

Codage de la parole en bande élargie basé sur la structure MBE

A. Amodio et G. Feng

Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/ Université Stendhal
B.P. 25
38040 Grenoble Cedex 09, France

RÉSUMÉ

Cet article présente l'adaptation à la bande élargie d'un codeur de parole MBE, initialement développé sur la bande téléphonique. Les contraintes de qualité et de débit, ainsi que les caractéristiques des signaux étant différentes sur la bande élargie, la structure du codeur est reconsidérée. Plusieurs améliorations, dont certaines ont déjà été proposées en bande téléphonique, comme la classification phonétique des trames ou la modélisation multi-harmonique du spectre sont apportées. Dans le but d'atteindre une plus grande qualité, particulièrement pour les voix aiguës, nous proposons aussi de modéliser et synthétiser l'erreur entre le spectre original et le spectre modélisé avec la technique MBE classique.

1 Introduction

Actuellement, un intérêt croissant est porté à la compression des signaux de parole sur la bande élargie. En effet, la bande élargie (50-7000Hz) permet, par rapport à la bande téléphonique (300-3400Hz) de transmettre une information plus riche car contenant le naturel de la voix, la sensation de présence du locuteur et de l'ambiance où il se trouve. Sur cette bande, plusieurs codeurs utilisant des stratégies différentes ont été développés pour des débits de 32 kbits/s et de qualité identique à celle du codeur G.722 normalisé en 1986 par l'ITU (ex CCITT) de débit 64 kbits/s et servant de référence sur cette bande. En même temps, la technique de codage MBE (Multi-Band Excitation), proposée par Griffin en 1987 [1], continue de prouver son efficacité et son intérêt sur la bande téléphonique, comme le montrent les nombreuses recherches menées, concernant par exemple la réduction du débit. C'est dans ce contexte particulier que nous nous proposons d'élaborer un codeur en bande élargie basé sur la technique de codage MBE, notre but ultime étant de construire un codeur de débit de 16 à 24 kbits/s, possédant une qualité comparable à celle du codeur G.722, et dans lequel sera intégré un modèle d'audition. Cet article présente la première étape de notre travail qui consiste en la détermination de la structure globale du codeur, c'est à dire en l'analyse des différents modules et algorithmes intervenant dans le codeur MBE et à l'analyse

ABSTRACT

This paper deals with the adaptation to wideband of the MBE coder which was initially developed for the telephone band. As the constraints of quality and bit rate for a wideband and a telephone band coder are different, and as the signal characteristics on these two bands are different too, we must reconsider the coder structure. Several improvements are proposed, some of which were already proposed for the telephone band such as the phonetic classification of the frames or the multi-harmonic modelling of the spectrum. We also propose in order to reach a good quality, especially for high frequency voices, to model and synthesize, as part of the signal, the initial error between the synthetic and original spectra.

de leur adéquation sur la bande élargie. Après un bref rappel sur la technique de codage MBE, nous exposerons les modifications, les choix et les améliorations qui ont été apportés et qui nous ont été dictés par l'analyse des principaux problèmes rencontrés lors de l'adaptation de cette technique sur la bande élargie.

2 Le codeur MBE

L'intérêt essentiel de la structure MBE est qu'elle permet, grâce à une modélisation efficace du spectre de parole, de développer des codeurs de très faible débit et faible complexité tout en conservant une bonne qualité de parole.

2.1 Modélisation du spectre

La transformée de Fourier $S_w(\omega)$ du signal $s_w(n)$ est modélisée par le produit d'une excitation spectrale $|E_w(\omega)|$ et d'une enveloppe spectrale $H_w(\omega)$.

$$s_w(n) = w(n)s(n) \quad (2.1)$$

et

$$\check{S}_w(\omega) = H_w(\omega)|E_w(\omega)| \quad (2.2)$$

La totalité de la bande élargie est divisée en N sous-bandes, N étant un nombre fixe ou pouvant varier avec le signal. L'excitation spectrale est modélisée par un spectre harmonique sur les sous-bandes voisées et par un spectre de bruit blanc sur les sous-bandes non voisées. L'enveloppe spectrale est représentée par M amplitudes A_m centrées sur les harmoniques de la fréquence fondamentale F_0 du signal sur la fenêtre analysée.

2.2 Analyse et synthèse

Plusieurs propositions concernant la réduction de complexité ou de débit ont été faites depuis la première présentation du codeur MBE par Griffin en 1987 [1]. Nous présentons ici la structure et les procédures initialement utilisées pour le codeur MBE dans la bande téléphonique [2].

Analyse

Les M amplitudes A_m , la fréquence fondamentale F_0 et les N décisions voisé/non voisé constituent les paramètres du codeur et sont estimées en minimisant l'erreur quadratique ϵ entre le spectre original et le spectre synthétisé.

Le spectre est d'abord considéré comme entièrement voisé et la fréquence F_0 est déterminée par minimisation de ϵ . Le spectre synthétique est alors calculé et les décisions voisé/non voisé sont déterminées par comparaison à un seuil de l'erreur, sur chaque sous-bande, entre le spectre original et le spectre voisé synthétisé. Enfin, les M amplitudes A_m sont estimées par minimisation de ϵ .

Synthèse

Le signal synthétisé est construit par sommation de fonctions sinus issues des raies spectrales des sous-bandes voisées dont les pulsations, amplitudes et phases sont obtenues à partir de F_0 et des complexes A_m , et d'un signal bruité construit par FFT inverse d'un spectre de bruit blanc mis en forme par les A_m des sous-bandes non voisées.

3 Problèmes sur la bande élargie

Le codeur MBE a prouvé son efficacité sur la bande téléphonique grâce à une modélisation efficace du spectre. Cette modélisation impose un découpage en sous-bandes uniquement voisées ou non voisées, et si la bande est voisée, à une modélisation par un spectre de raies centrées sur les harmoniques de F_0 . Aussi, on peut s'interroger sur l'adéquation de cette technique sur la bande élargie.

3.1 Spectre harmonique

Sur les sous-bandes voisées, le spectre de raies est modélisé par un spectre harmonique de fréquence F_0 . Cependant, si les raies du spectre original coïncident bien avec les harmoniques de F_0 pour les basses fréquences, elles

peuvent se décaler vers les hautes fréquences conduisant à deux problèmes majeurs. Tout d'abord, les raies spectrales ne coïncident plus avec les pics du spectre, le calcul des amplitudes A_m est erroné ce qui conduit à une distorsion d'amplitude pour les hautes fréquences. Ensuite, les décisions voisé/non voisé étant faites en estimant l'erreur entre le spectre original et le spectre harmonique, si les raies ne coïncident plus, l'erreur spectrale est grande alors que la bande est bien voisée. Ces erreurs dans les décisions voisé/non voisé conduisent à des dégradations très gênantes.

3.2 Le nombre d'harmoniques varie avec F_0

Le spectre d'un signal, sur une fenêtre d'analyse, est modélisé par un nombre variable d'harmoniques, la distance entre deux harmoniques étant égale à F_0 ; cette modélisation du spectre peut être vue comme une procédure d'échantillonnage à fréquence variable. Le nombre d'harmoniques modélisant le spectre est en moyenne deux fois plus grand pour une voix d'homme que pour une voix de femme alors que la quantité d'information est théoriquement équivalente. Aussi, lorsque la fréquence fondamentale du signal augmente, l'information pertinente doit être ailleurs que uniquement sur les pics spectraux : cette information n'est pas prise en compte dans le codeur MBE classique.

3.3 Qualité du codeur MBE sur la bande élargie

L'adaptation du codeur MBE à la bande élargie a conduit à des résultats très encourageants pour les voix d'hommes mais très décevants pour les voix de femmes. Ceci confirme que le dernier problème mentionné est un point critique du codeur.

4 Structure proposée

Le point sur lequel nous nous focalisons est celui de la qualité, les contraintes de débit et complexité étant pour l'instant relâchées.

4.1 Classification phonétique des trames

Les différentes sous-bandes du spectre sont déclarées voisées ou non voisées et une erreur dans ces déterminations conduit à des dégradations très audibles. De plus, quand une trame entièrement non voisée est traitée, la modélisation MBE ne présente plus aucun intérêt. Aussi, nous proposons d'introduire une classification phonétique des trames en « non voisée » pour lesquelles la totalité de la bande est non voisée et où le spectre est représenté par un nombre fixe d'amplitudes, ou « mixte » pour lesquelles le procédé MBE classique est appliquée. Les critères de discrimination utilisés sont l'énergie, le nombre de passage par zéro et le SFM (Spectral Flatness Measure). D'autres types de classification ont déjà été proposés pour les signaux de la bande téléphonique [3], [4].

4.2 Nouvelle modélisation du spectre de raies

L'analyse spectrale d'une trame voisée montre que l'harmonicité apparaît moins pour les hautes que pour les basses fréquences, conduisant aux problèmes cités plus haut. Comme les raies sont toujours présentes en haute fréquence, mais semblent centrées sur les harmoniques de $(F_0 + \Delta f)$, nous divisons le spectre en L sous-bandes et sur chaque sous-bande l , nous cherchons la fréquence $(F_0 + \Delta f_l)$ dont les harmoniques coïncident avec les raies du spectre original.

Détermination de F_0

La fréquence fondamentale F_0 est évaluée en minimisant par rapport à l'ensemble des fréquences, l'erreur quadratique ϵ entre les spectres original et synthétisé. Cette procédure étant complexe, Griffin [2] proposa d'évaluer d'abord F_0 par une méthode d'autocorrélation puis de l'appliquer comme procédure de « raffinement ». Nous utilisons cette démarche pour évaluer F_0 .

Détermination des L fréquences $(F_0 + \Delta f_l)$

Pour chaque sous-bande, nous cherchons, par la procédure de « raffinement », la fréquence optimale $(F_0 + \Delta f_l)$ dont les harmoniques coïncident avec les raies du signal original. Ceci conduit à une réduction notable de la distorsion spectrale, tout particulièrement pour les hautes fréquences comme le montre la figure 1. De plus, une erreur importante entre le spectre original et le spectre harmonique devient effectivement un indicateur de non voisement puisqu'une erreur importante ne peut plus provenir d'un décalage entre les raies spectrales.

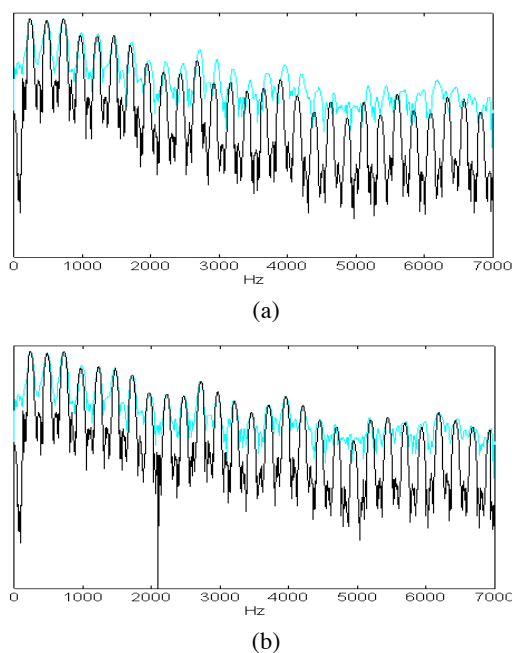


Fig. 1. (a) Spectre original (clair) et spectre modélisé avec un unique F_0 . (b) Spectre original et spectre modélisé avec 8 fréquences optimales $(F_0 + \Delta f_l)$ sur 8 sous-bandes.

Evaluation

La bande a été divisée en 8 sous-bandes de même taille. L'amélioration apportée est très nette et élimine du bruit aussi bien pour les voix d'hommes que pour les voix de femmes. Deux remarques importantes peuvent être faites concernant le débit. Tout d'abord, les L fréquences $(F_0 + \Delta f_l)$ étant très corrélées, seront quantifiées de manière différentielle de manière à ne pas augmenter trop le débit. De plus, la distorsion produisant le bruit le plus audible est la distorsion d'amplitude et non pas le décalage en fréquence. Si la procédure de détermination des fréquences optimales est appliquée uniquement durant l'étape d'analyse, permettant ainsi la détermination correcte des amplitudes des raies mais conduisant à leur positionnement erroné pour la synthèse (les raies sont alors positionnées sur les harmoniques de la fréquence F_0), le gain obtenu est tout à fait notable. Cette procédure permet d'augmenter la qualité tout en conservant un débit constant. Cette observation avait déjà été faite [5] : une modélisation bi-harmonique du spectre était proposée pour les codeurs MBE de la bande téléphonique. En considérant le débit visé nous pouvons optimiser le nombre L de bandes à considérer et décider si oui ou non les L fréquences correspondantes seront transmises.

4.3 Modélisation de l'erreur spectrale

La modélisation MBE du spectre de parole est très efficace sur la bande téléphonique. Cependant, pour la bande élargie, même avec l'introduction de la modélisation multi-harmonique du spectre, il semble que la structure MBE ne puisse fournir la qualité souhaitée.

Un problème théorique

Lorsque la fréquence fondamentale du signal augmente, le nombre d'harmoniques représentant le spectre et donc la quantité d'information diminue. Comme l'information n'est pas théoriquement moins présente dans une voix d'homme que dans une voix de femme, l'information pertinente doit se situer ailleurs que sur les raies spectrales. En fait, le spectre voisé pour une voix aiguë présente des raies moins nombreuses mais aussi moins fines que pour une voix grave. Alors, nous supposons que pour une voix grave l'information est uniquement concentrée sur les nombreuses raies spectrales alors que pour une voix aiguë l'information est diffusée sur et autour des raies. La structure MBE initiale permet de fournir une représentation adéquate du spectre pour des voix graves mais ne permet pas de modéliser toute l'information pertinente pour les voix aiguës. L'observation de l'erreur de modélisation spectrale pour une voix d'homme et une voix de femme permet de confirmer cette hypothèse (Fig.2.).

Proposition

Pour les voix aiguës, l'erreur de modélisation entre le spectre original et le spectre harmonique synthétisé étant

importante, nous proposons de la modéliser et de la transmettre au décodeur. Un signal temporel issu de cette modélisation est alors ajouté au signal synthétique au niveau du décodeur. Après la modélisation MBE du spectre original, nous notons $E_w(\omega)$ l'erreur spectrale entre signal original et le signal synthétisé.

$$E_w(\omega) = S_w(\omega) - \check{S}_w(\omega) \quad (4.1)$$

Comme l'erreur présente une structure quasi-harmonique, plus particulièrement pour les voix aiguës (Fig. 2.), nous modélisons ce spectre d'erreur par un spectre harmonique de fréquence F_i proche de F_o . La même procédure que celle utilisée pour modéliser le spectre original par un spectre harmonique est utilisée ; il en est de même pour l'étape de synthèse. Le signal synthétisé est à présent

$$\check{S}'_w(n) = \check{S}_w(n) + \hat{E}_w(n) \quad (4.2)$$

et l'erreur spectrale est alors

$$F_w(\omega) = S_w(\omega) - (\check{S}_w(\omega) + \hat{E}_w(\omega)) \quad (4.3)$$

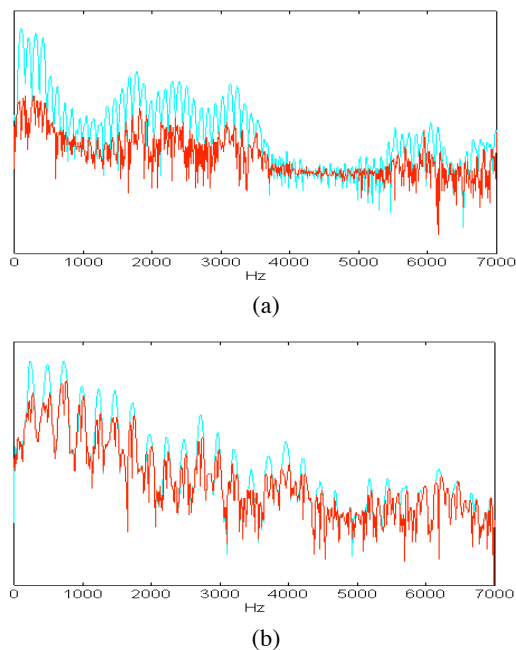


Fig. 2. Spectre original (clair) et erreur spectrale après la modélisation MBE pour une voix d'homme (a), pour une voix de femme (b).

Résultats

L'évaluation de l'énergie de l'erreur spectrale $E_w(\omega)$ et de l'erreur spectrale finale $F_w(\omega)$ pour différents signaux de parole montre qu'une amélioration non négligeable est apportée dans la modélisation du spectre. Une analyse plus détaillée montre que, comme nous le supposions, le gain est plus grand pour les voix aiguës. Théoriquement, le nombre d'harmoniques doit approximativement doubler lorsque l'erreur $E_w(\omega)$ est modélisée. Cependant, l'intérêt de cette méthode est qu'il est possible de choisir si oui ou non toutes les raies seront transmises au décodeur. Par exemple, lors du codage d'une voix grave, dans une sous-bande, si l'erreur

entre le spectre original et synthétisé est faible alors elle n'a pas besoin d'être modélisée et donc aucune amplitude n'est transmise au décodeur. Par contre, lors du codage d'une voix aiguë, le spectre d'erreur peut avoir une énergie très grande et dans ce cas, des amplitudes sont transmises pour modéliser le spectre d'erreur. Il est donc possible d'obtenir le même nombre d'amplitudes quelle que soit la fréquence fondamentale du signal.

Lorsque l'on compare les signaux obtenus grâce à cette nouvelle modélisation, le gain en qualité est plus important pour les voix aiguës que pour les voix graves. Des écoutes informelles ont permis de montrer que les voix aiguës sont maintenant plus naturelles et moins métalliques.

Des améliorations peuvent encore être apportées lors de la synthèse de l'erreur dans le domaine temporel. En effet, cette erreur est actuellement calculée avec interpolation entre les trames successives alors que ce signal n'a pas les propriétés du signal original.

5 Conclusion

Nous proposons une nouvelle structure pour un codeur en bande élargie basé sur la structure MBE qui semble prometteur. Les contraintes de débit n'ont pas été mentionnées et constitueront nos futures recherches tout comme l'introduction d'un modèle perceptuel.

6 Références

- [1] D.W. Griffin, "Multiband excitation vocoder", Ph.D. dissertation, M.I.T. Cambridge, MA, 1987.
- [2] D.W. Griffin and J.S. Lim, "Multiband Excitation vocoder", IEEE Trans. ASSP-36, no. 8, pp 1223-1235, A., 1997.
- [3] C.Garcia-Mateo, F.J. Casajus-Quiros, and L.A. Hernandez-Gomez, "Multi-band excitation coding of speech at 4.8 kbps", Pro. ICASSP, paper S1.4, 1990.
- [4] A. Das and A. Gersho, "Enhanced Multiband Excitation Coding of Speech at 2.4 kb/s with Phonetic Classification and variable Dimension VQ", Signal Processing VI, pp 943-946, 1994.
- [5] C.Garcia-Mateo, J.L. Alba-Castro, and E. R-Banga, "Speech Coding using Bi-harmonic spectral modeling", Signal Processing VII, pp 391-394, 1994.