Séparation de signaux de parole en enregistrement monaural

François Signol, Jean-Sylvain Liénard
LIMSI* CNRS
Université Paris 11, UMR 3251, France
@ : {francois.signol, jean-sylvain.lienard} @limsi.fr
☎ : +33 1 69 85 {81 25, 81 13}

JJCAAS

Résumé

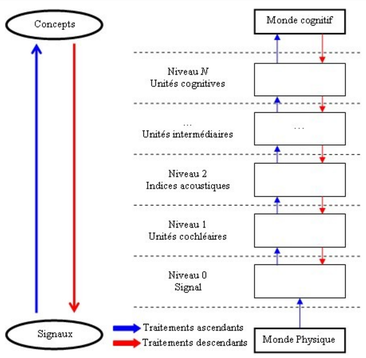
Le sujet qui nous occupe a pour objet la séparation de signaux de parole et relève de l'analyse computationnelle des scènes auditives. Il s'agit de mettre en place un système capable de séparer un mélange de deux signaux de parole naturelle à partir d'un enregistrement monocanal du mélange.
On envisage plusieurs niveaux de séparation :

1. Connaitre le nombre de locuteurs (0,1 ou 2) à chaque instant.
2. Suivre l'un ou l'autre des locuteurs dans le temps.
3. Fournir un signal estimé de chacun des locuteurs.

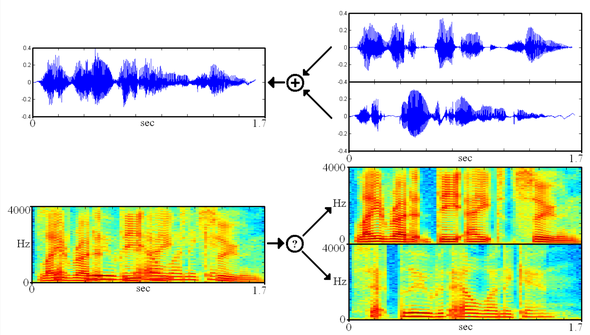
La séparation de parole n'est pas envisagée comme un processus relevant uniquement du traitement du signal. Elle est envisagée avec une dimension cognitive qui fait intervenir un modèle perceptif dans le traitement. La séparation de parole est un processus qui implique des traitements bas-niveau impliquant des techniques de traitement du signal complexes ainsi que des traitements haut-niveau amenant à l'utilisation de modèles de Markov afin de modéliser une certaine connaissance apprise des signaux.
La séparation bas-niveau s'appuie sur différents indices acoustiques du signal. Dans notre cas, il s'agit de la fréquence fondamentale, de l'énergie et de la forme globale du spectre. A chaque extrait du signal sont calculés et associés ces indices. Ces indices permettent de mettre ensemble les différents signaux constituants du mélange d'un extrait à l'autre en s'appuyant sur des propriétés de continuité spectrale ou continuité temporelle. Les connaissances haut-niveau du signal traduisent des connaissances du signal acquises par apprentissage. Le but recherché ou l'attention porté par d'un système cognitif se traduisent aussi par une action haut-niveau qui déterminera la séparation.

Perception auditive [1]

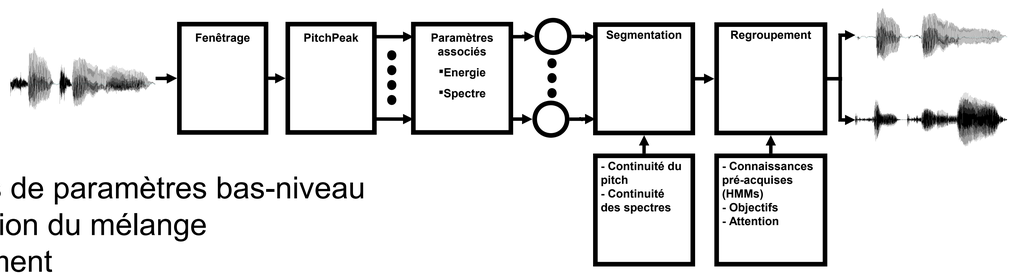
- Bottom-up
- Top-down



Mélange audio



Système de séparation [2][3][4][5]



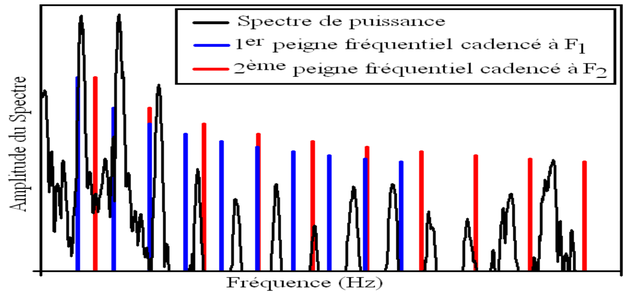
3 étapes :

1. Extractions de paramètres bas-niveau
2. Segmentation du mélange
3. Regroupement

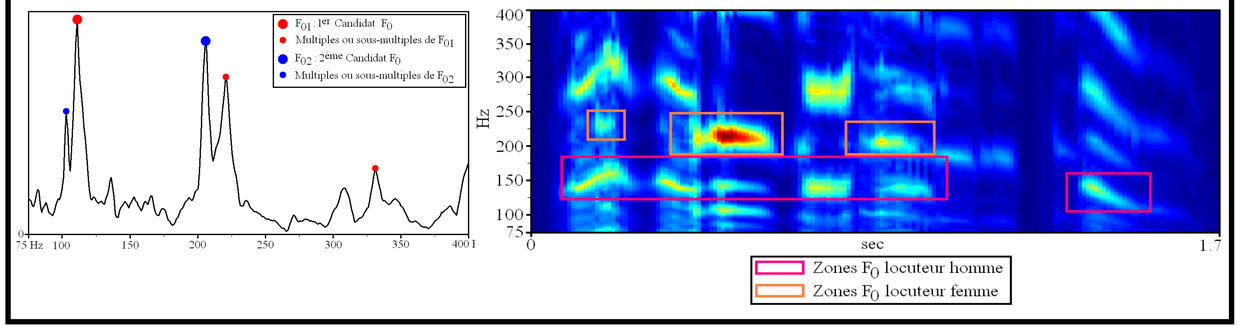
Détection multi-F0 : PitchPeak [6]

Principe de PitchPeak

- Méthode spectrale
- Peigne de Martin [7]
- Corrélation Spectre de puissance/Peigne
- Donne un nombre ***N*** de candidats F0

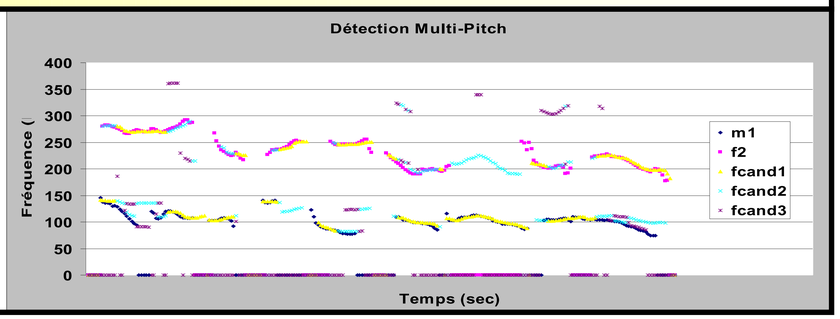


Exemple de PitchPeak



Résultats

- Trajectoires des F_0
- 3 premiers candidats
- Comparaison PRAAT/PitchPeak



Perspectives

- Suivi de Pitch multiple par Programmation Dynamique.
- Intégrer Onset/Offset pour les segments non voisés.
- Comptage du nombre de locuteur => Améliorations ASR
- Modélisation de connaissances haut-niveau par HMMs.

Références

[1] A. S. Bregman, *Auditory scene analysis.* Cambridge, MA : The MIT Press, 1990.
[2] J.P. Barker, M.P. Cooke, D.P.W. Ellis, *Decoding speech in the presence of other sources*. Speech Comm., **45,** 5-25, 2005.
[3] M. P. Cooke, *Modelling auditory processing and Organisation*. Cambridge, UK : Cambridge University Press, 1993.
[4] G. Hu, D. Wang, *An auditory scene analysis approach.* In Hansler E. and Schmidt G. (ed.), Topics in Acoustic Echo and Noise Control, Springer, Heidelberg, pp. 485-515.
[5] T.W. Parsons, *Separation of speech from interfering speech by means of harmonics selection*, JASA, vol. 60, No. 4, 911-918, 1976.
[6] A. de Cheveigné, *Multiple F0 estimation*, in Computational Auditory Scene Analysis: Principles and algorithms. D.L. Wang and G.J. Brown, IEEE Press / Wiley, 2006.
[7] P. Martin, *Comparison of pitch detection by cepstrum and cepstral comb analysis*, IEEE ICASSP, 180-183, 1982.

* ***L****aboratoire d'****I****nformatique pour la* ***M****écanique et les* ***S****ciences de l'****I****ngénieur*