *couvent* et ainsi produire la séquence de phonèmes /kuvãkuv/ et non /kuvãkuvã/.

Enfin, la dernière étape consiste à obtenir une représentation symbolique des annotations prosodiques de l'énoncé à synthétiser. Ces annotations, qui sont généralement sur l'accentuation, la courbe intonative et le schéma rythmique de l'énoncé, visent à fournir au module de traitements acoustiques des informations permettant de sélectionner des unités acoustiques ou de produire le signal de parole par un modèle paramétrique.

1.2.2 Synthèse par corpus

L'apparition de la synthèse par corpus s'est effectuée en deux temps. Tout d'abord, dans les années 1970, les systèmes par concaténation d'unités ont été conçus. Cette méthodologie repose sur une base contenant une unique occurrence associée à chaque unité de parole. La phase de synthèse consiste à concaténer ces occurrences sélectionnées en fonction de la consigne obtenue par l'étape de traitement linguistique. Un algorithme de traitement de signal est alors appliqué afin de plaquer une consigne prosodique. Généra-

lement, l'algorithme TD-PSOLA [Charpentier1989] est utilisé pour effectuer ce traitement et ainsi obtenir le signal de parole synthétisé. De plus, afin de réduire les artefacts au point de concaténation, l'unité utilisé dans ce type de système est, généralement, le diphone [6]. En effet, cette unité présente l'avantage de pouvoir effectuer une concaténation sur une zone stable : la zone centrale du phone qui est la moins affectée par les transitions.

A la fin des années 80, la prise en compte de plusieurs occurrences d'unités de taille variable a été introduite par [Sagisaka1988]. Cette méthode de synthèse, appelée synthèse par sélection d'unités ou synthèse par corpus, repose sur un corpus de parole annoté d'une durée de plusieurs heures. Au milieu des années 1990, l'article de Hunt et Black [Hunt1996], dans le cadre du système CHATR [Black1994], formalise le problème de la sélection d'unités comme la résolution d'un problème d'optimisation d'une fonction composée des coûts suivants :

- un coût cible qui permet d'évaluer la proximité d'une unité, dite candidate, par rapport à la consigne issue des traitements linguistiques et prosodiques du système TTS,
- un coût de concaténation qui permet d'estimer la distorsion obtenue au point de jonction de deux unités candidates.

Cette fonction de coût est une équation récurrente d'ordre 1 qui peut être résolu grâce au paradigme de la programmation dynamique. L'algorithme généralement utilisé est l'algorithme de Viterbi [Viterbi1967]. La dernière étape consiste alors, comme pour les systèmes par concaténation, à plaquer la consigne prosodique pour obtenir le signal de parole.

Par construction, la synthèse par corpus consiste donc à sélectionner les unités les plus longues possibles. Néanmoins, il est nécessaire de disposer d'unités élémentaires dont un nombre d'occurrences minimal est garanti. Plusieurs types d'unités ont donc été proposés allant de la trame [Hirai2004] à la phrase [Donovan1999] et différentes études, comme celle proposée dans [Kishore2003] ont été réalisées afin de comparer ces unités. Toutefois, comme pour la synthèse par concaténation, le diphone reste l'unité de prédilection.

Le principe de la synthèse par corpus est résumé par la figure 1.8. Sur cette figure nous distinguons les deux phases importantes de la synthèse par sélection :

1. En se basant sur une consigne de synthèse et un corpus de parole annoté, un graphe d'unités candidates est alors constitué (A),
2. les unités sont sélectionnées en appliquant l'algorithme de Viterbi et finalement concaténées pour obtenir la synthèse (B).

1.2.3 Synthèse paramétrique par HMM

Contrairement à la synthèse par corpus, la synthèse paramétrique repose sur l'utilisation des techniques de traitement de signal pour obtenir une représentation paramétrique

6. Un diphone est une unité qui s'étend sur deux phones consécutifs allant du milieu du premier au milieu du second.

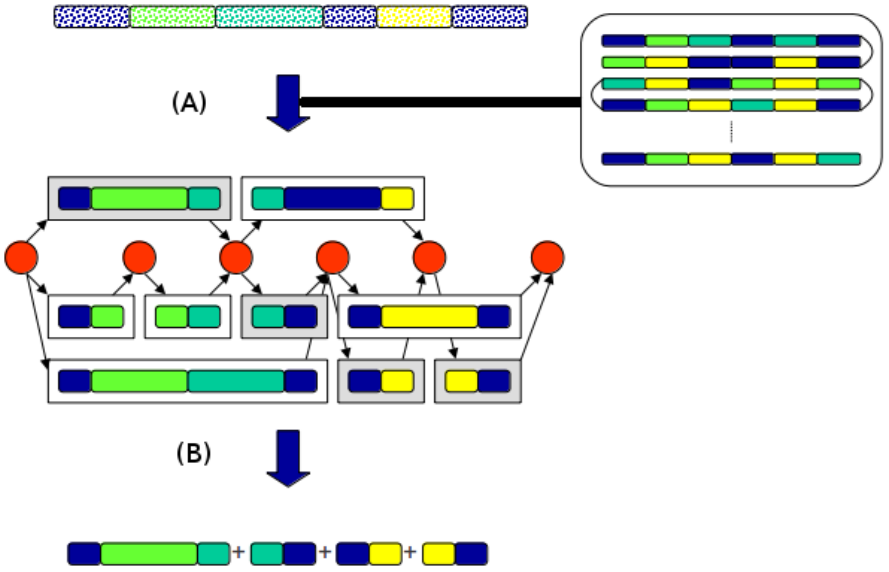


FIGURE 1.8 – Principe de la synthèse par corpus. Dans un premier, en se basant sur une consigne définie et un corpus de parole annoté, un graphe d'unités candidates est constitué (A). En utilisant un algorithme de Viterbi, les unités optimales (au sens des coûts définis dans le système de synthèse) sont sélectionnées puis concaténées (B).

du signal de parole. Les premiers systèmes de synthèse paramétrique modélisaient la parole selon un ensemble de règles. Chaque règle permettait, à partir de consignes phonético-prosodiques (séquence de phonèmes, la durée de chaque de phone à produire et la consigne mélodique), de déterminer les trajectoires des paramètres du modèle de réprésenation du signal. Ces systèmes sont également appelés systèmes de synthèse par formants car les règles utilisées permettaient généralement de modéliser l'évolution des formants. Parmi ces systèmes, le plus connu reste sans doute l'OVE (Orator Verbis Electris) de G. Fant [Fant70]. Grâce à l'évolution des technologies, le traitement de données massives a été rendu possible.

De nos jours, les systèmes actuels modélisent l'évolution des paramètres acoustiques par des modèles stochastiques. Parmi ces modèles, l'utilisation du HMM dans le cadre de la synthèse TTS a été proposée au milieu des années 1990 par R. Donovan [Donovan1996, Donovan1995] et K. Tokuda [Tokuda1995].

Le système présenté par R. Donovan dans [Donovan1996] se base sur HTK [Young1993, Young2005] (the HMM ToolKit), qui propose un ensemble d'outils pour l'utilisation des HMM dans le cadre de la reconnaissance de la parole. R. Donovan [Donovan1995] a adapté ces outils pour réaliser différentes expériences basées sur une paramétrisation du signal de parole en coefficient MFCC. Ces coefficients sont utilisés pour apprendre des HMM modélisant des phones en contexte. Un arbre de décision [Young1994] est ensuite construit

pour que, lors de la phase de synthèse, les phones en contexte, que l'on souhaite synthétiser et qui ne sont pas présents dans le corpus d'apprentissage, aient un modèle qui leur soit associé. Lors de la phase de synthèse, ces modèles permettent de prédire une séquence de coefficients LPC ainsi qu'une consigne de voisement qui leur est associée.

Le système présenté par K. Tokuda ET AL.dans [Tokuda1995a] repose sur le même paradigme que le système précédent. Toutefois l'apport de ce système est la prise en compte de la dynamique de premier ordre lors de la génération des coefficients acoustiques dont T. Masuko ET AL.[Masuko1996] ont montré l'impact positif sur la qualité des coefficients issus de la génération.

Ainsi, à l'heure actuelle, le système référent pour la synthèse HMM découle des travaux de K. Tokuda ET AL. [Tokuda1995a] et s'intitule HTS (pour HMM Speech Synthesis System, qui a été réduit en HMM Triple S puis en HTS). Ce système, développé par le laboratoire Nitech, se décline selon deux modes :

- La modélisation dite *dépendante du locuteur* [Zen2005, Zen2006] qui consiste à apprendre des modèles à partir d'un corpus dicté par un locuteur pour effectuer une synthèse dont les caractéristiques du signal obtenu seront propres à ce locuteur ;
- La modélisation dite *indépendante du locuteur* [Yamagishi2007a, Yamagishi2008] qui consiste à apprendre des modèles moyens à partir d'un corpus composé de multiples locuteurs. Ces modèles sont ensuite adaptés au locuteur cible en utilisant un corpus extrêmement réduit (selon [Yamagishi2008a] environ 6 minutes de parole suffisent pour effectuer l'adaptation)

Dans la suite du document, nous ne tiendrons compte que de la modélisation dépendante du locuteur dont l'architecture est illustrée par la figure 1.9. Néanmoins, le lecteur pourra se référer aux articles [Yamagishi2008, Yamagishi2005, Yamagishi2007] pour plus de détails concernant la synthèse HTS indépendante du locuteur.

Le système HTS repose sur une modélisation source/filtre, telle que nous l'avons vue à la section 1.1.3 de ce chapitre, pour représenter le signal de parole. Ainsi, pour effectuer un apprentissage, le système HTS utilise un corpus de parole annoté dont le signal est paramétré pour obtenir les coefficients suivants :

- La fréquence fondamentale ;
- Les coefficients MGC [Fukada1992] qui représentent le filtre ;
- Les coefficients d'apériodicité si le vocodeur STRAIGHT est utilisé pour extraire le F0 et obtenir le spectre.

À ces coefficients, qualifiés de statiques, s'ajoutent leurs dérivés de premier et second ordre.

En plus de ces paramètres, chaque segment est qualifié en utilisant un jeu de descripteurs, spécifique à une langue, qui permettent de prendre en compte le contexte de chacun de ces segments. Lors de la phase de synthèse, ce sont ces descripteurs, déterminés lors

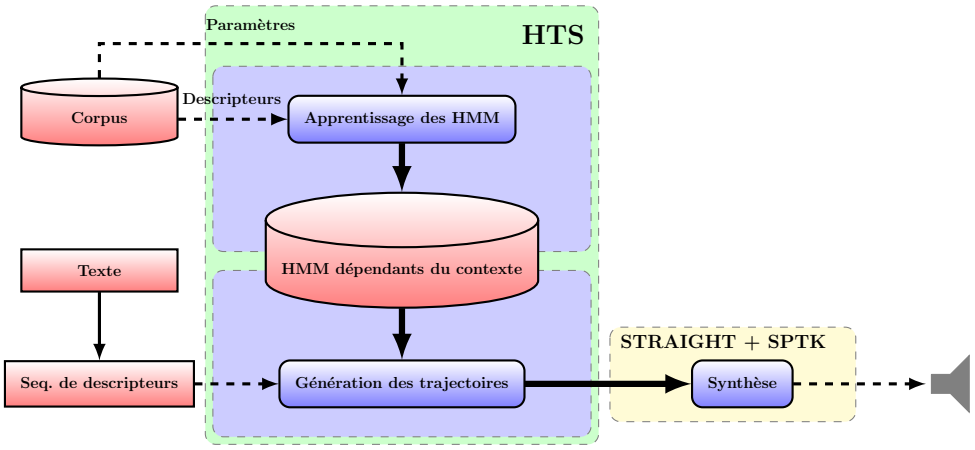


FIGURE 1.9 – Le système HTS : le corpus est constitué du signal paramétrisé (par les outils STRAIGHT [Kawahara1999, Kawahara2001] et SPTK [Fukada1992]) ainsi que les descripteurs permettant d'identifier un segment. En se basant sur ce corpus, les HMM sont appris. Lors de la phase de synthèse, la séquence de paramètres, correspondant à la séquence de descripteurs déterminés par les outils d'analyse linguistique, est générée. Les outils STRAIGHT et SPTK permettent d'obtenir le signal acoustique à partir des paramètres générés par HTS.

de la phase de traitements linguistiques, qui vont permettre de sélectionner les modèles adéquats.

Le système HTS utilise un ensemble de concepts et d'algorithmes issus du domaine de la reconnaissance de la parole pour apprendre les modèles. Parmi les plus importants, nous pouvons citer l'utilisation de modèles semi-Markoviens (HSMM) [Zen2005, Russell1985] qui permet de représenter la durée de séjour dans un état par une distribution gaussienne. Comme pour le système présenté par Donovan [Donovan1995], des arbres de décisions [Young1994] ont été utilisés afin de garantir la présence d'un modèle lors de la phase de synthèse.

D'autres concepts ont été introduits afin de répondre à des besoins spécifiques. Ainsi, afin de prendre en compte l'état voisé et l'état non-voisé d'une trame, les MSD (Multi-Space Distribution) ont été introduits. L'objectif des MSD [Tokuda1999, Tokuda2000a] est de proposer une représentation unique par des distributions associées aux valeurs de F0.

Bien que le système HTS utilise de nombreux concepts adaptés à la synthèse de parole, la paramétrisation implique une perte de qualité du signal de parole. Le timbre généré souffre d'un effet de bourdonnement et les différentes méthodes, disponibles à l'heure actuelle, pour paramétriser le signal ne permettent pas de résoudre complètement ce problème. Ainsi, lors des challenges Blizzard [Black2005, King2012] qui permettent d'évaluer les systèmes de synthèse à l'état de l'art, les systèmes de synthèse par corpus obtiennent globalement de meilleurs scores. Néanmoins, la paramétrisation permet de manipuler le signal plus simplement ce qui rend la synthèse paramétrique plus flexible que

les systèmes par corpus. Ainsi, à l'heure actuelle, la synthèse par HMM fait l'objet d'une attention particulière dans le domaine de la synthèse de la parole.

1.3 Conclusion

Dans ce chapitre, nous avons présenté brièvement le phénomène de la parole. Nous avons pu ainsi voir qu'il s'agit d'un phénomène complexe qui repose sur de nombreux mécanismes physiologique et cognitif. En présentant le modèle source/filtre, nous avons pu introduire les principales techniques de traitement de signal utilisées pour décrire le signal de parole. Nous avons mis en avant deux modèles, STRAIGHT et les coefficients MGC, qui permettent de paramétriser ce signal.

Dans un second temps, nous avons introduit les deux principales méthodes de synthèse TTS courantes à ce jour : la synthèse par corpus et la synthèse paramétrique basée sur des HMM. À l'issue de cette introduction, nous avons indiqué que le système HTS est un des systèmes de synthèse référent. Ce système étant notre objet d'étude, nous allons le détailler dans le chapitre suivant.

Chapitre 2

Système HTS - Présentation

Dans le chapitre précédent, nous avons présenté brièvement le domaine de la synthèse de la parole. À l'issue de ce chapitre, nous avons indiqué que deux méthodes dominent la synthèse TTS à l'heure actuelle, les systèmes par sélection d'unités et les systèmes paramétriques statistiques. Nos travaux se focalisent sur la synthèse paramétrique par HMM et plus précisément sur le système HTS.

Tout d'abord, le système HTS utilise les concepts issus de la suite logicielle HTK. Nous allons donc débuter ce chapitre par l'introduction des concepts proposés par cette

suite pour modéliser le signal de parole et expliciter les problèmes que posent l'utilisation de HTK pour effectuer une génération. Ensuite, nous présenterons les algorithmes de génération produisant la séquence de coefficients nécessaire au couple d'outils STRAIGHT [Kawahara1999] et SPTK [Fukada1992] pour produire un signal de parole. Dans un troisième temps, nous introduirons les modifications apportées aux modèles afin de les adapter à la synthèse de la parole. Nous détaillerons alors les processus d'apprentissage et de synthèse proposés par HTS pour, respectivement, obtenir les modèles et générer les coefficients à partir de ces modèles. La dernière section est plus spécifique à nos travaux et présente la configuration utilisée pour apprendre les modèles et générer les coefficients en utilisant le système HTS.

2.1 De HTK à HTS

Comme le système proposé par R. Donovan et décrit dans le chapitre précédent, HTS est un système de synthèse reposant sur la suite logicielle HTK proposée initialement dans le cadre de la reconnaissance de la parole. Dans cette section, nous allons présenter les concepts fondamentaux concernant les HMM. Nous poursuivrons par la présentation des adaptations effectuées dans la suite logicielle HTK pour modéliser un signal de parole par des modèles de Markov cachés. Nous expliciterons ensuite en quoi HTS diverge de HTK.

2.1.1 Introduction aux HMM

On considère une séquence d'observations $O = [o_1, \ldots, o_t, \ldots, o_T]$ de T trames. Chaque observation o_t à un instant discret t est un vecteur de dimension M décrivant des coefficients acoustiques auxquels peuvent être ajoutés les coefficients dynamiques de premier ou de second ordre. Un modèle de Markov caché (ou HMM), noté λ, est un modèle stochastique qui suppose que la probabilité d'observer le processus $\{o_t\}$ est conditionné par un processus non observable $\{q_t\}$ purement hypothétique et dont le rôle est de simplifier les dépendances entre les observations. $\{Q_t\}$ est une chaîne de Markov à espace d'états discret. L'observation de $O = [o_1, \ldots, o_t, \ldots, o_T]$ est conditionné par la chaîne $Q = [q_1, \ldots, q_T]$ où la variable aléatoire q_t prend une valeur dans un ensemble de S états possibles.

Un HMM λ est défini par :

$$
\begin{array}{lcll}
\lambda & = & (A, B, \pi) & \\
A & = & \{a_{i,j}\}, & \forall i, j \in [1..S] \\
B & = & \{b_j(o_t)\}, & \forall j \in [1..S], t \in [1..T] \\
\pi & = & \{\pi_i\}, & \forall i \in [1..S] \\
a_{ij} & = & P(q_t = j | q_{t-1} = i), & \forall i, j \in [1..S], t \in [2..T] \\
b_j(o_t) & = & P(o_t | q_t = j), & \forall j \in [1..S], t \in [1..T] \\
\pi_i & = & P(q_1 = i), & \forall i \in [1..S]
\end{array}
\tag{2.1}
$$

où a_{ij} correspond à la probabilité de transition entre l'état i et l'état j pour le processus caché. π_i est la probabilité que l'état i soit l'état initial du processus caché. Enfin $b_j(o_t)$ correspond à la probabilité d'émission de l'observation o_t conditionné la valeur de q_t, ici $q_t = j$. Dans le cadre de HTK, la densité de probabilité $b_j(o_t)$ est représenté par un mélange de lois normales, appelé mixture, comme suit :

$$b_j(o_t) = \sum_{k=1}^{N} w_k \mathcal{N}(o_t; \mu_k, \Sigma_k) \tag{2.2}$$

avec μ_k, Σ_k et w_k qui correspondent, respectivement, au vecteur moyenne de dimension M, à la matrice de covariance MxM et au poids associé à la composante k.

En se basant sur cette définition, [Rabiner1989] identifie trois problèmes reliés aux HMM :

1. déterminer la probabilité $P(O|\lambda)$ qu'une séquence O ait été produite par le HMM λ,
2. déterminer la séquence d'états Q maximisant la probablité conjointe de O et Q étant donné un modèle λ,
3. mettre à jour les paramètres A, B et π du HMM λ à partir des observations.

Déterminer $P(O|\lambda)$

Déterminer $P(O|\lambda)$ consiste à marginaliser la probablité du modèles conjoint, $P(O, Q|\lambda)$ sur l'ensemble des valeurs possibles du processus non observé.

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) \tag{2.3}$$

$$P(O|\lambda) = \sum_{Q} P(O|Q, \lambda) P(Q|\lambda) \tag{2.4}$$

$$= \sum_{q_1 \ldots, q_T} \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1}, q_t} \times b_{q_t}(o_t) \tag{2.5}$$

On pose $\alpha_t(i)$ la probabilité d'avoir émis, par le modèles λ, une séquence d'observation partielle $[o_1, \cdots, o_t]$ et d'avoir $q_t = i$. On considère également $\beta_t i$ la probabilité d'être à l'état $q_t = i$ à l'instant t puis d'émettre la fin de la séquence $(o_{t+1} \ldots o_T)$ par modèles λ. Ces probabilités probabilités s'écrivent donc :

$$\alpha_t(j) = P(o_1 \ldots o_t, q_t = j|\lambda) \tag{2.6}$$

$$\beta_t(i) = P(o_{t+1} \ldots o_T|q_t = i, \lambda) \tag{2.7}$$

Pour résoudre ces équations, il est nécessaire de procéder par récurrence (voir Annexe A.1)

L'algorithme *Forward-backward* [Baum1967] se base sur ces probabilités pour réécrire

l'équation (2.3) en :

$$P(O|\lambda) = \sum_{i=1}^{S} \alpha_T(i) = \sum_{i=1}^{S} \pi_i \beta_0(i) \tag{2.8}$$

L'algorithme *Forward-backward* est un algorithme qui s'appuie sur le paradigme de la programmation dynamique en tirant partie de la seule dépendance markovienne d'ordre 1 de Q.

La complexité de cet algorithme est en $O(S^2T)$ au lieu des $O(S^T)$ si l'évaluation avait été un calcul direct lors de la marginalisation.

Déterminer la séquence Q qui maximise $P(O, Q|\lambda)$

Pour déterminer la séquence d'états cachés Q maximisant $P(O, Q|\lambda)$, on peut introduire une variable intermédiaire $\delta_t(i)$ qui correspond à la probabilité du meilleur chemin aboutissant à l'état i tout en ayant observer la séquence $(o_1 \dots o_t)$:

$$\delta_t(i) = \max_{q_1,\dots,q_{t-1}} P(q_1, \dots, q_t = i, o_1, \dots, o_t|\lambda) \tag{2.9}$$

$delta_t(j)$ peut être déterminé à partir $\delta_{t-1}(i)$ de la manière suivante :

$$\delta_t(j) = [\max_i \delta_t(i) a_{ij}] * b_j(o_t) \tag{2.10}$$

S'agissant d'une récurrence d'ordre 1, le paradigme de programmation dynamique peut de nouveau être utilisé ici. Toutefois, il est nécessaire d'introduire une variable supplémentaire $\psi_t(j)$ qui permet de mémoriser l'état j utilisé pour conditionner l'observation o_t. Grâce à cette variable, l'algorithme de Viterbi [Viterbi1967], qui est appliqué pour résoudre ce problème, peut retrouver le meilleur chemin parcouru. L'algorithme de Viterbi est détaillé dans l'annexe A.2.

Estimation des paramètres du modèles λ

Le dernier problème identifié par [Rabiner1989] est l'estimation des paramètres (A, B, π) du modèle λ. Toutefois, il n'est pas possible d'estimer les paramètres directement car l'émission d'une séquence d'observation O dépend de la séquence d'états cachés Q.

Pour pouvoir estimer ces paramètres, il est nécessaire d'introduire deux nouvelles variables : $\gamma_t(i)$ qui identifie la probabilité d'être dans l'état à i à l'instant t en supposant la séquence d'observations O émise ; $\xi_t(i, j)$ qui représente la probabilité d'être à l'état i

l'instant t et à l'état j à l'instant $t+1$. Ces deux variables sont donc définies par :

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O,\lambda) \tag{2.11}$$
$$\gamma_t(i) = P(q_t = i|O,\lambda) \tag{2.12}$$
$$\tag{2.13}$$

Il est possible d'exprimer $\gamma_t(i)$ grâce à $\xi_t(i,j)$ en marginalisant $\xi_t(i,j)$ sur j :

$$\gamma_t(i) = \sum_{j=1}^{S} \xi_t(i,j) \tag{2.14}$$

Pour estimer les paramètres du HMM, L. Baum et L. Welch [Baum1970] ont proposé un algorithme itératif. Cette méthode suppose un modèle λ préalablement initialisé et utilise ce modèle pour déterminer un nouveau modèle $\overline{\lambda}$. Pour cela, il est nécessaire de disposer d'un ensemble de K séquences d'observations $\mathcal{O} = \{O^1 \dots O^K\}$. Pour obtenir les équations de réestimation, il est nécessaire de dériver la fonction auxiliaire suivante :

$$\mathcal{Q}(\lambda,\overline{\lambda}) = \sum_{k=1}^{K}\sum_{Q}\left[P(Q|O^k,\lambda) \times log[P(O^k,Q|\overline{\lambda})]\right] \tag{2.15}$$

Ces équations sont décrites dans l'annexe A.3.

En se basant sur ces équations de réestimation, le coeur de l'algorithme Baum-Welch se déroule en deux temps :

1. Estimer la log-vraisemblance de $P(O^k,Q^k|\lambda)$ (phase d'Estimation),
2. Mettre à jour les paramètres de manière à maximiser $P(\overline{\lambda}|O^k,Q^k)$ (phase de Maximisation).

Ces opérations sont effectuées jusqu'à ce que la différence entre $P(O^k,Q^k|\overline{\lambda})$ et $P(O^k, Q^k|\lambda)$ soit inférieure à un seuil qui doit être déterminé.

Il est important de noter que l'algorithme forward-backward n'aboutit qu'à un maximum local. L'apprentissage du modèle est donc sensible à la phase d'initialisation des paramètres.

2.1.2 Modélisation HTK

Initialement, la suite logicielle HTK a été conçue pour utiliser les modèles de Markov cachés dans le cadre de la reconnaissance de la parole. Plusieurs raffinements ont été apportés pour satisfaire les contraintes de ce domaine. Nous allons maintenant présenter les spécificités de la suite logicielle HTK utilisée par le système HTS. Pour plus de détails sur HTK, le lecteur pourra se référer à l'article introductif [Young1993] ainsi qu'au *HTK book*[Young2005].

Topologie des modèles

Pour pouvoir modéliser un signal de parole par un HMM, il est nécessaire de fournir deux types d'informations à HTK : l'ensemble des séquences d'observations $\{O^1, \ldots, O^K\}$, qui correspondent à des vecteurs de coefficients acoustiques permettant de décrire un signal de parole et qui sont généralement complétés par les informations de dynamique ; l'ensemble des séquences d'étiquettes phonétiques $\{E^1, \ldots, E^K\}$ associées à ces observations. Ces étiquettes sont obtenues grâce à un processus d'annotation, automatique ou manuel, comme par exemple le processus qui a été utilisé pour ces travaux et qui est présenté dans la section 5.2 du chapitre 5.

En tenant compte de propriétés inhérentes au signal de parole, deux caractéristiques de ce signal permettent de contraindre la topologie des modèles [Odell1995] : la parole correspond à une séquence de phones et chaque phone correspond à une séquence de trames ; il existe une durée minimale naturelle pour chaque phone. Ces contraintes aboutissent à utiliser, dans la majorité des cas, une topologie dite linéaire (de gauche à droite et sans saut, ou modèle de Bakis) illustrée par la figure 2.1. Pour ces nombreux systèmes de reconnaissance ainsi que pour notre étude, un HMM modélise un phone. Il est donc nécessaire de pouvoir concaténer différents HMM pour représenter un énoncé. Pour cela, HTK ajoute en début et en fin de HMM deux états non émetteurs qui servent uniquement à pouvoir ancrer deux modèles consécutifs.

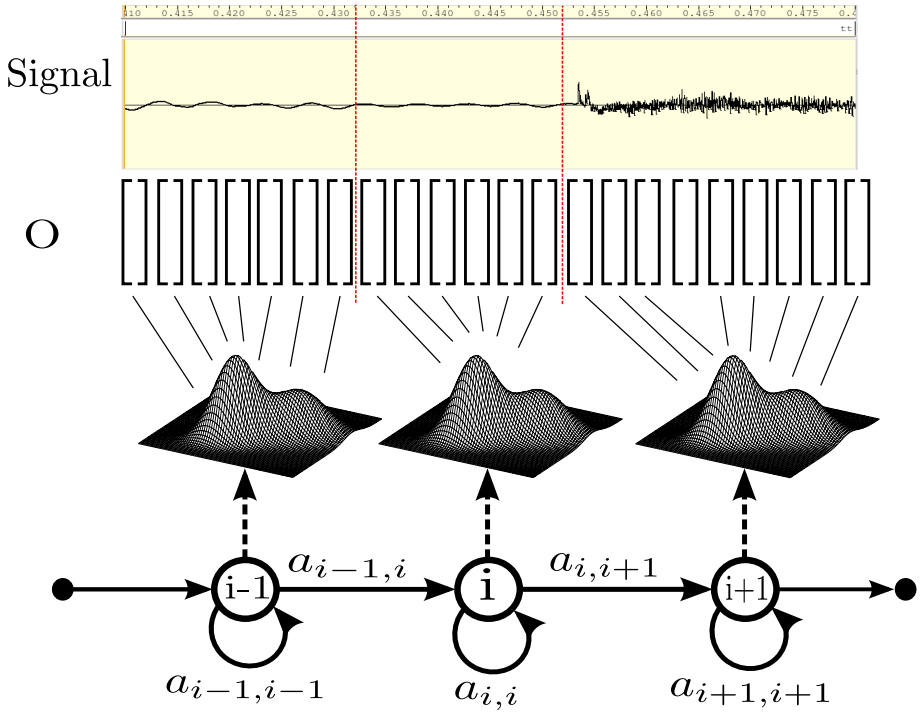


FIGURE 2.1 – Topologie de HMM couramment utilisés avec HTK pour modéliser un signal de parole. Si on considère un signal associé à un phone et la séquence d'observations O extraite de ce signal, un HMM repose sur une topologie linéaire pour modéliser ce signal en utilisant des lois normales comme probabilités d'émission.

HTK introduit la notion de flux afin de pouvoir considérer différentes parties d'un vecteur d'observation comme statistiquement indépendantes [Young2005].

Élagage

Lors de la phase d'apprentissage, l'estimation des variables α et β peut s'avérer coûteuse en temps et en espace. Néanmoins, il est possible de réduire l'espace de recherche en tenant compte de la topologie particulière des HMM de Bakis sans saut. En effet, cette topologie impose de parcourir tous les états du HMM. Il n'est donc pas possible que le premier état soit associé à la dernière trame dans le cadre d'un HMM composé d'au moins deux états émetteurs. En appliquant ce raisonnement sur l'ensemble des états, nous constatons qu'en réalité l'association entre les trames et les états de la phrase-HMM [1] forme un faisceau comme l'illustre la figure 2.2.

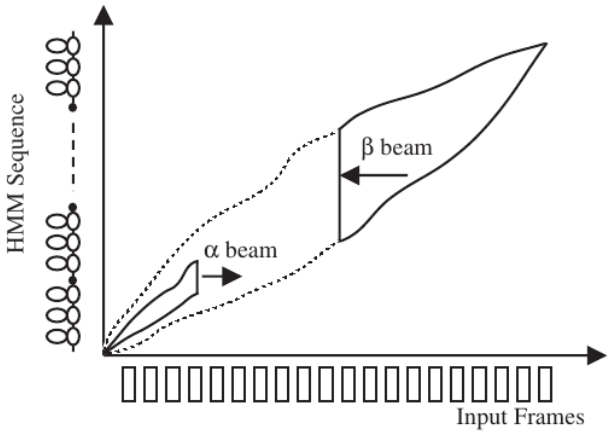


FIGURE 2.2 – Élagage lors du Forward-backward. Lors de la phase de calcul de β, une largeur maximale de faisceau est définie. Cette largeur est contrainte, en se basant sur les résultats obtenus lors de la phase Backward, pour déterminer α. Figure extraite de [Young1993].

Ainsi il est possible de réduire la complexité en temps et en espace pour déterminer $P(O|\lambda)$ en effectuant une opération d'élagage (pruning)[Young1993]. Pour cela, on considère la taille maximale F du faisceau. Lors de la phase *backward*[2], la taille du faisceau est limitée à F et $\overline{\beta}$. Pour le calcul de α, cette taille est encore réduite en tenant compte des résultats obtenus lors de la phase *backward*. De cette manière, les états dont la probabilité d'émission de l'observation o_t est très faible, à cause de la topologie du modèle, sont ignorés ce qui permet de réduire l'espace et le temps nécessaire pour déterminer α et β.

1. Une phrase-HMM correspond à la concaténation des HMM déterminés par la séquence d'étiquettes phonétiques associées.
2. HTK exécute le calcul de β avant le calcul de α

Arbre de décision et tying

Afin d'obtenir des modèles plus pertinents, le système HTK permet de prendre en compte le contexte linguistique associé à chaque segment utilisé pour effectuer l'apprentissage des HMM. Toutefois, la prise en compte du contexte conduit rapidement à une explosion combinatoire des paramètres. Pour pallier ce problème, HTK introduit des arbres de décision [Young1994] pour limiter le nombre de paramètres contextuels.

Les arbres de décision associés à un paramètre d'observation et à un état donné, dont un exemple est illustré par la figure 2.3, sont des arbres binaires respectant la topologie suivante :

- chaque **noeud** correspond à une propriété liée au contexte linguistique et prosodique du segment modélisé. À chaque propriété est associé un ensemble de valeurs qui définissent cette propriété. En fonction des valeurs de description, les paramètres acoustiques sont obtenus par descente dans l'arbre jusqu'à rencontrer une feuille.

- chaque **feuille** contient une distribution statistique. Pour aboutir à une feuille, il est nécessaire de valider un ensemble de caractéristiques linguistiques/prosodiques. Cet ensemble correspond au parcours dans l'arbre. Si l'on considère un ensemble de modèles liés à des observations dont les descripteurs valident cet ensemble de caractéristiques, nous obtenons alors un ensemble de distributions correspondant aux émissions de chacun des modèles pour l'état donné. La distribution associée à la feuille est déterminée à partir de cet ensemble grâce à un mécanisme de partage de distributions (ou *tying*).

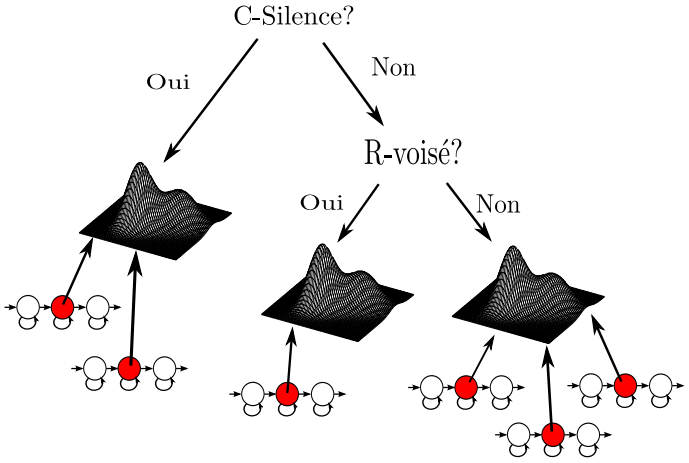


FIGURE 2.3 – Exemple d'arbre de décision. Si l'on suppose un arbre associé au second état des HMM, alors les feuilles de cet arbre correspondent à des distributions statistiques et, pour chaque HMM, la probabilité d'émission du second état est *liée* à l'une de ces distributions. La distribution est déterminée en fonction des descripteurs caractérisant les segments associés aux HMM.

2.1.3 De HTK à HTS

HTS utilise les paramètres de modèles HMM pour générer les coefficients acoustiques nécessaires à la synthèse du signal de parole[Tokuda1995].

La différence la plus importante se situe lors de la phase de synthèse. En effet, comme nous l'avons précédemment évoqué, l'objectif de cette phase est de générer les paramètres correspondant à un énoncé dont les descripteurs ont été obtenus lors de la phase de traitements linguistiques. Pour obtenir le modèle correspondant à l'énoncé que l'on souhaite synthétiser, deux étapes sont nécessaires.

Soit S le nombre de segments constituant l'énoncé. La première étape consiste à concaténer S structures de HMM pour obtenir la topologie du modèle correspondant à l'énoncé. Toutefois, ce modèle n'est pas complet car il manque les probabilités d'émission. Pour obtenir ces distributions, les arbres de décision sont utilisés. En effet, chacun des S segments composant l'énoncé est qualifié par les descripteurs obtenus lors de la phase de traitements linguistiques. En se basant sur ces descripteurs, les probabilités d'émission sont sélectionnées dans les arbres de décision pour compléter le modèle associé à l'énoncé. L'objectif du système de synthèse HTS est alors de déterminer la séquence d'observations $O^\star$ telle que :

$$O^\star = \underset{O}{\operatorname{argmax}}\ P(O|\lambda) \tag{2.16}$$

où λ correspond au HMM complet associé à l'énoncé.

Le système HTK a pour objectif de déterminer la séquence de descripteurs en ayant comme consigne une séquence de coefficients acoustiques décrivant le signal de parole. Au contraire, HTS utilise la séquence de descripteurs pour construire les modèles qui vont permettre de produire le signal de parole.

2.2 Génération des trajectoires

L'apport principal du système HTS est de pouvoir générer les coefficients acoustiques utilisés par les outils SPTK [Fukada1992] et STRAIGHT [Kawahara1999] pour synthétiser ensuite le signal de parole. Dans cette section nous allons présenter les équations utilisées pour effectuer la génération. Cette présentation s'effectue en deux temps : tout d'abord l'équation fondamentale, permettant de lier coefficients statiques et coefficients dynamiques ; puis la variance globale, mise en place pour pallier le problème de sur-lissage.

2.2.1 Vecteur d'observations

Dans le cadre du système HTS, chaque observation o_t, illustrée par la figure 2.4, est un vecteur composé de cinq blocs. La décomposition de o_t est rendue possible grâce au concept de flux proposé par HTK et présenté dans la section précédente. Cinq flux sont donc nécessaires :

- Le premier flux contient les coefficients MGC, tels que présentés dans la section 1.1.3 du chapitre précédent, ainsi que les coefficients dynamiques de premier et second ordre,
- Les trois flux suivants contiennent, respectivement, le F0, la dynamique de premier ordre et la dynamique de second ordre,
- Le dernier flux contient les coefficients d'apériodicité nécessaires au vocodeur STRAIGHT.

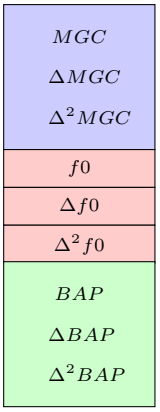


FIGURE 2.4 – Vecteur d'observations o_t utilisé par HTS. Figure inspirée de [Yoshimura1999]

2.2.2 Équation fondamentale

L'ensemble des apports effectués par le système HTS découle de l'équation linéaire suivante qui n'est que l'expression numérique d'une dérivée :

$$O = W \times C \tag{2.17}$$

où le vecteur C correspond aux coefficients statiques et O au vecteur d'observation pour les HMM (coefficients statiques et dynamiques). Enfin, W est une matrice de fenêtrage permettant d'obtenir les coefficients dynamiques à partir des coefficients statiques C. La forme de la matrice W [3] est fixe et peut être décrite par le système suivant (illustré

3. v_t^0 permet de conserver les coefficients statiques

figure 2.5) :

$$W = [v_1, v_2, \ldots, v_T]^\top \tag{2.18}$$

$$v_t = [v_t^0, v_t^1, v_t^2] \tag{2.19}$$

$$v_t^0 = [\underbrace{0, \ldots, 0}_{t-1}, 1, \underbrace{0, \ldots, 0}_{T-t}]^\top \tag{2.20}$$

$$v_t^1 = [\underbrace{0, \ldots, 0}_{t-L^1-1}, w_{-L^1}^1, \ldots, w_0^1, \ldots, w_{L^1}^1, \underbrace{0, \ldots, 0}_{T-(t+L^1)}]^\top \tag{2.21}$$

$$v_t^2 = [\underbrace{0, \ldots, 0}_{t-L^2-1}, w_{-L^2}^2, \ldots, w_0^2, \ldots, w_{L^2}^2, \underbrace{0, \ldots, 0}_{T-(t+L^2)}]^\top \tag{2.22}$$

$$\tag{2.23}$$

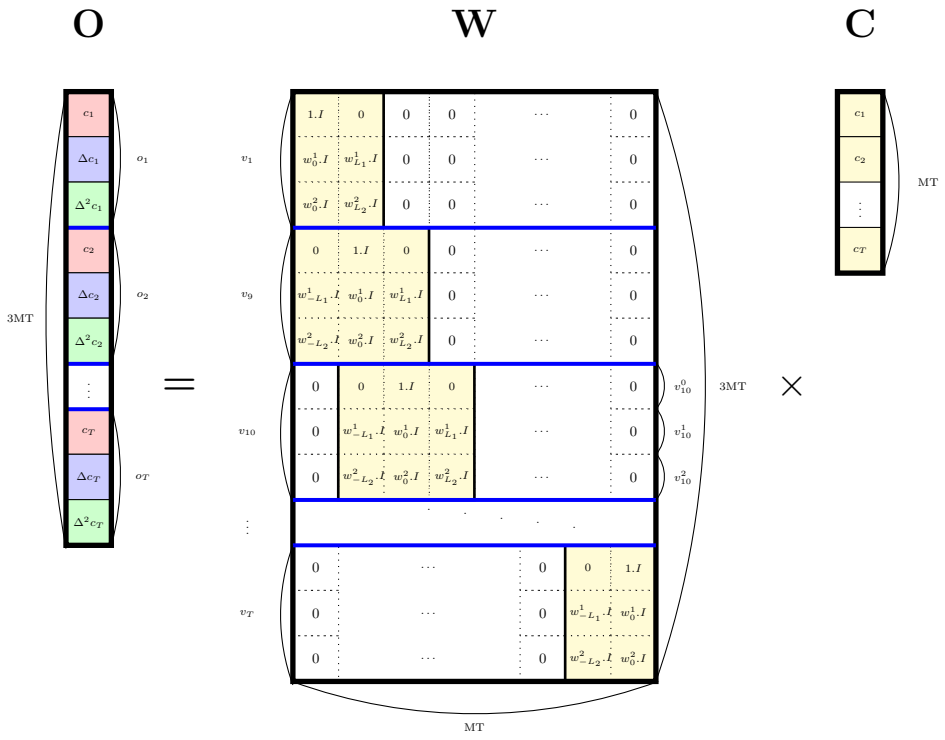


FIGURE 2.5 – Représentation de l'équation 2.18 : dans cet exemple $L_1 = L_2 = 1$ (opérateur de dérivation sur 3 points) et chaque case correspond à une matrice de taille $M \times M$ où M représente l'ordre des coefficients acoustiques. Figure inspirée de [Zen2007a]. (Un seul flux est représenté.)

Lors de la phase de génération de paramètres, en supposant la séquence d'états Q connue, [Tokuda1995a, Tokuda2000b] note que déterminer les trajectoires des coefficients revient à maximiser $P(O|Q, \lambda)$ où λ correspond à la phrase-HMM issue de la concaténation des HMM déterminés par la séquence de descripteurs obtenus à l'issue de l'analyse linguistique du texte à synthétiser. En prenant en compte la relation 2.17, et en définissant le critère suivant :

$$\frac{\partial log(P(WC|Q, \lambda))}{\partial C} = 0 \tag{2.24}$$

Maximiser $P(O|Q,\lambda)$ revient à résoudre le système d'équations suivant :

$$(W^\top \Sigma^{-1} W).C = W^\top \Sigma^{-1} \mu \tag{2.25}$$

où μ est un vecteur obtenu par la concaténation des vecteurs moyennes μ_t issus des états q_t tels que $q_t \in Q$; Σ correspond à la matrice de covariance obtenue par concaténation des matrices Σ_t associées aux états q_t issus de la séquence d'états Q. Par hypothèse, la séquence d'états Q étant connue, et la matrice de fenêtrage W étant fixe, la seule inconnue de cette équation est C, le vecteur de coefficients que l'on souhaite générer.

Néanmoins, dans la pratique, la séquence d'états Q n'est pas connue à l'avance. L'objectif consiste plutôt à déterminer le vecteur de coefficients C qui maximise $P(O|\lambda)$. Pour simplifier le problème, K. Tokuda ET AL. [Tokuda1995a, Tokuda1995] posent comme hypothèse :

$$P(O|\lambda) = \max_{Q} P(O,Q|\lambda) \tag{2.26}$$

En s'appuyant sur cette hypothèse, déterminer le vecteur de coefficients C optimal revient à maximiser la loi conjointe $P(O,Q|\lambda)$. Néanmoins, comme l'indique [Tokuda1995a], $P(O,Q|\lambda)$ peut être transformée en :

$$P(O,Q|\lambda) = P(Q|\lambda) \times P(O|Q,\lambda) \tag{2.27}$$

$P(Q|\lambda)$ ne dépendant pas de O, maximiser $P(O|\lambda)$ revient donc à maximiser $P(O|Q,\lambda)$.

Un algorithme a été mis au point pour résoudre l'équation 2.25 et est présenté dans [Tokuda1995a, Tokuda1995]. Cet algorithme consiste à estimer une trame à l'instant t telle que la mise à jour des paramètres μ_t et Σ_t liés à la composante i_t de la mixture reliée à l'état q_t implique la plus forte augmentation de $P(O,Q|\lambda)$. Si cette augmentation est suffisamment élevée, μ_t, Σ_t et C sont mis à jour sinon la procédure s'arrête et retourne la séquence de coefficients C. L'algorithme dépend donc fortement de la condition initiale : le choix de la séquence d'états Q, déterminée en utilisant les durées moyennes de séjour, qui doit être proche de l'optimal.

[Tokuda2000b] présente une approche différente qui suppose que la séquence d'états Q est cachée. Cela revient donc à vouloir maximiser $P(O|\lambda)$ sans poser une hypothèse particulière sur la séquence Q. L'algorithme repose sur une approche de type EM qui met à jour les paramètres des distributions associées au couple état/mixture (q_t, i_t) pour l'ensemble des trames T. De plus, depuis [Tokuda2000b], l'équation (2.25) est résolue par une décomposition de Cholesky qui, grâce à la structure particulière de la matrice W, permet de passer d'une complexité de $O(T^3M^3)$ à une complexité de $O(TM^3L^2)$ sachant que $L << T$.

Données : Une phrase-HMM λ
Résultat : Une séquence de coefficients C

$cur = 0$;
Définir une séquence d'état Q en utilisant les durées moyennes de λ;
Résoudre l'équation (2.25) pour déterminer C en utilisant Σ^{-1} et $\Sigma^{-1}M$;

répéter
  $prev = curr$;
  Déterminer $P(q_t = (q, i)|O, \lambda)$ et $curr = P(O|\lambda)$ via l'algorithme forward/backward;
  Déterminer $\overline{\Sigma^{-1}}$ en connaissant Σ^{-1};
  Déterminer $\overline{\Sigma^{-1}M}$ en connaissant $\Sigma^{-1}M$;
  Résoudre l'équation (2.25) pour déterminer $\overline{C}$ en utilisant $\overline{\Sigma^{-1}}$ et $\overline{\Sigma^{-1}M}$;

  $C = \overline{C}$;
  $\Sigma^{-1} = \overline{\Sigma^{-1}}$;
  $\Sigma^{-1}M = \overline{\Sigma^{-1}M}$;
jusqu'à $((curr - prev) \leq seuil)$;

Algorithme 1: Algorithme de synthèse proposé dans [Tokuda2000b]

2.2.3 Variance globale (GV)

L'algorithme précédent permet de générer un vecteur de coefficients C compatible avec une synthèse du signal de parole par un vocodeur (par exemple STRAIGHT). Cependant, expérimentalement, la variance des coefficients générés par HTS est souvent trop faible et le surlissage qui en résulte conduit à un signal de synthèse *étouffé*. Pour pallier ce défaut, la notion de variance globale [Toda2005] a été introduite. L'objectif de cette idée est d'estimer la variance intrinsèque des trames acoustiques d'un locuteur puis de l'utiliser, lors de la phase de génération, pour accroître artificiellement la variance des coefficients synthétisés.

La variance globale, associée aux vecteurs de coefficients C de dimension M est définie comme le vecteur $v(C) = [v(1), \ldots, v(m), \ldots, v(M)]^\top$ où :

$$v(m) = \frac{1}{T}\sum_{t=1}^{T}(C_t(m) - \overline{C}(m))^2 \tag{2.28}$$

$$\overline{C}(d) = \frac{1}{T}\sum_{\tau=1}^{T}C_\tau(m) \tag{2.29}$$

où T correspond au nombre de trames analysées pour un énoncé. Utiliser l'énoncé comme horizon de calcul constitue un compromis entre le nombre de vecteurs nécessaires pour déterminer la variance globale et le nombre de valeurs nécessaires à l'apprentissage d'une distribution gaussienne.

L'algorithme de génération présenté précédemment a donc été modifié pour prendre en compte cette variance globale [Toda2005a]. L'algorithme consiste à maximiser le critère $\mathcal{L}$ suivant :

$$\mathcal{L} = P(O|\lambda)^{\omega} \times P(v(C)|\lambda_v) \tag{2.30}$$

où $v(C)$ correspond à la variance globale de la séquence de coefficients C que l'on souhaite obtenir, λ_v la distribution modélisant la variance globale et ω une constante permettant de contrôler l'influence de la variance globale.

En utilisant une méthode de gradient, il est possible de déterminer C itérativement grâce à :

$$\overline{C^{(i+1)}} = C^{(i)} + \alpha \cdot \Delta C^{(i)} \tag{2.31}$$

où α correspond au pas utilisé par la méthode de gradient.

Deux méthodes du gradient sont proposées dans [Toda2005a] pour effectuer la génération de C en utilisant la variance globale : la descente de gradient, si l'on utilise uniquement la dérivée de premier ordre ; la méthode de Newton-Raphson si les dérivées de premier et second ordre sont prises en compte. ΔC est défini par :

$$\Delta C^{(i)} = \begin{cases} \left.\frac{\partial L}{\partial C}\right|_{C=C^{(i)}} & \text{, descente de gradient} \\ -\left(\frac{\partial^2 L}{\partial C \partial C^{\top}}\right)^{-1} \left.\frac{\partial L}{\partial C}\right|_{C=C^{(i)}} & \text{, méthode de Newton-Raphson} \end{cases} \tag{2.32}$$

Le processus de génération repose sur l'algorithme standard décrit précédemment (voir l'algorithme 1). La résolution de l'équation (2.30) est l'étape suivant la résolution de l'équation (2.25).

2.3 Modélisation

Les concepts utilisés lors de la phase de génération des coefficients acoustiques ayant été présentés, nous allons maintenant décrire les modifications apportées aux modèles pour qu'ils puissent être utilisés lors de cette phase.

2.3.1 Modélisation du F0

Pour obtenir une représentation unifiée du F0, HTS utilise des distributions dites multi-espaces [Tokuda2000a] (ou MSD). La particularité d'une telle distribution est de considérer qu'une variable aléatoire est en réalité constituée de deux informations : la dimension n de l'espace ayant comme support $\mathbb{R}$ et une valeur prise dans cet espace.

Dans le cadre de la représentation du F0, deux états sont à prendre en compte. Ainsi, les probabilités d'émission $b(o_t)$ sont définies de la manière suivante :

$$b(o_t) = \begin{cases} w_1 \mathcal{N}(V(o_t); \mu_{F0}, \Sigma_{F0}), & S(o_t) = \{1\}, \quad \text{Cas voisé} \\ w_2 \mathcal{N}(V(o_t); 0, 0), & S(o_t) = \{0\}, \quad \text{Cas non voisé} \end{cases} \tag{2.33}$$

où $S(o_t)$ correspond à la dimension n de l'espace associée à l'observation o_t ; $V(o_t)$ correspond à la valeur prise par o_t dans l'espace $\mathbb{R}^{S(o_t)}$. $\mathcal{N}(V(o_t); 0, 0)$ correspond à une distribution de Dirac centrée en 0 (la distribution sera centrée en $-1.e^{10}$ s'il s'agit de la modélisation du logarithme de la fréquence fondamentale).

En utilisant les MSD, l'estimation de $P(O|\lambda)$ devient alors :

$$P(O|\lambda) = \sum_{\forall Q,L} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t,l_t} \mathcal{N}_{q_t,l_t}(V(o_t), \mu_t, \Sigma_t) \tag{2.34}$$

où Q correspond à la séquence des états non observés de la chaîne markovienne et $L = [l1, \ldots, l_t, \ldots, l_T]$ avec $l_t = S(o_t)$.

Néanmoins, la limite principale d'une telle modélisation est la disjonction entre la distribution représentant la partie voisée et celle représentant la partie non voisée du F0. Cette disjonction impose que, lors de la phase d'estimation de l'algorithme EM, une trame contribue exclusivement à l'une des deux distributions. Lors des traitements de trames en frontière de voisement, l'algorithme devient sensible aux erreurs d'analyse du F_0. En couplant cela au fait que chaque paramètre est traité indépendamment, il est possible d'obtenir, lors de la phase de génération, un spectre incohérent avec le F0 (par exemple, un spectre lié à une trame voisée alors que le F0 est non voisé). Pour résoudre cette limite, d'autres modélisations [Latorre2011, yu2011] ont été proposées mais ne sont actuellement pas intégrées au système HTS.

2.3.2 Modélisation de la durée

Soit $D = (d_1, \ldots, d_j, \ldots, d_N)$ les durées, en nombre de trames, des N états composant la phrase-HMM nécessaire à la génération d'une phrase de synthèse. En supposant ces variables aléatoires indépendantes, $P(D|\lambda)$ s'écrit :

$$\log P(D|\lambda) = \sum_{j=1}^{N} \log P_j(d_j) \tag{2.35}$$

où la durée de séjour dans un état j est décrite par la relation suivante :

$$P_j(d_j) = a_{jj}^{d_j - 1}.(1 - a_{jj}) \tag{2.36}$$

L'objectif de l'algorithme de génération est de déterminer la séquence de durées maximisant $P(D|\lambda)$. En combinant la relation (2.35) et la relation (2.36), nous constatons que la séquence de durée optimale implique un temps de séjour d'une trame pour chaque état [Zen2004].

Pour résoudre cette difficulté, la durée doit être modélisée de manière explicite. [Yoshimura1998] utilise des lois normales. L'objectif est d'aboutir à une modélisation qui permet d'obtenir une séquence d'états Q maximisant la relation suivante :

$$\log P(Q|\lambda, T) = \sum_{j=1}^{N} P(d_j) \tag{2.37}$$

sous la contrainte :

$$T = \sum_{j=1}^{N} d_j \tag{2.38}$$

où d_j correspond à la durée associée à l'état j du HMM λ et T au nombre total de trames devant être générées. En modélisant $P(d_j)$, seulement lors de la phase de synthèse, par une loi normale de moyenne μ_j et de variance σ_j, la durée des états maximisant l'équation (2.37) est obtenue par :

$$d_j = \mu_j + \rho \cdot \sigma_j^2 \tag{2.39}$$

$$\rho = \frac{T - \sum_{j=1}^{N} \mu_j}{\sum_{j=1}^{J} \sigma_j^2} \tag{2.40}$$

Afin d'unifier la modélisation de la durée entre la phase de synthèse et la phase d'apprentissage, H. Zen ET AL. [Zen2004] ont proposé d'utiliser des modèles semi-markoviens, ou HSMM [Russell1985], où la durée de séjour dans un état est modélisée par une distribution gaussienne. Ainsi, lors de la phase d'apprentissage, les formules α et β ont du être adaptées pour prendre en compte cette nouvelle définition explicite de la durée :

$$\alpha_0(j) = \pi_j, \tag{2.41}$$

$$\alpha_t(j) = \sum_{d=1}^{t} \sum_{i=1, i\neq j}^{N} \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{\tau=t-d+1}^{t} b_j(\mathbf{o}_\tau), 1 \leq t \leq T \tag{2.42}$$

$$\beta_T(i) = 1 \tag{2.43}$$

$$\beta_t(i) = \sum_{d=1}^{T-t} \sum_{j=1, i\neq j}^{N} a_{ij} p_j(d) \prod_{\tau=t-d+1}^{t} b_j(\mathbf{o}_\tau) \beta_t(j), 0 \leq t < T \tag{2.44}$$

où la probabilité d'une durée de séjour d de l'état j est représentée par $p_j(d)$. En se basant

sur ces équations, $P(O|\lambda)$ s'écrit :

$$P(O|\lambda) = \sum_{j=1}^{N} \sum_{j=1, j\neq i}^{N} \sum_{d=1}^{t} \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{\tau=t-d+1}^{t} b_j(o_\tau) \beta_t(j) \tag{2.45}$$

Les modèles respectant la topologie des HSMM et dont les émissions contiennent des distributions MSD sont appelés MSD-HSMM.

2.3.3 Arbre de décision

HTS repose sur la caractérisation d'un segment acoustique par un ensemble conséquent de descripteurs (53 pour le jeu de descripteurs standard [Tokuda2000]). Il est, en pratique, impossible de définir un corpus couvrant l'ensemble de ces descripteurs en nombre suffisant. Comme nous l'avons indiqué dans l'introduction du système HTK (section 2.1.2 de ce chapitre), il est possible de définir un arbre de décision afin de garantir la présence d'un modèle pour chaque combinaison possible de descripteurs.

Dans le cadre du système HTS, un arbre de décision est associé à chaque état du HMM et chaque type de coefficient (MGC, F0, apériodicité). Un dernier arbre est associé à la durée des états. La figure 2.6 illustre la composition d'un modèle en utilisant différents arbres.

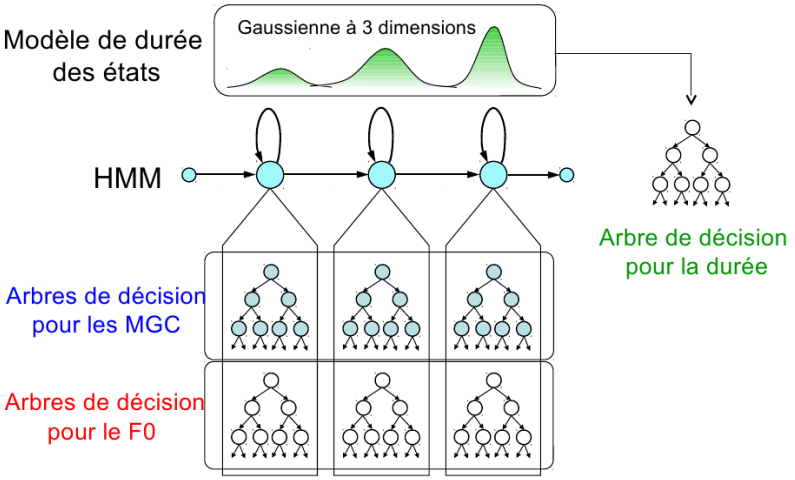


FIGURE 2.6 – Topologie des modèles utilisant les arbres de décisions. Figure extraite de [Tokuda2000] (l'apériodicité n'est pas représentée)

Afin de construire l'arbre de décision, un algorithme itératif (détaillé algorithme 2) consistant à effectuer des partitions est appliqué. Soit un arbre de décision U constitué de F feuilles. Une itération de l'algorithme consiste à déterminer une feuille S_f, avec $1 \leq f \leq F$, dont l'éclatement en deux partitions selon un descripteur q (S_{fq}^{+} et S_{fq}^{-}) permettrait d'améliorer la vraisemblance des données d'apprentissage. On note U' l'arbre

résultant de cette transformation. Pour limiter la profondeur de l'arbre et ainsi conserver une capacité de généralisation sur un ensemble autre que l'ensemble d'apprentissage, un critère de parcimonie doit être appliqué. HTS utilise un critère de type MDL, ou Minimum Description Length, [Shinoda2000].

Données : un ensemble de questions avec les descripteurs valides + l'ensemble des distributions associées aux descripteurs les identifiant
Résultat : L'arbre de décision optimal, $U^\star$

Définir un arbre initial $U = \{S_0\}$ contenant l'ensemble des lois normales obtenues après apprentissage HTK;
répéter
  Trouver la feuille S_m de l'arbre U et la question q qui maximise une fonction de coût $|\theta_m(q)|$;
  si $\theta_m(q) < 0$ **alors**
    Scinder S_m en deux partitions en utilisant q pour obtenir U';
    $U = U'$;
  fin
jusqu'à $\theta_m(q) >= 0$;
$U^\star = U$;

Algorithme 2: Algorithme de construction d'un arbre de décision

Tout d'abord, la différence $\theta_f(q)$, permettant de quantifier l'apport de l'éclatement de la feuille f de l'arbre U en utilisant la question q, est définie comme suit :

$$\theta_f(q) = D(U') - D(U) \tag{2.46}$$

avec $D(U)$ correspondant à la longueur de description de l'arbre U et définie par :

$$D(U) \equiv -\mathcal{L}(U) + LF\log(G) + C \tag{2.47}$$

où L correspond à la dimension des vecteurs d'observation. En définissant $\zeta_t(f)$ comme la probabilité *a posteriori* d'utiliser la distribution issue de la partition f à l'instant t, on définit $\Gamma_m = \sum_{t=1}^{T} \zeta_t(f)$ comme le taux d'utilisation global de la distribution f ; G est alors défini par $G = \sum_{f=1}^{F} \Gamma_f$. C est considéré ici comme une valeur constante. $\mathcal{L}(U)$ correspond à la log-vraisemblance de l'arbre U obtenue de la manière suivante :

$$\mathcal{L}(U) \simeq \sum_{f=1}^{F}\sum_{t=1}^{T} \zeta_t(m) log\mathcal{N}_f(\mathbf{o}_t; \mu_f, \Sigma_f) \tag{2.48}$$

avec μ_f et Σ_f respectivement la moyenne et la matrice de covariance de la distribution m.

Par construction $D(U')$ est définie par :

$$D(U') \triangleq -\mathcal{L}(U') + L(F+1)log(G) + C \tag{2.49}$$

La figure 2.7 illustre l'ensemble des concepts introduits pour la construction d'un arbre de décision en utilisant le critère MDL dont l'objectif est de déterminer une taille d'arbre qui offre un compromis entre la précision de la modélisation et la parcimonie de description du modèle. La qualité du modèle est représentée par $\mathcal{L}(U)$ et la longueur de description du modèle par $LF\log(G) + C$. Le critère permet ainsi de d'obtenir le nombre de feuilles F qui minimise la longueur de description $D(U)$ associée à l'arbre U.

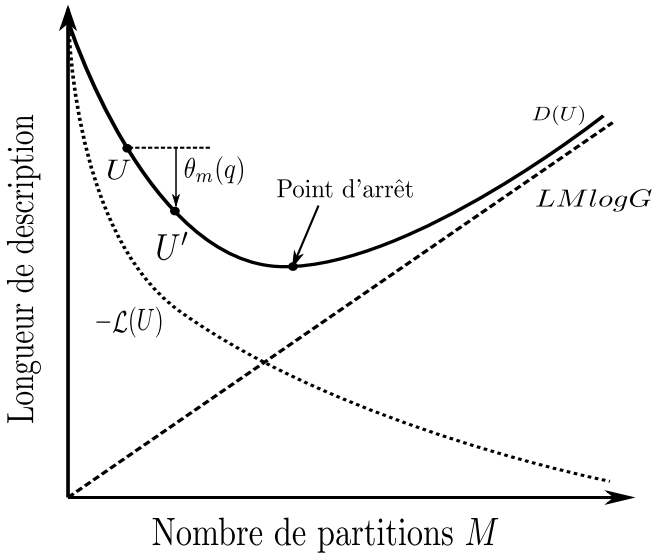


FIGURE 2.7 – Illustration du critère MDL. Figure inspirée de [Yamagishi2006a]

2.3.4 Quelques évolutions majeures

Actuellement, le système HTS fait l'objet d'une attention particulière et plusieurs évolutions du système, recensées dans [Zen2009], ont été proposées. Dans cette partie, nous nous focalisons sur deux évolutions majeures pour l'apprentissage des modèles. Tout d'abord, nous présenterons les *trajectory*-HMM dont l'apport est d'apprendre les paramètres non plus en fonction de O mais de C en se basant sur la relation (2.17). Ensuite, nous introduirons le critère MGE (Minimum Generation Error), remplaçant le critère de maximum de vraisemblance.

Les *trajectory*-HMM

Lors de la phase de synthèse, la relation (2.17) permet d'exprimer les coefficients statiques C que l'on souhaite générer en fonction des coefficients statiques et dynamiques issus des observations O. Néanmoins, lors de la phase d'apprentissage, cette relation n'est pas prise en compte. En effet, les coefficients dynamiques sont extraits en amont de la phase d'apprentissage effectuée par HTS. Ainsi, les paramètres des MSD-HSMM sont mis à jour en ne considérant pas la contrainte liant les coefficients statiques aux coefficients dynamiques.

Afin de pallier cette différence entre la phase de synthèse et la phase d'apprentissage, les *Trajectory-HMM* sont introduits dans [Zen2007a]. Pour cela, il faut partir du constat que l'équation suivante n'est pas valide :

$$C^\star = \underset{C}{\operatorname{argmax}}\{P(WC|\lambda, T)\} \tag{2.50}$$

où l'on cherche la séquence de coefficients $C^\star$ maximisant $P(O|\lambda, T) = P(WC|\lambda, T)$ pour un énoncé composé de T trames.

Cela est dû au fait que la relation suivante n'est pas vérifiée :

$$\int_{\mathbb{R}^{MT}} \mathcal{N}(WC|\mu, \Sigma)dC = 1 \tag{2.51}$$

où C correspond au vecteur (colonne) de coefficients de dimension MT (M représente la dimension du vecteur de coefficients associé à une seule trame). μ désigne le vecteur colonne de dimension $3MT$ contenant les T vecteurs espérances et Σ est la matrice obtenue par concaténation des matrices de covariance associées à la séquence d'états Q.

Les *trajectory*-HMM reposent sur l'introduction d'un coefficient de normalisation, Z, pour valider la relation suivante :

$$\int_{\mathbb{R}^{MT}} \frac{1}{Z}\mathcal{N}(WC|\mu, \Sigma)dC = 1 \tag{2.52}$$

En définissant Z de la manière suivante, l'équation (2.52) peut être considérée comme valide :

$$Z = \int_{\mathbb{R}^{MT}} \mathcal{N}(Wc|\mu, \Sigma)dc \tag{2.53}$$

L'introduction de ce coefficient permet, lors de la phase d'apprentissage, de prendre en compte explicitement la relation entre coefficients dynamiques et coefficients statiques. Néanmoins, bien que cette modification des modèles semble améliorer significativement certains résultats [Shannon2011], elle n'est actuellement pas intégrée au système HTS.

Le critère MGE

L'utilisation du critère de maximum de vraisemblance lors de la phase d'apprentissage est classique pour des applications en reconnaissance de la parole. Couplé à un modèle de langage, l'objectif sera, à partir d'une séquence d'observations acoustiques, d'obtenir une séquence des modèles les plus probables. L'utilisation des HMM en synthèse est de nature différente, puisque ces modèles sont utilisés comme générateurs d'observations. Le critère de maximum de vraisemblance porte sur la pertinence des données par rapport à un modèle et non par rapport aux données générées via ce modèle.

Un nouveau critère a été défini pour la phase d'apprentissage : le critère d'erreur mini-

male de génération [Wu2006] (Minimum Generation Error ou MGE). Ce critère consiste à mettre en place une phase de génération implicite et à en vérifier la pertinence en respectant la contrainte de la synthèse : générer des vecteurs acoustiques les plus proches possibles de ceux extraits du signal naturel. Ce critère repose donc sur la définition d'une distance $D(C, \tilde{C}(\lambda, Q))$ entre la séquence de paramètres originaux C et la séquence de paramètres générés $\tilde{C}(\lambda, Q)$ où λ et Q sont connus.

Actuellement deux distances ont été expérimentées pour le critère MGE : la distance euclidienne [Wu2006] et une distorsion spectrale [Wu2008]. Néanmoins, cette dernière distance a été analysée pour le cas où le spectre était représenté par des coefficients LSP [4]. Le critère MGE utilisant la distance euclidienne a été intégré au système HTS à partir de la version 2.2 [Oura2011].

2.4 Processus d'apprentissage

Le processus d'apprentissage, illustré figure 2.8, se décompose en trois étapes. La première étape (non représentée sur la figure) est effectuée avant l'utilisation du système HTS et consiste à obtenir les vecteurs d'observations o_t qui vont permettre d'apprendre les modèles. Comme nous l'avons précisé précédemment, cette étape est effectuée par STRAIGHT [Kawahara1999] et SPTK [Fukada1992] ; les dynamiques de premier et de second ordre sont calculées en utilisant une matrice W prédéfinie. La seconde étape correspond à l'adaptation de la topologie des modèles pour les transformer en MSD-HSMM. Pour effectuer cela, HTS fonctionne en deux temps : l'apprentissage de modèles HMM, utilisant des distributions MSD, puis la transformation des MSD-HMM en MSD-HSMM. Enfin, la dernière étape consiste à prendre en compte les contextes et effectuer le calcul de l'arbre de décision.

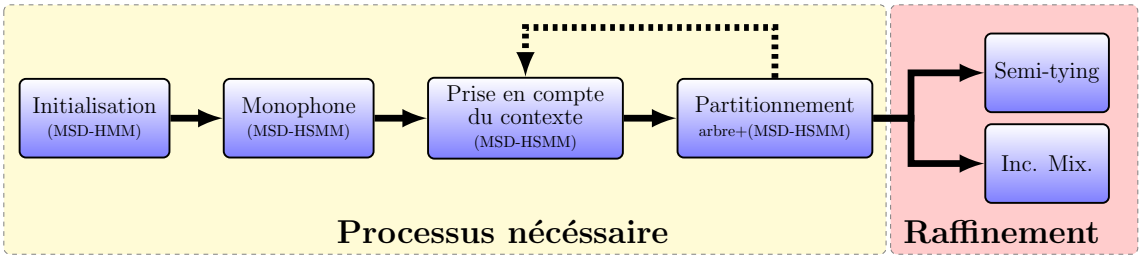


FIGURE 2.8 – Séquence d'apprentissage du système HTS

À ce stade, le système HTS permet d'obtenir des coefficients qui permettent de synthétiser un signal associé à n'importe quel énoncé. Néanmoins, une étape supplémentaire peut être appliquée : soit la transformation de la probabilité d'émission d'une gaussienne en une mixture composée de deux gaussiennes ; soit l'utilisation de distributions dites STC [5]. Dans le cadre de l'utilisation de distributions STC, la matrice de covariance associée à chaque

4. Line Spectral Pairs
5. Semi-Tied Component

probabilité d'émission est obtenue par la multiplication de deux matrices : une matrice diagonale, spécifique à chaque distribution gaussienne, et une matrice complète commune à l'ensemble des distributions gaussiennes issues d'une même partition. Néanmoins, ces étapes sont coûteuses en temps et peuvent introduire des artefacts. Elles ne seront donc pas détaillées dans ce document car elles n'influent pas sur le coeur de nos travaux.

2.4.1 Initialisation de la structure des modèles

L'apprentissage effectué par HTS, à partir de HMM issus de HTK, débute par l'initialisation de la structure des MSD-HSMM.

Cette initialisation s'effectue en deux étapes.

Tout d'abord, afin de pouvoir prendre en compte les valeurs de F0, les modèles utilisés sont des MSD-HMM. Cela implique que lors de la première étape de l'initialisation des modèles, la durée est toujours modélisée par une loi géométrique. Cette étape consiste à estimer les paramètres MSD-HMM en considérant chaque segment sans tenir compte du contexte acoustique. Cette étape s'effectue en trois temps. Tout d'abord, les paramètres des distributions gaussiennes de chaque état, pour un type de coefficients, sont initialisés en utilisant la moyenne et la variance calculées sur l'ensemble des observations associées à ce type de coefficients. Ensuite, une phase de ré-estimation basée sur une segmentation est effectuée par l'algorithme de Viterbi. Enfin, cette étape se conclut par la ré-estimation des paramètres de chaque MSD-HMM via un algorithme Baum-Welch.

Le seconde étape consiste à prendre en compte la durée et à transformer les MSD-HMM en MSD-HSMM. Pour cela, les paramètres de la distribution gaussienne de dimension S, permettant de modéliser la durée de séjour dans les S états du MSD-HSMM, sont initialisés en utilisant les probabilités de transition des modèles issus de l'étape précédente. Les probabilités de transition sont ensuite définies de la manière suivante afin de respecter la topologie de Bakis :

$$a_{ij} = \begin{cases} 1, & j = i+1 \\ 0, & sinon \end{cases} \tag{2.54}$$

De plus, à partir de cette étape et dans la suite du processus, la ré-estimation des paramètres est effectuée en utilisant l'algorithme Baum-Welch à l'échelle de l'énoncé. Pour cela, l'ensemble des MSD-HSMM, correspondant aux segments phonétiques présents dans l'énoncé associé à la séquence d'observation O, sont concaténés et l'algorithme Baum-Welch est appliqué sur la phrase-HMM ainsi obtenue.

2.4.2 Prise en compte des contextes

La seconde phase de la procédure d'apprentissage est la prise en compte des contextes linguistique et prosodique. Cette phase, qui nécessite également deux étapes, débute par la duplication des MSD-HSMM appris à l'étape précédente afin d'obtenir les modèles correspondant aux contextes présents dans le corpus d'apprentissage. Lors de cette étape, le nombre de descripteurs utilisés pour qualifier un segment acoustique est influent. En effet, le nombre de segments utilisés dépend directement du nombre de descripteurs nécessaires à la qualification du segment. En considérant le nombre de descripteurs utilisés dans le jeu de descripteurs standard, le nombre de segments associés à un label en contexte n'est pas suffisant pour obtenir une estimation fiable. Ainsi, la première étape de cette phase permet simplement d'introduire un peu de variabilité dans les modèles en vue de l'étape de partitionnement.

L'étape suivante est donc la construction de l'arbre de décision basé sur le critère MDL. Toutefois, à l'issue de cette étape, les arbres obtenus ne peuvent être considérés comme optimaux. En effet, les modèles utilisés pour déterminer ces arbres ont été mis à jour en n'utilisant que très peu d'observations. Néanmoins, le nombre d'observations associées à chaque partition, peut être considéré comme suffisant pour obtenir une réestimation fiable. Ainsi, après avoir reconstruit les modèles en contexte en se basant sur les distributions issues des premiers arbres, de nouveaux arbres de décision sont construits. Les paramètres sont ensuite ré-estimés pour obtenir les modèles finaux.

Cette phase est centrale pour nos travaux car le choix des descripteurs influe directement sur les modèles obtenus. Comme nous avons pu le voir, à corpus constant, le choix d'un jeu de descripteurs aura deux conséquences. La première conséquence porte sur la pertinence des modèles en contexte réestimés au vu du nombre de segments associés à ces modèles. Plus le jeu de descripteurs sera complexe, moins le nombre de segments associés aux modèles en contexte sera élevé. La seconde conséquence porte sur la complexité des arbres de décision. En effet, plus le jeu de descripteurs sera complexe, plus, dans la limite définie par le critère MDL, le nombre de partition sera élevé. Cela implique que les modèles devraient être plus précis mais le temps de calcul nécessaire à leur estimation sera plus important. Ces deux cas montrent que le choix d'un jeu de descripteurs influence fortement la modélisation effectuée par HTS.

2.5 Processus de synthèse

La génération des paramètres acoustiques repose sur le processus décrit par la figure 2.9. L'objectif est, à partir d'une séquence de descripteurs, d'obtenir la séquence de coefficients acoustiques nécessaires au couple d'outils STRAIGHT et SPTK pour synthétiser le signal de parole. Ce processus s'effectue en trois étapes.

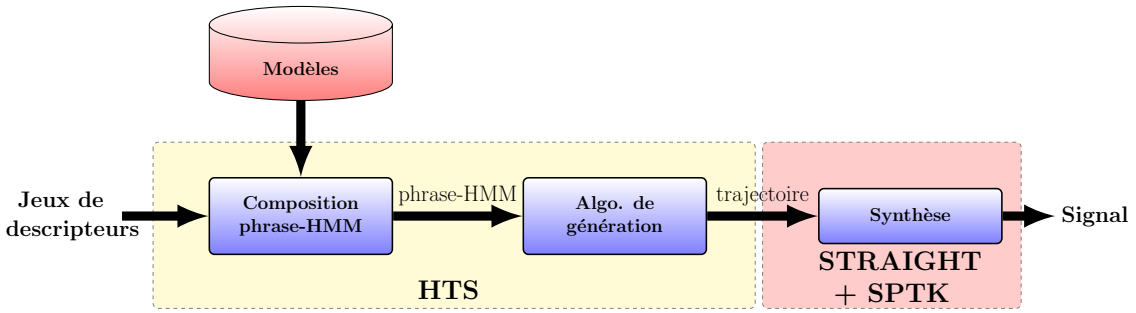


FIGURE 2.9 – Processus de génération des paramètres. Tout d'abord, à partir d'une séquence de descripteurs, une phrase-HMM est constituée à partir de la structure imposée des modèles et des distributions issues des arbres de décision grâce aux descripteurs. Un algorithme de génération est ensuite utilisé pour obtenir la séquence des coefficients acoustiques nécessaires à la génération du signal.

La première étape du processus consiste à obtenir une phrase-HMM associée à l'énoncé que l'on souhaite synthétiser. La constitution de ce modèle repose sur la structure des modèles qui est imposée ainsi que sur les arbres de décision. En effet, après avoir concaténé autant de MSD-HSMM que l'énoncé contient de segments, les paramètres des distributions sont déterminés en parcourant l'arbre de décision associé à chaque état et chaque paramètre. Ce parcours est rendu possible par la qualification de chaque segment par une combinaison de descripteurs compatible avec le jeu de descripteurs utilisé lors de la phase d'apprentissage.

L'algorithme de génération, basée sur la variance globale et permettant d'obtenir la trajectoire pour chacun des paramètres acoustiques, est alors employé. Trois modes de génération sont disponibles pour générer les trajectoires acoustiques. Le premier consiste à supposer que la séquence d'états et de gaussiennes est cachée ; le second consiste à supposer que la séquence d'états est connue mais qu'il reste à déterminer quelles gaussiennes utiliser ; le dernier mode suppose la séquence d'états et de gaussiennes connue. De fait, si les mixtures n'ont qu'une seule composante, le second mode est identique au troisième. Le troisième mode consiste, en pratique, à n'effectuer qu'une seule itération de l'algorithme 1.

La dernière étape consiste à générer la forme d'onde acoustique du signal de parole à partir des trajectoires obtenues précédemment par application du vocodeur STRAIGHT.

2.6 Paramétrisation du corpus et configuration de HTS

Jusqu'à présent, nous avons présenté le système HTS et les concepts utilisés par ce système pour apprendre des modèles en vue d'une synthèse. Nous allons maintenant présenter la configuration utilisée pour réaliser les travaux présentés dans ce document. Cette configuration se décompose en deux parties : la configuration des outils qui permettent de paramétriser le signal et la configuration du système HTS lui-même. Sauf

mention explicite, les configurations que nous avons utilisées sont standard et sont celles de la démonstration proposée par les concepteurs du système HTS et associée à la version 2.1.1 de ce système [Hts211].

2.6.1 Paramétrisation du signal

En pré-requis, il est nécessaire d'obtenir que les signaux soient échantillonnés à 16kHz et ne possèdent qu'un seul canal.

En se basant sur ces signaux, le premier outil utilisé, STRAIGHT (version v40-007-d), permet d'obtenir le F0, le spectre et l'apériodicité. Pour l'ensemble de ces paramètres acoustiques, le décalage de trame utilisé est de 5ms. 60Hz-300Hz a été définie comme plage de valeurs de F0 valides pour STRAIGHT. Cette plage englobe la plage 80Hz-200Hz caractéristique de la voix d'homme (indiquée section 1.1.1 du chapitre 1). La dimension des FFT utilisées par STRAIGHT pour effectuer les extractions est de 512 points.

En réalité le système HTS n'apprend pas les valeurs du F0 sur une échelle linéaire mais logarithmique. La constante $-1e+10$ est utilisée pour représenter $log(0)$ et ainsi permettre la représentation des zones non voisées. De plus, l'apériodicité extraite par STRAIGHT est découpée en cinq bandes de fréquence ($0-63$Hz, $64-127$Hz, $128-255$Hz, $256-383$Hz et $384-512$Hz), une valeur moyenne est calculée pour chacune de ces bandes.

Enfin, les coefficients spectraux obtenus par STRAIGHT sont ensuite convertis en coefficients MGC, d'ordre 39, grâce à la suite logicielle SPTK v3.5 [Sptk]. De plus, comme cela a été indiqué dans la section 1.1.3, la valeur du coefficient α dépend de la fréquence d'échantillonnage. Pour une fréquence d'échantillonnage de 16kHz, le coefficient α a pour valeur 0.42 [Imai1983].

2.6.2 Configuration de HTS

La configuration du système HTS débute par la définition de la topologie des modèles MSD-HSMM. Dans le cadre de nos travaux, les modèles appris sont des MSD-HSMM à 5 états émetteurs. Pour être cohérent avec le pas d'analyse de 5 ms fixé lors du calcul des vecteurs acoustiques, la durée minimale évaluée par un état est de 5 ms.

Lors de la phase d'apprentissage, une seule ré-estimation est effectuée pour chaque phase du processus d'apprentissage. Afin d'optimiser la ré-estimation, la largeur du faisceau parcouru par l'algorithme Forward-backward est contrainte. En début de réestimation, la largeur du faisceau est limitée à 1500 états alignés. Si l'apprentissage ne converge pas, cette taille est augmentée en utilisant un pas de 100 dans la limite de 5000 états alignés sur une trame. Enfin, le seuil de variance est de 0.01 pour l'ensemble des paramètres.

La configuration associée à la phase de génération fait intervenir la variance globale (décrite section 2.2.3). La méthode de *Newton-Raphson* est utilisée pour déterminer les paramètres respectant le critère décrit par l'équation (2.30) avec un facteur de convergence de 10^{-4} et un nombre maximum de 50 itérations. Enfin, comme nous l'avons vu dans ce chapitre, HTS propose trois modes de génération. Le temps de génération entre les modes 1 et 3 diffère fortement. Le mode de génération ne dépendant pas du jeu de descripteurs utilisé, nous avons opté pour utiliser le troisième mode de génération : celui qui suppose une séquence d'états Q connue et qui maximise $P(O|Q, \lambda)$.

2.7 Conclusion

Dans ce chapitre nous avons détaillé les concepts et les processus utilisés par HTS pour pouvoir produire des modèles dans l'optique de générer un signal de parole de synthèse. Nous avons également présenté les algorithmes utilisés lors de la phase de génération.

Au cours de cette présentation, nous avons mis en avant le fait que le choix d'un jeu de descripteurs influe sur l'étape de la prise en compte des contextes linguistiques et prosodiques. Nos travaux portant sur l'influence des descripteurs sur la modélisation effectuée par HTS, la qualité de cette modélisation impacte directement la qualité de la synthèse obtenue. Dans le prochain chapitre, nous allons analyser les différents jeux de descripteurs proposés par différents travaux publiés sur HTS.

Chapitre 3

Système HTS - Jeux de descripteurs

Dans le chapitre précédent, nous avons présenté le système HTS et nous avons vu que ce système nécessite de qualifier un segment par un ensemble de descripteurs linguistique et prosodique. L'objet des travaux présentés dans ce document concerne l'évaluation de l'influence de ces descripteurs sur la qualité de système HTS dans le cadre de la langue française. Pour réaliser cette étude, il est donc nécessaire de disposer d'un jeu de descrip-

teurs spécifique pour le français et, à l'heure actuelle, aucun jeu de descripteurs n'a été publié pour cette langue.

Avant de préciser un jeu de descripteurs pour le français, nous présentons une étude des jeux de descripteurs proposés pour effectuer une synthèse HTS dans des langues diverses. Cette étude a consisté à comparer les différences entre les jeux de descripteurs proposés par rapport au jeu standard [Tokuda2000] défini pour l'anglais. Nous avons complété cette étude par le recensement des descripteurs utilisés dans les modules de prédiction de prosodie et les systèmes de synthèse par sélection pour le français. Grâce à cela, nous avons pu définir un jeu de descripteurs spécifique au français pour le système HTS. Dans la dernière section de ce chapitre, nous présenterons les études proposées pour analyser l'influence des descripteurs sur la synthèse effectuée par HTS.

3.1 Jeu de descripteurs proposé pour l'anglais

Le premier jeu de descripteurs publié concerne l'anglais [Tokuda2002]. Ce jeu de descripteurs a depuis été complété pour obtenir ce que nous identifions comme le jeu de descripteurs standard décrit dans [Zen2009] (Ce jeu de descripteurs est résumé dans l'annexe C.1).

Comme nous l'avons vu dans le premier chapitre, un système de synthèse TTS nécessite deux phases. En se basant sur le jeu de descripteurs proposé dans [Tokuda2002], le système Festival [Taylor1998] est utilisé pour réaliser la première étape : obtenir la séquence de descripteurs associés à un énoncé. Le système HTS est ensuite utilisé pour effectuer la synthèse proprement dite.

Dans cette section, nous allons présenter les propriétés composant le jeu de descripteurs standard et l'influence de l'utilisation de Festival sur ces propriétés.

3.1.1 Description à l'échelle du phonème

La description d'un segment à l'échelle du phonème est constituée, en majeure partie, de la séquence de cinq labels phonétiques (le label peut également indiquer un NSS [1] ou bien une position inconnue [2]) dont l'étiquette centrale de cette séquence est le label phonétique du segment courant. Il s'agit d'une première évolution par rapport au jeu de descripteurs originel car, dans [Tokuda2002], seules trois étiquettes (phonèmes précédent-courant-suivant) étaient utilisées. Les questions utilisées pour construire l'arbre de décision consistent alors à balayer l'ensemble des catégories phonologiques (comme le lieu ou le mode d'articulation par exemple). De plus, les syllabes de l'anglais respectent, en grande

1. Non-Speech-Sound, ce sont des évènements acoustiques qui ne correspondent pas à de la parole, comme une pause, un bruit, une aspiration
2. nécessaire pour associer un label au segment précédent le premier segment par exemple

majorité, la structure CV [3] voir CVC (70% selon Dominguez ET AL. [Dominguez1997]). Ainsi, l'utilisation d'une séquence de cinq phonèmes permet également de couvrir, implicitement, une syllabe.

Si le segment est un phone, deux descripteurs sont ajoutés pour déterminer la position de ce phone par rapport au début et à la fin de la syllabe à laquelle il appartient. Plus généralement, dans le jeu de descripteurs standard, les informations de position d'un élément donné sont définies par rapport au début et à la fin de l'élément du niveau phonologique immédiatement supérieur.

3.1.2 Description à l'échelle de la syllabe

Au niveau de la syllabe, la fenêtre se réduit à trois éléments : syllabe précédente, syllabe courante et syllabe suivante. Trois descripteurs sont alors utilisés pour qualifier ces éléments : deux descripteurs sont liés à l'accentuation et le dernier correspond à la longueur de la syllabe en nombre de phones.

HTS distingue deux types d'accentuation : l'accent lexical (*stressed syllable*) et l'accent tonique (*accented syllable*). D'après le manuel de Festival [Black2002], l'accent lexical est obtenu en utilisant les règles issues d'un lexique et le mot auquel la syllabe est liée. L'accent tonique en revanche résulte de l'acoustique observée. Toujours d'après ce manuel, le système Festival considère que la syllabe possède un accent tonique si celle-ci est liée à un évènement intonatif. Ces évènements sont déterminés par le modèle Tilt [Taylor2000].

La syllabe courante fait l'objet d'une description plus complète. Des informations permettent de situer la syllabe, à laquelle le segment appartient, au sein du mot, de la phrase et de l'énoncé. Ces informations sont complétées par des descripteurs permettant de situer la syllabe en fonction des syllabes accentuées. Par exemple, l'un des descripteurs représente le nombre de syllabes entre la syllabe courante et la prochaine syllabe considérée comme ayant un accent lexical. Enfin, le dernier descripteur, pris en compte pour la syllabe courante, est le label phonétique du noyau de cette syllabe.

3.1.3 Description à l'échelle du mot

La description du segment à l'échelle du mot est constituée de la position du mot au sein de la phrase, de la taille du mot en nombre de syllabes, de l'étiquette grammaticale du mot ainsi que des informations de positionnement par rapport aux mots qualifiés de signifiant.

L'étiquette grammaticale d'un mot est définie par Festival en utilisant un ensemble de règles qui permettent de distinguer les catégories suivantes : les auxiliaires de temps,

3. Consonne-Voyelle

les auxiliaires modaux, les conjonctions, les déterminants, les pronoms personnels, les pronoms interrogatifs, le mot *to*, les prépositions et la ponctuation. Une dernière catégorie a été ajoutée pour prendre en compte les mots signifiants. L'étiquette grammaticale du mot permet de déterminer si un mot est fonctionnel et, s'il ne l'est pas, de spécifier plus précisément sa catégorie grammaticale.

3.1.4 Description à l'échelle de la phrase et à l'échelle de l'énoncé

Une phrase est décrite en fonction du nombre de mots qui la constituent, du nombre de syllabes ainsi que de sa position dans l'énoncé. À ces informations s'ajoute une étiquette TOBI spécifique à la prosodie de fin de phrase.

TOBI (TOnes and Break Indices) est un formalisme standard, proposée par Silverman [Silverman1992] qui s'appuie sur les travaux de J. Pierrehumbert [Pierrehumbert1990], pour annoter symboliquement la prosodie d'un énoncé. Cette annotation se définit selon deux axes. Le premier axe consiste à décrire la courbe du F0 par une séquence de symboles mélodiques. Pour cela, trois symboles sont utilisés : `L` (tonalité basse), `H` (tonalité haute) et `%` (marqueur de début ou de fin d'énoncé). Par exemple, la séquence `L-H%` correspond à une montée de F0 à la fin d'un énoncé. Le second axe consiste a décrire les indices de rupture, les frontières entre mots, avec une échelle allant de 0 (frontière clitique) à 4 (fin de phrase).

HTS n'utilise que l'annotation de courbe mélodique et se restreint aux trois cas suivants : `L-L%`, `L-H%` et `H-H%`. Deux symboles sont ajoutés `NONE` et `0` pour indiquer l'absence d'indicateur, dans le premier cas si le segment est un phone et dans le second si le segment est une pause. Pour obtenir ces informations, le système Festival produit cette étiquette en utilisant le modèle Tilt [Taylor2000].

Enfin, la dernière échelle utilisée pour décrire un segment est l'énoncé. Cette dernière échelle est constituée de trois descripteurs : le nombres de syllabes, de mots et de phrases dans l'énoncé. Dans la version 2.1.1 de HTS [Oura2010], un ensemble de questions liées à ces descripteurs a été défini pour construire un arbre de décision spécifique à la modélisation de la variance globale.

3.2 Jeux de descripteurs proposés pour d'autres langues

Dans [Zen2009], H. Zen ET AL.proposent un recensement des langues pour lesquelles le système HTS a été utilisé pour effectuer une synthèse. Ce recensement constitue un point de départ pour pouvoir analyser les jeux de descripteurs proposés dans le cas de langues autres que l'anglais (un résumé est présenté dans les tableaux de l'annexe C.3). Nous allons analyser ces jeux de descripteurs par rapport au jeu de descripteurs standard afin

de pouvoir isoler les descripteurs communs pouvant être intégrés au jeu proposé pour le français. Cette analyse se fera de l'échelle du phonème à l'échelle de l'énoncé. Les échelles de description non présentes dans le jeu de descripteurs standard seront analysées en fin de section.

3.2.1 Description à l'échelle du phonème

À l'échelle du phonème, la description d'un segment acoustique est très proche de celle proposée dans [Tokuda2002]. Quelques différences subsistent néanmoins.

En effet, pour l'allemand, [Krstulovic2007] propose en plus des descripteurs habituels, le type de position. Ce descripteur permet de déterminer si le phonème est la seule composante de la syllabe ou bien s'il fait parti de l'amorce, du noyau ou de la coda. Pour le basque [Erro2010], un descripteur supplémentaire au jeu standard permet de déterminer la position du phonème entre deux pauses. Des descripteurs peuvent également être enlevés, comme pour l'espagnol [Bonafonte2008], où la position du phonème dans la syllabe n'a pas été intégrée dans le jeu de descripteurs.

En revanche, pour la majorité des langues, seuls l'alphabet et l'horizon (contexte de 5 phones ou bien contexte de 3 phones) varient.

3.2.2 Description à l'échelle de la syllabe

Lors de la présentation du jeu de descripteurs standard, nous avons vu que deux types d'accents étaient pris en compte : l'accent lexical et l'accent tonique. Les descripteurs liés à l'accentuation de la syllabe sont propres à la langue et plusieurs cas se distinguent : les jeux de descripteurs associés à certaines langues ne prennent pas en compte l'accent lexical (comme le suédois [Lundgren2005] par exemple), l'accent tonique (comme le portugais brésilien [Maia2003] par exemple) ou bien les deux (comme le finnois [Silen2008]). Pour d'autres langues, les descripteurs sont adaptés pour l'accent lexical. Dans ce dernier cas, le domaine de valeurs, des descripteurs liés à l'accent lexical, peut être identique au domaine du jeu de descripteur standard mais les outils permettant d'obtenir ces valeurs restent propres à la langue cible.

De plus, pour certaines langues, la syllabe est remplacée par un support plus adapté. Ainsi, par exemple, le japonais repose sur la more [4]. Cette information, bien que pouvant être pertinente pour l'anglais, n'a pas été intégrée au jeu de descripteurs standard. Ceci démontre que la définition d'un jeu de descripteurs consiste avant tout à effectuer un choix parmi un ensemble de propriétés adaptées à la langue cible.

4. La more est une unité phonologique permettant de quantifier la durée de la syllabe. La more peut être calculée sur le noyau et la coda ou bien sur le noyau seul selon les langues. Dans le cas du japonais, une voyelle courte implique que la syllabe est monomoraïc ; une syllabe contenant une diphtongue sera bimoraïc. La coda n'est pas prise en compte dans le calcul pour cette langue.

3.2.3 Description à l'échelle du mot

Comme nous l'avons vu dans la description du jeu de descripteurs standard, la particularité des descripteurs à l'échelle du mot réside dans la notion d'étiquette grammaticale. Cette information permet de séparer deux catégories de mots : les mots fonctionnels de la langue (dont le descripteur *gpos* détaille le type de fonction) et les mots lexicaux.

Certains jeux de descripteurs, comme celui proposé pour le basque [Erro2010], simplifient cette distinction jusqu'à n'obtenir que deux valeurs pour ce descripteur (fonctionnel ou lexical). Néanmoins, deux tendances générales se distinguent : soit l'étiquette grammaticale n'est pas utilisée comme descripteur (c'est le cas du suédois, du mandarin, du portugais européen, du finnois, de l'espagnol, du grec et croate) ; soit les valeurs sont adaptées à la langue (pour les autres langues).

L'étiquette grammaticale constitue un repère important pour l'information de position du mot courant, dans l'énoncé, en fonction du nombre de mots lexicaux. Néanmoins, les descripteurs liés à cette propriété ne sont généralement pas utilisés (seuls le basque [Erro2010], l'allemand [Krstulovic2007] et le portugais brésilien [Maia2003] utilisent cette information).

3.2.4 Description à l'échelle de la phrase et à l'échelle de l'énoncé

L'échelle de la phrase ne contient qu'un descripteur qui n'est pas une information de position : l'étiquette TOBI qui permet de décrire le contour mélodique de la fin de la phrase. Cette étiquette n'est pas utilisée pour la majorité des langues (seul l'allemand [Krstulovic2007] utilise ce descripteur en l'appliquant également à la phrase précédente, la phrase courante et la phrase suivante).

La différence principale entre la majorité des jeux de descripteurs et le jeu de descripteurs standard est l'absence ou non de l'étiquette TOBI. Toutefois, certains jeux de descripteurs introduisent de nouvelles informations. Ainsi le japonais [Oura2011a] introduit un descripteur permettant d'indiquer si la phrase est une phrase interrogative ou non. Ce descripteur joue, dans une moindre mesure, le même rôle que le descripteur TOBI car la courbe intonative évolue différemment selon que la phrase est interrogative ou ne l'est pas.

En ce qui concerne la description d'un segment à l'échelle de l'énoncé, la majorité des langues utilisent les descripteurs standard avec quelques adaptations. Néanmoins ces adaptations restent minimes (le thaïlandais [Chomphan2007] n'utilise pas le nombre de phrases, le japonais [Oura2011a], et l'allemand [Krstulovic2007] prennent en compte des horizons qui leur sont propres). Enfin, le seul descripteur, qui ne soit pas lié à une information de position, à avoir été ajouté est le type d'émotion pour le basque [Erro2010].

3.2.5 Prise en compte de nouvelles échelles de description

Enfin, parmi les jeux de descripteurs qui ont été proposés, de nouvelles échelles ont été introduites. Deux cas de figure se distinguent. Le premier consiste simplement à changer la sémantique d'une échelle de description. Par exemple, pour le japonais [Oura2011a], la phrase est considéré comme un groupe accentuel.

Le second cas consiste à ajouter une nouvelle échelle pour intégrer des descripteurs de nature différente. Parmi l'ensemble des jeux de descripteurs analysés, seuls l'espagnol [Bonafonte2008], le japonais [Oura2011a], l'allemand [Krstulovic2007] et le basque [Erro2010] sont dans ce cas. L'objectif peut être d'introduire des informations d'un niveau qui n'est pas présent dans le jeu de descripteurs standard. C'est le cas, par exemple, du jeu de descripteurs proposé pour l'allemand qui introduit des informations liées à la ponctuation. L'objectif peut être également d'enrichir la description associée à un niveau déjà présent dans le jeu de descripteurs standard. Par exemple, dans le jeu de descripteurs standard, le label phonétique permet également d'identifier si le segment est une pause et de quel type de pause il s'agit. En revanche, le jeu de descripteurs proposé pour le basque introduit un échelle spécifique à la pause. Cette échelle complète le label associé au segment décrit en ajoutant, entre autre, le nombre de segments entre le segment courant et la prochaine pause. Toutefois, parmi les jeux de descripteurs recensés, l'introduction d'une nouvelle échelle n'est pas spécifique à une langue. Ceci implique que l'ensemble des échelles rencontrées peut être utilisé pour n'importe quelle langue cible.

3.2.6 Bilan

À l'issue de l'analyse de ces différents jeux de descripteurs, nous pouvons en tirer les conclusions suivantes. Tout d'abord, le formalisme utilisé pour décrire les segments est souple et suffisamment général pour qu'aucune adaptation du système ne soit nécessaire pour obtenir des modèles propres à une langue.

Dans un second temps, l'influence du jeu de descripteurs standard sur la définition des jeux propres à d'autres langues est indéniable. En effet, pour une grande majorité de langues analysées, les seules adaptations effectuées ont consisté à ignorer des descripteurs ou bien à compléter le jeu par des descripteurs de nature proche, par exemple le nombre de phones par énoncé pour le basque [Erro2010]. Néanmoins, certains nouveaux descripteurs pris en compte intègrent des informations de nature différente par rapport aux descripteurs d'origines. Ceci est le cas, par exemple, de la prise en compte de la fréquence d'apparition de l'unigramme qui a été ajouté à l'échelle du mot (jeu de descripteurs proposé pour l'allemand [Krstulovic2007]).

Enfin, la dernière conclusion importante concerne le nombre de descripteurs utilisés. Ce nombre varie fortement selon les langues allant du triphonème (pour le Croate [Ipsic2006])

à un jeu comportant environ 70 descripteurs (pour l'allemand [Krstulovic2007]).

3.3 Jeux de descripteurs pour le français

Jusqu'à présent, nous avons abordé les jeux de descripteurs pour HTS et les différentes études concernant l'influence de ces jeux sur la synthèse. Lors de notre recensement bibliographique nous n'avons pas rencontré de jeu de descripteurs proposé spécifiquement pour le français. Afin de définir un jeu de descripteurs, pour effectuer une synthèse en français par le système HTS, nous allons recenser les différentes caractéristiques linguistiques et prosodiques utilisées dans les systèmes de prédiction de prosodie spécifiques au français.

3.3.1 Descripteurs utilisés en sélection d'unités et en prédiction de prosodie

Les modules de prédiction de prosodie, qui permettent de prédire la fréquence fondamentale ainsi que la durée pour une séquence d'unités à sélectionner, reposent sur un ensemble de règles linguistiques. Ainsi, pour le système proposé dans les années 1980 par le CNET [Moulines1990], comme pour le système MBROLA [Dutoit1996], la règle principale utilisée pour la prédiction du F0 concerne l'accentuation de la syllabe. À cette propriété s'ajoute l'information de position du patron d'intonation [5] dans la phrase pour le système MBROLA[Malfrere1998] et l'information de position de la consonne dans le mot pour le système proposé par le CNET [Sorin1984, sorin1987].

Pour ces mêmes systèmes de synthèse par sélection d'unités, la durée fait l'objet d'une description plus complète. En effet, des informations de position ont été prises en compte sur diverses échelles (comme la position du phonème dans la syllabe ou bien dans le mot par exemple). À cela s'ajoute la spécification du contexte phonémique ainsi que la définition du type de syllabe. Enfin, les segments sont également décrits relativement aux pauses. Ainsi, le système proposé par le CNET [Moulines1990] utilise également la position par rapport à la pause, la longueur des pauses ainsi que le type de frontières [Bartkova1987, sorin1987].

Plus récemment, des systèmes de synthèse par sélection ont été développés pour le français. Généralement, le coût cible utilisé par ces systèmes est basé sur des descripteurs linguistiques. Ceci est le cas pour le système proposé par le LIMSI [Prudon2002] qui intègre des informations de position du phone dans le mot et la syllabe. Pour ce système, le seul descripteur non lié au phone est la propriété de signifiance du mot.

L'ensemble des descripteurs, présentés dans cette section, est donc proche du jeu standard proposé par les concepteurs du système HTS. De plus, bien que certaines propriétés

5. Pour le système MBROLA, un patron d'annotation correspond à une clef qui contient les descripteurs et un ensemble de courbes de F_0 cibles

ne soient pas explicitement présentes dans le jeu de descripteur HTS, ils peuvent être déduits d'autres descripteurs. Par exemple, la position de la dernière syllabe accentuée peut se rapprocher du descripteur indiquant le nombre de syllabes entre la syllabe courante et la dernière syllabe considérée comme accentuée.

3.3.2 Jeu de descripteurs proposé

En se basant sur les descripteurs décrits précédemment, nous avons défini un jeu de descripteurs pour pouvoir effectuer une synthèse en langue française via le système HTS. Ce jeu, comme la majorité de ceux proposés pour les autres langues, reste proche du jeu de descripteurs standard. L'annexe C présente en détail ce jeu de descripteurs.

Le premier changement apporté par rapport au jeu de descripteurs standard concerne l'accentuation au niveau de la syllabe. Dans le cadre de nos travaux, l'accent lexical n'a pas été pris en compte. De plus, comme pour la plupart des jeux de descripteurs, l'étiquette TOBI a également été ignorée.

Dans un second temps, d'autres descripteurs ont été adaptés. Tout d'abord, la prédiction de la proéminence d'une syllabe, que nous associons à l'accent tonique de cette syllabe, a été effectuée en utilisant les règles décrites dans [Simon2008]. Ensuite, à l'échelle du mot, la propriété de signifiance est déterminée en se basant sur les règles décrites par le tableau 3.1 sachant qu'un mot clitique est considéré comme un mot non-signifiant.

Enfin, un dernier descripteur a été ajouté. Ce descripteur est spécifique au corpus que nous utilisons et qui sera décrit dans le chapitre suivant. En effet, lors de nos expériences, nous avons constaté que certaines voyelles ne possédaient aucune trame considérée comme voisée. Cet état étant incohérent avec la définition d'une voyelle, nous avons introduit un descripteur qui permet d'identifier ces segments et ainsi de les isoler en utilisant l'arbre de décision. Lors de la phase de synthèse, pour chaque voyelle, nous imposons la valeur 1 à ce descripteur pour indiquer au système HTS de sélectionner uniquement les modèles appris sur des voyelles considérées comme voisées.

Ainsi, à l'issue de cette définition, nous obtenons un ensemble de 44 descripteurs, décrits dans l'annexe C.2, pour qualifier le contexte d'un segment acoustique.

3.4 Évaluation des jeux de descripteurs sur la synthèse HTS

Les conclusions sur la comparaison des jeux de descripteurs, effectuée dans la section 3.2, posent la question de l'influence des descripteurs utilisés sur la modélisation effectuée par HTS. En effet, en considérant des extrêmes, peut-on supposer que le jeu de descripteurs utilisé pour l'allemand [Krstulovic2007] est plus pertinent que le jeu de

Cat. gram.	Sous-cat.	Acc.	Distinction
Verbe		NC	
Nom		NC	
Adjectif		NC	
Adverbe	sauf *ne*	NC	
Numéral		NC	
Mot-phrase		NC	
Pronom	personnel	C	je, tu, on, le, la, me, te, se, ...
		NC	moi, toi, eux, soi
		?	nous, vous, elle, elles, lui, leur, ...
	possessif	NC	
	indéfini	NC	
	interrogatif	NC	
	relatif	C	que
		NC	qui
	démonstratif	C	ce, c'
		NC	ceci, ça, ...
Déterminant	article	C	
	interrogatif	C	
	possessif	C	
	démonstratif	C	
	indéfini	NC	
	prédéterminant	NC	
Conjonction		?	
Préposition		?	

TABLE 3.1 – Association entre une étiquette grammaticale et le caractère clitique d'un mot auquel est associée cette étiquette. (C=clitique, NC=non-clitique, ?=peut contenir des syllabes accentuées sous certaines conditions). Ce tableau est extrait de [Mertens2001].

descripteurs utilisé pour le croate [Ipsic2006] ?

À l'heure actuelle nous n'avons identifié que deux études qui analysent l'influence des descripteurs sur la synthèse effectuée par HTS. La première a été proposée par O. Watts ET AL. [Watts2010] et se focalise sur les descripteurs prosodiques. La seconde, proposée par S. Yokomizo ET AL. [Yokomizo2010], a pour objectif la définition d'un jeu de descripteurs minimal.

3.4.1 Étude des descripteurs prosodiques

En 2010, O. Watts ET AL. [Watts2010] ont proposé une étude visant à évaluer l'impact des descripteurs prosodiques sur la modélisation effectuée par HTS. Ces descripteurs, qui peuvent être définis manuellement ou automatiquement, sont liés à : l'accent tonique de la syllabe, l'étiquette grammaticale du mot et la courbe mélodique de fin de phrase. L'objectif de cette étude est de déterminer si la modélisation, effectuée par HTS, est impactée selon qu'on utilise une information obtenue manuellement ou automatiquement.

Apprentissage Synthèse	Manuel Manuel Manuel (G)	Manuel Auto Mixte (M)	Auto Auto Auto (A)
Descripteurs			
Lex. POS Phrase TOBI	G1	M1	A1
Lex. POS Phrase	G2	M2	A2
Lex. POS	G3	M3	A3
Lex.	G4	M4	A4

TABLE 3.2 – Systèmes analysés dans l'étude présentée dans [Watts2010]

Pour atteindre cet objectif, les auteurs ont appris les modèles en utilisant 12 configurations différentes qui sont présentées dans le tableau 3.2. Comme le montre ce tableau, l'ensemble des descripteurs peut se lire selon deux axes. Le premier consiste à regrouper les configurations selon le mode de définition des labels. Dans ce cas, trois ensembles se distinguent : les labels sont obtenus manuellement (GX) ; les labels sont obtenus automatiquement (AX) ; les labels d'apprentissage sont obtenus manuellement et les labels utilisés pour la phase de synthèse sont obtenus de manière automatique (MX). Le second axe consiste à regrouper les configurations selon les catégories utilisées pour définir les labels. Pour cela, quatre ensembles ont été proposés en partant du label complet (étiqueté sous la forme X1) jusqu'à celui ne contenant que les informations concernant le phonème et les accents lexicaux (étiqueté sous la forme X4). Pour passer de l'ensemble complet à l'ensemble contenant le moins d'information, il faut ignorer, successivement, les descripteurs liés à l'accentuation tonique et à l'étiquette TOBI, les descripteurs liés aux frontières de phrases et puis ceux liés à l'étiquette grammaticale.

En découpant les configurations selon ces deux axes, les auteurs peuvent comparer l'influence des descripteurs prosodiques ainsi que l'influence de l'utilisation des annotations, obtenues par un processus automatique, sur la modélisation effectuée par HTS. Pour pouvoir déterminer ces influences, deux méthodes ont été proposées. En considérant les arbres de décision associés au F0, la première méthode consiste à comparer visuellement [6] les taux d'utilisation de chaque catégorie linguistique et prosodique auxquelles appartiennent les nœuds composant ces arbres. Cette méthode permet donc d'illustrer l'importance accordée à une catégorie par le système HTS. La seconde méthode consiste à effectuer une série de tests subjectifs (de type AB) afin de comparer différentes combinaisons de systèmes.

En comparant les proportions d'utilisation des nœuds, les auteurs aboutissent à deux conclusions. Tout d'abord, les labels utilisés lors de la phase d'apprentissage sont plus influents que ceux utilisés lors de la phase de synthèse sur le taux d'utilisation d'un nœud. Cette conclusion se base sur le constat que les proportions associées aux systèmes identifiés par M et G sont globalement proches comparés aux systèmes identifiés par A. La seconde conclusion concerne la compensation entre les catégories de descripteurs. En effet, si une catégorie de descripteurs n'est pas utilisée, les descripteurs d'une autre catégorie tendent à avoir plus d'importance (par exemple l'augmentation de l'utilisation des descripteurs

6. Un niveau de gris est associé à une catégorie en fonction de son taux d'utilisation

liés à l'étiquette grammaticale lorsque les descripteurs d'informations liées à la phrase ne sont plus présents).

En utilisant les évaluations subjectives, les auteurs ont pu comparer l'ensemble des synthèses effectuées en utilisant des labels manuels (G1 à G4) puis les synthèses effectuées en faisant varier le type des labels utilisés (G1, M1 et A1). Les résultats obtenus montrent que G1 est préféré à G4 et qu'il s'agit de la seule différence significative. Cela permet aux auteurs de formuler l'hypothèse qu'ignorer des descripteurs linguistiques aura plus d'impact sur la modélisation effectuée par HTS si ces descripteurs ont été déterminés manuellement plutôt qu'automatiquement.

3.4.2 Définition d'un jeu de descripteur minimal

La seconde étude que nous avons recensée a été proposée par S. Yokomizo ET AL. [Yokomizo2010] et a pour objectif la réduction du nombre de descripteurs utilisés pour caractériser un segment. Le constat, sur lequel se basent les auteurs, est que l'utilisation d'une cinquantaine de descripteurs nécessite un temps de calcul élevé lors de la phase de construction de l'arbre de décision. Ainsi, l'objectif des auteurs est de déterminer un jeu de descripteurs minimal afin de réduire le temps d'apprentissage des modèles.

Afin de réaliser cette étude, les auteurs ont sélectionné deux corpus : le corpus CMU ARCTIC [Kominek2003], composé de six locuteurs non-professionnels (quatre hommes et deux femmes), pour l'anglais et le corpus ATR [Kurematsu1990], composé de dix locuteurs professionnels (six hommes et quatre femmes), pour le japonais. Les auteurs ont ensuite décomposé le jeu de descripteurs standard et le jeu associé au japonais tel que cela est décrit, respectivement, dans les tableaux 3.3 et 3.4.

Pour les deux langues, trois paramètres acoustiques ont été évalués :

- Les coefficients MGC en utilisant une distance mel-cepstrale ;
- Les valeurs de F0 en utilisant une erreur RMS [7] ;
- La durée en utilisant également une erreur RMS.

Pour l'ensemble de ces mesures et pour chaque corpus, le résultat correspond à la moyenne des scores obtenus pour l'ensemble des locuteurs.

Tout d'abord, pour l'anglais, les auteurs ont défini un jeu de descripteur optimal en comparant les valeurs de RMS obtenues pour chacun des descripteurs. Les descripteurs, utilisés pour définir ce jeu, sont indiqués dans la seconde colonne du tableau 3.3. Grâce aux mesures objectives, les auteurs montrent que les jeux de descripteurs optimal, minimal (constitué unique de la séquence de cinq phonèmes) et complet aboutissent à des coefficients générés très proches. Ce constat est accentué par le résultat de l'étude sub-

7. Root Mean Square

Éch.	Sel. ?	description
Phonème		Position du phonème dans la syllabe (début, fin)
Syllabe		Syllabe (préc., cour., suiv.) avec accent lexical ?
		Syllabe (préc., cour., suiv.) avec accent tonique ?
		Nombre de phonèmes de la syllabe (préc., cour., suiv.)
		Position de la syllabe dans le mot (début, fin)
		Position de la syllabe dans la phrase (début, fin)
	X	Nombre de syllabes avec accent lexical (avant, après) la syllabe dans la phrase
	X	Nombre de syllabes avec accent tonique (avant, après) la syllabe dans la phrase
	X	Nombre de syllabes entre la (dernière, courante) syllabe avec accent lexical auto-f jusque la (courante, prochaine)(avant, après) la syllabe avec accent lexical
	X	Nombre de syllabes entre la (denière, courante) syllabe avec accent tonique jusque la (courante, prochaine)(avant, après) la syllabe avec accent tonique
	X	Voyelle de la syllabe
Mot		L'étiquette grammaticale du mot (préc., cour., suiv.)
	X	Nombre de syllabes dans le mot (préc., cour., suiv.)
	X	Position du mot dans la phrase (début, fin)
	X	Nombres de mots signifiants (avant, après) le mots dans la phrase
	X	Nombres de mots entre le (dernier, courant) mot signifiant jusqu'au (prochain courant) mot signifiant
Phrase	X	Nombre de syllabes dans la phrase (préc., cour., suiv.)
	X	Nombre de mots dans la phrase (préc., cour., suiv.)
	X	Position de la phrase dans l'énoncé (début, fin)
	X	Étiquette TOBI de fin de phrase
Énoncé	X	Nombre de syllabes dans l'énoncé
		Nombre de mots dans l'énoncé
		Nombre de phrases dans l'énoncé

TABLE 3.3 – Décomposition du jeu de descripteur standard. La seconde colonne permet d'identifier les descripteurs utilisés dans le jeu optimal proposé pour l'anglais. Tableau extrait de [Yokomizo2010]

Identifiant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Accentuation									X	X	X	X	X	X	X	X
Éti. Gram.					X	X	X	X					X	X	X	X
Syntagme			X	X			X	X			X	X			X	X
Taille(phrase)		X		X		X		X		X		X		X		X

TABLE 3.4 – Les différentes combinaisons analysées pour le jeu de descripteurs pour le japonais. La colonne 10 correspond à la combinaison optimal. Tableau extrait de [Yokomizo2010]

jective. En effet, un test de préférence de type ABX, entre le jeu de descripteurs optimal et le jeu de descripteurs complet, a été effectué. Les résultats obtenus montrent que les synthèses sont considérées équivalentes malgré le nombre de descripteurs plus élevé dans le jeu complet.

L'analyse effectuée pour le jeu de descripteurs japonais est semblable à celle décrite

précédemment. En revanche, plutôt que de calculer une distance associée à un descripteur, l'analyse effectuée se base sur la combinaison de descripteurs. Le tableau 3.4 présente l'ensemble des combinaisons analysées. En comparant les distances obtenues pour chacune des combinaisons, les auteurs ont déterminé que le jeu de descripteur optimal correspond à la combinaison 10. En effet, l'information d'accentuation réduit fortement la dégradation de modélisation du f0 (la RMS chute d'environ 100 cent). De plus, les auteurs notent que les résultats obtenus pour la durée dépendent fortement du locuteur sans indiquer plus d'informations. De même que pour l'anglais, les résultats obtenus, à l'issue de la comparaison entre le jeu de descripteurs optimal et le jeu de descripteurs complet, restent proches.

Enfin, en comparant les temps de calcul obtenus entre les jeux de descripteurs optimal et complet, une diminution d'environ 30% est constatée, et ceci pour les deux langues. Les auteurs concluent donc qu'il est possible, en utilisant le jeu de descripteurs simplifiés, de réduire fortement le temps de calcul tout en conservant une qualité de modélisation constante.

3.4.3 Bilan et positionnement

Les études que nous avons présentées dans cette section ont pour objet l'évaluation de l'impact des descripteurs sur la modélisation. Néanmoins, ces deux études ne se focalisent pas sur les mêmes buts. Tout d'abord, l'étude proposée par S. Yokimizo ET AL. [Yokomizo2010] a pour objectif la définition d'un jeu de descripteurs minimal. Ensuite, l'étude proposée par O. Watts ET AL. [Watts2010] a pour objectif de déterminer l'influence de la qualité de l'annotation (automatique par rapport à manuelle) sur la synthèse effectuée par HTS.

Nos travaux ont pour objectif de déterminer en quoi les descripteurs influencent les modèles. Ainsi, bien qu'étant proche de ce que nous souhaitons faire, les études présentées précédemment n'ont pas le même but. Toutefois, nous avons considérés les deux études précédentes comme des références auxquelles nous devions relier nos résultats.

3.5 Conclusion

Dans ce chapitre, nous avons présenté le jeu de descripteurs standard utilisé pour effectuer une synthèse en utilisant le système HTS. Ce jeu étant spécifique à l'anglais, nous avons ensuite effectué un recensement de jeux de descripteurs utilisés pour d'autres langues. En confrontant ces jeux de descripteurs, nous avons déterminé un ensemble de propriétés communes pour l'ensemble des langues.

Aucun jeu de descripteurs n'ayant été publié pour effectuer une synthèse en français via

le système HTS, nous avons ensuite examiner les descripteurs proposés pour les systèmes de prédiction de prosodie et les systèmes de synthèse par corpus dans cette langue. En se basant sur l'ensemble des descripteurs présentés pour HTS et pour le français, nous avons pu définir un jeu de descripteurs pour effectuer une synthèse HTS en français. Ce jeu de descripteurs reste, comme pour la majorité des langues, très proche du jeu de descripteurs standard.

Enfin, la dernière section de ce chapitre présentaient les études publiées qui ont pour objet l'évaluation de l'impact des descripteurs sur la modélisation effectuée par HTS. Cette section nous a également permis de situer la problématique des travaux présentés dans ce document par rapport à ces études.

Conclusion de la première partie

Dans cette première partie, nous avons exposé le cadre dans lequel se situent les travaux présentés dans ce document. Pour cela, nous avons procédé en trois temps.

Dans le premier chapitre, nous avons présenté le phénomène de la parole puis la synthèse de la parole à partir du texte en nous focalisant sur les deux méthodes les plus importantes dans ce domaine à l'heure actuelle. Ce chapitre a permis d'introduire les notions utilisées pour définir un jeu de descripteurs pour le français, ceci ayant été effectué dans le chapitre 3. Ce chapitre nous a également permis de situer le système HTS dans le domaine de la synthèse TTS.

Le second chapitre est consacré à la présentation des concepts utilisés par HTS pour modéliser et produire un signal de parole. Dans ce chapitre, nous avons mis en avant les étapes du processus d'apprentissage (définition des modèles en contexte et construction de l'arbre de décision) où le choix de descripteurs influe. Ces étapes sont donc centrales à nos travaux dont l'objectif est d'évaluer l'influence d'un descripteur sur la synthèse, et donc la modélisation effectuée par HTS. À la fin de ce chapitre, nous avons également présenté la configuration du système HTS et des outils d'extraction (STRAIGHT et SPTK) que nous avons utilisés pour effectuer les expériences décrites dans ce document.

Le dernier chapitre s'est focalisé sur la définition d'un jeu de descripteurs adapté au français en vue d'effectuer une synthèse grâce au système HTS. Pour cela, nous avons procédé en trois étapes. Nous avons tout d'abord introduit le jeu de descripteurs standard qui est considéré comme la référence. Nous avons ensuite comparé les jeux de descripteurs d'autres langues avec cette référence. Lors de cette comparaison, nous avons pu déterminer les descripteurs les plus utilisés et devant être intégrés au jeu de descripteurs proposé pour le français. Enfin, en complétant notre analyse par celle des descripteurs proposés pour les modules de prédiction de prosodie et les systèmes de synthèse par sélection adaptés au français, nous avons déterminé un jeu de descripteurs pour cette langue. Ce chapitre s'est clos par la présentation de deux études dédiées à l'influence des descripteurs sur la synthèse effectuée par HTS. Toutefois, lors de cette présentation, il a été montré que l'objectif de ces études concernent l'influence des outils utilisés pour obtenir les descripteurs ou bien l'influence des descripteurs sur la qualité de la synthèse plutôt que l'analyse de l'influence des descripteurs sur les modèles. Ces études restent néanmoins des références auxquelles

nous devons nous comparer.

Dans la partie suivante, nous présenterons les protocoles que nous avons mis au point pour pouvoir compléter ces études et effectuer une analyse complète de l'influence des descripteurs sur la synthèse. Nous présenterons également les données expérimentales que nous avons utilisées.

Deuxième partie

Évaluation HTS
-
Méthodologie et données expérimentales

Introduction à la deuxième partie

Dans la première partie de ce document, nous avons présenté le système HTS et nous avons exposé la problématique des travaux de thèse : comprendre l'influence des descripteurs sur la modélisation HTS, et par conséquence sur la synthèse. De plus, nous avons focalisé nos travaux sur l'étude des descripteurs pour le français et présenté un jeu de descripteurs pour notre étude.

Afin de répondre à cette problématique, nous avons mis en place deux protocoles expérimentaux et nous avons constitué un corpus en nous basant sur un processus automatique. L'objectif de la seconde partie, composée de deux chapitres, est de présenter ces protocoles et les données utilisées pour réaliser les expériences. Pour cela, la seconde partie se décompose en deux chapitres.

Le premier chapitre (intitulé « Méthodologie d'évaluation », page 75) présente deux protocoles mis en place pour évaluer la modélisation effectuée par HTS. Le premier protocole consiste à évaluer l'espace acoustique généré par HTS en modélisant cet espace par un GMM. Le second protocole consiste à calculer une « distance » entre les coefficients extraits du signal naturel et ceux générés par HTS pour chaque trame. Ces écarts sont ensuite combinés pour déterminer une valeur représentant la dégradation obtenue à l'issue de la génération effectuée par HTS par rapport aux données extraites du signal naturel. En guidant cette combinaison nous pouvons déterminer cette dégradation pour des points précis et ainsi évaluer de manière plus fine la modélisation effectuée par le système HTS.

Le second chapitre de cette partie (intitulé « Données expérimentales », page 89) présente les données utilisées pour réaliser les expériences. Ce chapitre se découpe en quatre points. Tout d'abord, la structure de données, utilisée pour représenter le corpus et organiser les annotations, et le processus d'annotation automatique utilisé pour obtenir ces annotations sont présentés. La section suivante présente le corpus ainsi obtenu en détail. Enfin, la dernière section du chapitre présente les jeux de descripteurs évalués et permet de poser les conditions d'évaluation.

Chapitre 4

Méthodologie d'évaluation

Le choix d'un jeu de descripteurs impacte la modélisation effectuée par le système HTS et plus spécifiquement sur les étapes de prise en compte du contexte et du partitionnement. Afin d'évaluer l'influence des descripteurs sur la modélisation effectuée par HTS, et donc sur ces étapes plus spécifiquement, nous avons mis en place deux protocoles d'évaluation qui sont présentés dans ce chapitre.

Le premier protocole, présenté dans une première section, a pour objectif d'évaluer l'espace acoustique généré par HTS en représentant cet espace par un GMM [1]. Le second protocole complète cette première évaluation en permettant de pouvoir focaliser l'analyse sur des points précis. Pour permettre cela, ce second protocole repose sur le calcul d'écarts entre les vecteurs acoustiques générés par HTS et ceux extraits du signal naturel pour l'ensemble des trames alignées. L'application de ce second protocole est possible car HTS permet d'effectuer une génération en imposant la durée des modèles.

1. Gaussian Mixture Model, autrement appelé mélange gaussien

4.1 Évaluation par GMM

Au milieu des années 1990, les GMM ont été introduits dans le domaine de l'identification du locuteur par Douglas Reynolds [Reynolds1995] pour modéliser l'espace acoustique propre à un locuteur. La méthode présentée dans [Reynolds1995] consiste à calculer la log-vraisemblance d'un ensemble de trames pour le GMM associé à un locuteur donné et à déterminer si ces trames ont été émises par ce locuteur ou non. Plus tard, Yannis Stylianou [Stylianou1998] adapte cette méthode pour la conversion de voix. Les GMM y sont utilisés pour modéliser les enveloppes spectrales des locuteurs sources et cibles. Une fonction de transformation permet alors de passer d'un GMM à un autre.

En faisant l'hypothèse que les GMM capturent convenablement l'espace acoustique associé à un locuteur, nous pouvons adapter la méthode proposée dans [Reynolds1995] pour évaluer l'influence des descripteurs sur la modélisation d'un type de coefficients (MGC, BAP, F0, durée) effectué par HTS. Chaque GMM sera caractéristique d'une voix et nous chercherons à mesurer la proximité entre une voix HTS et une voix naturelle.

Ainsi, contrairement à l'utilisation classique d'un modèle GMM, nous allons faire varier les modèles. En effet, chaque GMM modélise l'espace associé à un paramètre acoustique qui a été produit par HTS selon un jeu de descripteurs de donné. Pour comparer ces modèles, nous utilisons un corpus de vecteurs acoustiques, correspondant au paramètre acoustique évalué, constant. Ces vecteurs acoustiques sont extraits du signal naturel. En déterminant la log-vraisemblance de ces vecteurs sur les modèles, nous pouvons mesurer la qualité du GMM sur les données extraites du signal naturel. Puisque nous faisons l'hypothèse que l'espace acoustique est convenablement capturer par GMM, ce protocole permet donc d'évaluer la qualité de l'espace acoustique produit par le système HTS.

Le protocole d'évaluation proposé nécessite trois étapes. La première étape, intitulée *Préparation des données*, consiste à obtenir les coefficients que l'on souhaite évaluer. La seconde étape, intitulée *Apprentissage des modèles*, utilise ces coefficients pour apprendre un modèle GMM permettant de représenter l'espace des vecteurs acoustiques générés. La dernière étape, intitulée *Évaluation des modèles*, consiste à utiliser un ensemble de vecteurs acoustiques extraits du signal naturel pour évaluer le GMM. Par l'hypothèse que nous avons posé, nous supposons qu'évaluer le GMM HTS est équivalent à évaluer l'espace de vecteurs acoustiques générés par HTS et donc en première approximation la qualité de modélisation segmentale d'une parole de synthèse effectuée par HTS.

4.1.1 Préparation des données

La première étape consiste à obtenir les données indispensables à l'évaluation. Pour cela, il est nécessaire de disposer de 3 corpus d'énoncés disjoints :

– le corpus d'apprentissage A

– le corpus de validation V

– le corpus de test T

L'ensemble des énoncés de ces corpus sont convertis en labels HTS selon le jeu de descripteurs k.

À chacun de ces corpus d'énoncés est associé un corpus de vecteurs acoustiques extraits du signal naturel en utilisant les outils STRAIGHT et SPTK et en se basant sur la configuration du système HTS présenté dans la section 2.6.2 du chapitre 2. Ces corpus de coefficients sont identifiés par :

– le corpus d'apprentissage $A_{a/s}$

– le corpus de validation $V_{a/s}$

– le corpus de test $T_{a/s}$

On note a/s le procédé d'analyse/synthèse sur la voix naturelle opéré par STRAIGHT.

La dernière partie de cette étape consiste à générer, à l'aide du système HTS calibré selon le jeu de descripteurs k, trois nouveaux corpus de coefficients correspondant respectivement aux énoncés des corpus A, V et T :

– le corpus d'apprentissage A_k

– le corpus de validation V_k

– le corpus de test T_k

La figure 4.1 résume la phase de préparation de données.

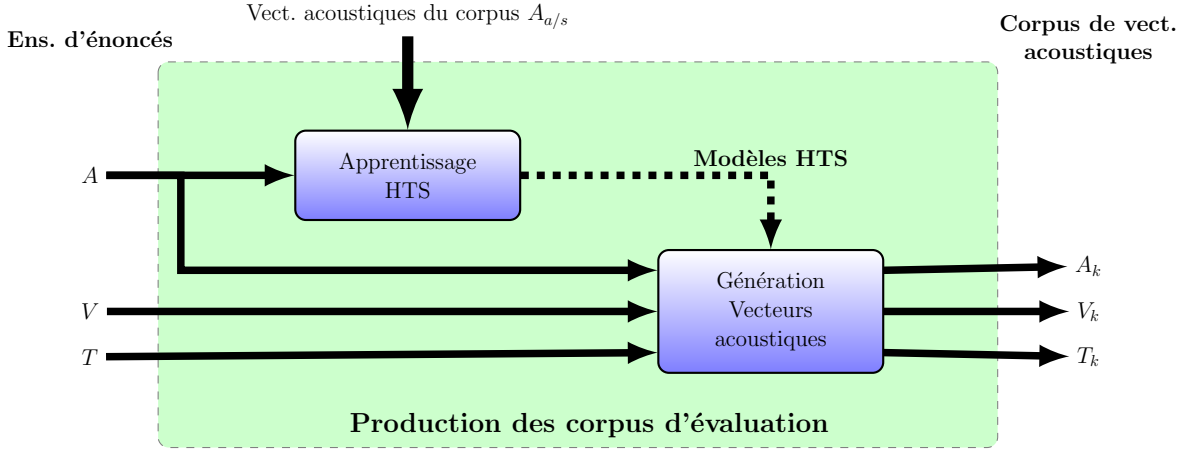


FIGURE 4.1 – Étape 1 - Préparation des données pour le protocole d'évaluation basé sur les GMM. En utilisant les énoncés (étiquetés selon le jeu de descripteurs k) issus du corpus A et les vecteurs acoustiques du corpus $A_{a/s}$, les modèles HTS sont appris. En utilisant ces modèles et les énoncés des corpus A, V et T, les corpus de vecteurs acoustiques A_k, V_k et T_k sont produits.

À l'issue de cette étape, nous obtenons un ensemble de 6 corpus de vecteurs acoustiques qui seront utilisés dans la suite du protocole pour effectuer l'évaluation. Les corpus A_k et

V_k sont utilisés dans la seconde étape pour apprendre les GMM et le corpus $T_{a/s}$ est utilisé dans la dernière étape comme corpus de test. Ce corpus va permettre de déterminer quelle est la vraisemblance du GMM associée au jeu de descripteurs k et, par extension, quelle est la pertinence de la modélisation effectuée par HTS en utilisant ce jeu de descripteurs.

En utilisant maintenant les corpus $A_{a/s}$ et $V_{a/s}$, nous pouvons apprendre un GMM modélisant l'espace acoustique des données extraites du signal naturel. La log-vraisemblance du corpus $T_{a/s}$ sur ce GMM correspond donc à un optimal et nous permet d'obtenir une référence à laquelle la dégradation peut être comparée.

De plus, bien qu'il ne soit pas nécessaire pour le déroulement du protocole, le corpus T_k est utile pour valider un GMM. En effet, si les log-vraisemblances des corpus T_k, A_k et V_k diffèrent significativement c'est parce que le GMM, représentant l'espace des vecteurs acoustiques générés par HTS selon le jeu de descripteurs k, n'est pas conforme à cet espace.

4.1.2 Apprentissage des modèles GMM

En utilisant les corpus A_k et V_k, l'objectif de cette étape est d'apprendre un modèle GMM, noté $\mathcal{M}_k$, représentant l'espace des coefficients générés par HTS selon le jeu de descripteurs k. $\mathcal{M}_k$ est caractérisé par un ensemble de N_k distributions gaussiennes et chaque composante $n \in [1..N_k]$ est caractérisée par le triplet $(\mu_n, \Sigma_n, \omega_n)$ où

- μ_n représente la moyenne de la distribution n
- Σ_n représente la matrice de covariance de la distribution n
- ω_n le poids associé à la distribution n.

L'objectif de cette étape est donc de déterminer ce triplet pour chacune des N_k distributions. Ainsi, cet apprentissage est effectué grâce au corpus A_k en utilisant un algorithme de type E.M. (critère du maximum de vraisemblance).

L'utilisation de cet algorithme suppose que le nombre N_k de composantes est fixé. Pour modéliser au plus près l'espace des vecteurs acoustiques générés par HTS, nous souhaitons calibrer, de manière automatique, ce nombre de composantes. Pour cela, nous allons utiliser le corpus de validation V_k afin de mettre en place une procédure itérative, résumée dans l'algorithme 3, basée sur la détection d'une situation de sur-apprentissage.

Tout d'abord, nous définissons une situation de sur-apprentissage par la relation suivante :

$$LL(V_k; \mathcal{M}_k(N_k)) << LL(A_k; \mathcal{M}_k(N_k)) \tag{4.1}$$

où $\mathcal{M}_k(N_k)$ désigne le modèle $\mathcal{M}_k$ à N_k composantes et $LL(V_k; \mathcal{M}_k(N_k))$ la log-vraisemblance des données V_k pour le modèle $\mathcal{M}_k(N_k)$. Ainsi $\mathcal{M}_k$ est appris sur A_k mais son nombre de composantes est déterminé en utilisant le corpus V_k.

Afin de pouvoir automatiser cette détection, une marge de tolérance ϵ est introduite. Ainsi, la relation (4.1) est réécrite de la manière suivante :

$$\Big[LL_{min}(A_k; \mathcal{M}_k(N_k)) - LL_{max}(V_k; \mathcal{M}_k(N_k))\Big] > \epsilon \tag{4.2}$$

Le nombre de vecteurs acoustiques présents dans V_k permet de calculer un intervalle de confiance. Ainsi, $LL_{min}(A_k; \mathcal{M}_k(N_k))$ correspond à la borne basse de l'intervalle de confiance pour la log-vraisemblance du corpus A_k sur le modèle $\mathcal{M}_k(N_k)$ et $LL_{max}(V_k; \mathcal{M}_k(N_k))$ correspond à la borne haute de l'intervalle de confiance pour la log-vraisemblance du corpus V_k sur ce même modèle.

Pour obtenir le nombre optimal de composantes $N_k^\star$, les GMM $\mathcal{M}_k(N_k)$ sont appris en considérant N_k de la forme 2^i. i est une variable qui est incrémentée progressivement tant que la condition (4.2) n'est pas satisfaite. $N_k^\star$ correspond à la valeur maximum N_k ne validant pas (4.2). La marge de tolérance ϵ constitue un paramètre de la méthode et doit être défini.

Données : un corpus de coefficients A_k, un corpus de coefficients V_k et une marge ϵ
Résultat : Le GMM $\mathcal{M}_k$ appris et le nombre de composantes $N_k^\star$

i=1;
répéter
  Apprendre $\mathcal{M}_k(2^i)$ en utilisant A_k par l'algorithme E.M.;
  Déterminer $LL(A_k; \mathcal{M}_k(2^i))$ et $LL(V_k; \mathcal{M}_k(2^i))$;
  i = i + 1;
jusqu'à *(critère 4.2 validé)*;

$(N_k^\star) = 2^{i-1}$;
$\mathcal{M}_k = \mathcal{M}_k(2^{i-1})$;

Algorithme 3: Algorithme d'apprentissage du modèle $\mathcal{M}_k$

De plus, il peut arriver que la dimension des données implique, lors de la phase d'apprentissage, des problèmes de stabilité numérique. Ces problèmes se traduisent généralement par une incapacité à inverser les matrices de covariance lors de la phase d'estimation de l'algorithme E.M. Afin de stabiliser les calculs, nous avons réduit la dimension des vecteurs acoustiques par une analyse en composantes principales (ACP). Comme nous l'avons indiqué précédemment, notre objectif est de modéliser au mieux l'espace des vecteurs acoustiques générés par HTS. Nous proposons donc de déterminer une ACP, notée Π_k, sur chaque corpus A_k et permettant d'expliquer 95% de la variance initiale.

Ainsi, si les paramètres acoustiques évalués sont de dimension élevée alors l'utilisation de Π_k est nécessaire. Dans ce cas, elle est appliquée à A_k et V_k pour obtenir, respectivement, $\Pi_k(A_k)$ et $\Pi_k(V_k)$. $\mathcal{M}_k$ est dans ce cas appris en utilisant les données issues de $\Pi_k(A_k)$ et $\Pi_k(V_k)$.

L'ensemble des traitements effectués lors de cette étape est résumé figure 4.2.

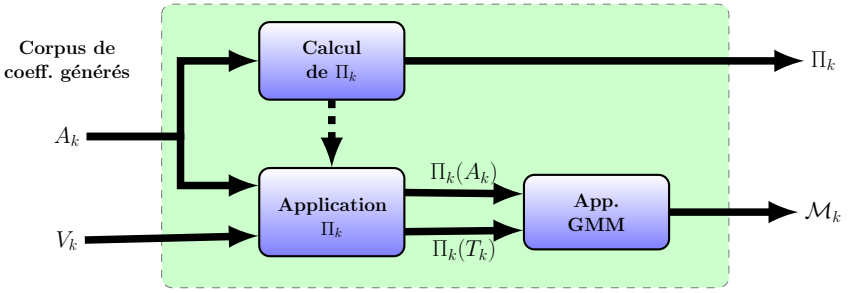


FIGURE 4.2 – Étape 2 - Apprentissage des modèles GMM

4.1.3 Évaluation des modèles

La dernière étape du protocole d'évaluation consiste à évaluer l'espace capturé par le GMM $\mathcal{M}_k(n_k^\star)$ à l'aide du corpus de test $T_{a/s}$. Pour cela, la log-vraisemblance du corpus de référence $T_{a/s}$ est calculée pour le GMM $\mathcal{M}_k(n_k^\star)$. Cette quantité est notée $LL(T_{a/s}; \mathcal{M}_k(n_k^\star))$. De même, $LL(T_{a/s}; \mathcal{M}_{a/s}(n_{a/S}^\star))$ désigne la log-vraisemblance de $T_{a/s}$ pour le $\mathcal{M}_{a/s}(n_{a/s}^\star)$ correspondant à l'espace acoustique de référence, et constitue une borne haute à l'ensemble des valeurs de log-vraisemblance calculées.

Dans le cas où la dimension des vecteurs acoustiques utilisées implique un calcul d'ACP Π_k, cette transformation est appliquée au corpus $T_{a/s}$ pour obtenir le corpus $\Pi_k(T_{a/s})$. La log-vraisemblance est alors calculée sur le corpus $\Pi_k(T_{a/s})$ mais, par hypothèse, nous supposons que cette vraisemblance est représentative de la vraisemblance $LL(T_{a/s}; \mathcal{M}_k(n_k^\star))$.

Ainsi, Les écarts entre les vraisemblances obtenues permettent de quantifier les *distances* entre les espaces des coefficients générés par HTS et l'espace des coefficients associés au signal naturel après un traitement d'analyse effectué par STRAIGHT et SPTK.

La figure 4.3 résume l'étape d'évaluation des modèles.

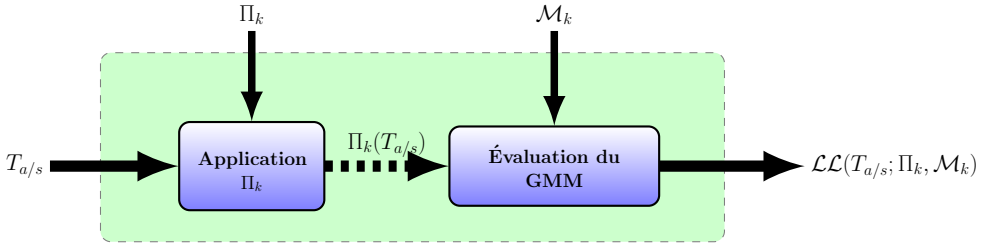


FIGURE 4.3 – Étape 3 - Évaluation de l'espace acoustique modélisé par un GMM $\mathcal{M}_k$ pour un corpus de référence $T_{a/s}$

4.2 Évaluation non paramétrique

L'avantage de l'approche précédente est qu'elle caractérise un espace acosutique globalement, maintenant elle reste sensible au nombre de vecteurs utilisées lors de la phase d'apprentissage et indirectement au nombre de composantes. Pour augmenter la finesse de la modélisation, on souhaite en effet un nombre important de composantes.

Afin d'affiner notre analyse de l'influence d'un jeu de descripteurs sur la modélisation effectuée par HTS, un second protocole expérimental a été défini. Ce protocole est de nature locale et non-paramétrique, il repose sur le calcul d'une distance entre les vecteurs acoustiques générés par HTS et ceux extraits du signal naturel. En effectuant un changement d'échelle puis un partitionnement, il devient alors possible d'évaluer la pertinence de la modélisation effectuée par HTS pour un ensemble de modèles restreints. Le processus d'évaluation proposé est illustré par la figure 4.4.

Comme l'indique cette figure, ce protocole repose sur trois étapes. La première (*Appariement*) consiste à calculer, pour chaque trame, un vecteur d'écarts entre les vecteurs acoustiques générés C_{gen} et ceux extraits du signal naturel C_{ori}. Une fois l'ensemble des vecteurs d'écarts obtenu, la seconde étape (*Passage à l'échelle*) permet d'obtenir un vecteur d'erreurs représentatif pour chaque segment d'échelle supérieure (un segment défini comme un phone par exemple). Un partitionnement de cet ensemble est ensuite effectué lors de la troisième étape (*Partitionnement*). La quatrième étape (*Normalisation*) doit permettre de calculer pour chaque vecteur d'erreurs un scalaire représentant la dégradation associée à chaque segment. Cette étape peut être réalisée en parallèle de l'étape de partitionnement et fait partie du bloc dédié au passage de l'échelle de la trame à l'échelle supérieure. Néanmoins, pour simplifier l'implémentation, elle a été placée à la suite de l'étape de partitionnement. La dernière étape (*Combinaison*) consiste à déterminer un écart représentatif de la dégradation obtenue à l'issue de la génération effectuée par HTS par rapport aux données extraites du signal naturel et ceci pour chaque partition. Dans cette section nous allons maintenant détailler l'ensemble de ces étapes dans l'ordre chronologique des traitements.

4.2.1 Appariement

Tout d'abord, nous devons supposer deux ensembles alignés de vecteurs de coefficients. Le premier, C_{Ori}, correspond aux coefficients extraits du signal naturel et le second, C_{Syn}, correspond aux coefficients générés par HTS. En se basant sur ces séquences, la première étape consiste à calculer, pour chaque trame t, un vecteur d'écarts ve_t, entre les vecteurs $C_{Ori}(t)$ et $C_{Syn}(t)$, tel que :

$$ve_t = D(C_{Ori}(t), C_{Syn}(t)) \tag{4.3}$$

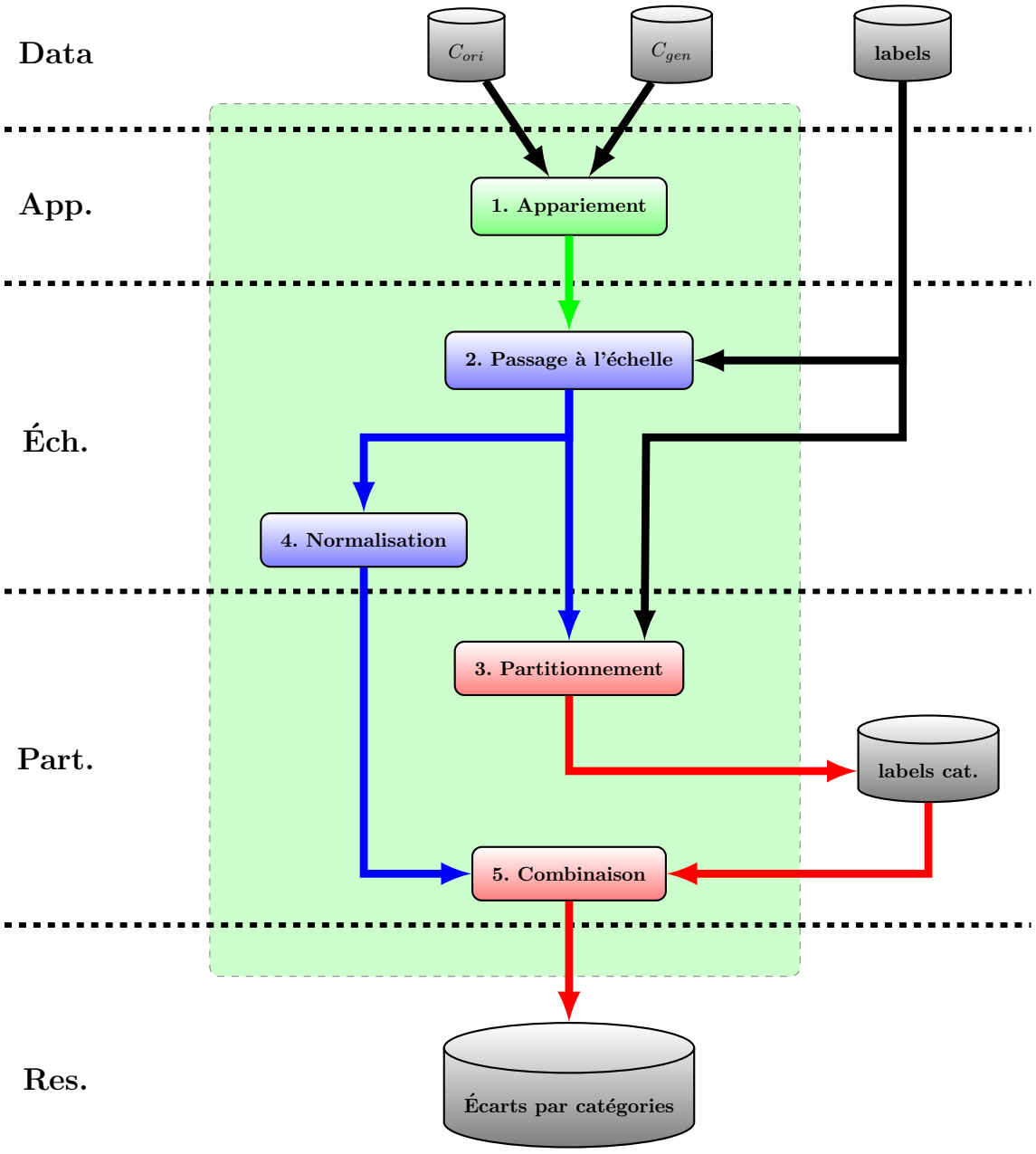


FIGURE 4.4 – Représentation schématique du protocole d'évaluation non-paramétrique. Cinq blocs se distinguent. Le bloc *Données* (en noir) identifie les données nécessaires pour appliquer le protocole. Le bloc *App.* (Appariement, en vert) identifie les opérations nécessaires pour obtenir un écart représentatif à l'échelle de la trame. Le bloc *Éch.* (Changement d'échelle, en bleu) identifie les opérations nécessaires pour obtenir une erreur représentative à une échelle supérieure. Le bloc *Part.* (Partitionnement, en rouge) correspond aux opérations nécessaires pour déterminer un écart représentatif pour chaque ensemble de segments. Comme nous le verrons, ces ensembles peuvent être prédéfinis ou déterminés par le protocole. Le dernier bloc *Res.* (Résultat, en noir) correspond aux résultats (écarts représentatifs + identifiants des ensembles) obtenus à l'issue du protocole.

où D correspond à l'une des fonctions suivantes :

– l'écart relatif entre $C_{Ori}(t)$ et $C_{Syn}(t)$:

$$D(C_{Ori}(t), C_{Syn}(t)) = C_{Syn}(t) - C_{Ori}(t) \tag{4.4}$$

– l'écart absolu $C_{Syn}(t)$ et $C_{Ori}(t)$:

$$D(C_{Ori}(t), C_{Syn}(t)) = \left| C_{Syn}(t) - C_{Ori}(t) \right| \tag{4.5}$$

Dans le cadre de l'évaluation du F0, D peut correspondre, également, à l'une des fonctions suivantes :

– l'écart en cent pour l'analyse du F_0 :

$$D(C_{Ori}(t), C_{Syn}(t)) = 1200 * log_2 \left(\frac{C_{Syn}(t)}{C_{Ori}(t)} \right) \tag{4.6}$$

– l'erreur de voisement :

$$D(C_{Ori}(t), C_{Syn}(t)) = \begin{cases} 0, & si\ C_{Ori}(t) = C_{Syn}(t) = 0 \\ 0, & si\ C_{Ori}(t) \neq 0\ et\ C_{Syn}(t) \neq 0 \\ 1, & sinon \end{cases} \tag{4.7}$$

La dernière mesure a été introduite pour pallier une adaptation nécessaire à l'utilisation des trois autres fonctions possibles : les trames non-voisées sont ignorées. En effet, prendre en compte les trames non-voisées impliquerait soit une distance, pour une même trame, nulle (pas d'erreur de voisement) soit une distance trop élevée. Dans ce second cas, la distance obtenue entre coefficients voisés associés à une même trame serait négligeable.

À l'issue de la première étape, le vecteur d'écarts ve_t est de même dimension que les vecteurs $C_{Ori}(t)$ et $C_{Syn}(t)$. Comme précédemment, nous notons M cette dimension. Le tableau 4.1 résume les mesures d'écarts utilisées en fonction du type de coefficients évalué.

Paramètre	**Éc. abs.**	**Éc. rel.**	**Éc. cent**	**Er. Voisement**
F0	X	X	X	X
MGC	X	X		
BAP	X	X		
Durée	X	X		

TABLE 4.1 – Distance utilisée en fonction du type de coefficients analysé

4.2.2 Changement de niveau de représentation

Soit h un élément de l'ensemble des horizons H tel que, pour ce protocole, H = {trame, phone, syllabe, énoncé}. Nous pouvons définir un segment s_h (noté s à partir de la prochaine étape du protocole) comme une séquence d'indices de trame où $deb(s_h)$ identifie la première trame et $fin(s_h)$ la dernière trame du segment. Ainsi, la seconde étape

du protocole consiste, à partir de l'ensemble des vecteurs d'écarts ve_t, de déterminer pour chaque segment s_h un vecteur d'écarts représentatif $ver(s)$.

Pour cela, il est nécessaire de définir l'occurrence $o(s_h)$ telle que :

$$o(s_h) = (ve_{deb(s_h)}, \ldots, ve_{fin(s_h)}) \tag{4.8}$$

l'occurrence $o(s_h)$ est une sous séquence des vecteurs d'écarts.

Le tableau 4.2 résume l'ensemble des horizons utilisés pour chacun des types de coefficients évalués. Une case vide indique que l'horizon n'est pas pris en compte et une case grise que l'horizon n'existe pas pour le paramètre acoustique associé. L'horizon *phrase* pour la durée correspond au débit syllabique tel que présenté dans la section du chapitre 1.

Paramètre	**Trame**	**Phone**	**Syllabe**	**Phrase**
F0	X	X		
MGC	X	X		
BAP	X	X		
Durée		X	X	X (Déb. syl.)

TABLE 4.2 – Horizons utilisés en fonction du type de coefficients analysés. Une case vide indique que l'horizon n'est pas pris en compte et une case grise indique que l'horizon n'existe pas pour ce type de coefficients.

En supposant que chaque occurrence $o(s_h)$ est connue, nous pouvons déterminer le vecteur d'écart représentatif du segment s_h en utilisant l'une des fonctions suivantes :

– la moyenne des écarts :

$$ver(s_h) = \frac{1}{fin(s_h) - deb(s_h) + 1} \left(\sum_{t=deb(s_h)}^{fin(s_h)} (ve_t) \right) \tag{4.9}$$

– l'écart central :

$$ver(s_h) = ve_{(deb(s_h)+fin(s_h))/2} \tag{4.10}$$

Le tableau 4.3 indique quelle fonction est utilisée pour chacun des types de vecteurs acoustiques et horizons analysés. L'horizon de la trame est particulier en ce sens qu'au maximum un seul vecteur d'écart peut être associé à une trame. Ainsi, pour cet horizon. Une case vide indique que l'horizon n'est pas évalué. Une case grise indique qu'aucune des fonctions, permettant de déterminer un vecteur d'écart représentatif, n'est appliquée.

4.2.3 Partitionnement

À l'issue de l'étape précédente, nous obtenons pour chaque segment du corpus un vecteur d'écarts. La troisième étape du protocole consiste à regrouper ces vecteurs d'écarts en fonction de caractéristiques linguistique ou phonologique.

Paramètre	Trame	Phone	Syllabe	Phrase
F0		Ce ou Mo	Ce ou Mo	
MGC		Ce ou Mo		
BAP		Ce ou Mo		
Durée				Ce ou Mo

TABLE 4.3 – Fonction utilisée pour calculer un vecteur d'écarts représentatif, pour chaque segment, en utilisant l'ensemble des vecteurs d'écarts associés aux trames composant ces segments (Ce=vecteur d'écarts associé à la trame centrale, Mo=moyenne des vecteurs d'écarts). Une case vide indique que pour le paramètre évalué, l'horizon associé ne sera pas analysé ; une case grise qu'aucune fonction n'est utilisée.

Pour atteindre cet objectif, nous définissons une catégorie comme un ensemble de propriétés à valider. Supposons J le nombre de catégories. À chacune d'elles est associée un identifiant A_h^j où $j \in [1..J]$ et h correspond à l'horizon temporel. Cet horizon est nécessaire car l'identifiant de la catégorie est associé à chaque segment s d'horizon h.Enfin, pour chaque segment s, nous définissons $b(s)$, l'application qui permet de déterminer la catégorie, identifiée par A_h^j, à laquelle le segment appartient :

$$b(s) = A_h^j \tag{4.11}$$

Le tableau 4.4 permet d'identifier les différentes partitions utilisées lors de l'application de ce protocole. Nous avons restreint l'horizon au phone uniquement. Ainsi, trois cas de partitionnement ont été réalisés. Deux partitionnements effectués reposent sur des catégories déterminées de manière automatique. Le premier ne prend en compte que l'identifiant phonétique du segment s, le second, en revanche, considère un contexte de deux phones (labels des segments $s-1$, s et $s+1$) pour caractériser le segment s [2]. Le dernier partitionnement effectué considère que les catégories ont été définies manuellement. Quatre catégories se distinguent dans ce cas : les voyelles, les consonnes voisées, les consonnes non-voisées et les NSS. Ce choix découle des premiers résultats obtenus par l'application de ce protocole et sera ainsi expliqué dans le chapitre 7.

4.2.4 Normalisation

Comme nous l'avons indiqué lors de la présentation générale du protocole, la quatrième étape consiste à déterminer, à partir d'un vecteur d'écart, un scalaire. Ce scalaire représente la dégradation entre les vecteurs acoustiques générés par HTS et ceux extraits du signal naturel. Cette étape peut être réalisée en parallèle de l'étape de partitionnement.

Afin de déterminer la dégradation associée à chaque segment, que nous identifions par

2. Chaque énoncé est considéré comme isolé afin de pouvoir utiliser les labels HTS. Ainsi la valeur indéfinie est appliquée au segment suivant le dernier segment de chaque énoncé et au segment précédent le premier segment de chaque énoncé.

Paramètre	Phone	Phone en contexte	Cat. imp.
F0	X	X	X
MGC	X	X	X
BAP	X	X	X
Durée	X	X	X

TABLE 4.4 – Méthodes de partitionnement utilisées en fonction du paramètre analysé. Trois cas de partitionnement ont été appliqués. Le partitionnement par arbre de décision a été effectué en utilisant deux contextes : le label phonétique du segment et les labels phonétiques de 3 phones dont le label central correspond au segment. Le dernier cas de partitionnement utilisé (colonne Cat. imp.) consiste à identifier le segment comme faisant partie de l'une des catégories suivantes : voyelle, consonne voisée, consonne non-voisée, NSS.

$ne(s)$, nous avons utilisé la norme euclidienne :

$$ne(s) = \sqrt{\sum_{m=1}^{M} (ver^m(s))^2} \tag{4.12}$$

4.2.5 Combinaison

À l'issue des étapes de partitionnement et de normalisation, nous obtenons, pour chaque segment s, un scalaire $ne(s)$, représentant la dégradation de la génération effectuée par le système de HTS par rapport aux vecteurs acoustiques extraits du signal naturel, ainsi qu'un identifiant de catégorie obtenu par l'application $b(s)$ telle que définie par l'équation (4.11).

Notons $\mathcal{S}$ l'ensemble des segments du corpus utilisé pour l'évaluation. En utilisant l'application réciproque $b^{-1}(A_h^j)$ nous obtenons un sous-ensemble de $\mathcal{S}$ qui correspond au segment appartenant à la catégorie A_h^j. Nous supposons que ce sous-ensemble, noté $\mathcal{S}_j$ est de cardinalité S.

La dernière étape du protocole consiste à utiliser ces informations pour déterminer la dégradation, de la génération effectuée par HTS par rapport aux vecteurs acoustiques extraits du signal naturel, pour l'ensemble des J catégories. Nous souhaitons déterminer me_j tel que :

$$me_j = E(ne(\mathcal{S}_j(1), \ldots, \mathcal{S}_j(S))) \tag{4.13}$$

où B désigne l'ensemble des segments ayant pour identifiant celui de la catégorie A_h^j. Enfin, E correspond à une fonction permettant de déterminer la dégradation issue de la modélisation effectuée par HTS.

Pour ce protocole, nous distinguerons trois fonctions :

– la moyenne

$$E(ne(\mathcal{S}_j(1),\ldots,\mathcal{S}_j(S))) = \frac{1}{S}\sum_{s=1}^{S} ne_h(\mathcal{S}_j(s)) \tag{4.14}$$

– la variance

$$E(ne(\mathcal{S}_j(1),\ldots,\mathcal{S}_j(S))) = \frac{1}{S}\sum_{s=1}^{S}\left[ne(\mathcal{S}_j(s)) - \frac{\sum_{r=1}^{S} ne_h(\mathcal{S}_j(r))}{S}\right]^2 \tag{4.15}$$

– la RMS

$$E(ne(\mathcal{S}_j(1),\ldots,\mathcal{S}_j(S))) = \sqrt{\frac{\sum_{s=1}^{S} ne(\mathcal{S}_j(s))^2}{S}} \tag{4.16}$$

4.3 Conclusion

Dans ce chapitre, nous avons présenté deux protocoles expérimentaux permettant d'évaluer la qualité des vecteurs acoustiques générés par HTS par rapport à ceux extraits du signal naturel. Le premier consiste à modéliser par un GMM l'espace des vecteurs acoustiques générés. En utilisant un corpus de référence composé de coefficients extraits du signal naturel, la pertinence de cet espace peut être évaluée. Néanmoins, ce protocole offre un point de vu global sur les espaces acoustiques et ne permet pas de contrôler précisément des dégradations au niveau de la trame. Afin de pouvoir effectuer une analyse plus locale et surtout ciblée, un second protocole a été mis en place. Ce protocole consiste à calculer un écart représentatif, entre les vecteurs acoustiques générés par HTS et ceux extraits du signal naturel, pour un ensemble de catégories qui peuvent être définies manuellement ou automatiquement.

Par ces deux protocoles, nous possédons les outils nécessaires pour effectuer l'évaluation objective de l'impact des descripteurs sur la modélisation effectuée par HTS. Notre objectif est de chercher à comprendre quelle est cette influence et comment sélectionner un jeu de descripteurs optimal pour la synthèse du français. Il nous faut donc à présent définir un corpus de parole annoté. Les descripteurs seront déterminés à partir de ces annotations. Le prochain chapitre est consacré à la présentation du corpus de parole utilisé pour effectuer les évaluations ainsi qu'aux jeux de descripteurs qui seront évalués.

Chapitre 5

Données expérimentales

Pour évaluer l'influence des descripteurs sur la modélisation effectuée par HTS, il est nécessaire de disposer d'un corpus de parole annotée. L'ensemble des annotations doit permettre d'obtenir les descripteurs utilisés en synthèse. L'objectif de ce chapitre est de présenter ce corpus ainsi que l'ensemble des données expérimentales utilisées pour effectuer les évaluations.

Ce chapitre débute par la présentation de la structure de données ROOTS qui a été conçue pour pouvoir stocker l'ensemble des annotations en assurant des relations cohérentes entre tous les niveaux de description des énoncés. En se basant sur la structure ROOTS, et

en tirant ainsi avantage de ses spécificités, un processus d'annotation automatique complet a été mis en place. Ce processus, utilisé pour obtenir le corpus, est présenté dans la seconde section de ce chapitre. La troisième section est consacrée à la présentation du corpus utilisé pour les évaluations. Enfin, la dernière section de ce chapitre présente les jeux de descripteurs évalués.

5.1 Représentation des données par ROOTS

Pour annoter un corpus, il est nécessaire d'utiliser un ensemble d'outils complémentaires qui représentent rarement les données observées sous un même format. La diversité des outils et des formats associés implique le besoin de définir des systèmes cohérents d'annotation de corpus. L'objectif est de pouvoir croiser les informations fournies par divers outils analysant les différents niveaux de description de la parole en s'affranchissant de la complexité des relations qui lient les différents niveaux de description.

Au début des années 2000, plusieurs systèmes de représentation ont été publiés. En 2001, S. Bird [Bird2001] propose une représentation sous forme de graphes composés d'annotations directement ancrées sur le signal indépendamment les unes des autres. Les graphes d'annotation ont été implantés dans deux applications ATLAS [Bird2000] et AGTK. Cassidy et al. [Cassidy1996, Cassidy2001] ont de leur côté développé un outil dont le premier objectif était de fournir des méthodes d'interrogation et d'analyse de corpus. Leur outil, Emu, repose sur une structure organisée en niveaux d'annotations composées d'éléments (tokens) associés ou non à une information temporelle ; une relation de dominance, de séquencement ou d'association permet de relier les divers éléments de cette structure.

Une autre structure a également été proposée au début des années 2000 : la structure HRG (Heterogeneous relation graphe). Contrairement aux structures précédentes, HRG veut apporter une réponse à la représentation homogène des annotations en vue d'une utilisation dans un système de synthèse. Ainsi, répondant à un souci de rapidité et de sécurité, ce système permet de représenter une information linguistique sous forme de graphes dont les nœuds ne contiennent pas d'information mais sont reliés aux entités linguistiques et dont les arcs définissent les liens entre ces items. Toute redondance d'information est supprimée grâce à cette structure et à l'emploi massif de pointeurs garantissant ainsi la mise à jour et la cohérence des informations. Dans [Rojc2007], une structure de données de type HRG est combinée à des machines à états finis. Les auteurs mettent en avant une séparation claire entre ce qui est dépendant et indépendant de la langue. Enfin, l'IRCAM [Veaux2008], dans sa plateforme IRCAMCORPUSTOOLS, organise la représentation des annotations en deux classes : l'une apporte le contenu des annotations et l'autre les relations hiérarchiques et/ou séquentielles qui existent entre elles.

Pour réaliser nos travaux, nous souhaitions disposer d'une structure permettant d'ac-

céder à l'ensemble des informations et des liens entre différentes informations possibles pour décrire un énoncé de parole. L'utilisation de la structure choisie, bien que spécifique à la parole, ne doit pas se limiter à l'annotation d'un signal acoustique. Elle doit pouvoir, par exemple, servir pour stocker les annotations liées uniquement à un texte. La structure d'annotation la mieux adaptée au contexte de nos travaux nous a semblé être HRG. Cependant, dans notre cas et contrairement à HRG, nous souhaitons favoriser l'exhaustivité des annotations et la versatilité des accès au détriment de la rapidité d'accès.

Pour satisfaire l'ensemble des contraintes précédentes, le système d'annotation d'énoncés ROOTS [Barbot2011] a été conçu. Il permet de décrire un corpus de parole en se basant sur le paradigme objet. L'objectif est de conserver la totalité des informations ainsi que l'ensemble des liens entre ces informations. Pour cela, ROOTS repose sur quatre concepts principaux, illustrés par la figure 5.1 :

- Des *items* permettent de décrire des éléments d'annotation ;
- Des *séquences* structurent les items dans le temps.
- Des *relations* lient les items de deux séquences différentes ;
- Un *énoncé* rassemble un ensemble de séquences et de relations.

Afin d'alléger la figure 5.1, les items ne sont pas représentés mais inclus au sein d'une séquence.

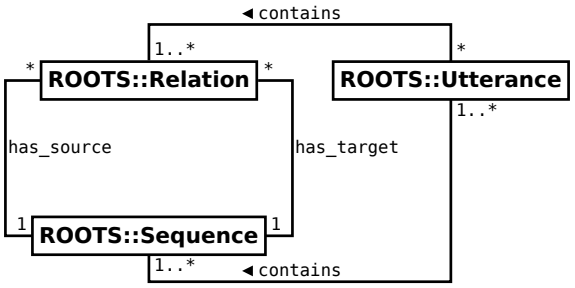


FIGURE 5.1 – Spécification fonctionnelle ROOTS

5.1.1 Item et séquences

Pour représenter un niveau d'annotation, ROOTS utilise le concept d'item. Un item correspond à une instance d'une classe modélisant un type d'annotation (un segment acoustique, un phone, une syllable, un syntagme, etc.). En se basant sur le paradigme objet, il est possible de définir simplement des nouveaux items par mécanisme d'héritage et de polymorphisme. Par exemple, un article défini, qui est un déterminant particulier, est décrit par une classe héritant de la classe représentant le déterminant.

Dans ROOTS, les items ne peuvent exister qu'au sein d'une séquence (éventuellement une séquence d'un seul item). La séquence permet de relier des items en utilisant comme support le temps. Certains items particuliers comme les syllables ou les arbres syntaxiques sont particuliers car ils ne peuvent exister sans items élémentaires (l'item syllable s'appuie

sur des items phonèmes, l'item syntaxe s'appuie sur des items mots). Ces items seront appelés items composés.

Séquence

Une séquence est un objet qui permet de stocker les items ROOTS en respectant plusieurs contraintes :

- les items sont ordonnés en se basant sur leur ordonnancement temporel ;
- une séquence est homogène : tous les items contenus dérivent d'une même classe. Une séquence peut donc contenir des allophones ou des mots mais pas les deux en même temps. Une séquence peut néanmoins contenir un article et un adjectif indéfini car ces deux items sont des déterminants et dérivent donc d'une classe mère identique (la classe *Determiner*). Cette restriction volontaire permet d'assurer que des items avec des sémantiques différentes ne sont pas mélangés.

La classe *Sequence* factorise les opérations de parcours et d'accès aux éléments et fournit donc une interface commune aux classes concrètes qui en dérivent. Ces dernières implantent la notion de séquence d'éléments pour chaque type d'éléments existants. C'est par ce moyen que les séquences sont homogènes.

Ces différentes classes sont naturellement reliées au type d'item qu'elles font intervenir. De nouveau, l'aspect générique de cette représentation permet de conserver une sémantique cohérente aux séquences en garantissant l'homogénéité des items qu'elles contiennent, mais permet également de diversifier ces derniers en spécialisant la classe de chaque item. Par exemple, une séquence de segments acoustiques pourrait très bien faire intervenir des segments de signal brut représentés par une séquence d'échantillons et des segments de signal représentés par un modèle d'analyse/synthèse.

Item composé

Il existe des items particuliers, comme les syllabes ou les arbres syntaxiques, qui ne peuvent exister sans des items plus simples, respectivement les allophones/phonèmes et les mots. Afin de respecter ces contraintes, ROOTS introduit, de manière analogue à HRG, la notion d'item composé.

Un item composé ne peut exister seul. En effet, les éléments qui le composent sont des références à des items d'une autre séquence. Un item composé qualifie donc un item en le déclarant partie d'une structure (arbre syntaxique, arbre syllabique). Ainsi, il ne peut y avoir de relations impliquant les items référencés (les feuilles de l'arbre syntaxique/syllabique) ; les relations doivent utiliser la séquence d'items contenant les items référencés.

Pour le moment, les items composés implémentés dans ROOTS sont des arbres syllabiques et syntaxiques. La figure 5.2 illustre le principe des items composés et des séquences : la séquence du haut représente une séquence de syllabes dont chaque item référence des phones contenus dans la séquence illustrée en bas.

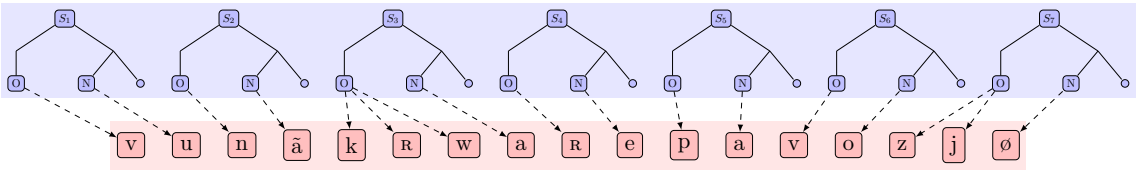


FIGURE 5.2 – Illustration des concepts de séquence et d'item composé pour la phrase *vous n'en croirez pas vos yeux*. La séquence du haut représente une séquence de syllabes. La séquence du bas représente une séquence de phones. Chaque syllabe référence un ensemble de phones. Ce référencement est indiqué par les liens en pointillés. Cette figure a été générée par ROOTS.

5.1.2 Relations

L'objectif d'une relation est de lier les items ROOTS de deux séquences de classes différentes. Par définition, une relation est donc de nature hétérogène (non binaire), mais elle est construite entre des séquences associées à un même énoncé.

Dans ROOTS, la classe *Relation* factorise les opérations d'accès aux indices des séquences mises en relation ainsi que les opérations d'existence d'une relation entre deux items. Techniquement, la relation entre deux séquences d'items, a et b, est stockée sous la forme d'une matrice R_a^b. L'élément d'indice (i, j) de la matrice est un entier représentant un booléen qui indique si l'item a_i est en relation avec l'item b_j. Du fait de l'ordonnancement des indices selon le référentiel temporel, la matrice d'une relation entre deux séquences fait apparaître une structure particulière : les entrées égales à 1 forment un faisceau orienté dans le sens du temps pour les deux séquences. La relation réciproque entre les items de b et a s'obtient aisément par la transposition de la matrice R_a^b. Un objet Relation stocke également les références des séquences impliquées dans la relation. La figure 5.3 illustre le concept de relation entre deux séquences de mots issues de deux outils différents.

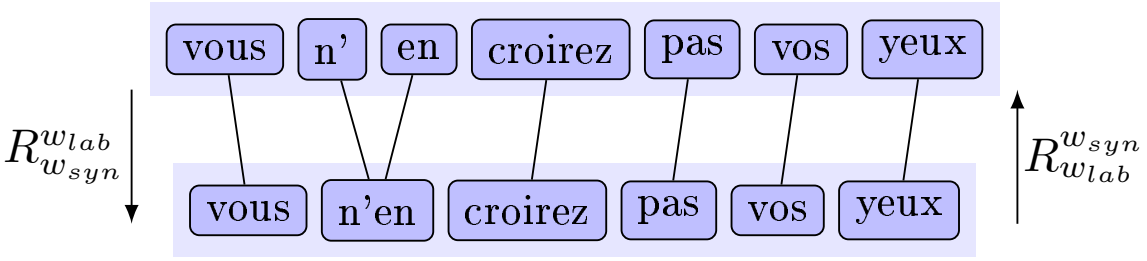


FIGURE 5.3 – Exemple de relation entre deux séquences dans ROOTS pour la phrase *vous n'en croirez pas vos yeux*. La séquence de mots du haut est obtenue par un analyseur syntaxique et la séquence du bas correspond au texte original. Cet exemple se focalise sur le lien entre le mot *n'en* que l'analyseur divise en deux mots : *n* et *en*

Pour des cas simples, la relation correspond à une fonction ou une application : cela se traduit par la présence d'au plus un (resp. d'un seul) élément égal à 1 par ligne dans la matrice de relation pour une fonction (resp. une application). La figure 5.3 illustre ce cas sur un exemple de relation entre une séquence w_{syn} de mots issue d'une analyse syntaxique et une séquence w_{lab} de mots issue d'un énoncé. On peut remarquer qu'au mot *n'en* de l'énoncé, contraction de *ne* et *en*, correspond *n* et *en* du côté de l'analyse syntaxique. La matrice associée à cette relation s'écrit :

$$R^{w_{lab}}_{w_{syn}} = \begin{array}{c} \\ \text{vous} \\ \text{n} \\ \text{en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{array} \begin{array}{c} \begin{array}{cccccc} \textit{vous} & \textit{n'en} & \textit{croirez} & \textit{pas} & \textit{vos} & \textit{yeux} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{5.1}$$

La relation est ici une application.

Cependant, selon les relations et les séquences réalisées, une relation est généralement plus complexe. Un item d'une séquence source peut être en relation avec plusieurs items de la séquence cible (comme, par exemple, pour la relation réciproque à celle présentée par la figure 5.3. Si l'on considère la relation définie entre une séquence de mots et une séquence de syllabes, un mot est généralement en relation avec (car composé de) plusieurs syllabes. Par exemple, dans le cas de l'énoncé *Vous n'en croirez pas vous yeux*, on peut supposer la relation associant la séquence w_{lab} de 6 mots (illustrée par la figure 5.3) à la séquence s de 7 syllabes (illustrée par la figure 5.2) décrite par :

$$R^{s}_{w_{lab}} = \begin{array}{c} \\ \text{vous} \\ \text{n'en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{array} \begin{array}{c} \begin{array}{ccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \tag{5.2}$$

Il est également possible d'associer plusieurs items cibles pour un item source en utilisant ce mécanisme. Ainsi, ROOTS permet de couvrir un maximum de cas et délègue à l'utilisateur la sémantique de la relation à implanter en fonction de l'application souhaitée. Par exemple, le sens donné à une relation entre deux séquences de mots (ce peut être comme le cas présenté précédemment, un lien entre des outils d'analyse linguistique qui ont des stratégies de parsing différentes).

Afin de compléter l'objet relation, l'item ROOTS contient une propriété appelée link forward qui permet de connaître les relations dans lesquelles un item est impliqué. Grâce à cette propriété et à la structure de la relation, il est possible de connaître rapidement quel est l'ensemble des items reliés à un item donné par l'intermédiaire de relations entre séquences.

Compte-tenu de la structure particulière d'une matrice de relation, il est souvent possible composer des relations R_a^b et R_b^c où les séquences d'items a, b et c sont associées à un même énoncé à l'aide d'un produit matriciel de type $R_a^b R_b^c$: c'est en particulier toujours le cas lorsque la première relation R_a^b correspond à une fonction. Par exemple, la relation $R_{w_{syn}}^{w_{lab}}$, entre la séquence w_{syn} de mots issue d'une analyse syntaxique et la séquence w_{lab} présentée figure 5.3, peut être composée avec la relation $R_{w_{lab}}^s$, entre la séquence w_{lab} de mots et la séquence s de syllabes (décrite équation 5.1), pour obtenir la relation entre les mots *syntaxiques* et les syllabes qui est décrite par :

$$\begin{aligned} R_{w_{syn}}^s &= R_{w_{syn}}^{w_{lab}} R_{w_{lab}}^s \\ &= (R_{w_{lab}}^{w_{syn}})^\top R_{w_{lab}}^s \\ &= \begin{array}{c} \\ \text{vous} \\ \text{n} \\ \text{en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{array} \begin{array}{c} \begin{array}{ccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \end{aligned}$$

Si la composition de relations est possible, il faut néanmoins rester prudent car une relation ainsi obtenue dépend des relations intervenant dans la composition. Cela est dû à l'absence *de réversibilité* de la relation. Par exemple, en considérant la matrice de relation $R_{w_{syn}}^{w_{lab}}$ décrite par la matrice (5.1), la relation entre la séquence de mots w_{lab} et w_{syn}, est donnée par la matrice

$$R_{w_{lab}}^{w_{syn}} = \left(R_{w_{syn}}^{w_{lab}}\right)^T .$$

Le produit $R_{w_{syn}}^{w_{lab}} R_{w_{lab}}^{w_{syn}}$ ne correspondant donc pas à la matrice identité, les relations entre la séquence de mots w_{lab} et séquence de syllabes s par

– la matrice $R_{w_{lab}}^s$ décrite par (5.2)

– la matrice de composition $R_{w_{lab}}^{w_{syn}} R_{w_{syn}}^s$

sont différentes. Ainsi, il est nécessaire, avant de créer une nouvelle relation par composition, de bien définir la sémantique de la relation souhaitée.

5.1.3 Énoncé (utterance)

Les objets ROOTS de classe *Utterance* doivent être vus comme des conteneurs regroupant l'interprétation multi-niveaux d'une même réalisation. Ainsi, un énoncé va contenir des séquences d'items issus d'une même réalisation et les objets relations liant ces séquences.

Un objet *utterance* est donc une structure de données qui n'apporte pas de sémantique particulière. Cet aspect est à la charge de l'utilisateur de l'utterance qui choisira de nommer les séquences selon le sens qu'il leur donne.

Nous pouvons reprendre l'exemple de la figure 5.3. Cette figure représente une utterance dans laquelle l'utilisateur a choisi de mettre en correspondance deux interprétations linguistiques d'une même phrase. Une relation existe entre les items de ces deux séquences, c'est une relation d'inclusion qu'on peut obtenir par une distance lexicographique. L'utilisateur peut apporter du sens à cette relation en donnant aux séquences des labels comme par exemple *Séquence issue du logiciel A/B*.

5.2 Processus d'annotation

A l'aide de ROOTS, un processus d'annotation automatique complet a été mis en place pour remplir cette structure [Boeffard2012, boeffard2012a]. Cette chaîne d'annotation permet, à partir d'un signal audio et du texte qui lui est associé, d'obtenir un corpus complet annoté sur différents niveaux.

Tout d'abord, la mise en place du procédé d'annotation doit respecter un certain nombre de contraintes dictées par l'usage du corpus annoté et la maîtrise des performances. La première contrainte concerne le texte qui devra être conservé sous sa forme originale ; les écarts de lecture devront être signalés par des balises. La seconde concerne le découpage du corpus. Notre objectif étant d'annoter un corpus sur différents niveaux, nous sommes amenés à traiter des fragments de parole ou de texte de taille variable. En effet il est souhaitable pour une analyse syntaxique de disposer d'une phrase complète alors qu'une segmentation en phones est plus efficace sur des extraits courts. Le texte et les plages d'enregistrement seront donc découpés avec une granularité suffisamment fine pour pouvoir travailler sur des fragments courts. Ces fragments pourront être réunis afin de former une phrase ou un fragment plus long à condition de toujours conserver la cohérence du texte. Enfin, pour garantir une annotation de qualité, il faut garder la possibilité d'une intervention manuelle en tenant compte des indicateurs de confiance fournis par les outils intervenant dans le processus.

La chaîne d'annotation, présentée sur la figure 5.4, est constituée de deux étapes. La première consiste à fractionner l'enregistrement de plusieurs heures de parole et d'y asso-

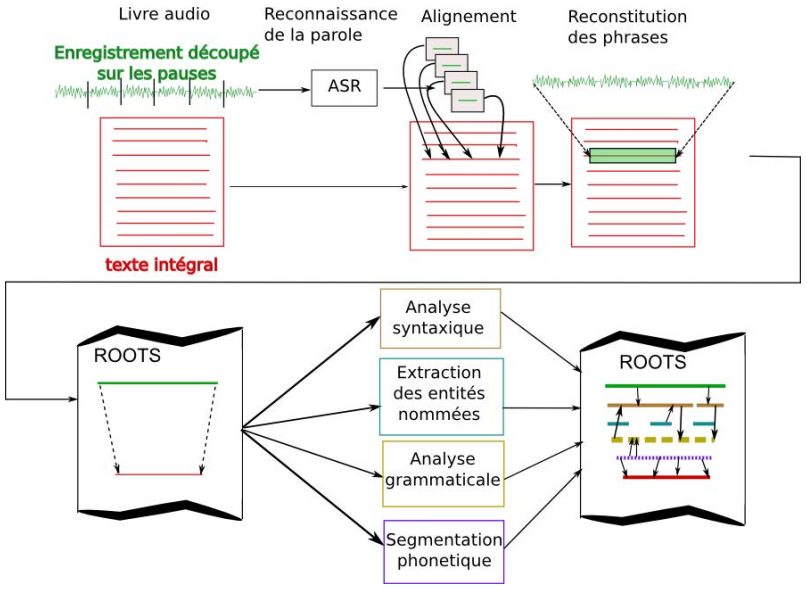


FIGURE 5.4 – Processus d'annotation d'un livre-audio

cier le texte correspondant. Cette étape, d'autant plus coûteuse en temps que le découpage du signal est fin, nécessite le recours à un système de reconnaissance de la parole (ASR) pour retrouver dans le texte complet l'ancrage de la transcription associée au signal. Les extraits sont ensuite regroupés pour reconstituer les phrases du texte original. La deuxième étape concerne l'annotation des phrases du texte et du signal mis en correspondance. La représentation des données par ROOTS est réalisée dès l'étape de découpage par exemple d'un livre-audio en phrases. Elle est enrichie au fur et à mesure de l'annotation des données par l'ajout de nouvelles séquences de description et de relations entre ces séquences.

5.2.1 Découpage du signal de parole et alignement avec le texte

La première étape, concernant à obtenir un alignement entre le texte et le signal, a fait l'objet de plusieurs travaux. [Braunschweiler2010] propose un système automatique alignant des zones de textes d'un livre-audio pour des applications en synthèse de la parole à partir du texte. L'objectif pour d'autres était de faire face à des transcriptions approximatives [Tao2010] ou d'effectuer un alignement sans découper le texte [Moreno2009, prahallad2007]. Dans notre cas, le texte devra être découpé pour effectuer les différentes analyses et pour faciliter sa représentation, en particulier dans l'hypothèse d'une vérification manuelle.

Nous avons choisi d'effectuer l'alignement du texte et du son en 3 étapes :

1. Découpage de l'enregistrement sur des pauses ;
2. Reconnaissance du texte associé à chaque fragment sonore par un système de reconnaissance automatique de la parole ;
3. Alignement entre le texte reconnu et le texte original.

Le découpage de l'enregistrement repose sur l'observation des niveaux d'énergie et sur la longueur des silences. Les seuils sont fixés selon le débit du locuteur et les niveaux d'enregistrement. L'idéal est d'obtenir un fragment sonore en dessous de la phrase, permettant ainsi de reconstituer les phrases tout en gardant des points d'ancrage à l'intérieur de chaque phrase dans le cas des phrases longues.

En utilisant un système de reconnaissance, un texte est obtenu à partir du signal acoustique. Ce texte, qui généralement diffère du texte original, permet d'obtenir un pivot entre le signal et le texte original. Pour cela, un alignement est effectué entre les deux textes. Les extraits reconnus sont traités dans l'ordre du texte ce qui permet, si un extrait est présent plusieurs fois dans le texte, de n'effectuer l'analyse qu'une seule fois et d'utiliser le résultat obtenu pour les autres occurrences. Lorsque la position du premier extrait dans le texte original est déterminée, il est associé au fichier sonore puis supprimé du texte original. L'opération est reproduite pour les extraits suivants jusqu'à la fin du texte. De plus, en comparant le texte original et le texte issu de la reconnaissance, il est possible de mesurer le taux d'erreurs de reconnaissance des mots. Ces erreurs peuvent être dues à une défaillance du système de reconnaissance ou à une lecture erronée du texte (mauvaise prononciation ou modification du texte). Les écarts entre texte reconnu et texte original peuvent être signalés afin de permettre à un opérateur de contrôler les passages concernés.

Enfin pour conserver au mieux la structure du texte, les extraits sont regroupés en phrases en respectant les ponctuations majeures. Lorsqu'un extrait n'est pas terminé par une ponctuation forte, il est simplement regroupé avec l'extrait suivant. Cependant l'information sur la frontière entre les deux extraits est conservée.

5.2.2 Annotation

La deuxième étape consiste à obtenir les annotations sur différents corpus en se basant sur le texte ainsi que le signal précédemment alignés et stockés dans un ensemble d'énoncés Roots.

Pour réaliser cette étape, l'ensemble des outils nécessaires sont appelés les uns indépendamment des autres. En effet, si on suppose une séquence pivot, chaque annotation peut être obtenue séparément. La mise en commun des différents annotations étant réalisée en utilisant la séquence pivot comme support. Roots permet de réaliser cela grâce à l'isolement d'une annotation dans une séquence dédiée et à la relation qui permet de lier la séquence de l'annotation avec la séquence pivot. La structure Roots étant sauvée au format XML, le chargement d'un énoncé stocké dans plusieurs fichiers a été implanté. Ceci apporte deux avantages : la possibilité de ne charger qu'un sous-ensemble de séquences nécessaires à un traitement donné ; la possibilité de pouvoir également distribuer le processus d'annotation. En dernier lieu, il est nécessaire de définir les interfaces permettant d'importer le résultat obtenu par un outil dans la structure Roots.

Les informations obtenues à chaque analyse sont intégrées à l'énoncé ROOTS affinant ainsi la description du corpus et permettant d'établir de nouvelles relations entre les éléments des différentes annotations.

5.3 Présentation du corpus CORDIAL

Les protocoles d'évaluation étant présentés, il nous reste à détailler le corpus utilisé pour effectuer l'analyse. L'objectif de cette partie est de décrire le corpus, que nous appellerons corpus CORDIAL, afin de tenir compte de ces particularités lors de l'analyse des résultats.

5.3.1 Statistiques générales

Le corpus CORDIAL est résumé globalement par le tableau 5.1 qui présente le nombre total d'occurrences pour chacun des horizons utilisés pour qualifier un segment acoustique.

Unité	phone/NSS	syllabe	mot	syntagme	énoncé
Nb. d'occ.	419742	165320	104731	4138	3339

TABLE 5.1 – Résumé du nombre d'occurrences par unité linguistique.

En détaillant les occurrences pour l'ensemble des horizons, comme présenté dans le tableau 5.2, nous pouvons observer que le corpus CORDIAL est un corpus varié. En effet, non seulement l'écart entre les valeurs minimales et les valeurs maximales est important mais les écarts-types sont généralement élevés montrant une dispersion forte entre les occurrences. Parmi ces statistiques, des cas particuliers existent. Par exemple, quelques énoncés ne sont constitués que d'un seul mot, d'une seule syllabe et d'une séquence de quatre labels phonétiques (par exemple, l'énoncé *Ah !* dont la séquence phonologique associée est *start-insp-aa-end*). Bien que présents dans les statistiques du corpus, ces énoncés seront écartés lors de la phase d'apprentissage des modèles HTS et lors de la phase d'évaluation.

En appliquant une analyse analogue pour les durées, dont le résumé est présenté dans le tableau 5.3, les conclusions sont identiques.

5.3.2 Définition des sous-corpus

Pour réaliser les expériences, dont les protocoles ont été décrits dans le chapitre précédent, il est nécessaire d'extraire un ensemble de corpus disjoints à partir corpus CORDIAL. En effet, afin de pouvoir comparer les données générées par HTS, il est nécessaire de disposer de signaux qui ne sont pas utilisés pour l'apprentissage des modèles d'HTS mais qui sont issus du même corpus.

	Unité	Nb. de tailles $\neq$	taille min.	taille max.	Nb. min occ/taille	Nb. max occ/taille	Nb. moy. occ/taille	σ
S	ph.	8	1	8	1	96003	20665	34277.25
M	syl.	6	1	6	49	61419	17455.17	24294.72
	ph.	15	1	15	2	33658	6982.07	9432.23
SY	mots	146	1	199	1	132	28.34	33.10
	syl.	181	1	236	1	93	22.86	25.78
	ph.	357	1	548	1	44	11.59	11.52
U	g.s	10	1	10	1	2795		
	mots	167	2	258	1	73	19.99	21.69
	syl.	206	1	341	1	62	16.21	17.27
	ph.	393	1	772	1	38	8.50	7.86

TABLE 5.2 – Statistiques du nombre d'occurrences sur les syllabes, mots et énoncés du corpus CORDIAL en fonction de leur taille. La première colonne représente les unités linguistiques analysées : S=Syllabe, M=Mots, SY=syntagme et U=énoncé. Ainsi, une ligne du tableau permet de définir la composition en fonction des unités linguistiques de *niveau inférieur*. Les statistiques associées au niveau phonologique ne portent que sur les informations relatives aux phonèmes (les NSS sont ignorés). Cette description est une adaptation de celle présentée dans [Francois2001]

	Durée min (s)	Durée max (s)	Durée moyenne (s)	σ
Ph. / NSS	0.03	2.12	0.72	0.43
Syllabes	0.03	0.86	0.38	0.21
Mots	0.03	1.63	0.62	0.36
Syntagmes	0.07	63.76	12.15	9.40
Énoncé	0.66	82.41	14.37	10.63

TABLE 5.3 – Statistiques sur les durées des phonèmes, syllabes et mots du corpus CORDIAL.

Comme nous le verrons par la suite, nous devons sélectionner trois sous-ensembles d'énoncés pour former les corpus d'apprentissage, de validation et de test. Le corpus d'apprentissage sert à apprendre les modèles et le corpus de test va permettre d'effectuer l'analyse. Le corpus de validation sera décrit ultérieurement.

La sélection des corpus a été effectuée de manière aléatoire en respectant les contraintes suivantes :

- les corpus doivent être disjoints (un énoncé ne peut être présent dans deux corpus) ;
- la durée de chaque corpus a été imposée : environ 1h pour le corpus d'apprentissage, 10min pour les corpus de validation et de test. Ces tailles ont été sélectionnées pour rester comparables à la démonstration fournie par les concepteurs du système HTS et ainsi vérifier, subjectivement et dans une moindre mesure, que la synthèse réalisée est cohérente.

De la même manière que dans le chapitre précédent, nous allons décrire ces corpus et les caractériser en fonction du corpus global en se focalisant sur les horizons suivants : le phonème, la syllabe et le mot.

5.3.3 Focus sur les phonèmes et les NSS

Le premier horizon utilisé est le phonème. La figure 5.5 permet de comparer les distributions phonémiques des différents corpus. La figure se découpe en trois parties : en haut à gauche, la distribution des voyelles et semi-voyelles, en haut à droite la distribution des NSS et enfin en bas, la distribution des consonnes. Pour un phonème, nous avons, de gauche à droite, son taux de représentation dans le corpus global, le corpus d'apprentissage, le corpus de test et enfin le corpus de validation.

La figure montre que, malgré quelques différences, les distributions des quatre corpus sont proches. On peut donc supposer que les trois sous-corpus disjoints sont représentatifs, de part leurs contenus phonologiques. Cela permet également de supposer que l'analyse qui sera effectuée dans les chapitres suivants serait identique si un autre tirage aléatoire avait été effectué.

La plupart des différences concerne le corpus de test. En effet, les phonèmes /kk/, /ss/ et /tt/ y sont significativement sous-représentés (plus de 1% d'écart). Cela peut poser problème si le nombre de trames associées aux phonèmes sous-représentés est insuffisant pour effectuer une analyse statistique. Néanmoins, le nombre de segments associés à ces phonèmes est supérieur à 150 ce qui constitue un nombre de représentants raisonnable pour une telle analyse. Ainsi, nous pouvons supposer que la sous-représentation de ces phonèmes a peu d'influence sur les analyses qui sont décrites dans la partie suivante.

Enfin, il existe deux phonèmes particuliers : le phonème /ng/ qui n'est présent dans aucun des sous-corpus et le phonème /gn/ qui est peu présent (4 exemplaires dans le corpus test). Aucune conclusion ne pourra être émise pour ces deux phonèmes.

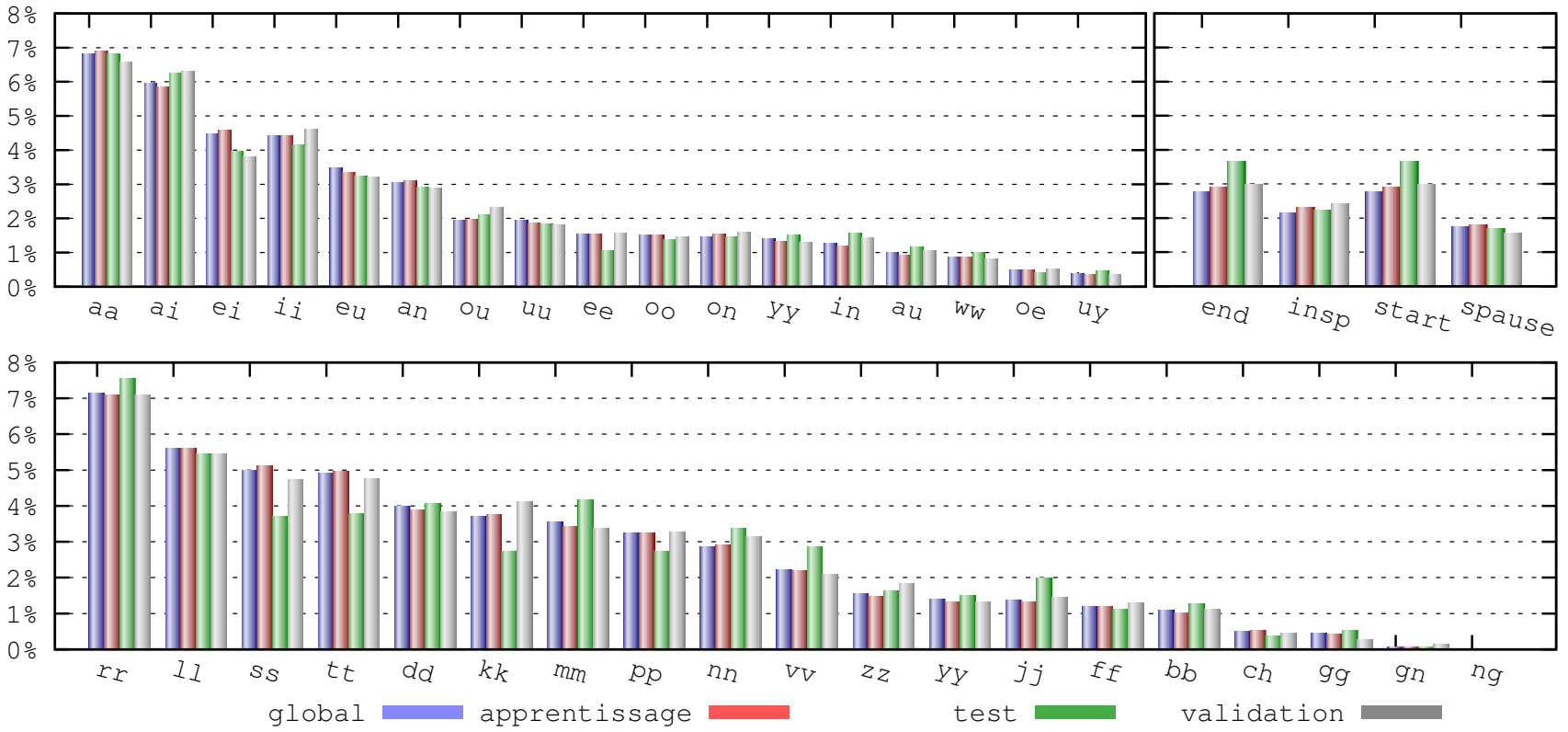


FIGURE 5.5 – Comparaison de la distribution des phonèmes pour le corpus global (en bleu), le corpus d'apprentissage (en rouge), le corpus de test (en vert) et le corpus de validation (en gris). L'axe des abscisses correspond aux phonèmes classés en fonction de leur taux de représentation dans le corpus global. Le graphe en bas illustre la distribution des consonnes, le graphe en haut à gauche illustre la distribution des voyelles et semi-voyelles, le graphe en haut à droite illustre la distribution des NSS (segments acoustiques hors parole).

5.3.4 Focus sur les syllabes

Pour analyser les syllabes du corpus CORDIAL, nous avons déterminer les distributions de structures syllabiques. Ces distributions sont illustrées dans la figure 5.6.

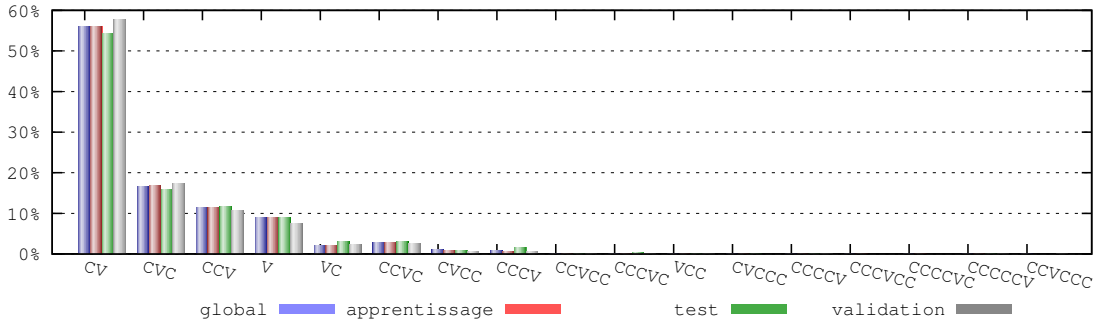


FIGURE 5.6 – Distribution des syllabes pour les corpus global (en bleu), d'apprentissage (en rouge), de test (en vert) et de validation (en gris) en fonction de leur structure.

Toute d'abord, les distributions associées au corpus CORDIAL et au sous-corpus utilisés sont proches du français. En effet, [Leon1992] indique que les trois structures dominantes du français sont CV (59.9%), CVC (17.1%) et CCV (14.1%). Dans un second temps, nous pouvons constater que les distributions sont relativement homogènes. Ceci indique que les résultats obtenus lors de l'évaluation ne présentent pas de biais concernant les descripteurs associés à la syllabe.

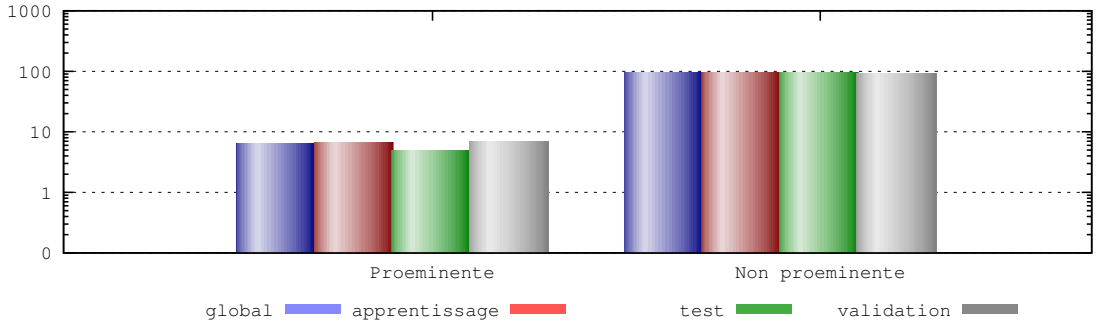


FIGURE 5.7 – Distribution des syllabes pour les corpus global (en bleu), d'apprentissage (en rouge), de test (en vert) et de validation (en gris) en fonction de leur proéminence.

De plus, en ce qui concerne la description d'une syllabe, le jeu de descripteurs proposé pour le français utilise la notion de proéminence. Pour compléter l'analyse du corpus CORDIAL, la figure 5.7 présente les distributions des syllabes en fonction de leur propriété de proéminence. Cette figure confirme que le découpage en sous corpus ne biaise pas la représentation syllabique car les distributions présentées dans cette figure sont également homogènes.

5.3.5 Focus sur les mots

Le dernier horizon analysé concerne le mot. Tout d'abord, la figure 5.8 présente la distribution des occurrences de mots en fonction du nombre de phonèmes qui les composent. Le premier constat est une homogénéité de taille de mots, en phonèmes, entre les différents corpus. Quelques différences subsistent mais elles sont, en proportion, mineures.

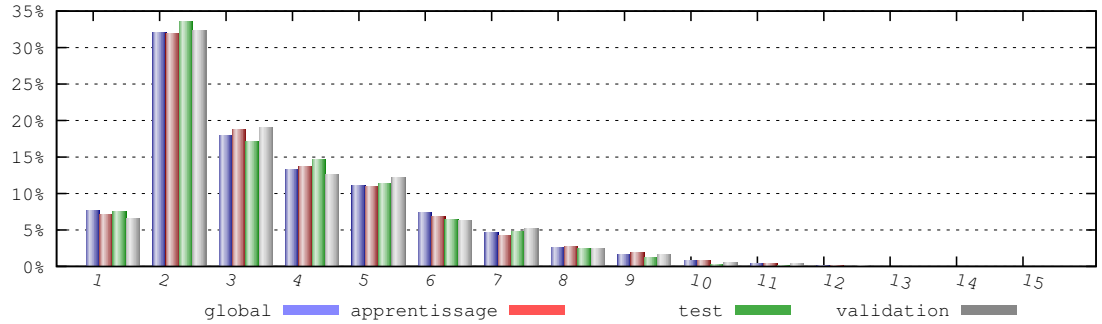


FIGURE 5.8 – Distribution des mots pour les corpus global (en bleu), d'apprentissage (en rouge), de test (en vert) et de validation (en gris) en fonction de leur taille en nombre de phonèmes.

Concernant le jeu des descripteurs proposés pour le français, la signifiance d'un mot est également prise en compte. Il est donc nécessaire de comparer les corpus en fonction de cette propriété ici réduite au fait que le mot soit un mot grammatical ou non. Ainsi, la figure 5.9 présente la distribution des mots en fonction de leur signifiance. Les résultats obtenus sont homogènes ce qui implique que l'analyse de la description à l'horizon d'un mot n'est pas biaisée. De plus, en comparant les proportions entre les trois catégories, nous constatons que la propriété de signifiance est discriminante. En effet, la moitié des mots présents dans le corpus sont considérés comme signifiants. L'autre moitié se répartie équitablement entre les mots non-signifiants et les mots pour lesquelles cette propriété n'est pas définie.

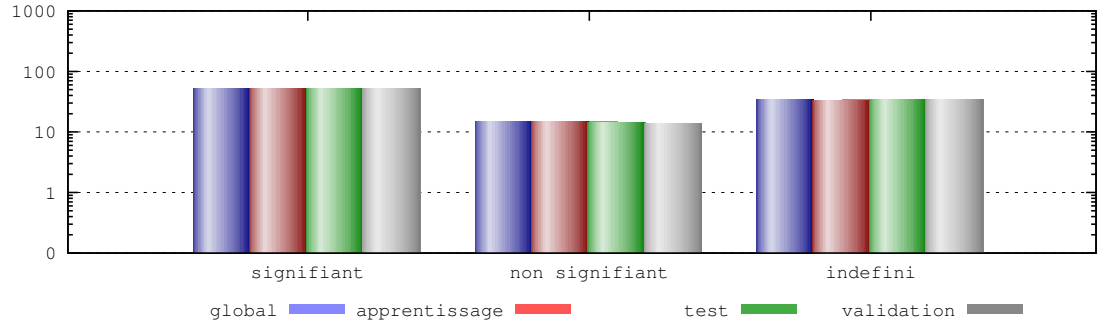


FIGURE 5.9 – Distribution des mots pour les corpus global (en bleu), d'apprentissage (en rouge), de test (en vert) et de validation (en gris) en fonction de leur propriété de signifiance.

5.4 Jeux de descripteurs évalués

Le choix des jeux de descripteurs qui seront évalués repose sur deux hypothèses :

– pour un horizon donné, les informations de position et les informations prosodiques sont supposées obtenues indépendamment les unes des autes,

– les informations à un horizon donné complètent les informations des horizons inférieurs.

En se basant sur ces deux hypothèses, nous avons catégoriser les jeux de descripteurs pour obtenir le découpage présenté dans le tableau 5.4. De cette manière, nous avons défini les jeux présentés dans la figure 5.5. Ce sont ces jeux de descripteurs qui seront évalués.

	Caté.	Description
Pho.		Label phonétique du segment courant
		Labels phonétiques des segments précédent/suivant
		Labels phonétiques des segments précédent-précédent/suivant-suivant
Syllabe		Nb. phones + pos. du phone courant dans la syl.
		Position de la syllabe dans le mot
		Position de la syllabe dans le syntagme
		Information d'accentuation
		Nb. de syl. depuis la dernière accent./syl. cour. jusque la syl. cour./prochaine acc.
		Nb. de syllabes acc. avant/après la syllabe courante dans le syntagme
		Voyelle de la syllabe
Mot		Nb. de syllabes dans le mot
		Position du mot dans le syntagme
		Tag grammatical du mot
		Nb. de mots depuis le dernier sign./mot cour. jusqu'au mot cour./prochain sign.
		Nb. de mots signifiant avant/après le mot cour. dans le syntagme
Synt.		Nb. de syl. dans le syntagme
		Nb. de mots dans le syntagme
		Pos. du syntagme dans l'énoncé

TABLE 5.4 – Catégorisation des descripteurs. Chaque couleur permet d'identifier une catégorie qui correspond à un couple (horizon, nature) permettant de qualifier un descripteurs. Les horizons sont indiqués dans la première colonne. Un descripteur peut être de nature positionnelle ou prosodique. Plus de détails sont disponibles dans l'annexe C.2.

5.4.1 Corpus

Le choix d'un jeu de descripteurs impacte directement sur le nombre de segments associés à une séquence de labels. En effet, plus le nombre de descripteurs utilisés est important, moins il y aura de représentants. Afin d'évaluer l'importance de ce phénomène dans notre étude, la figure 5.10 illustre l'évolution du nombre de segments moyens associés à un label en fonction du jeu de descripteurs.

Cette figure illustre le problème de dispersion de données du à l'utilisation de nombreux descripteurs. En effet, pour l'ensemble des jeux de descripteurs évalués excepté `p1`, un

Id	Signification	Code couleur
p1	Ph. courant	
p3	Ph. en contexte (3 ph)	
p5	Ph. en contexte (5 ph)	
p5-sy_pos	Inf. position de la syl	
p5-sy_accent	Inf. prosodique de la syl	
p5-sy_full	Inf. complète de la syl	
p5-w_pos	Inf. position du mot	
p5-w_content	Inf. prosodique du mot	
p5-w_full	Inf. complète du mot	
p5-s_pos	Inf. position du Syntagme	

TABLE 5.5 – Présentation des jeux de descripteurs utilisés dans les expériences. Les couleurs, présentés dans le tableau 5.4, sont utilisées pour indiquer les descripteurs d'un jeu donné.

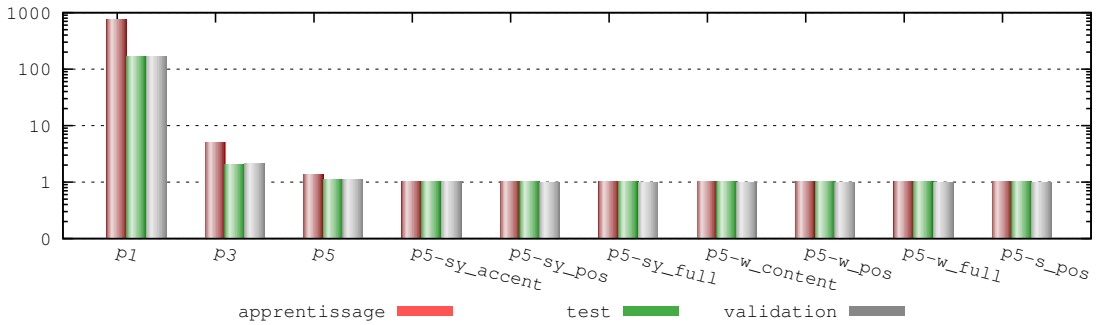


FIGURE 5.10 – Nombre moyen de segments associés à un label pour le corpus d'apprentissage (en rouge), le corpus de test (en vert) et le corpus de validation (en gris)

seul segment acoustique est associé à un label en contexte. Cette figure démontre ainsi la nécessité de l'utilisation des arbres de décision par HTS.

5.4.2 Arbres de décision et HTS

Jusqu'à présent nous avons proposé une caractérisation des corpus et des jeux de descripteurs indépendamment du système HTS. Il est néanmoins possible de jauger l'influence des descripteurs en utilisant des arbres de décision. Pour cela, il suffit d'analyser la proportion d'utilisation d'une catégorie de questions comme cela a été effectué dans [Watts2010].

Un taux élevé d'utilisation de noeuds associés à une catégorie de questions peut avoir deux implications :

- la catégorie comporte beaucoup de noeuds dans l'arbre (indicateur quantitatif),
- la catégorie comporte des noeuds proches de la racine de l'arbre (indicateur qualitatif).

Bien qu'il ne soit pas possible de discriminer ces indicateurs sans traitement supplémentaire,

cette proportion permet de se faire une première idée de l'influence d'une catégorie de descripteurs sur la modélisation effectuée par HTS.

La figure 5.11 illustre l'application de cette méthode pour le corpus d'apprentissage sur les quatre paramètres évalués. L'axe des abscisses est associé à la catégorie des questions utilisées, l'axe des ordonnées aux jeux de descripteurs évalués et le niveau de gris à la proportion d'utilisation d'une catégorie de questions pour un jeu de descripteurs.

Le premier point mis en avant par la figure 5.11 concerne la différence de traitement des paramètres. Contrairement à la durée et au F0, les arbres de décisions associés aux coefficients MGC et aux coefficients d'apériodicité semblent accorder peu d'importance aux descripteurs dont l'horizon est au-delà de la syllabe. Ce résultat indique que les descripteurs liés aux horizons *mot* et *syntagme* ne sont pas déterminants pour la génération du spectre et de l'apériodicité. En revanche, la modélisation du F0 et de la durée semble nécessiter l'utilisation de descripteurs d'horizons plus variés. Ainsi, en considérant l'ensemble des paramètres, la construction des arbres de décision semblent nécessiter l'ensemble des catégories de questions.

Il faut également noter la substitution d'un catégorie de noeuds par une autre. Ceci est particulièrement visible pour les catégories de position pour les horizons *syllabes* et *mots* pour le paramètre de durée. Ainsi, la proportion d'utilisation des questions liées à ces catégories, pour les jeux de descripteurs `p5-w_full` (toutes les informations sauf les informations de position à l'horizon du syntagme) et `p5-s_pos` (toutes les informations disponibles), est quasi-nulle alors que, pour les horizons inférieurs, son utilisation semble au même niveau que les autres catégories.

De même que pour le corpus d'apprentissage, les figures 5.12 et 5.13 illustrent les proportions pour, respectivement, le corpus de test et le corpus de validation. Contrairement à la phase d'apprentissage, l'utilisation de l'arbre consiste à sélectionner des distributions pour construire les modèles. Ainsi, ces figures permettent d'évaluer l'influence d'une catégorie de questions, et donc de descripteurs, directement sur la génération de paramètres.

Ces figures respectent la même topologie que la figure 5.11 pour le corpus de test et de validation. De plus, lors de la phase d'analyse des distributions des sous-corpus, il a été mis en avant une homogénéité forte entre les corpus de test, de validation et d'apprentissage. Ces deux constats impliquent que l'arbre de décision, obtenu lors de la phase d'apprentissage, est un bon compromis entre les différents corpus.

5.4.3 Cohérence STRAIGHT

Un dernier point à vérifier concerne la fréquence fondamentale. En effet, la nature de la modélisation de ce paramètre peut aboutir à la génération de valeurs ne respectant pas

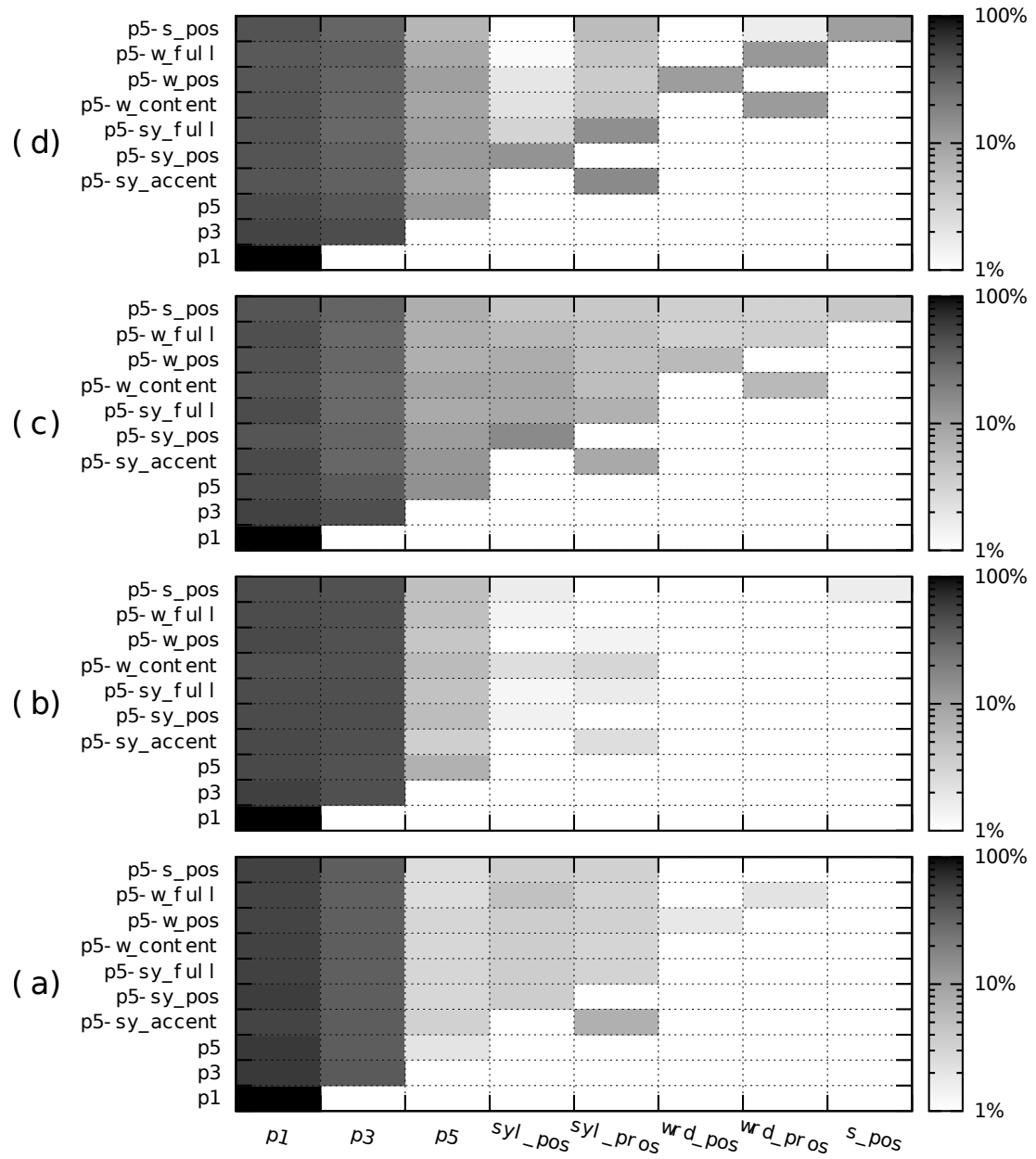


FIGURE 5.11 – Proportion des catégories de questions utilisées pour caractériser les segments du corpus d'apprentissage pour les coefficients MGC (a), les coefficients d'apériodicité (b), le log(F0) (c) et la durée (d). Cette proportion est calculée pour chacun des jeux de descripteurs. L'axe des abscisses identifie les catégories de questions : `p1`, `p3`, `p5` correspondent aux questions liées à l'échelle du segment phonétique et en tenant compte, respectivement des horizons 0, 1 et 2 ; `syl_pos`, `wrd_pos` et `s_pos` correspondent aux informations de position pour, respectivement, les horizons de la syllabe, du mot et de la phrase ; `syl_pros` et `wrd_pros` correspondent aux informations d'accentuation pour, respectivement, les horizons de la syllabe et du mot.

les plages précédemment indiquées. Par exemple, si la moyenne de la distribution associée à la dérivée d'ordre 1 d'une trame a une forte valeur absolue et que la distribution devant générer la valeur F0 de la trame suivante a une espérance proche d'une de ces limites, l'algorithme de génération peut générer une valeur F0 hors limite. Il est donc important

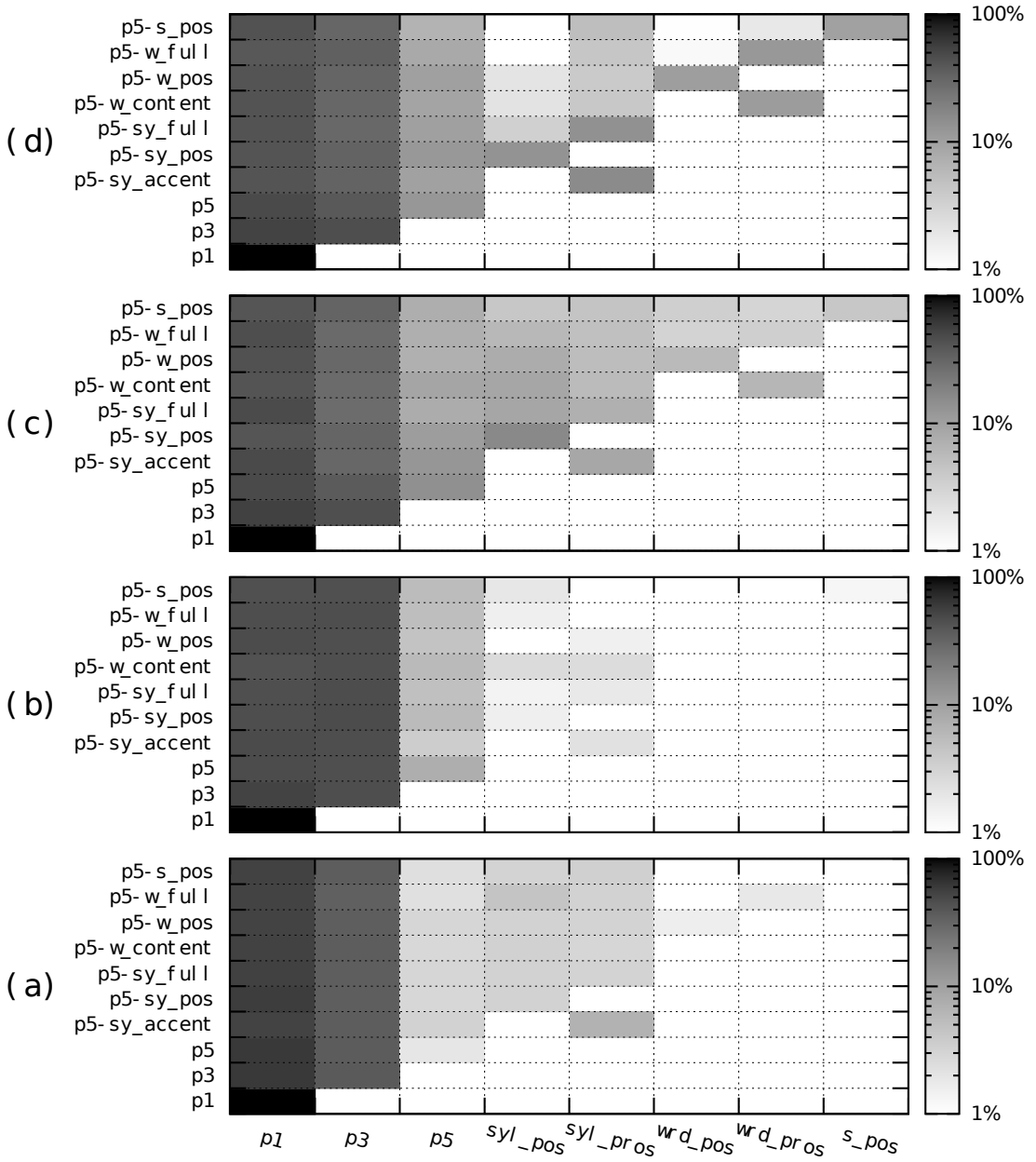


FIGURE 5.12 – Proportion des catégories de questions utilisées pour caractériser les segments du corpus de test pour les coefficients MGC (a), les coefficients d'apériodicité (b), le log(F0) (c) et la durée (d). Cette proportion est calculée pour chacun des jeux de descripteurs et représentée par un niveau de gris à l'échelle logarithmique. L'axe des abscisses identifie les catégories de questions : `p1`, `p3`, `p5` correspondent aux questions liées à l'échelle du segment phonétique et en tenant compte, respectivement des horizons 0, 1 et 2 ; `syl_pos`, `wrd_pos` et `s_pos` correspondent aux informations de position pour, respectivement, les horizons de la syllabe, du mot et de la phrase ; `syl_pros` et `wrd_pros` correspondent aux informations d'accentuation pour, respectivement, les horizons de la syllabe et du mot.

de vérifier la proportion des trames hors-limites.

La figure 5.14 résume les proportions obtenues selon quatre catégories : les valeurs

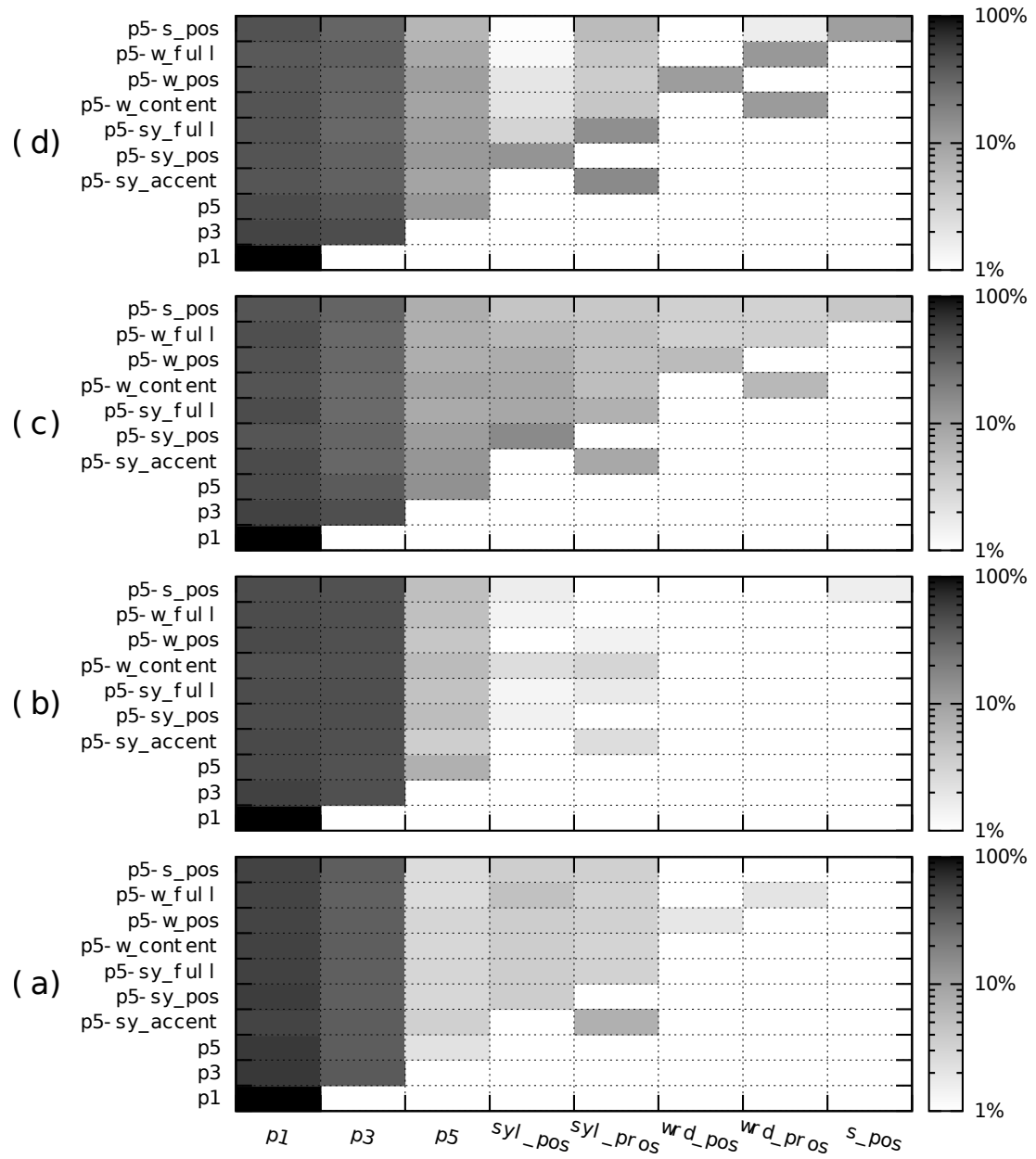


FIGURE 5.13 – Proportion des catégories de questions utilisées pour caractériser les segments du corpus de validation pour les coefficients MGC (a), les coefficients d'apériodicité (b), le log(F0) (c) et la durée (d). Cette proportion est calculée pour chacun des jeux de descripteurs. L'axe des abscisses identifie les catégories de questions : `p1`, `p3`, `p5` correspondent aux questions liées à l'échelle du segment phonétique et en tenant compte, respectivement des horizons 0, 1 et 2 ; `syl_pos`, `wrd_pos` et `s_pos` correspondent informations de position pour, respectivement, les horizons de la syllabe, du mot et de la phrase ; `syl_pros` et `wrd_pros` correspondent aux informations d'accentuation pour, respectivement, les horizons de la syllabe et du mot.

F0 non-voisées, les valeurs F0 voisées mais inférieures au minimum autorisé, les valeurs F0 voisées valides et les valeurs F0 voisées mais supérieures au maximum autorisé. Les résultats présentés montrent que les proportions associées aux plages non valides sont très proches et faibles ($< 1\%$) sauf pour le jeu `p5-sy_accent` ou cela représente environ 3%

quant aux valeurs F0 supérieures au maximum autorisé.

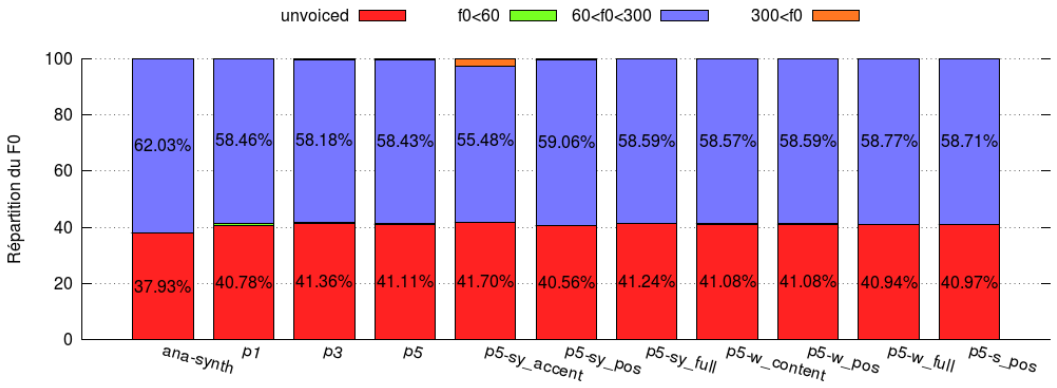


FIGURE 5.14 – Répartition des valeurs du F0 générées pour chacun des jeux de descripteurs analysés.

5.5 Conclusion

Dans ce chapitre, nous avons présenté le corpus CORDIAL ainsi que les jeux de descripteurs qui sont utilisés pour effectuer les évaluations. Cette présentation s'est déroulée en plusieurs étapes.

Tout d'abord, nous avons introduit la structure de données ROOTS qui permet de représenter le corpus. L'apport majeur de cette structure de données est l'introduction de la notion de composition de matrices. Ce concept permet de déterminer la totalité des liens entre les différentes annotations en se basant sur un ensemble de relations minimal. Nous avons ensuite présenté le processus d'annotation automatique qui se base sur les propriétés de la structure ROOTS.

La troisième partie de ce chapitre est consacrée à la présentation du corpus CORDIAL ainsi que des *sous-corpus* utilisés pour les évaluations. Pour cela, nous avons présenté des statistiques permettant de qualifier l'ensemble de ces corpus. De cette manière, nous avons ainsi pu valider que chaque corpus, utilisé lors de l'évaluation, ne comporte pas de trait particulier qui pourrait biaiser l'analyse.

Enfin, nous avons présenté les jeux de descripteurs que nous avons évalués lors de nos expériences. Nous avons également proposé une première analyse basée sur les arbres de décision déterminés par HTS. Cette analyse préliminaire nous fournit une première mesure de l'influence d'un descripteur sur la modélisation effectuée par HTS. Toutefois, cette mesure ne permet pas d'évaluer la qualité de la synthèse effectuée par HTS et l'objectif de nos travaux est de pouvoir effectuer cette évaluation.

Conclusion à la deuxième partie

Cette partie a permis de présenter les méthodes d'évaluation et les données expérimentales qui vont nous permettre de pouvoir analyser l'influence des descripteurs sur la modélisation et la synthèse effectuée par HTS.

Dans le premier chapitre de cette partie, nous avons présenté deux protocoles permettant d'analyser l'influence des descripteurs sur la modélisation effectuée par HTS. Le premier protocole consiste à modéliser l'espace acoustique généré par HTS par un GMM. En calculant la log-vraisemblance des vecteurs acoustiques extraits du signal naturel sur le GMM des vecteurs générés par HTS. Nous obtenons ainsi une quantification de la dégradation issue de la modélisation effectuée par HTS. Néanmoins, l'application de ce protocole nécessite un ensemble important de vecteurs pour la phase d'apprentissage du GMM. Nous proposons un second protocole qui permet de calculer un écart entre trames alignées. En effectuant un changement d'échelle et en rassemblant les écarts selon les catégories que nous souhaitons analyser, nous pouvons évaluer la modélisation effectuée par HTS à une échelle plus locale.

Nous avons utilisé un corpus de parole en langue française. Ce corpus et le processus d'annotation automatique associé ont été présentés dans le second chapitre de cette partie. Les jeux de descripteurs qui seront évalués ont également été exposés. Les sous-corpus pour effectuer les expériences ont été analysés.

À l'issue de cette partie, nous avons donc présenté l'ensemble des données nécessaires pour nos évaluations. Les résultats obtenus sont exposés et analysés dans la partie suivante.

Troisième partie

Évaluation HTS

-

Résultats pour le français

Introduction à la troisième partie

Dans la seconde partie de ce document, nous avons présenté les protocoles expérimentaux conçus pour évaluer l'impact des descripteurs sur la synthèse effectuée par HTS. Nous avons également présenté le corpus utilisé pour effectuer ces évaluations. Dans cette partie, nous allons présenter les résultats obtenus après application de ces protocoles ainsi que les résultats d'évaluations subjectives permettant de valider les résultats obtenus par les analyses objectives. Cette partie se décompose en trois chapitres.

Le premier chapitre (intitulé « Évaluation objective - Évaluation par GMM », page 119) présente les résultats obtenus en appliquant le protocole modélisant l'espace acoustique généré par HTS par un GMM. Le protocole a été appliqué pour évaluer quatre types de coefficients : la fréquence fondamentale, les coefficients MGC, la durée ainsi que les coefficients d'apériodicité.

Cette évaluation restant globale, le second chapitre (intitulé « Évaluation objective - Évaluation non paramétrique », page 137), présente les résultats obtenus par le second protocole pour les quatre types de coefficients précédemment mentionnés. L'analyse effectuée pour l'ensemble de ces paramètres acoustiques est donc plus locale et a pour objectif de déterminer plus précisément l'influence des descripteurs utilisés sur les modèles.

Le dernier chapitre de cette partie (intitulé « Évaluation subjective », page 163) est consacré aux évaluations subjectives conduites pour confronter les résultats obtenus par les protocoles présentés dans la partie précédente. La première évaluation subjective consiste à évaluer globalement la synthèse effectuée par HTS selon différents jeux de descripteurs afin d'éprouver les résultats obtenus par les évaluations objectives. La seconde évaluation subjective a pour objectif de comparer la dégradation due à la synthèse par rapport au signal naturel. Le dernier test subjectif consiste à évaluer globalement la synthèse effectuée par HTS sur des énoncés qui ne sont pas issus du corpus Cordial.

Chapitre 6

Évaluation objective - Évaluation par GMM

Le protocole d'évaluation objective basé sur la modélisation de l'espace acoustique par un GMM a été appliqué afin d'analyser l'impact d'un jeu de descripteurs sur la synthèse effectuée par HTS. Ce protocole a été appliqué aux quatre paramètres utilisés par ce système : les coefficients MGC, le F0, les coefficients d'apériodicité et la durée. Afin d'évaluer les trois premiers paramètres indépendamment de la durée, une première génération a été effectuée en imposant la durée des modèles. Une seconde génération, sans contrainte, a ensuite été effectuée afin d'obtenir une durée estimée par HTS. Les coefficients MGC, le F0 et les coefficients d'apériodicité étant isolés dans des flux HTK distincts, nous pouvons évaluer chacun des paramètres acoustiques indépendemment les uns des autres.

Comme cela a été indiqué précédemment, nous avons généré trois des quatre paramètres en imposant la durée que doit couvrir le HMM. Nous supposons qu'à l'issue de la génération, l'énoncé généré et l'énoncé *original* sont donc alignés. Ce chapitre débute par la présentation d'une étude préliminaire effectuée pour valider cet alignement. Ce chapitre se poursuit par la présentation des résultats de l'évaluation pour les coefficients MGC, le F0, l'apériodicité et enfin la durée.

6.1 Étude préliminaire

L'objectif de la méthode est d'évaluer objectivement chacun des paramètres générés par HTS. Pour cela, il est nécessaire de s'assurer de l'indépendance de chacun de ces paramètres. Intrinsèquement, le système HTS voit la représentation spectrale, la fréquence fondamentale et l'apériodicité comme des paramètres indépendants. Néanmoins, la trajectoire de ces paramètres est conditionnée par la chaîne des états non observés. Le nombre de ces états est directement relié à la durée des HMM utilisés et HTS permet de générer une trajectoire en imposant la durée du phone/HMM.

L'étape de segmentation utilisée lors du processus d'annotation du corpus permet d'obtenir une durée pour chaque segment acoustique. Pour le corpus d'analyse-synthèse, nous disposons donc d'une durée de référence à l'horizon du phone. Lors de la phase de génération, nous avons indiqué à HTS que la durée des phones à produire doit correspondre à cette référence. Toutefois, il est nécessaire que la durée réellement produite par HTS ne diverge pas de cette durée de référence.

En effet, le système HTS permet d'utiliser la durée issue des fichiers de labels pour imposer la durée des modèles utilisés lors de la phase de génération. Néanmoins, la topologie des modèles impose une durée minimale de 25ms [1] pour chaque segment. De plus, la génération se fait via le paramètre ρ, permettant de contrôler le débit syllabique (voir section 2.3.2 du chapitre 2), plutôt que par l'utilisation des durées directement. Ainsi, à cause de ces deux points, la durée générée peut différer de la consigne définie dans les fichiers de labels HTS.

Pour évaluer l'influence de cette possible divergence, nous avons déterminé trois mesures. La première correspond à l'erreur moyenne en nombre de trames. La deuxième correspond au pourcentage de segments acoustiques alignés en autorisant une trame de décalage. La dernière mesure correspond au pourcentage d'énoncés dont la durée de référence et la durée générée diffèrent de plus de 5 trames. Le tableau 6.1 résume les résultats obtenus.

Les résultats présentés dans la première colonne montrent que la taille moyenne du décalage entre la durée de référence et celle générée par HTS est inférieure à 1 trame. En autorisant une trame de décalage par segment, on constate que plus de 95% des segments sont alignés. En autorisant cinq trames de décalage par phrase, seules quelques phrases ne sont pas alignées ce qui tend à indiquer que HTS compense le décalage obtenu à l'horizon du segment de manière globale.

Plusieurs hypothèses peuvent être émises sur ces décalages :

– des erreurs d'arrondis : contrairement à la durée de référence du segment, HTS ne peut générer que des durées multiples de 5ms ;

1. Les modèles appris sont des MSD-HSMM composés de 5 émetteurs et le décalage de trame utilisé est de 5ms.

	err. moy. (nb. trames)	% seg. ali.	% énoncé non al.
p1	0.75	96.71	0.00
p3	0.39	96.09	1.32
p5	0.68	96.96	0.66
p5-sy_accent	0.54	96.73	0.66
p5-sy_pos	0.76	96.99	0.66
p5-sy_full	0.76	96.99	0.66
p5-w_content	0.63	96.87	0.66
p5-w_pos	0.75	96.96	0.66
p5-w_full	0.73	97.09	0.66
p5-s_pos	0.75	96.90	0.66

TABLE 6.1 – Validation de la durée générée en mode *durée imposée*. La première colonne présente l'erreur moyenne en nombre de trames pour l'ensemble du corpus de test. La seconde colonne présente le pourcentage de segments alignés à une trame près. La troisième colonne présente le pourcentage d'énoncés qui ne sont pas alignés malgré un delta de 5 trames autorisées.

– la durée du segment original est inférieure à la durée minimale d'un segment que peut modéliser HTS : la topologie des modèles HTS impose le passage par l'ensemble des états et l'émission d'au moins une valeur. Cela implique donc que la durée générée est forcément supérieure ou égale au produit du nombre d'états par la durée d'un état émis : soit dans notre cas 25ms.

Néanmoins, ces résultats montrent que la durée générée en mode *durée imposée* par HTS peut être considérée comme fiable. De plus, comme nous l'avons indiqué dans la section 2.2.1 du chapitre 2, la modélisation utilisée consiste à associer à chaque type de coefficients (MGC, BAP et F0) un flux dédié. Ainsi, lors de la phase de synthèse, HTS génère indépendamment chacun de ces flux. Nous pouvons donc supposer qu'en générant les coefficients MGC, le F0 et les coefficients BAP pour la durée imposée et qu'en générant séparément la durée, les valeurs obtenues pour un type de coefficient ne dépendent pas d'un autre type de coefficient.

6.2 Évaluation du F0

Le premier paramètre évalué est la fréquence fondamentale. Ce paramètre étant mono-dimensionnel, nous avons décidé de le compléter en prenant en compte la dynamique de premier ordre. L'objectif est d'intégrer la notion d'évolution temporelle qui est inhérente à une mélodie. En revanche, nous avons ignoré la dynamique de second ordre. En effet, une étude réalisée par Y. Chen ET AL.[Chen2010a], qui avait pour objectif d'évaluer l'apport de la dérivée seconde sur la génération, a montré que cette dérivée n'avait qu'un impact à court terme et permettait de lisser la courbe de F0. En ignorant cette composante, la génération aboutit à une courbe en *dents de scie* telle que l'illustre la figure 6.1.

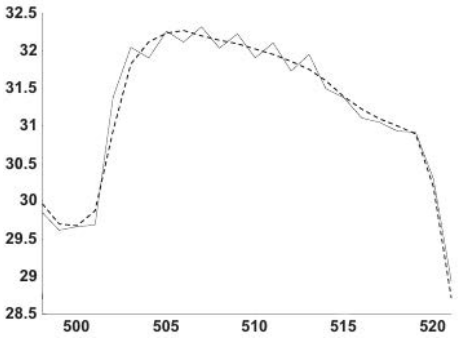


FIGURE 6.1 – Influence de la dérivée seconde sur la génération effectuée par HTS. En abscisse sont indiqués les indices de trames et en ordonnée l'amplitude en décibel. La courbe en pointillé prend en compte la dérivée seconde et la courbe en trait plein est obtenue en ignorant cette dérivée. Figure extraite de [Chen2010a].

Ainsi, pour effectuer l'évaluation du F0, nous proposons donc d'utiliser les vecteurs de coefficients v tels que :

$$v = [F0, \Delta F0] \tag{6.1}$$

où seules les trames voisées ($F0 \neq 0$), et dont la dynamique peut être définie [2], sont prises en compte. Les résultats de l'évaluation du F0 obtenus en appliquant le protocole d'évaluation par GMM sont présentés figure 6.2.

Si l'on considère le corpus de test $T_{a/s}$, il est naturellement le plus vraisemblable pour le modèle $\mathcal{M}_{a/s}$. Par contre, c'est par rapport au modèle $\mathcal{M}_{p1}$ que les éléments de $T_{a/s}$ sont les moins vraisemblables. Afin de quantifier l'amélioration de cette vraisemblance relativement aux modèles associés aux autres jeux de descripteurs, un ratio r, défini ci-dessous, est calculé et indiqué dans la figure 6.2 pour chaque jeu de descripteurs autres que `a/s` et `p1` :

$$r = \frac{LL(T_{a/s}; \mathcal{M}_k) - LL(T_{a/s}; \mathcal{M}_{a/s})}{LL(T_{a/s}; \mathcal{M}_{p1}) - LL(T_{a/s}; \mathcal{M}_{a/s})} * 100 \tag{6.2}$$

où $k \notin \{\texttt{a/s}, \texttt{p1}\}$.

Grâce à cette figure, nous pouvons constater que le nombre de composantes $n_k^\star$ de chaque GMM, présenté en abscisse secondaire, varie de 64 à 256. Lors de la phase d'estimation des paramètres par l'algorithme E.M., les variances de certaines composantes étaient trop faibles pour que la matrice de covariance associée puisse être inversée. En l'état actuel, il est difficile de pouvoir conclure car le nombre de composantes varie fortement entre les différents GMM. Ainsi, afin de pouvoir croiser les résultats, nous avons contraint le nombre de composantes au nombre minimal possible, à savoir 64 par GMM. Nous obtenons alors les résultats présentés dans la figure 6.3.

Tout d'abord, nous pouvons constater que les résultats illustrés par les figures 6.2 et 6.3 sont extrêmement proches. En effet, l'écart maximum, entre les log-vraisemblances $LL(A_k; \mathcal{M}_k(n_k^\star))$, $LL(T_k; \mathcal{M}_k(n_k^\star))$ et $LL(T_{a/s}; \mathcal{M}_k(n_k^\star))$ dans le cadre où le nombre de

2. Si une trame se situe en frontière de voisement, la dynamique ne peut pas être définie.

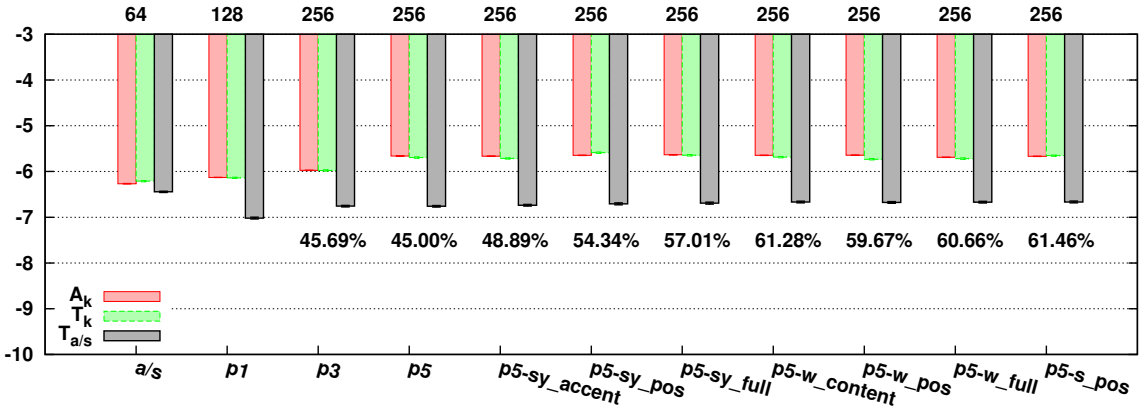


FIGURE 6.2 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace du F0. Le pourcentage d'amélioration de la log-vraisemblance, apporté par le jeu de descripteurs k, par rapport au jeu p1 est indiqué en dessous de la barre associée à k. Enfin, les intervalles de confiance à 95% sont représentés.

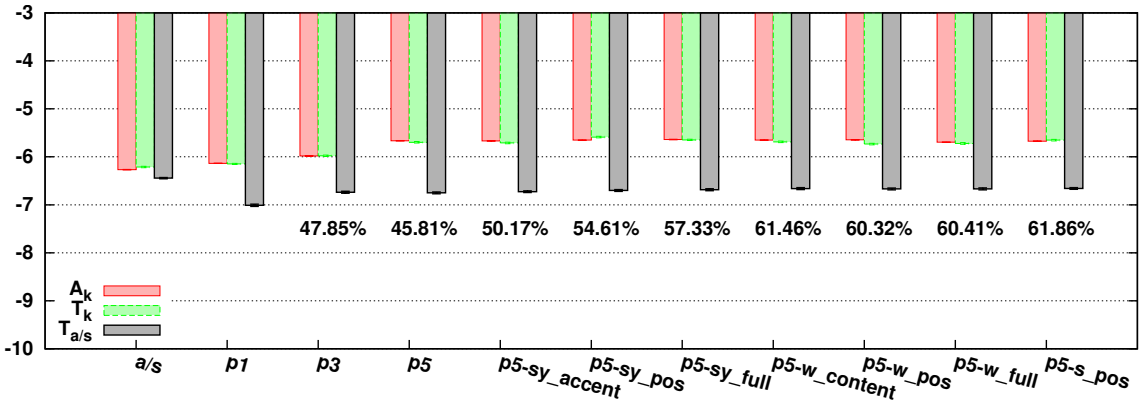


FIGURE 6.3 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace du F0 en limitant le nombre de composantes à 64. Le pourcentage d'amélioration de la log-vraisemblance, apporté par le jeu de descripteurs k, par rapport au jeu `p1` est indiqué en dessous de la barre associée à k. Enfin, les intervalles de confiance à 95% sont représentés.

composantes $n_k^\star$ est imposé et celui où il n'est pas imposé, est de 0.02. Cette stagnation des log-vraisemblances montre que la prise en compte de plus de 64 composantes n'améliore pas significativement la modélisation de l'espace acoustique effectuée par un GMM.

Lorsque l'on considère les intervalles de confiance pour $LL(T_{a/s}; \mathcal{M}_k(n_k^\star))$, pour chaque jeu de descripteurs $k \neq a/s$, deux améliorations significatives se distinguent. La première, et sans doute la plus significative, est apportée par la prise en compte du contexte phonétique direct. En effet, l'utilisation du jeu de descripteurs `p3` réduit d'environ 50% l'écart entre la log-vraisemblance relative à $\mathcal{M}_{p1}$ et celle associée à $\mathcal{M}_{a/s}$. Une seconde amélioration de la log-vraisemblance se distingue lors de la prise en compte des informations prosodiques (`p5-sy_accent`) ou de position (`p5-sy_pos`) au niveau de la syllabe par rapport au jeu de descripteurs `p5`. Bien que le ratio indiqué montre une amélioration de 4% entre `p5-sy_full` et `p5-w_content`, les intervalles de confiance montrent que cette

différence n'est pas significative. En réalité, à partir du jeu de descripteurs `p5-sy_full`, les améliorations ne sont pas significatives. À l'issue de cette analyse, il semble donc que le jeu `p5-sy_full` offre le meilleur compromis entre le nombre de descripteurs nécessaires pour qualifier un segment acoustique et la qualité de la modélisation du F0 effectuée par HTS compte tenu de ce protocole.

Enfin, les NSS constituent une partie importante du corpus global (voir figure 5.5 du chapitre 5). Toutefois, il s'agit de segments particuliers qui ne sont généralement pas utilisés lors de la phase de synthèse. Il nous semble important d'effectuer une évaluation sans tenir compte de ces segments pour déterminer la qualité de modélisation des phones en contexte. Les résultats illustrés par la figure 6.4 correspondent à l'application du protocole en ignorant les NSS présents dans les corpus lors de l'apprentissage des GMM et lors de l'évaluation.

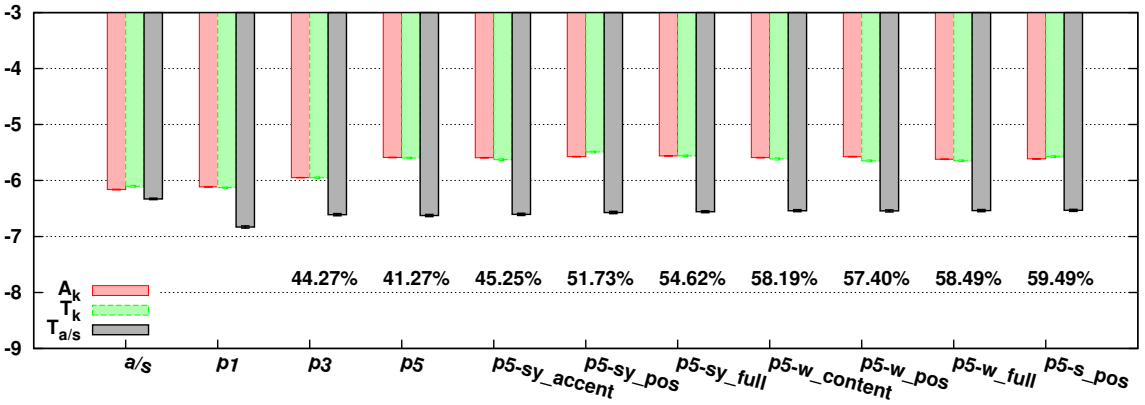


FIGURE 6.4 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace du F0 en limitant le nombre de composantes à 64 et en ignorant les NSS. Le pourcentage d'amélioration de la log-vraisemblance, apporté par le jeu de descripteurs k, par rapport au jeu `p1` est indiqué en dessous de la barre associée à k. Enfin, les intervalles de confiance à 95% sont représentés.

Ces résultats sont proches de ceux qui ont été obtenus en prenant en compte les NSS. Ceci indique que l'utilisation d'un descripteur plutôt qu'un autre n'influe pas sur la modélisation des NSS. Ce résultat était attendu puisque seules les trames voisées sont prises en compte lors de cette évaluation et que, par définition, les trames associées au NSS sont majoritairement non voisées. Ainsi, à l'issue de cette évaluation, le jeu de descripteurs `p5-sy_full` est considéré comme le jeu optimal pour modéliser la prosodie.

6.3 Évaluation de la modélisation spectrale

Le premier paramètre que nous avons évalué sont les coefficients MGC. HTS utilise un vecteur de 40 coefficients MGC pour modéliser le filtre associé au conduit vocal. Néanmoins, comme nous l'avons indiqué lors de la présentation des coefficients MGC dans la section 1.1.3 du chapitre 1.1, la première dimension correspond à l'énergie du

signal. Pour évaluer la modélisation spectrale, nous allons donc appliquer le protocole d'évaluation en tenant compte de l'énergie dans un premier temps et en ignorant l'énergie dans un second temps.

De plus, comme nous l'avons vu dans la section 4.1.2 du chapitre 4, la dimension élevée de ces vecteurs peut entraîner des problèmes de stabilité numérique. Pour évaluer les coefficients MGC, nous avons donc calculer la transformation Π_k telle que définie dans le protocole.

6.3.1 Validation de l'ACP

Avant d'utiliser une ACP Π_k, déterminée pour chaque corpus A_k, il est nécessaire de s'assurer que cette transformation appliquée sur le corpus $T_{a/s}$ n'aura aucune influence sur les résultats obtenus. En effet, lors de la phase 3 du protocole (section 4.1.3 du chapitre 4), la comparaison se fera en utilisant le corpus $\Pi_k(T_{a/s})$. Si la dégradation des données de $T_{a/s}$, due à l'application de Π_k, varie selon le jeu de descripteurs k utilisé, alors nous pourrons difficilement comparer les quantités $LL(\Pi_k(T_{a/s}); \mathcal{M}_k)$ et, en conséquence, les espaces générés par HTS en fonction des jeux de descripteurs.

Afin d'évaluer la dégradation inhérente à l'application de l'ACP, nous avons reconstruit pour chaque élément s de $T_{a/s}$ « l'image réciproque de $\Pi_k(s)$ par Π_k » en effectuant le changement de base inverse et en complétant les coordonnées du vecteur $\Pi_k(s)$ relativement aux derniers axes principaux par des valeurs nulles. Par abus de notation, l'ensemble des éléments ainsi obtenus est noté de la façon suivante :

$$T'_{a/s} = \Pi_k^{-1}(\Pi_k(T_{a/s}))\,. \tag{6.3}$$

Nous avons ensuite calculé une distance entre les corpus $T_{a/s}$et $T'_{a/s}$ basée sur la distance euclidienne en décibel de la façon suivante [3] :

$$d(T'_{a/s}, T_{a/s}) = \left(\frac{10}{ln(10)}\sqrt{2.0}\right) * \frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{\sum_{n=1}^{N}(T'_{a/s}(t,n) - T_{a/s}(t,n))^2}\right) \tag{6.4}$$

où $T_{a/s}(t,n)$ désigne la n-ème coordonnée de la t-ème trame de $T_{a/s}$ (respectivement $T'_{a/s}(t,n)$ pour $T'_{a/s}$), T le nombre total de trames dans $T_{a/s}$ et N la dimension des vecteurs MGC.

Nous évaluerons trois cas d'application de l'ACP :

– La prise en compte de l'ensemble des coefficients de $T_{a/s}$;

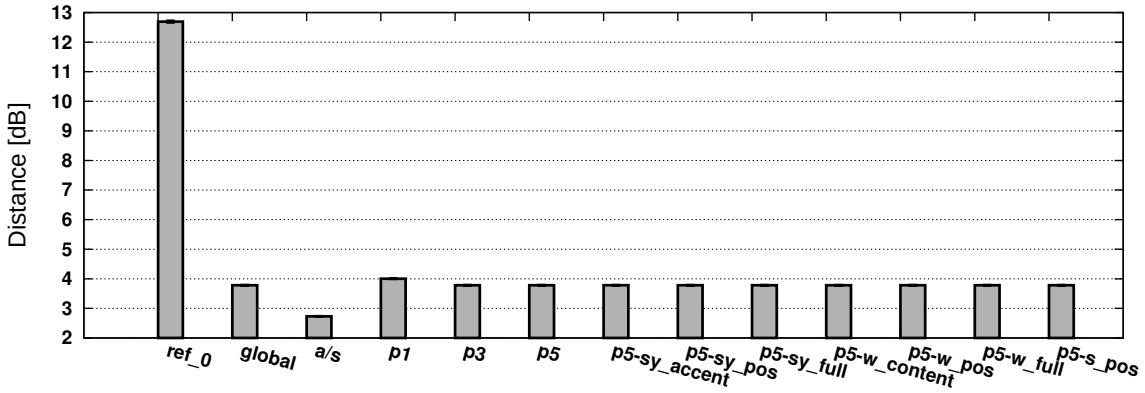
– La prise en compte des coefficients de $T_{a/s}$ associés à des trames qui ne sont pas

3. Dans le cas où l'énergie n'est pas prise en compte, cette équation correspond à la distance mel-cepstrale proposée dans SPTK

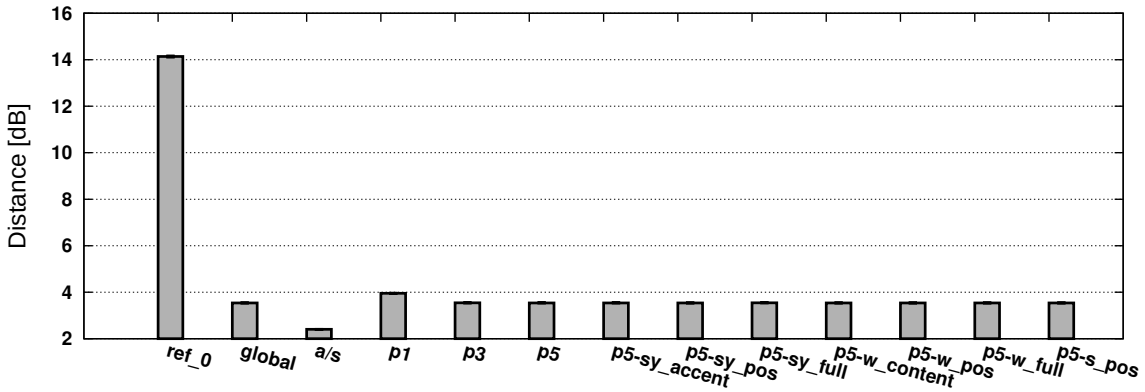
étiquetées comme des NSS ;

– La configuration précédente en ignorant la première dimension qui correspond à l'énergie du signal.

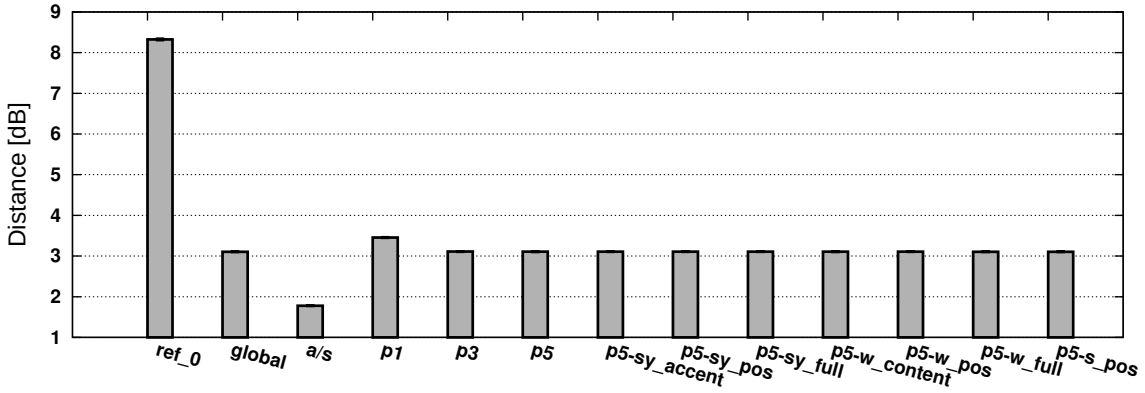
Les résultats obtenus sont illustrés figure 6.5.



(a) Configuration standard



(b) Configuration sans NSS



(c) Configuration sans NSS et sans énergie

FIGURE 6.5 – Erreur obtenue par l'application de la transformation Π_k sur le corpus de test $T_{a/s}$. La colonne *global* correspond à la transformation Π. ref_0 correspond à la distance $d(0, T_{a/s})$. Les intervalles de confiance représentés sont à 95%.

Cette figure montre que les résultats diffèrent selon les jeux de descripteurs pour l'ensemble des configurations avec une accentuation visible lorsque l'énergie n'est pas prise en compte. En effet, `a/s` obtient une dégradation significativement plus faible ce qui est

naturellement attendu. De plus, `p1` obtient une dégradation significativement plus élevée et les résultats associés aux autres jeux de descripteurs sont équivalents.

Deux dégradations supplémentaires ont été intégrées : la première (`ref_0`) correspond au calcul de $d(T_{a/s}, 0)$ et permet d'obtenir une référence à laquelle nous pouvons comparer les dégradations ; la seconde (`global`) correspond à la dégradation $d(T'_{a/s}, T_{a/s})$ où $T'_{a/s}$ a été déterminée en utilisant la transformation Π calculée sur l'ensemble des corpus A_k.

En se basant sur ces deux distorsions, nous pouvons compléter nos résultats. Tout d'abord, les différences constatées entre les distorsions sont faibles relativement à la distance par rapport à 0. Ensuite, l'application d'une ACP globale aboutit à une distorsion semblable à celles associées aux jeux de descripteurs autres que `a/s` et `p1`. Ceci peut indiquer que la masse des données associées à ces jeux de descripteurs couvre l'influence des données associées à `a/s` et `p1`. Néanmoins, afin de garantir que l'utilisation de l'ACP ne fausse pas les analyses ultérieures, nous décidons d'utiliser la transformation Π calculée sur les trames issues de l'ensemble des corpus A_k.

6.3.2 Résultat de l'évaluation

L'évaluation de la modélisation des coefficients MGC aboutit aux résultats présentés dans la figure 6.6. Comme pour le F0, nous avons contraint le nombre de composantes des GMM pour pouvoir comparer les log-vraisemblances obtenues. Ainsi chaque GMM utilisé pour obtenir ces résultats contient 256 composantes.

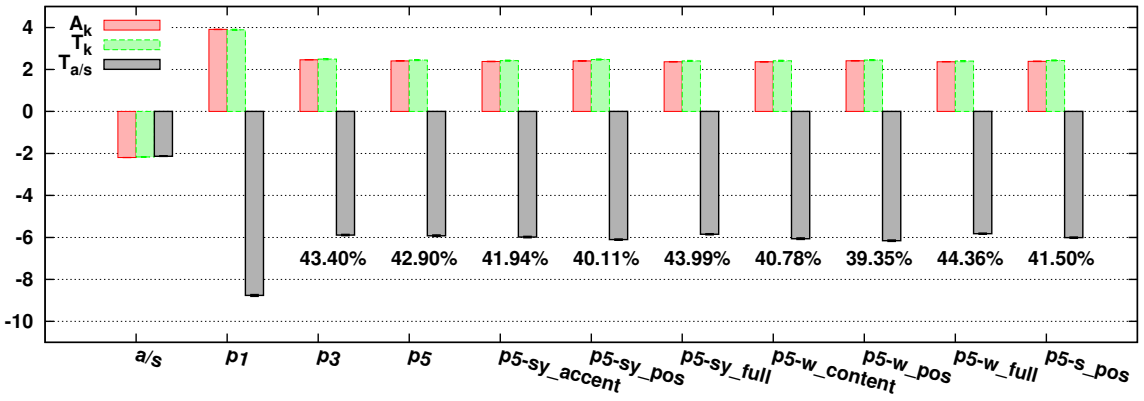


FIGURE 6.6 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace spectral

Tout d'abord, pour tout $k \in \{a/s, \ldots, p5\_full\}$, nous pouvons remarquer que les log-vraisemblances des corpus A_k et T_k relativement à $\mathcal{M}_k$ sont quasi-identiques. Cela valide le fait que, pour chacun des GMM, il n'y a pas de situation de sur-apprentissage. Nous observons également que les quantités $LL(A_{a/s}; \mathcal{M}_{a/s})$ et $LL(T_{a/s}; \mathcal{M}_{a/s})$ sont plus faibles que $LL(A_k; \mathcal{M}_k)$ et $LL(T_k; \mathcal{M}_k)$ pour $k \neq a/s$. Le nombre de composantes n étant identique, la génération des paramètres spectraux réalisée par HTS semble donc réduire la

variabilité des paramètres par rapport à ceux extraits du signal original malgré l'utilisation de la variance globale. Pour vérifier cela, la moyenne des variances, notée $\nu(\Sigma_k)$, a été calculée pour chacun des GMM. Les résultats tableau 6.2 présentent le ratio entre $\nu(\Sigma_k)$ et $\nu(\Sigma_{a/s})$ ainsi que la moyenne des variances.

Jeu de desc.	a/s	p1	p3	p5	p5-sy_accent	p5-sy_pos
Var. moyenne	0.0219	0.0036	0.0053	0.0052	0.0054	0.0053
Ratio % $\mathcal{M}_{a/s}$	1	13	6.9	6.9	6.8	6.9

Jeu de desc.	p5-sy_full	p5-w_content	p5-w_pos	p5-w_full	p5-s_pos
Var. moyenne	0.0053	0.0055	0.0055	0.0054	0.0054
Ratio % $\mathcal{M}_{a/s}$	6.9	6.7	6.8	6.7	6.7

TABLE 6.2 – Variance moyenne par jeux de descripteurs et ratio

Ainsi, les variances du GMM $\mathcal{M}_{p1}$ sont, en moyenne, 13 fois plus faibles que celles de $\mathcal{M}_{a/s}$ et les variances associées aux autres GMM sont 7 fois plus faibles que celles de $\mathcal{M}_{a/s}$ ce qui confirme une perte de variabilité due à la modélisation effectuée par HTS.

D'autre part, si l'on considère le corpus de test $T_{a/s}$, les données de ce copus sont naturellement les plus vraisemblables pour $\mathcal{M}_{a/s}$ et les moins vraisemblables pour $\mathcal{M}_{p1}$: la caractérisation d'un segment, par sa seule étiquette phonologique, est insuffisante pour produire un espace acoustique pour lequel les données de test, issues du processus d'analyse/synthèse, seraient vraisemblables.

Les résultats précédents prennent en compte les NSS (Non-Speech-Sounds) qui représentent, comme l'illustre la figure 5.5 du chapitre 5, environ 10% des corpus de test et d'apprentissage. Ces segments sont particuliers dans la mesure où ils correspondent en majorité à des silences. Afin d'évaluer la modélisation du spectre en ignorant ces segments particuliers, nous avons appliqué le même protocole d'évaluation en excluant les trames associées aux NSS. Les résultats sont illustrés figure 6.7.

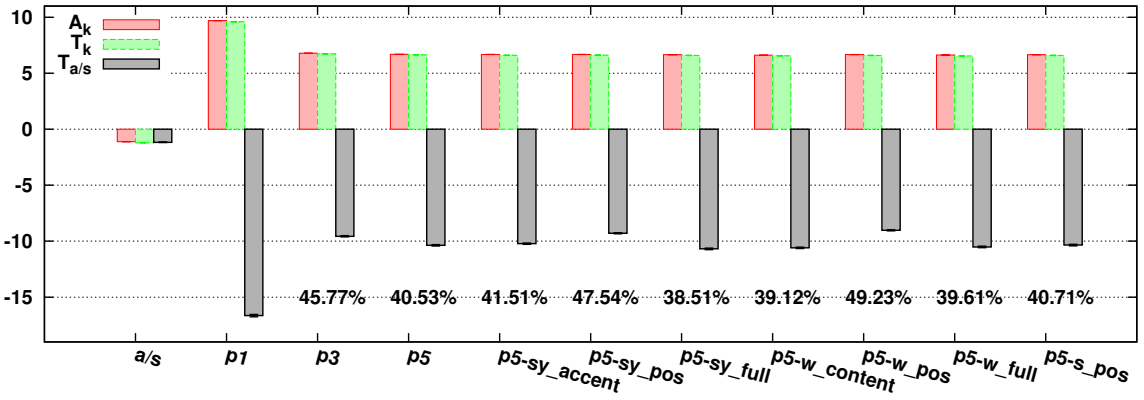


FIGURE 6.7 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace spectral sans tenir compte des NSS

En comparant le comportement des log-vraisemblances obtenues, les conclusions concernant le corpus $T_{a/s}$ dépourvu de NSS semblent identiques à celles émises lors de la prise

en compte des NSS. Les NSS forment une classe de bruits (majoritairement des pauses et des bruits d'aspiration) avec des enveloppes spectrales relativement homogènes entre elles. Les différences entre les jeux de descripteurs semblent porter principalement sur la modélisation des phones.

Enfin, par définition (cf. section 1.1.3 du chapitre 1), la première dimension d'un vecteur de coefficients MGC correspond à l'énergie. Une différence sur l'amplitude du signal de synthèse n'impliquant pas une mauvaise représentation de la structure du phone, nous avons appliqué l'évaluation par GMM en ignorant cette première dimension liée à l'énergie. Nous obtenons les résultats illustrés figure 6.8.

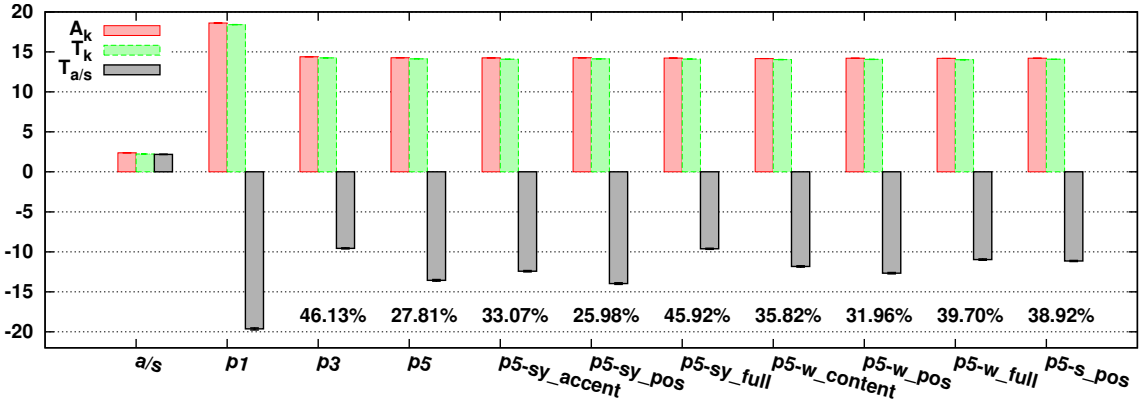


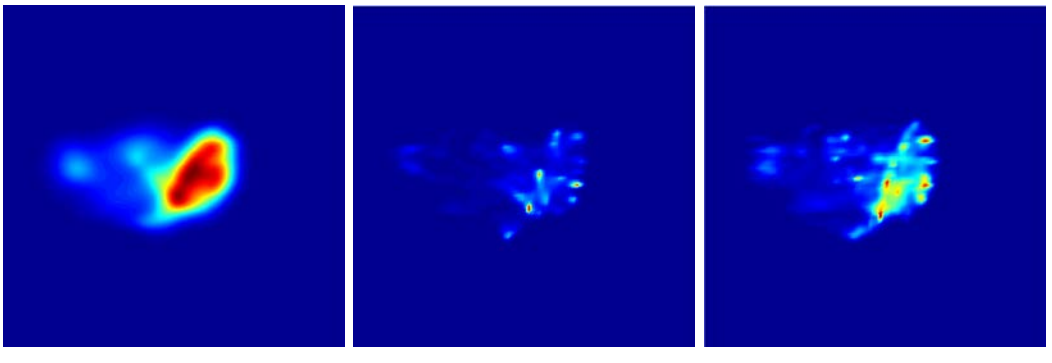
FIGURE 6.8 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de l'espace spectral sans tenir compte des NSS et en ignorant le coefficient d'énergie

Les résultats obtenus suivent la même tendance que lorsque l'énergie était prise en compte. Néanmoins, certaines dégradations sont accentuées. En effet, en comparant les log-vraisemblances $LL(T_{a/s}; \mathcal{M}_{p3})$ et $LL(T_{a/s}; \mathcal{M}_{p5})$, nous pouvons remarquer que le jeu de descripteurs `p5` aboutit à une modélisation des coefficients MGC moins pertinente que `p3`. Afin de visualiser et comparer plus aisément les différents GMM $\mathcal{M}_k$, une projection de ces derniers est effectuée relativement à un repère de $\mathbb{R}^2$ et illustrée figure 6.9.

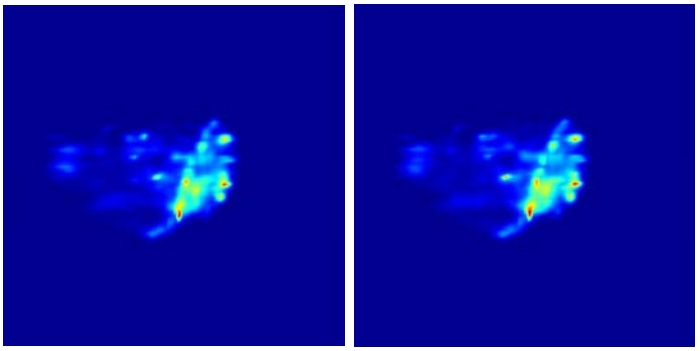
Ce repère de projection, commun à tous les GMM, est calculé de la façon suivante :

- On considère un mélange de gaussiennes composé de l'ensemble des GMM $\mathcal{M}_k$. La pondération associée à chaque $\mathcal{M}_k$ est proportionnelle à la taille du corpus d'apprentissage A_k. Étant donné que pour tout jeu de descripteurs k, A_k est de même taille que $A_{a/s}$, cette pondération des GMM est donc uniforme. Le GMM résultant est noté $\mathcal{M}$,
- la distribution gaussienne $\mathcal{N}(\mu, \Sigma)$ la plus proche de $\mathcal{M}$ au sens des moindres carrés est calculée,
- les deux premiers vecteurs propres orthonormés de Σ (correspondants aux plus grandes valeurs propres) forment le nouveau repère dans lequel sont projetés les GMM $\mathcal{M}_k$.

En visualisant les espaces acoustiques projetés, nous constatons une nette différence



(a) a/s (b) p1 (c) p3



(d) p5 (e) p5-s_pos

FIGURE 6.9 – Visualisation de l'espace des enveloppes spectrales MGC modélisé par un GMM et projeté en 2D. Les GMM ont été appris en ignorant les NSS et le coefficient d'énergie. Un repère unique global a été déterminé de façon a intégrer la *couverture* de l'ensemble des espaces décrits par les différents GMM. Chaque GMM a été projeté dans ce repère, seules les deux premières dimensions sont retenues.

entre l'espace associé à l'analyse synthèse et ceux obtenus par génération HTS. En effet, on observe une variance plus large des composantes de $\mathcal{M}_{a/s}$ avec un *recouvrement* important des distributions ; la zone de plan à *forte vraisemblance* pour $\mathcal{M}_{a/s}$ est étendue et homogène. A contrario, pour $\mathcal{M}_{a/s}$, les variances sont très faibles et les *pics* des distributions associées sont relativement isolés. L'espace acoustique associé au jeu de descripteurs `p1` est plus hétérogène que celui correspondant aux données issues de l'analyse/synthèse, bien que leurs périmètres semblent identiques. Enfin, en utilisant le jeu de descripteurs `p5`, nous nous situons dans un cas intermédiaire entre `p1` et `p3` ce qui peut expliquer la différence constatée entre $LL(T_{a/s}; \mathcal{M}_{p3})$ et $LL(T_{a/s}; \mathcal{M}_{p5})$. Néanmoins, en utilisant la méthodologie actuelle, il nous semble difficile d'analyser plus loin cette différence de modélisation.

6.4 Évaluation de la durée

Le dernier paramètre évalué est la durée des segments acoustiques (phones ou NSS). Comme cela a été indiqué dans l'étude préliminaire, il faut tenir compte du fait que la configuration du système HTS que nous avons utilisée ne peut générer que des durées

multiples de 5ms. Il est donc impossible pour HTS de modéliser exactement la durée naturelle. Dans le cadre de ces expériences, cette contrainte ne constitue pas un obstacle car nous ne souhaitons pas déterminer une mesure absolue de la qualité de modélisation effectuée par HTS mais simplement comparer différentes générations effectuées par HTS. La contrainte précédente s'appliquant pour l'ensemble des modèles appris à l'aide des différents jeux de descripteurs, les résultats obtenus à l'issue de notre protocole restent valides ; les résultats sont illustrés figure 6.10.

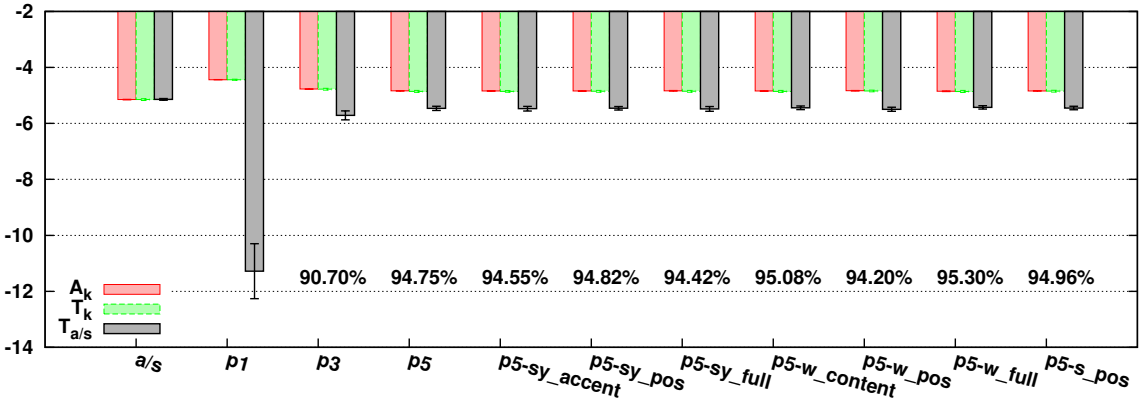


FIGURE 6.10 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de la durée.

L'ensemble des log-vraisemblances concernant les jeux de descripteurs autres que `p1` sont proches de $LL(T_{a/s}; \mathcal{M}_{a/s})$ qui constitue la borne haute. De plus, bien que quelques variations existent, les intervalles de confiance associés à chacune des log-vraisemblances $LL(T_{a/s}; \mathcal{M}_k)$, où $k \notin a/s, p1$ se chevauchent. Ces variations ne sont pas significatives. Enfin, en ignorant les NSS, nous obtenons les résultats présentés dans la figure 6.11.

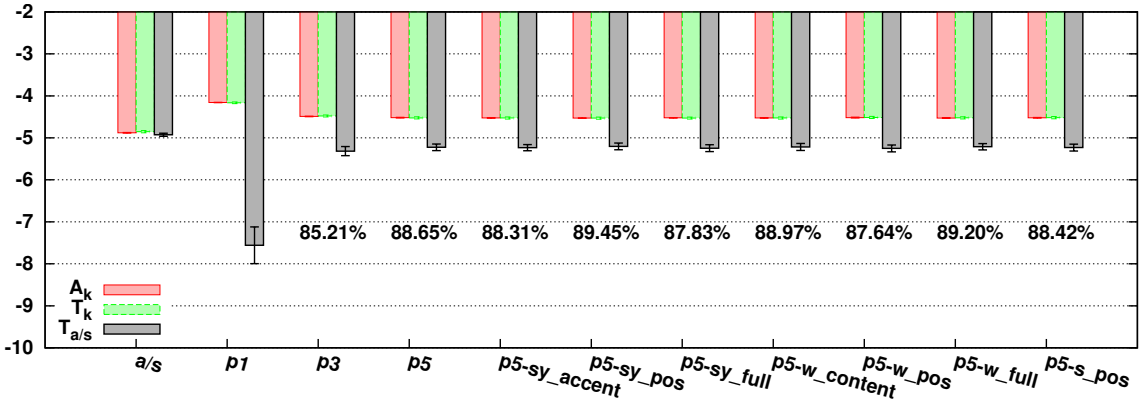


FIGURE 6.11 – Résultat du protocole d'évaluation objective basé sur la modélisation GMM de la durée en ignorant les NSS.

Les résultats obtenus confirment ceux présentés lors de la prise en compte des NSS. Néanmoins, le nombre de composantes est limité à 2 par GMM. La durée étant une donnée scalaire, il est possible de représenter la distribution des durées des phones du corpus $A_{a/s}$ dépourvue des NSS ainsi que les modélisations par HTS. Ceci est illustré sur la figure 6.12.

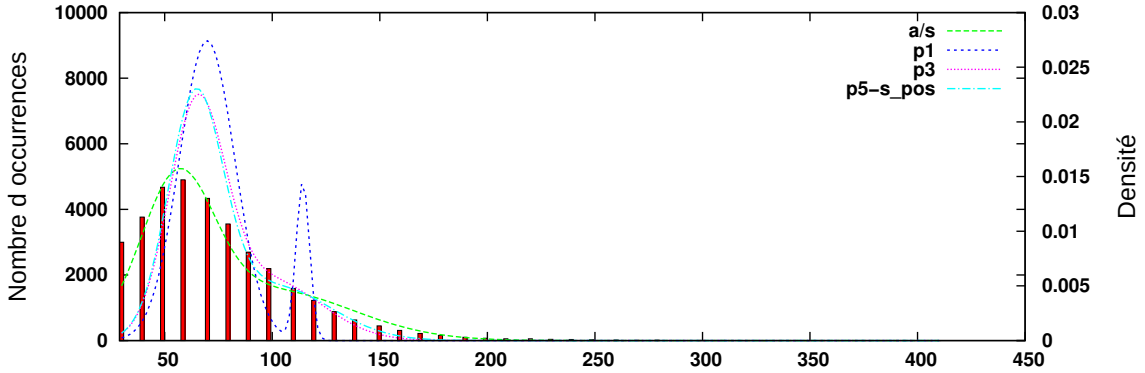


FIGURE 6.12 – Distribution de la durée des phones du corpus d'apprentissage $A_{a/s}$ dépourvu de NSS et représentation des GMM utilisés lors de l'évaluation pour un sous-ensemble des jeux de descripteurs.

Cette figure montre que les durées naturelles ne forment qu'un seul mode ce qui explique pourquoi le nombre de composantes des GMM est faible. Néanmoins, le GMM $\mathcal{M}_{p1}$ indique que la génération effectuée par HTS en utilisant le jeu de descripteurs `p1` aboutit à deux partitions bien distinctes : la première centrée sur 70 ms et la seconde sur 120 ms.

Ainsi, d'après l'ensemble des résultats obtenus à l'issue de l'évaluation de la durée, nous pouvons conclure que le jeu de descripteurs optimal pour modéliser ce paramètre acoustique est le jeu de descripteurs `p3`.

6.5 Évaluation de l'apériodicité

Le dernier paramètre est l'apériodicité qui, contrairement aux paramètres précédents, reste spécifique à STRAIGHT. Comme nous l'avons indiqué précédemment (section 2.6), l'apériodicité est représentée par un vecteur de 5 coefficients. Chacun de ces coefficients est associé à une bande de fréquences, chaque fréquence de cette bande étant émise proportionnellement au coefficient d'apériodicité. L'évaluation par GMM consiste donc à modéliser l'espace de l'apériodicité en se basant sur ce vecteur de 5 coefficients. Suite à l'application du protocole sur différents GMM, le nombre de composantes a été imposé à 256 (cf. section 6.2). Les résultats obtenus sont illustrés par la figure 6.13.

Comme précédemment, la log-vraisemblance associée à la modélisation GMM de l'espace associé au jeu de descripteurs `p1` est la plus faible, ce qui confirme d'ailleurs que ce jeu de descripteurs n'est pertinent pour aucun des paramètres évalués. En comparant les log-vraisemblances associées aux autres jeux de descripteurs, deux jeux de descripteurs semblent améliorer significativement la modélisation de l'apériodicité : `p5-sy_full` et `p5-w_content` qui ne sont pas significativement différents entre eux. De plus, comme pour la modélisation du spectre, l'utilisation du jeu de descripteurs `p5` semble dégrader la modélisation de l'apériodicité par rapport à l'utilisation de`p3` au regard de l'évaluation par

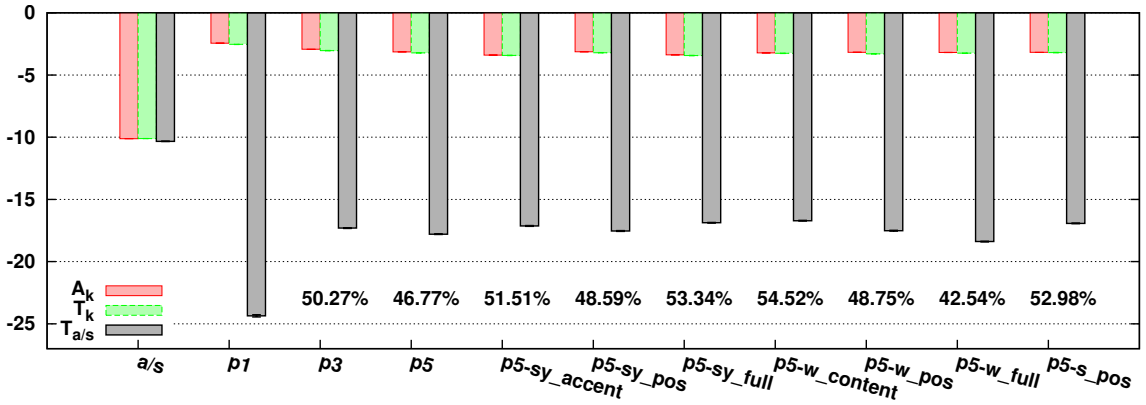


FIGURE 6.13 – Résultat du protocole d'évaluation objective basé sur la modélisation de l'espace de l'apériodicité en utilisant GMM

GMM. Afin de compléter l'analyse, nous avons appliqué le protocole en ignorant les NSS. Nous obtenons alors les résultats présentés dans la figure 6.14 où chaque GMM contient également 256 composantes.

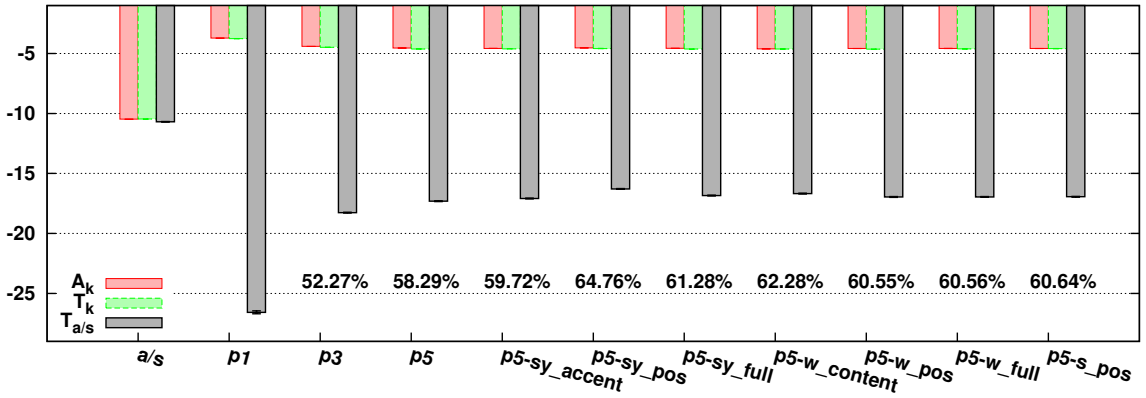
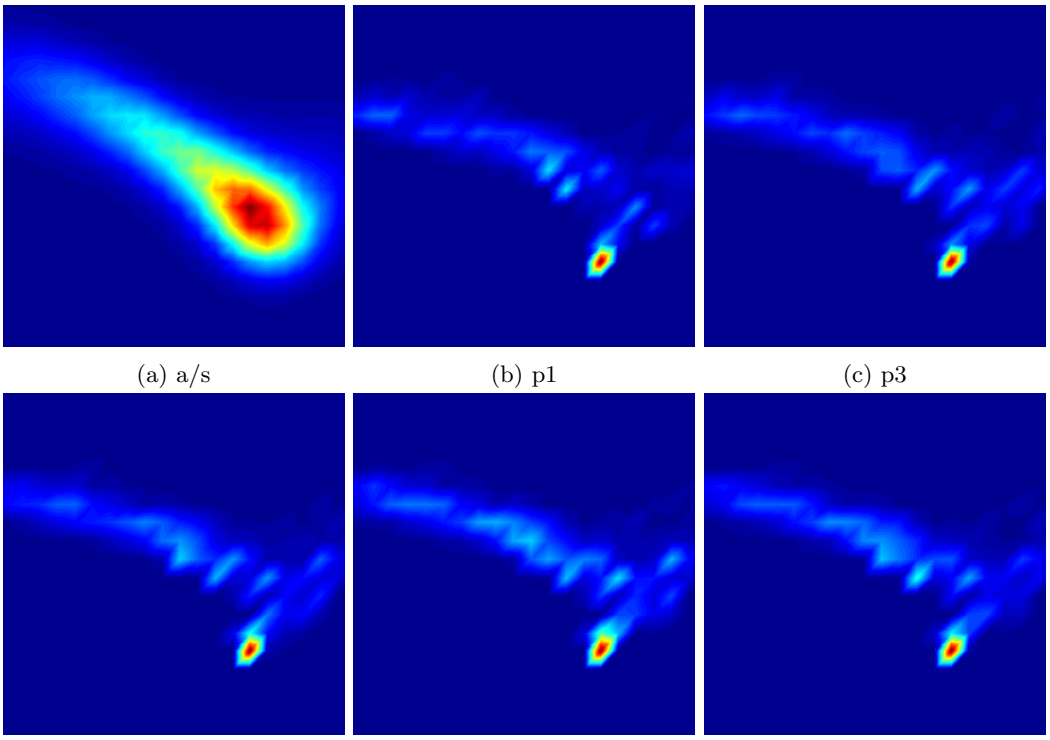


FIGURE 6.14 – Résultat du protocole d'évaluation objective basé sur la modélisation de l'espace de l'apériodicité en utilisant GMM. Les GMM sont appris sur le corpus A_k dépourvu de NSS et les NSS du corpus $T_{a/s}$ ont été ignorés lors du calcul de la log-vraisemblance

Les résultats obtenus diffèrent légèrement de ceux obtenus précédemment. En effet, le jeu de descripteurs `p5-sy_pos` obtient un ratio d'amélioration plus élevé et peut être considéré comme le jeu de descripteurs optimal en ne tenant plus compte de la modélisation des NSS. Afin de pouvoir analyser la différence constatée entre les jeux de descripteurs `p3`, `p5`, `p5-sy_pos` et `p5-sy_full`, nous avons représenté les espaces de coefficients d'apériodicité de manière analogue à la représentation des espaces associés aux coefficients MGC précédemment. Ces espaces sont illustrés dans la figure 6.15.

Cette figure confirme une différence, entre les espaces des coefficients d'apériodicité générés par HTS et celui issu de l'analyse synthèse, dont la nature est équivalente à celle constatée pour les coefficients MGC : les domaines des coefficients coïncident mais la génération HTS aboutit à des gaussiennes de variance plus faible. Néanmoins, cette figure



(a) a/s (b) p1 (c) p3

(d) p5 (e) p5-sy_pos (f) p5-sy_full

FIGURE 6.15 – Visualisation de l'espace de l'apériodicité modélisé par un GMM. Les GMM ont été appris en ignorant les NSS et le coefficient d'énergie. Un repère global a été déterminé de façon a maximiser la *couverture* de l'ensemble des espaces décrits par les GMM. Chaque GMM a été projeté dans ce repère et seules les deux premières dimensions sont retenues.

montre également qu'il y a très peu de différences entre les espaces générés par HTS : du jeu `p1` au jeu `p5-sy_pos`, nous pouvons constater une homogénéisation progressive entres les composantes du GMM. Toutefois entre `p5-sy_pos` et `p5-sy_full`, les différences restent marginales.

En conclusion, il semble difficile de conclure sur l'amélioration apportée par `p5-sy_pos` par rapport à `p5` sur la modélisation de l'apériodicité. Néanmoins, dans l'ensemble des analyses effectuées, nous constatons qu'utiliser des descripteurs d'horizons plus élevés que la syllabe ne permet pas réellement d'améliorer la modélisation.

6.6 Bilan et conclusion

À l'issue de ces expériences, nous avons mis en avant la différence de qualité de modélisation entre le jeu de descripteurs `p1` et les autres jeux de descripteurs pour l'ensemble des paramètres générés par HTS (MGC, F0, durée, apériodicité). En ce qui concerne les coefficients MGC et la durée, le protocole indique qu'utiliser uniquement le contexte phonétique direct (jeu de descripteurs `p3`) suffit. En revanche, dans le cadre de ce proto-

cole, le F0 et l'apériodicité nécessitent davantage de descripteurs notamment au niveau de la syllabe. Le jeu de descripteurs `p5-sy_full` est alors considéré comme le plus pertinent pour ces deux paramètres.

Selon notre évaluation basée sur une modélisation des espaces de paramètres acoustiques par des GMM, l'utilisation de descripteurs au delà de la syllabe ne permet pas d'améliorer significativement la modélisation effectuée par HTS. Néanmoins, cette évaluation ne permet pas de déterminer l'influence des descripteurs sur les modèles appris par HTS. En effet, il est nécessaire de disposer d'un ensemble de données important pour apprendre un GMM pertinent. Analyser précisément les modèles réduit fortement le nombre de vecteurs de coefficients disponibles pour cet apprentissage ce qui exclu l'emploi de ce protocole pour des analyses plus locales.

Le chapitre suivant présente les résultats obtenus par un protocole moins global en effectuant des mesures de distance entre vecteurs acoustiques appariées (protocole présenté au chapitre 4)

Chapitre 7

Évaluation objective - Évaluation non paramétrique

Au cours du chapitre précédent, nous avons pu comparer l'influence des descripteurs sur l'espace modélisé par HTS pour chacun des paramètres acoustiques. Nous avons constaté que le jeu de descripteurs `p1` se démarquait dans un sens défavorable des autres jeux de descripteurs pour l'ensemble des coefficients analysés. De plus, à l'issue du chapitre précédent, le jeu de descripteurs considéré comme optimal est le jeu `p5-sy_full`. Pour le F0 et l'apériodicité, il est nécessaire d'inclure les informations à l'horizon de la syllabe pour obtenir les vraisemblances les plus fortes.

Dans ce chapitre, nous allons présenter les résultats obtenus en appliquant le protocole d'évaluation non-paramétrique. Ce protocole a été présenté dans la section 4.2 du chapitre 4. L'utilisation de ce protocole va permettre, en premier lieu, d'éprouver les résultats obtenus précédemment. En effet, par opposition à l'évaluation par GMM, qui est nécessairement globale, l'évaluation par distance repose sur un appariement entre trames et permet d'effectuer une évaluation plus locale. De plus, l'avantage de ce protocole par rapport à l'évaluation par GMM est qu'il n'est pas dépendant d'une étape d'apprentissage donc insensible à une quantité de données statistiquement significative pour obtenir un résultat fiable.

Afin de présenter les résultats obtenus, nous suivons le même plan que le chapitre précédent : les résultats associés au F0, aux coefficients MGC puis à la durée seront tout d'abord analysés. À l'issue de l'analyse des coefficients d'apériodicité, spécifiques à STRAIGHT, un bilan de l'évaluation basée sur un calcul de distance sera effectué.

7.1 Évaluation de la modélisation du F0

Le premier type de coefficients que nous allons évaluer est le F0. Comme précédemment, nous ne considérons que des couples de trames voisées. Néanmoins, contrairement à l'évaluation par GMM, nous pouvons déterminer les erreurs de voisement (prédiction d'une trame voisée alors que la trame devrait être non voisée et réciproquement).

7.1.1 Résultats globaux

Afin d'évaluer globalement la modélisation du F0, nous avons tout d'abord appliquer les fonctions permettant de déterminer la distance entre les valeurs de F0 générées et celles associées au signal naturel (voir 4.2.5 du chapitre 4 page 75) sur l'ensemble des trames. Nous mettons en œuvre trois types de mesure permettant de quantifier l'erreur globale entre la génération effectuée par HTS et les coefficients acoustiques extraits du signal naturel : l'erreur RMS en Hertz, l'erreur RMS en cent utilisée par S. Yokomizo ET AL. [Yokomizo2010] dans leur étude présentée section 3.4.2 du chapitre 3 ; l'écart moyen relatif en Hertz qui permet de déterminer si la synthèse est, en moyenne, plus aiguë ou plus grave que le signal original.

Tout d'abord, nous allons nous focaliser sur les résultats du calcul de la RMS en Hertz. Ces résultats sont présentés figure 7.1 et désignent, comme jeu de descripteurs optimal, les jeux `p5-sy_accent` et `p5-sy_full` ce qui coïncide avec les résultats obtenus lors de l'évaluation par GMM. Ce sont les deux seuls jeux de descripteurs dont la RMS est significativement plus faible que la RMS associée à `p5`. Néanmoins, une différence importante apparaît si l'on compare les résultats associés au jeu `p1` qui, contrairement à l'évaluation GMM, n'est pas considéré comme plus mauvais que les autres jeux de

descripteurs. En passant à une échelle logarithmique, nous obtenons les résultats présentés figure 7.2.

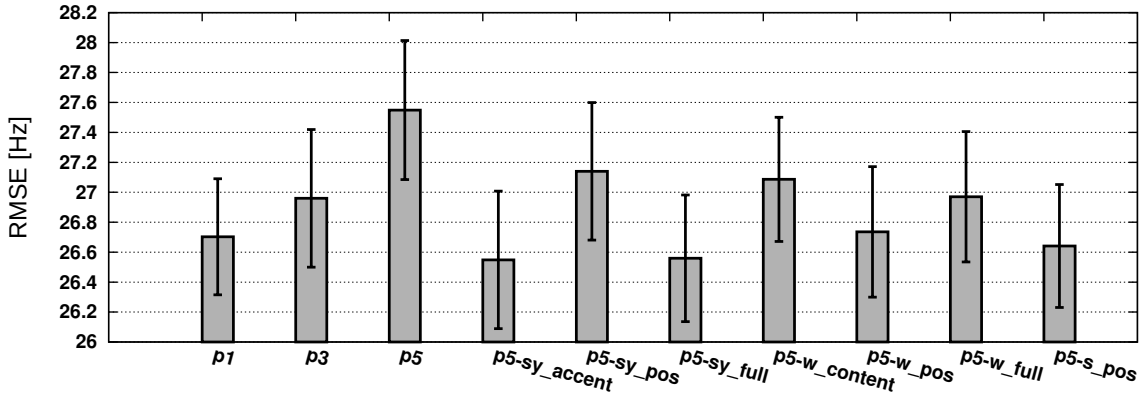


FIGURE 7.1 – RMS en Hertz entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour l'ensemble des trames du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

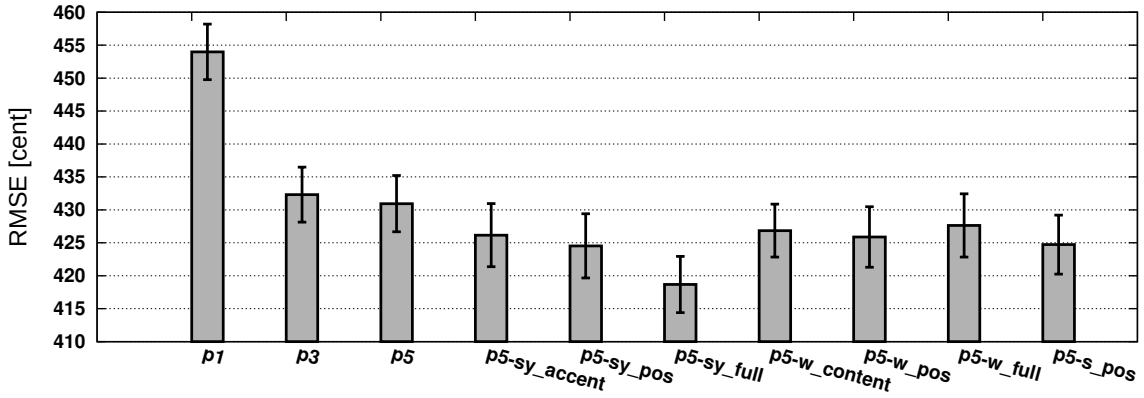


FIGURE 7.2 – RMS en cent entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour l'ensemble des trames du corpus de test. Les intervalles de confiance correspondent au niveau de confiance de 95%.

Tout d'abord, comme il s'agit de la mesure employée par S. Yokomizo ET AL.[Yokomizo2010], nous cherchons à comparer nos résultats à cette référence. S. Yokomizo ET AL. indiquent que l'erreur RMS obtenue pour leurs expériences varie entre 350 cents (pour l'équivalent de `p5` sur le corpus japonais) à environ 227 cents (pour différents cas sur le corpus anglais). Notre corpus obtient des valeurs de RMS comprises entre 410 et 460 cents ce qui est supérieur. Néanmoins cette différence peut provenir de la différence entre les corpus utilisés. Les corpus utilisés dans [Yokomizo2010] sont le corpus CMU ARCTIC [Kominek2003] pour l'anglais et le corpus ATR [Kurematsu1990] pour le japonais. Ces deux corpus diffèrent du corpus CORDIAL par la langue mais également par le locuteur. En effet, le corpus CMU ARCTIC est composé de six locuteurs non professionnels (quatre hommes et deux femmes) et le corpus ATR est composé de dix locuteurs professionnels (six hommes et quatre femmes) alors que le corpus CORDIAL ne contient qu'un seul locuteur. Les scores présentés dans [Yokomizo2010], pour chaque corpus, correspondent

aux moyennes des scores obtenus pour l'ensemble des locuteurs. De plus, contrairement à Cordial, ces deux corpus ne sont pas considérés comme des corpus de parole expressive.

En comparant les valeurs de RMS obtenues pour chaque jeu de descripteurs, nous remarquons une structure proche des résultats obtenus lors de l'évaluation par GMM : le jeu de descripteurs `p1` est considéré comme le plus mauvais, l'introduction des informations à l'horizon de la syllabe, et plus spécifiquement des descripteurs présents dans `p5-sy_full`, permet de franchir un nouveau palier concernant la qualité de modélisation du F0 dans HTS. Néanmoins, contrairement aux résultats de l'évaluation par GMM, l'introduction de descripteurs avec un horizon supérieur à la syllabe aboutit à des distances plus élevées ce qui indique une dégradation de la modélisation à une échelle plus locale.

Enfin, la répartition des erreurs RMS en fonction du jeu de descripteurs a évolué en passant à l'échelle cent. La différence entre les deux échelles est l'application d'un logarithme. Ainsi, nous pouvons émettre l'hypothèse que le jeu de descripteurs `p1` aboutit à une modélisation moins pertinente pour les basses et moyennes fréquences. En revanche, sur l'échelle des Hertz, la RMS associée au jeu de descripteurs `p5-sy_full` est plus faible que celle associée au jeu `p5`. Cela implique donc que l'utilisation du jeu `p5-sy_full` permet d'obtenir une meilleure modélisation du F0, uniformément selon les fréquences que le jeu `p5`. À noter que dans le domaine hertzien, le jeu de descripteurs `p5-sy_accent` obtient des résultats équivalents au jeu de descripteurs `p5-sy_full`. Ainsi, comme pour le jeu `p1`, nous pouvons conclure que les différences de modélisation entre l'utilisation du jeu `p5-sy_accent` et `p5-sy_full` se situent dans les basses et moyennes fréquences. En ignorant les NSS, ces résultats n'évoluent pas comme l'illustre la figure 7.3.

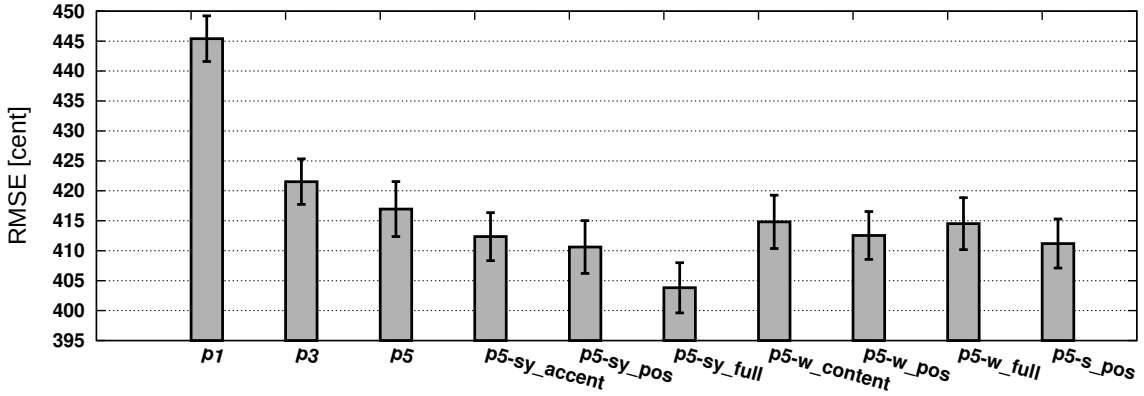


Figure 7.3 – RMS en Hertz entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées comme faisant partie d'un NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

En calculant la moyenne des écarts relatifs, entre les coefficients générés par HTS et ceux extraits du signal naturel, nous obtenons les résultats présentés dans la figure7.4. En comparant ces écarts, nous constatons que les variations sont faibles voires, pour la majorité des jeux de descripteurs, non-significatives. Toutefois, l'écart relatif moyen est positif ce qui indique que la synthèse produite par HTS est plus aiguë.

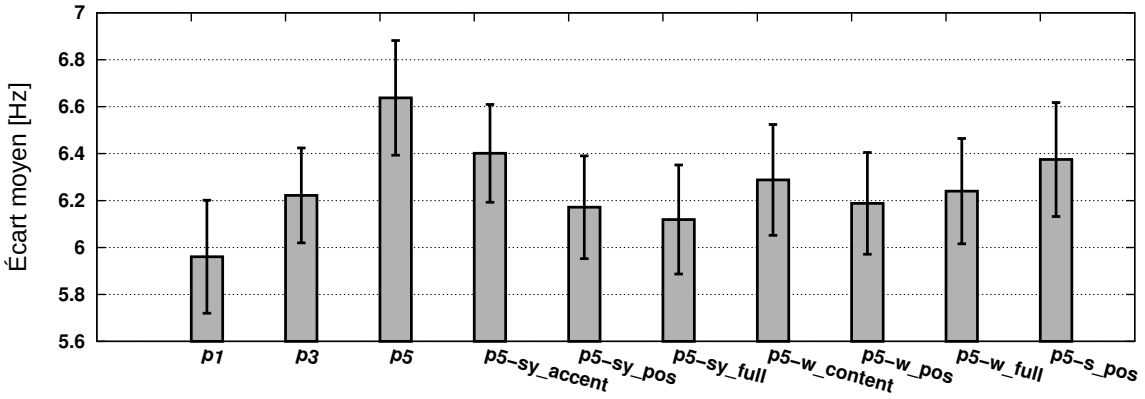


FIGURE 7.4 – Moyenne des écarts relatifs entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées comme faisant partie d'un NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

De plus, ces écarts se situent aux alentours du 6Hz ce qui est supérieur au seuil différentiel (JND, Just Noticeable Difference) qui est d'environ 3 ou 4Hz pour des fréquences inférieures à 500 Hz. En complétant ces résultats par la moyenne des écarts absolus, présentée dans la figure 7.5, nous pouvons émettre l'hypothèse que ces écarts devraient être perceptibles.

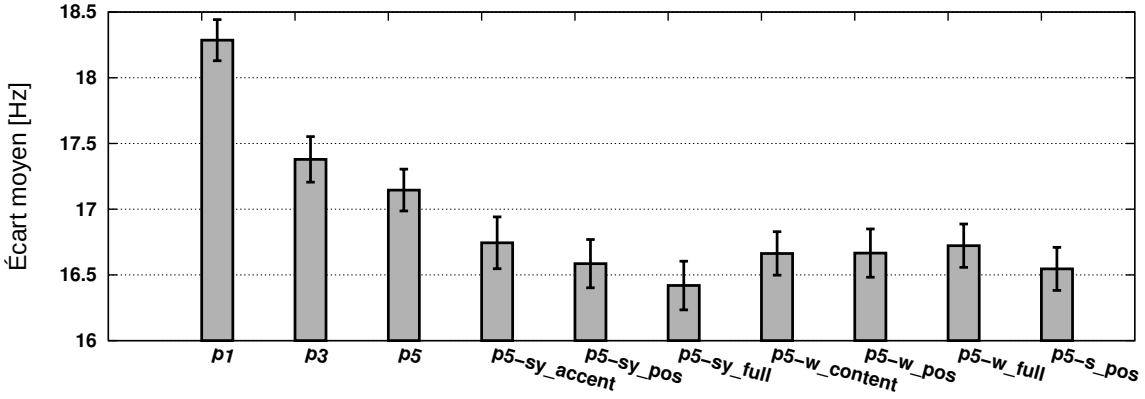


FIGURE 7.5 – Moyenne des écarts absolus entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées comme faisant partie d'un NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

Enfin, un indicatif important de la qualité de modélisation du F0 est le taux d'erreurs de voisement. Ce taux est obtenu en calculant la moyenne des trames qui présentent une erreur de voisement telle qu'elle a été définie lors de la présentation de la première étape (section 4.2.1 du chapitre 4). La figure 7.6 illustre les résultats obtenus.

D'après ces résultats, l'utilisation du jeu `p1` aboutit à un taux d'erreurs de voisement d'environ 15% ce qui correspond à 2% d'erreurs de voisement supplémentaires par rapport aux autres jeux de descripteurs. Les jeux de descripteurs `p3` et `p5-sy_pos` permettent d'obtenir moins d'erreurs de voisement que la majorité des autres jeux de descripteurs.

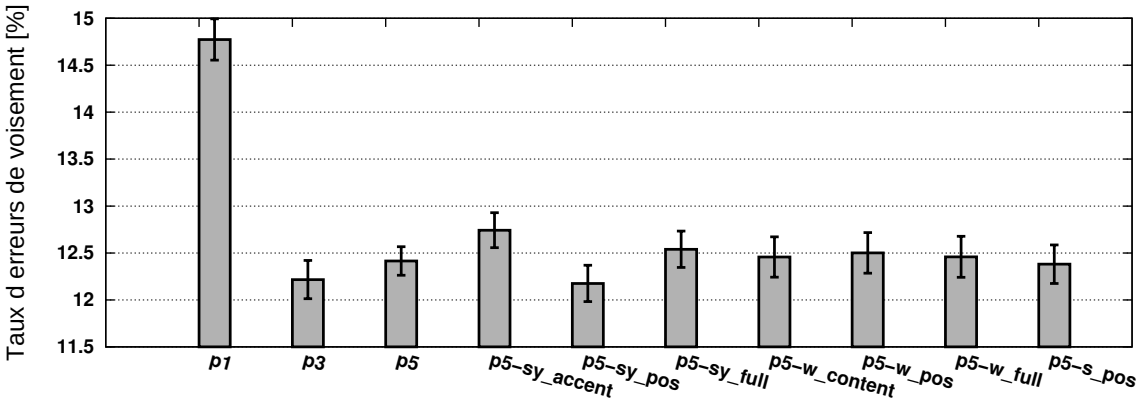


FIGURE 7.6 – Taux d'erreurs de voisement entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour l'ensemble des trames du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

Ces constats se confirment en ignorant les NSS comme le montre la figure 7.7

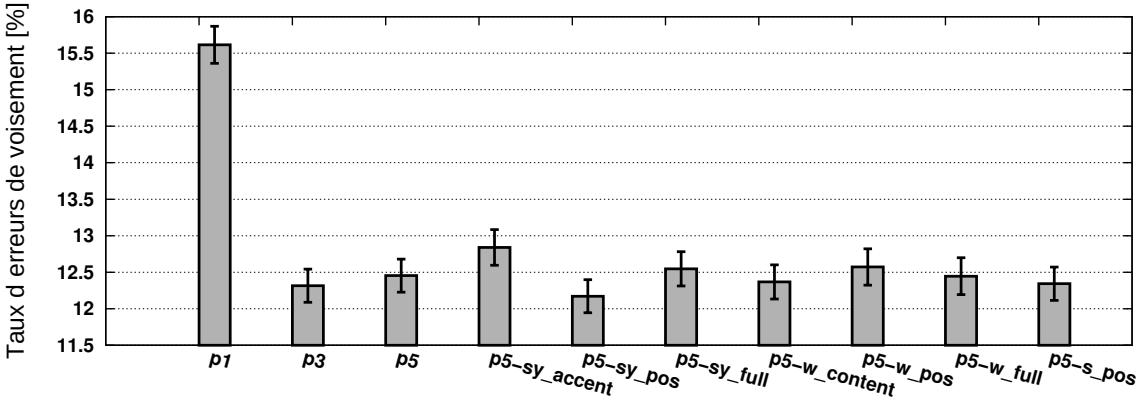


FIGURE 7.7 – Taux d'erreurs de voisement entre les valeurs de F0 générées par HTS selon un jeu de descripteurs (précisé en abscisse) et celles extraites du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées comme faisant parti d'un NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

Ainsi, le jeu de descripteurs `p1` se distingue nettement, et cela pour la majorité des mesures employées, des autres jeux de descripteurs pour modéliser le F0. En revanche, selon la métrique utilisée, des disparités existent entre les autres jeux de descripteurs comme, par exemple, le fait que l'écart en cent permette de discriminer le jeu `p5-sy_full` du jeu de descripteurs `p3`. En effet, cette distinction n'est pas présente si l'écart est déterminé sur l'échelle des Hertz. Toutefois, nous obtenons une conclusion identique à celle obtenue lors de l'évaluation par GMM : le jeu de descripteurs `p5-sy_full` correspond, globalement, au jeu optimal pour modéliser le F0.

7.1.2 Résultats par catégorie de voisement

Afin d'affiner notre analyse, nous avons appliquer le protocole d'évaluation à l'horizon du phone, comme le permet la seconde étape du protocole (décrite section 4.2.2 du chapitre 4). De plus, lors de l'étape de partitionnement (décrite section 4.2.3 du même chapitre), nous avons regroupé les écarts représentatifs obtenus à l'issue des étapes précédentes par catégorie de voisement. L'influence des NSS ayant déjà été analysée, nous allons ignorer ces segments jusqu'à la fin de cette section. Nous avons finalement retenu trois catégories de voisement : les consonnes voisées, les consonnes non-voisées et les voyelles. La figure 7.8 illustrent les résultats obtenus en calculant une RMS en cent pour chaque catégorie de voisement.

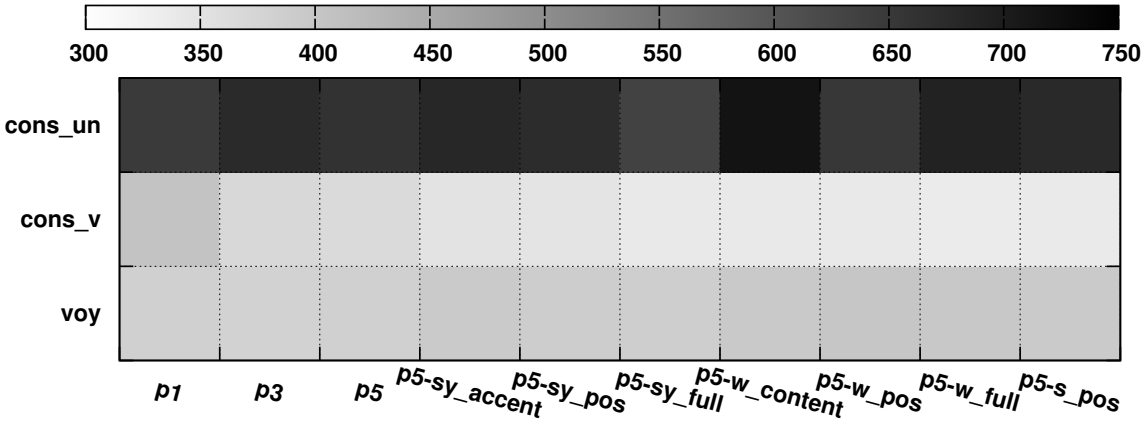


FIGURE 7.8 – RMS en cent, entre le F0 généré par HTS et celui issu de l'analyse effectuée par STRAIGHT, en fonction du jeu de descripteur utilisé et la catégorie de voisement. L'écart représentatif correspond à l'écart moyen pour chaque segment phonétique. Le jeu de descripteurs est indiqué en abscisse et la catégorie en ordonnée.

Ces résultats indiquent, tout d'abord, une nette différence entre les consonnes non-voisées et les autres catégories. Ces résultats peuvent s'expliquer par la nature même des consonnes non-voisées qui implique que des valeurs non nulles de F0 peuvent se situer en frontière de phone. Bien que d'après la figure 7.8, les jeux de descripteurs `p5-sy_full` et `p5-w_pos` améliorent la modélisation des consonnes non-voisées, les intervalles de confiance (non représentés ici) se chevauchent, indiquant que ces améliorations ne sont pas statistiquement significatives. les résultats présentés dans la figure 7.9 propose un partitionnement plus fin intégrant le mode d'articulation.

Ces résultats confirment la mauvaise modélisation d'une catégorie : les plosives non-voisées. L'amélioration apportée par `p5-sy_full` par rapport aux autres jeux de descripteurs concerne également cette catégorie. Une autre catégorie se distingue : les fricatives non voisées. En prenant en compte les intervalles de confiance, comme l'illustre la figure 7.10, nous constatons que ces différences sont statistiquement significatives [1]. Pour

1. La diphtongue est particulière car il y a peu d'exemples de ce phone dans le corpus de test et la variance des RMS est plus élevée que pour les autres phones. Ceci explique la taille de l'intervalle de confiance associé à cette catégorie.

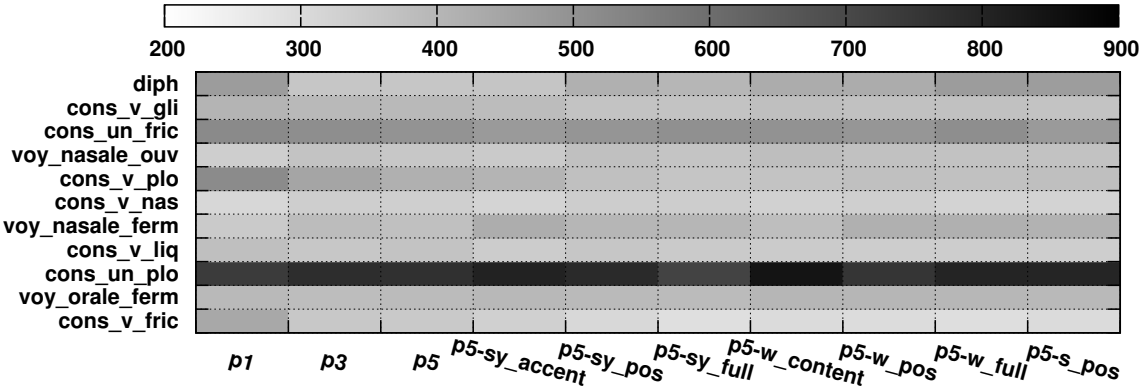


FIGURE 7.9 – RMS en cent entre le F0 généré par HTS et celui issu de l'analyse effectuée par STRAIGHT, en fonction du jeu de descripteurs et de la catégorie phonétique déterminée à partir du mode d'articulation. L'écart représentatif correspond à l'écart moyen pour chaque segment phonétique. Le jeu de descripteurs est indiqué en abscisse et la catégorie en ordonnée.

compléter cette analyse, nous avons déterminé le taux d'erreurs de voisement en fonction du mode d'articulation. Ces résultats sont présentés dans la figure 7.11.

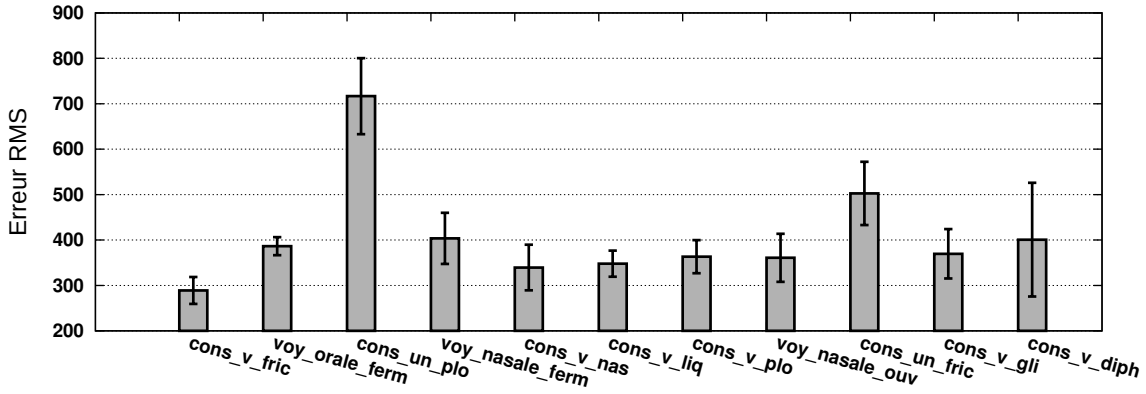


FIGURE 7.10 – RMS en cent, pour chaque catégorie phonétique, entre le F0 généré par HTS en utilisant le jeu de descripteurs p5-sy_full et celui extrait du signal naturel.

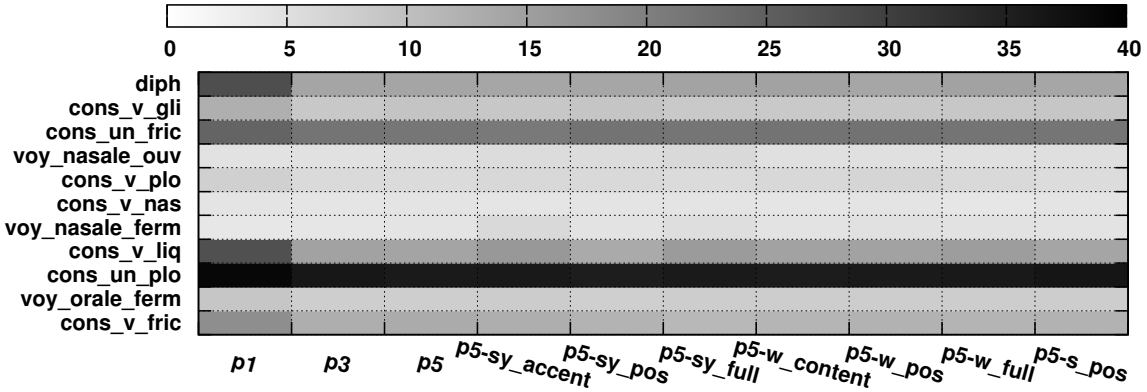


FIGURE 7.11 – Taux d'erreurs de voisement entre le F0 généré par HTS et celui issu de l'analyse effectuée par STRAIGHT, en fonction du jeu de descripteurs et de la catégorie phonétique déterminée à partir du mode d'articulation. L'écart représentatif correspond à l'écart moyen pour chaque segment phonétique. Le jeu de descripteurs est indiqué en abscisse et la catégorie en ordonnée.

Les résultats obtenus suivent la même tendance que ceux présentés figure 7.9 : les plosives non-voisées aboutissent à un taux d'erreurs de voisement beaucoup plus élevé (environ 35%) et la génération du F0 basée sur le jeu de descripteurs `p1` augmente le nombre d'erreurs de voisement sur davantage de catégories.

En conclusion, nous constatons donc que certaines catégories phonétiques sont mieux modélisées que d'autres. De plus, aucun des descripteurs introduits dans les différents jeux proposés ne permet d'homogénéiser la qualité de la modélisation en fonction des catégories phonétiques. Toutefois, comme nous l'avons vu lors de la présentation des MSD (section 2.3.1 du chapitre 2), la sensibilité des MSD aux frontières de voisement est un phénomène qui a déjà été identifié. Ainsi, pour expliquer les résultats obtenus pour les phones non-voisés, trois hypothèses peuvent s'appliquer dans le contexte de cette évaluation. Tout d'abord, la segmentation ayant été effectuée de manière automatique, les erreurs de positionnement de frontières influent sur la modélisation effectuée par HTS. Toutefois, nous ne savons pas à quel point cette influence dégrade la modélisation effectuée par HTS. La seconde hypothèse est la suivante : les résultats obtenus ne sont dus qu'à l'utilisation des MSD qui pénalise la modélisation de ces catégories phonétiques. Enfin, la troisième hypothèse consiste à supposer qu'aucun des descripteurs utilisés ne permet de capter les spécificités de ces phones. Dans l'état actuel de nos travaux, nous ne pouvons malheureusement conclure sur une des hypothèses émises car nous ne disposons pas de corpus aligné manuellement et nous n'utilisons que des modèles MSD.

7.1.3 Bilan de l'évaluation pour le F0

Lors de l'évaluation du F0, plusieurs résultats ont été mis en avant. Tout d'abord, le jeu de descripteurs `p1` aboutit à une modélisation du F0 moins pertinente dans la majorité des cas analysés. En revanche, il n'existe que très peu de variations entre les autres jeux

de descripteurs, hormis le `p5-sy_full` qui obtient une RMS globale plus faible.

Un autre résultat important de cette évaluation est la disjonction entre les catégories de voisement. En effet, d'après nos analyses, la modélisation du F0 pour les segments étiquetés comme non-voisés peut être considérée comme moins pertinente que la modélisation pour les segments étiquetés comme voisés. Ainsi, les consonnes non voisées et plus spécifiquement les plosives non-voisées, ont un taux d'erreurs de voisement ainsi qu'une valeur de RMS plus élevés. De plus, l'amélioration apportée par `p5-sy_full` par rapport aux autres jeux de descripteurs se situe principalement sur la modélisation de cette catégorie de phones. Néanmoins, nous pouvons conclure de ces résultats qu'aucun jeu de descripteurs ne permet réellement de capturer les spécificités du F0 pour les modèles étiquetés comme non-voisés et ainsi obtenir une modélisation de qualité équivalente aux modèles étiquetés voisés.

7.2 Évaluation de la modélisation spectrale

Après avoir évalué l'influence des descripteurs sur la modélisation du F0, nous allons maintenant analyser l'influence des descripteurs sur la modélisation des coefficients MGC. Dans le chapitre précédent (section 6.3.2), seul le jeu *p1* se distinguait de manière défavorable. La modélisation spectrale associée à ce jeu de descripteurs était considérée comme la moins pertinente. Toutefois, nous avons pu constater que prendre en compte des descripteurs d'horizon supérieur au contexte phonétique direct aboutissait parfois à des log-vraisemblances plus faibles. Dans cette partie, nous allons non seulement voir si le protocole basé sur les distances aboutit à la même conclusion mais également vérifier si cela se confirme pour l'ensemble des catégories phonétiques.

7.2.1 Résultats globaux

Tout d'abord le protocole a été appliqué à l'échelle de la trame acoustique en utilisant une mesure RMS. Les résultats obtenus sont présentés figure 7.12.

Ces résultats sont cohérents avec ceux obtenus lors de l'application du protocole d'évaluation par GMM. En effet, seul le jeu de descripteurs `p1` se distingue en ayant des distorsions plus élevées que les autres jeux de descripteurs. De même que pour l'évaluation basée sur la modélisation de l'espace spectrale par des GMM, nous avons appliqué cette méthode en ignorant les NSS. Les résultats obtenus sont présentés figure 7.13. Nous avons calculé la distance mel-cepstrale en ignorant la première dimension liée à l'énergie. Nous obtenons les résultats illustrés figure 7.14 [2].

Comme précédemment le jeu de descripteurs `p1` se distingue en étant plus mauvais

2. L'unité est différente de la RMS car, comme l'indique l'équation (6.4), un facteur permettant de passer en décibel a été appliqué.

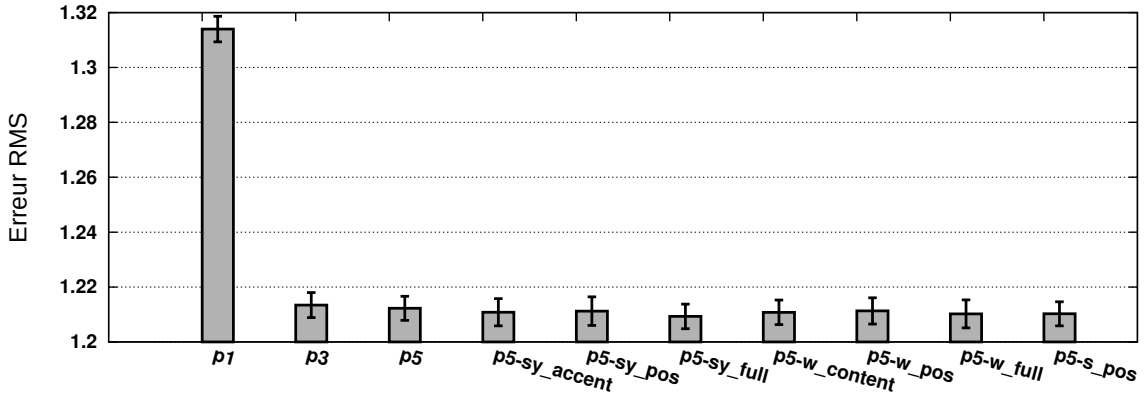


FIGURE 7.12 – RMS entre les coefficients MGC générés par HTS selon un jeu de descripteurs (précisé en abscisse) et les coefficients MGC extraits du signal naturel pour l'ensemble des trames du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

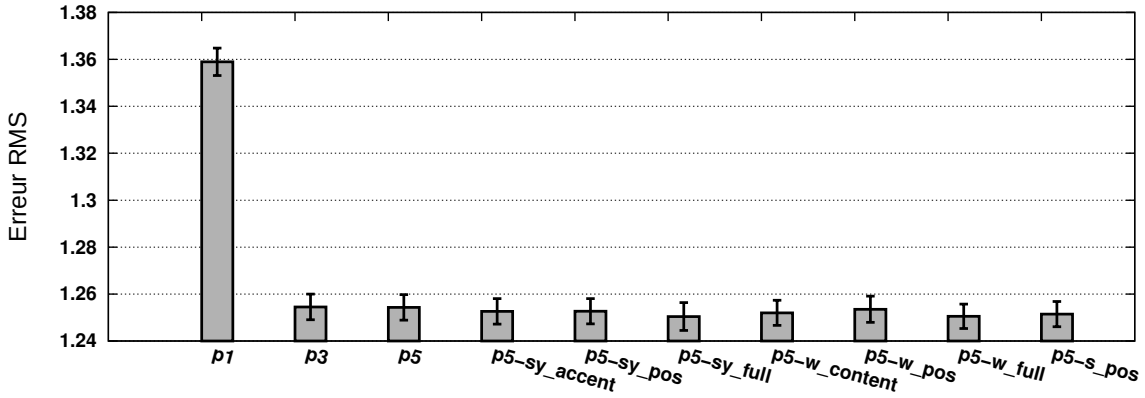


FIGURE 7.13 – RMS entre les coefficients MGC générés par HTS selon un jeu de descripteurs (précisé en abscisse) et les coefficients MGC extraits du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

et les autres jeux de descripteurs aboutissent à des résultats équivalents les uns par rapport aux autres. De plus, comme nous l'avons indiqué au chapitre 3, S. Yokomizo ET AL.[Yokomizo2010] ont utilisé la distance mel-cepstrale pour effectuer leur étude. Nous pouvons ainsi comparer les résultats que nous avons obtenus avec leurs résultats. Tout d'abord, pour l'évaluation effectuée en utilisant le corpus ARCTIC (pour l'anglais), les auteurs indiquent une distance cepstrale d'environ 6.94dB. En ce qui concerne l'évaluation effectuée sur le corpus ATR (pour le japonais), les distances mel-cepstrales obtenues se situent aux alentours de 4.8dB. Dans le cadre de notre étude, nous obtenons des distances mel-cepstrales comprises entre 5.2dB et 5.7dB. Nos résultats semblent cohérents avec ceux présentés dans [Yokomizo2010].

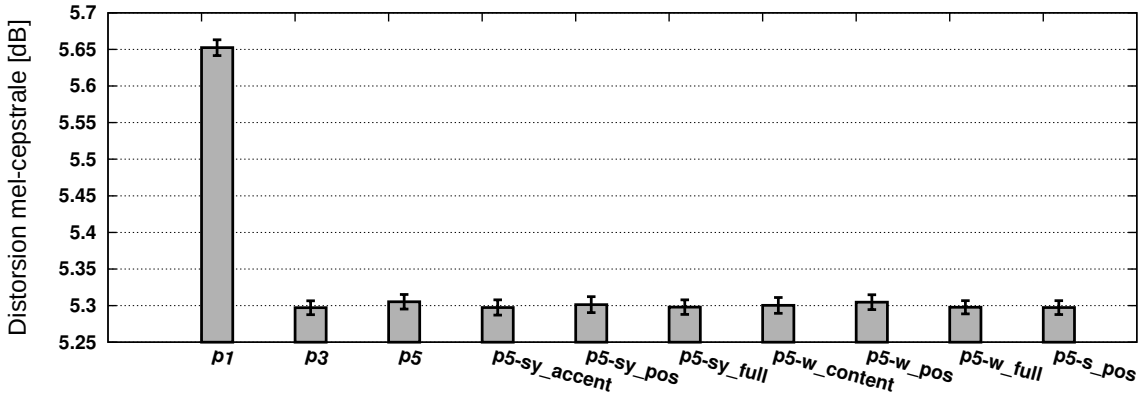


FIGURE 7.14 – Distance mel-cepstrale entre les coefficients MGC par HTS selon un jeu de descripteurs (précisé en abscisse) et les coefficients MGC extraits du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%. Contrairement à la RMS, la distance mel-cepstrale ne tient pas compte de la composante énergie.

7.2.2 Résultats par catégorie de voisement

Afin d'affiner l'analyse effectuée sur la modélisation du spectre, nous pouvons calculer la distance cepstrale moyenne associée à chaque catégorie de voisement. La figure 7.15 illustre les résultats obtenus.

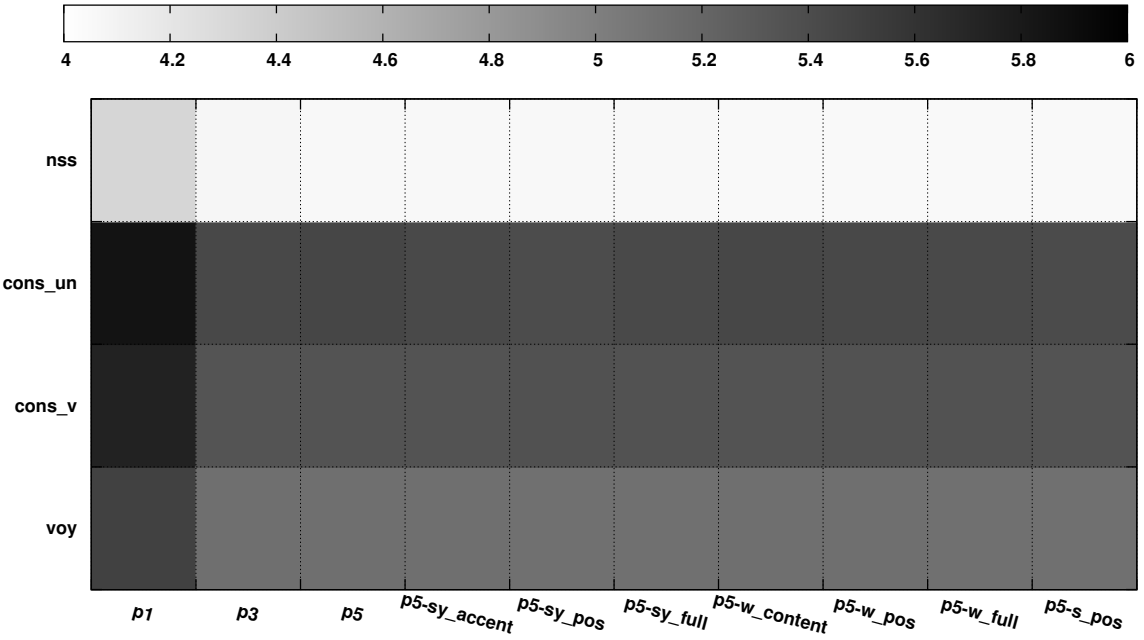


FIGURE 7.15 – Distance cepstrale moyenne déterminée entre les coefficients MGC générés par HTS et ceux extraits du signal naturel en fonction du jeu de descripteurs et de la catégorie de voisement. Le jeu de descripteurs est indiqué en abscisse et la catégorie de voisement en ordonnée.

Les résultats illustrés montrent que l'amélioration de modélisation apportée par le jeu `p3` par rapport à `p1` concerne l'ensemble des catégories y compris les NSS. De plus, cette

catégorie obtient la distorsion cepstrale la plus faible ce qui est cohérent avec la nature de ces segments. Par la suite, nous avons ignoré les NSS afin d'évaluer plus précisément les divergences constatées entre les autres catégories de voisement. Les résultats sont présentés figure 7.16.

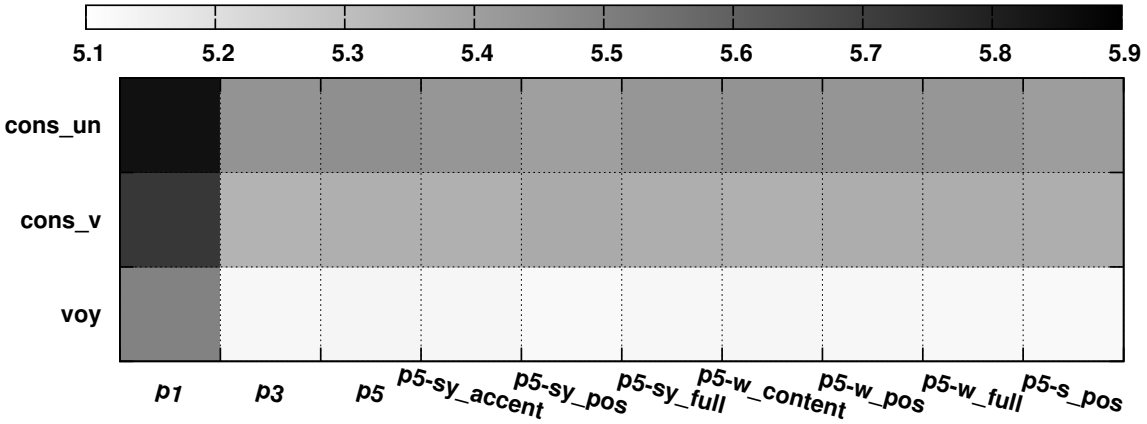


FIGURE 7.16 – Distance cepstrale moyenne déterminée entre les coefficients MGC générés par HTS et ceux extraits du signal naturel en fonction du jeu de descripteurs et de la catégorie de voisement. Les NSS ne sont pas pris en compte sur cette figure. Le jeu de descripteurs est indiqué en abscisse et la catégorie de voisement en ordonnée.

Ces résultats confirment que prendre en compte des descripteurs d'horizon plus élevé que le contexte phonétique direct ne permet pas d'améliorer la modélisation du spectre pour l'ensemble des catégories de voisement. De plus, pour l'ensemble des jeux de descripteurs évalués, les voyelles obtiennent des écarts plus faibles que les consonnes. Afin de déterminer si ce constat s'appliquer à l'ensemble des voyelles, la distance mel-cepstrale a été déterminée pour chaque catégorie phonétique définie en fonction du mode d'articulation. Les résultats sont présentés dans la figure 7.17.

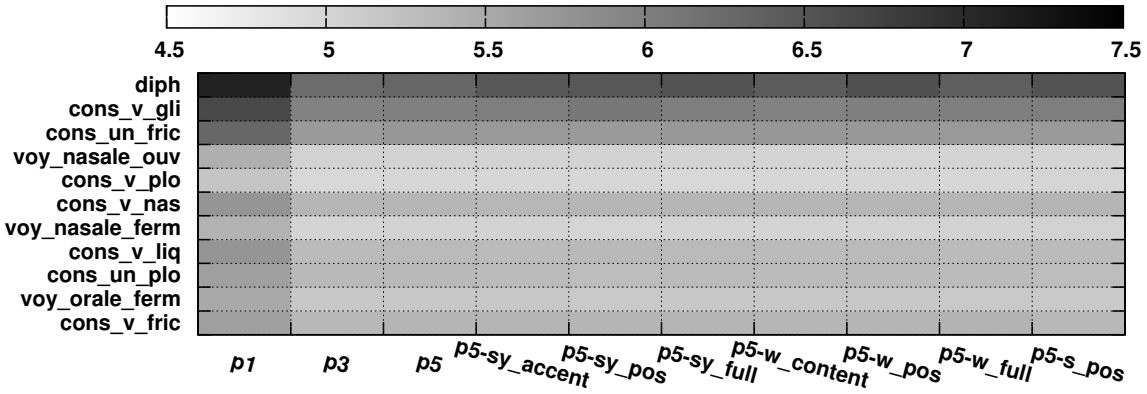


FIGURE 7.17 – Distance cepstrale moyenne déterminée entre les coefficients MGC générés par HTS et ceux extraits du signal naturel en fonction du jeu de descripteurs et de la catégorie phonétique du segment. Les NSS ne sont pas pris en compte dans cette figure. Le jeu de descripteurs est indiqué en abscisse et la catégorie de phonétique en ordonnée.

En analysant ces résultats, nous pouvons constater une disjonction entre certaines catégories phonétiques définies en se basant sur le mode d'articulation. Les intervalles de

confiance, non représentés ici, permettent de distinguer trois blocs : les voyelles couplées aux plosives voisées, la diphtongue couplée aux glissantes et aux fricatives non-voisées, puis les autres consonnes. Il semble qu'aucun jeu de descripteurs ne permette de combler l'écart entre ces blocs. Le fait que les distorsions mel-cepstrales associées à la diphtongue soient plus élevées que la majorité des distorsions obtenues pour les autres catégories de phones peut s'expliquer par un plus faible nombre de données d'apprentissage des modèles HTS. Cependant, il n'est pas possible d'avancer ce type d'argument pour les autres catégories. Ainsi, nous pouvons supposer qu'aucun descripteur ne permette à HTS de capter les spécificités propres à certains phones.

Enfin, le dernier point important à soulever concerne la métrique. En effet, nous avons opté pour l'utilisation de la distance cepstrale qui ignore l'énergie. En appliquant une RMS sur l'ensemble des coordonnées des vecteurs MGC, nous obtenons les résultats présentés dans la figure 7.18.

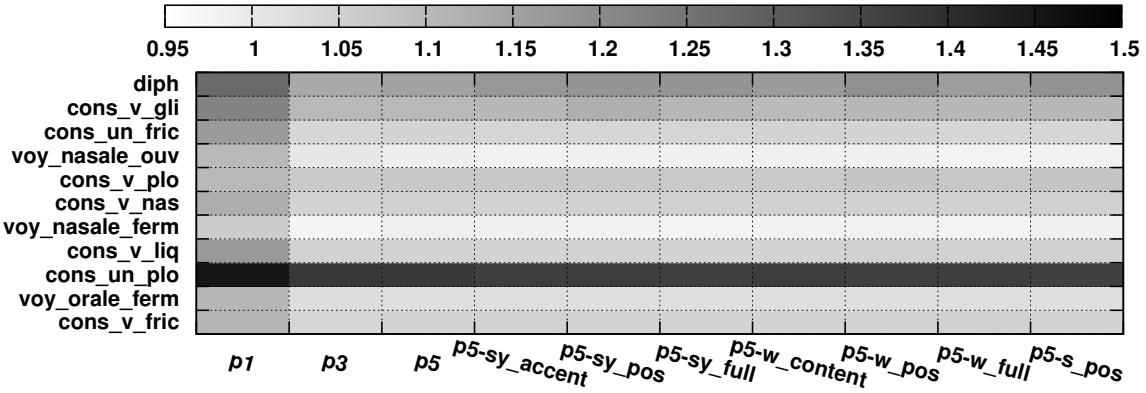


FIGURE 7.18 – RMS moyenne déterminée entre les coefficients MGC générés par HTS et ceux extraits du signal naturel en fonction du jeu de descripteurs et de la catégorie phonétique du segment. Les NSS ne sont pas pris en compte dans cette figure. Le jeu de descripteurs est indiqué en abscisse et la catégorie de phonétique en ordonnée.

Ces résultats diffèrent fortement et tendent à indiquer que les plosives non-voisées sont moins bien modélisées. Néanmoins en comparant ces résultats avec ceux de la figure 7.17, nous constatons que seule l'énergie associée à ces phones permet de les distinguer des autres phones. Ceci indique que, malgré une différence de volume, ces phones sont aussi bien modélisés que la plupart d'entre eux. Ceci confirme également que le choix de la métrique à utiliser pour appliquer le protocole peut influencer les résultats et leur interprétation.

7.2.3 Bilan de l'évaluation pour le paramètre MGC

À l'issue de l'application du protocole sur les coefficients MGC, nous pouvons proposer deux conclusions.

Tout d'abord, la seule amélioration de la qualité de la modélisation est apportée lors de la prise en compte du contexte phonétique directe (jeu de descripteurs `p3`) par rapport

au jeu de descripteurs `p1`. Comme nous avons pu le voir, cette amélioration est globale à l'ensemble des catégories phonétiques. De plus, utiliser un jeu de descripteurs qui contient plus d'informations que celles présentes dans le jeu `p3` n'améliore pas la modélisation du spectre. Ce résultat est important car il indique qu'en réalité le choix d'un jeu de descripteurs ne doit pas être guidé par la modélisation du spectre.

La seconde conclusion concerne la différence de modélisation entre les catégories de voisement. Si l'on tient compte de l'énergie, les consonnes plosives non-voisées se distinguent des autres catégories avec des écarts plus élevés. Toutefois, cette différence s'estompe dès que l'énergie est ignorée. D'après les analyses effectuées, les voyelles sont mieux modélisées que les consonnes et les consonnes les moins bien modélisées sont les fricatives non-voisées ainsi que les glissantes voisées.

7.3 Évaluation de la modélisation de la durée

Le troisième paramètre acoustique que nous allons analyser est la durée. Lors de l'évaluation par GMM, nous avons constaté que seul le jeu de descripteurs `p1` se distinguait en ayant une log-vraisemblance plus faible par rapport à l'ensemble des autres jeux de descripteurs. Comme nous l'avons indiqué précédemment, la durée est une composante particulière en ce sens que la configuration du système HTS, imposant la génération d'une trame toutes les 5 ms, empêche mécaniquement une modélisation fine de ce paramètre.

En calculant une distance entre la durée des phones [3] présents dans le corpus $T_{a/s}$ et les durées produites par HTS, nous allons pouvoir déterminer dans quelle proportion le jeu de descripteurs `p1` dégrade la durée.

7.3.1 Résultats globaux

Tout d'abord, nous avons calculé une RMS en millisecondes afin de se comparer aux résultats obtenus par S. Yokomizo ET AL. [Yokomizo2010] et H. Silén [Silen2010]. Les travaux présentés dans [Silen2010] ont pour objectif de comparer différentes modélisations de la durée par rapport à la modélisation standard utilisée dans HTS. Pour effectuer cette comparaison, une erreur RMS et un coefficient de corrélation entre la durée générée et la durée originale ont été calculés. Dans le cadre de notre étude, nous souhaitons comparer nos résultats aux valeurs de RMS présentées dans cet article. Les résultats que nous avons obtenus sont illustrés figure 7.19.

Les résultats obtenus montrent qu'aucune différence significative ne permet de distinguer un jeu de descripteurs par rapport à un autre. Toutefois, nous pouvons distinguer une nette tendance indiquant que le jeu de descripteurs `p1` dégrade davantage la modélisation

3. la durée obtenue par segmentation du signal naturel n'est connue que pour cette échelle

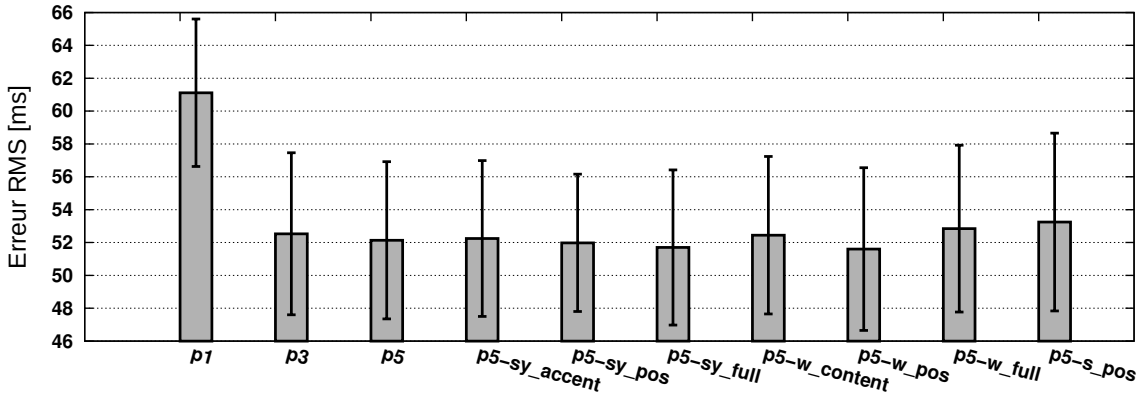


FIGURE 7.19 – RMS en ms entre la durée générée par HTS selon un jeu de descripteurs (précisé en abscisse) et la durée obtenue à l'issue de la segmentation à l'issue de la segmentation pour l'ensemble des segments du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

de la durée. Cette tendance est cohérente avec les résultats obtenus lors de l'évaluation par GMM. Cela peut indiquer que les durées produites par ce jeu de descripteurs, bien que peu vraisemblables, ne soient en réalité pas si éloignées des durées originales.

De plus, l'ensemble des valeurs de RMS obtenues se situent aux alentours de 50ms. Toutefois, les erreurs RMS présentées dans [Yokomizo2010] et [Silen2010] sont respectivement d'environ 30ms (corpus japonais)/40ms (corpus anglais) et 25ms (corpus finnois)/30ms (corpus anglais) ; nos erreurs sont plus élevées. La particularité de notre corpus concerne les NSS qui peuvent être des segments de longue durée. En ignorant les NSS, nous obtenons les résultats, présentés figure 7.20, équivalents à ceux de l'état de l'art.

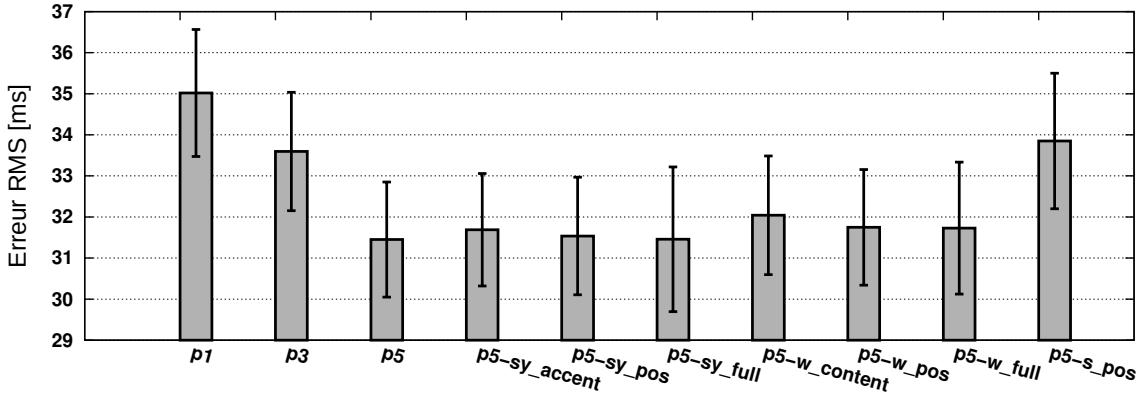


FIGURE 7.20 – RMS en ms entre la durée générée par HTS selon un jeu de descripteurs (précisé en abscisse) et la durée obtenue à l'issue de la segmentation pour l'ensemble des phones du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

Ces résultats sont cohérents avec ceux obtenus lors de l'évaluation par GMM (et illustrés figure 6.11) à deux exceptions près : `p3` et `p5-s_pos`. les RMS associées à ces jeux de descripteurs ne sont pas significativement différentes de celles obtenues par `p1`.

Les résultats que nous obtenons se situent dans les intervalles présentés dans [Yokomizo2010] et [Silen2010]. En comparant les erreurs RMS obtenues pour chaque jeu de descripteurs, nous constatons que le jeu `p1` est toujours considéré comme le moins pertinent. Toutefois, les jeux `p3` et `p5-s_pos` ont une erreur RMS légèrement plus élevée que les autres jeux de descripteurs qui se situent tous aux alentours de 31ms. Néanmoins, ces différences n'étant pas statistiquement significatives, nous ne pouvons réellement conclure sur ces jeux de descripteurs.

Pour compléter cette étude globale, un écart relatif moyen entre la durée générée et la durée issue de la segmentation a été calculé pour chacun des jeux de descripteurs sans tenir compte des NSS. Les résultats obtenus sont présentés figure 7.21.

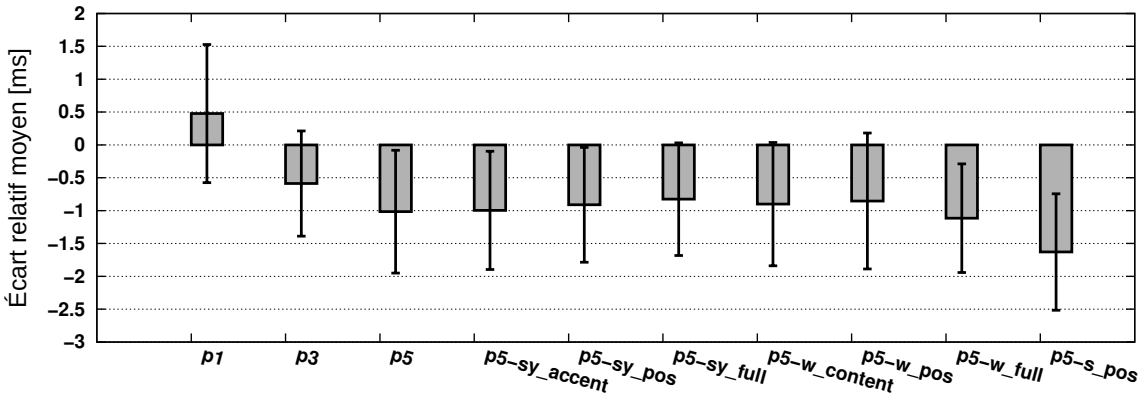


FIGURE 7.21 – Moyenne des écarts relatifs entre la durée générée par par HTS selon un jeu de descripteurs (précisé en abscisse) et la durée obtenue à l'issue de la segmentation pour l'ensemble des phones du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95% et les NSS ne sont pas pris en compte.

Ces résultats montrent que les écarts semblent se compenser et aboutissent à une durée moyenne de -0.5ms, ce qui est inférieur à la durée d'une trame. Les résultats présentés figure 7.21 indiquent que les segments ne sont, en moyenne, ni plus courts ni plus longs car l'écart maximal, qui est de 2.5ms, est inférieur au pas d'analyse temporel d'une trame.

7.3.2 Résultats par catégorie de voisement

Pour compléter l'analyse de la durée, une RMS à l'horizon du phone a été calculée pour chaque catégorie de voisement. Les résultats sont présentés figure 7.22.

Ces résultats indiquent que la durée des voyelles est celle qui obtient le plus d'erreurs et qui explique les différences entres les RMS présentées figure 7.20. En revanche le jeu de descripteurs `p5-s_pos` aboutit à des durées générées plus éloignées des durées observées dans le corpus $T_{a/s}$ pour l'ensemble des catégories de voisement. En considérant les résultats, figure 7.23, présentant les valeurs de RMS en tenant compte de l'intervalle de confiance pour les consonnes voisées, nous constatons que ces différences ne sont statistiquement pas significatives. La largeur de l'intervalle de confiance associé à `p5-s_pos` montre toutefois

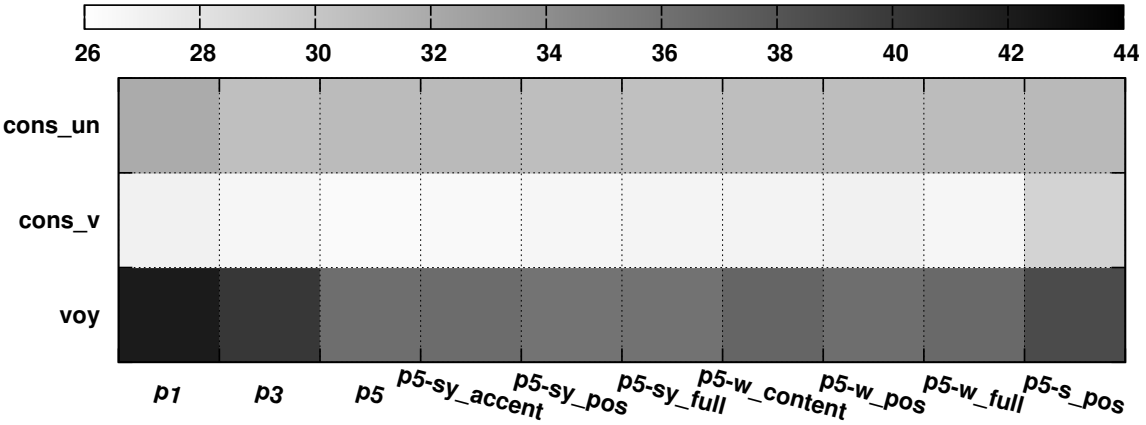


FIGURE 7.22 – RMS entre la durée générée par HTS et la durée obtenue à l'issue de la segmentation des phones du corpus de test en fonction du jeu de descripteurs et de la catégorie de voisement. Le jeu de descripteurs est indiqué en abscisse et la catégorie de voisement en ordonnée.

que la génération de la durée basée sur ce jeu de descripteurs est plus variable que pour les autres jeux de descripteurs.

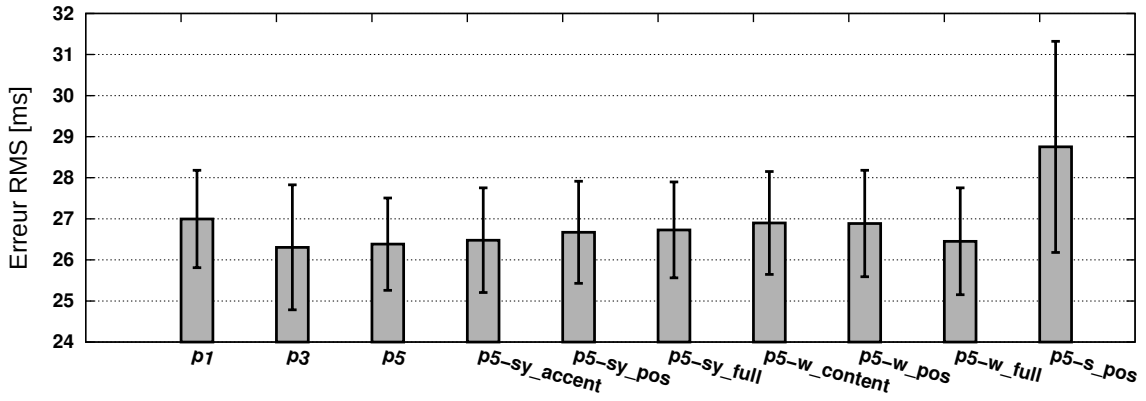


FIGURE 7.23 – RMS en ms entre la durée générée par HTS selon un jeu de descripteurs (précisé en abscisse) et la durée obtenue à l'issue de la segmentation pour les consonnes non-voisées du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance 95%.

7.3.3 Résultats par label phonétique

En affinant l'analyse par un calcul de la RMS pour chaque label phonétique, nous obtenons les résultats présentés figure 7.24.

Les résultats obtenus montrent que la modélisation de la durée la moins pertinente est pour le phone *oe* et ceci quel que soit le jeu de descripteurs. Toutefois, comme le montre la figure 7.25, cette différence n'est pas significative. De plus, la taille de l'intervalle de confiance peut s'expliquer par le nombre d'occurrences du phone *oe* dans le corpus de test (27 occurrences soit 0.42% du corpus) et du corpus d'apprentissage (188 occurrences soit

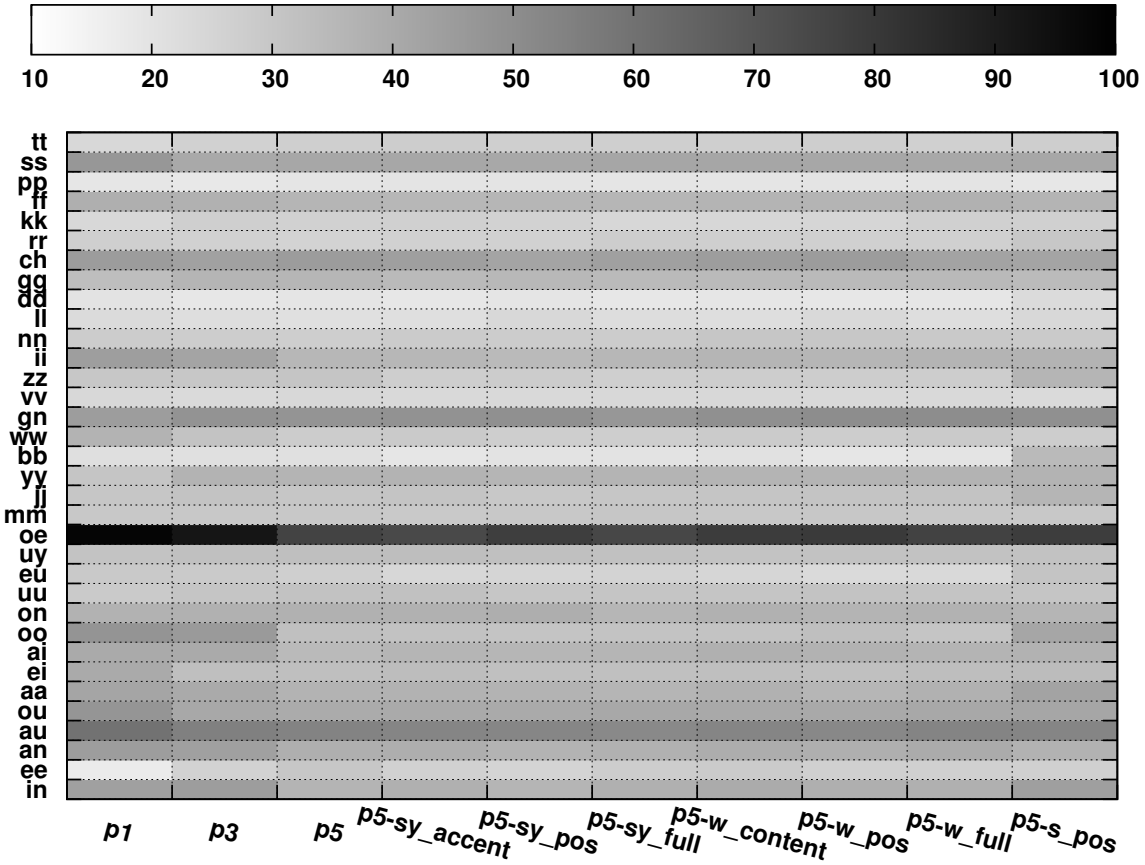


FIGURE 7.24 – RMS en ms entre la durée générée par HTS selon un jeu de descripteurs (précisé en abscisse) et la durée obtenue à l'issue de la segmentation, en fonction du jeu de descripteurs et du label phonétique. Le jeu de descripteurs est indiqué en abscisse et le label phonétique en ordonnée.

0.48% du corpus) ainsi que par la variabilité de la durée de ce phone dans les données naturelles (l'écart-type associé à *oe* est de 97ms contre 38ms pour l'ensemble des phones du corpus de test, et de 61ms contre 35ms pour l'ensemble des phones du corpus d'apprentissage).

7.3.4 Résultats pour le débit syllabique

Nous avons calculé le débit syllabique de chaque énoncé du corpus de test pour chacun des jeux de descripteurs. Nous avons comparé chacun des débits syllabiques produits par HTS au débit syllabique original. Le débit syllabique original est, en moyenne, de 5 syllabes par seconde. Ceci implique que le locuteur a un débit lent car d'après l'étude de B. Zellner [Zellner1998], le débit syllabique du français se situe entre 4 et 7 syllabes par secondes. En calculant un écart relatif moyen entre les différents débits nous obtenons les résultats présentés figure 7.26. Ces résultats montrent que le débit syllabique ne permet pas de discriminer un jeu particulier de descripteurs.

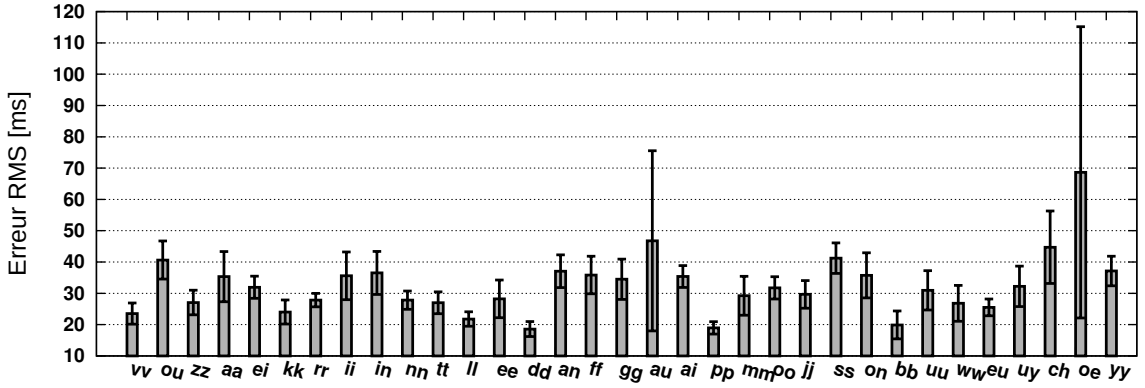


FIGURE 7.25 – RMS en ms entre la durées générée par HTS selon le jeu de descripteurs `p5-sy_full` et la durée obtenue à l'issue de la segmentation pour l'ensemble des phones du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance 95%. et l'axe des abscisses contient les labels phonétiques.

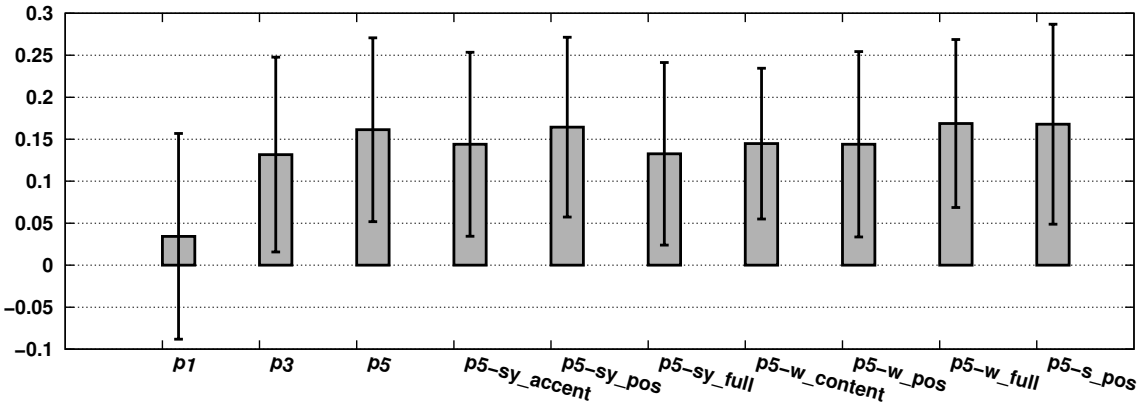


FIGURE 7.26 – Moyenne des écarts relatifs entre les débits syllabiques générés par HTS et le débit syllabique original pour un même segment phonétique. L'axe des abscisses correspond aux jeux de descripteurs. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

7.3.5 Bilan de l'évaluation pour le paramètre de durée

À l'issue de l'évaluation de la modélisation de la durée, deux constats peuvent être mis en avant. Tout d'abord, le jeu de descripteurs considéré comme optimal pour modéliser la durée est le jeu de descripteurs `p5`. L'intervalle de confiance de la RMS associée au jeu `p3` chevauche celui de la RMS associée au jeu `p1` et le jeu de descripteurs `p1` obtient la RMS moyenne la plus élevée. Cette mesure est la seule nous permettant de distinguer les jeux de descripteurs. Nous pouvons alors considérer que le jeu de descripteurs optimal pour modéliser la durée est le jeu `p5`.

En comparant les catégories de voisement, nous avons mis en avant que les voyelles obtiennent des écarts plus élevés que les consonnes. Toutefois, ce résultat s'applique en tenant compte de l'ensemble des voyelles. La RMS associée à la voyelle /oe/ qui semble ressortir des tableaux n'est, en définitive, pas significativement différente des RMS associées

aux autres voyelles.

7.4 Évaluation de la modélisation de l'apériodicité

Le dernier paramètre acoustique analysé dans ce chapitre est l'apériodicité. Dans le chapitre précédent (section 6.5), l'évaluation paramétrique des coefficients d'apériodicité ont montré une variabilité plus forte entre les log-vraisemblances associées aux différents jeux de descripteurs évalués que pour les autres types de coefficients (MGC, F0 et durée). Néanmoins, lors de l'évaluation objective par GMM, aucun jeu de descripteurs, hormis `p1`, ne se distinguait de manière significative.

Dans cette section, nous allons tout d'abord voir si l'application du protocole confirme ou infirme les résultats obtenus lors de l'évaluation par GMM. Comme précédemment, nous compléterons des analyses globales par une étude à l'échelle du phone.

7.4.1 Résultats globaux

En utilisant une mesure RMS pour déterminer un écart représentatif pour chaque jeu de descripteurs, nous obtenons les résultats illustrés figure 7.27.

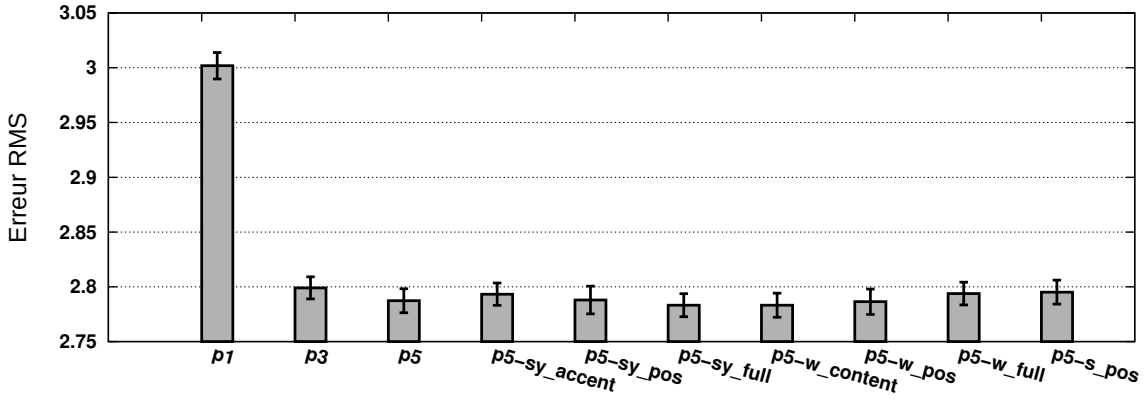


FIGURE 7.27 – RMS entre les coefficients d'apériodicité générés par HTS selon le jeu de descripteurs spécifié en abscisse et ceux extraits du signal naturel pour l'ensemble des trames du corpus de test. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

Ces résultats montrent que la valeur d'apériodicité générée par HTS en utilisant le jeu de descripteurs `p1` est la plus éloignée de l'apériodicité extraite du signal naturel par STRAIGHT. De plus, sans pour autant être statistiquement significatives, nous pouvons remarquer quelques variations entre les autres jeux de descripteurs. Ces résultats sont donc proches de ceux obtenus lors de l'évaluation par GMM (voir la figure 6.13). En ignorant les NSS, les erreurs RMS obtenues sont illustrées par la figure 7.28. Ces erreurs confirment les tendances.

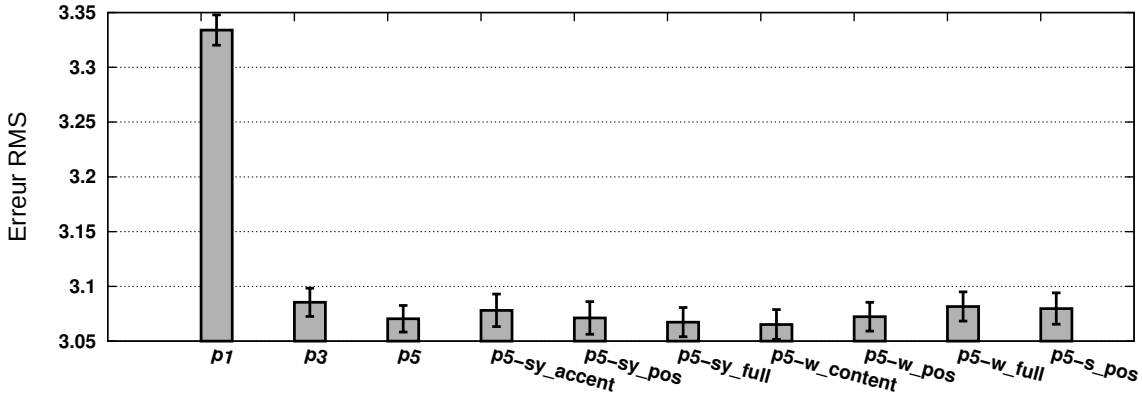


FIGURE 7.28 – RMS entre les coefficients d'apériodicité générés par HTS selon le jeu de descripteurs spécifié en abscisse et ceux extraits du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

H. Silén ET AL.[Silen2011] ont proposé une méthode pour prédire les coefficients d'apériodicité en se basant uniquement sur les coefficients spectraux. Pour comparer les coefficients d'apériodicité qu'ils ont obtenus en utilisant leur méthode par rapport à ceux déterminés par la version standard de HTS, H. Silén ET AL. ont calculé une erreur RMS pour chaque bande d'apériodicité. Afin de nous comparer à ces résultats, nous avons appliqué la même méthodologie ; les valeurs RMS ainsi calculées sont présentées figure 7.29.

Dans leur étude, les auteurs de [Silen2011] mentionnent des RMS aux alentours de 4 pour les deux premiers coefficients, 3 pour le troisième, 2 pour les coefficients d'ordre 4 et 5. Les erreurs RMS obtenues pour le corpus CORDIAL sont comparables à ces valeurs, ce qui permet de supposer la cohérence des résultats obtenus.

7.4.2 Résultats par catégorie de voisement

En appliquant l'analyse à l'horizon du phone en fonction des catégories de voisement, nous obtenons les résultats présentés figure 7.30.

Ces résultats montrent que l'amélioration de la modélisation, introduite par le jeu de descripteurs `p3` par rapport au jeu `p1`, se situe sur les consonnes voisées et les voyelles. De plus, pour l'ensemble des jeux de descripteurs hormis `p1`, les résultats sont identiques. Enfin, en comparant les valeurs de RMS pour un même jeu de descripteurs nous constatons que les voyelles apportent des valeurs plus élevées que pour les consonnes. En appliquant le protocole en fonction des modes d'articulation, nous obtenons les résultats illustrés figure 7.31.

Ces résultats confirment que l'amélioration apportée par `p3` par rapport à `p1` concerne l'ensemble des catégories exceptées pour les fricatives. La seule variation constatée sur les autres descripteurs concernent la diphtongue qui est très peu représentée dans le corpus

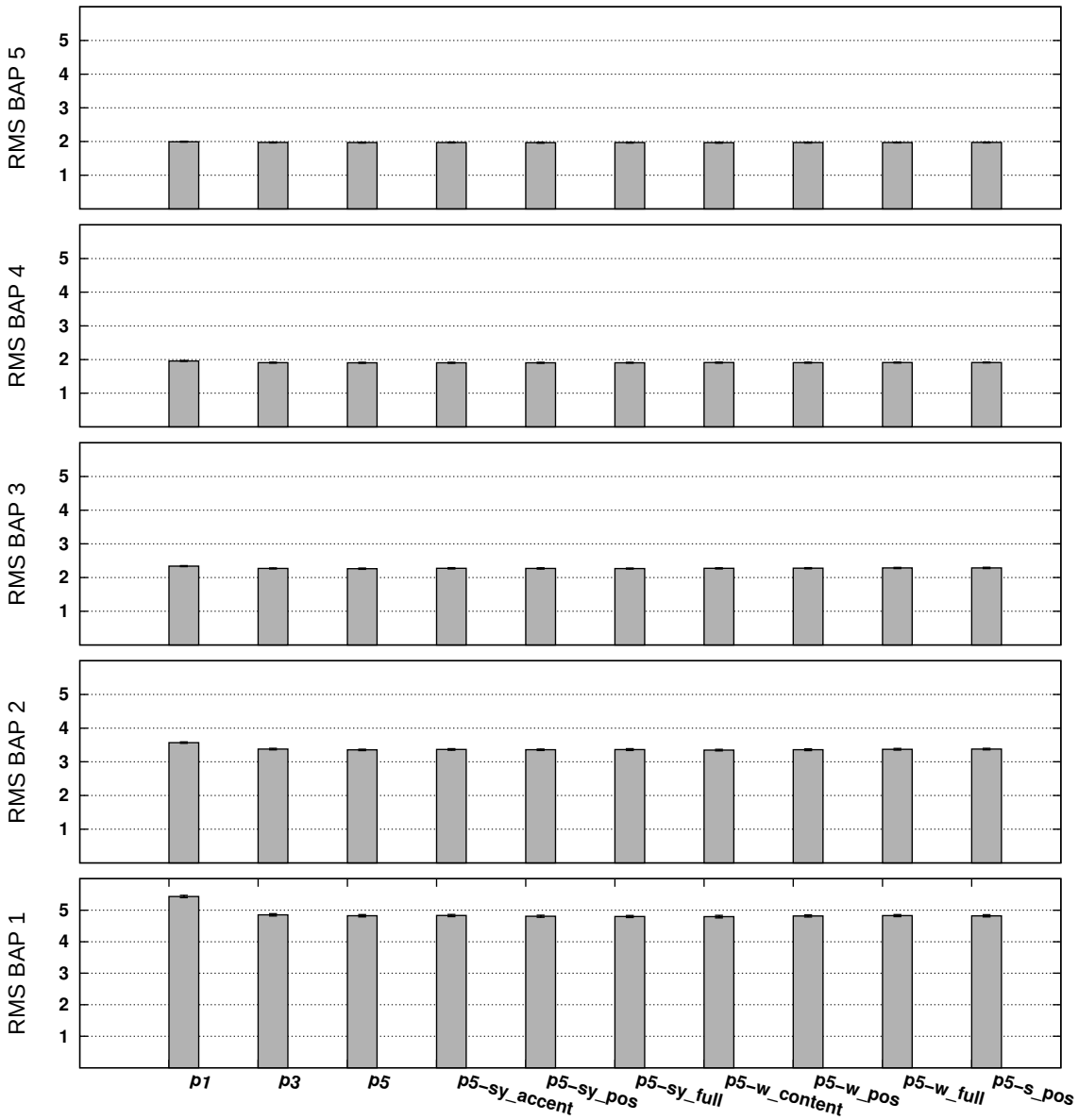


FIGURE 7.29 – RMS par dimension du vecteur de coefficients d'apériodicité entre les coefficients d'apériodicité générés par HTS selon le jeu de descripteurs spécifié en abscisse et ceux extraits du signal naturel pour les trames du corpus de test qui ne sont pas étiquetées NSS. Les intervalles de confiance correspondent à un niveau de confiance de 95%.

d'apprentissage et le corpus de test. Enfin, en comparant les catégories phonétiques, nous remarquons que les fricatives se distinguent en obtenant une RMS plus faible et que l'apériodicité associée aux voyelles est considérée comme moins pertinente pour l'ensemble des voyelles.

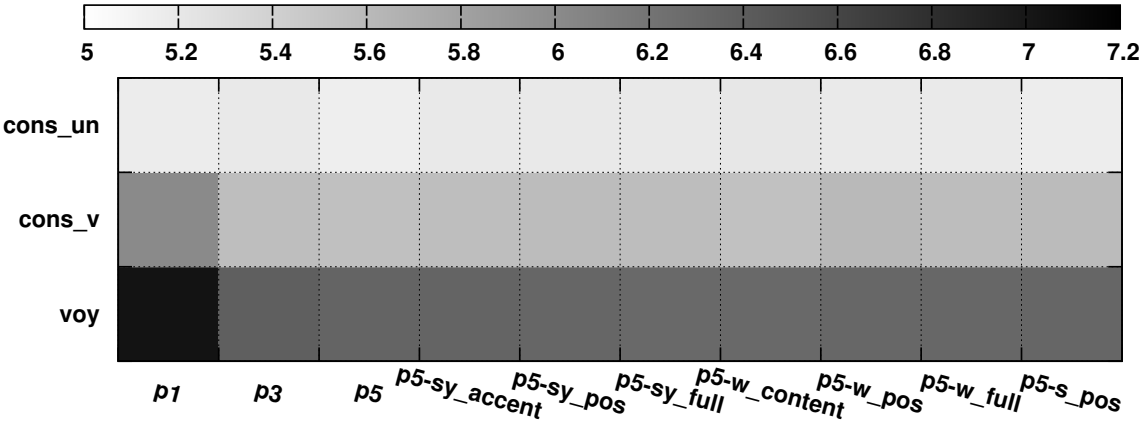


FIGURE 7.30 – RMS, entre les coefficients d'apériodicité générés par HTS et les coefficients d'apériodicité extraits du signal naturel pour une même trame, en fonction du jeu de descripteurs et de la catégorie de voisement. Le jeu de descripteurs est indiqué en abscisse et la catégorie de voisement en ordonnée.

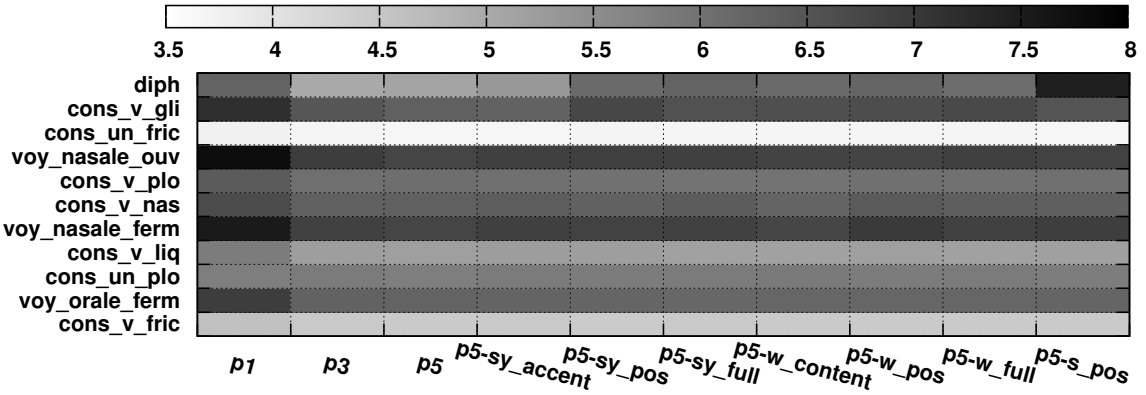


FIGURE 7.31 – RMS, entre les coefficients d'apériodicité générés par HTS et les coefficients d'apériodicité extraits du signal naturel pour une même trame, en fonction du jeu de descripteurs et de la catégorie de phonétique. Le jeu de descripteurs est indiqué en abscisse et la catégorie de phonétique en ordonnée.

7.4.3 Bilan de l'évaluation pour le paramètre d'apériodicité

Les expériences menées sur l'apériodicité ont montré que le jeu de descripteurs `p1` aboutit à une modélisation de l'apériodicité moins pertinente que les autres jeux de descripteurs. Les autres jeux obtiennent des résultats similaires les uns par rapport aux autres malgré de légères tendances non significatives.

En analysant les résultats par catégories de phones, les écarts les plus élevés sont obtenus par les voyelles et ceci pour la majorité des voyelles. De plus les fricatives, voisées et non voisées, obtiennent les écarts les plus faibles.

7.5 Bilan et conclusion

Dans ce chapitre, nous avons effectué une analyse complète et précise de l'influence des descripteurs sur la modélisation, effectuée par HTS, des quatre paramètres acoustiques : F0, coefficients MGC, coefficients d'apériodicité et durée.

À l'issue de ces analyses, nous pouvons conclure, comme pour le chapitre précédent, que l'utilisation de la seule étiquette, permettant d'identifier le segment courant, ne suffit pas. En effet, pour l'ensemble des analyses effectuées, les résultats associés au jeu de descripteurs `p1` montrent une dégradation plus forte que pour les autres jeux de descripteurs. Les résultats obtenus lors de l'évaluation non-paramétrique indique que le jeu de descripteurs optimal est `p5-sy_full`. Pour aboutir à cette conclusion, le F0 est le paramètre acoustique déterminant car, si l'on considère la durée, les coefficients MGC ou bien l'apériodicité, l'utilisation des informations phonétiques suffit à obtenir une modélisation dont la qualité est équivalente à celle obtenue avec des descripteurs de niveaux supérieurs.

Dans un second temps, nous avons montré que pour l'ensemble des paramètres acoustiques évalués, des différences de qualité de modélisation subsistent entre les catégories de voisement de phones. En effet, pour le F0 et les coefficients MGC, les consonnes non-voisées ont des écarts plus élevés que les voyelles. En revanche, pour la durée et les coefficients d'apériodicité, nous constatons le contraire. De plus, malgré l'introduction de nouveaux descripteurs, ces différences ne sont pas comblées. Ceci confirme que l'introduction d'un descripteur permet d'améliorer globalement la qualité de modélisation d'un type de coefficient.

Chapitre 8

Évaluation subjective

Afin de situer les résultats obtenus par les évaluations objectives décrites dans ce document, des évaluations subjectives ont été réalisées. Les deux évaluations objectives ont désigné comme jeux de descripteurs optimaux le jeu `p3`, pour la durée et les coefficients MGC, ainsi que `p5-sy_full` pour les coefficients d'apériodicité et pour le F0. Le jeu de descripteurs `p5-sy_full` serait le jeu de descripteurs à privilégier car il permet d'obtenir le meilleur compromis entre le nombre de descripteurs utilisés et la qualité de modélisation de l'ensemble des types de coefficients. Les résultats présentés dans les chapitres précédents ont également montré que le jeu de descripteurs `p1` est considéré comme insuffisant pour obtenir une bonne modélisation.

L'objectif des évaluations subjectives étant de confronter les résultats obtenus à la perception humaine, nous avons réalisé trois évaluations différentes. La première a consisté à déterminer une qualité globale en comparant des notes d'évaluation afin de les confronter aux résultats des évaluations objectives. La seconde évaluation a pour objectif de quantifier la dégradation globale obtenue à l'issue d'une synthèse HTS par rapport au signal naturel. Pour la troisième évaluation, les modèles sont identiques mais les énoncés synthétisés sont

différents. L'objectif de la dernière évaluation est de déterminer l'influence des descripteurs sur des énoncés qui ne sont pas dans le registre linguistique du corpus d'apprentissage des modèles HTS.

8.1 Évaluation subjective globale

La première évaluation subjective conduite a pour objectif de mettre à l'épreuve les résultats obtenus à l'issue des évaluations objectives. Conduire une évaluation subjective ayant un coût humain non-négligeable, il n'est cependant pas possible de valider chacun des points analysés lors du chapitre précédent.

8.1.1 Données évaluées

Pour effectuer cette évaluation, nous utilisons un sous-ensemble des énoncés issus du corpus de test. Sept systèmes ont été évalués :

- le signal naturel qui correspond à la référence absolue ;
- le signal issu de l'analyse/synthèse, effectuée par STRAIGHT et SPTK, qui correspond à la référence pour HTS ;
- le signal issu de la génération effectuée par HTS en utilisant le jeu de descripteurs `p1` qui correspond à la configuration minimale et qui devrait donc produire le signal le moins bien perçu.
- le signal issu de la génération effectuée par HTS en utilisant le jeu de descripteurs `p3` qui correspond à la configuration optimale pour la modélisation du spectre selon les résultats des évaluations objectives ;
- le signal issu de la génération effectuée par HTS en utilisant le jeu de descripteurs `p5` qui permet de pouvoir évaluer l'apport du contexte phonétique d'horizon deux par rapport au contexte phonétique directe ;
- le signal issu de la génération effectuée par HTS en utilisant le jeu de descripteurs `p5-sy_full` qui correspond à la configuration optimale selon les évaluations objectives ;
- le signal issu de la génération effectuée par HTS en utilisant le jeu de descripteurs `p5-s_pos` qui correspond à la configuration la plus complète ;

Pour chacun de ces 7 systèmes, 31 stimuli, correspondant à des énoncés dont la durée est comprise entre 3 et 8 secondes, ont été sélectionnés. Cette contrainte a été définie afin de s'assurer que les échantillons puissent être suffisamment longs pour pouvoir percevoir leur qualité mais leur durée doit être suffisamment homogène afin de pouvoir être comparables. La durée moyenne des stimuli est d'environ 6 secondes. Enfin, les énoncés ont été sélectionnés dans le corpus de test utilisé pour effectuer les évaluations objectives

des chapitres précédents et dont les principales caractéristiques sont présentées dans la section 5.3 du chapitre 5.

8.1.2 Protocole

Le protocole d'évaluation utilisé est une évaluation subjective permettant d'obtenir un score absolu de type MOS (Mean Opinion Score).

En se basant sur ces modèles, la question posée aux testeurs est la suivante :

« *Comment jugez-vous globalement la qualité de ce que vous venez d'entendre ?* »

Les réponses possibles étaient alors :

– Excellente (5)

– Bonne (4)

– Passable (3)

– Médiocre (2)

– Mauvaise (1)

En ce qui concerne le déroulement, 10 auditeurs (experts dans le domaine de la parole) ont réalisé le test qui comportait 112 étapes (dont 7 d'introduction pour lesquelles les résultats n'ont pas été pris en compte). Un carré latin a été défini pour générer les instances de test et a permis de faire évaluer 30 stimuli, par système, par exactement 5 auditeurs différents.

8.1.3 Résultats

Les résultats obtenus à l'issue de cette évaluation sont illustrés dans la figure 8.1.

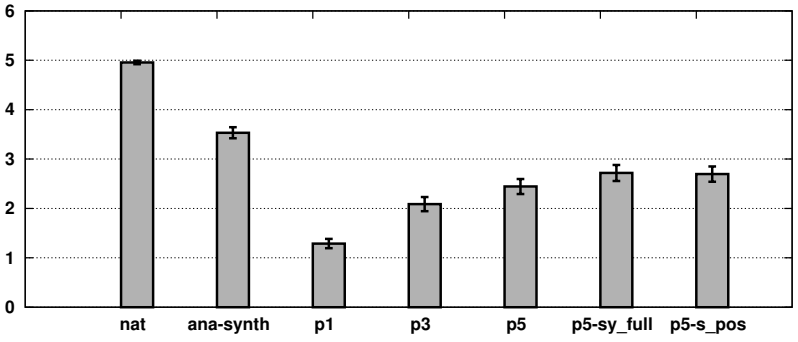


FIGURE 8.1 – Résultats du test MOS

En comparant les résultats obtenus par l'analyse/synthèse, par rapport à ceux obtenus par le signal naturel, nous pouvons constater qu'une dégradation, due à la paramétrisation STRAGIHT a été perçue par les auditeurs. Parmi les signaux générés par HTS, nous

distinguons trois paliers : le signal obtenu grâce au jeu de descripteurs `p1` est le plus mauvais, qualifié de “médiocre” ; une amélioration nette est perçue grâce à l'introduction du contexte direct par le jeu `p3`, la qualité des signaux testés étant jugée comme passable ; les jeux de descripteurs `p5`, `p5-sy_full` et `p5-s_pos` obtiennent des notes plus élevées mais sont en moyenne considérés comme passables.

Si on suit cette évaluation subjective, une amélioration de la synthèse a été constatée entre les jeux de descripteurs `p3`, `p5` et `p5-sy_full`. Si la nette amélioration constatée entre `p1` et `p3` peut être considérée comme une amélioration sur l'ensemble des types de coefficients, la différence entre les autres paliers est moins significative. Ainsi, soit l'amélioration, liée à ces paliers, est moins significative pour l'ensemble des paramètres acoustiques soit seule la qualité d'une partie de ces paramètres acoustiques s'est améliorée. Bien qu'une évaluation type MOS ne nous permette pas de départager les deux hypothèses, elle ne contredit pas les résultats obtenus par les évaluations objectives.

8.2 Évaluation subjective de la dégradation

Une seconde évaluation subjective a été conduite afin de mesurer la dégradation globale due à HTS. Pour cela, nous avons décidé d'effectuer un test de type différentiel afin de quantifier la dégradation du signal issu du système HTS par rapport au signal naturel. Ainsi, après avoir présenté les données utilisées et le protocole appliqué, nous exposerons les résultats obtenus.

8.2.1 Données évaluées

Pour effectuer cette évaluation, nous avons sélectionné les six systèmes suivants :

- le signal naturel qui constitue la référence ;
- le signal naturel dégradé suivant le procédé MNRU [1] avec un ratio Q de 20dB. Ce ratio a été déterminé manuellement afin de pouvoir obtenir un ancrage pour séparer la qualité de la synthèse effectuée par STRAIGHT et HTS ;
- le signal issu de l'analyse/synthèse effectuée par STRAIGHT et SPTK ;
- le signal issu de la synthèse effectuée par le jeu de descripteur `p5-sy_accent`. Ce jeu de descripteurs a été sélectionné car il correspond au jeu directement « inférieur » à `p5-sy_full` qui contient uniquement des informations prosodiques. En utilisant ce jeu de descripteurs, nous cherchons à évaluer l'apport d'informations qu'un utilisateur pourrait contrôler directement, ici les informations d'accentuation, de la prosodie ;

1. Le MNRU (Modulated Noise Reference Unit) consiste à introduire un bruit aléatoire sur le signal d'origine. Le ratio Q détermine le rapport entre la puissance du bruit et celle du signal vocal. Ce ratio est un paramètre de la méthode.

– le signal issu de la synthèse effectuée par le jeu de descripteur `p5-s_pos` qui contient l'ensemble des 44 descripteurs proposés pour le français et décrit dans la section 3.3 du chapitre 3 page 53 ;
– le signal issu de la synthèse effectuée par le jeu de descripteurs `p5` qui contient l'ensemble des informations phonétiques.

Pour chacun de ces six systèmes, 31 stimuli, correspondant à des énoncés extraits du corpus de test utilisé pour l'évaluation précédente, ont été sélectionnés.

8.2.2 Protocole

Le protocole d'évaluation est un test de dégradation, décrit par la norme p. 800 de l'ITU-T [ITU-T1996], qui permet d'obtenir un score DMOS (Differential Mean Opinion Score). Le score obtenu permet de quantifier la dégradation constatée entre un signal référent, dans notre cas le signal naturel, et un signal dégradé.

Pour guider l'utilisateur, la question suivante lui a été posée :

> « *Comment appréciez-vous la dégradation de l'échantillon dégradé par rapport à l'échantillon référence ?* »

Puis, après avoir écouté les deux échantillons, les auditeurs ont eu le choix entre les réponses suivantes :

– Dégradation inaudible (5)
– Dégradation audible mais pas gênante (4)
– Dégradation un peu gênante (3)
– Dégradation gênante (2)
– Dégradation très gênante (1)

En ce qui concerne le déroulement, 10 auditeurs (experts dans le domaine de la parole) ont réalisé le test qui comportait 60 étapes (dont 6 d'introduction pour lesquelles les résultats n'ont pas été pris en compte). En utilisant un carré latin pour générer les instances de test, cela a permis de faire évaluer 30 stimuli, par système, par exactement 3 auditeurs différents.

8.2.3 Résultats

Les résultats obtenus à l'issue de l'évaluation sont illustrés figure 8.2.

Tout d'abord, le résultat associé au signal naturel sont cohérents : tous les auditeurs, pour tous les échantillons correspondant au signal naturel, n'ont noté aucune dégradation. De plus, les échantillons dégradés avec l'algorithme MNRU ont tous obtenu une note

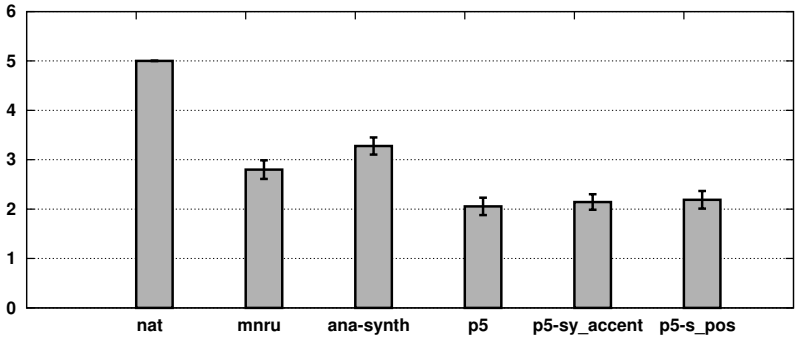


FIGURE 8.2 – Résultats du test DMOS.

moyenne plus faible que ceux associés au signal naturel. Cela nous permet de définir une dégradation référence à laquelle nous pourrons comparer les résultats obtenus par synthèse.

Dans un deuxième temps, les échantillons associés au système STRAIGHT sont classés entre le échantillons de signaux naturels et ceux dégradés par le MNRU. Nous pouvons donc considérer que la dégradation effectuée par STRAIGHT est plus faible que l'ajout d'un bruit de 20dB.

Enfin l'ensemble des résultats obtenus pour les échantillons obtenus par synthèse via le système HTS sont en deça de ceux des échantillons obtenus par analyse/synthèse ainsi que ceux obtenus après application du MNRU. Cela indique qu'une dégradation est due à la modélisation et la génération effectuées par le système HTS. En comparant les résultats de ces échantillons au score obtenu par le MNRU, nous pouvons considérer que la dégradation après utilisation du couple STRAIGHT/HTS est plus forte que l'application d'un bruit de 20dB.

Enfin, en comparant les scores obtenus par les synthèses HTS, nous ne constatons pas de différence significative malgré une légère tendance plaçant `p5-sy_accent` au dessus de `p5`. Cette évaluation basée sur la dégradation globale ne permet pas de discriminer l'influence du spectre et de la prosodie. Ainsi, il est possible que, lors de cette évaluation, la dégradation du spectre ait eu plus d'influence que celle de la prosodie. Toutefois ces résultats sont cohérents avec ceux obtenus par les évaluations objectives.

8.3 Évaluation subjective sur des énoncés différents

Le dernier test d'écoute a pour objectif d'évaluer la synthèse effectuée par HTS en utilisant des énoncés qui ne sont pas issus du même texte que les énoncés utilisés pour apprendre les modèles. Les énoncés que nous avons utilisés lors de la phase de synthèse sont des phrases phonétiquement équilibrées qui ont été proposées par P. Combescure [combescure1980]. Puisque le locuteur n'a pas émis ces énoncés, nous ne possédons pas

de signal naturel associé à ces énoncés. Nous ne pouvons donc pas effectuer un test de dégradation. Nous avons opté pour une évaluation absolue (permettant d'obtenir un score MOS) telle que décrite dans la norme p-800 de l'ITU-T [ITU-T1996]. Après avoir présenté le protocole, nous exposerons les résultats obtenus à l'issue de cette évaluation.

8.3.1 Données évaluées et protocoles

Comme précédemment, six systèmes ont été proposés pour effectuer cette évaluation. Ces systèmes sont les suivants :

- le signal naturel ;
- le signal naturel dégradé suivant le procédé MNRU avec un ratio de 20dB ;
- le signal issu de l'analyse/synthèse effectuée par STRAIGHT et SPTK.
- le signal issu de la synthèse effectuée par le jeu de descripteurs `p3` afin de déterminer si en réduisant l'horizon de description phonémique nous dégradons la modélisation.
- le signal issu de la synthèse effectuée par le jeu de descripteurs `p5` ;
- le signal issu de la synthèse effectuée par le jeu de descripteurs `p5-s_pos*` qui correspond au jeu `p5-s_pos` sans les informations d'accentuation au niveau de la syllabe car nous ne disposions pas de modèles prosodiques ce qui nous empêche de prédire l'accentuation à partir du texte ;

Enfin, pour chacun de ces six systèmes, 31 stimuli ont été sélectionnés pour une durée moyenne d'environ 3 secondes. En effet, la synthèse effectuée par HTS pour les énoncés de P. Combescure ont une durée plus faible (entre 2 et 3 secondes pour l'ensemble des 100 phrases synthétisées) que les échantillons issus du corpus Cordial (entre 2 et 8 secondes). Nous avons donc sélectionné des échantillons dont la durée est proche de 3 secondes.

En ce qui concerne le déroulement, 10 auditeurs (qui ne travaillent pas dans le domaine de la parole) ont réalisé le test qui comportait 66 étapes (dont 6 d'introduction pour lesquelles les résultats n'ont pas été pris en compte). En utilisant un carré latin pour générer les instances de test, cela a permis de faire évaluer 30 échantillons, par système, par exactement 6 auditeurs différents.

8.3.2 Résultats

Les résultats obtenus à l'issue de cette évaluation sont illustrés figure 8.3.

Ces résultats sont proches de ceux obtenus lors des précédentes évaluations subjectives : le signal naturel obtient la meilleure note et le signal dégradé par le MNRU avec une valeur Q de 20dB permet de discriminer l'analyse/synthèse de la synthèse HTS.

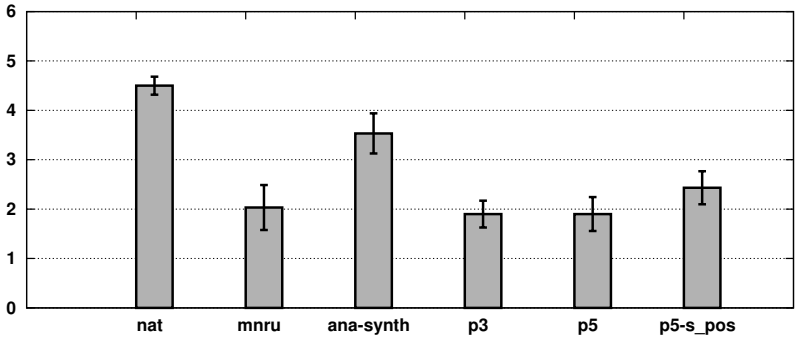


FIGURE 8.3 – Résultat du test MOS sur le corpus Combescure.

En comparant les signaux produits par HTS, nous constatons que les jeux `p3` et `p5` sont considérés comme équivalents. Le jeu de descripteurs `p5-s_pos*` obtient un score plus élevé sans toutefois que cela soit statistiquement significatif. Ces résultats sont donc cohérents avec l'ensemble des résultats obtenus lors des évaluations subjectives et ceci malgré un changement de structure linguistique du corpus.

8.4 Conclusion

Dans ce chapitre, nous avons présenté les protocoles et résultats de trois évaluations subjectives. Les deux premières évaluations, réalisées par des experts en traitement automatique de la parole, ont consisté, respectivement, à évaluer globalement la synthèse effectuée par HTS et à quantifier la dégradation obtenue à l'issue de la synthèse HTS par rapport au signal naturel. La dernière évaluation, réalisée par des personnes qui ne sont pas du domaine, a consisté à estimer la qualité absolue et globale de la synthèse effectuée par HTS sur des énoncés qui ne font pas partie du corpus CORDIAL.

À l'issue de ces évaluations, nous avons constaté que les résultats obtenus confirmaient ceux obtenus par les évaluations subjectives et décrits dans les deux chapitres précédents. En effet, le jeu de descripteurs `p1` ne permet pas à HTS d'effectuer une modélisation pertinente. Dans le cadre de la première évaluation subjective, le jeu de descripteurs `p5-sy_full` ne distingue pas du jeu de descripteurs `p5-s_pos` ce qui confirme que `p5-sy_full` est le jeu de descripteurs optimal. En effet, pour une qualité de synthèse équivalente au jeu de descripteurs `p5-s_pos`, le jeu `p5-sy_full` utilise 20 descripteurs de moins. De plus, dans le second test, l'amélioration de la synthèse associée au jeu de descripteurs `p5-s_pos` est faible comparée au jeu de descripteurs `p5-sy_accent`. Bien que le protocole diffère, il semble que l'amélioration de la modélisation apportée par `p5-sy_full` par rapport à `p5-sy_accent` soit limitée ce qui confirme, une fois de plus, les résultats obtenus lors des évaluations objectives. Enfin, en effectuant une évaluation utilisant des énoncés issus d'un corpus différent du corpus CORDIAL, nous avons confirmé que cette tendance se généralise quels que soient les énoncés utilisés.

Toutefois, les évaluations subjectives restent globales et ne permettent pas de déterminer si la dégradation d'un paramètre acoustique est plus influente que la dégradation d'un autre paramètre acoustique, comme par exemple la dégradation du spectre par rapport à la dégradation du F0. Pour obtenir ces informations, il serait nécessaire de réaliser des évaluations subjectives plus poussées.

Conclusion de la troisième partie

Dans cette troisième partie, nous avons exposé et analysé les résultats obtenus par les protocoles présentés dans le chapitre 4. Ces analyses ont été complétées par la présentation des résultats obtenus lors de trois évaluations subjectives.

Dans le premier chapitre de cette partie, nous avons analysé les résultats issus de l'application du protocole d'évaluation basé sur les GMM. Pour chacun des quatre paramètres acoustiques, nous avons comparé l'influence des différents jeux de descripteurs, présentés dans la section 5.4 page 105 du chapitre 5, sur la qualité de modélisation de l'espace acoustique généré par HTS. À l'issue de ces évaluations, nous avons constaté que le jeu de descripteurs `p5-sy_full` pouvait être considéré comme le jeu de descripteurs optimal car ce jeu permet d'obtenir la meilleure modélisation pour chacun des types de coefficients avec un nombre de descripteurs réduits. En effet, ce jeu de descripteurs ne contient que les descripteurs associés aux horizons phonémiques et syllabiques. Ceci permet de n'utiliser qu'une vingtaine de descripteurs contenant les labels phonétiques, les informations de position de la syllabe et du phonème et enfin les informations de proéminences. Les résultats indiquent également que pour les coefficients MGC et la durée, utiliser des descripteurs supplémentaires par rapport au jeu `p3` n'améliore pas la qualité de la synthèse.

Dans le second chapitre de cette partie, nous avons analysé les résultats obtenus suite à l'application du protocole basé sur le calcul de distances entre trames appariées. Pour l'ensemble des paramètres acoustiques analysés, les résultats obtenus à l'horizon de la trame confirme ceux de l'évaluation basée sur la modélisation de l'espace acoustique par GMM. En élargissant l'horizon au phone, nous avons pu analyser précisément l'influence des descripteurs sur la modélisation effectuée par HTS. Nous avons ainsi observé des divergences dans la qualité de modélisation suivant les catégories phonétiques. Enfin, nous avons montré que malgré l'ajout de nouveaux de descripteurs, ces différences de qualité de modélisation ne se résorbaient pas, l'amélioration constatée entre deux jeux de descripteurs étant généralement globale à l'ensemble des modèles.

Dans le dernier chapitre, nous avons présenté les évaluations subjectives effectuées pour valider les résultats des évaluations objectives puis nous avons analysé les résultats obtenus. L'ensemble des évaluations subjectives confirment les résultats des deux protocoles présentés dans ce document. La synthèse effectuée par HTS étant toujours considérée

comme de moins bonne qualité que le signal d'analyse/synthèse, les évaluations subjectives montrent que la modélisation effectuée par le système peut être améliorée. Toutefois, ces évaluations ayant également montré qu'une dégradation due à la paramétrisation du signal était perceptible, pour améliorer la qualité de synthèse il semble également nécessaire d'améliorer la qualité de la paramétrisation du signal.

Conclusion

Conclusion

Les travaux de thèse présentés dans ce document ont pour objectif l'évaluation de l'influence des descripteurs, utilisés pour caractériser le contexte d'un segment acoustique, sur la modélisation effectuée par le système de synthèse HTS dans le cadre du français.

Contributions

Partant de ce constat, nous avons proposé deux méthodes d'évaluation objective (présentées chapitre 4 page 75).

La première méthode consiste à modéliser l'espace acoustique, associé à un paramètre acoustique (MGC, F0, durée ou apériodicité), par un mélange de gaussiennes (GMM). En se basant sur un ensemble de vecteurs de coefficients associés au paramètre acoustique évalué, la nouveauté de la méthode présentée consiste non pas à évaluer la pertinence de cet ensemble, comme c'est le cas classiquement, mais à utiliser cet ensemble de données comme référence et ainsi évaluer la pertinence d'un jeu de modèles différents.

Néanmoins, la limite principale de cette méthode est la masse de données nécessaires à l'apprentissage des GMM. Il est nécessaire de disposer de suffisamment de vecteurs dans le corpus d'apprentissage, pour obtenir un GMM dont le nombre de composantes soit suffisamment élevé pour couvrir l'espace à modéliser. Il est également nécessaire de disposer d'un nombre de trames suffisant dans le corpus de référence afin d'obtenir une vraisemblance moyenne, et donc une estimation fiable de la qualité de l'espace évalué.

Pour compléter cette première approche, une seconde méthode est proposée et repose sur un calcul de distance guidé par l'horizon temporel, la caractéristique et le paramètre acoustique que l'on souhaite évaluer. En calculant une distance entre les vecteurs acoustiques générés par HTS et ceux extraits du signal naturel, il n'est plus nécessaire d'apprendre un modèle statistique pour effectuer l'évaluation. Ce protocole écarte la notion de corpus d'apprentissage et permet de lever la contrainte liée à ce corpus. Néanmoins, la contrainte liée au corpus de test demeure. Il est donc important, lors de l'analyse des résultats obtenus par ce protocole, de tenir compte de la spécificité de ce corpus.

En appliquant ces deux protocoles à une modélisation d'une voix française par HTS, nous avons analysé différentes configurations des jeux de descripteurs et évalué l'influence de ces descripteurs sur la qualité du signal de synthèse. Néanmoins, les résultats de ces protocoles ne montrent pas d'apport significatif des descripteurs autres que l'apport effectué par la prise en compte du contexte phonétique direct du segment. Ces résultats ont d'ailleurs été confirmés par les évaluations subjectives présentées au cours du dernier chapitre de ce document.

Nous avons pu découvrir que la qualité de modélisation n'était pas homogène. En effet, selon les paramètres et selon les caractéristiques phonétiques des segments modélisés, les résultats obtenus varient. Ainsi, nous avons pu mettre en avant que pour le corpus analysé, les consonnes non-voisées apportaient un taux d'erreurs de voisement d'environ sept fois supérieur à celui des voyelles.

Perspectives

Les travaux de thèse présentés dans ce document se focalisent sur l'évaluation de l'influence des descripteurs sur la modélisation. Néanmoins, seuls un corpus et une langue ont été analysés ce qui implique que les conclusion obtenues sont propres à ce corpus. Pour compléter cette étude, le premier point intéressant serait d'appliquer les protocoles décrits dans ce document sur un ensemble plus variés de langues et de corpus afin de pouvoir déterminer si les conclusions sont globales ou spécifiques au corpus utilisé.

De plus, les méthodes proposées ne sont pas spécifiques à l'évaluation de l'influence des descripteurs. En supposant un jeu de descripteurs constant, ces méthodes peuvent donc être utilisées pour effectuer des évaluations préliminaires. Ainsi, l'étude des points suivants peuvent compléter nos travaux afin de pouvoir mieux comprendre le fonctionnement du système HTS :

- Déterminer l'influence des algorithmes utilisés. Jusqu'à présent, les différentes avancées ont été jugées globalement. Quelques études, comme celle présentée par M. Shannon [Shannon2011], s'intéressent à l'influence de l'introduction d'un concept sur la modélisation. En utilisant les deux méthodes proposées, il devrait être possible de pouvoir étendre ces études pour comprendre en quoi le concept permet d'améliorer la modélisation ;
- Déterminer l'impact des locuteurs sur la modélisation effectuée par HTS. Pour cela, afin de ne considérer que le locuteur, il est nécessaire de disposer de corpus parallèles ;
- Déterminer quelles sont les différences entre la modélisation indépendante et la modélisation dépendante du locuteur. D'après une étude proposée par J. Yamagishi ET AL. [Yamagishi2008a], la modélisation indépendante du locuteur est plus robuste que la modélisation dépendante du locuteur. Cela est du au fait que ce type de modélisation repose sur des modèles moyens appris sur une masse importante de

données. La phase d'adaptation consiste à mettre à jour ces modèles, en utilisant un corpus réduit associé à un locuteur cible, pour pouvoir effectuer une synthèse. Néanmoins, il peut être intéressant de comprendre l'influence des modèles moyens par rapport aux données utilisées pour effectuer l'adaptation sur les modèles finaux.

Annexes

Annexe A

HMM

Dans cette annexe nous allons rappeler les équations fondamentales liées au HMM de manière à compléter les notions introduites chapitre 2. Cette annexe s'appuie en majeure partie sur [Rabiner1989], [Cornuejols2011] et [Young2005].

A.1 Forward-Backward

Pour calculer $P(O|\lambda)$, l'algorithme « Forward-Backward » s'appuie sur un principe de programmation dynamique. Cet algorithme utilise deux variables principales : α et β.

La variable $\alpha_t(i)$ permet d'estimer la probabilité que le modèle λ ait émis la sous-séquence d'observations $(o_1, \ldots, o_t)$ et que l'émission de l'observation o_t est assurée par l'état i. Le calcul des probabilités $\alpha_t(i)$, qui correspond donc à la phase « forward », s'effectue par récurrence de la manière suivante :

$$\begin{aligned} \alpha_1(i) &= \pi_i b_i(o_1), & 1 \leq i \leq N \\ \alpha_t(j) &= \left(\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right) b_j(o_t), & 2 \leq t \leq T, 1 \leq j \leq N \end{aligned} \tag{A.1}$$

La variable $\beta_t(i)$ permet d'estimer la probabilité que le modèle λ ait émis la séquence partielle $(o_{t+1}, \ldots, o_T)$ et que l'émission de l'observation o_t soit conditionnée par $q_t = i$. Le calcul de $\beta_t(i)$, qui correspond donc à la phase « backward », s'effectue par récurrence de la manière suivante :

$$\begin{aligned} \beta_T(i) &= 1, & 1 \leq i \leq N \\ \beta_{t-1}(i) &= \sum_{j=1}^{N} \left(a_{ij} b_j(o_t) \beta_t(j) \right), & 1 \leq t \leq T, 1 \leq i \leq N \end{aligned} \tag{A.2}$$

Les valeurs de α et β entrent dans l'estimation de plusieurs probabilités.

L'expression $P(O|\lambda)$ s'écrit :

$$P(O|\lambda) = \sum_{q_T} P(o_1, \cdots, o_T, q_T) \tag{A.3}$$

$$= \sum_{i=1}^{N} \alpha_T(i) \tag{A.4}$$

ou encore :

$$P(O|\lambda) = \sum_{q_1} P(o_1|q1)P(q_1)P(o_2, \cdots, o_T|q_1) \tag{A.5}$$

$$= \sum_{i=1}^{N} \pi_i \beta_0(i) \tag{A.6}$$

La probabilité *a posteriori* d'observation d'une variable cachée q_t étant connues les observations O s'écrit :

$$P(q_t = i|O, \lambda) = \frac{P(o_1, \cdots, o_T, q_t|\lambda)}{P(o_1, \cdots, o_T|\lambda)} \tag{A.7}$$

$$= \frac{P(o_1, \cdots, o_T, q_t|\lambda)}{\sum_{q_t} P(o_1, \cdots, o_T, q_t|\lambda)} \tag{A.8}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(i)\beta_t(j)} \tag{A.9}$$

La probabilité *a posteriori* $P(q_t = i|O, \lambda)$ est souvent notée $\gamma_t(i)$.

A.2 Algorithme de Viterbi

L'algorithme de Viterbi, qui permet d'obtenir la séquence d'états Q maximisant $P(O, Q|\lambda)$, repose sur la définition de deux variables $\delta_t(i)$ et $\psi_t(i)$. $\delta_t(i)$ représente la probabilité que le modèle λ ait émis la séquence d'observation $(o_1, \ldots, o_t)$ en arrivant à l'état i. $\psi_t(i)$ est le chemin optimal pour arriver à l'état i à l'instant t.

L'initialisation de l'algorithme de Viterbi repose sur les deux équations suivantes :

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(o_i), \quad 1 \leq i \leq N \\ \psi_1(i) &= 0 \end{aligned} \tag{A.10}$$

La phase de récurrence consiste à déterminer la probabilité la plus élevée d'arriver

à l'état j à l'instant t en ne connaissant que les probabilités associées à l'instant $t-1$ (variable δ) puis à sauvegarder l'état optimal (variable ψ) :

$$\begin{array}{rcll} \delta_t(j) & = & \max\limits_i \ (\delta_{t-1}(i)a_{ij}) \times b_j(o_t) & 2 \leq t \leq T \\ & & & 1 \leq j \leq N \\ \psi_t(j) & = & \underset{1 \leq i \leq N}{\operatorname{argmax}} (\delta_{t-1}(i)a_{ij}) & 2 \leq t \leq T \\ & & & 1 \leq j \leq N \end{array} \tag{A.11}$$

L'algorithme s'arrête lorsque la dernière observation, o_T, est émise :

$$\begin{array}{rcl} P_T^\star & = & \max\limits_{1 \leq i \leq N} \big(\delta_T(i)\big) \\ Q_T^\star & = & \underset{1 \leq i \leq N}{\operatorname{argmax}} \big(\delta_T(i)\big) \end{array} \tag{A.12}$$

Pour retrouver le chemin optimal, on effectue un « backtrack » rendu possible grâce à la variable ψ qui a permis de mémoriser ce chemin :

$$q_t^\star \ = \ \psi_{t+1}(q_{t+1}^\star) \quad \forall t \in [T-1, \ldots, 1] \tag{A.13}$$

A.3 Algorithme de Baum-Welch

En supposant un jeu de paramètres connu, modèle $\lambda = (\pi, A, B)$, cet algorithme permet d'estimer de nouveaux paramètres, modèle $\overline{\lambda} = (\overline{\pi}, \overline{A}, \overline{B})$. Cet algorithme d'estimation itératif repose sur une estimation EM composé de deux étapes principales :

1. Etape E (Expectation) : inférer la probabilité *a posteriori* $\gamma_t^k(i)$ et *a posteriori* conjointe $\xi_t^k(i,j)$ en utilisant les paramètres du HMM λ et les observations $O^1, \cdots, O^k$, chacune de longueur $T(k)$.
2. Etape M (Maximization) : calculer de nouveaux paramètres, modèle $\overline{\lambda}$, par maximum de vraisemblance.

Ces deux étapes sont itérées jusqu'à convergence. Il y a convergence lorsque la différence entre $P(\mathcal{O}|\overline{\lambda})$ et $P(\mathcal{O}|\lambda)$ est inférieure à un seuil fixé où $\mathcal{O} = (O^1, \ldots, O^K)$ correspond à un ensemble de K séquences d'observations.

A.3.1 Phase E, inférence

Pour alléger l'écriture nous ignorons dans cette sous-section l'indice k qui correspond à l'indice de la séquence utilisée.

La variable $\xi_t(i,j)$ définit la probabilité *a posteriori* d'être à l'état i à l'instant t et à

l'état j à l'instant suivant en sachant que la séquence d'observation O a été observée :

$$\xi_t(i,j) = P(q_{t-1} = i, q_t = j|O, \lambda) \tag{A.14}$$

qui, en utilisant la règle de Bayes, devient :

$$\begin{aligned} \xi_t(i,j) &= \frac{P(O|q_{t-1} = i, q_t = j, \lambda)P(q_{t-1} = i, q_t = j|\lambda)}{P(O|\lambda)} && \text{(A.15)} \\ &= \frac{\alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j)}{\sum_{l=1}^{N}\sum_{m=1}^{N}\alpha_{t-1}(l)a_{lm}b_m(o_t)\beta_t(m)} && \text{(A.16)} \\ & && \text{(A.17)} \end{aligned}$$

$\gamma_t(i)$ qui correspond au posterior $P(q_t = i|O, \lambda)$ peut se rééecrire en fonction de $\xi_t(i,j)$:

$$\begin{aligned} \gamma_t(i) &= P(q_t = i|O, \lambda) && \text{(A.18)} \\ &= \sum_{j=1}^{N}\xi_t(i,j) && \text{(A.19)} \end{aligned}$$

A.3.2 Phase M, estimation

Une fois les variables $\gamma_t^k(i)$ et $\xi_t^k(i,j)$ déterminées, les équations de réestimation suivantes sont utilisées pour calculer les paramètres de $\overline{\lambda}$:

$$\begin{aligned} \overline{\pi}_i &= \sum_{k=1}^{K}\frac{1}{P(O^k|\lambda)} * \gamma_1^k(i) && \text{(A.20)} \\ \overline{a}_{ij} &= \sum_{k=1}^{K}\frac{1}{P(O^k|\lambda)} * \frac{\sum_{t=1}^{T-1}\xi_t^k(i,j)}{\sum_{t=1}^{T-1}\gamma_t^k(i)} && \text{(A.21)} \\ & && \text{(A.22)} \end{aligned}$$

Dans le cadre de la modélisation proposée par HTK, les probabilités d'émission $b(.)$ correspondent à des mixtures de gaussiennes. Nous imposerons néanmoins à HTS l'utilisation d'une seule gaussienne, ce qui réduit l'estimation à deux paramètres : la moyenne $\overline{\mu_i}$ (équation (A.23)), la covariance $\overline{\Sigma_i}$ (équation (A.24)) de la distribution associée à l'état i.

$$\overline{\mu_i} = \sum_{k=1}^{K}\frac{1}{P(O^k|\lambda)} * \frac{\sum_{t=1}^{T-1}\gamma_t^k(i) * o_t^k}{\sum_{t=1}^{T-1}\gamma_t^k(i)} \tag{A.23}$$

$$\overline{\Sigma_i} = \sum_{k=1}^{K}\frac{1}{P(O^k|\lambda)} * \frac{\sum_{t=1}^{T-1}\gamma_t^k(i) * \left[(o_t^k - \overline{\mu_i})(o_t^k - \overline{\mu_i})^\top\right]}{\sum_{t=1}^{T-1}\gamma_t^k(i)} \tag{A.24}$$

Annexe B

Alphabets phonémiques

Consonnes	
IPA	**Liaphon**
p	pp
t	tt
k	kk
b	bb
d	dd
g	gg
m	mm
n	nn
ɲ	gn
ŋ	ng
f	ff
s	ss
z	zz
v	vv
ʃ	ch
ʒ	jj
l	ll
ʀ	rr

Voyelles	
IPA	**Liaphon**
i	ii
y	uu
u	ou
e	ei
ø	eu
o	au
ɛ	ai
ə	ee
œ	oe
ɔ	oo
a	aa
ɛ̃	in
õ	on
ã	an
ũ	un

Semi-voyelles	
IPA	**Liaphon**
w	ww
ɥ	uy
j	yy

Annexe C

Jeux de descripteurs

C.1 Jeu de descripteurs standard

C.1.1 Topologie de labellisation d'un segment acoustique

```
p1^p2-p3+p4=p5@p6_p7
/A:a1_a2_a3/B:b1-b2-b3@b4-b5&b6-b7#b8-b9$b10-b11!b12-b13;b14-b15|b16/C:c1+c2+c3
/D:d1_d2/E:e1+e2@e3+e4&e5+e6#e7+e8/F:f1_f2
/G:g1_g2/H:h1=h2@h3=h4|h5/I:i1_i2
/J:j1+j2-j3
```

C.1.2 Présentation des descripteurs

Idx	Id	Description
1	p1	phonème précédent-précédent
2	p2	phonème précédent
3	p3	phonème courant
4	p4	phonème suivant
5	p5	phonème suivant-suivant
6	p6	position du phonème courant dans la syllabe courante (du début)
7	p7	position du phonème courant dans la syllabe courante (de la fin)
8	a1	est ce que la syllabe précédente est ≪ stressed ≫ ?
9	a2	est ce que la syllabe précédente est ≪ accented ≫ ?
10	a3	nombre de phonèmes dans la syllabe précédente

Idx	Id	Description
11	b1	est ce que la syllabe courante est ≪ stressed ≫ ?
12	b2	est ce que la syllabe courante est ≪ accented ≫ ?
13	b3	nombre de phonèmes dans la syllabe courante
14	b4	position de la syllabe courante dans le mot courant (du début)
15	b5	position de la syllabe courante dans le mot courant (de la fin)
16	b6	position de la syllabe courante dans la phrase courante (du début)
17	b7	position de la syllabe courante dans la phrase courante (de la fin)
18	b8	nombre de syllabes ≪ stressed ≫ avant la syllabe courante dans la phrase courante
19	b9	nombre de syllabes ≪ stressed ≫ après la syllabe courante dans la phrase courante
20	b10	nombre de syllabes ≪ accented ≫ avant la syllabe courante dans la phrase courante
21	b11	nombre de syllabes ≪ accented ≫ après la syllabe courante dans la phrase courante
22	b12	nombre de syllabes à partir de la dernière syllabe ≪ stressed ≫ jusque la syllabe courante
23	b13	nombre de syllabes à partir de la syllabe courante jusque la prochaine syllabe ≪ stressed ≫
24	b14	nombre de syllabes à partir de la dernière syllabe ≪ accented ≫ jusque la syllabe courante
25	b15	nombre de syllabes à partir de la syllabe courante jusque la prochaine syllabe ≪ accented ≫
26	b16	label de la voyelle de la syllabe courante
27	c1	est ce que la syllabe suivante est ≪ stressed ≫ ?
28	c2	est ce que la syllabe suivante est ≪ accented ≫ ?
29	c3	nombre de phonèmes dans la syllabe suivante
30	d1	tag grammatical estimé du mot précédent
31	d2	nombre de syllabes dans le mot précédent
31	e1	tag grammatical estimé du mot courant
32	e2	nombre de syllabes dans le mot courant
33	e3	position du mot courant dans la phrase courante (du début)
34	e4	position du mot courant dans la phrase courante (de la fin)
35	e5	nombre de mots signifiants avant le mot courant dans la phrase courante
36	e6	nombre de mots signifiants après le mot courant dans la phrase courante
37	e7	nombre de mots à partir du dernier mot ”accentuable” jusqu'au mot courant
38	e8	nombre de mots à partir du mot courant jusqu'au prochain mot ”accentuable”
39	f1	tag grammatical estimé du mot suivant
40	f2	nombre de syllabes dans le mot suivant
41	g1	nombre de syllabes dans la phrase précédente
42	g2	nombre de mots dans la phrase précédente
43	h1	nombre de syllabes dans la phrase courante
44	h2	nombre de mots dans la phrase courante
45	h2	position de la phrase dans l'énoncé (du début)
46	h2	position de la phrase dans l'énoncé (de la fin)
47	h2	Tag ToBi de fin de phrase
48	i1	nombre de syllabes dans la phrase suivante
49	i2	nombre de mots dans la phrase suivante
50	j1	nombre de syllabes dans l'énoncé
51	j2	nombre de mots dans l'énoncé
52	j3	nombre de phrases dans l'énoncé

C.2 Jeu de descripteurs proposé

C.2.1 Format de label

```
p1^p2-p3+p4=p5@p6_p7
/A:a1_a2_a3/B:b1-b2-b3@b4-b5&b6-b7#b8-b9$b10-b11!b12-b13;b14-b15|b16/C:c1+c2+c3
/D:d1_d2/E:e1+e2@e3+e4&e5+e6#e7+e8/F:f1_f2
/G:g1_g2/H:h1=h2@h3=h4|h5/I:i1_i2
/J:j1+j2-j3/Z:z1
```

C.2.2 Présentation des descripteurs

Idx	Id	Description
1	p1	phonème précédent-précédent
2	p2	phonème précédent
3	p3	phonème courant
4	p4	phonème suivant
5	p5	phonème suivant-suivant
6	p6	position du phonème courant dans la syllabe courante (du début)
7	p7	position du phonème courant dans la syllabe courante (de la fin)
8	a1	
9	a2	est ce que la syllabe précédente est proéminente ?
10	a3	nombre de phonèmes dans la syllabe précédente
11	b1	
12	b2	est ce que la syllabe courante est proéminente ?
13	b3	nombre de phonèmes dans la syllabe courante
14	b4	position de la syllabe courante dans le mot courant (du début)
15	b5	position de la syllabe courante dans le mot courant (de la fin)
16	b6	position de la syllabe courante dans le syntagme courant (du début)
17	b7	position de la syllabe courante dans le syntagme courant (de la fin)
18	b8	
19	b9	
20	b10	nombre de syllabes proéminentes avant la syllabe courante dans le syntagme courant
21	b11	nombre de syllabes proéminentes après la syllabe courante dans le syntagme courant
22	b12	
23	b13	
24	b14	nombre de syllabes à partir de la dernière syllabe proéminente jusque la syllabe courante
25	b15	nombre de syllabes à partir de la syllabe courante jusque la prochaine syllabe proéminente
26	b16	label de la voyelle de la syllabe courante
27	c1	
28	c2	est ce que la syllabe suivante est proéminente ?
29	c3	nombre de phonèmes dans la syllabe suivante

Idx	Id	Description
30	d1	tag grammatical estimé du mot précédent
31	d2	nombre de syllabes dans le mot précédent
31	e1	tag grammatical estimé du mot courant
32	e2	nombre de syllabes dans le mot courant
33	e3	position du mot courant dans le syntagme courant (du début)
34	e4	position du mot courant dans le syntagme courant (de la fin)
35	e5	nombre de mots signifiants avant le mot courant dans le syntagme courant
36	e6	nombre de mots signifiants après le mot courant dans le syntagme courant
37	e7	nombre de mots à partir du dernier mot signifiant jusqu'au mot courant
38	e8	nombre de mots à partir du mot courant jusqu'au prochain mot signifiant
39	f1	tag grammatical estimé du mot suivant
40	f2	nombre de syllabes dans le mot suivant
41	g1	nombre de syllabes dans le syntagme précédent
42	g2	nombre de mots dans le syntagme précédent
43	h1	nombre de syllabes dans le syntagme courant
44	h2	nombre de mots dans le syntagme courant
45	h2	position du syntagme dans l'énoncé (du début)
46	h2	position du syntagme dans l'énoncé (de la fin)
47	h2	
48	i1	nombre de syllabes dans le syntagme suivant
49	i2	nombre de mots dans le syntagme suivant
50	j1	nombre de syllabes dans l'énoncé
51	j2	nombre de mots dans l'énoncé
52	j3	nombre de syntagmes dans l'énoncé
53	z1	Est ce que le segment est une voyelle voisée, non voisée ou n'est pas une voyelle

C.3 Comparaison des jeux de descripteurs

Langue	Phonème	Syllabe	Mots	Phrase	Énoncé	Niveau supplémentaire
Standard [Tokuda2002, Zen2009]	Label 5-phone Position dans syl.	Taille en ph. {P,C,S} Stressed {P,C,S} Acccented {P,C,S} Position%stressed Position%accented Dist {P,S} stressed Dist {P,S} accented Position dans mot Position dans phrase Voyelle	Et. gram. {P,C,S} Taille en syl. {P,C,S} Position%content Dist {P,S}%content Position dans phrase	Taille en syl. {P,C,S} Taille en mots {P,C,S} Position dans énoncé Et. TOBI	Taille en syl. Taille en mots Taille en phrases	
Japonais [Oura2011a]		Syllabe ⇒ More - Stressed {P,C,S} - Acccented {P,C,S} - Position%stressed - Position%accented - Dist {P,S} stressed - Dist {P,S} accented - Position dans phrase - Voyelle	- Taille en syl. {P,C,S} - Position%content - Dist {P,S}%content + Forme inf. {P,C,S} + Type conj. {P,C,S}	Phrase ⇒ accent phrase - Et. TOBI + Type accentuation {P,C,S} + Interrogative ? {P,C,S} + Pause entre la phr. P et la C. ? + Pause entre la phr. C et la S ? + Position (en phrase) + Position (en mora)	- Taille en mots + Taille en GS	+ GS + Taille en phrases (par phrase) {P,C,S} + Taille en phrases (par more) {P,C,S}
Portugais brésilien [Maia2003]		- Acccented {P,C,S} - Position%accented - Dist {P,S} accented	+ Interrogation ?	- Et. TOBI		
Thaïlandais [Chomphan2007]	5-Phone ⇒ 3-phone	- Stressed {P,C,S} - Acccented {P,C,S} - Position%stressed - Position%accented - Dist {P,S} stressed - Dist {P,S} accented - Position dans phrase - Voyelle	- Position%content - Dist {P,S}%content	- Et. TOBI		
Suédois [Lundgren2005]		- Stressed {P,C,S} - Position%stressed - Dist {P,S} stressed - Voyelle	- Et. gram. {P,C,S} - Position%content - Dist {P,S}%content	- Et. TOBI		
Finnois [Silen2008]	5-Phone ⇒ 3-phone	- Stressed {P,C,S} - Acccented {P,C,S} - Position%stressed - Position%accented - Dist {P,S} stressed - Dist {P,S} accented	- Et. gram. {P,C,S} - Position%content - Dist {P,S}%content	- Et. TOBI		

Langue	Phonème	Syllabe	Mots	Phrase	Énoncé	Niveau supplémentaire
Standard [Tokuda2002, Zen2009]	Label 5-phone Position dans syl.	Taille en ph. {P,C,S} Stressed {P,C,S} Acccented {P,C,S} Position%stressed Position%accented Dist {P,S} stressed Dist {P,S} accented Position dans mot Position dans phrase Voyelle	Et. gram. {P,C,S} Taille en syl. {P,C,S} Position%content Dist {P,S}%content Position dans phrase	Taille en syl. {P,C,S} Taille en mots {P,C,S} Position dans énoncé Et. TOBI	Taille en syl. Taille en mots Taille en phrases	
Mandarin [Qian2006]	5-Phone ⇒ 3-phone	- Stressed {P,C,S} - Acccented {P,C,S} - Position%stressed - Position%accented - Dist {P,S} stressed - Dist {P,S} accented - Voyelle + Label du ton {P,C,S}	- Et. gram. {P,C,S} - Taille en syl. {P,C,S} - Position%content - Dist {P,S}%content - Position dans phrase + Indices de coupure	Phrase ⇒ phrase ≪ GS ≫ - Taille en mots {P,C,S} - Et. TOBI		
Espagnol [Bonafonte2008]	- Position dans syl.	- Taille en ph. {P,C,S} - Acccented {P,C,S} - Position%accented - Dist {P,S} accented + Première/dernière syllabe ?	- Position%content - Dist {P,S}%content + Premier/dernier mot ?	- Et. TOBI		+ Groupe accentuel + taille en ph. {P,C,S} + premier/dernier groupe ? + Type d'accent + Groupe intonatif + premier/dernier groupe ? + Type d'intonation
Allemand [Krstulovic2007]	+ Constituante syl. ?	+ Taille en seg. + is break ? {P,C}	+ Taille en seg. + Fréquence unigram. (éch. log.)	+ Ét. TOBI accent {P,S,S-S} + Ét. TOBI endtone {P,S,S-S}	+ Taille en ponctuations	+ Échelle ponctuation + type ponct. {P,C,S} + dist.%{P,S}ponc en mots + Échelle segment + position dans syllabe + position dans mot
Portugais européen [Barros2005]		- Acccented {P,C,S} - Position%accented - Dist {P,S} accented	- Et. gram. {P,C,S} - Position%content - Dist {P,S}%content + Interrogation ?	- Et. TOBI		

Langue	Phonème	Syllabe	Mots	Phrase	Énoncé	Niveau supplémentaire
Standard [Tokuda2002, Zen2009]	Label 5-phone Position dans syl.	Taille en ph. {P,C,S} Stressed {P,C,S} Acccented {P,C,S} Position%stressed Position%accented Dist {P,S} stressed Dist {P,S} accented Position dans mot Position dans phrase Voyelle	Et. gram. {P,C,S} Taille en syl. {P,C,S} Position%content Dist {P,S}%content Position dans phrase	Taille en syl. {P,C,S} Taille en mots {P,C,S} Position dans énoncé Et. TOBI	Taille en syl. Taille en mots Taille en phrases	
Grec [Karabetsos2008]		- Acccented {P,C,S} - Position%accented - Dist {P,S} accented - Voyelle + Nombre de voyelles	- Et. gram. {P,C,S{ - Position%content - Dist {P,S}%content	- Et. TOBI		
Tchèque [Hanzlivcek2010]	5-Phone ⇒ 3-phone - Position dans syl.	- Taille en ph. {P,C,S} - Stressed {P,C,S} - Acccented {P,C,S} - Position%stressed - Position%accented - Dist {P,S} stressed - Dist {P,S} accented - Position dans phrase - Voyelle	- Et. gram. {P,C,S} - Taille en syl. {P,C,S} - Position%content - Dist {P,S}%content	- Position dans énoncé		
Croate [Ipsic2006]	5-Phone ⇒ 3-phone - Position dans syl.					
Basque [Erro2010]	+ Dist {P,S} pause	+ Dist {P,S} pause	+ Dist {P,S} pause	Phrase ⇒ groupe d'accentuation + Dist {P,S} pause - Et. TOBI	+ Taille en phones + Taille en pauses + Émotion + Type de l'énoncé	+ Échelle pause + Type de pause {P,S} + Position%pause

Bibliographie

[Barbot2011] N. Barbot, V. Barreaud, O. Boëffard, L. Charonnat, A. Delhay, S. Le Maguer, D. Lolive, et al. Towards a versatile multi-layered description of speech corpora using algebraic relations. In *Conference of the International Speech Communication Association (Interspeech)*, pages 1501–1504, 2011.

[Barros2005] M.J. Barros, R. Maia, K. Tokuda, F. Resende, and D. Freitas. Hmm-based european portuguese tts system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2581–2584, 2005.

[Bartkova1987] K. Bartkova and C. Sorin. Predictive model of segmental duration in french. In *109th ASA Meeting*, 1985.

[Baum1967] Leonard E Baum and JA Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3) :360–363, 1967.

[Baum1970] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1) :pp. 164–171, 1970.

[Bird2000] S. Bird, D. Day, J. Garofolo, J. Henderson, C. Laprun, and M. Liberman. ATLAS : A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1699–1706, 2000.

[Bird2001] Steven Bird and Mark Liberman. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2) :23–60, 2001.

[Black1994] A.W. Black and P. Taylor. Chatr : a generic speech synthesis system. In *Proceedings of the conference on Computational linguistics*, pages 983–986. Association for Computational Linguistics, 1994.

[Black2002] Alan Black, Paul Taylor, and Richard Caley. The Festival Speech Synthesis System - System documentation. http://festvox.org/docs/manual-1.4.3/festival_toc.html.

[Black2005] A.W. Black and K. Tokuda. The blizzard challenge-2005 : Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 2005.

[Boeffard2012] Olivier Boeffard, Laure Charonnat, Sébastien Le Maguer, and Damien Lolive. Towards fully automatic annotation of audio books for tts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[Boite2000] R Boite. *Traitement de la parole*. Presses polytechniques et universitaires romandes, Lausanne, 2000.

[Bonafonte2008] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez. Corpus and voices for catalan speech synthesis. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3325–3329, 2008.

[Braunschweiler2010] N. Braunschweiler, M.J.F. Gales, and S. Buchholz. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Proc. of Interspeech*, pages 2222–2225, 2010.

[Calliope1989] Calliope. *La Parole et son traitement automatique*. Masson, Paris {;;Milan} {;;Barcelone}, 1989.

[Cassidy1996] Steve Cassidy and Jonathan Harrington. Emu : An enhanced hierarchical speech data management system. *Proc. of the 6th Australian Int. Speech Science and Technology Conf.*, pages 361–366, 1996.

[Cassidy2001] S Cassidy. Multi-level annotation in the Emu speech database management system. *Speech Communication*, 33(1-2) :61–77, January 2001.

[Charpentier1989] Francis Charpentier and Eric Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2013–2019, 1989.

[Chen2010a] Yi-ning Chen, Zhi-jie Yan, and Frank K Soong. A Perceptual Study of Acceleration Parameters in HMM-Based TTS. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, number September, pages 426–429, 2010.

[Chomphan2007] S. Chomphan and T. Kobayashi. Design of tree-based context clustering for an hmm-based thai speech synthesis system. In *Proceedings of the Speech Synthesis Workshop (SSW)*, volume 22, pages 160–165, 2007.

[Cornuejols2011] Antoine Cornuéjols and Laurent Miclet. *Apprentissage artificiel*. Eyrolles, 2011.

[Delattre1966] Pierre Delattre. No Title. *The french review*, 40(1) :1–14, 1966.

[Dicristo2000] Albert Di Cristo. Interpréter la prosodie. In *XXIIèmes Journées d'Études sur la Parole*, Aussois, France, 2000.

[Dominguez1997] A. Domínguez and M.Y.C. DE VEGA. Lexical inhibition from syllabic units in visual word recognition. *Language and Cognitive Processes*, 12 :401–422, 1997.

[Donovan1995] R.E. Donovan and P.C. Woodland. Automatic speech synthesiser parameter estimation using hmms. In *Proceedings of the international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 640 –643, may 1995.

[Donovan1996] Robert Donovan. *Trainable speech synthesis*. PhD thesis, Cambridge University, 1996.

[Donovan1999] RE Donovan, M. Franz, JS Sorensen, and S. Roukos. Phrase splicing and variable substitution using the ibm trainable speech synthesis system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, 1999.

[Dudley1939] Homer Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11 :169–177, 1939.

[Dutoit1996] Thierry Dutoit, Vincent Pagel, Nicolas Pierret, François Bataille, and Olivier Van der Vrecken. The mbrola project : Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1393–1396. IEEE, 1996.

[Erro2010] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez. Hmm-based speech synthesis in basque language using hts. *Proceedings of the FALA*, 2010.

[Fant70] Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[Fonagy2003] Iván Fónagy. Des fonctions de l'intonation : essai de synthèse. *Flambeau*, 29 :1–20, 2003.

[Francois2001] Hélène François. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. PhD thesis, Université de Rennes 1, 2002.

[Fukada1992] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptative Algorithm for mel-cepstral analysis of speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140, 1992.

[Hanzlivcek2010] Z. Hanzlíček. Czech hmm-based speech synthesis. In *Proceedings of the Text, Speech and Dialogue Conference (TSD)*, pages 291–298. Springer, 2010.

[Hirai2004] Toshio Hirai and Seiichi Tenpaku. Using 5 ms segments in concatenative speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[Hts211] Hts 2.1.1. http://hts.sp.nitech.ac.jp/?Release%20Archive#haba8935.

[Hunt1996] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 373–376, 1996.

[ITU-T1996] ITU-T. P800 : Methods for objective and subjective assessment of quality. Technical report, 1996.

[Imai1983] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (mlsa) filter for speech synthesis. *Electronics and Communications in Japan (Part I : Communications)*, 66(2) :10–18, 1983.

[Imai1988] S. Imai. Unbiased estimator of log spectrum and its application to speech signal processing. *Proceedings of EURASIP*, 1988.

[Ipsic2006] I. Ipsic and S. Martincic-Ipsic. Croatian hmm-based speech synthesis. *Journal of Computing and Information Technology*, 14(4) :307–313, 2006.

[Kamina2006] Pierre Kamina. *Carnet d'anatomie, tome 2 : tête, cou, dos*. Paris : Maloine, 2006.

[Karabetsos2008] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis. Hmm-based speech synthesis for the greek language. In *Proceedings of the Text, Speech and Dialogue Conference (TSD)*, pages 349–356. Springer, 2008.

[Kawahara1999] Hideki Kawahara, Ikuyo Masuda-katsuse, and Alain De Cheveign. Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency- based F0 extraction : Possible role of a repetitive structure in sounds 1. *Speech Communication*, 27 :187–207, 1999.

[Kawahara2001] Hideki Kawahara, Jo Estill, and Osamu Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proceedings of the Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[King2012] Simon King and Vasilis Karaiskos. The blizzard challenge 2012.

[Kishore2003] SP Kishore and A.W. Black. Unit size in unit selection speech synthesis. In *Proceedings of EUROSPEECH*, volume 2003, pages 1317–1320, 2003.

[Kominek2003] John Kominek and Alan W Black. CMU ARCTIC 0.95. Technical report, 2003.

[Krstulovic2007] S. Krstulovic, A. Hunecke, and M. Schröder. An hmm-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 7. Citeseer, 2007.

[Kurematsu1990] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. {ATR} japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9(4) :357 – 363, 1990.

[Latorre2011] J. Latorre, M.J.F. Gales, S. Buchholz, K. Knill, M. Tamurd, Y. Ohtani, and M. Akamine. Continuous f0 in the source-excitation generation for hmm-based tts : Do we need voiced/unvoiced classification ? In *Proceedings of the international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4724–4727. IEEE, 2011.

[Leon1992] Léon Pierre. *Phonétisme et prononciations du français*. Nathan-Université, 1992.

[Lundgren2005] A. Lundgren. An hmm-based text-to-speech system applied to swedish. Master's thesis, Royal Institute of Technology (KTH), 2005.

[Maia2003] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. Resende Jr. Towards the development of a brazilian portuguese text-to-speech system based on hmm. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2465–2468. Citeseer, 2003.

[Malfrere1998] F. Malfrère, T. Dutoit, and P. Mertens. Fully automatic prosody generator for text-to-speech. In *Proceedings of the International Conference on Spoken Language Processing*, volume 98, 1998.

[Masuko1996] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Imai Satoshi. Speech synthesis using HMMs with dynamic features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 389–392, Atlanta, Georgia, USA, 1996.

[Mertens1992] Piet Mertens. L'accentuation des syllabes contiguës. *ITL. Review of Applied Linguistics*, 95(95-96) :145–165, 1992.

[Mertens1993] P Mertens. Accentuation, intonation et morphosyntaxe. Technical Report 26, 1993.

[Mertens2001] Piet Mertens, Jean-Philippe Goldman, and Arnaud Gaudinat. La synthèse de l'intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues (TAL)*, 42(1) :145–192, 2001.

[Moreno2009] P.J. Moreno and C. Alberti. A factor automaton approach for the forced alignment of long speech recordings. In *Proc. of IEEE ICASSP*, pages 4869–4872, 2009.

[Moulines1990] E. Moulines, F. Emerard, D. Larreur, JL Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, and C. Sorin. A real-time french text-to-speech system generating high-quality synthetic speech. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 309–312. IEEE, 1990.

[Odell1995] Julian James Odell. *The use of context in large vocabulary speech recognition*. PhD thesis, Citeseer, 1995.

[Oura2010] Keiichiro Oura. List of modifications made in HTS (for version 2.1.1). pages 1–13, 2010.

[Oura2011] Keiichiro Oura. List of modifications made in HTS (for version 2.2), 2011.

[Oura2011a] Keiichiro Oura. An example of context-dependent label format for hmm-based speech synthesis in japanese.

[Pierrehumbert1990] J. Pierrehumbert. The meaning of intonational contours in the interpretation of discourse janet pierrehumbert and julia hirschberg. *Intentions in communication*, page 271, 1990.

[Prudon2002] R. Prudon, C. d'Alessandro, and P.B. de Mareuil. Prosody synthesis by unit selection and transplantation on diphones. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 119–122. IEEE, 2002.

[Qian2006] Y. Qian, F. Soong, Y. Chen, and M. Chu. An hmm-based mandarin chinese text-to-speech system. *Chinese Spoken Language Processing*, pages 223–232, 2006.

[Rabiner1989] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[Reynolds1995] Douglas A Reynolds. Speaker identification and verification using Gaussian mixture speaker models. 17 :91–108, 1995.

[Robel05] A. Röbel. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proceedings of the 8th International Conference on Digital Audio Effects*, pages 30–35, 2005.

[Rojc2007] Matej Rojc and Zdravko Kačič. Time and space-efficient architecture for a corpus-based text-to-speech synthesis system. *Speech Communication*, 49(3) :230–249, 2007.

[Rothenberg2008] Martin Rothenberg. The source-filter model lives (if you are careful). In *Voice Foundation 37th Annual Symposium*, 2008.

[Russell1985] Martin Russell and Roger Moore. Explicit modelling of state occupancy in Hidden Markov Models for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5–8, 1985.

[Sagisaka1988] Yoshinori Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 679–682, 1988.

[Shannon2011] Matt Shannon, Heiga Zen, and William Byrne. The Effect of Using Normalized Models in Statistical Speech Synthesis. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pages 121–124, 2011.

[Shinoda2000] Koichi Shinoda and Takao Wanabe. MDL-based context-dependent subword modeling for speech recognition. *Acoustical Science and Technology (AST)*, 21(2) :79–86, 2000.

[Silen2008] H. Silen, E. Helander, J. Nurminen, and M. Gabbouj. Evaluation of finnish unit selection and hmm-based speech synthesis. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, 2008.

[Silen2010] Hanna Silén, Elina Helander, Jani Nurminen, and Moncef Gabbouj. Analysis of Duration Prediction Accuracy in HMM-Based Speech Synthesis. In *Proceedings of speech prosody*, 2010.

[Silen2011] Hanna Silen, Elina Helander, and Moncef Gabbouj. Prediction of Voice Aperiodicity Based on Spectral Representations in HMM Speech Synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 105–108, 2011.

[Silverman1992] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi : A standard for labeling english prosody. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 867–870, 1992.

[Simon2008] A.C. Simon, M. Avanzi, J.P. Goldman, et al. La détection des proéminences syllabiques. un aller-retour entre l'annotation manuelle et le traitement automatique. In *Proceedings of Congrès Mondial de Linguistique Française*, pages 1673–1686, 2008.

[Sorin1984] C. Sorin, M. Stella, and A. Aggoun. Règles prosodiques et synthèse de parole "multi-style". In *Symposium Franco-Soviétique sur le Dialogue Home-Machine*, 1984.

[Sptk] Sptk 3.5. http://sp-tk.sourceforge.net/.

[Stevens1937] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3) :185–190, 1937.

[Stylianou1998] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2) :131–142, mar 1998.

[Tao2010] Y. Tao, L. Xueqing, and W. Bian. A dynamic alignment algorithm for imperfect speech and transcript. *Computer Science and Information Systems*, 7(1) :75–84, 2010.

[Taylor1998] Paul Taylor, Alan W Black, and Richard Caley. The architecture of the Festival speech synthesis system. In *Proceedings of the ISCA Speech Synthesis Workshop (SSW)*, 1998.

[Taylor2000] P. Taylor. Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, 107 :1697, 2000.

[Taylor2009] Paul Taylor. *Text-to-speech synthesis*. Cambridge University Press, Cambridge {UK} {;;New} York, 2009.

[Toda2005] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Spectral Conversion Based on Maximum Likelehood Estimation Considering Global Variance of Converted Parameter. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 9–12. ICASSP, 2005.

[Toda2005a] Tomoki Toda and Keiichi Tokuda. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pages 2801–2804, 2005.

[Tokuda1995] Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, and Satoshi Imai. An algorithm for speech parameter generation from continuous mixture hmms with dynamic features. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 1995.

[Tokuda1995a] Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech parameter generation from HMM using dynamic features. In *Proceedings of the International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 660–663, 1995.

[Tokuda1999] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 229–232, 1999.

[Tokuda2000] Keiichi Tokuda, Heiga Zen, and Alan W Black. An hmm-based speech synthesis system applied to english. In *Proceedings of the Speech Synthesis Workshop (SSW)*, pages 2–5, 2002.

[Tokuda2000a] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Multi-Space Probability Distribution HMM. *IEICE Transactions on Information and Systems*, E85-D(3) :455–464, 2000.

[Tokuda2000b] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *Proceedings of the International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 1315–1318, 2000.

[Tokuda2002] Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-based speech synthesis system applied to English. In *IEEE Workshop 2002*, 2002.

[Veaux2008] C. Veaux, G. Beller, D. Schwarz, and X. Rodet. Ircamcorpustools : an extensible plateform for speech corpora exploitation, à paraître dans. In *Proceedings of the Language Resources and Evaluation Conference*, 2008.

[Viterbi1967] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2) :260–269, 1967.

[Watts2010] Oliver Watts, Junichi Yamagishi, and Simon King. The Role of Higher-Level Linguistic Features in HMM-Based Speech Synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 841–844, 2010.

[Wu2006] Yi-jian Wu and Ren-hua Wang. Minimum generation error training for hmm-based speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 89–92, 2006.

[Wu2008] Yi-jian Wu and Keiichi Tokuda. Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2008.

[Yamagishi2005] Junichi Yamagishi and Takao Kobayashi. Adaptative training for hidden semi-Markov model. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 365–368, Philadelphia, USA, 2005.

[Yamagishi2006a] Junichi Yamagishi. An introduction to hmm-based speech synthesis. Technical report, Tokyo Institute of Technology, 2006.

[Yamagishi2007] Junichi Yamagishi, Kobayashi Takao, Steve Renals, Simon King, Heiga Zen, Tomoki Toda, and Keiichi Tokuda. Improved Average-Voice-based Speech Synthesis Using Gender-Mixed Modeling and a Parameter Generation Algorithm Considering GV. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pages 125–130, 2007.

[Yamagishi2007a] Junichi Yamagishi, Heiga Zen, Tomoki Toda, and Keiichi Tokuda. Speaker-Independent HMM-based Speech Synthesis System - HTS-2007 System for the Blizzard Challenge 2007. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 2007.

[Yamagishi2008] Junichi Yamagishi, Heiga Zen, Yi-Jian Wu, Tomoki Toda, and Keiichi Tokuda. The HTS-2008 System : Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge. In *Blizzard Challenge 2008*, 2008.

[Yamagishi2008a] J. Yamagishi, Z. Ling, and S. King. Robustness of hmm-based speech synthesis. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, volume 8, pages 581–584, 2008.

[Yokomizo2010] Shuji Yokomizo, Takashi Nose, and Takao Kobayashi. Evaluation of Prosodic Contextual Factors for HMM-Based Speech Synthesis. In *proceedings of Interspeech*, pages 430–433, 2010.

[Yoshimura1998] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Duration Modeling For HMM-Based Speech Synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 29–32, 1998.

[Yoshimura1999] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of the European Conference on Speech Communicationand Technology (Eurospeech)*, volume 5, pages 2347–2350, Budapest, Hungary, 1999.

[Yoshimura2001] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Mixed Excitation for HMM-based Speech Synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2263–2266, 2001.

[Young1993] S.J. Young and S. Young. The htk hidden markov model toolkit : Design and philosophy. *Department of Engineering, Cambridge University, UK, Tech. Rep. TR*, 153, 1993.

[Young1994] Steve J Young, Julian J Odell, and Phil C Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology (HLT)*, pages 307–312, Morristown, New Jersey, USA, 1994. Association for Computational Linguistics.

[Young2005] Steve Young, Gunnar Everman, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book*. Number July 2000. 2005.

[Zellner1998] Brigitte Zellner. *Caractérisation et prédiction du débit de parole en français*. PhD thesis, Université de Lausanne, 1998.

[Zen2004] Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Hidden semi-markov model based speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1397–1400, 2004.

[Zen2005] Heiga Zen and Tomoki Toda. An overview of Nitech HMM-based speech synthesis system for blizzard challenge 2005. In *Proceedings of the 9th European Conference on Speech Communicationand Technology (Eurospeech)*, Lisbon, Portugal, 2005.

[Zen2006] Heiga Zen, Tomoki Toda, and Keiichi Tokuda. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard challenge 2006. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*. Nitech, 2006.

[Zen2007a] Heiga Zen, Keiichi Tokuda, and Tadashi Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, 21(1) :153–173, 2007.

[Zen2009] H. Zen, K. Tokuda, and A.W. Black. Review : Statistical parametric speech synthesis. *Speech Communication*, 51(11) :1039–1064, 2009.

[boeffard2012a] Olivier Boëffard, Laure Charonnat, Sébastien Le Maguer, Damien Lolive, and Gaëlle Vidal. Vers une annotation automatique de corpus audio pour la synthèse de parole. In *Proceedings of the Joint Conference JEP-TALN-RECITAL*, pages 731–738, Grenoble, France, June 2012. ATALA/AFCP.

[combescure1980] P. Combescure. Phrases phonétiquement équilibrées. *CNET Lannion, Recherches acoustiques*, 6 :45–62, 1980.

[harris1978] Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1) :51–83, 1978.

[prahallad2007] K. Prahallad, A.R. Toth, and A.W. Black. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases. In *Proc. of Interspeech*, pages 2901–2904, 2007.

[sorin1987] C. Sorin, D. Larreur, and R. Llorca. A rhythm-based prosodic parser for text-to-speech systems in french. *XIème Congrès International des Sciences Phonétiques*, pages 125–128, 1987.

[yu2011] K. Yu and S. Young. Continuous f0 modeling for hmm based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5) :1071–1079, 2011.

Résumé

Les travaux présentés dans cette thèse se situent dans le cadre de la synthèse de la parole à partir du texte et, plus précisément, dans le cadre de la synthèse paramétrique utilisant des règles statistiques. Nous nous intéressons à l'influence des descripteurs linguistiques utilisés pour caractériser un signal de parole sur la modélisation effectuée dans le système de synthèse statistique HTS. Pour cela, deux méthodologies d'évaluation objective sont présentées. La première repose sur une modélisation de l'espace acoustique, généré par HTS par des mélanges gaussiens (GMM). En utilisant ensuite un ensemble de signaux de parole de référence, il est possible de comparer les GMM entre eux et ainsi les espaces acoustiques générés par les différentes configurations de HTS. La seconde méthodologie proposée repose sur le calcul de distances entre trames acoustiques appariées pour pouvoir évaluer la modélisation effectuée par HTS de manière plus locale. Cette seconde méthodologie permet de compléter les diverses analyses en contrôlant notamment les ensembles de données générées et évaluées. Les résultats obtenus selon ces deux méthodologies, et confirmés par des évaluations subjectives, indiquent que l'utilisation d'un ensemble complexe de descripteurs linguistiques n'aboutit pas nécessairement à une meilleure modélisation et peut s'avérer contre-productif sur la qualité du signal de synthèse produit.

Mots-clefs : Informatique, Traitement automatique de la parole, Synthèse de la parole à partir du texte, HTS

Abstract

The work presented in this thesis is about TTS speech synthesis and, more particularly, about statistical speech synthesis for French. We present an analysis on the impact of the linguistic contextual factors on the synthesis achieved by the HTS statistical speech synthesis system. To conduct the experiments, two objective evaluation protocols are proposed. The first one uses Gaussian mixture models (GMM) to represent the acoustical space produced by HTS according to a contextual feature set. By using a constant reference set of natural speech stimuli, GMM can be compared between themselves and consequently acoustic spaces generated by HTS. The second objective evaluation that we propose is based on pairwise distances between natural speech and synthetic speech generated by HTS. Results obtained by both protocols, and confirmed by subjective evaluations, show that using a large set of contextual factors does not necessarily improve the modeling and could be counter-productive on the speech quality.

Keywords : Computer science, Speech processing, Text-to-Speech synthesis, HTS