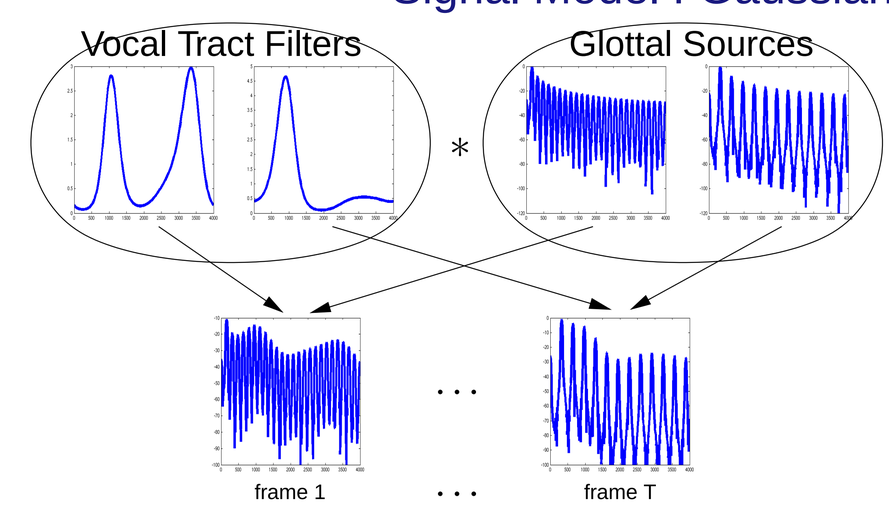
Automatic Extraction of the Main Melody in Music Signals

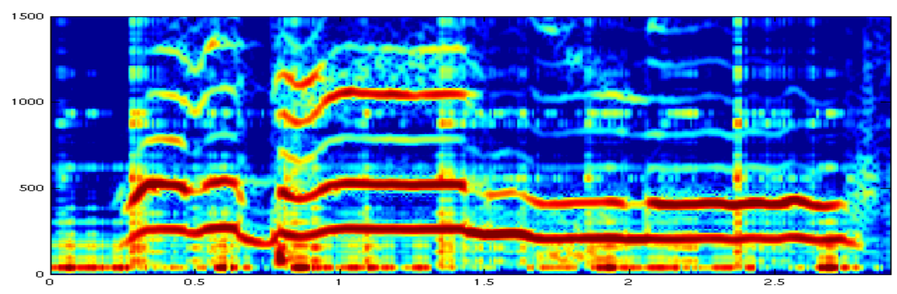
Jean-Louis Durrieu Gaël Richard Bertrand David
GET-ENST (Télécom Paris)
contact : durrieu@enst.fr http ://www.enst.fr/~durrieu

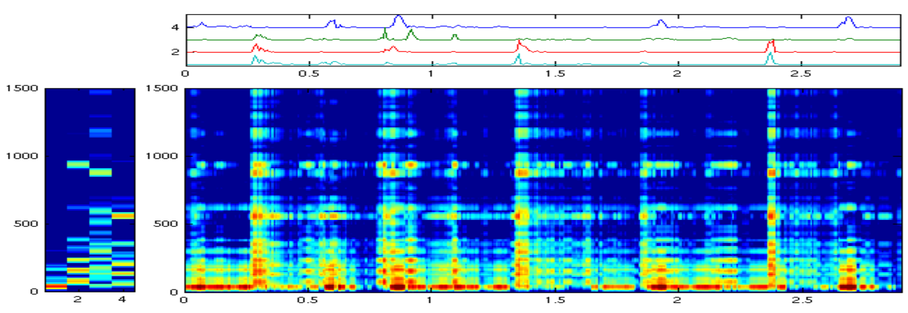
Introduction

– Extraction of the **melody sung by a human performer** and its transcription into a **musical score**
– Signal spectrum model which handles the voice spectrum as a **GMM** and decomposes the background music spectra on **elements of a dictionnary**
– The voice model also allows to directly extract the **melody line**

Signal Model : Gaussian Mixture and Non-Negative Matrix Factorization







$\{V(f,t)\}_{f,t}$: Short Time Fourier Transform (STFT) of the **Vocal signal**, **Gaussian Mixture Model (GMM)** (as in [1]) combined with a **source-filter model**.

For a frame t :

$$p(\{V(f,t)\}_f) = \sum_{k,f_0} \omega_k \nu_{f_0} p(\{V(f,t)\}_f | k, f_0)$$

where $p(V(f,t)|k,f_0) = \frac{1}{\pi\sigma_k^2(f)\sigma_{f_0}^2(f)} \exp\left(-\frac{|V(f,t)|^2}{\sigma_k^2(f)\sigma_{f_0}^2(f)}\right)$,

k : index for the vocal tract filter
f_0 : index for the glottal source $\Leftrightarrow$ musical note

$\downarrow$

The signal is considered as **the sum of the vocal and music signals** :

$$X(f,t) = V(f,t) + M(f,t)$$

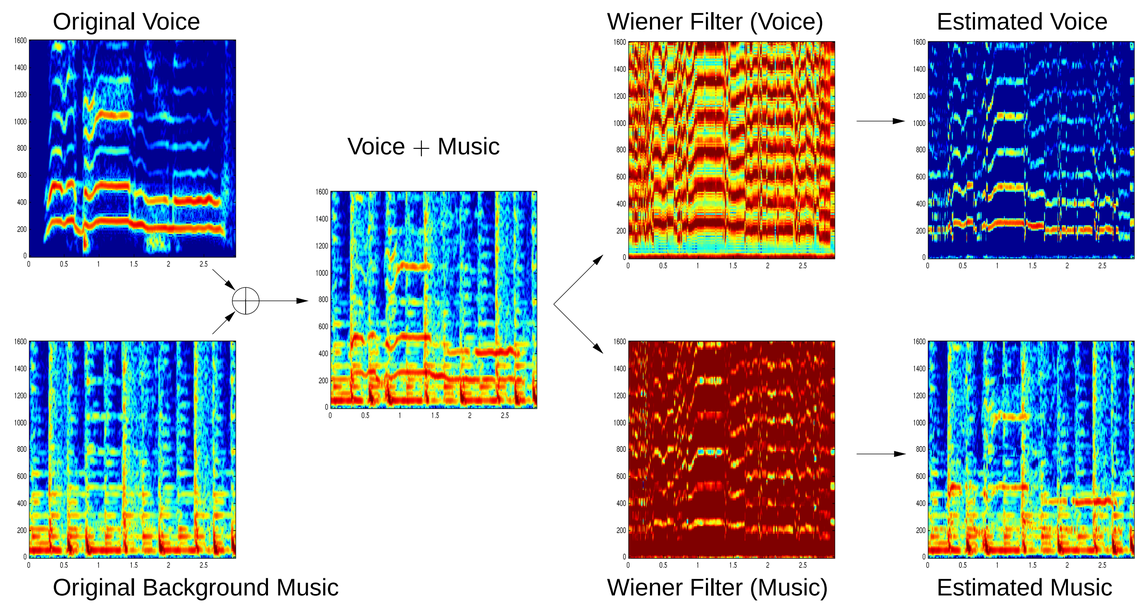
$\uparrow$

$\{M(f,t)\}_{f,t}$: STFT of the **Music signal**, sum of Gaussian spectra. For each frame t and frequency f :

$$M(f,t) = \sum_r a(r,t) b_r(f,t)$$

with $b_r(f,t) \sim \mathcal{N}_c(0, \sigma_r^2(f))$, such that $M(f,t) \sim \mathcal{N}_c(0, \sum_r a^2(r,t)\sigma_r^2(f))$

Estimating $a^2(r,t)$ and $\sigma_r^2(f)$ by Maximum Likelihood $\Longleftrightarrow$ **Non-negative Matrix Factorization (NMF)** of the spectrogram $|X(f,t)|^2$ for the Itakura-Saito divergence.



Preliminary Results and Experiments

EM-like algorithm : Estimate the parameters. **Voice/music separation** system as follows :
– Elements of the “music” dictionnary learned on non-vocal parts of the song (NMF with IS divergence)
– Joint estimation of the voice and music from the mix : fixed $\sigma_{f_0}^2(f)$, activation factors $a^2(r,t)$ and the filters $\sigma_k^2(f)$ (cf. [2]) learned on the vocal parts
– **Wiener filters** to extract the estimated parts (cf. [1])
– Estimation of the sung melody thanks to the parameters (a posteriori probabilities), leading to the desired **music score**

Results :
– Estimation of the voice rather satisfying, some artifacts
– Background music also well estimated, with a voice part attenuated but still present

Discussions and Perspectives

– Less complex music part than in [2], however able to better explain the musical phenomena. More complex vocal part, with an explicit model of notes (glottal sources)
– Promising preliminary tests (SIR $\simeq 17$ and SAR $\simeq 1$, cf. [3]), building databases for further evaluations of the source separation and the melody recognition performances
– Need for a vocal/non-vocal segmentation pre-processing

Références

[1] Laurent Benaroya and Frédéric Bimbot. Wiener based source separation with hmm/gmm using a single sensor. *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 957–961, April 2003.

[2] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 90–93, 16-19 Oct. 2005.

[3] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech and Audio Proc*, 2005.