Bases théoriques et expérimentales pour une nouvelle méthode de séparation des composantes pseudo-harmoniques et bruitées de la parole

Laurent GIRIN

Institut de la Communication Parlée, INPG/Univ. Stendhal/CNRS UMR 5009
46 av. Félix Viallet, 38040 Grenoble, France
Tél: +33 476 57 47 15, fax: +33 476 57 47 10, email : girin@icp.inpg.fr
Web : http://www.icp.inpg.fr/~girin/

ABSTRACT

In this paper, the problem of separating the harmonic and noise components of speech signals is addressed. A new method is proposed, based on two specific processes dedicated to better take into account the non-stationarity of speech signals: first, a period-scaled synchronous analysis of spectral parameters (amplitudes and phases) is done, referring to the Fourier series expansion of the signal, as opposed here to the typically used Short-Term Fourier Transform (STFT). Second, the separation itself is based on a low-pass filtering of the parameters trajectory. Preliminary experiments on synthetic speech show that the proposed method has the potential to significantly outperform a reference method based on STFT: Signal-to-error ratio gains of 5 dB are typically obtained. Conditions to go beyond the theoretical framework towards more practical applications on real speech signals are discussed.

1. INTRODUCTION

Les composantes du signal de parole peuvent être grossièrement classifiées en deux catégories, selon la nature de la source vocale : d'un côté les composantes *harmoniques* (H) sont générées par les vibrations des cordes vocales, et d'un autre côté, les composantes *bruitées* (B) sont générées par une source de bruit fricatif, plosif ou d'aspiration [1]. Comme les sources H/B peuvent être simultanées, ces composantes sont souvent mélangées au niveau de la réalisation acoustique du signal. Pour un son donné, la contribution respective de ces composantes peut être quantifiée par l'estimation d'un rapport (de puissance) harmonique à bruit (RHB) (voir une revue dans [2]). Plus difficile, la *séparation* complète des composantes harmoniques et bruitées du signal mixte est un challenge important dans un certain nombre d'applications de traitement de la parole [3]–[5] (et aussi de la musique [6]). L'objectif est d'obtenir deux signaux à partir du signal original : un signal estimé complètement voisé et un autre complètement non-voisé, tels que la somme des deux soit égale au signal original. Ces deux signaux peuvent alors être séparément analysés, modélisés, et modifiés, en particulier pour la synthèse [7], le codage [8], et l'étude fondamentale de la production de parole.

Plusieurs méthodes ont été proposées dans la littérature pour l'estimation du RHB [2], et pour la séparation H/B [3]–[6]. Les méthodes en fréquence sont quasiment toutes basées sur la transformée de Fourier à court-terme (TFCT) pour l'analyse et la synthèse : grosso modo, les pics dominants du spectre sont supposés correspondre aux harmoniques, et les régions du spectre « irrégulières » ou entre les pics sont supposés correspondre aux composantes de bruit. Une telle approche est limitée par un facteur crucial : la parole est un signal localement quasi-stationnaire, et non strictement stationnaire. Ceci signifie que les composantes harmoniques et bruitées évoluent continûment dans le temps, plus ou moins lentement, et c'est une difficulté majeure pour une méthode précise de séparation H/B de ne pas considérer l'évolution des harmoniques d'une période à la suivante comme une partie des composantes bruitées [2]. Pourtant, dans la littérature, les fenêtres d'analyse-synthèse comprennent plusieurs périodes de signal, et la TFCT est intrinsèquement un processus moyenneur qui ne capture pas les différences précises entre ces périodes mais qui extrait plutôt leur caractéristiques communes pour les identifier à des composantes harmoniques constantes sur la fenêtre.

Dans cette étude, nous posons les bases d'une nouvelle méthode de séparation H/B travaillant à l'échelle de la période de signal, de façon à suivre précisément l'évolution des paramètres du signal d'une période à l'autre. C'est une méthode travaillant à la fois en temps et en fréquence, car elle se fonde sur la décomposition en séries de Fourier de chaque période de signal (au lieu de l'habituelle TFCT). La séparation résulte d'un filtrage des trajectoires des paramètres analysés. C'est pourquoi la méthode s'appelle Séparation H/B par filtrage des trajectoires de paramètres spectraux période-synchrones (FTPS2).

Cet article est organisé de la façon suivante. La méthode de séparation H/B est présentée en Section 2. La méthodologie de test est donnée à la Section 3. Résultats et perspectives sont données aux Sections 4 et 5.

2. LA METHODE FTPS2

2.1. Principe général

Soit un signal mixte voisé / non voisé, comprenant un grand nombre K de (pseudo-)périodes $s_k(n)$. Chacune de ces K périodes est décomposée au sens des séries de Fourier réelles, comme une somme de cosinus harmoniques :

$$s_k(n)=\sum_{i=1}^{I} A_i^k \cos\left(i\omega_0^k n+\theta_i^k\right) \quad k = 1 \text{ à } K \qquad (1)$$

Le signal complet est donc représenté par I jeux de K amplitudes A_i^k et phases relatives θ_i^k ($i = 1$ à I, $k = 1$ à K), plus un jeu de la fréquence fondamentale ω_0^k, $k = 1$ à K. Pour un signal pseudo-périodique, l'évolution des amplitudes et des phases d'une période à la suivante doit être plutôt lente ou « lisse », à cause de la nature

déterministe du signal. Au contraire, les composantes apériodiques/bruitées ont une nature aléatoire, et les paramètres spectraux associés (en particulier les phases) doivent varier de façon beaucoup plus importante [8]. Puisque les paramètres sont extraits sur un signal mixte H/B, leur évolution prend typiquement la forme d'une trajectoire de fond lente/lisse, supposée due aux composantes pseudo-harmoniques, à laquelle se superpose un bruit de type additif, supposé dû aux composantes bruitées. Par conséquent, l'extraction du signal harmonique à partir du signal mixte se fait en extrayant la trajectoire de fond lisse des paramètres par un filtrage passe-bas, et en l'identifiant à celle des paramètres des composantes harmoniques. Le signal harmonique estimé est alors généré en appliquant (1) avec les paramètres filtrés à la place des paramètres initiaux. Finalement, le signal de bruit est estimé en soustrayant le signal harmonique estimé au signal mixte. Il est important de noter que cette technique est en fait équivalente à un moyennage *glissant* qui respecte la dynamique des paramètres *à l'échelle de la période*. Ceci s'oppose au moyennage brut de la TFCT déjà mentionné. On cherche dans cette étude à retrouver la « vraie » trajectoire des paramètres harmoniques à partir de mesures perturbées par les composantes bruitées, et à l'inverse de la resynthèse par TFCT inverse, la méthode proposée garantit de reconstruire un signal harmonique estimé qui évolue d'une période à l'autre.

2.2. Détails techniques

Analyse des paramètres : la méthode proposée suppose que le signal à traiter est d'abord segmenté en périodes successives. Dans cette étude, les tests portent sur des signaux synthétiques (voir Section 3.1). Les frontières de périodes (*pitch-marks*) sont donc exactement contrôlées et utilisées dans le processus d'analyse. Dans l'optique d'une extension future aux signaux réels, différentes méthodes peuvent être utilisées pour estimer automatiquement les *pitch-marks*. Clairement, la précision de la méthode de séparation H/B dépend fortement de celle de cette estimation. Nous n'insistons pas sur ce point dans cet article car nous focalisons sur les bases théoriques de la méthode et sur de premières confirmations expérimentales du bien-fondé de ces bases. Cependant, des solutions pour dépasser cette difficulté sont proposées à la Section 5. Ainsi, dans cette étude, ω_0^k est donnée directement par l'inverse de la longueur de la période k. Puis, étant donnée ω_0^k, les amplitudes A_i^k et les phases θ_i^k sont estimées en utilisant la procédure de George et Smith de [9]. Cette procédure est une minimisation itérative au sens des moindres carrés de l'erreur entre le modèle harmonique de (1) et le signal. Elle permet d'obtenir une estimation précise des paramètres avec un très faible coût de calcul.

Régularisation de la phase: les mesures de phase sont obtenues modulo 2π. Il faut donc d'abord s'assurer qu'aucun saut de 2π artificiel ne vient perturber leur trajectoire « naturelle ». Pour cela, une procédure de régularisation de ces mesures le long de l'axe temporel est appliquée. Elle consiste à ajouter ou retrancher itérativement 2π à chaque mesure de phase si cela permet de diminuer la variance du vecteur compilant les mesures. Comme, la trajectoire de fond évolue au cours du temps, la variance est calculée avec une fenêtre glissante (typiquement, quatre périodes peuvent être utilisées) et plusieurs passes peuvent être effectuée. A la fin de cette procédure, les mesures sont parfaitement homogènes, bien que toujours bruitées.

Filtrage des paramètres : comme expliqué en Section 2.1, l'étape suivante composant le cœur de la méthode est le filtrage passe-bas des trajectoires des paramètres spectraux (amplitudes et phases). Des tests pilotes ont montrés qu'une large gamme de filtres très simples (*i.e.*, FIR avec peu de coefficients) fournissait des résultats assez proches. Dans les expériences de la Section 4, on utilise un filtre FIR à 10 coefficients avec une fréquence de coupure de 0.1 obtenu par la méthode de fenêtrage avec une fenêtre rectangulaire. Ce filtre est appliqué en mode *forward-backward* (filtrage à phase nulle), de façon à ce que les paramètres filtrés et non filtrés soient synchrones, et qu'il en soit de même pour les signaux H/B séparés et le signal mixte original.

Ré-estimation de l'amplitude : En pratique, on observe que les trajectoires des paramètres d'amplitude sont généralement plus bruitées que celles des paramètres de phase. Par conséquent, la méthode a été raffinée avec une ré-estimation des amplitudes après le filtrage des paramètres de phase. Ceci est fait avec une version simplifiée de la procédure d'analyse des moindres carrés itératifs utilisée précédemment, avec la phase maintenant fixée à la valeur obtenue après filtrage. Les amplitudes ainsi ré-estimées sont ensuite filtrées par le filtre passe-bas.

3. METHODOLOGIE DE TEST

3.1. Génération de signaux synthétiques

Des signaux synthétique mixtes voisés / non voisés sont générés de façon à ce que les « vraies » parties harmoniques et bruitées soient séparément disponibles. Ceci permet de calculer des mesures objectives de séparation, telles que le rapport signal à erreur (RSE) [2][4] que nous utilisons par la suite (voir la sous-section 3.3). Les signaux synthétiques consistent en différentes versions des voyelles /a/ d'une voix masculine et /i/ d'une voix féminine, prononcées de manière soutenue ($K = 300$) et échantillonnées à 48kHz (avec une bande passante limitée à 8 kHz). Leur génération suit la méthodologie utilisée dans [2][4]. Un train d'ondes glottales suivant le modèle de Rosenberg [10] est utilisé comme source harmonique. Un bruit blanc gaussien est utilisé comme source de bruit. Il est éventuellement modulé en amplitude par le train d'ondes glottales pour plus de naturel [7]. Les deux sources alimentent un filtre numérique tout-pôles modélisant le conduit vocal. Ce filtre résulte de l'analyse LPC à l'ordre 50 d'un signal réel produit par un locuteur masculin pour le /a/ et par une locutrice pour le /i/. Des filtres du premier ordre pour la pré-emphase et la simulation de radiation labiale sont aussi utilisés pour mieux caler le spectre du signal synthétique sur celui des signaux réels et permettre un son plus naturel. Le signal mixte est obtenu en sommant les deux signaux filtrés et centrés avec différents RHB dans la gamme –10 dB à 30 dB. Enfin, pour évaluer la robustesse de la méthode proposée sur des signaux non-stationnaires et plus proches du naturel, de la prosodie est générée par modulation de la fréquence fondamentale de la source glottale selon la formule :

$$\omega_0^k = \alpha + \beta \cos\left(2\pi \frac{3k}{K}\right) + \gamma \frac{k^2}{K^2} \qquad (2)$$

Le terme en cosinus assure trois cycles mélodiques et le terme quadratique assure une montée rapide à la fin de la voyelle. A la Section 4, on fournit des résultats pour une fréquence fondamentale fixe (*i.e.*, pour /a/, $\alpha = 130$, $\beta = \gamma = 0$; pour /i/, $\alpha = 280$, $\beta = \gamma = 0$), pour une intonation « normale » (*i.e.*, pour /a/, $\alpha = 130$, $\beta = 10$, $\gamma = 20$; pour /i/, $\alpha = 250$, $\beta = 10$, $\gamma = 20$), et pour une intonation « exagérée » (*i.e.*, pour /a/, $\alpha = 110$, $\beta = 30$, $\gamma = 100$; pour /i/, $\alpha = 200$, $\beta = 30$, $\gamma = 200$) (toutes les valeurs sont en Hz).

3.2. Une méthode de référence basée sur la TFCT

Pour évaluer comparativement notre méthode, nous avons implanté la méthode *Pitch-Scaled Harmonic Filter* (PSHF) de [4]. Cette méthode a été choisie car 1) elle est bien représentative des méthodes basées sur la TFCT 2) elle est relativement simple à implanter par rapport à d'autres méthodes [5] 3) son évaluation sur des signaux synthétiques a fourni des scores de RSE de référence. Son principe est de calculer des spectres successifs par TFCT sur exactement quatre périodes du signal mixte, de façon à ce que les pics des harmoniques soient supposés être localisés tous les quatre canaux spectraux de la TFCT et donc facilement isolés par un filtre peigne. Quatre périodes du signal harmonique estimé sont alors générées par TFCT inverse du spectre filtré par le peigne. Le signal harmonique complet est reconstruit par sommation pondérée des estimations successives. En le soustrayant au signal mixte, on obtient le signal de bruit estimé.

3.3. Mesures de Rapport Signal à Erreur (RSE)

L'évaluation objective de la séparation H/B est faite par calcul de rapport signal à erreur (RSE) qui peut être fait aussi bien pour le signal harmonique estimé que pour le signal de bruit estimé. Notons RSE_H le rapport de puissance entre le signal harmonique « vrai » et sa différence d'avec le signal harmonique estimé. De même, notons RSE_B le même rapport pour le signal de bruit. Comme le signal de bruit estimé est obtenu en soustrayant le signal harmonique estimé au signal mixte, on a : $RSE_H = RSE_B + RHB$. Par conséquent, par la suite nous ne considérons que RSE_B (noté simplement RSE), du fait qu'il s'est révélé quasiment constant en fonction du RHB dans [4].

4. RESULTATS

4.1. Rapports Signal à Erreur

La Figure 1 montre les RSE obtenus sur les voyelles de test, avec les deux méthodes, FTPS2 et PSHF, et pour les trois contours de ω_0. Les principaux résultats sont les suivants :

- Les deux méthodes fournissent des résultats remarquablement stables sur une large gamme de RHB : les RSE sont quasi-constants de –10 à environ 15 dB de RHB dans presque tous les cas. Pour la méthode PSHF, le RSE est d'environ 5 dB (de 5 à 5.4 dB dans la gamme de –10 à 15 dB de RHB, selon les conditions) et ce résultat est très cohérent avec ceux de [4], qui donnent une valeur stable typique de 5 dB.
- Les performances obtenues avec la méthode FTPS2 dépassent largement cette référence de 5 dB. Pour –10 à 15 dB de RHB, les valeurs de RSE sont toutes autour de 9.5 dB pour /a/ (sauf pour l'intonation exagérée) et autour de 10.5 dB pour /i/ (du moins pour l'intonation normale et exagérée; de façon surprenante, une valeur légèrement inférieure de 10 dB est obtenue pour ω_0 constant). Ainsi, la méthode FTPS2 à un gain de l'ordre de 4 à 5.5 dB par rapport à la méthode PSHF, selon les conditions. Un exemple typique de résultat est donné sur la Figure 2.
- Les performances des deux méthodes chutent pour un RHB supérieur à 15 dB, et plus la variation d'intonation est forte, plus la dégradation est forte. Ce résultat n'est pas surprenant : plus la partie bruitée du signal est faible, plus elle est difficile à séparer de la partie harmonique, et l'augmentation de la non-stationnarité du signal rend la tâche encore plus difficile. On peut aussi remarquer que les deux voyelles offrent une robustesse différente à ces dégradations, mais une discussion sur l'influence de facteurs phonétiques va au-delà du cadre de cet article. Notons que même dans des conditions difficiles, l'avantage de la méthode FTPS2 sur la méthode PSHF reste toujours supérieur à 4 dB, sauf pour /i/ avec intonation exagérée à 25–30 dB, où « seulement » 3.7 et 3.1 dB de gain sont obtenus. Pour les autres conditions, ce gain est typiquement de 5 dB, et il peut même aller au-delà, par exemple 8 dB pour /a/ avec ω_0 fixe à 30 dB.
- Finalement, on peut noter que les résultats obtenus avec ou sans la modulation de la source de bruit par la source glottale sont toujours très proches. C'est pourquoi on présente sur la Figure 1 seulement les résultats obtenus sans modulation. De façon complémentaire, sur l'exemple de la Figure 2, le signal de bruit est modulé. Ces résultats semblent indiquer que les deux méthodes sont assez robustes par rapport aux non-stationarités de la source de bruit. Ce point est important en regard de l'application de la séparation H/B sur des signaux réels et devra être étudié plus en détails dans le futur.

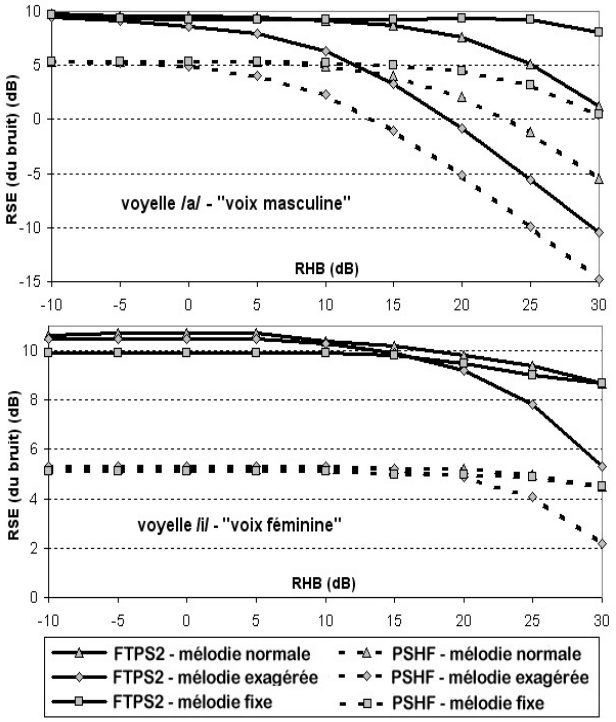

Figure 1 : RSE (du signal de bruit) en fonction du RHB.

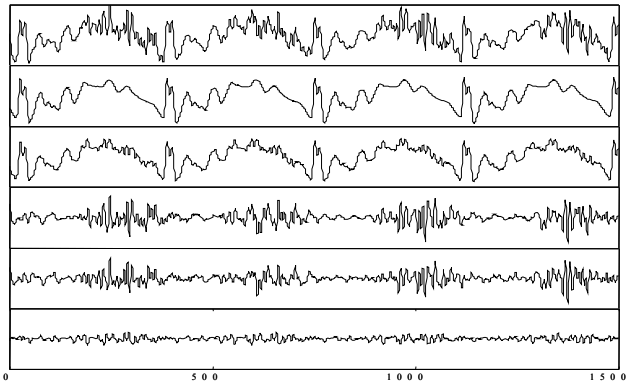

Figure 2 : Exemple de séparation H/B avec la méthode FTPS2 : voyelle /a/ avec modulation de la source de bruit et RHB = 0 dB. De haut en bas : signal mixte, signal H « vrai » et estimé, signal B « vrai » et estimé, différence entre signal « vrai » et estimé. L'axe des Y est arbitraire mais homogène entre les différentes figures. On obtient ici RSE = 9 dB.

4.2. Tests d'écoute informels

L'écoute des signaux a confirmé les bonnes performances de la méthode FTPS2, et l'amélioration par rapport à la méthode PSHF. Pour des RHB de 0 à 30 dB, le signal harmonique estimé avec la méthode FTPS2 n'est généralement pas distinguable du « vrai » signal harmonique, alors qu'il reste généralement un résidu de bruit significatif dans le signal harmonique estimé avec la méthode PSHF. La méthode PSHF semble bien souffrir du fait que les valeurs échantillonnées tous les quatre canaux de TFCT ne correspondent pas forcément exactement aux pics harmoniques si le signal est non-stationnaire. Pour des RHB faibles, les signaux séparés sont généralement de moins bonne qualité, c'est-à-dire moins proches des « vrais » signaux harmonique et de bruit, avec des qualités différentes pour les deux méthodes. Tous les signaux testés (« vrais » H et B, mixtes et séparés) sont disponibles online : www.icp.inpg.fr/~girin/HNS/HNS_demo.zip. Le lecteur est invité à se faire son propre jugement.

5. DISCUSSION

Bien qu'encourageants, les résultats précédents doivent être considérés prudemment, à cause de la dépendance de la méthode sur la précision des *pitch-marks* déjà mentionnée. En particulier, on peut s'attendre à ce que les mesures de phase des harmoniques de rang élevé soient significativement perturbées par les imprécisions du *pitch-marking*, puisque la variation de phase est égale à l'intégration temporelle de la fréquence. Cependant, ces limites doivent être fortement pondérées par deux points qui constituent le noyau de nos travaux actuels :

- D'abord, le filtrage passe-bas des paramètres spectraux pourrait intrinsèquement compenser ce bruit de mesure additif. En d'autres termes, le filtrage pourrait éliminer à la fois le bruit dû aux composantes bruitées de la parole et le bruit dû aux imprécisions de l'analyse. Ceci est vrai tant que la somme des contributions de ces bruits n'empêchent pas la trajectoire de fond des paramètres de phase d'émerger. Une étude plus poussée est nécessaire pour clarifier ce point. En particulier, l'influence de l'estimation automatique des *pitch-marks* doit être analysée, ainsi que les interactions entre les deux sources de bruit (mesures et parole elle-même).
- Ensuite, les *phases relatives* considérées dans cette étude peuvent être remplacées par les *phases absolues*, c'est-à-dire les valeurs de phase résultant de l'intégration temporelle des fréquences. En effet, à l'inverse des trajectoires de phase relative, les trajectoires de phase absolue peuvent être reconstruite à partir de mesures effectuées à des instants arbitraires. A l'inverse de la procédure de régularisation de la Section 2.2, cette reconstruction des trajectoires de phase absolue nécessite une procédure duale de dépliement [11] assez simple à implanter. Ainsi, l'estimation d'une trajectoire de phase absolue lisse à partir de mesures bruitées doit pouvoir conduire à un résultat au moins équivalent, avec l'avantage déterminant de ne pas dépendre des instants de mesure, comme les *pitch-marks* dans l'étude présente (par contre, la taille de la fenêtre d'analyse doit rester de l'ordre d'une période de signal pour capturer finement son évolution). Notons que le lissage des trajectoires de phase absolues peut aussi être obtenu par un filtrage passe-bas, mais aussi par des techniques alternatives comme la modélisation à long terme proposée dans [12].

6. REFERENCES

1. Stevens, K. N. *Acoustic phonetics*, MIT Press, Cambridge, MA, 1998.
2. Murphy, P. Perturbation-free measurements of the harmonics-to-noise ratio in voice signals using pitch-synchronous harmonic analysis, *J. Acoust. Soc. Am.*, 105(5), 2866-2880, 1999.
3. Stylianou, Y. Decomposition of speech signals into a deterministic and a stochastic part, *Proc. Int. Conf. Spoken Language Proc.*, Philadelphia, 1996.
4. Jackson, P. & Shadle, C. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech, *IEEE Trans. Speech Audio Proc*., 9(7), 713-726, 2001.
5. Yegnanarayana, B., d'Alessandro, C. & Darsinos, V. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, 6(1), 1-11, 1998.
6. Serra, X. & Smith, J. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic + stochastic decomposition, *Comp. Music J*., 14(4), 12-24, 1990.
7. Hermes, D. J. Synthesis of breathy vowels: Some research methods, *Speech Communication*, 10, 497-502, 1991.
8. Kang, G. & Everett, S. Improvement of the excitation source in the narrow-band linear prediction vocoder, *IEEE Trans. Acoust. Speech Sig. Proc.*, 33(2), 377-386, 1985.
9. George, E. & Smith, M, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.,* **5**(5), 389-406, 1997.
10. Rosenberg, A. E., Effect of glottal pulse shape on the quality of natural vowels, *J. Acoust. Soc. Am*. 49(2), 583-590, 1971.
11. R. J. McAulay & T. F. Quatieri, Speech analysis/ synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, 34(4), 744-754, 1986.
12. L. Girin, M. Firouzmand & S. Marchand, "Long term modeling of phase trajectories within the speech sinusoidal model framework," *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, 2004.