Evaluation d'un codeur de parole à très bas débit

Marc PADELLINI[1,2], Geneviève BAUDOIN[1], François CAPMAN[2]

[1] ESIEE, ESYCOM Telecommunications Systems LaboratoryBP 99, 93162 Noisy-Le-Grand, CEDEX, France.
[2]THALES COMMUNICATIONS, 160, Bd de Valmy, BP 82, 92704 Colombes, CEDEX, France.
marc.padellini@fr.thalesgroup.com

Résumé – Les approches classiques de codage de la parole à la trame ne permettent pas d'assurer une qualité de restitution suffisante à des débits inférieurs à 600 bits/s. En dessous, une nouvelle classe de codeurs dite à très bas débit doit être utilisée pour permettre d'exploiter la corrélation entre les trames successives afin de maintenir une bonne intelligibilité de la parole. Cet article présente un schéma complet de codage à très bas débit à 500 bits/s, qui exploite une base de donnée de segments de parole déterminés de manière entièrement automatique. Une évaluation complète de cette approche est réalisée suivant trois tests différents (MOS, intelligibilité, intelligibilité « segmentale ») conjointement avec trois autres codeurs dont le débit se situe entre 3600 bits/s et 600 bits/s. Les tests montrent qu'une bonne intelligibilité ainsi qu'une bonne qualité peuvent être atteintes grâce à ce type d'approche.

Abstract *– Classical frame based speech coding is lacking of quality at rates under 600 bits/s. Below, a new type of coders must be used to achieve sufficient intelligibility, by exploiting the correlation between successive frames. This paper presents a very low bit rate speech coder at 500 bits/s, resorting on a speech segment database designed by a completely unsupervised process. A three test evaluation on MOS, intelligibility and "segmental" intelligibility has been performed jointly with three other coders operating at a bit rate ranging between 3600 bits/s and 600 bits/s. This evaluation reveals that good intelligibility as well as good quality can be achieved by using segment indexing techniques.*

1. Introduction

Les codeurs de parole à très bas débit standardisés comme le NATO STANAG 4591 à 1200 bits/s utilisent des schémas classiques de codage à la trame de type MELP (Mixed Excitation Linear Predictive vocoder), dont le fonctionnement a été étendu à des bas débits en optimisant la quantification de trames successives. Ceci montre les limites du codage de la parole par trames. En effet, on peut dire de manière générale qu'en dessous de 1000 bits/s l'intelligibilité de ce type de codeur est fortement dégradée.

Ces vingt dernières années, plusieurs approches de codage à très bas débit ont été proposées afin d'exploiter la corrélation entre plusieurs trames successives de parole. Hoshiya et al. [2] proposent de combiner reconnaissance et synthèse de la parole en utilisant des HMM (Hidden Markov Models). Le codage est similaire à un système de reconnaissance vocale : une séquence de phonèmes est déterminée à l'aide de HMM. Cette séquence permet de restituer la parole en concaténant les modèles HMM correspondants, puis en générant une suite de coefficients cepstraux à partir de leur densité de probabilité d'émission. Ces coefficients permettent d'obtenir une approximation de l'évolution de l'enveloppe spectrale. Ce type d'approche permet d'atteindre des débits de l'ordre de 150 bits/s, cependant la qualité de ce type de codeur semble encore limitée [2].

Un autre approche consiste à transmettre directement l'évolution de l'enveloppe spectrale en utilisant par exemple un modèle polynomial sur des paramètres proches des formants [3]. Une bonne intelligibilité est atteinte entre 550 et 650 bits/s. Mais l'enveloppe spectrale n'est pas précise à cause du faible nombre de paramètres mis en jeu, ce qui limite la qualité de restitution. En revanche ce type de schéma permet un codage indépendant du locuteur.

Enfin une dernière approche s'inspire des techniques de reconnaissance et de synthèse vocale TTS (Text to Speech Synthesis). Le codeur reconnaît une séquence d'unités acoustiques dans le signal original. Pour chaque segment, il transmet le symbole correspondant à l'unité reconnue ainsi que des paramètres auxiliaires tels que les contours de fréquence fondamentale et d'énergie ainsi que la longueur du segment. La synthèse se fait par concaténation de représentants des unités élémentaires. Ces représentants (formes d'onde temporelles) sont stockés dans une base de données ou corpus de parole dont la taille est similaire à celle des systèmes TTS.

Ce type de codage utilise des phonèmes comme unités acoustiques. Potentiellement une bonne qualité peut être atteinte pour des débits de l'ordre de 400 bits/s [4] mais l'apprentissage doit être supervisé. D'autres types d'unités acoustiques peuvent être utilisées pour résoudre ce problème. Pour cela, des classes d'unités sont déterminées automatiquement par des méthodes purement statistiques. Une qualité se rapprochant d'un MELP à 2400 bits/s a été atteinte par KS. Lee et R. Cox [5] pour un débit de 580 bits/s. Le codeur étudié dans ce papier appartient à cette famille de codeurs, ils permettent d'offrir un codage large bande du signal de parole pour des débits inférieurs à 1 Kbits.

Cet article présente un schéma de codage très bas débit par indexation d'unités de taille variable appelé VLBR (Very Low Bit Rate coder). Il a été développé dans le cadre du projet RNRT SYMPATEX et a été comparé à différents codeurs bas débits dans une évaluation réalisée par le

CNAM. Les différents codeurs faisant l'objet de l'évaluation sont présentés dans la partie 3. Le protocole de test employé ainsi que les résultats obtenus sont présentés dans la partie 4.

2. Présentation du codeur VLBR

Ce schéma de codage se place dans la continuité des travaux présentés dans [6]. Il utilise une approche de reconnaissance par HMM (Hidden Markov Model) d'unités élémentaires de parole dans le codeur et une approche de synthèse par corpus dans le décodeur [11]. Ces unités correspondent à des segments de parole où le signal est stable. Le système utilise un corpus de parole de grande taille dans lequel le codeur recherche le segment le plus proche du segment à coder et dans lequel le décodeur lit le segment de synthèse à concaténer.

2.1 Phase d'apprentissage

Une phase d'entraînement des HMM est préalablement nécessaire pour déterminer les unités élémentaires de parole. Pour cela 64 HMM sont entraînés de manière itérative. Leur topologie est la suivante : trois états, une évolution de type gauche-droite avec une gaussienne modélisant la probabilité d'émission. Pour initialiser l'apprentissage, une première transcription est nécessaire. Elle est obtenue en deux étapes, par une étape de segmentation, puis une étape de classification des segments. Le corpus de parole est segmenté en réalisant un dendrogramme qui regroupe les trames successives les plus proches afin de former des segments ancrés sur des zones stables du signal. Puis ces segments sont classés en minimisant la distance cumulée entre les segments et 64 centroïdes obtenus par quantification vectorielle du corpus. Après l'entraînement des HMM, un algorithme de Viterbi est utilisé pour segmenter et classer conjointement l'ensemble du corpus. Tous les segments de même classe et de même sous-classe sont regroupés dans une base de donnée. La sous-classe représente le contexte d'une unité, elle est définie par la classe de l'unité la précédent dans le temps.

2.2 Phase de codage

Le codeur est représenté sur la figure 2. Il effectue l'extraction au fil de l'eau de 16 paramètres cepstraux de type LPCC (Linear Predictive Cepstral Coefficients), ainsi que le calcul de l'énergie sur des fenêtres de 20ms avec 10ms de recouvrement. Un algorithme de Viterbi à horizon limité est utilisé pour segmenter et classer conjointement les segments à l'aide des HMM.

2.2.1 Sélection des unités

Pour pouvoir sélectionner l'unité de la base la plus proche de l'unité à coder, un processus de sélection à deux étapes a été proposé dans [7]. Tout d'abord les 16 unités - de la même classe et de la même sous-classe - les plus proches en terme de pitch moyen sont présélectionnées. Puis l'unité qui combine la meilleure corrélation avec l'unité à coder est sélectionnée. Une corrélation cumulée sur le profil de pitch, le profil d'énergie et l'enveloppe spectrale moyenne est utilisée comme critère de sélection. Ceci permet de garder l'ensemble des unités du corpus tout en ayant une quantité fixe de bits alloués à l'index de l'unité sélectionnée.

2.2.2 Correction de la prosodie

Un modèle appelé HSD pour (Harmonic Stochastic Model) est utilisé pour manipuler l'unité sélectionnée. C'est un modèle de type Harmonique plus Bruit basé sur [8]. Il permet de réaliser des transformations sur l'échelle de temps et sur le pitch avec une bonne qualité. Deux paramètres de correction de la prosodie sont calculés en normalisant linéairement la longueur de l'unité sélectionnée avec celle de l'unité à coder:

- Un paramètre de correction linéaire permet de corriger la pente du profil de pitch.
- Un gain permet de corriger le niveau moyen du profil d'énergie.

Ces paramètres sont quantifiés sur une échelle logarithmique.

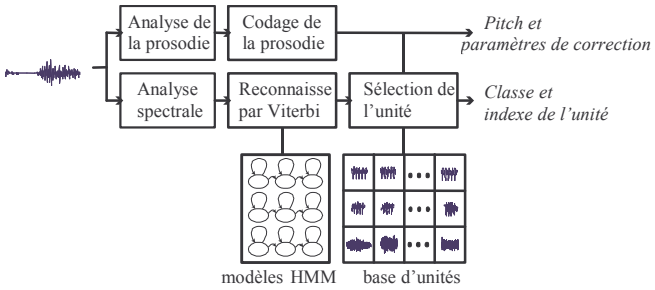

FIG. 2: Principe de codage VLBR.

2.3 Phase de décodage

Pendant la phase de décodage, les unités sont récupérées dans la base d'unités grâce à l'indice de classe, l'index de l'unité et le pitch moyen transmis. Les unités sont corrigées grâce aux paramètres de correction de la prosodie, puis elles sont concaténées pour restituer le signal de parole.

2.4 Table d'allocation des bits

Le débit moyen obtenu est d'environ 500 bits/s avec 170 bits/s réservés aux indices de classes et 330 bits/s réservés au codage de la prosodie. La table d'allocation des bits est présentée sur le tableau 1. La version testée ici est dépendante du locuteur et traite des signaux en bande élargie.

TAB 1: Table d'allocation des bits pour une unité

Paramètre	**Nombre de Bits alloués**
Classe (64)	6 bits
Index de l'unité (16)	4 bits
Total Information spectrale	*10 bits par segment*
Longueur du segment (3-18)	4 bits
Pitch moyen	5 bits
Correction du pitch	5 bits
Correction de l'énergie	5 bits
Total Information Prosodie	*19 bits/segment*
Total	**29 bits/segment**

3. Présentation des codeurs évalués

Cinq codeurs ont été retenus pour l'évaluation :

- Le codage PCM (Pulse Code Modulation) à une fréquence d'échantillonnage de 16000Hz car il permet

de constituer une référence du signal de parole non dégradé.

- L'analyse/synthèse harmonique (HSD) car elle correspond à la qualité maximale que l'ont peut espérer obtenir avec le codeur VLBR puisqu'elle est mise en jeu dans la manipulation des unités.
- Le codeur HSX-WB (Harmonic Stochastic eXcitation Wide-Band) au débit de 3600 bit/s. Il a été développé par THALES Communications sur la base décrite dans [9] et étendu à un fonctionnement en bande élargie [50 7000Hz]. Il offre une très bonne qualité sonore et une complexité de traitement réduite face au MELP.
- Le codeur MELPe (Enhanced Mixed excitation Linear Predictive vocoder) décrit dans [1] standardisé NATO STANAG-4591, à un débit de 2400 bit/s. Le codage porte sur le signal de parole en bande étroite [50 4000Hz].
- Une version du codeur MELP à 600 bit/s. Elle a été retenue pour une prochaine standardisation STANAG 4591.

4. Evaluation

Le codeur VLBR a d'abord été entraîné durant une phase d'apprentissage (Cf 2.1) séparément sur 10 locuteurs de la base BREF [10]. La quantité de parole utilisée est de l'ordre d'une heure de parole. 70 phrases par locuteur ont été réservées pour constituer des données de test.

4.1.1 Matériel de parole utilisé :

Sont codés par les 5 codeurs de l'évaluation des phrases qui constitueront 3 matériels de test:

- 70 phrases prononcées par 10 locuteurs. Ils ont été sélectionnés pour leur diversité dans la base de corpus lus de parole BREF (voir [10]). (M1)
- 40 phrases n'ayant pas de sens logique, extraites du corpus CNAM lues par 1 locuteur. (M2)
- 8 séquences de 28 mots lues par 1 locuteur. (M3)

Treize auditeurs ont été recrutés parmi des étudiants faisant des études supérieures. Trois tests différents ont été réalisés sur les quatre codeurs lors de l'évaluation .

4.1.2 Test subjectif MOS :

Le matériel de test (M1) est utilisé sur 13 auditeurs. Le test porte sur deux critères: la qualité globale et l'intelligibilité. 50 phrases sont diffusées (une phrase est tirée aléatoirement pour chaque locuteur et chaque codeur). Les auditeurs attribuent une note de 1 à 5 à chacun des extraits sonores et ce sur un critère seulement.

Pour tester la fatigue auditive des auditeurs, le test est doublé en alternant de critère. Ainsi les 13 auditeurs ont participé à 4 séances d'écoute.

Les résultats obtenus sont présentés sur la Figure 3 et 4, avec des intervalles de confiance à 95%. Les deux critères donnent le même classement, cependant la courbe d'intelligibilité est simplement translatée vers le haut ce qui montre que les codeurs sont plus performants en intelligibilité.

Le codeur VLBR surpasse le MELP à 600 bits/s mais reste de mauvaise qualité, alors que le MELPe à 2400 bits/s est de qualité moyenne. Le codeur HSX-WB et l'Analyse/Synthèse HSD sont perçus de manière équivalente et sont de bonne qualité.

4.1.3 Test d'intelligibilité par retranscription :

Le matériel de test (M2) est utilisé pour 11 auditeurs. Les locuteurs doivent retranscrire ces 40 phrases qui n'ont pas de sens logique. Il n'y a pas d'a priori possible, l'ordre des phrases et le codeur utilisé sont déterminés aléatoirement. Sur la Figure 5 est représenté le nombre de phonèmes non ou mal reconnus par phrase.

Ce test place le VLBR et le MELPe à 2400 bits/s à des niveaux comparables de performance en terme d'intelligibilité, et devant le MELP à 600 bits/s. En revanche le codeur HSX-WB et l'Analyse/Synthèse HSD donnent des niveaux comparables à la référence non dégradée.

4.1.4 Test d'intelligibilité « segmental » :

Le matériel de test (M3) est utilisé sur 11 auditeurs. C'est un test de rimes portant sur des mots monosyllabiques de type consonne/voyelle/consonne. L'auditeur entend un mot et doit choisir le mot qu'il a reconnu parmi deux mots qui lui sont présentés visuellement sur une feuille. Une séquence de 28 mots issue du même codeur doit être reconnue. Les codeurs et les séquences sont alternées aléatoirement de manière à ce que l'auditeur ait à reconnaître les 8 séquences de mots codés par les 6 codeurs, c'est à dire 48 séquences. Le nombre d'erreurs par séquence est représenté Figure 6.

Ce test place le VLBR et le MELPe à 2400 bits/s à des niveaux comparables (recouvrement de leurs intervalles de confiance), et devant le MELP à 600 bits/s. Le HSX-WB est proche de l'Analyse/Synthèse HSD mais les auditeurs le distinguent clairement par rapport à l'original.

5. Conclusion

Cette article a présenté les performances générales de codeurs à bas débit. Il montre que le codeur VLBR permet d'atteindre des niveaux d'intelligibilité comparables à ceux du codeur MELPe à 2400 bits/s pour un débit 5 fois plus faible. Ceci s'explique par l'apport du codage en bande élargie et par la représentation de l'évolution spectrale des trames par segments. Le codeur HSX-WB à 3600 bps bénéficie aussi de l'apport de la bande élargie ce qui lui confère une très bonne qualité globale de codage.

En revanche la qualité globale du VLBR reste à améliorer pour concurrencer les autres approches de codage et justifier les contraintes qui lui sont liées. Elle est toutefois légèrement supérieure à celle du codeur MELP à 600 bits/s. La concaténation sur des unités de type diphones devrait pouvoir réduire les discontinuités propres à ce type d'approche. Des travaux sur la robustesse du codeur VLBR dans des environnements bruyant sont en cours ainsi que son extension à un codage indépendant du locuteur.

Ce travail a été réalisé dans le cadre du projet RNRT SYMPATEX.

Références

[1] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J.S. Collura. *A 1200/2400 bps coding suite based on MELP*. SCW'02, pp 90-92, 2002.

[2] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. *Improving the performance of HMM-based very low bit rate speech coding.* Proc. ICASSP, pp800-803, 2003

[3] P. Zolgaghari, T. Robinson. *Speech coding using mixture of gaussians polynomial model.* Proc. Eurospeech. 1999

[4] M. Felici, Al. *Very low bit rate speech coding using a diphone-based recognition and synthesis approach.* Trans. IEE Electronic letters 9(34), pp 859-860, 1998.

[5] K.S. Lee, R.V. Cox. *A segmental speech coder based on a concatenative TTS*. Trans. Speech Communication 38, pp 89-100. 2002.

[6] J.Cernocky, G.Baudoin, G.Chollet, *Segmental vocoder – going beyond the phonetic approach*, Proc. ICASSP-98, pp. 605-608, 1998.J. Cernocky.

[7] M. Padellini, F. Capman, G. Baudoin. *Very low bit rate (VLBR) speech coding around 500 bits/sec*. Proc. EUSIPCO, Vienne, 2004.

[8] I. Stylianou, *Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole et du Locuteur.* Thèse ENST, 1996.

[9] P. Gournay, F. Chartier. *A 1200 bits/s speech coder for very low bit rate communications*. IEEE Workshop on Signal Processing Systems (SiPS), Boston, 1998.

[10] L.F.Lamel, J.L.Gauvain, M.Eskenazi, *BREF, a large vocabulary spoken corpus for French*, Proc. EUROSPEECH-91, Genoa, Italie, 1991.

[11] G. Baudoin, F. El-chami**,** "Corpus based very low bit rate speech coding", *ICASSP'03 International Conference on Acoustics, Speech & Signal Processing,* 6 - 10 Avril 2003 Hong Kong.

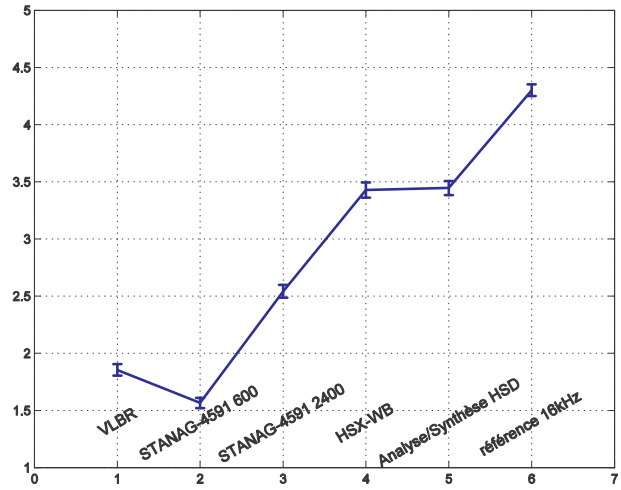

FIG. 3: Courbe test MOS, Critère: qualité globale

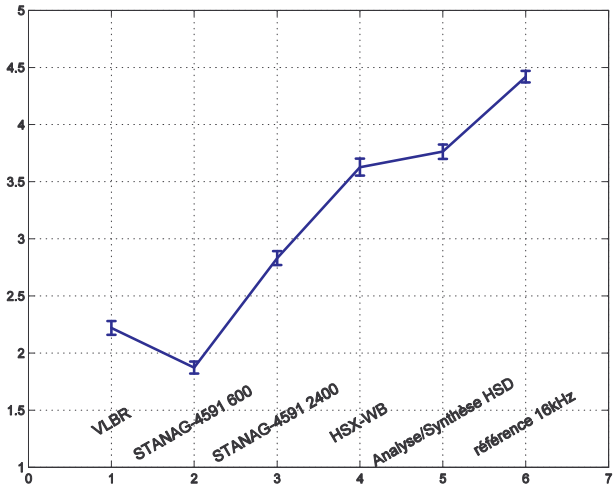

FIG. 4: Courbe test MOS, Critère: intelligibilité

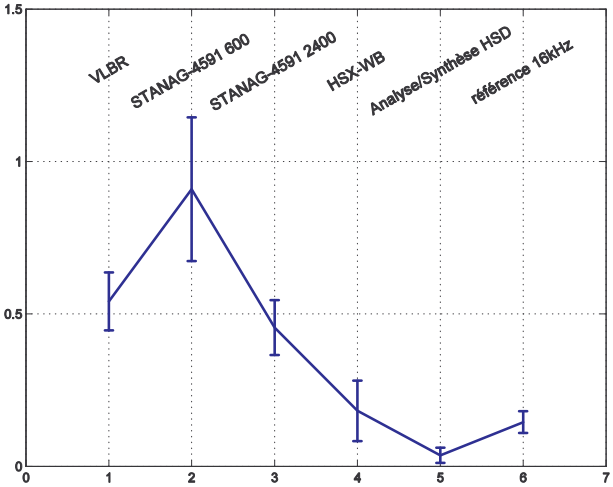

FIG. 5: Courbe relative au nombre d'erreur de transcription de phonèmes par phrase

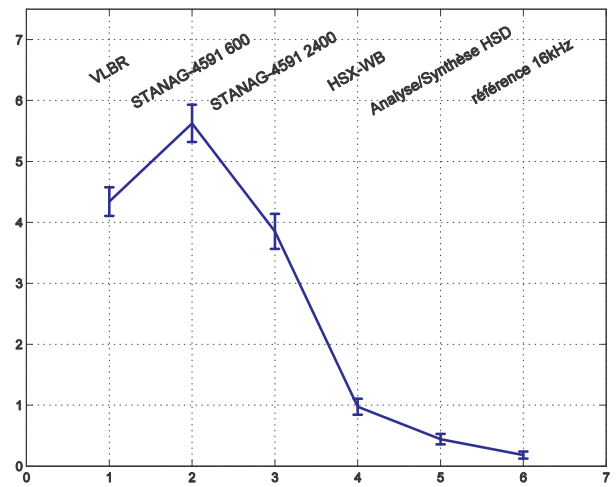

FIG. 6: Courbe relative au nombre d'erreur de reconnaissance par séquence de test.