Évaluation segmentale du système de synthèse HTS pour le français

Sébastien Le Maguer Nelly Barbot Olivier Boeffard
Université de Rennes 1, Irisa, Lannion, France
{sebastien.le_maguer, nelly.barbot, olivier.boeffard}@irisa.fr

RÉSUMÉ

HTS est un système paramétrique de synthèse de la parole qui repose sur l'usage de modèles de Markov cachés (HMM). Ce système est de plus en plus diffusé dans le domaine de la synthèse de la parole. Très peu d'études ont cependant été effectuées pour analyser l'influence des paramètres de ce système sur la qualité de la voix de synthèse. L'objectif de cet article est de proposer une évaluation objective de la qualité de la synthèse réalisée par le système dans le but de vérifier l'impact des multiples descriptions sur la qualité de la synthèse. Nos travaux ne concernent que l'analyse du continuum spectral de la parole générée et s'appliquent au français. Nous proposons d'utiliser des GMM pour mesurer les dégradations par rapport à un système de synthèse de référence. Nous proposons enfin un test d'écoute de manière à calibrer notre mesure objective. Les expériences proposées montrent que l'apport des descripteurs autres que la séquence *phonème précédent-courant-suivant* ne semble pas significatif sur la modélisation du spectre.

ABSTRACT

Segmental evaluation of HTS

HTS is a parametric speech synthesis system based on the use of Hidden Markov Models (HMM). HTS is now widely used but very few studies have been conducted to analyze the influence of parameters on the quality of the synthetic speech. The aim of this paper is to provide an objective evaluation of the quality of the synthetic speech produced by HTS in order to assess the influence of multiple descriptions. Our study concerns only the speech spectrum analysis and is applied to French. We propose to use GMM to measure the degradations introduced by HTS compared to reference voice. Finally, we propose a listening test in order to calibrate our objective measure. This method indicates that using other descriptors than the *previous-current-next* phoneme sequence does not improve significantly the modelisation of the spectrum.

MOTS-CLÉS : HTS, qualité segmentale, évaluation, GMM.

KEYWORDS: HTS, evaluation, segmental quality, GMM, spectral features.

1 Introduction

Au cours des années 2000, le système HTS (Zen et Toda, 2005) est devenu populaire dans le domaine de la synthèse de la parole. Il s'agit d'un système paramétrique qui repose sur des modèles de Markov cachés. Le signal de parole est créé à l'aide d'un modèle acoustique dont les paramètres évoluent au cours du temps (le plus souvent sur une base centiseconde). Les modèles acoustiques les plus couramment utilisés sont des filtres MLSA (Fukada *et al.*, 1992)

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 569–576,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

ou plus récemment le modèle STRAIGHT. Pour une phrase à synthétiser, l'évolution temporelle des paramètres de ces modèles acoustiques est déterminée par un HMM au niveau de la phrase dont les observations recouvrent des informations spectrales et prosodiques (f0 et durée). Le principe de HTS est d'utiliser ces HMM dans un mode génératif (et non pas comme classifieur) pour retrouver une séquence plausible de paramètres pour les modèles acoustiques.

Pour HTS, un segment acoustique correspond à un phone en contexte qualifié par un ensemble de descripteurs (Tokuda *et al.*, 2002). Ces descripteurs permettent de tenir compte du contexte phonétique, phonologique, prosodique et linguistique du segment observé. Ils sont utilisés par le système pour regrouper les segments proches et contribuent ainsi à l'apprentissage des modèles.

Les systèmes de synthèse de type HTS sont régulièrement évalués lors du Challenge Blizzard (King et Karaiskos, 2010). Il s'agit principalement de mesures d'intelligibilité et de qualité obtenues par des tests d'écoute. À notre connaissance, peu d'études ont cherché à mesurer l'impact du choix des descripteurs au niveau de la modélisation HMM sur la qualité de la voix de synthèse produite. On peut citer (Chen *et al.*, 2010) qui étudie l'influence des paramètres dits d'accélération en calculant un jeu de distances entre un énoncé généré à partir de modèles incluant des paramètres d'accélération et le même énoncé généré sans tenir compte de ces paramètres. (Silén *et al.*, 2010) étudie la prédiction de durée des segments acoustiques par le système HTS par la mesure d'une erreur de type RMS et d'un coefficient de corrélation entre un énoncé synthétique et l'énoncé original correspondant. La seule étude concernant l'influence des descripteurs sur la qualité de la synthèse a été proposée par (Yokomizo *et al.*, 2010) qui décrit une évaluation de descripteurs de nature prosodique. Cette évaluation est effectuée sur les langues anglaise et japonaise.

L'objectif de cet article est de proposer un protocole d'évaluation de la synthèse obtenue par le système HTS pour la langue française en effectuant un apprentissage dépendant du locuteur. L'évaluation porte exclusivement sur la qualité segmentale du signal de synthèse. Ainsi, les facteurs prosodiques ont été neutralisés dans cette évaluation. Le protocole repose sur une modélisation de l'espace acoustique à l'aide de mixtures de gaussiennes (GMM). Par analogie avec les travaux en transformation de voix, nous faisons l'hypothèse que l'espace acoustique d'un locuteur, qu'il soit naturel ou résultant d'une voix de synthèse, peut être capturé par un GMM. La comparaison d'espace acoustique a pour avantage, sur le calcul d'une distance entre phrase générée, d'être indépendant d'un algorithme d'alignement. Cela permet d'évaluer la qualité de la modélisation spectrale effectuée par HTS découplée de la modélisation de la durée. Ainsi, à chaque configuration particulière du système HTS correspond une voix de synthèse différente reliée à un modèle. En croisant les différents GMM sur les différents corpus, il devient possible d'obtenir une mesure de proximité entre voix. L'objectif est d'observer si des combinaisons de paramètres HTS éloignent ou rapprochent certaines voix entre elles.

L'architecture du système HTS sera brièvement présentée dans la section 2 puis le protocole d'évaluation sera détaillé section 3. Une évaluation subjective sera présentée section 4 afin de valider la pertinence du protocole de mesure objective sur la qualité perçue.

2 Le système HTS

Dans cet article, la version du système HTS utilisée correspond à l'architecture présentée au challenge blizzard de 2005 (Zen et Toda, 2005), plus précisément HTS 2.1.1. Cette architecture

permet d'apprendre des modèles dépendants du locuteur.

HTS repose sur une modélisation statistique de type HMM. Un ensemble de HMM décrit les caractéristiques de variables aléatoires définies ici par des observations acoustiques (spectre, f0, la durée des segments). Par un apprentissage de type supervisé, les HMM mettent en relation des descripteurs de nature linguistique (syntaxe, grammaire, prosodie, phonologie) et une observation de vecteurs acoustiques.

Le vecteur des observations est directement lié au choix du modèle acoustique utilisé pour générer le signal de parole. Il est constitué des coefficients MGC (Mel Generalised-cepstral Coefficients) (Fukada *et al.*, 1992) représentant le spectre, des valeurs de la fréquence fondamentale F0 ainsi que l'apériodicité définis par le modèle acoustique STRAIGHT (Kawahara *et al.*, 1999). À ces coefficients statiques sont associés des coefficients dynamiques (dérivées première et seconde) (Tokuda *et al.*, 2000).

La supervision des HMM est mise en œuvre grâce à un ensemble de descripteurs obtenus par une analyse phonologique, prosodique et linguistique du segment en contexte. Les concepteurs du système HTS ont proposé un ensemble de descripteurs pour la langue anglaise (Tokuda *et al.*, 2002). Cette caractérisation est essentielle car, pour générer un signal, le système en situation de synthèse à partir du texte ne disposera que de ces informations.

Dans l'absolu, il serait nécessaire d'adapter l'ensemble des descripteurs au cas particulier d'une nouvelle langue. Pour une synthèse en français, nous avons repris les descripteurs utilisés pour l'anglais. Les valeurs de ces descripteurs, telles que les étiquettes grammaticales ou encore des informations d'accentuation sont obtenues par des outils d'analyse linguistique propres au français.

L'ensemble des combinaisons de descripteurs étant en pratique impossible à obtenir, une étape de classification des HMM est effectuée lors de la phase d'apprentissage par application d'un arbre de décision (Young *et al.*, 1994). Les nœuds de l'arbre correspondent à des questions permettant de séparer les valeurs d'un descripteur en deux sous-ensembles selon le retour booléen de la question appliquée sur le vecteur de description du HMM. L'objectif est d'aboutir à un partitionnement de l'espace des paramètres des HMM ; chacune des classes obtenues est étiquetée par une information issue du vecteur de description. Les feuilles de l'arbre contiennent les paramètres des modèles gaussiens estimés à partir des vecteurs acoustiques. Lors de la construction de l'arbre, l'application d'un critère MDL (Shinoda et Watanabe, 2000) évite un phénomène de sur-apprentissage.

La phase de génération des paramètres acoustiques s'occupe d'obtenir une séquence de coefficients statiques. Pour chaque segment à synthétiser, il est nécessaire d'extraire une séquence de descripteurs à partir du texte. À partir de la séquence des descripteurs, des feuilles dans l'arbre de décision sont sélectionnées et permettent de récupérer les paramètres des distributions gaussiennes qui seront utilisées pour la génération des paramètres acoustiques. Pour une phrase à synthétiser, un macro-HMM est construit par la mise bout-à-bout de modèles de phones en contexte.

(Tokuda *et al.*, 2000) propose différents algorithmes de génération de paramètres acoustiques à partir du macro-HMM du niveau phrase, nous avons fait le choix d'utiliser l'algorithme maximisant $P(O|Q,\lambda)$ où O correspond aux vecteurs d'observations, λ au vecteur de paramètres des modèles HMM et Q à la séquence d'états du macro-HMM.

3 Evaluation objective

Le protocole proposé a pour but d'étudier l'influence de différents descripteurs sur l'espace acoustique produit par un système HTS mono-locuteur et d'évaluer sa proximité avec l'espace acoustique associé au signal naturel. L'hypothèse centrale de cette méthodologie est d'affirmer que si une configuration du système HTS dégrade la qualité du signal de synthèse au niveau spectral, la vraisemblance des données acoustiques produites sur un modèle de référence de type GMM devrait aussi se dégrader. Comme la vraisemblance d'un GMM dépend à la fois du modèle et des données, nous proposons de conserver comme référentiel un même corpus de test.

Les jeux de descripteurs considérés dans cet article sont présentés Table 1 : les jeux p1, p3 et p5 correspondent à la prise en compte des étiquettes phonétiques du phone et de son contexte correspondant à son horizon proche (0, 1 ou 2 phones) ; les jeux $p3\_full$ et $p5\_full$ permettent d'étudier l'apport des informations prosodiques et linguistiques. L'espace acoustique du locuteur estimé à partir de signaux d'analyse/synthèse servira de référentiel. Ce cas particulier, noté a/s, correspond à une non utilisation de HTS (il s'agit du cas favorable pour les expérimentations). Par la suite, les notations $A_{a/s}$, $V_{a/s}$ et $T_{a/s}$ désigneront trois ensembles de vecteurs acoustiques issus de signaux d'analyse/synthèse, correspondants à des ensembles d'énoncés deux à deux disjoints.

Identifiant	**Description**
a/s	modèle analyse/synthèse de STRAIGHT sans utiliser HTS
p1	phonème courant seulement
p3	phonème en contexte d'horizon 1 (phonèmes précédent, courant et suivant)
p5	phonème en contexte d'horizon 2
p3_full	p3 + informations prosodiques et linguistiques
p5_full	p5 + informations prosodiques et linguistiques

TABLE 1 – Jeux de descripteurs

Pour chaque jeu de descripteurs k (mis à part $k = a/s$), l'apprentissage du système HTS est effectué sur le corpus $A_{a/s}$ en tenant uniquement compte du jeu k. Un corpus A_k, respectivement V_k et T_k, de vecteurs acoustiques est ensuite produit par HTS et correspond aux mêmes énoncés que $A_{a/s}$, respectivement $V_{a/s}$ et $T_{a/s}$.

Afin de modéliser et comparer les espaces acoustiques associés aux vecteurs issus de HTS et aux vecteurs issus directement de l'analyse/synthèse, l'apprentissage de GMM $\mathscr{M}_k$ sur A_k est effectué à l'aide d'un algorithme EM, pour tout $k \in \{a/s, \ldots, p5\_full\}$. Le nombre de composantes de $\mathscr{M}_k$ est réglé à l'aide du corpus de validation V_k. Enfin, on cherche à déterminer les log-vraisemblances croisées entre corpus et modèles : $LL(A_k|\mathscr{M}_k)$, $LL(T_k|\mathscr{M}_k)$ et $LL(T_{a/s}|\mathscr{M}_k)$.

3.1 Modélisation GMM

Pour tout $k \in \{a/s, \ldots, p5\_full\}$, chaque vecteur de A_k correspond à la partie spectrale d'une trame et est représenté par 40 coefficients MGC. Afin d'assurer une bonne stabilité numérique lors de l'apprentissage du GMM $\mathscr{M}_k$, une réduction de la dimensionnalité est d'abord effectuée à l'aide d'une analyse en composantes principales (PCA) appliquée sur les vecteurs du corpus A_k. Le

nombre de vecteurs propres représentant la partie spectrale est choisi conformément à un seuil de variance expliquée d'au moins 95%. Afin de comparer des données homogènes, la transformation linéaire $\mathscr{T}_k$ issue de cette PCA est également appliquée aux vecteurs du corpus de validation V_k, ainsi qu'aux vecteurs des corpus de test T_k et $T_{a/s}$ lors du calcul de leur log-vraisemblance relativement à $\mathscr{M}_k$. Malgré l'application de la PCA, on conservera les notations A_k, V_k et T_k.

Le nombre de composantes n du GMM $\mathscr{M}_k(n)$ est déterminé à l'aide du corpus de validation : pour $i \in [1..9]$, le modèle $\mathscr{M}_k(2^i)$ est appris sur A_k et la log-vraisemblance des éléments du corpus d'apprentissage, $LL(\mathscr{M}_k(n)|A_k)$, et du corpus de validation, $LL(\mathscr{M}_k(n)|V_k)$, sont calculées. Les matrices de covariance des composantes gaussiennes sont diagonales. Une situation de sur-apprentissage est détectée si $LL(\mathscr{M}_k(n)|V_k) << LL(\mathscr{M}_k(n)|A_k)$. On choisit pour valeur optimale de n, $n^\star$, le nombre maximal 2^i tel que $LL(\mathscr{M}_k(n)|V_k) \simeq LL(\mathscr{M}_k(n)|A_k)$.

Une fois le nombre optimal de classes déterminé, les log-vraisemblances des données des corpus de test ($LL(\mathscr{M}_k(n^\star)|T_{a/s})$ et $LL(\mathscr{M}_k(n^\star)|T_k)$ sont calculées et permettent d'évaluer la proximité entre l'espace acoustique généré par HTS et le signal d'analyse/synthèse de référence.

3.2 Protocole expérimental et résultats

Le corpus de parole utilisé pour effectuer l'évaluation repose sur une lecture neutre. Les coefficients ont été extraits d'un signal monophonique échantillonné à 16kHz. Pour réaliser les stimuli, les données sont réparties aléatoirement sur les trois corpus : le corpus d'apprentissage est constitué de 1000 phrases (environ 1h), les corpus de validation et de test de 120 phrases (environ 6min) chacun. Les log-vraisemblances $LL(A_k|\mathscr{M}_k)$, $LL(T_k|\mathscr{M}_k)$, $LL(T_{a/s}|\mathscr{M}_k)$ sont calculées conformément à la méthodologie présentée dans le précédent paragraphe. Les résultats obtenus sont présentés figure 1.

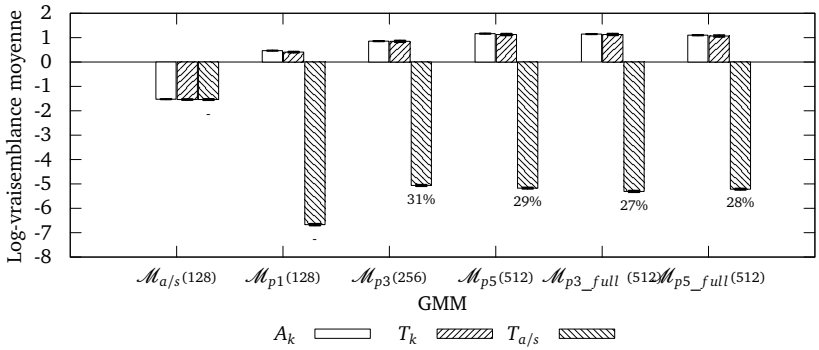
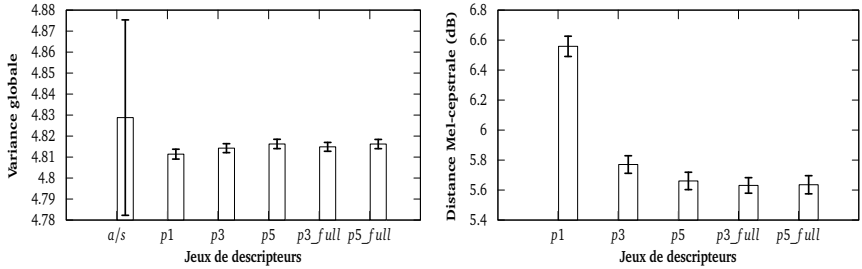


FIGURE 1 – Log-vraisemblances de A_k, T_k et $T_{a/s}$ pour le modèle $\mathscr{M}_k$, avec $k \in \{a/s, \ldots, p5\_full\}$ et leurs intervalles de confiance au niveau 95%. Le nombre de composantes de $\mathscr{M}_k$ est indiqué entre parenthèses. Les pourcentages indiquent les taux d'amélioration $(LL(T_{a/s}|\mathscr{M}_k) - LL(T_{a/s}|\mathscr{M}_{p1}))/(LL(T_{a/s}|\mathscr{M}_{a/s}) - LL(T_{a/s}|\mathscr{M}_{p1}))$ apportés par chaque jeu relativement à $p1$.

Tout d'abord, pour tout $k \in \{a/s, \ldots, p5\_full\}$, on remarque que les log-vraisemblances des corpus A_k et T_k relativement à $\mathscr{M}_k$ sont cohérentes. On observe également que les quantités $LL(A_{a/s}|\mathscr{M}_{a/s})$ et $LL(T_{a/s}|\mathscr{M}_{a/s})$ sont plus faibles que $LL(A_k|\mathscr{M}_k)$ et $LL(T_k|\mathscr{M}_k)$ pour $k \neq a/s$. En fixant le nombre n de composantes pour l'ensemble des modèles (on ignore ainsi l'étape de calibrage des modèles) le même phénomène est observé et cela pour n allant de 128 à 512. La génération des paramètres spectraux réalisée par HTS semble donc réduire la variabilité des paramètres, par rapport à ceux extraits du signal original. Afin d'étayer cette interprétation, la variance globale (Toda et Tokuda, 2007) a été calculée sur chaque corpus d'apprentissage A_k. La figure 2(a) montre ainsi que la variance globale des données extraites du signal est plus variable que celle obtenue pour chacun des corpus générés par HTS.



(a) Comparaison de la variabilité des corpus d'apprentissage (b) Comparaison des corpus via une distance mel-cepstrale

FIGURE 2 – Pour chaque jeu de descripteurs k, la figure (a) indique la variance globale moyenne des vecteurs par phrase dans A_k ainsi que la variance associée, la figure (b) illustre la distance mel-ceptrale moyenne (en dB) entre les vecteurs de T_k (issus de HTS) et ceux de $T_{a/s}$ (issus de l'analyse/synthèse) ainsi que l'intervalle de confiance à 95% correspondant.

D'autre part, si l'on considère le corpus de test $T_{a/s}$, ses données sont naturellement les plus vraisemblables pour $\mathscr{M}_{a/s}$ et les moins vraisemblables pour $\mathscr{M}_{p1}$: 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 signal d'analyse/synthèse, seraient vraisemblables. Enfin, l'utilisation de descripteurs autres que des attributs de séquencement *phonème précédent, phonème courant et phonème suivant* semble peu pertinente pour modéliser le spectre. L'augmentation du nombre de descripteurs accroît mécaniquement le nombre de modèles ainsi que la complexité du partitionnement opéré par l'arbre de classification. L'utilisation de descripteurs peu corrélés de manière directe à une certaine variabilité de l'acoustique peut donc s'avérer contre-productive. Pour comparer nos résultats à ceux obtenus par (Yokomizo *et al.*, 2010), une distance mel-cepstrale a été calculée entre les phrases issues de chaque corpus de test généré par HTS et celles du corpus de test dont les coefficients sont issus de l'analyse/synthèse. Les résultats sont présentés dans la figure 2(b), La distorsion spectrale moyenne obtenue pour le jeux de descripteurs $p5\_full$ est comparable à celle identifiée par l'identifiant "fullcontext" dans (Yokomizo *et al.*, 2010). De plus, les résultats obtenus par calcul d'une distance cepstrale corroborent ceux décrits précédement pour la méthode basée sur les GMM.

4 Validation subjective

Afin de valider les résultats de l'évaluation objective décrite ci-dessus, une évaluation subjective a été réalisée. Par souci de cohérence, il est nécessaire de n'évaluer que la qualité des coefficients spectraux. Pour répondre à cet objectif, le signal synthétisé est obtenu en utilisant les coefficients spectraux générés par HTS couplés au F0 et à l'apériodicité extraits du signal naturel par STRAIGHT. Lors de la phase de génération des paramètres, la durée naturelle des phones a été imposée à HTS. La méthode d'évaluation subjective choisie est un test de type ACR(ITU-T, 1996) avec une mesure MOS. Cinq notes allant de *mauvais* (1) à *excellent* (5) permettent de qualifier un signal perçu par un auditeur. Le test porte sur la qualité globale du système. L'intitulé de la question proposée à l'auditeur était la suivante : "Comment jugez-vous la qualité de l'échantillon sonore que vous venez d'écouter ?"

Les 7 auditeurs sont des experts dans le domaine du traitement de la parole. Parmi les énoncés prononcés par le locuteur, 21 énoncés sont tirés au hasard dans le corpus de test. Des signaux correspondants sont donc construits à partir des coefficients spectraux associés dans T_k pour $k \in \{a/s, \ldots, p5\_full\}$. Ainsi, à chaque énoncé correspond 6 signaux, un par système testé. Un énoncé est choisi au hasard parmi les 21 initiaux et les 6 signaux associés constituent la phase d'introduction au test d'écoute, qui ne sera pas prise en compte dans les résultats. La durée globale du test est d'environ 40 minutes. Les résultats sont présentés figure 3 : l'axe des ordonnées présente le score MOS moyen obtenu selon le jeu de descripteurs indiqué en abscisse.

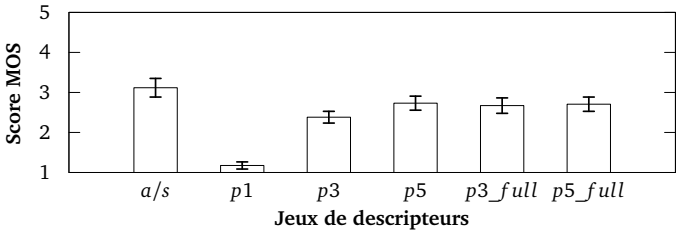


FIGURE 3 – Résultat de l'évaluation subjective, test ACR sur une échelle MOS.

On constate tout d'abord que les notes obtenues sur des échantillons d'analyse/synthèse sont relativement faibles. Cela semble indiquer que la paramétrisation du spectre introduit par le modèle STRAIGHT peut impliquer une dégradation de la qualité du signal de synthèse. L'intervalle de confiance associé à l'analyse/synthèse chevauche ceux de $p5$, $p3\_full$ et $p5\_full$. Au delà de ces observations, l'évaluation subjective confirme les résultats de l'évaluation objective : le jeu de descripteurs $p1$ obtient le plus mauvais score et les autres jeux ne sont pas significativement différents.

5 Conclusion

Dans cet article, nous avons présenté un protocole pour mesurer objectivement des dégradations pouvant être apportées par un système de synthèse de type HTS. L'usage de GMM permet de

modéliser les espaces acoustiques des différentes voix de synthèse. Une validation utilisant un corpus de test de référence sur différents modèles permet de quantifier les dégradations expliquées par l'usage de tel ou tel descripteur dans la modélisation HTS. Le résultat surprenant au premier abord est qu'un simple modèle qui n'utilise que des étiquettes phonétiques sur un horizon de 5 phones fait aussi bien que l'ensemble des descripteurs communément admis (étiquettes grammaticales, syntaxe, syllabes, positions respectives des éléments sur leur niveau de description, etc.). Cette mesure objective est confirmée par un test subjectif. En terme de perspectives, nous comptons appliquer ce scénario pour mesurer automatiquement la qualité de diverses combinaisons de descripteurs sur des panels de voix différentes de manière à confirmer expérimentalement le choix pertinent d'un vecteur de description pour une synthèse de type HTS (notamment par la prise en compte conjointe de critères segmentaux et prosodiques).

Références

CHEN, Y.-n., YAN, Z.-j. et SOONG, F. K. (2010). A Perceptual Study of Acceleration Parameters in HMM-Based TTS. *In proceedings of Interspeech*.

FUKADA, T., TOKUDA, K., KOBAYASHI, T. et IMAI, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. *In proceedings of ICASSP*, pages 137–140.

ITU-T (1996). P800 : Methods for objective and subjective assessment of quality. Rapport technique.

KAWAHARA, H., MASUDA-KATSUSE, I. et DE CHEVEIGNÉ, A. (1999). 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. *Speech Communication*, 27:187–207.

KING, S. et KARAISKOS, V. (2010). The Blizzard Challenge 2010.

SHINODA, K. et WATANABE, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *Journal of the Acoustical Society of japan*, 21(2):79–86.

SILÉN, H., HELANDER, E., NURMINEN, J. et GABBOUJ, M. (2010). Analysis of Duration Prediction Accuracy in HMM-Based Speech Synthesis. *In proceedings of Speech Prosody*.

TODA, T. et TOKUDA, K. (2007). A speech parameter generation algorithm considering global variance for hmm-based speech synthesis. *IEICE Transactions*, pages 816–124.

TOKUDA, K., YOSHIMURA, T., MASUKO, T., KOBAYASHI, T. et KITAMURA, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *In proceedings of ICASSP*, pages 1315–1318.

TOKUDA, K., ZEN, H. et BLACK, A. W. (2002). An hmm-based speech synthesis system applied to english. *In proceedings of ICASSP*, pages 227–230.

YOKOMIZO, S., NOSE, T. et KOBAYASHI, T. (2010). Evaluation of Prosodic Contextual Factors for HMM-Based Speech Synthesis. *In proceedings of Interspeech*, pages 430–433.

YOUNG, S. J., ODELL, J. J. et WOODLAND, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *In proceedings of HLT*, pages 307–3012.

ZEN, H. et TODA, T. (2005). An overview of Nitech HMM-based speech synthesis system for blizzard challenge 2005. *In proceedings of Eurospeech*, pages 1957–1960.