Codage de la parole à très bas débit par indexation d'unités de taille variable

M. Padellini[1], G. Baudoin[2], F. Capman[1]

(1) Thales Communications, 66 rue du fosse blanc – BP 156, 92231 Gennevilliers Cedex
Mel : padellim@esiee.fr

(2) ESIEE, signal processing and telecommunication department, BP 99, cedex 93162 Noisy Le Grand
Tél.: ++33 1 45 92 06 46 - Fax: ++33 1 45 92 66 99
Mél: g.baudoin@esiee.fr

ABSTRACT

Ce papier présente un schéma de codage de la parole par indexation utilisant des techniques de reconnaissance et de synthèse par corpus. Cette méthode a permis d'atteindre des débits inférieurs aux codeurs basés sur des modèles paramétriques de la parole. Au cours de cette présentation, les améliorations nécessaires pour une utilisation pratique de ce type d'approche seront abordées.

1. INTRODUCTION

Actuellement, les techniques de codage de la parole bas débit visent à caractériser le signal par un modèle paramétrique de la parole. Ces approches permettent d'atteindre des débits compris entre 800 bits/sec et 4800 bits/sec. Pour permettre de franchir cette limite, une nouvelle approche est nécessaire. Elle a été initiée par les travaux de Cernocky [Cer98], et s'inspire des progrès réalisés en reconnaissance et en synthèse vocale.

L'idée principale consiste à segmenter le signal de parole et à extraire le signal de parole aux frontières de ces segments pour constituer des unités de synthèse. Une fois classées et regroupées dans une base de données, le signal de parole peut alors être caractérisé par une suite d'indexes faisant référence aux unités de la base. Le signal de parole peut être reconstruit en concaténant les unités référencées par le signal codé à partir de la base d'apprentissage.

Pour pouvoir réaliser une compression, la base de données doit être de taille limitée. On la limite donc aux données d'apprentissage. La segmentation du signal et la classification des unités sont réalisées par un système de reconnaissance de parole. Ces tâches sont entièrement automatisées grâce à une modélisation statistique HMM (Hidden Markov Models) des segments. Cependant, la limitation de la taille de la base d'apprentissage ne permet pas de caractériser toutes les variations d'intonation que peut réaliser un locuteur. Pour palier à ce problème, la prosodie est extraite du signal à coder et est transmise avec les indexes. Elle permet lors du décodage de modifier les unités pour restituer l'intonation du locuteur.

Cette approche permet d'avoir un débit moyen de 500 bits/s pour un échantillonnage de la parole à 16Khz. Dans la **Section 2**, nous verrons les problèmes liés à la phase d'apprentissage, qui est faite de manière entièrement automatique pour permettre une utilisation pratique du codeur et facilement extensible à d'autres locuteurs. Dans **la Section 3** nous présenterons le schéma général de codage et de décodage et les problèmes liés à la sélection d'unités.

2. PHASE D'APPRENTISSAGE

La phase d'apprentissage se divise en deux étapes. La première consiste à créer un modèle de reconnaissance de type HMM à partir des données paramétrées de parole. La seconde étape consiste à classer les données d'apprentissage automatiquement et à les organiser dans la base de données.

2.1 Constitution d'un modèle HMM

Pour pouvoir entraîner les modèles de Markov, le signal de parole est paramétré ce qui permet d'aborder le signal d'une manière plus physique. La paramétrisation s'inspire des modèles de production de parole initialement développés pour coder la voix. D'autre part, des modèles de Markov sont utilisés conjointement pour modéliser statistiquement la variabilité des segments de parole. Ils s'apparentent aux modèles de langages utilisés dans les systèmes de reconnaissance vocale.

Paramètres de reconnaissance

Le choix des paramètres représentant le signal de parole peut avoir beaucoup d'incidences sur les performances et la robustesse du système au bruit. Pour l'instant, le système de reconnaissance est basé sur des coefficients LSF (Line Spectral Frequencies) et d'énergie. D'autres choix de paramètres peuvent être mieux adaptés à la reconnaissance et s'inspirent du système de perception humain. Des paramètres comme les MFCC (Mel Frequency Cepstral Coefficients) ou bien les coefficients PLP (Perceptual Linear Predictive) doivent être étudiés pour ce système. En ce qui concerne la robustesse au bruit, une simple soustraction du cepstre moyen est appliquée pour normaliser les données vis à vis des variations aux conditions d'enregistrement. D'autres procédés de

normalisation sont à prévoir : le filtrage RASTA développé par Hemansky [Her94] ou bien l'analyse en composante principale en sont des exemples. Ils ont permis d'améliorer sensiblement la robustesse au bruit des systèmes de reconnaissance.

Segmentation

Les modèles de Markov sont estimés à partir des paramètres extraits des données d'apprentissage. Ils doivent rendrent compte de la variabilité de différentes séquences de paramètres (segments de paroles). Ces modèles étant statistiques, une grande quantité de données est nécessaire pour pouvoir les entraîner. Environ une heure de parole par locuteur est utilisée comme données d'apprentissage, provenant du corpus lu de langue française BREF.

Pour pouvoir réaliser l'estimation des modèles, il faut au préalable avoir une première segmentation du signal de parole pour pouvoir initialiser les modèles. Un critère de segmentation doit donc être fixé. Le choix s'est porté sur la décomposition temporelle développée par Atal [Ata83]. Cette méthode a pour but de décomposer le signal de parole en des fonctions localisées d'interpolation, déterminées par un centre de masse et un coefficient de pondération. Le but de l'algorithme est de réduire l'erreur quadratique entre le signal reconstruit à partir des fonctions d'interpolation et le signal d'origine. Le centre de masse se situe sur des zones de stabilité du signal, où il y a peu de variations. La segmentation est réalisée aujourd'hui à l'intersection de deux fonctions d'interpolation. Un segment représentera donc une zone stable du signal : un phone. Une fois la base de données entièrement segmentée par cette méthode, une quantification vectorielle est effectuée sur les vecteurs milieux des segments. Une première classification des segments permet alors d'initialiser les modèles de Markov de manière itérative.

Le choix du nombre de modèles HMM reste cependant délicat. Une correspondance phonétique pourrait être visée mais le critère de stabilité que l'on s'est fixé ne permettra pas à priori d'obtenir une telle correspondance. Une étude est à faire sur l'impact du nombre de classes ainsi que sur la taille des segments sur les performances du système. De plus, la segmentation étant réalisée sur des zones instables, il serait peut être plus judicieux de resegmenter le signal sur des zones stables pour permettre d'effectuer un lissage des segments lors de la synthèse. Enfin, d'autres critères de segmentation peuvent être choisis. Par exemple, Lee [Lee02] a utilisé une segmentation automatique par VLSQ pour réaliser un codeur très bas débit. Cette méthode est en quelque sorte une quantification matricielle à taille variable. Les segments peuvent être classés et ne nécessitent plus de modèles HMM.

2.2 Constitution de la base de données

Une fois les modèles HMM estimés, les données d'apprentissage peuvent être re-segmentées et classées (Figure 1). Chaque segment est associé à une classe et une sous-classe correspondant à la classe du segment qui lui précède. De cette manière, le contexte de chaque séquence extraite est conservé.

Le problème principal lié à la base de données reste sa taille. Des méthodes classiques paramétriques de compression de voix pourraient réduire sensiblement sa taille. D'autre part, une approche visant à simplifier la base pourrait permettre de garder uniquement les segments les plus représentatifs. Enfin, afin d'accélérer la phase d'apprentissage, les phrases d'apprentissage pourraient être optimisées de manière à ce qu'elles soient phonétiquement équilibrées et permette de couvrir rapidement l'ensemble des phones que le locuteur est capable de produire.

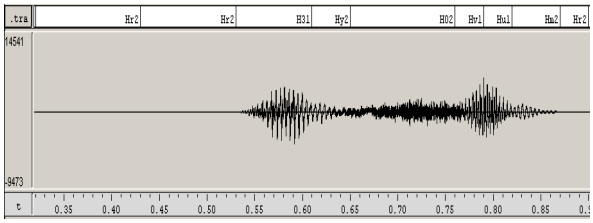

Figure 1: Segmentation et classification d'un signal de parole par modèles HMM.

3. PHASE DE CODAGE, DECODAGE

Une fois la phase d'apprentissage terminée, le processus de codage doit tirer au mieux parti de la base de données. Le signal de parole est segmenté par un algorithme de Viterbi utilisant les modèles HMM estimés. Chaque segment reconnu est associé à l'unité de synthèse de la base de données qui permettra de se rapprocher le plus possible du signal d'origine. Pour cela, l'unité de synthèse sélectionnée est modifiée pour rendre les variations prosodiques du locuteur. Enfin, les indices des classes et les informations prosodiques sont quantifiées pour pouvoir être envoyées sur le canal de transmission. A l'inverse, le processus de décodage consiste à déquantifier les informations de prosodie et à appliquer les mêmes modifications sur les unités de synthèses sélectionnées.

Modèle harmonique plus bruit

Pour pouvoir sélectionner l'unité représentant au mieux le segment à coder, une analyse harmonique du signal de parole est nécessaire. Cette analyse devra permettre de rendre les modifications des segments naturelles à l'oreille vis à vis du pitch ou des changements d'échelle de temps. L'analyse par coefficients LSF n'est pas adaptée à cette application. On rencontre le même genre de problèmes en synthèse par corpus. En effet, pour avoir une voix synthétique ayant la plus grande variabilité possible, des modifications sur les unités du corpus sont nécessaire pour pouvoir rendre une intonation naturelle.

Beaucoup de systèmes utilisent un modèle harmonique du signal de parole auquel vient se rajouter une composante de bruit plus ou moins grande suivant le voisement des sons. Stylianou [Sty96] a développé le modèle HNM basé sur cette approche (Harmonic plus Noise Model). C'est ce modèle qui est utilisé dans le système de codage.

Sélection des unités

Une analyse HNM est réalisée sur les segments. Sachant la classe et la sous-classe du segment à coder, le segment le plus proche doit être sélectionné. Le choix est fait parmi l'ensemble des représentants (unités de synthèse) de cette sous-classe. Mais on doit aussi sélectionner l'unité qui permettra la meilleur concaténation lors du décodage. Hunt [Hun96] a introduit des coûts de concaténation permettant de sélectionner l'unité qui permettra la concaténation la plus naturelle lors de la synthèse. Pour cela il introduit un coût visant à pénaliser les unités dont le pitch moyen est différent du pitch du segment à coder, et à encourager les unités qui se suivent dans les données d'apprentissage. Ainsi des passages naturels de coarticulations peuvent être rendus.

Décodage

Chaque unité de synthèse est concaténée et modifiée de manière à se rapprocher le plus possible des informations de prosodie transmises. Le décodage des unités se fait actuellement sans lissage à cause d'une segmentation sur des zones instables du signal.

4. CONCLUSION

Ce schéma de codage a été testé sur neufs locuteurs de la base BREF et est en cours d'évaluation. De nombreuses améliorations restent à faire, notamment pour permettre une utilisation dans un environnement bruité ou avec des locuteurs multiples. Pour palier à ce problème un système d'adaptation au locuteur doit être développé. Dans le cadre de la conversion de voix, Stylianou [Sty96b] a proposé une méthode de transformation de voix basée sur un modèle GMM (Gaussian Mixture Model) des locuteurs. Cette méthode nécessite d'avoir des phrases identiques, lues par chaque locuteur pour pouvoir fonctionner. Elle a été développée dans le but de créer de nouvelles voix de synthèse et ne convient pas à une adaptation progressive lors d'une conversation téléphonique par exemple. De plus, les modèles HMM doivent aussi être adaptés pour permettre de réaliser une bonne segmentation du signal. Les techniques adoptées en reconnaissance de la parole pourraient être utilisées dans le système. Un grand nombre d'améliorations doivent encore être réalisées pour que ce schéma de codage soit utilisable dans des conditions réelles d'utilisation.

BIBLIOGRAPHIE

[Cer98] Cernocky J. (1998), "Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification", PhD Thesis, Université Paris XI, Orsay.

[Her94] Hermansky H. (1994) " RASTA Processing of speech", IEEE Transactions on Speech and Audio processing, pp. 578-589.

[Ata83] Atal B. (1983) " Efficient coding of LPC parameters by temporal decomposition", Proc. IEEE ICASSP, pp 81-84.

[Lee02] K.S. Lee, R.V. Cox (2002), "A segmental speech coder based on a concatenative TTS", Speech Communication 38, pp 89-100.

[Sty96] Stylianou Y. (1996), "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD Thesis, Ecole Nationale Supèrieure des Télécommunications.

[Hun96] A.J. Hunt, A.W. Black (1996), "Unit selection in a concatenative speech synthesis syste using a large speech database", Proc. IEEE ICASSP.

[Sty96b] Stylianou Y., Cappé O. (1996) "Statistical methods for voice quality transformation", Proc. IEEE ICASSP.