SINGLE SENSOR SINGER/MUSIC SEPARATION USING A SOURCE/FILTER MODEL OF THE SINGER VOICE

Jean-Louis Durrieu, Gaël Richard, Bertrand David
TELECOM ParisTech / CNRS LTCI - 46, rue Barrault - 75634 Paris Cedex 13 - France - e-mail: durrieu@enst.fr

Introduction

- **Single-sensor singer/music separation**: separating the singer voice from the background polyphonic music on audio signals;
- **Proposed method**: applying a source/filter model to the vocal part and estimating its sequence of fundamental frequencies.

SIGNAL MODEL

Assumptions on the signal:

- **2 sources**: singer voice v and background music m, observed signal x such that: $x = v + m$,
- Wide sense (local) **stationarity**: analysis based on the short time Fourier transform (STFT) X,
- **Proper Gaussian** centered random variables: $Y \sim \mathcal{N}_c(0, \sigma_Y)$

Source/filter singer voice model

Source/Filter model for the voice:

- Dictionary of **fixed glottal source PSDs** σ_{f_0} (fig. 1),
  - KLGLOTT model: spectral "combs"
  - Fundamental frequencies between 100 and 800 Hz, 48 notes per octaves, $N_{\text{notes}} = 145$ combs,
  - No model for unvoiced part of singer signal,
  - $f_0 \in [1, N_{\text{notes}}]$.

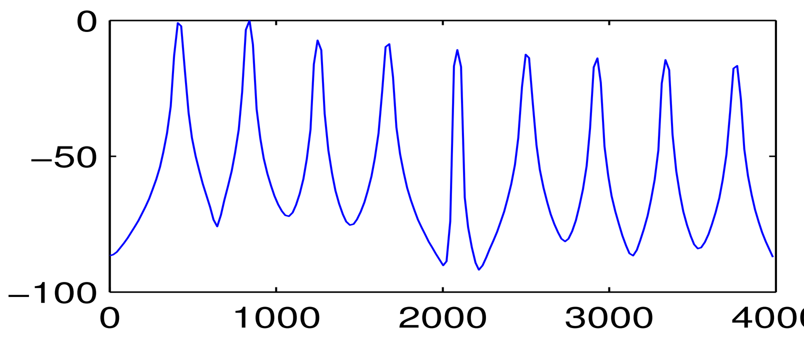

fig. 1 - σ_{f_0} (dB)

- Dictionary of **vocal tract filters** σ_k (fig. 2),
  - Each σ_k characteristic of 1 vowel (in theory),
  - $K = 9$ filters to be estimated, $k \in [1, K]$,
  - No constraints on estimation of $\sigma_k \to$ not accurate.

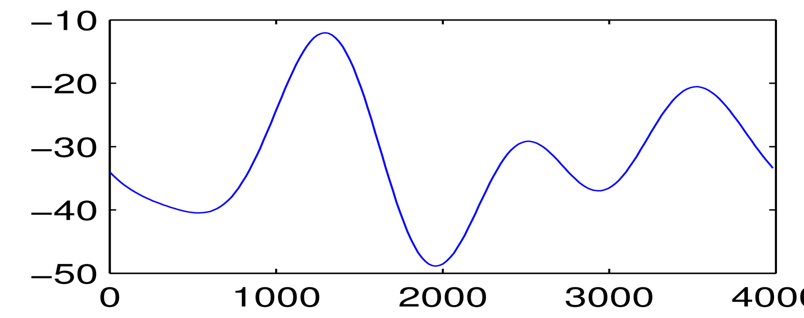

fig. 2 - σ_k (dB)

- Resulting prototype **PSD of the voice** at frequency bin f, for a given source/filter couple (k, f_0) (fig. 3): $\sigma_k(f) \times \sigma_{f_0}(f)$

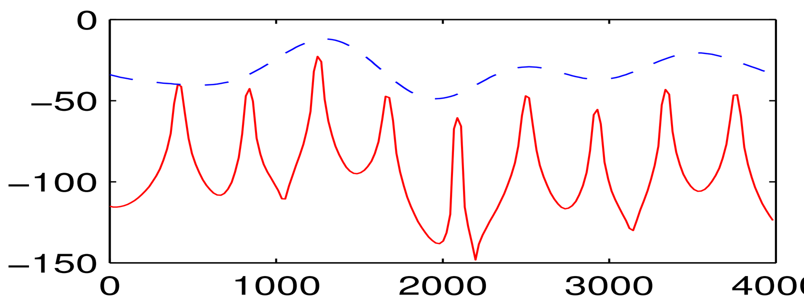

fig. 3 - (red) $\sigma_k \times \sigma_{f_0}$, (dotted line) σ_k (dB)

Instantaneous Mixture Model (IMM):

- $a_k(t)$ and $a_{f_0}(t)$ amplitude coefficients for filter k and source f_0,
- Each couple (k, f_0) always "active".

$$V(f,t) \sim \mathcal{N}_c(0, \underbrace{\sum_k a_k(t)\sigma_k(f)}_{V_K(f,t)} \times \underbrace{\sum_{f_0} a_{f_0}(t)\sigma_{f_0}(f)}_{V_{F_0}(f,t)})$$

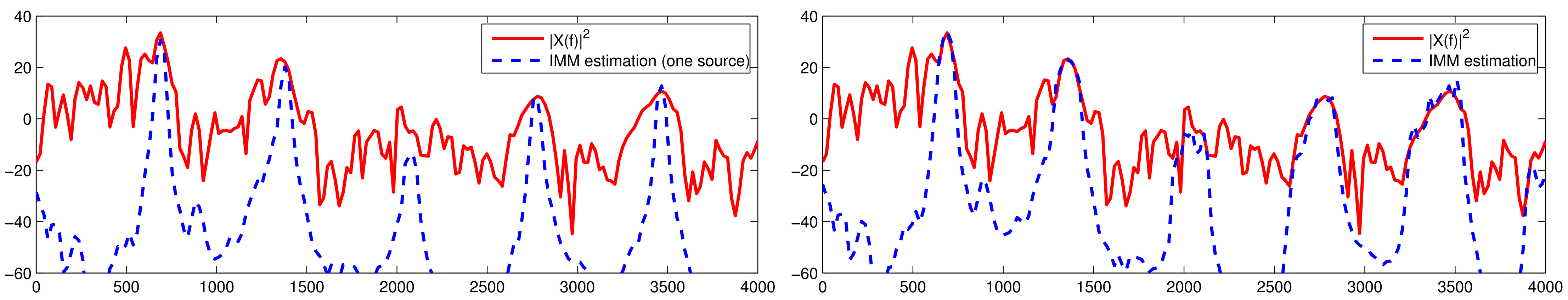

fig. 4 Frame of a singer "chirp" on polyphonic music: advantage of multiple-source model

Background music model

Instantaneous mixture of R Gaussian independent sources, with variances σ_r:

$$M(f,t) \sim \mathcal{N}_c(0, \underbrace{\sum_{r=1}^{R} a_r(t)\sigma_r(f)}_{D_R(f,t)})$$

Mixture signal

Instantaneous mixture of the two original sources: $X = V + M \Longrightarrow$

$X(f,t) \sim \mathcal{N}_c(0, D(f,t))$ with:
$D(f,t) = V_K(f,t) \times V_{F_0}(f,t) + D_R(f,t)$

SYSTEM OUTLINE

1. **ML estimation of** a_r, a_{f_0}, σ_k, a_r, σ_r: multiplicative gradient approach,
2. **Melody line** $F_0(t)$ inference: Viterbi smoothing on $a_{f_0}(t)$ [1]
3. **Re-estimation of the parameters**: ML initialized with modified amplitude glottal source coefficients $\tilde{a}_{f_0}(t)$ such that $\forall t$, $\tilde{a}_{f_0}(t) = a_{f_0}(t)$, if $f_0 = F_0(t)$ and 0 otherwise.
4. **Computation of the separated signals** $\hat{v}$ and $\hat{m}$: Wiener filters and Overlap-Add.

RESULTS

BSS EVAL criteria

- **Different contributions** in separated signals:

$$\hat{v} = \underbrace{\alpha_v v}_{s_{target}} + \underbrace{\beta_m m}_{e_{interference}} + e_{artefact}$$

- Normalized criteria computed from **Source-to-Distortion/Interference/Artifact-Ratio (SDR/SIR/SAR)**:

$$\text{SDR} = 20\log_{10}\left(\frac{\|s_{target}\|}{\|e_{interference} + e_{artefact}\|}\right)$$

$$\text{SIR} = 20\log_{10}\left(\frac{\|s_{target}\|}{\|e_{interference}\|}\right)$$

$$\text{SAR} = 20\log_{10}\left(\frac{\|s_{target} + e_{interference}\|}{\|e_{artefact}\|}\right)$$

Synthetic data

Synthetized audio from 200 MIDI files, melody played by an oboe:

	$\hat{v}$			$\hat{m}$		
	SDR	SIR	SAR	SDR	SIR	SAR
1st est.	10.04	24.34	8.76	7.51	15.48	12.45
2d est.	12.92	25.91	11.56	10.38	25.82	14.06

Real data

10 "pop" songs, with/without vocal/non-vocal segmentation [2]

	$\hat{v}$			$\hat{m}$		
	SDR	SIR	SAR	SDR	SIR	SAR
no vocal/non-vocal segmentation:						
1st est.	3.73	12.08	0.39	0.7	5.9	9.87
2d est.	6.42	14.82	2.37	1.58	12.78	8.44
manual v/n-v segmentation:						
1st est.	6.98	22.03	1.34	3.13	6.08	13.92
2d est.	10.71	25.01	4.93	5.66	13.96	12.81

Conclusions and Perspectives

- **Results at the state of the art**, with good perceptual results,
- IMM drawbacks balanced by **re-estimation of parameters**,
- **Bayesian framework** allowing model refinements: temporal and spectral regularization of the parameters, e.g. ARMA models on σ_k, HMM on $a_{f_0}(t)$ etc.

[1] J.-L. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. *ICASSP*, 2008.
[2] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs. *IEEE Trans. on ASLP*, 2007.

Acoustics'08, Paris, June 29 - July 4, 2008