

Modélisation 2D (« fréquence-temps ») des amplitudes spectrales

Mohammad Firouzmand & Laurent Girin

ICP – INPG/Univ. Stendhal/CNRS
B.P.25-38040 Grenoble France
{[@icp.inpg.fr](mailto:ginin, firouz)}

Abstract

This paper presents a method for modeling the spectral amplitude parameters of speech signals in “two dimensions” (2D). It consists in two cascaded modeling: the first one along the frequency axis is usual, since it consists in modeling the log-scaled spectral envelope with a sum of Discrete Cosine (DC) functions. The second one, along the time axis, consists in modeling the trajectory of the envelope DC parameters by another similar DC model. An iterative algorithm that optimally fits this 2D-model, taking into account perceptual criterions, is proposed. This approach is shown to provide an efficient representation of speech spectral amplitude parameters in terms of coefficient rates, while providing good signal quality, opening new perspectives in very-low bit-rate speech coding.

1. Introduction

Le modèle sinusoïdal de la parole (SMS) [1] a été largement étudié depuis les années 80 et appliqué avec succès à un grand nombre d'applications, tel que le codage et la transformation de la parole [2-4]. Il consiste à modéliser le signal comme une somme de I sinusoïdes:

$$s(n) = \sum_{i=1}^I A_i(n) \cos[\theta_i(n)] \quad \text{avec} \quad \theta_i(n) = \sum_{k=0}^n \omega_i(k) + \theta_i(0) \quad (1)$$

Les paramètres du SMS, amplitudes $A_i(n)$, fréquences $\omega_i(n)$ et phases $\theta_i(n)$, évoluent lentement au cours de temps. Un système d'analyse-synthèse basé sur ce modèle repose sur la mesure « à court terme » (CT) de ces paramètres aux centres de trames d'analyse consécutives, puis l'interpolation CT des valeurs mesurées consécutives afin de reconstruire le signal entier [1]. « Court terme » dénote habituellement une durée d'environ 10-30ms.

Dans une série d'études récentes [5][6], nous avons proposé de modéliser les paramètres du SMS à « long terme » (LT). Ceci signifie que les trajectoires temporelles de ces paramètres ont été modélisées sur de longues sections de parole, au-delà de la longueur de trame usuelle en analyse-synthèse à court terme. Par exemple, dans ces études, un modèle LT unique a été employé pour coder chaque trajectoire des paramètres sur chaque section de parole entièrement voisée. Une telle approche a été appliquée sur les trajectoires d'amplitude [5][6] et de phase [6] en utilisant un modèle en Cosinus Discrets (MCD) comme modèle LT. Un algorithme itératif incluant des contraintes perceptives a permis d'estimer conjointement l'ordre optimal du modèle et ses coefficients. Le nombre de ces coefficients a été réduit significativement par rapport à la modélisation CT, ce qui ouvre une nouvelle voie pour le codage de parole/audio à très bas débit. Dans ce papier, nous étendons et raffinons

cette approche en ajoutant une nouvelle étape de modélisation le long de l'axe des fréquences avant de considérer l'axe du temps : nous modélisons d'abord l'enveloppe des amplitudes par un modèle en Cosinus Discrets, comme proposé dans [7], mais avec une version raffinée qui tient compte des contraintes perceptives, inspirée de [5]. Puis, nous appliquons un deuxième MCD sur l'axe du temps pour modéliser à long terme la trajectoire des coefficients résultant de la modélisation d'enveloppe. La modélisation de l'enveloppe spectrale effectuée avant la modélisation à LT permet : 1) de résoudre le problème de « taille variable » des jeux de paramètres d'amplitudes d'une trame d'analyse à l'autre, dû aux variations de la fréquence fondamentale (et à la présence de composantes de bruit), car l'enveloppe est modélisée en utilisant un ordre fixe sur la section de parole considérée à LT ; 2) de réduire la taille de ces jeux de paramètres avant la modélisation à LT, puisque l'ordre du modèle d'enveloppe est généralement très inférieur au nombre d'amplitudes mesurées ; c'est un point important dans l'optique de l'utilisation de cette modélisation 2D pour le codage de parole à très bas débit ; et 3) de s'adapter aux transformations du signal en fréquence, telle que le *pitch-scaling* par exemple. Notons que dans cet article, nous considérons seulement la modélisation 2D des paramètres d'amplitude. Comme un des buts sous-jacents de cette étude est de fournir des outils efficaces pour un codeur à très bas débit, l'information de phase est réduite à la trajectoire de fréquence fondamentale pour les sections considérées qui sont toutes voisées. La modélisation à long terme de la trajectoire de fréquence fondamentale peut être réalisée en parallèle, comme dans [6].

Ce papier est organisé comme suit. Le processus complet d'analyse-modélisation-synthèse est décrit dans la Section 2, incluant la description du MCD, l'approche 2D, et l'algorithme d'ajustement. Des résultats sont donnés dans la Section 3.

2. Modélisation 2D des amplitudes

Nous supposons que le signal de parole est d'abord segmenté en parties voisées et non voisées par des classificateurs habituels (non décrits ici), et nous considérons ici le problème de modélisation en 2D des paramètres d'amplitude d'une section voisée entière $s(n)$, où $n = 0$ à N . Nous présentons d'abord le processus d'analyse des amplitudes, puis le modèle 2D et ensuite l'algorithme qui est utilisé pour adapter le modèle aux mesures d'amplitude.

2.1. Analyse

La première étape du processus est d'extraire les jeux successifs de paramètres d'amplitude devant être modélisés en 2D. Bien que le processus de modélisation 2D est un processus à LT selon l'axe du temps, l'analyse de chaque jeu d'amplitude est fournie sur une base habituelle à court terme.

Les expériences décrites dans cet article ont été conduites avec une analyse pitch-synchrone, pour être cohérentes avec nos études précédentes [5][6], mais des techniques d'analyse plus habituelles, par exemple basées sur la Transformée de Fourier à court terme, peuvent être utilisées. Ainsi, les signaux sont d'abord « pitch-marqués » en employant le logiciel Praat [8]. Cela signifie que les frontières des quasi-périodes du signal ont été automatiquement détectées et ces quasi-périodes sont utilisées comme trame d'analyse. La fréquence fondamentale ω_0^k est alors directement donnée par l'inverse de la période. Puis, étant donnée ω_0^k , les I_k amplitudes A_i^k correspondant à chaque harmonique à la fréquence $i \omega_0^k$ et mesurées au centre n_k de chaque période, ont été estimées en utilisant le procédé employé par George et Smith dans [4]. Pour la suite, il est important de noter que I_k dépend de ω_0^k . L'estimation des amplitudes est basée sur la minimisation de l'erreur entre le modèle sinusoïdal harmonique et le signal selon un critère des moindres carrés (MMSE). Cette procédure fournit une estimation des paramètres très précise avec un coût de calcul très bas.

2.2. Modélisation de l'enveloppe spectrale

Dans nos études précédentes, nous avons considéré la modélisation des amplitudes directement selon l'axe du temps [5]. Donc, à la fin du processus d'analyse, les amplitudes étaient réordonnées comme I jeux de K amplitudes (' dénote le vecteur/matrice transposé):

$$A_i = [A_i^1 \ A_i^2 \ \dots \ A_i^K]^t, \quad i = 1 \text{ à } I,$$

Et la trajectoire résultante était modélisée par un MCD. Alternativement, dans la présente étude, les amplitudes sont d'abord ordonnées comme couramment selon l'axe des fréquences, comme K jeux de I_k valeurs :

$$A^k = [A_1^k \ A_2^k \ \dots \ A_{I_k}^k]^t, \quad k = 1 \text{ à } K$$

Chaque jeu de paramètres est alors remplacé par un modèle d'enveloppe spectral. Nous employons le MCD de [8] qui consiste à modéliser l'enveloppe spectrale (en échelle log) par une somme de fonctions cosinus :

$$A_{DCM}^k(\omega) = d_0^k + 2 \sum_{m=1}^M d_m^k \cos(2\pi m \omega) \quad (2)$$

où M est un nombre entier positif qui est l'ordre du modèle. Bien que le nombre d'harmoniques puisse varier d'une trame à l'autre, M a une valeur fixe pour toutes les trames d'une section voisée modélisée à LT. Nous verrons dans la Section 2.4 comment M est estimé pour chaque section. Une fois M estimé, le vecteur D^k des coefficients du modèle est estimé par une version pondérée de la procédure MMSE d'ajustement du MCD avec le jeu d'amplitudes mesurées (en échelle log), minimisant l'erreur :

$$\mathcal{E}_k = \sum_{i=1}^{I_k} w_i^k \|A_i^k - A_{DCM}^k(i\omega_0^k)\|^2 \quad (3)$$

Si on dénote par M_k la matrice $I_k \times (M+1)$ de terme général $\cos(2\pi m i \omega_0^k)$ (avec un facteur 2 pour les colonnes $m > 1$), et si on dénote par W_k la matrice diagonale qui contient les poids de (3) mis au carré sur sa diagonale (ce poids sont estimés à partir de contraintes perceptives dans l'algorithme de la Section 2.4), on a :

$$D^k = (M_k^t W_k M_k)^{-1} M_k^t W_k A^k \quad (4)$$

2.3. Modélisation des trajectoires d'enveloppe

Une fois que la modélisation spectrale a été faite pour toutes les trames de la séquence de parole considérée (section voisée), la seconde étape de modélisation est la modélisation de la trajectoire temporelle des coefficients d_m^k du modèle d'enveloppe. Ainsi, ces coefficients sont d'abord ré-ordonnés le long de l'axe du temps comme $M+1$ jeux de K -vecteurs (comme fait directement pour les amplitudes dans [4]) :

$$D_m = [d_m^1 \ d_m^2 \ \dots \ d_m^K]^t \quad \text{pour } m = 0 \text{ à } M$$

Un deuxième modèle MCD est alors appliqué sur chacune des $M+1$ trajectoires, selon la même approche que celle utilisée dans [5] pour modéliser les amplitudes à LT :

$$d_m^{DCM}(n) = c_0^m + 2 \sum_{p=0}^{P_m} c_p^m \cos\left(2\pi p \frac{n}{2N}\right) \quad \text{pour } m = 0 \text{ à } M \quad (5)$$

Comme pour la modélisation d'enveloppe, les coefficients du modèle sont estimés par minimisation des moindres carrés pondérés (rappelons que les index n_k sont les centres des K trames d'analyse) :

$$\mathcal{E}_m = \sum_{k=1}^K w_m^k \|d_m^k - d_m^{DCM}(n_k)\|^2 \quad (6)$$

Si M_m dénote ici la matrice $K \times (P_m+1)$ de terme général $m_{kp} = \cos(2\pi p n_k / 2N)$, et si W_m dénote la matrice diagonale qui contient sur sa diagonale les poids de (6) mis au carré, le vecteur des coefficients du modèle est donné de façon similaire à (4) par :

$$C^m = (M_m^t W_m M_m)^{-1} M_m^t W_m D_m \quad (7)$$

Cependant, dans ce cas, l'ordre du modèle P_m dépend de la section de parole considérée et potentiellement du rang m du coefficient de l'enveloppe. En effet, l'évolution de l'enveloppe spectrale peut varier considérablement, par exemple selon la longueur de la section considérée, la séquence de phonèmes, le locuteur, la prosodie, etc. Nous présentons ci-dessous l'algorithme que nous utilisons pour estimer conjointement l'ordre P_m , les poids qui sont utilisés dans (4) et (7), et également l'ordre M du modèle d'enveloppe. Notons que le modèle 2D proposé peut être efficacement exploité pour le codage de parole à très bas débit si dans la pratique on constate que l'ordre estimé P_m est significativement inférieur à K . Ce point sera discuté plus en détail dans la Section 4.

2.4. Estimation de l'ordre et algorithme d'adaptation

Par simplicité, dans cette étude, nous prenons le même ordre $P_m = P$ pour tous les vecteurs D_m , $m = 0$ à M . L'algorithme ci-dessous peut être raffiné pour garder la possibilité d'un ordre spécifique pour chaque rang de coefficient. L'algorithme est divisé en deux parties: La première partie consiste à régler M de façon à conjointement adapter K modèles d'enveloppe optimaux aux K jeux d'amplitudes mesurées selon un critère perceptuel moyen. Ces modèles d'enveloppe seront utilisés comme une référence pour la deuxième partie de l'algorithme qui traite de la dimension temporelle, incluant une estimation itérative de l'ordre P .

Première partie de l'algorithme : modélisation des enveloppes (M est initialisé à une valeur arbitraire, typiquement 10)

- 1) Pour chaque index du temps k , $k = 1$ à K , calculer le seuil de masquage fréquentiel global $T^k(\omega)$ associé au vecteur d'amplitude A^k en employant le modèle de [9].
- 2) Initialiser K vecteurs de poids W_k , chacun de longueur I_k , avec tous les éléments à un. Itérer alors le processus suivant de l'étape 3 à l'étape 5, jusqu'à ce que chaque ratio R^k de l'étape 5 soit maximisé.
- 3) Calculer K vecteurs MCD d'enveloppe avec (4).
- 4) Pour chaque trame k , augmenter les poids où l'erreur de modélisation dépasse le seuil masquage, selon (*square* et *max* dénotent les fonctions carré et maximum (élément-par-élément), *diag* dénote la fonction qui produit une matrice diagonale à partir d'un vecteur, les éléments du vecteur étant mis sur la diagonale) :
$$E_i^k = \frac{1}{2} \text{square}(A_i^k - M_k D^k)$$

$$dW_k = \text{diag}(\max(E_i^k - T^k(i\omega_0^k), 0))$$

$$W_k = W_k + dW_k / \max(dW_k)$$
- 5) Calculer le pourcentage R^k des éléments nuls de dW_k .
- 6) Une fois que tous les R^k sont maximisés, calculer la valeur moyenne R_{min} de ces rapports. Si R_{min} est supérieur à un rapport prédéfini R_{target} (en général 0,8 à 0,9), M est diminué de 1, sinon, M est augmenté de 1. Retourner ensuite à l'étape 2 jusqu'à stabilisation de M .

Deuxième partie de l'algorithme : modélisation des enveloppes au cours de temps

- 7) Initialiser P à une valeur arbitraire significativement inférieure à K , par exemple, la partie entière de $K/4$.
- 8) Pour $m=0$ à M , calculer le modèle LT (2^{ème} MCD) de trajectoire des coefficients d'enveloppe (1^{er} MCD) avec (7).
- 9) Décoder les K modèles d'enveloppe à partir des $M+1$ MCD à long terme par : $\hat{D}_m = M_m C^m$ (équivalente à (5) appliquée aux instants n_k).
- 10) Réordonner les coefficients d'enveloppe décodés selon l'axe des fréquences et decoder les amplitudes par : $\hat{A}^k = M_k \hat{D}^k$ (équivalent à (2) appliquée aux fréquences harmoniques et avec les coefficients d'enveloppe décodés).
- 11) Calculer le rapport R_{min} de l'étape 6 en remplaçant les amplitudes modélisées à l'étape 6 (modélisation « 1D ») par celles issues de l'étape 10 (modélisation « 2D »). Si R_{min} dépasse un rapport prédéfini (typiquement $0.75 \times R_{target}$), P est diminué de 1, sinon P est augmenté de 1. Retourner ensuite à l'étape 8 jusqu'à stabilisation de P .

Notons que dans cette étude, tous les poids fonctions du temps sont mis à un, *i.e.*, aucune pondération n'est en réalité effectuée le long de l'axe de temps dans la deuxième partie de l'algorithme. Un ajustement itératif des poids semblable à ce qui a été fait dans la première partie de l'algorithme et dans nos études précédentes [5][6] pourrait probablement aboutir à diminuer encore P . Cependant, dans la plupart des expériences que nous avons faites, ce raffinement augmente considérablement la complexité de calcul pour une

amélioration faible (dans la plupart des cas, la valeur optimale de P est obtenue à la première itération sur les poids). De nouveaux travaux doivent être menés concernant ce point.

2.5. Synthèse

La synthèse est réalisée en appliquant d'abord l'interpolation linéaire entre les amplitudes (ramenées à l'échelle linéaire) résultant de l'algorithme ci-dessus. Cette interpolation inclut le processus habituel de « naissance et de mort » pour les harmoniques qui dépassent la fréquence de Nyquist [1]. Notons qu'un codeur de parole à bas débit utilisant la méthode proposée doit coder la trajectoire de fréquence fondamentale qui est employée dans l'étape 10. L'équation (1) est ensuite utilisée pour produire le signal de synthèse (fréquences et phases mesurées sont interpolées en utilisant la procédure classique de [1], puisque dans cette étude on s'intéresse seulement à la modélisation 2D des amplitudes). Dans cette étude, le processus d'analyse-modélisation-synthèse concerne seulement les parties voisées de parole. Les sections non voisées sont conservées telles quelles et sont concaténées avec les sections voisées modélisées avec une pondération locale pour éviter des artefacts audibles [4].

3. Expériences

Dans cette section, nous décrivons un ensemble d'expériences qui ont été conduites pour évaluer la modélisation des paramètres d'amplitudes par le modèle 2D présenté. Nous avons utilisé des signaux de parole échantillonnés à 8 kHz et produits par 12 locuteurs (six masculins et six féminins). Un total d'environ 3500 segments voisés de différentes tailles ont été modélisés, représentant plus de 13 minutes de parole.

3.1. Précision de la modélisation

D'abord, l'algorithme s'adapte généralement correctement aux différentes configurations des sections modélisées. Par exemple, l'ordre M peut varier avec des petites valeurs (par exemple 4) pour des spectres assez « pauvres », des valeurs habituelles pour coder la parole féminine (en général 10-11) et masculine (typiquement 15-16, voir [7]), et plus pour des spectres « riches ». Le long de l'axe de temps, l'ordre P varie également beaucoup, selon la longueur et le contenu de la section de parole modélisée. Généralement, le modèle 2D fournit des valeurs d'amplitudes qui sont assez proches des amplitudes mesurées. Ceci est garanti par la contrainte perceptuelle qui guide le comportement de l'algorithme d'ajustement : à la fin de l'algorithme, $0.75 \times R_{target}$ pourcents des amplitudes modélisées sont assurées de vérifier cette contrainte (l'erreur de modélisation est au-dessous du modèle de seuil de masquage, et on s'attend ainsi à ce qu'elle soit inaudible). Le réglage de R_{target} à 0,75 assure qu'au moins la moitié des amplitudes sont correctement modélisées selon la contrainte perceptuelle. Les expériences ont montré qu'il n'est pas très efficace d'essayer d'augmenter ce rapport global. En effet, étonnamment, la deuxième étape de modélisation (au cours de temps) fournit généralement des trajectoires d'amplitudes assez « bruitées », qui aboutissent à une diminution de la qualité du signal de synthèse. Les trajectoires modélisées par le MCD à long terme ont bien la propriété intrinsèque d'être lisses, puisque le modèle est composé de fonctions de type cosinus, elles-mêmes lisses. Cependant, des trajectoires lisses de coefficients d'enveloppe ne fournissent

pas nécessairement une suite d'enveloppes spectrales évoluant de façon régulière (et ainsi pour les amplitudes). En d'autres termes, des variations régulières des coefficients d'enveloppe entre deux trames consécutives peuvent produire des discontinuités gênantes pour les trajectoires d'amplitudes. Par conséquent, plutôt que de modifier profondément la méthode proposée, un post-filtrage est appliqué sur les trajectoires d'amplitudes pour les lisser. Un filtre médian simple s'est montré efficace pour régulariser les trajectoires et permettre la synthèse de signaux de bonne qualité (voir la Figure 1).

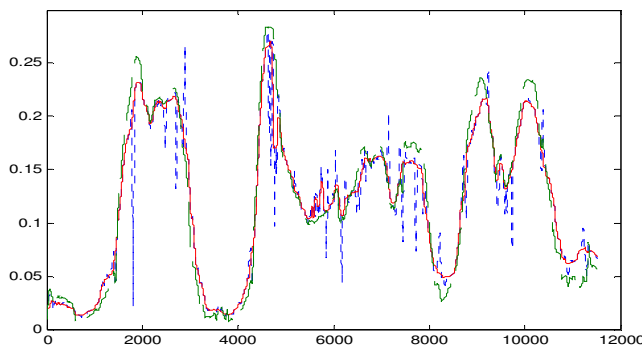


Figure 1 : Trajectoire des amplitudes mesurées (tirets verts), modélisées en 2D avant (pointillés bleus) et après (continu rouge) post-filtrage, pour la 1^{ère} harmonique d'une longue séquence voisée de voix de femme (1,33s à 8 kHz, $K=408$).

3.2. Débit des coefficients

Dans cette section, nous donnons des exemples de débits moyens pour les coefficients du modèle 2D (*i.e.*, ceux de C^m , soit l'information à transmettre dans un système de communication pour coder les amplitudes spectrales avec la technique proposée). Ces débits ont été calculés sur les 13 minutes de parole testées. Avec le rapport-cible global réglé à 90%, nous obtenons un débit moyen de 341 coefficients/s pour les voix féminines, et 404 coefficients/s pour les voix masculines. Ces débits peuvent être comparés à ceux d'une approche habituelle à court terme avec une fenêtre de 20ms : si on suppose que des spectres de voix de femme sont codés avec 11 coefficients MCD en moyenne (pour 8 kHz de parole), et que ceux de voix d'homme exigent 17 de ces coefficients, nous obtenons respectivement $11 \times 50 = 550$ coefficients/s et $17 \times 50 = 850$ coefficients/s. En faisant la moyenne, nous obtenons environ 370 coefficients/s pour le modèle 2D contre 700 coefficients/s pour l'approche habituelle « 1-D ». Ainsi, la stratégie de modélisation à long terme permet de diminuer de façon significative le débit des coefficients (d'un facteur 2 dans cette expérience). Notons que, dans ces expériences, les signaux ont été modélisés avec les deux approches 1D et 2D, et les débits mentionnés ci-dessus fournissent généralement des signaux de qualité semblable (voir ci-dessous).

3.3. Qualité du signal

Deux sujets avec une audition normale ont intensivement écouté les signaux synthétisés. Comme mentionné avant, quand le modèle 2D est appliqué directement, la qualité du signal synthétisé est diminuée par le bruit de modélisation sur les trajectoires d'amplitude. Cependant, l'utilisation d'un filtre médian permet de résoudre ce problème de

discontinuité des amplitudes et la qualité du signal synthétisé est alors fortement améliorée sans exiger d'information supplémentaire. Pour des valeurs du rapport-cible d'environ 50%, le signal synthétisé est de bonne qualité, et assez proche de l'original. Si le rapport diminue, les trajectoires des coefficients du MCD à LT sont « simplifiées » (puisque la contrainte sur l'ordre P est relâchée), et il en est de même pour les trajectoires de l'enveloppe spectrale (après post-traitement pour le lissage additionnel). Ainsi, le signal restitué, bien qu'ayant une bonne sonorité, s'éloigne du signal original : il tend vers une version hypo-articulée de ce dernier. Ceci confirme l'importance de considérer l'aspect dynamique (temporel) dans la modélisation du signal de parole.

4. Conclusion

Ce travail a confirmé la robustesse et la généralité du modèle en Cosinus Discrets, adéquat pour modéliser l'enveloppe spectrale (comme déjà montré dans [7]) et la trajectoire temporelle de paramètres (comme déjà montré dans [5]). Réunir ces deux aspects dans un modèle 2D a abouti à de nouvelles avancées. Dans cette étude, la raison principale de cette efficacité est la variabilité intrinsèque du débit : l'ordre M du modèle d'enveloppe et l'ordre P du modèle temporel sont tous les deux ajustés sur les caractéristiques locales du signal. L'approche 2D et l'algorithme associé peuvent permettre de réduire significativement le nombre de coefficients pour la représentation des amplitudes spectrales, tout en préservant une qualité raisonnable pour les signaux synthétisés. Les travaux futurs concerneront l'utilisation de l'approche proposée dans un codeur de parole à très bas débit sans contraintes sur le délai de codage-décodage.

5. Références

1. R. J. McAulay & T. F. Quatieri, Speech analysis/ synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, **34**(4), 1986, pp. 744-754.
2. T. F. Quatieri & R. J. McAulay, Shape invariant time-scale and pitch modification of speech, *IEEE Trans. Signal Proc.*, **40**(3), 1992, pp. 497-510.
3. R. J. McAulay & T. F. Quatieri, Sinusoidal coding, in *Speech coding and synthesis*, (W. B. Kleijn & K. K. Paliwal, eds), ch. 4, Elsevier, 1995.
4. E. B. George & M. J. T. Smith, Speech analysis/ synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.*, **5**(5), 1997, pp. 389-406.
5. M. Firouzmand & L. Girin, Perceptually weighted long-term modeling of sinusoidal speech amplitude trajectories, *Proc. IEEE Int. Conf. on Acoustics, Speech & Signal Proc. (ICASSP 2005)*, Philadelphia, USA, 2005.
6. L. Girin, M. Firouzmand & S. Marchand, Perceptual long-term variable-rate sinusoidal modeling of speech, *submitted to IEEE Trans. Speech and Audio Proc.*, 2005.
7. O. Cappé, J. Laroche & E. Moulines, Regularized estimation of cepstrum envelope from discrete frequency points, *Proc. IEEE Workshop Applications Signal Proc. Audio Acoustics (WASPAA)*, 1995.
8. www.praat.org
9. ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s, Part 3: Audio, 1992.