Estimation d'enveloppes spectrales contraintes temporellement pour la conversion de voix

Elizabeth Godoy, Olivier Rosec, Thierry Chonavel

Estimation d'enveloppes spectrales contraintes temporellement pour la conversion de voix

Elizabeth Godoy [1], *Olivier Rosec*[1], *Thierry Chonavel* [2]

[1] Orange Labs, Lannion, France
[2] Telecom Bretagne, Signal & Communication Department, Brest, France
{elizabeth.godo, olivier.rosec}@orange-ftgroup.com,thierry.chonavel@telecom-bretagne.eu

ABSTRACT

This paper presents a new approach to estimating the speech spectral envelope that is adapted for Voice Conversion (VC). In particular, we represent the spectral envelope as a sum of peaks that evolve smoothly in time, within a phoneme. We highlight important properties of our proposed spectral envelope estimation and illustrate its potential for use in a VC context. We analyse natural speech using the proposed methods and we compare results with those from a more traditional frame-by-frame cepstrum-based analysis. Subjective comparisons of synthesized speech quality, as well as implications of this work in future research are also discussed.
Keywords: spectral envelope, voice conversion

1. INTRODUCTION

L'estimation de l'enveloppe spectrale est un sujet récurrent en traitement de la parole de part l'importance que revêt cette information dans des applications telles que le codage de la parole ou encore la transformation et la conversion de voix. La pratique courante consiste à effectuer cette estimation trame par trame en considérant chacune de ces trames isolément. Pour ce faire, des méthodes à base de prédiction linéaire ou de cepstre sont couramment employées en vue de générer des paramètres d'intérêt tels que les LSF (Line Spectral Frequencies) ou les coefficients cepstraux [1]. Cette stratégie est adaptée dans de nombreux contextes applicatifs, mais elle trouve néanmoins ses limites dans des applications nécessitant des transformations de l'enveloppe spectrale variant dans le temps [2]. C'est par exemple le cas en conversion de voix où l'on cherche à modifier l'enveloppe spectrale d'un locuteur source de telle sorte que le signal résultant semble avoir été prononcé par le locuteur cible désiré. L'un des problèmes cruciaux dans un tel cadre est de préserver la cohérence temporelle du signal converti. Le processus de conversion de voix est classiquement découpé en trois étapes : premièrement l'analyse visant à extraire des paramètres d'intérêt (*e.g.* l'enveloppe spectrale), deuxièmement l'apprentissage d'une fonction de conversion établissant le lien entre les espaces acoustiques des locuteurs source et cible et troisièmement la transformation proprement dite. Des travaux récents ont porté sur la restitution de trajectoires continues via des méthodes d'interpolation [3]. Ces travaux se concentrent uniquement sur les étapes d'apprentissage et de transformation sans remettre en cause la phase d'analyse.

Dans cet article, nous suggérons d'introduire des contraintes lors de l'analyse de façon à extraire des paramètres d'enveloppe spectrale variant continûment et dont l'évolution temporelle peut être facilement contrôlée lors de l'étape de conversion. Plus précisément, sur la base d'une représentation de l'enveloppe spectrale par somme de pics gaussiens [4], nous prenons en compte les dépendances temporelles entre trames au sein de chaque phone afin de modéliser des trajectoires de pics spectraux. Une telle représentation revêt plusieurs propriétés intéressantes. Tout d'abord, les pics spectraux obtenus sont fortement liés aux formants, ce qui les rend pertinent du point de vue de la perception. En second lieu, cette représentation offre une grande flexibilité dans un contexte de modification de signaux de parole, car elle permet un contrôle fin de la position de la largeur et de l'amplitude de chacun des pics spectraux. Enfin, la modélisation de trajectoires spectrales est cruciale pour pouvoir rendre compte de l'évolution des résonnances du conduit vocal, celle-ci se faisant généralement de manière lisse pour la plupart des sons voisés (ceux-là mêmes qui portent une grande partie de l'identité vocale). En résumé, nous proposons une nouvelle méthode d'estimation de l'enveloppe spectrale adaptée à la conversion de voix en ce sens qu'elle fournit des paramètres liés à la perception, localisés en fréquences et dont l'évolution temporelle peut être contrôlée.

Dans cet article, nous examinons également le potentiel de notre méthode d'analyse à discriminer des événements acoustiques. Ceci est particulièrement important en conversion de voix afin de faire ressortir des différences de timbre pertinentes entre deux locuteurs. Pour cela nous suggérons une métrique basée sur la distance entre les amplitudes des pics spectraux.

L'article est structuré comme suit. La section 2 présente le modèle de pics spectraux utilisé et la méthode d'analyse proposée. Dans la section 3, nous analysons le potentiel de l'approche à effectuer une analyse-synthèse de haute qualité et examinons également ses capacités en terme de discrimation d'événements acoustiques. La section 4 conclut nos propos et fournit des perspectives à cette étude.

2. ESTIMATION DE L'ENVELOPPE SPECTRALE CONTRAINTE TEMPORELLEMENT

2.1. Modélisation des pics spectraux

Notre modélisation est basée sur l'hypothèse que l'amplitude de l'enveloppe spectrale à la fréquence *f*, $S_i(f)$, de la trame d'indice *i* s'écrit comme la somme de M_i gaussiennes :

$$S_i(f) = \sum_{m=1}^{M_i} a_m^i N(f; \mu_m^i, \sigma_m^i)$$

Les paramètres $\{a_m^i, \mu_m^i, \sigma_m^i\}$ désignent respectivement l'amplitude, la position et la variance du pic d'indice *m*. Cette modélisation est à rapprocher de celle proposée dans [4]. Mais à la difference des travaux reportés dans [4] et [5], notre analyse est basée sur la transformée de Fourier Discrète (TFD) plutôt que sur des versions de l'enveloppe spectrale obtenue par lissage spectral, prédiction linéaire ou modélisation cepstrale qui peuvent d'emblée altérer les caractéristiques spectrales du signal. De plus, notre méthode d'estimation des paramètres diffère radicalement de la méthode basée sur l'algorithme EM décrite dans [4]. Nous suggérons ici d'estimer les positions et amplitudes des pics par le biais d'un algorithme de détection de pics utilisant un masque fréquentiel dont la taille est ajustée en fonction du support fréquentiel considéré et de la fréquence fondamentale. Cette taille variable permet une bonne résolution fréquentielle sur les basses fréquences (particulièrement importante sur le plan de la perception) tout en évitant de modéliser les harmoniques elles-mêmes. Le nombre de pics détectés dépend de ce masque est varie donc selon la trame analysée. Nous autorisons cependant un nombre maximum de 20 pics par trame. Une fois que les positions et amplitudes des pics ont été déterminées, nous calculons la variance de chaque pic en considérant la partie du spectre au voisinage immédiat du pic en question. De cette manière nous évitons les interférences entre pics qui pourraient conduire à des sur-estimation des amplitudes. Plus précisément, entre deux pics nous déterminons le point du spectre généré par les deux pics de manière équiprobable. Ce point permet de calculer les variances de l'échantillon à droite et à gauche respectivement des pics de gauche et droite. La variance d'un pic est alors la moyenne de ces variances à gauche et à droite. Un soin tout particulier doit être accordé au premier pic spectral pour ne pas que sa variance soit influencée par les valeurs de la TFD en deçà de f0. La variance est ainsi calculée en assurant une interpolation lisse entre les amplitudes des premier et deuxième pics.

La Figure 1 montre les spectres d'amplitude obtenus par TFD, par notre méthode et par une modélisation par cepstre discret d'ordre 40 et calculé sur une échelle Bark conformément à [6]. Les deux approches permettent de capturer la forme globale de l'enveloppe spectrale, la modélisation cepstrale se révélant davantage apte à modéliser les détails. Pour pouvoir apprécier les différences entre les deux méthodes, il convient tout d'abord de noter que la méthode du cepstre discret n'introduit pas de fortes contraintes sur la forme de l'enveloppe spectrale. Notre modélisation cherche quant à elle à capturer les pics spectraux principaux puis à estimer des trajectoires entre ces pics. Ainsi, nous sacrifions donc une certaine précision à l'échelle de la trame, dans le but de faire émerger une structure spectrale à l'échelle phonémique.

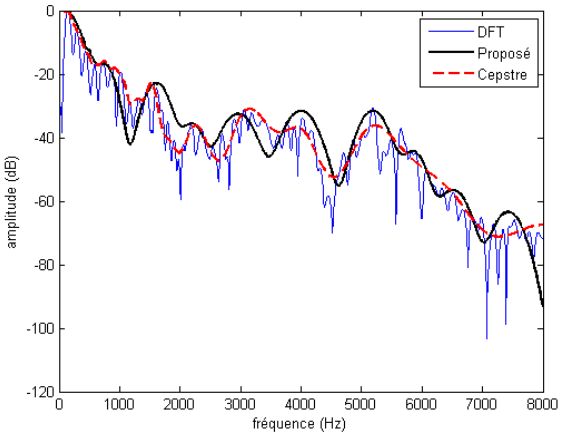

Figure 1: Spectre d'amplitude d'une trame de parole obtenu par TFD (en bleu) et envelopes spectrales estimées par la méthode du cepstre (en pointillé rouge) discret et la méthode proposée (en noir).

2.2. Suivi de l'évolution des pics spectraux à l'échelle phonémique

Un point clé de notre méthode est de contraindre le processus d'estimation de l'enveloppe spectrale par la prise en compte explicite de l'évolution temporelle des pics spectraux à l'échelle d'un phone. Pour cela, nous supposons qu'une segmentation associée au signal de parole est disponible. L'hypothèse sous-jacente est que les pics spectraux ne doivent pas varier trop drastiquement d'une trame à l'autre pour un phone donné. Nous introduisons donc une étape de suivi des pics spectraux préalablement déterminés.

Nous nous intéressons tout d'abord à la partie stable d'un phone. Cette zone est définie pour la majorité des phonèmes (à l'exception toutefois des consonnes plosives), et peut être localisée au voisinage du milieu du phone. Nous définissons alors arbitrairement cette partie stable comme le triplet de trames constitué de la trame centrale d'un phonème et de ses voisines immédiates. Notons que si le phone est déclaré voisé nous effectuons une analyse pitch-synchrone ; dans le cas contraire l'analyse est faite avec un pas de *10* ms. L'intérêt de cette partie stable est que le signal peut y être considéré comme stationnaire et que les effets de coarticulation y sont minimums. L'analyse y est donc a priori plus fiable. Nous suggérons alors d'exploiter ces propriétés afin d'obtenir des points d'ancrage pertinents pour guider notre

procédure d'analyse. Nous commençons alors par une analyse de ces trois trames prises individuellement. Nous sélectionnons les deux trames les plus proches sur la base d'une distance euclidienne entre spectres d'amplitude déduits des représentations spectrales et nous alignons les pics de ces deux trames. Cet alignement est réalisé en minimisant localement la différence entre les positions fréquentielles des pics. La moyenne des paramètres de ces pics alignés définit alors les paramètres de la trame centrale du phone. Ensuite, nous faisons de part et d'autre de cette trame stable une analyse de proche en proche jusqu'aux frontières de phone. Pour une trame donnée, cette analyse est contrainte de manière à ne sélectionner que les pics suffisamment proches de ceux détectés pour la trame adjacente précédemment analysée. Cette méthode d'analyse assure ainsi une évolution graduelle des paramètres à l'échelle d'un phone, tout en privilégiant les parties sur lesquelles l'estimation de l'enveloppe spectrale est la plus fiable, *i.e.* les parties stables.

La figure 2 montre une vue 3D de la séquence de spectres d'amplitude déduits des enveloppes spectrales estimées sur un exemple de réalisation acoustique du phonème 'A'. Notons que nous n'avons représenté que la partie du spectre comprise entre 0 et 4 kHz. Sur cette figure nous observons que la modélisation proposée fournit une représentation temps-fréquence lisse permettant clairement d'identifier des trajectoires spectrales sur l'ensemble du phone. A l'inverse, la modélisation par cepstre discret met en évidence des variations sporadiques et discontinues de l'enveloppe spectrale.

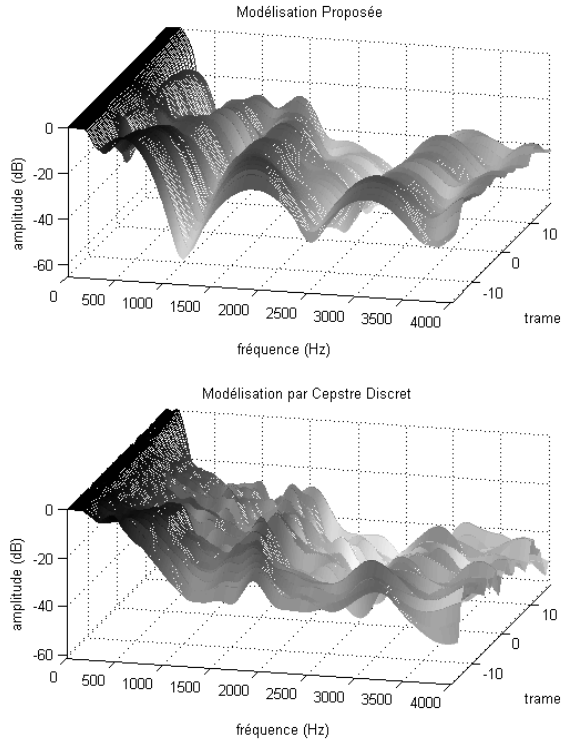

Figure 2: 3-D plots of the amplitude-normalized spectral envelopes from an instance of the phoneme 'A.' Frames are shown in sequence wrt the phoneme center at 0.

3. ILLUSTRATION DU POTENTIEL DE LA METHODE PROPOSEE

3.1. Qualité accessible en analyse-synthèse

Dans cette section nous appliquons notre nouvelle méthode d'analyse de l'enveloppe spectrale dans un contexte d'analyse-synthèse. Notre expérience porte sur 4 corpus de parole (i.e. 4 voix, 2 féminines et 2 masculines) utilisés dans le système de synthèse vocale Baratinoo de France Télécom. Ces corpus ont été enregistrés en studio par des locuteurs professionnels. Les signaux sont échantillonnés à 16 kHz, étiquetés, segmentés en phones et pitch-marqués. Pour l'analyse et la resynthèse des signaux nous utilisons un modèle HNM (Harmonic plus Noise Model) [6]. Dans cette expérience, nous avons fixé la fréquence maximale de voisement des trames voisées à 8 kHz. Les amplitudes des harmoniques sont obtenues par échantillonnage de l'enveloppe spectrale. Les phases utilisées à la synthèse sont celle estimées lors de l'analyse HNM. Cette méthodologie nous permet ainsi de comparer uniquement l'effet de l'estimation de l'enveloppe spectrale sur le résultat de la synthèse. Sur cette base nous avons donc procédé à un test d'écoute informel dont l'objectif était de comparer la méthode d'estimation proposée à celle du cepstre discret. Ce test d'écoute a montré une légère dégradation de la qualité en analyse-synthèse. Une telle dégradation était prévisible dans la mesure où, comme mentionné précédemment, l'enveloppe estimée par notre méthode est moins précise. Des ajustements de phases seraient alors nécessaires pour rendre plus cohérente cette information de phase avec le spectre d'amplitude estimé par notre méthode. Des tests supplémentaires faisant intervenir des modifications de timbre variant dans le temps sont également nécessaires pour pouvoir mieux apprécier le potentiel de notre méthode en tant que brique d'analyse-modification-synthèse. Un objectif est bien entendu de tester cette approche dans le cadre de la conversion de voix.

3.2. Caractérisation des espaces acoustiques

Comme mentionné précédemment, un point clé en conversion de voix est de déterminer le lien entre les espaces acoustiques des locuteurs source et cible. Bien entendu, pour qu'un tel lien puisse être établi, il faut pouvoir disposer d'un espace de représentation adéquat, c'est-à-dire permettant de regrouper des événements acoustiques perçus comme similaires et donc également capable d'offrir une discrimination claire entre sons perçus différemment.

L'une des motivations de ce travail était qu'en imposant une continuité temporelle dès l'étape d'analyse, nous favoriserons de fait le regroupement de trames adjacentes au sein de classes acoustiques similaires, tout en garantissant une évolution temporelle suffisamment lisse du degré d'appartenance à ces classes acoustiques. Ainsi, indépendamment de la méthode utilisée pour la conversion de voix (quantification vectorielle, mélange de

gaussiennes, réseaux de neurones, …) il apparaît crucial de mesurer les capacités de classification et de discrimination des modèles sous-jacents.

Pour cela des métriques adaptées doivent être mises en œuvre, car c'est au final ces métriques qui seront utilisées dans les étapes d'apprentissage voire de conversion elle-même. Dans le cas d'une modélisation cepstrale, la distance classiquement employée est la distance euclidienne entre vecteurs de coefficients cepstraux [6]. Notre représentation est de taille variable, ce qui proscrit l'utilisation de mesures euclidiennes pour comparer deux trames acoustiques. C'est pourquoi nous proposons une mesure adaptée à notre espace de représentation de l'enveloppe spectrale.

Le calcul de la distance entre deux trames se fait en deux étapes. La première étape vise à aligner les pics spectraux des deux trames de façon à disposer d'un espace de représentation commun. Cet alignement est réalisé via l'algorithme mentionné en section 2.2. Dans un second temps, la distance euclidienne entre les vecteurs composés des log-amplitudes des pics alignés est calculée.

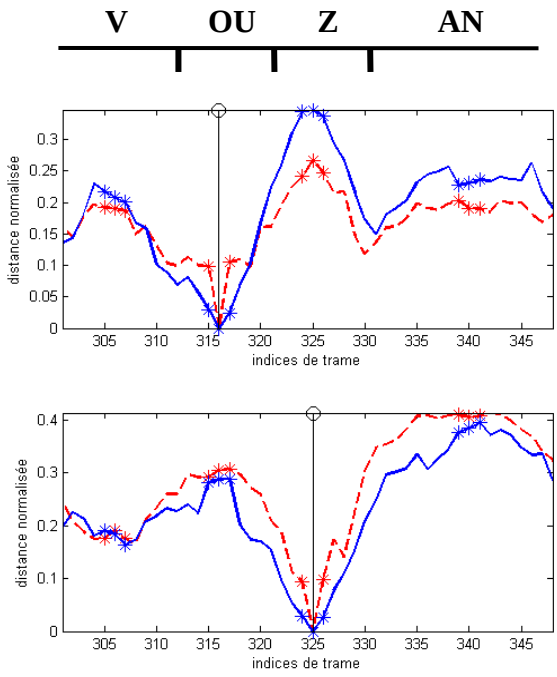

Figure 3: Distances normalisées pour la séquence phonétique 'V-OU-Z-AN'. Les trames de référence sont indiquées par une barre verticale et les trames stables sont marquées par une astérisque. La métrique proposée (en bleu) est comparée à la distance euclidienne entre coeeficients cepstraux (courbe rouge en tirets).

La Figure 3 montre un exemple d'utilisation de cette métrique pour comparer une trame de référence à l'ensemble des trames correspondant à la séquence phonétique 'V-OU-Z-AN'. La distance proposée est comparée à une distance cepstrale classique. Notons que ces deux distances sont normalisées par l'énergie de la trame de référence. Les courbes obtenues font apparaître un comportement similaire dans les deux cas. Des travaux futurs consisteront à comparer de manière plus approfondie ces deux types de métriques, et ceci notamment dans un contexte d'apprentissage pour la conversion de voix.

4. Conclusions et Perspectives

Dans cet article, nous avons proposé une nouvelle méthode d'estimation de l'enveloppe spectrale de signaux de parole basée sur une modélisation et un suivi de l'évolution temporelle de pics spectraux. Nous avons illustré quelques propriétés intéressantes de cette méthode, notamment le fait qu'elle fournisse une enveloppe spectrale évoluant de manière lisse et régulière à l'échelle d'un phone. Nous avons également introduit une métrique permettant la comparaison de deux trames sur la base de la représentation proposée et comparé notre analyse à une approche cepstrale plus traditionnelle.

Les avantages de notre méthode d'estimation de l'enveloppe spectrale devraient apparaître de manière plus évidente dans un contexte de conversion de voix. En particulier, nos prochains travaux viseront à explorer les capacités de la méthode proposée à modéliser les espaces acoustiques des locuteurs source et cible.

Références

[1] Turk, O., and Arslan, L., "Robust processing techniques for voice conversion," Computer Speech and Language 20, 2006, 441-467.

[2] *Springer Handbook of Speech Processing*, Editors Benesty, J., Sondhi M. & Huang Y., Springer, 2008.

[3] Nguyen, B. and Akagi, M., "Spectral Modification for Voice gender Conversion Using Temporal Decomposition," Journal of Signal Processing, Vol. 11, No. 4, pp. 333-336, July 2007.

[4] Zolfaghari, P., Watanabe, S., Nakamura, A. and Katagiri, S. "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," in Proc of ICASSP '04, pp. 553-556.

[5] Nguyen, B., "Studies on Spectral Modification in Voice Transformation," Ph.D. diss, Japan Advanced Institute of Science and Technology, March 2009.

[6] Stylianou, Y. "Harmonic Plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. diss., ENST, Paris, France, Jan. 1996.