Influence de la modélisation spectrale sur les performances d'un système de conversion de voix

Taoufik En-Najjary [1], Olivier Rosec [1] et Thierry Chonavel [2]

[1] France Telecom R&D DIH/IPS, 2 avenue Pierre Marzin, 22 307 Lannion Cedex

{taoufik.ennajjary, olivier.rosec}@francetelecom.com

[2] ENST Bretagne, département SC, BP 832, 29285 Brest Cedex

Thierry.Chonavel@enst-bretagne.fr

Résumé – La conversion de voix est une technique qui consiste à modifier le signal de parole d'un locuteur de référence appelé aussi locuteur source, d'une façon telle qu'il semble, à l'écoute, être prononcé par le locuteur désiré. Dans ce papier, nous étudions l'influence de la modélisation spectrale sur la qualité de la conversion du timbre. Nous comparons, dans le cadre de la conversion par GMM, les modélisations par cepstre discret et par paramètres LSF. Des tests objectifs montrent que l'utilisation des paramètres LSF conduit à de meilleurs résultats de conversion.

Abstract *– Voice conversion is a technique which aims to modify the speech signal uttered by a so called source speaker so that it is perceived as if it was uttered by the desired speaker. In this paper, we study the influence of the spectral modelling on the quality of timbre conversion. In the framework of GMM-based conversion, we compare discrete cepstrum and LSF parameter modelling. Objective tests show that the use of LSF parameters leads to better conversion results.*

1. Introduction

La conversion de voix est une technique qui consiste à modifier le signal de parole d'un locuteur de référence appelé aussi locuteur source, d'une façon telle qu'il semble, à l'écoute, être prononcé par le locuteur désiré, dénommé locuteur cible. Pour cela, un apprentissage est mené sur un enregistrement restreint des locuteurs source et cible afin de déterminer une fonction de transformation qui sera ensuite appliquée au locuteur de référence et réalisera ainsi la conversion de voix. Cette technologie, appliquée dans le cadre de la synthèse de la parole par corpus, offre un moyen simple de diversifier les voix de synthèse en limitant les opérations d'enregistrement et surtout de vérification d'un corpus de parole entier.

Les performances d'un système de conversion de voix dépendent de deux facteurs : d'une part la nature des paramètres transformés et d'autre part la technique de conversion utilisée. La plupart des travaux menés dans le domaine traitent essentiellement de la transformation du timbre et de nombreuses méthodes ont été développées [1, 2, 3, 4, 5]. A ces modifications d'enveloppe spectrale sont généralement associées des modifications de pitch allant d'une simple mise à l'échelle à une véritable prédiction de contours de pitch [6].

Dans cet article, nous nous restreignons à la modification du timbre par GMM (Gaussian Mixture Model). Cette technique a été appliquée indépendamment à la modification du cepstre discret [4] et des paramètres LSF [5], sans qu'aucune comparaison véritable n'ait été effectuée. Nous nous proposons donc d'analyser les performances accessibles par ces deux types de modélisation, en nous appuyant notamment sur une mesure de distorsion spectrale.

Dans la section 2, nous décrivons succinctement le principe de la conversion par GMM dans le cadre d'une procédure d'analyse-synthèse basée sur le modèle HNM [4]. Les performances obtenues par ces deux types de paramètres sont détaillées en section 3 et une conclusion est faite en section 4.

2. Transformation du timbre

2.1 Analyse

Le modèle HNM [4] a montré son utilité en synthèse de la parole, dans la mesure où il permet d'effectuer des modifications prosodiques (nécessaires pour adapter le rythme et la hauteur de voix de la parole synthétique), voire spectrales, de haute qualité. Etant donnée une trame de parole voisée, ce modèle sépare le spectre en deux parties délimitées par une fréquence de coupure Fc appelée fréquence maximale de voisement. La partie basse du spectre est approximée par une somme de sinusoïdes harmoniquement reliées. Au-delà de Fc, le spectre est modélisé par un filtre LPC excité par un bruit blanc gaussien.

Ce modèle HNM a également été utilisé dans le cadre de la conversion de voix. Le mode d'implémentation proposé dans [4] suppose une fréquence maximale de voisement constante et détermine une fonction de conversion sur [0,Fc]. Des filtres correctifs permettent ensuite de transformer la partie bruitée.

Dans cet article, nous déterminons une fonction de conversion valable sur l'ensemble du spectre. Pour cela, nous utilisons un modèle harmonique. Les amplitudes a_k obtenues lors de cette décomposition harmonique sont

utilisées, comme dans le cas de l'analyse HNM, pour estimer les paramètres de l'enveloppe spectrale.

2.1.1 Cepstre discret régularisé [7]

Les coefficients du cepstre discret $c = [c_1 c_2 \ldots c_p]^T$ où p est l'ordre du cepstre sont obtenus par la minimisation du critère des moindres carrés régularisé suivant :

$$\varepsilon_r = \sum_{k=1}^{L} \left\| \log a_k - \log|S(f_k;c)| \right\|^2 + \lambda R[S(f;c)], \quad (1)$$

où l'amplitude du spectre $S(f;c)$ est reliée aux coefficients cepstraux par

$$\log|S(f;c)| = c_0 + 2\sum_{i=1}^{p} c_i \cos(2\pi f i) \quad (2)$$

et où $R[S(f;c)]$ est une fonction destinée à pénaliser les variations rapides de l'enveloppe spectrale, λ étant le paramètre de régularisation associé.

2.1.2 Les coefficients LSF

Le spectre est tout d'abord sur-échantillonné en utilisant une interpolation cubique [8]. Ensuite, à partir de la densité spectrale de puissance, les coefficients d'autocorrélation sont calculés par FFT inverse. Puis, les coefficients LPC sont obtenus par application de l'algorithme de Levinson-Durbin. Ces derniers sont finalement convertis en coefficients LSF [9,10].

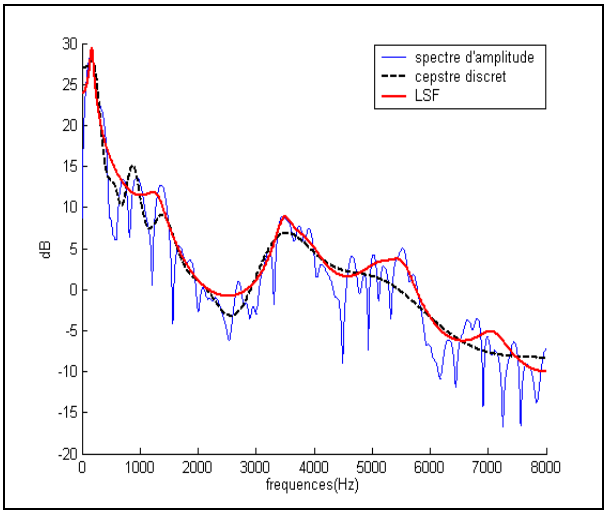

FIG. 1 : Exemple de spectre d'amplitude d'une trame voisé, et d'enveloppes spectrales obtenues avec le cepstre discret et les LSF

2.2 Conversion du timbre par GMM

Formellement, la densité de probabilité d'une variable aléatoire z suivant un modèle GMM d'ordre Q s'écrit :

$$p(z) = \sum_{i=1}^{Q} \alpha_i N(z;\mu_i;\Sigma_i), \sum_{i=1}^{Q} \alpha_i = 1, \ \alpha_i \geq 0, \quad (3)$$

où $N(z;\mu,\Sigma)$ est la densité de probabilité de la loi normale de moyenne μ et de matrice de covariance Σ, et où les α_i sont les coefficients du mélange (α_i est la probabilité *a priori* que z soit généré par la $i^{ème}$ composante gaussienne). Les paramètres GMM (α,μ,Σ) de la densité $p(z)$ sont estimés par un algorithme EM [11] initialisé à l'aide d'une technique classique de quantification vectorielle.

Soient $x = [x_1 x_2 \ldots x_N]$ une suite de vecteurs spectraux caractérisant une séquence de parole prononcée par le locuteur source et $y = [y_1 y_2 \ldots y_N]$ la séquence des vecteurs spectraux correspondant au même contenu acoustique prononcé par le locuteur cible. Le but est alors d'estimer une fonction F qui permet de faire le lien entre les vecteurs source et cible. Pour cela, deux méthodes employant le modèle GMM ont été développées.

Dans la première [4], le modèle GMM modélise uniquement la distribution de probabilité de la source. La fonction de transformation proposée est exprimée de la forme suivante :

$$F(x_i) = \sum_{q=1}^{M} P(q|x_i)\left[\nu_q + \Gamma_q \Sigma_q^{-1}(x_i - \mu_q)\right]. \quad (4)$$

Les paramètres ν_q et Γ_q sont déterminés en minimisant la distance quadratique moyenne entre les vecteurs de référence transformés et les vecteurs cible donnée par :

$$\varepsilon = \sum_{i=1}^{N} \|y_i - F(x_i)\|^2, \quad (5)$$

où x_i et y_i désignent respectivement les vecteurs sources et cibles préalablement alignés par DTW.

Dans une deuxième étude [5], les vecteurs de la source et de la cible sont combinés dans $z = [x^T y^T]^T$, puis on estime la distribution jointe des paramètres des locuteurs source et cible $p(x,y)$. La fonction de conversion est donnée par l'espérance conditionnelle :

$$\begin{aligned} \hat{y}_{CE} &= F(x) = E[y|x] \\ &= \sum_{i=1}^{Q} h_i(x)\left[\mu_i^y + \Sigma_i^{xy}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)\right] \end{aligned} \quad (6)$$

où $$h_i(x) = \frac{\alpha_i N(x;\mu_i^x,\Sigma_i^{xx})}{\sum_{j=1}^{Q} \alpha_i N(x;\mu_i^x,\Sigma_i^{xx})} \quad (7)$$

est la probabilité *a posteriori* que x soit généré par la $i^{éme}$ gaussienne,

avec $$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \quad \text{et} \quad \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}.$$

D'un point de vue théorique les deux approches conduisent au même estimateur. Cependant, la deuxième méthode est plus stable numériquement, notamment lorsque l'ordre des modèles GMM augmente, et offre par ailleurs une complexité beaucoup moins élevée que la première [5].

3. Résultats

3.1 Mise en oeuvre

Dans les expériences décrites ci-dessous, nous employons quatre bases de données de parole échantillonnées à 16kHz et correspondant à deux voix d'homme et deux voix de femme. Chaque base acoustique est constituée d'environ 1200 diphones, couvrant ainsi l'ensemble des diphones de la langue française.

Les données sont analysées de manière pitch synchrone et pour les zones non-voisées, des trames de 10 ms sont utilisées. Les vecteurs LSF ou cepstraux d'ordre 16 sont obtenus conformément aux procédures d'analyse décrites précédemment. Puis, un appariement des trames acoustiques par alignement dynamique est effectué. Lors de cette opération, des contraintes sont introduites de manière à respecter les marques de segmentation en phones et diphones. Notons de plus que, dans cette étude, seules sont prises en compte les trames voisées. Au final, les bases d'apprentissage contiennent environ 25000 vecteurs pour les voix d'hommes et de l'ordre de 15000 pour les voix de femmes.

Pour l'application de la conversion de voix dans le cadre d'un système de synthèse par concaténation, le but est de convertir les enveloppes spectrales de toute la base de référence. Dans le cas des bases de diphones étudiées ici, il est nécessaire de disposer d'un volume de données suffisant, pour que l'apprentissage puisse se faire de manière correcte. C'est pourquoi, l'apprentissage est effectué sur l'ensemble de la base de diphones.

L'apprentissage est mené en utilisant des matrices de covariance pleines. Pour empêcher des singularités, une petite valeur a été ajoutée aux éléments diagonaux des matrices de covariance après chaque itération. Pour chaque base d'apprentissage, nous faisons varier le nombre de composantes du mélange en considérant les puissances de 2 comprises entre 8 et 128. Lors de ces expériences, 20 itérations sont jugées suffisantes pour atteindre la convergence de l'algorithme EM.

3.2 Evaluation

Afin d'évaluer objectivement les conversions effectuées, il est nécessaire de définir une mesure qui puisse permettre de comparer les deux modélisations. Pour cela, nous utilisons la mesure de distorsion spectrale moyenne normalisée suivante :

$$SD = \frac{\sum_{n=1}^{N} \left\| P_{dB}(y_n) - P_{dB}(\hat{y}_n) \right\|_2}{\sum_{n=1}^{N} \left\| P_{dB}(y_n) - P_{dB}(x_n) \right\|_2} \qquad (8)$$

où $P_{dB}(x)$ désigne l'enveloppe spectrale issue de x exprimée en dB.

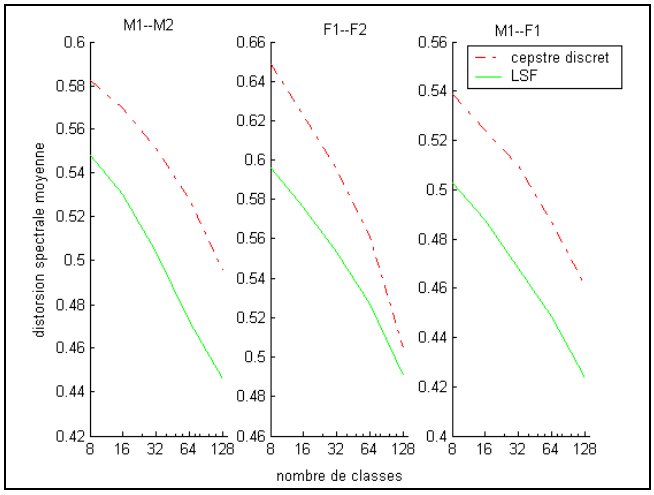

FIG. 2 : Distorsion spectrale moyenne entre enveloppes cible et convertie pour les transformations homme-homme, femme-femme et homme-femme : cepstre en traits mixtes, LSF en traits pleins.

La figure 2 présente les résultats obtenus avec le cepstre discret et les paramètres LSF en faisant varier l'ordre du mélange de gaussiennes de 8 à 128. Les courbes ont des allures similaires et font apparaître une diminution de la distorsion moyenne lorsque le nombre de classes augmente. On note cependant une erreur supérieure dans le cas du cepstre discret, pour chacune des 3 configurations de conversion testées. A titre d'exemple, pour un mélange d'ordre 128, la réduction varie de 2,8% pour une conversion femme-femme à 9,9% pour une conversion homme-homme. Ce résultat peut s'expliquer par le fait que les LSF sont reliés aux formants et parviennent ainsi à mieux capturer et modifier les informations pertinentes de l'enveloppe spectrale de signaux de parole.

4. Conclusion

Les comparaisons menées dans cette étude ont montré que la modélisation par les paramètres LSF conduit à des résultats de conversion meilleurs que la modélisation par le cepstre discret. D'autres expériences devront être menées, notamment sur des corpus de parole plus importants, afin de vérifier la validité de ces résultats.

Il serait également pertinent de mesurer, sur le plan de la perception, la différence entre les deux modélisations, par le biais de tests subjectifs. Pour que cette évaluation subjective puisse être faite, il est cependant nécessaire de pouvoir modifier de manière conjointe le timbre le pitch et le rythme d'élocution. L'intégration de ces différentes fonctionnalités dans un système de conversion complet fait partie de nos travaux à venir.

Références

[1] Mr. Abe, S. Nakamura, K Shikano and H. Kuwabara, "Voice conversion through vector quantization", Proceedings of IEEE ICASSP, pp. 655-658, 1988.

[2] H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique", Speech Communications, vol. 11, pp. 175-187, 1995.

[3] N. Iwahashi and Y. Sagisaka, "Speech spectrum transformation by speaker interpolation", Proceedings of IEEE ICASSP, 1994.

[4] Y. Stylianou, "Harmonic plus Noise Model for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, Ecole Nationale Supérieure des Telecommunications, Paris, France, 1996.

[5] A. Kain and Mr. Macon, "Text-to-speech voice adaptation from sparse training dated", Proceedings of ICSLP 1998.

[6] T. En-Najjary, O. Rosec and T. Chonavel, "A new method for pitch prediction and its application in voice conversion", Proceedings of Eurospeech 2003.

[7] O. Cappé, J. Laroche and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points", IEEE ASSP Workshop on application of signal processing to audio and acoustics, Mohong, 1995.

[8] M. Unser, A. Adroubi, and M. Eden " B-spilne signal processing ", *IEEE Transactions on Speech and Audio Processing vol. 41,* 2 (February 1993), 821-833.

[9] R. J. Mc Aulay, and T. F. Quatieri, "Sinusoidal coding", In Speech coding and synthesis, W.B. Kleijn and K.K. Paliwal, Eds. Elsevier Science, Amsterdam, Holland, 1995, ch. 4, pp. 121-173.

[10] F. Itakura, "Frequency line spectrum representation of linear predictive coefficient of speech signals", Journal of the Acoustical Society of America vol.57, S35(A),1975.

[11] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society Serie B, vol. 39, pp. 1-38, 1977.