Séparation des sources acoustiques de la parole : description de la méthode X-GLOS

Cette section décrit le principe de la méthode X-GLOS [19] pour séparer les sources glottiques des sources supraglottiques de la parole.

Modélisation des sources

Le signal de la parole $s(t)$ est modélisé comme la somme d'une composante périodique $s_p(t)$ et d'une composante apériodique (bruitée) $s_n(t)$:

$$s(t) = s_p(t) + s_n(t) \quad (1)$$

La composante périodique est une somme de M sinusoïdes :

$$s_p(t) = \sum_{m=1}^{M} b_m(t) e^{i\phi_m(t)}, \quad (2)$$

où $b_m(t)$ est l'amplitude des sinusoïdes, et où les phases $\phi_m(t)$ sont données par :

$$\phi_m(t) = 2\pi \int_0^t m f_0(\tau) d\tau + \phi_{0m}. \quad (3)$$

Pour la suite de l'article, nous considérons des signaux discrets $s[n]$, échantillonnés uniformément à la fréquence d'échantillonnage F_e.

Parmi les méthodes précédentes de décomposition périodique/apériodique des signaux de parole [20, 21], l'analyse est effectuée sur des courtes fenêtres temporelles recouvrantes dans lesquelles le signal est considéré comme stationnaire. La séparation des sources est alors effectuée indépendamment pour chacune des fenêtres d'analyse. À l'instar de ces méthodes, nous conservons le principe du découpage en fenêtres d'analyse, mais nous nous affranchissons de l'hypothèse de stationnarité du signal de parole qui est une hypothèse très forte. En effet, le signal de parole est par définition non-stationnaire, même à l'intérieur de courtes fenêtres d'analyse. Les principales variations concernent la fréquence fondamentale et l'amplitude des sinusoïdes, du fait des phénomènes de jitter (variation de la période fondamentale d'un cycle à l'autre) et de shimmer (variation d'amplitude d'un cycle à l'autre). La méthode X-GLOS prend en compte le jitter, qui est un indice acoustique important de la parole expressive. Il est notamment utilisé pour détecter des voix pathologiques [22].

Pour la méthode X-GLOS, le signal est donc découpé en fenêtre de Hann de longueur L, et la séparation est appliquée à chacune des trames temporelles d'indice k :

$s_k[n] = s[n]h_k[n]$,

où $h_k[n] = h[n-ka]$ et $a = L/4$. Les estimations des composantes périodiques et apériodiques sont alors respectivement notées $\tilde{s}_{k,p}$ et $\tilde{s}_{k,n}$, et sont définies par :

$$s_k[n] = \tilde{s}_{k,p}[n] + \tilde{s}_{k,n}[n], \quad (4)$$

Au sein d'une trame temporelle, nous considèrerons l'amplitude des sinusoïdes comme constantes, ce qui revient à réécrire les équations (2) et (3) par :

$$s_{k,p}[l] = h_k[l] \sum_{m=1}^{M} b_m e^{i\phi_m[l]}, \quad (5)$$

et

$$\phi_m[l] = \frac{1}{F_e} \sum_{i=0}^{l} f_m[l], \quad (6)$$

où l'indice l est l'indice de l'échantillon au sein de la fenêtre k de longueur L et $f_m[l]$ est la fréquence instantanée de la m^{ieme} sinusoïde.

Cette expression peut se mettre sous forme matricielle, ce qui donne pour le signal de parole fenêtré :

$$\boldsymbol{s}_k = \boldsymbol{V}_k \boldsymbol{b}_k + \boldsymbol{s}_{k,n} = \boldsymbol{s}_{k,p} + \boldsymbol{s}_{k,n}, \quad (7)$$

où, $\mathbf{s}_k \in \mathbb{R}^{L\times 1}$, $\mathbf{s}_{k,p} \in \mathbb{R}^{L\times 1}$, et $\mathbf{s}_{k,n} \in \mathbb{R}^{L\times 1}$ sont les vecteurs contenant les L échantillons du signal de parole, des contributions de la source voisée, et des contributions de la source bruitée, respectivement,

$\mathbf{b}_k \in \mathbb{C}^{M\times 1} = [b_1, b_2, \ldots, b_M]^T$ est le vecteur contenant les amplitudes complexes des M sinusoïdes, et où

$\mathbf{V}_k \in \mathbb{C}^{L\times M}$ est une base du sous-espace signal engendré par les sinusoïdes de fréquence variable, donnée par :

$$\mathbf{V}_k = \mathbf{H}\begin{bmatrix} e^{j\phi_1(0)} & e^{j\phi_2(0)} & \dots & e^{j\phi_M(0)} \\ e^{j\phi_1(1)} & e^{j\phi_2(1)} & \dots & e^{j\phi_M(1)} \\ e^{j\phi_1(2)} & e^{j\phi_2(2)} & \dots & e^{j\phi_M(2)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\phi_1(L-1)} & e^{j\phi_2(L-1)} & \dots & e^{j\phi_M(L-1)} \end{bmatrix}, \quad (8)$$

où **H** est la matrice diagonale contenant les échantillons de la fenêtre de Hann *h*.

Le principe de la méthode consiste alors à estimer la composante périodique $\tilde{\mathbf{s}}_{k,p} = \tilde{\mathbf{V}}_k \tilde{\mathbf{b}}_k$ en estimant les fréquences instantanées des partiels réellement activés dans le but de construire une estimation de la base du sous-espace signal périodique $\tilde{\mathbf{V}}_k$, et ensuite d'estimer les amplitudes complexes $\tilde{\boldsymbol{b}}_k$. À partir de l'estimation $\tilde{\mathbf{s}}_{k,p}$, il vient alors que l'estimation de la composante apériodique $\tilde{\mathbf{s}}_{k,n}$ est :

$$\tilde{\mathbf{s}}_{k,n} = \mathbf{s}_k - \tilde{\mathbf{s}}_{k,p}.$$

Dans l'article, à moins d'être clairement spécifié, l'indice *k* est retiré des notations par souci de clarté.

Principe de la méthode

La première étape consiste à estimer la fréquence fondamentale. De nombreuses techniques existent pour cela [23, 24] et peuvent très bien s'adapter à X-GLOS. Nous avons néanmoins utilisé notre propre méthode, détaillée dans [25], qui s'appuie sur le calcul d'un périodogramme cumulé pondéré. Ensuite, il nous faut déterminer de manière précise les fréquences des partiels activés, dans le but de construire la base de l'espace signal périodique $\boldsymbol{V}_k$ de l'équation (8) puis de déterminer les amplitudes complexes $\boldsymbol{b}_k$.

Estimation du sous-espace signal périodique

Pour chaque fenêtre d'analyse pour laquelle du voisement a été détecté, à savoir $f_0 > 0$, la fréquence des partiels activés est estimée par l'intermédiaire de la méthode QIFFT (*Quadratically Interpolated FFT*) [26], qui consiste à estimer les paramètres de sinusoïdes de manière analytique à partir de l'interpolation polynomiale de second ordre du logarithme de l'amplitude de la FFT autour des pics spectraux. Pour une estimation non-biaisée, la QIFFT est utilisée sur une FFT du signal fenêtré par une fenêtre Gaussienne, étant donné que le logarithme de l'amplitude d'une telle fenêtre est réellement un polynôme du second ordre.

En considérant la source voisée comme strictement périodique en première approximation, les pics spectraux devraient être positionnés à des fréquences multiples de la fondamentale. Si ce n'est pas le cas, à savoir si le m^{ième} pic spectral est situé en dehors d'un certain intervalle autour de mf_0, le partiel est considéré comme non-activé, et ne sera pas pris en compte dans le calcul de la matrice $\boldsymbol{V}_k$. Pour cette étude, l'intervalle est choisi arbitrairement comme étant à ± 5 Hz. L'harmonique de référence $f_{m\text{-}1}$ est mis à jour à chaque incrément pour éviter la dispersion de l'erreur sur l'estimation de la fréquence fondamentale pour les partiels d'ordre élevé.

Du fait d'une pente spectrale relativement forte de la source voisée [27], et dans des conditions fortement bruitées, les contributions de la source voisée sont très peu énergétiques devant la densité spectrale du bruit, qui est importante en moyenne et haute fréquence [28]. La fréquence critique au-delà de laquelle les contributions voisées sont considérées comme négligeables est parfois appelée MVF (*Maximum Voiced Frequency*) [24]. Dans le but d'éviter d'incorporer des composantes de bruit dans le sous-espace signal périodique, nous proposons d'effectuer une estimation grossière de la MVF et de considérer l'ensemble des composantes fréquentielles situées au-delà comme des composantes de bruits.

Une fois que les fréquences des partiels activés ont été estimées, la base du sous-espace signal périodique est calculée à l'aide de l'équation (8).

Estimation des amplitudes complexes

Le problème de l'estimation de l'amplitude de sinusoïdes en présence de bruit a été traité par le passé (voir [29] pour un bref état de l'art). La méthode classique des moindres carrés est efficace et très peu coûteuse en temps de calcul. Dans notre cas, l'estimation des amplitudes $\boldsymbol{b}_k$ se fait à l'aide de l'opération suivante :

$$\mathbf{b}_k = \mathbf{V}^{\dagger}\mathbf{s}_k, \quad (10)$$

où $(\,.\,)^{\dagger}$ est l'opérateur pseudo-inverse de Moore-Penrose.

Validation

Les performances de la méthode ont été mesurées à l'aide de signaux de parole simulés. A partir de la connaissance des fonctions d'aire de différentes fricatives, obtenues par IRM (voir [30] pour plus de détails), les signaux des contributions de la source voisée et de la source de bruit de friction ont été simulés séparément à l'aide d'une méthode de synthèse numérique de parole adaptée [5]. Les fonctions d'aire concernent deux points d'articulation, alvéolaire (/z,s/) et post-alvéolaire (/ʒ,ʃ/). Afin d'évaluer la robustesse de la méthode aux non-stationnarités des contributions de la source voisée, un jitter artificiel a été introduit en modifiant la période fondamentale nominale T_0 à chaque instant par une valeur aléatoire suivant une loi normale, de variance adaptée au niveau de jitter que l'on souhaite simuler. Le signal de mélange est ensuite la somme de la source voisée et de la source de bruit qui sont pondérées par un facteur α permettant de régler le niveau relatif des deux sources. Celui-ci est mesuré à l'aide du quotient de voisement, noté *VQ*, qui quantifie le pourcentage des contributions de la source voisée (glottique) dans le signal acoustique produit, à savoir :

$$VQ = 100 \times \frac{||s_p(t)||^2}{||s(t)||^2}. \quad (11)$$

Une valeur de 100% indique un signal purement voisé, une valeur de 0% indique un signal purement non-voisé (apériodique), et une valeur de 50% indique que les contributions de la source glottique et de la source supraglottique possèdent la même énergie.

Pour des soucis de comparaison avec la méthode de référence PSHF [21], nous avons utilisé le même indicateur de performance, à savoir le SER (*Signal-to-Error Ratio*) :

$$\eta_p = 10\log_{10}\left(\frac{||\bar{s}_n||_2^2}{||e||_2^2}\right) \quad (dB)\,, \tag{12}$$

$$\eta_n = 10\log_{10}\left(\frac{||\bar{s}_p||_2^2}{||e||_2^2}\right) \quad (dB)\,, \tag{13}$$

où $e = \tilde{s}_p - \bar{s}_p$ est l'erreur d'estimation.

Pour des raisons de clarté et de concision, nous n'exposons qu'une partie des tests de performance. Le lecteur pourra consulter notre article complet [19] pour les résultats complémentaires.

Comparaison avec PSHF

Nous comparons dans un premier temps notre méthode avec la méthode PSHF [21]. La figure 6 affiche les performances des deux méthodes pour un signal de fréquence fondamentale 120 Hz, sans jitter, pour une fricative post-alvéolaire (en haut) et alvéolaire (en bas), et en fonction du quotient de voisement *VQ*. La figure montre que X-GLOS obtient de meilleurs résultats, à la fois en conditions faiblement et fortement bruitées. Sans filtrage MVF (ligne pleine en bleu, '□'), le gain de performance avec X-GLOS en comparaison avec PSHF est compris entre 2,5 et 5 dB à la fois pour les estimateurs des composantes voisées et bruitées. Le point d'articulation n'a que très peu d'influence sur la performance de la méthode. La méthode semble donc très peu sensible à la coloration du bruit.

Lorsque le filtrage MVF est appliqué, la performance de X-GLOS est améliorée de manière significative. La méthode devient alors très peu sensible au niveau de bruit : entre la condition non-bruitée (*VQ* = 95%) et la condition non-voisée (*VQ* = 5%), le SER de l'estimation de la composante apériodique est diminué de 24 dB avec un filtrage au premier zéro, et de seulement 12 dB avec un filtrage au deuxième zéro, alors que la chute est de 41 dB sans filtrage. Pour la composante périodique, l'application du filtrage améliore la performance de 20 dB pour un filtrage au premier zéro, et de 30 à 40 dB avec un filtrage au deuxième zéro en condition peu voisée, en comparaison avec la performance sans filtrage.

Effet du jitter

Ici, nous mesurons la performance et la robustesse de X-GLOS aux variations de fréquence fondamentale à travers le niveau de jitter. Ces performances sont représentées à la figure 7 pour différentes valeurs de jitter, à savoir 0,5%, 1%, 1,5%, et 3%. Nous mesurons également l'apport du modèle non-stationnaire, par rapport à un modèle localement stationnaire. La performance de l'estimateur apériodique h_n est systématiquement meilleure dans le cas non-stationnaire par rapport au modèle localement stationnaire. C'est également le cas pour l'estimateur périodique h_p. De manière intéressante, alors que la performance avec un modèle localement stationnaire s'abaisse de manière significative avec une augmentation du jitter, cela est beaucoup moins vrai avec le modèle non-stationnaire. L'indicateur de performance h_n perd 4 dB quand le jitter augmente de 0,5 à 3% avec le modèle non-stationnaire, alors qu'il chute de 7 dB avec le modèle localement stationnaire. Par conséquent, l'introduction des variations locales de fréquence dans le modèle de signal effectuée par la méthode X-GLOS permet de rendre la méthode de séparation robuste au jitter de la parole naturelle.

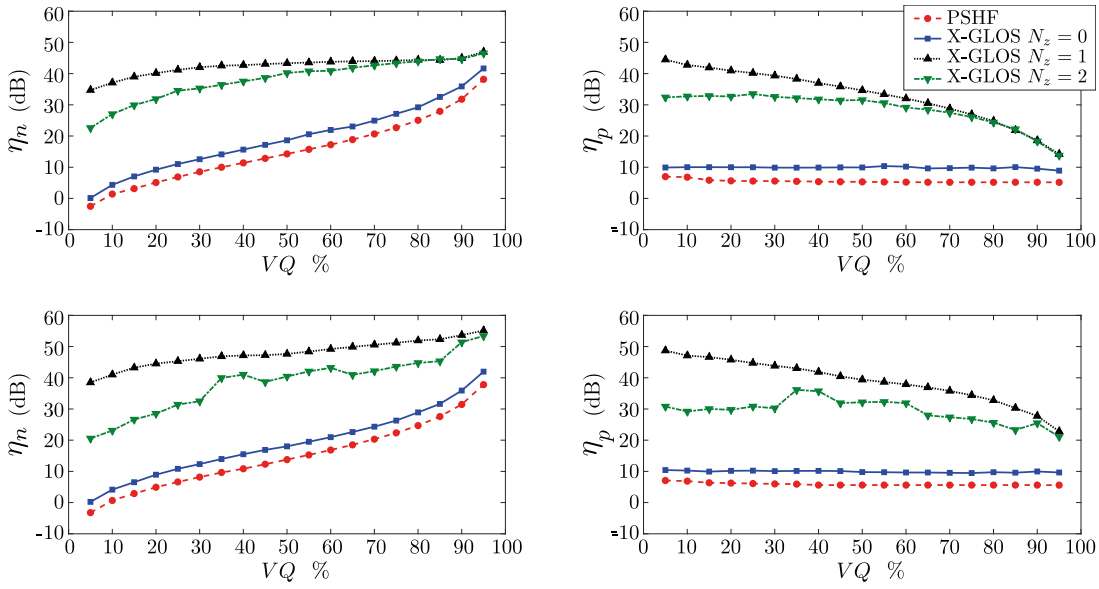

Fig. 6: Performance de X-GLOS sans filtrage MVF (ligne pleine en bleu, □), avec un filtrage MVF au premier zéro (ligne pointillée en noir, Δ), et filtrage MVF au deuxième zéro (pointillée en vert, ▽), pour les composantes apériodique (gauche) et périodique (droite). La performance de PSHF est également tracée pour comparaison (ligne traitillée en rouge, o). (Haut) fricative post-alvéolaire, (bas) fricative alvéolaire

X-GLOS performance without MVF filtering (solid line blue, □), with MVF filtering at first zero (dotted line black, Δ), and MVF filtering at the second zero (dash-dotted line green, ▽), for aperiodic components (left) and periodic (right). The performance of PSHF is plotted for comparison (dashed line red, o). (Top) postalveolar fricative, (bottom) alveolar fricative

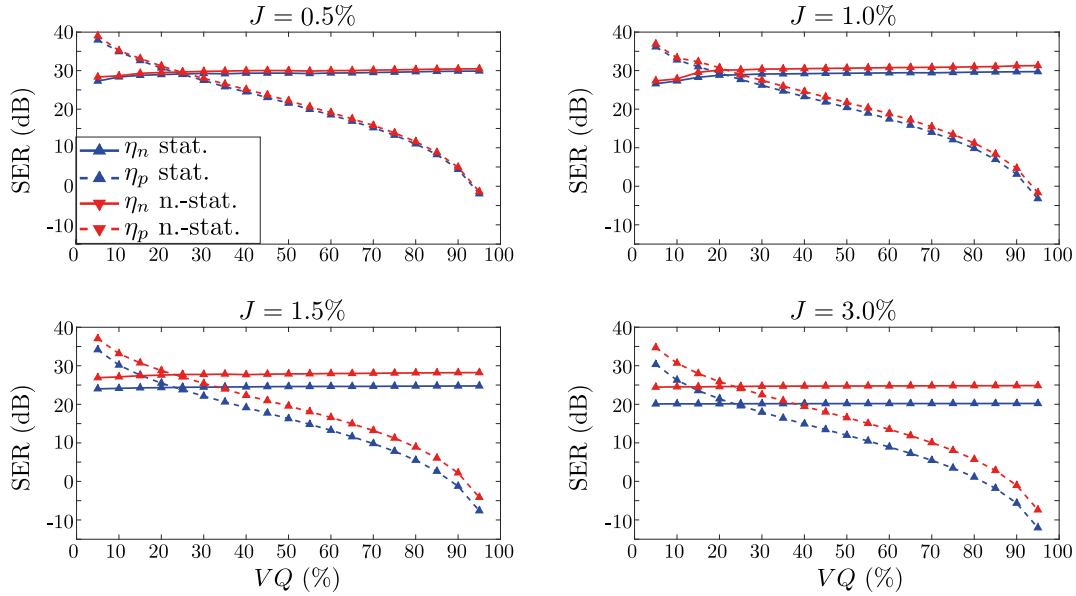

Fig. 7 : **Performances de X-GLOS avec un filtrage MVF au deuxième zéro pour l'estimateur apériodique (ligne pleine) et périodique (ligne traitillée), avec un modèle stationnaire (Δ), et avec le modèle non-stationnaire (Δ), pour différentes valeurs de jitter (0,5%, 1%, 1,5%, et 3%). Par soucis de concision, seule la fricative post-alvéolaire est représentée**
Performance of X-GLOS with second zero MVF filtering for the aperiodic estimate (solid line) and periodic estimate (dashed line), with a stationary model (Δ), and with the non-stationary model (Δ), for different jitter values (0.5%, 1%, 1.5%, and 3%). For the sake of brevity, only the postalveolar fricative is shown

Références bibliographiques

[1] P. Perrier, Y. Payan, S. I. Buchaillard, M. A. Nazari et M. Chabanas, «Biomechanical models to study speech,» Faits de langues, vol. 37, pp. 155-171, Jul 2011

[2] P. Birkholz, D. Jackèl et B. J. Kröger, «Construction and control of a three-dimensional vocal tract model,» chez Proc. Intl. Conf. Acoust., Spch., and Sig. Proc. (ICASSP 2006), 2006

[3] Y. Laprie, B. Vaxelaire et M. Cadot, «Geometric articulatory model adapted to the production of consonants,» chez 10th International Seminar on Speech Production (ISSP), Köln, Allemagne, 2014

[4] B. H. Story, «Phrase-level speech simulation with an airway modulation model of speech production,» Computer Speech & Language, vol. 27(4), pp. 989-1010, 2013

[5] B. Elie et Y. Laprie, «Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink,» Speech Communication , vol. 82, pp. 85-96, 2016

[6] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum et C. Neuschaefer-Rube, «Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis,» Computer Speech & Language, vol. 41, pp. 116-127, 2017

[7] K. Ishizaka et J. L. Flanagan, «Synthesis of voiced sounds from a two-mass model of the vocal cords,» Bell Syst. Tech. J., vol. 51(6), pp. 1233-1268, 1972

[8] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis et A. Hirschberg, «A symetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prothesis design,» Acta Acustica, vol. 84, pp. 1135-1150, 1998

[9] R. S. McGowan, «Tongue-tip trills and vocal-tract wall compliance,» Journal of the Acoustical Society of America, vol. 91(5), pp. 2903-2910, 1992

[10] B. Elie et Y. Laprie, «Simulating alveolar trills using a two-mass model of the tongue tip,» The Journal of the Acoustical Society of America, vol. 142, n° %15, pp. 3245-3256, 2017

[11] I. P. Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999

[12] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton et K. Kurowski, «Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters,» The Journal of the Acoustical Society of America, vol. 91(5), pp. 2979-3000, 1992

[13] L. M. Jesus et C. H. Shadle, «A parametric study of the spectral characteristics of European Portuguese fricatives,» Journal of Phonetics, vol. 30, n° %13, pp. 437-464, 2002

[14] O. Dmitrieva, A. Jongman et J. Sereno, «Phonological neutralization by native and non-native speakers: The case of Russian final devoicing,» Journal of phonetics, vol. 38, n° %13, pp. 483-492, 2010

[15] D. Pape, L. M. Jesus et P. Birkholz, «Intervocalic fricative perception in European Portuguese: An articulatory synthesis study,» Speech Communication, vol. 74, pp. 93-103, 2015

[16] S. S. Narayanan, A. A. Alwan et K. Haker, «An articulatory study of fricative consonants using magnetic resonance imaging,» The Journal of the Acoustical Society of America, vol. 98, n° %13, pp. 1325-1347, 1995

[17] C. H. Shadle, M. I. Proctor, K. Iskarous et M. A. Berezina, «Revisiting the role of the sublingual cavity in the /s/-/ʃ/ distinction.,» Journal of the Acoustical Society of America, vol. 125(4), pp. 2569-2569, 2009

[18] B. Elie et Y. Laprie, «Acoustic impact of the gradual glottal abduction degree on the production of fricatives: A numerical study,» The Journal of the Acoustical Society of America, vol. 142, n° %13, pp. 1303-1317, 2017

[19] B. Elie et G. Chardon, «Glottal/Supraglottal Source Separation in Fricatives Based on Non-Stationnary Signal Subspace Estimation,» 2018

[20] B. Yegnanarayana, C. d'Alessandro et V. Darsinos, «An iterative algorithm for decomposition of speech signals into periodic and aperiodic components,» Speech and Audio Processing, IEEE Transactions on, vol. 6, n° %11, pp. 1-11, 1998

[21] P. J. B. Jackson et C. Shadle, «Pitch-scaled estimation of simultaneous voiced and turbulence noise components in speech,» IEEE Trans. Speech Audio Process., vol. 9(7), pp. 713-726, 2001

[22] P. Lieberman, «Some acoustic measures of the fundamental periodicity of normal and pathologic larynges,» The Journal of the Acoustical Society of America, vol. 35, n° %13, pp. 344-353, 1963

[23] A. de Cheveigné et H. Kawahara, «YIN, a fundamental frequency estimator for speech and music,» The Journal of the Acoustical Society of America, vol. 11(4)1, pp. 1917-1930, 2002

[24] T. Drugman et A. Alwan, «Joint robust voicing detection and pitch estimation based on residual harmonics,» chez Twelfth Annual Conference of the International Speech Communication Association, 2011

[25] B. Elie et G. Chardon, «Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives,» chez Proceedings of the 22th International Congress on Acoustics, 2016

[26] M. Abe et J. O. S. III, «AM/FM rate estimation for time-varying sinusoidal modeling,» chez Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Philadelphia, PA, 2005

[27] J. Sundberg et J. Gauffin, «Waveform and spectrum of the glottal voice source,» Frontiers of speech communication research, pp. 301-320, 1979

[28] C. H. Shadle, Articulatory-Acoustic relationships in fricative consonants, Kluwer academic publishers, Dordrecht, pp. 187-209, 1990

[29] P. Stoica, H. Li et J. Li, «Amplitude estimation of sinusoidal signals: survey, new results, and an application,» IEEE Transactions on Signal Processing, vol. 48, n° %12, pp. 338-352, 2000

[30] B. Elie et Y. Laprie, «Glottal opening and strategies of production of fricatives,» chez Interspeech 2017, 2017

[31] D. Fohr, O. Mella et D. Jouvet, «De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée,» chez 8es Journées Internationales de Linguistique de Corpus (JLC2015), 2015