Modèles Sinusoïdaux Étendus pour le Codage Audio

Rémy Boyer[1], Slim Essid[2], Karim Abed-Meraim[2], Nicolas Moreau[2]

[1]Université de Sherbrooke, Département de Génie Électrique et de Génie Informatique
2500, boulevard de l'Université, Canada J1K 2R1

[2]GET-ENST, Département Traitement du Signal et des Images (TSI)
46, rue Barrault, 75013, Paris, France

Remy.Boyer@USherbrooke.ca, Essid/Abed/Moreau@tsi.enst.fr

Résumé − Dans cet article, on commence par faire un bref panorama de quelques extensions du modèle sinusoïdal. Ensuite, dans une optique de codage du signal audio, on retient deux représentations, nommées modèle sinusoïdal amorti exponentiellement et modèle sinusoïdal amorti et retardé. On montre alors leur utilité vis-à-vis de phénomènes audio identifiés (transitoires, pseudo-stationnaires, ...). En outre, on propose un algorithme d'estimation des paramètres de modèle alliant une approche Haute-Résolution et un schéma par déflation. Finalement, nous montrons en quoi ces deux modèles sont des solutions viables en tant que "briques de base" dans une architecture de codage sinusoïdal audio.

Abstract − In this paper, we present a short overview of several extended sinusoidal models. Two particular, models are then considered for the purpose of audio coding, namely the Exponentially Damped Sinusoidal model and the Damped and Delayed Sinusoidal model and we show their efficiency for the task of modeling special audio signal behavior (transient, pseudo-stationary, ...). Additionally, we expose an estimation algorithm based on a High-Resolution approach which is exploited in a deflation scheme. Finally, we propose these models as a basic tool in the heart of a sinusoidal audio coding architecture.

1 Introduction

La contribution de ce travail s'inscrit dans la philosophie des modèles sinusoïdaux initiés au MIT [1] au début des années 80 pour coder de la parole en bande téléphonique. De telles modélisations ne se sont toutefois pas limitées à la parole puisqu'elles ont été largement exploitées, à Stanford, dans les travaux de X. Serra [2] dans un schéma d'analyse/synthèse de signaux musicaux reposant sur un modèle dit "Sinusoïdes + Bruit".
Des modifications du modèle "Sinusoïdes + Bruit" ont été proposées donnant lieu à des modèles "Sinusoïdes + Transitoires + Bruit" [3, 4] afin de mieux représenter les signaux transitoires tels que les attaques ou évanouissements de sons. Ces modifications ont pour but d'améliorer la représentation des signaux à variations temporelles rapides.
La majorité des techniques paramétriques actuelles reposent sur l'utilisation du modèle sinusoïdal stationnaire [1]. Or, les signaux transitoires audio sont en général large bande et à support temporel étroit. Cette observation est en complète contradiction avec le "comportement" temps-fréquence du modèle sinusoïdal (large en temps et étroit en fréquence). Il en découle une inadaptation structurelle de ce modèle aux signaux brefs et à variations rapides.
Dans cet article, on présente brièvement plusieurs extensions du modèle sinusoïdal stationnaire que l'on regroupe sous le terme générique de *modèles sinusoïdaux étendus*. Ensuite, parmi ces derniers, on sélectionne deux modèles, étant à notre avis, particulièrement adaptés pour une representation compacte et générique des signaux audio. Ces deux extensions sont **(1)** le modèle sinusoïdal amorti exponentiellement (EDS) et **(2)** le modèle sinusoïdal amorti et retardé par paquets (PDDS). Ensuite, on propose un algorithme d'estimation des paramètres de modèle alliant une approche Haute-Résolution et un schéma par déflation. Finalement, on compare ces deux modèles du point de vue des performances de modélisation et de la complexité algorithmique et sur un exemple de signal percussif typique.

1. Amplitudes, phases et pulsations supposées constantes ou à variations lentes devant la durée d'analyse N.

2 Modèles sinusoïdaux étendus

Les modèles sinusoïdaux étendus s'expriment comme le produit d'un fenêtrage temporel et d'un terme d'oscillation. On donne sur la figure 1 une vision synthétique de ces modèles :

1. Le modèle sinusoïdal [1, 2],
2. le modèle sinusoïdal à fenêtrage exponentiel global [6] (une unique fenêtre pour une somme de composantes sinusoïdales),
3. le modèle sinusoïdal à fenêtrage exponentiel [5, 6], noté EDS (un fenêtrage par forme d'onde),
4. le modèle sinusoïdal à fenêtrage exponentiel retardé par *paquets*, noté PDDS (un unique retard pour une somme de composantes EDS),
5. le modèle sinusoïdal à fenêtrage exponentiel retardé, noté DDS (un retard par composante EDS).

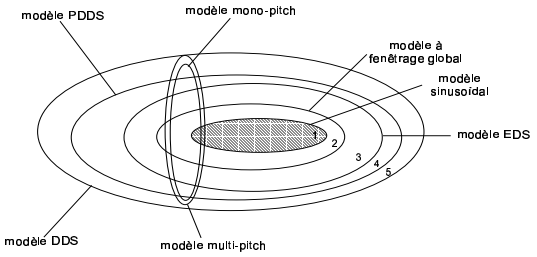


FIG. 1 – *Représentation synthétique des modèles sinusoïdaux étendus.*

2.1 Modèles sinusoïdaux étendus adaptés au codage audio

Les modèles sinusoïdaux et sinusoïdaux à fenêtrage global ne sont pas développés ici car il est maintenant établi que ceux-ci ne permettent pas d'obtenir des performances élevés (en terme de compacité de la représentation) sur la diversité des signaux audio et en particuliers sur les signaux musicaux transitoires. Ces approches sont en général dédiées à la modélisation (efficace) des signaux de parole [1, 6] et de certains signaux musicaux quasi-harmoniques [2].
De même, le modèle DDS est volontairement passé sous silence car malgré le fait que celui-ci permet de modéliser correctement un ensemble large de signaux audio, son grand nombre de paramètres (un retard par forme d'onde) conduit à des représentations non-compactes [4]. Enfin pour terminer, on peut noter que ces modèles peuvent être aussi considérés dans leur version mono ou multi-pitch. Cependant, ce choix implique une harmonicité du signal analysé, ce qui est adapté à la parole mais pas à certains phénomènes musicaux identifiés. Dans cet article, nous focaliserons, donc, notre propos sur les deux modèles restants : EDS et PDDS. Ces deux approches sont, à notre avis, des solutions viables pour une représentation compacte et générique d'un grand nombre de phénomènes audio.

2.2 Le modèle sinusoïdal amorti exponentiellement (EDS)

On définit le modèle sinusoïdal amorti exponentiellement (EDS) par l'expression suivante :

$$X(n) = \sum_{\ell=1}^{M} a_\ell e^{d_\ell n} \ . \ \cos[\omega_\ell n + \phi_\ell] \qquad (1)$$

où $0 \leq n \leq N-1$ et $\{a_\ell,d_\ell,\omega_\ell,\phi_\ell\}_{1\leq\ell\leq M}$ sont respectivement les amplitudes réelles, les facteurs d'amortissement réels, les pulsations, les phases et M est le nombre de sinusoïdes amorties. On peut voir que pour des valeurs $|d_\ell|$ grandes, l'amplitude de ce modèle s'autorise des variations rapides au regard de la durée d'analyse N.

2.2.1 Avantages et limitations du modèle EDS

L'utilisation du modèle EDS dans la communauté audio est assez récente et prometteuse [4, 5, 6]. Il est maintenant admis que ce modèle conduit à des représentations compactes de nombreux signaux audio de type pseudo-stationnaires ou transitoires "doux". La raison principale est que ce modèle, bénéficiant de facteurs d'amortissement, permet de contraindre son support temporel en *début* et *fin* du segment d'analyse.
Inversement, si on considère une analyse par segmentation régulière de signaux fortement transitoires (batterie, castagnettes, gong, ...), caractérisant bon nombre de signaux musicaux, le modèle EDS perd de son efficacité puisque de tels signaux n'ont aucune raison "d'apparaître" uniquement en début ou fin du segment d'analyse. Ces considérations seront renforcées dans la section 2.4 par l'étude d'une attaque de castagnettes.
Afin de surmonter ce problème, il semble alors naturel de modifier le modèle EDS en lui ajoutant un paramètre de retard et un échelon de Heaviside afin de pouvoir translater librement la forme d'onde du modèle au sein de la trame d'analyse. Le modèle résultant est nommé modèle sinusoïdal amorti et retardé par paquets et est présenté dans la section suivante.

2.3 Le modèle sinusoïdal amorti et retardé par paquets (PDDS)

L'expression du modèle sinusoïdal retardé par paquets (PDDS) est donnée par :

$$\bar{X}(n) = \sum_{k=0}^{K} \hat{x}_k(n) \qquad (2)$$

où K ($\geq$ 1) est le nombre de transitoires présents sur le segment analysé. Par convention, on pose $\bar{X}(n) = X(n)$ pour $K = 0$. On définit le k-ième "paquet" de sinusoïdes par :

$$\hat{x}_k(n) = \sum_{\ell=1}^{M_k} a_{\ell,k} e^{d_{\ell,k}(n-t_k)} \ . \ \cos[\omega_{\ell,k}(n-t_k)+\phi_{\ell,k}]\psi(n-t_k)$$

où M_k (respectivement t_k) est l'ordre partiel de modélisation (respectivement l'instant d'apparition ou retard) du k-ème transitoire et $\psi(n)$ est l'échelon de Heaviside. On définit l'ordre total de modélisation par $M = \sum_{k=0}^{K} M_k$. Contrairement au modèle EDS qui permet d'avoir des supports temporels réduits *seulement* en début et fin du segment d'analyse, le modèle PDDS, bénéficiant d'un paramètre de retard, réalise un pavage non-contraint du plan Temps-Fréquence et est donc mieux adapté à la modélisation de signaux audio présentant des retards et supports quelconques.

2.3.1 Modélisation des signaux fortement transitoires

On peut voir sur la figure 1 que le modèle PDDS est une généralisation du modèle EDS puisque pour des retards nuls, il se réduit au modèle EDS. Il présente donc les mêmes avantages que le modèle EDS sur les signaux quasi-stationnaires et transitoires "doux" (*cf.* paragraphe 2.2.1). En outre, le modèle PDDS cherche à exploiter au mieux certaines connaissances *a priori* sur les signaux percussifs. On fait alors deux hypothèses réalistes : **(H.1)** Un

son percussif peut être vu comme une somme de composantes EDS ayant un retard identique et **(H.2)** Deux transitoires sont séparés d'une distance suffisante ou/et ont une décroissance suffisamment rapide pour pouvoir estimer les temps d'arrivée en se basant uniquement sur la variation de l'enveloppe du signal.

2.3.2 Segmentation dynamique et non-recouvrante

En accord avec **(H.2)**, les retards sont estimés directement sur l'enveloppe du signal à l'aide d'une recherche de maxima dans la variation de l'enveloppe d'amplitude du signal [4]. La connaissance des $\{t_k\}_{0\leq k\leq K}$ permet de réaliser une segmentation *dynamique* et *non-recouvrante* de la trame courante où $\mathcal{T}_k = [t_k, \ldots, t_{k+1} - 1]$. Contrairement à l'approche par recouvrement, ce type de segmentation permet de limiter la propagation des erreurs d'estimation au travers des segments.

2.3.3 Justification d'une approche HR

En préliminaire, notons que le signal $\hat{x}_k(n + t_k)$ est un modèle EDS d'ordre M_k et de longueur $N - t_k$. De plus, ayant la connaissance des $\{t_k\}$, il est alors possible d'exploiter les méthodes à Haute-Résolution (HR) basées sur les décompositions en "sous-espace" (ESPRIT, Matrix-Pencil, ...) sur chaque $\mathcal{T}_k$. L'utilisation de ces approches se justifie par le fait que les segments $\{\mathcal{T}_k\}$ peuvent être de longueur réduite B_k du fait : (**1**) de la proximité de deux transitoires successifs et (**2**) de la distance entre le début (respectivement fin) de la trame et le retard du premier (respectivement dernier) transitoire sur la trame courante. La résolution fréquentielle, en $O(B_k^{-1})$, est donc élevée (supérieure à 150 Hz pour une attaque de castagnettes échantillonnée à 44.1 kHz) et interdit toutes approches d'estimation des paramètres du modèle basées sur une analyse de Fourier Court Terme.

2.3.4 Choix de l'algorithme d'estimation

Dans [4], il est proposé trois algorithmes HR d'estimation des paramètres d'un modèle PDDS. Deux réalisent une estimation conjointe des paramètres de modèle de l'ensemble des signaux transitoires présents sur la trame courante. Le troisième exploite quant à lui, une approche HR alliée à un schéma par déflation. On expose ici ce dernier car il présente une robustesse accrue aux erreurs d'estimation et évite le problème délicat de l'appariement des composantes. Plus précisément, cet algorithme est constitué de quatre phases :

• La première étape est l'estimation des ordres M_k. On doit apporter le plus grand soin à la modélisation temporelle de la phase d'attaque et de décroissance du son. On décide alors de déterminer les ordres de manière empirique : soit $\gamma \in \mathbb{R}^+$, on pose $M_k = \lceil \gamma \,.\, \varepsilon_k \rceil$ où $\lceil . \rceil$ dénote la partie entière supérieure et ε_k est la puissance du signal transitoire sur $\mathcal{T}_k$. Par suite, on détermine le coefficient de proportionnalité selon $\gamma = \frac{M+1}{\sum_k \varepsilon_k}$.

• La deuxième étape, nommée *analyse PDDS*, est l'estimation des paramètres $\{a_{\ell,k}, d_{\ell,k}, \omega_{\ell,k}, \phi_{\ell,k}\}_{1\leq\ell\leq M_k}$ du transitoire présent sur le segment $\mathcal{T}_k$. L'algorithme utilisé pour réaliser cette opération est exposé dans le tableau 1.

(1)	Construction de la matrice (Hankel) de données : $\boldsymbol{H}_k = \begin{pmatrix} x(t_k) & x(t_k+1) & \ldots & x(\tau_k) \\ x(t_k+1) & x(t_k+2) & \ldots & x(\tau_k+1) \\ \vdots & \vdots & & \vdots \\ x(\tau_k) & x(\tau_k+1) & \ldots & x(t_{k+1}-1) \end{pmatrix}$ où $\tau_k = \left\lceil \frac{t_k+t_{k+1}-1}{2} \right\rceil$
(2)	Calcul de la SVD (Singular Value Decomposition) : $\boldsymbol{H}_k = \sum_{r=1}^{\tau_k - t_k + 1} \sigma_r \,.\, \boldsymbol{u}_r . \boldsymbol{v}_r^T$
(3)	Détermination de la base signal : $\boldsymbol{U}^{(2M_k)} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{2M_k}]$
(4)	Extraction des pôles : $\{z_{1,k}, z_{1,k}^*, \ldots, z_{M_k,k}, z_{M_k,k}^*\} = \lambda\left\{\boldsymbol{U}_{\downarrow}^{(2M_k)\dagger} . \boldsymbol{U}_{\uparrow}^{(2M_k)}\right\}$ où $z_{\ell,k} = e^{d_{\ell,k} + i\omega_{\ell,k}}$.

TAB. 1 – *Méthode d'estimation HR. ↓ (respectivement ↑) signifie que la dernière (respectivement première) ligne est supprimée, [.] note la partie entière, † indique l'opération de pseudo-inverse et λ{.} représente le spectre de valeurs propres.*

• La troisième étape, nommée *synthèse PDDS*, construit à partir des paramètres estimés à l'étape d'analyse précédente, le k-ième paquet de sinusoïdes $\hat{\boldsymbol{x}}_k$ de N échantillons selon :

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{J}}_k \,.\, \boldsymbol{Z}_k^{(N-t_k)} \,.\, \boldsymbol{Z}_k^{(B_k)\dagger} \,.\, \boldsymbol{J}_k \,.\, \boldsymbol{x} \qquad (3)$$

où les matrices $\bar{\boldsymbol{J}}_k$, de dimensions $N \times (N - t_k)$, et $\boldsymbol{J}_k$, de dimensions $B_k \times N$, sont respectivement définies afin d'ajouter t_k lignes de "0" et de sélectionner les échantillons correspondants au segment $\mathcal{T}_k$. Enfin, on note la matrice de Vandermonde de a lignes par :

$$\boldsymbol{Z}_k^{(a)} = \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ z_{1,k} & z_{1,k}^* & \ldots & z_{M_k,k} & z_{M_k,k}^* \\ \vdots & \vdots & & \vdots & \vdots \\ z_{1,k}^{a-1} & z_{1,k}^{*(a-1)} & \ldots & z_{M_k,k}^{a-1} & z_{M_k,k}^{*(a-1)} \end{pmatrix}$$

• La dernière étape, nommée *procédure de déflation*, retire $\hat{\boldsymbol{x}}_k$ au dernier signal résiduel $\boldsymbol{r}_k = \boldsymbol{r}_{k-1} - \hat{\boldsymbol{x}}_k$ avec pour initialisation $\boldsymbol{r}_0 = \boldsymbol{x}$ (signal audio de trame courante). Au final, le signal $\boldsymbol{x}$ est approximé selon $\boldsymbol{x} \approx \sum_{k=0}^{K} \hat{\boldsymbol{x}}_k$.

2.4 Comparaison des modèles EDS et PDDS

Dans cette partie, on expose les points forts du modèle PDDS vis-à-vis du modèle EDS sur une attaque de castagnettes (*cf.* figure 2).

2.4.1 Exemple d'un signal transitoire typique

A l'analyse de la figure 3-(a), on peut observer les défauts typiques générés par le modèle EDS sur signaux fortement transitoires, *i.e.* **(1)** signal de *pré-écho* (surplus d'énergie en amont de l'attaque) et **(2)** *manque de dynamique* au niveau de l'attaque. Ces artéfacts de modélisation sont tout particulièrement gênants à l'écoute et apparaissent même pour des ordres de modélisation faibles (*cf.*

figure 3-(a) pour $M = 3$). A l'inverse, la figure 3-(b), nous montre que ces phénomènes sont efficacement combattus et on peut constater la très bonne restitution de la forme d'onde de l'attaque de castagnettes par le modèle PDDS.

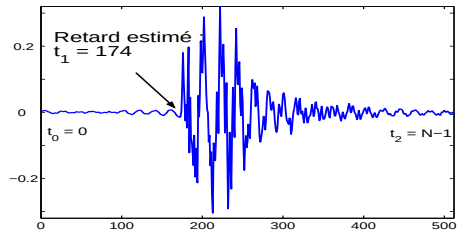
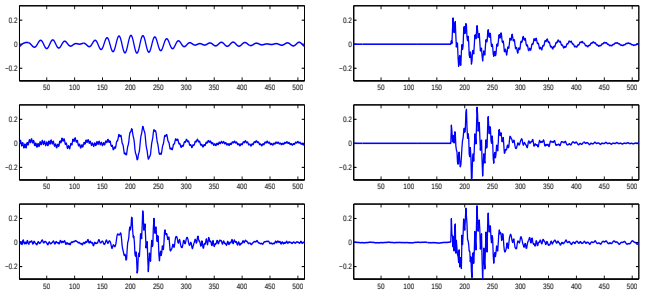


FIG. 2 – *Attaque de castagnettes. Forme d'onde.*



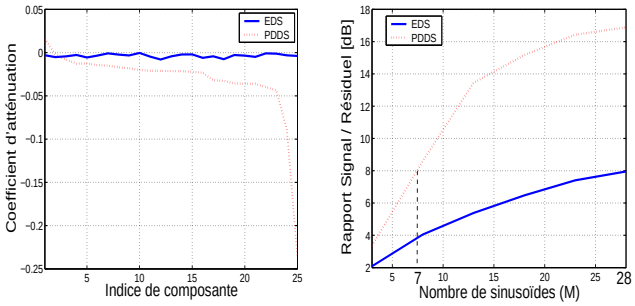
(a) *(b)*

FIG. 3 – *Modélisation progressive (de haut en bas $M = 3; 7; 28$), (a) modélisation EDS, (b) modélisation PDDS.*

La figure 4-(a) représente la valeur numérique de l'atténuation estimée en fonction de l'indice de la composante. On peut voir que le modèle EDS présente des atténuations faibles. On est proche du "comportement" d'un modèle sinusoïdal à amplitude constante. Ceci explique la relative inefficacité de ce modèle sur signaux transitoires. A l'inverse, les fortes atténuations du modèle PDDS permettent de dire que ce paramètre "joue bien son rôle" de représentation de la partie de décroissance du signal transitoire. En remarquant que le coût de codage d'un modèle EDS et d'un modèle PDDS est quasi-identique et en se référant à la figure 4-(b), on peut conclure que les performances de représentation de la forme d'onde temporelle par le modèle EDS sont intrinsèquement limitées alors qu'il est possible d'obtenir des RSR (Rapport Signal sur Résiduel) élevés pour le modèle PDDS. En guise d'illustration, on notera qu'il faut 7 composantes (31 paramètres) pour le modèle PDDS contre 28 (112 paramètres) pour le modèle EDS, pour un RSR de 8 dB. A l'inverse pour environ 112 paramètres PDDS, le RSR est proche de 17 dB.

2.4.2 Complexité algorithmique

La coût de l'algorithme HR exposé dans le tableau 1 est dominé par le coût de sa SVD, soit $O(N^3)$. Celui-ci peut être réduit en $O(NM^2)$ si on utilise un algorithme rapide d'extraction de la base signal de dimension M [4]. C'est aussi le coût dominant pour l'estimation des paramètres du modèle EDS. En ce qui concerne le modèle PDDS, le coût algorithmique est en général inférieur et égal à $\sum_k O(B_k M_k^2)$.



(a) *(b)*

FIG. 4 – *(a) Valeurs de l'atténuation ($M = 25$), (b) RSR [dB] Vs. nombre de sinusoïdes (M).*

3 Conclusion

Dans cette contribution, on présente l'intérêt qu'il peut exister à exploiter des extensions du modèle sinusoïdal afin de représenter plus efficacement certains signaux tels que les signaux fortement transitoires. Ces derniers sont traditionnellement le point faible des approches sinusoïdales, générant un ensemble d'artefacts très gênant à l'écoute. Ces dégradations sont largement atténuées par l'utilisation du modèle sinusoïdal amorti exponentiellement. Cependant, ce dernier reste perfectible sur les signaux fortement transitoires comme les sons percussifs. On présente alors une évolution profonde du modèle sinusoïdal amorti exponentiellement en intégrant un paramètre de retard permettant de translater au sein de la trame courante des "paquets" de sinusoïdes. En outre, on montre l'intérêt d'exploiter une approche rapide d'estimation Haute-Résolution dans le cadre de la modélisation d'événements audio brefs dans le temps. Finalement, le modèle PDDS est une solution viable en tant que "brique de base" dans une architecture de codage sinusoïdal audio.

Références

[1] R.J. McAulay and T.F. Quatieri, "Speech analysis & synthesis based on a sinusoidal representation", *IEEE Trans. on ASSP*, Vol. 34, No. 4, August 1986.

[2] X. Serra and J. Smith III, "Spectral Modeling Synthesis : A Sound System Based on a Deterministic plus Stochastic Decomposition", *Computer Music Journal*, Vol. 14, No. 4, Winter 1990.

[3] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, Stanford University, 1998. August, 1998.

[4] R. Boyer, *Modélisation et Codage de Signaux Audio par Extension du Modèle Sinusoïdal - Représentations Compactes des Signaux à Variations Rapides*, Thèse de doctorat, ENST, Paris, 2002.

[5] P. Lemmerling, I. Dologlou and S. Van Huffel, "Speech Compression based on exact modeling and Structured Total Least Norm optimization", *Proc. of IEEE Int. Conf. Signal Processing*, May, 1998.

[6] J. Jensen, *Sinusoidal Models for Speech Signal Processing*, PhD Thesis, University of Aalborg (CPK), August, 2000.