Un Algorithme de Réduction de la Réverbération de Signaux Issus du Vocoder de Phase

Joseph di Martino, Yves Laprie

LORIA
B.P. 239 54506 Vandoeuvre-lès-Nancy FRANCE
Tél.: ++33 (0)3 83 59 20 36 - Fax: ++33 (0)3 83 41 30 79
Mél: {jdm, laprie}@loria.fr
http://www.loria.fr/~jdm

Résumé

Time-Scale modifications of speech signals, based on frequency-domain techniques are hampered by an important artifact called phasiness. This artifact corresponds to the destruction of the shape of the original signal, i.e. the de-synchronisation between the phases of frequency components. This paper describes an algorithm that preserves the shape invariance of speech signals in the context of the phase vocoder. At ICASSP'2001 we presented a first version of this work where phases were corrected at the onsets of the voiced portions of the speech signals. In this study, we extended the previous work by allowing the algorithm to synchronize and correct the phases at regular intervals of the voiced segments of speech signals. Due to our algorithm, modified signals, even for large expansion factors, are of high quality and almost exempt of phasiness. A demonstration is proposed at the web page: **http://www.loria.fr/~jdm/PhaseVocoder/index.html** where several audio files can be down-loaded.

1. Introduction

Les techniques de compression/expansion de signaux de parole sont étudiées depuis plus d'une dizaine d'années. Elles peuvent se diviser en deux catégories suivant que les modifications du signal se font directement au niveau temporel ou au contraire au niveau fréquentiel. Les méthodes dites temporelles utilisent une technique de modification bien connue: PSOLA [1] [3]. Les méthodes fréquentielles, quant à elles, nécessitent généralement l'utilisation d'un vocoder de phase [4]. Les techniques de modification temporelles et fréquentielles ont des avantages et des inconvénients qui les différencient. Les méthodes temporelles généralement ne permettent pas l'utilisation des facteurs d'expansion importants. Elles ont leurs propres artefacts comme la «modulation de tempo» ou le phénomène perturbateur de «gazouillement». En revanche, elles ont l'avantage de nécessiter peu de temps de calcul. Les méthodes fréquentielles par contre sont plus gourmandes en temps de calcul. Elles ont leur propre artefact qui est bien connu sous la dénomination anglo-saxonne de «phasiness» ou réverbération. Elles ne nécessitent pas, comme pour les méthodes temporelles, la connaissance préalable du pitch, ce qui est un gros avantage. Laroche explique dans [2] les raisons pour lesquelles les signaux issus du vocoder de phase sont colorés par un bruit de réverbération, et il présente une méthode qui apporte une réponse partielle à ce problème. Dans [12] Quatieri et al. ont aussi proposé une méthode basée sur le même principe pour supprimer cet artefact acoustique.

Dans cet article nous ne traiterons que des méthodes fréquentielles fondées sur la notion de vocoder de phase. Nous expliciterons en détail l'algorithme de synchronisation et de correction des phases conçu pour éliminer la réverbération. Et surtout, nous détaillerons la méthode qui nous a permis d'effectuer une synchronisation itérée régulièrement dans le temps sur les portions voisées des signaux de parole.

2. Le Vocoder de Phase de Portnoff-Seneff

Dans cette section nous allons décrire rapidement le vocoder de phase de Portnoff-Seneff [4] [5] [6] [7]. Ce vocoder est régi par les équations d'analyse/synthèse suivantes:

l'équation d'Analyse:

$$Y(n,\omega_k) = A(\beta n,\omega_k) exp[j\upsilon(\beta n,\omega_k)/\beta] \qquad (1)$$

l'équation de Synthèse:

$$s(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(n,\omega_k) * (-1)^k \qquad (2)$$

où $\omega_k = 2\pi k/N$, N correspond au nombre de points de la DFT et β est le facteur d'expansion/compression.

$A(n,\omega_k)$ et $\upsilon(n,\omega_k)$ correspondent respectivement à l'amplitude spectrale et à la phase dépliée du signal original au temps n et à la fréquence ω_k. Pour que l'équation (2) soit valide, les discontinuités de π doivent être restaurées au niveau de la phase dépliée ([5] page 568). De bons algorithmes de dépliement de la phase peuvent être trouvés dans ([8], page 508) et [6].

L'équation (2) permet de calculer la valeur de l'échantillon synthétisé situé au milieu de la fenêtre d'analyse. C'est la raison pour laquelle, les coefficients DFT modifiés sont sommés en opposition de phase. Comme dans [5], le système décrit par la Fig. 1 opère à la fréquence d'échantillonage. Cela signifie que chaque opérateur de la Fig. 1 effectue une opération pour chaque

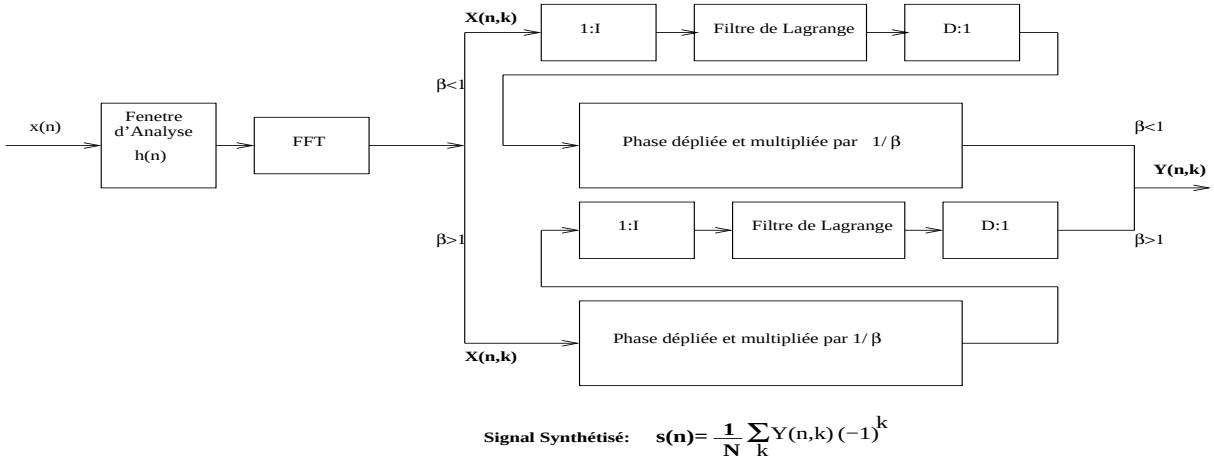

FIG. 1 – *Une description schématique du vocoder de phase. 1:I représente l'interpolation et ajoute I-1 zéros entre 2 échantillons; D:1 représente l'opération de décimation et sélectionne 1 échantillon sur D. Le rapport D/I approxime β.*

échantillon de parole.

Lors de la transformation d'un signal de parole il est nécessaire d'interpoler les coefficients de la transformée de Fourier. Pour cela, nous avons opté pour un filtre de Lagrange [9] parce qu'il préserve les points originaux (i.e. les coefficents DFT sous forme de partie réelle et imaginaire) et n'introduit aucune distorsion de phase.

3. L'algorithme de Correction des Phases

Dans la mesure où l'algorithme de dépliement de la phase fournit de bons résultats, les équations du vocoder précédemment décrites permettent d'obtenir des signaux compressés ou ralentis d'excellente qualité. Nous avons vérifié expérimentalement que l'algorithme de modification de signaux décrit par les équations (1) et (2) engendre des signaux d'excellente qualité, même pour des facteurs d'expansion importants. Toutefois les signaux obtenus, surtout pour des facteurs d'expansion élevés, ne sont pas exempts de réverbération. Afin d'éliminer cet artefact, nous avons décidé d'implémenter un algorithme de correction des phases qui maintient la forme du signal dilaté par rapport au signal original dans les portions voisées.

La technique de correction des phases que nous allons décrire, peut être appliquée, à toutes les périodes de pitch, ou bien au début d'un groupe de périodes de pitch, ou de manière extrême, au début de chaque région de voisement. Nous pensons que faire une synchronisation des phases toutes les périodes de pitch nécessiterait une trop grande puissance de calcul. Faire une synchronisation au début de chaque région de voisement, comme nous l'avons proposé en [7], peut nécessiter beaucoup de place mémoire et de plus, la qualité du signal obtenu varie en fonction de la longueur des segments de parole auxquels notre technique de synchronisation est appliquée. Ce sont les raisons pour lesquelles, nous avons opté, dans cette étude, pour une correction par groupe de périodes de pitch correspondant à un intervalle de temps constant.

Afin de pouvoir mettre en œuvre notre algorithme de correction de phases, nous avions besoin d'une méthode capable de détecter les «marques de pitch» au niveau du signal, correspondant généralement à des extrema du signal vocal. Pour ce faire nous avons utilisé un algorithme de détection de marques —ADM— que nous avons développé par ailleurs [10].

Soient $x(n)$ et $s(n)$ le signal analysé et modifié respectivement. Soit t_{n_0} un instant donné par l'ADM. Afin de préserver la propriété d'invariance de forme nous voulons que $x(t_{n_0}) = s(t_{n_0}/\beta)$. Afin d'atteindre ce but, il est possible d'introduire un décalage de phase ϕ_k à chaque canal k. En conséquence, le signal synthétisé est donné par:

$$s(n) = \frac{1}{N} \quad [A(\beta n, 0)e^{j\phi_0(n)} + \sum_{k=1}^{N/2-1} 2A(\beta n, \omega_k)cos(v(\beta n, \omega_k)/\beta + \phi_k) + A(\beta n, N/2)e^{j\phi_{N/2}(n)}] \tag{3}$$

$\phi_0(n)$, la phase de la composante continue $A(\beta n, 0)$ est égale à 0 ou π et par conséquent, nous avons choisi de ne pas la corriger. Le dernier terme $A(\beta n, N/2)$ n'a pas été pris en considération dans notre étude car son amplitude est négligeable.

Quatieri et McAuley ont proposé une expression similaire (voir [13] page 382) pour $s(n)$ dans le cadre bien connu du modèle sinusoïdal [11]. Toutefois, une inconsistance apparaît quand la propriété d'invariance doit être préservée plus d'une fois. Afin d'éliminer ce problème Quatieri et al. [14], ont proposé une approche sous-bande.

Dans le cadre de notre étude, nous avons décidé de ne pas utiliser le concept sous-bande. Nous avons opté pour une technique de minimisation des moindres carrés. Soient t_{n_i} les instants donnés par l'ADM appliqué au signal analysé x. Nous proposons de trouver le vecteur de décalage Φ tel que:

$$\Phi = Argmin_{\phi} E(\phi) \tag{4}$$

avec

$$E(\phi) = \sum_{n_i}[x(t_{n_i}) - \frac{1}{N}A(\beta t_{n'_i}, 0)e^{j\phi_0} - \frac{2}{N}\sum_{k=1}^{N/2-1} A(\beta t_{n'_i}, \omega_k) \times cos((v(\beta t_{n'_i}, \omega_k)/\beta + \phi_k))]^2 \quad (5)$$

où ϕ est le vecteur des ϕ_k. Il faut noter aussi que dans l'équation précédente, $t_{n'_i}$ est le temps pour le signal correspondant à t_{n_i} pour le signal modifié.

Il est important de mentionner que la propriété d'invariance est déterminée à partir du signal original.

Trouver la solution optimale de l'équation (4) n'est pas un problème simple du fait du caractère non-linéaire de l'équation. Pour résoudre l'équation (4) nous n'avons pas utilisé d'algorithmes de résolution non-linéaire, car ils sont généralement coûteux en temps de calcul et aussi parce qu'ils n'assurent pas de trouver une solution optimale. C'est la raison pour laquelle nous nous avons utilisé un algorithme d'optimisation itératif. Cet algorithme peut être résumé ainsi:

initialiser tous les ϕ_k pour $k > 0$ à 0
répéter
  i=1
  répéter
    résoudre l'équation (4) par rapport à ϕ_i en supposant que tous les ϕ_k, $k \neq i$ sont constants, en utilisant une technique simplifiée de recuit simulé.
    i = i + 1
  jusqu'à ce que i > K_0
jusqu'à ce que $E(\phi)$ soit stable

Généralement, les amplitudes des composantes $A(\beta t_{n_i}, \omega_k)$ diminuent avec k. La correction des phases, par conséquent, est plus importante pour de petites valeurs de k qui concentre la plus grande partie de l'énergie des signaux de parole. Nous avons pris le parti d'utiliser cette propriété pour limiter la correction des phases aux K_0 premières valeurs. Cela permet de réduire le temps de calcul sans compromettre les résultats.

Nous avons choisi une version simplifiée de l'algorithme de recuit simulé parce que $E(\phi)$ présente très peu de minima, qui par ailleurs sont bien distincts lorsqu'on fait varier ce dernier par rapport à une phase bien précise ϕ_i. En pratique, cet algorithme converge dans tous les cas et préserve de manière adéquate la propriété d'invariance de forme.

4. Mise en Oeuvre de l'algorithme de Synchronisation

Dans [7] nous ne synchronisions les phases qu'au début des zones voisées. Cette technique avait deux inconvénients majeurs. Elle pouvait nécessiter une place mémoire importante, proportionnelle à la taille de la zone voisée traitée et de plus la qualité du son ralenti était moindre lorsqu'il contenait de longs segments voisées. Afin de nous affranchir de ces inconvénients nous avons implémenté une méthode ad-hoc permettant d'effectuer une synchronisation répétée. Le danger d'une telle technique est qu'elle peut engendrer des bruits parasites aux voisinages des instants de synchronisation parce qu'il risque d'y avoir un problème de raccordement des phases.

Il faut noter que dans ce paragraphe tous les temps sont relatifs à l'échelle des temps du signal analysé. Soit t_o le nouvel instant de synchronisation fourni par l'ADM. Soit Φ_{curr} le nouveau vecteur de synchronisation. Et soit Φ_{prev} le vecteur de synchronisation qui prévalait avant t_o. Pour synthétiser le signal ralenti juste avant t_0, nous avons utilisé comme vecteur de synchronisation Φ_{prev}. Pour synthétiser le signal ralenti de t_o à $t_o + TS$ où TS est un seuil temporel que nous avons fixé à $16\,ms$ —il est impératif que ce seuil ne soit pas trop petit pour éviter toute forme de distorsion auditive— nous avons utilisé des vecteurs de décalage interpolés linéairement:

$$\Phi_t = \Phi_{curr} + \frac{t - t_o - TS}{TS}(\Phi_{curr} - \Phi_{prev}) \quad (6)$$

Enfin pour synthétiser le signal ralenti après $t_o + TS$, nous avons utilisé Φ_{curr} comme vecteur de décalage.

5. Quelques Détails d'Implémentation

Les vecteurs de coefficients DFT du vocoder de phase sont stockés dans deux buffers qui se remplissent alternativement en fonction du temps. Nous avons utilisé cette technique afin de diminuer la place mémoire nécessaire pour réaliser le vocoder de phase. Ainsi il n'est pas indispensable de stocker tous les vecteurs DFT. Deux buffers sont aussi nécessaires, pour réaliser l'opération de convolution entre les coefficients DFT et la réponse impulsionnelle du filtre de Lagrange. À ce sujet, il est important de mentionner qu'un soin tout particulier doit être apporté, pour réaliser la convolution exacte surtout en début de buffer courant, en tenant compte des derniers vecteurs du buffer précédent.

6. Résultats Expérimentaux

Grâce aux techniques que nous venons de présenter, nous avons pu éliminer pour une grande partie le problème du «phasiness» qui corrompt les signaux issus du vocoder de phase. La démonstration que nous proposons dans la page web :
http://www.loria.fr/~jdm/PhaseVocoder/index.html
montre clairement que pour des facteurs d'expansion allant jusqu'à 2 la réverbération est pratiquement éliminée. Au delà, une certaine forme de réverbération résiduelle persiste toutefois.

7. Discussion

Un des systèmes qui rivalise avec le notre est le système proposé par Quatieri et al. [12]. Les deux systèmes utilisent le même principe, à savoir, la préservation de l'invariance de forme. Mais les avantages et les inconvénients des deux systèmes ne sont pas

identiques. Le système proposé par Quatieri nécessite peu de puissance de calcul. En revanche pour des facteurs d'expansion élevés, les sons vocaux perdent leur caractère naturel. Une sorte d'effet «d'ivresse» apparaît dans les signaux synthétisés lorsque des facteurs d'expansion importants sont utilisés. En ce qui concerne notre système, la charge de calcul est importante parce que chaque échantillon donne lieu à un calcul de transformée de Fourier. Mais la qualité et le caractère naturel des sons sont parfaitement maintenus, même pour des facteurs d'expansion pouvant aller jusqu'à 5.

8. Conclusion

L'algorithme de vocoder de phase que nous avons modifié élimine très efficacement la réverbération des signaux de parole ralentis. Pour cela nous avons utilisé une méthode de correction de phase qui est répétée à intervalles de temps réguliers sur les régions voisées du signal de parole. Cette méthode donne d'excellents résultats dans le cas d'un ralentissement de la parole. Nous travaillons maintenant sur la modification simultanée du débit et de la fréquence fondamentale en réduisant les artéfacts acoustiques.

Références

[1] E. Moulines and J. Laroche, *"Non Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech"*, Speech Communication, Vol. 16, pp. 175-205, February 1995.

[2] J. Laroche, *"Improved Phase Vocoder Time-Scale Modification of Audio"*, IEEE Transactions on Speech and Audio Processing , Vol. 7, NO. 3, pp. 323-332, May 1999.

[3] J. Laroche, *"Time and Pitch Scale Modification of Audio Signals"*, in Applications of Digital Signal Processing to Audio and Acoustics, Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer, 1998.

[4] R. Portnoff, *"Time-Scale Modifications of Speech Signals Based on Short-Time Fourier Analysis"*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 29, pp. 374-390, 1981.

[5] S. Seneff, *"System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction"* IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-30, NO. 4, pp. 566-578, August 1982.

[6] J. di Martino, *"Speech Synthesis Using Phase Vocoder techniques"*, Proceedings of the 5th European Conference on Speech Communication and Technology - Eurospeech , Rhodes (Greece) , Sept. 1997 Eurospeech-97.

[7] J. di Martino and Y. Laprie, *"Suppression of Phasiness for Time-Scale Modifications of Speech Signals Based on a Shape Invariance Property"*, Proceedings of the International Conference on Acoustics, Speech and Signal procession, Salt-Lake City, USA, 2001.

[8] A. V. Oppenheim and R. W. Schaffer, *"Digital Signal Processing"*, Englewood Cliffs, NJ: Prentice-Hall, 1975.

[9] R. W. Schaffer and L.R. Rabiner, *"A Digital Signal Processing Approach to Interpolation"*, Proceedings of the IEEE, Vol. 61, No. 6, pp. 692-702, June 1973.

[10] Y. Laprie and V. Colotte, *"Automatic Pitch Marking for Speech Transformations Via TD-PSOLA"*, Proceedings of the IX European Signal Processing Conference, Rhodes (EUSIPCO), Greece, 1998.

[11] R. J. McAuley and T. F. Quatieri, *"Speech Analysis/Synthesis Based on a Sinusoidal Representation"*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, NO. 4, pp. 744-754, August 1986.

[12] T. F. Quatieri and R. J. McAuley, *"Shape Invariant Time-Scale and Pitch Modification of Speech"*, IEEE Transactions on Signal Processing, Vol. 40, NO. 3, pp. 497-510, March 1992.

[13] T. F. Quatieri and R. J. McAuley, *"Audio Signal Processing Based on Sinusoidal Analysis/Synthesis"*, in Applications of Digital Signal Processing to Audio and Acoustics, Kahrs and K. brandenburg, Eds. Boston, MA: Kluwer, 1998.

[14] T.F. Quatieri, R.B. Dunn, and T.E. Hanna, *"A Sub-band Approach to Time-Scale Expansion of Complex Acoustic Signals"*, IEEE Transactions on Acoustics, Speech, and Signal Processing, VOL. 3, NO. 6, pp. 515-519, November 1995.