"Robust Automatic Speech Recognition through on-line Semi Blind Source Extraction"

Francesco Nesta, Marco Matassoni

{nesta, matassoni}@fbk.eu

Fondazione Bruno Kessler-Irst, Trento (ITALY)

For contacts:
http://shine.fbk.eu/people/nesta
nesta@fbk.eu

Introduction

- Robust ASR has the goal to mimic the natural ability of humans to understand and recognize speech in adverse conditions, such as the case of speech contaminated by multiple competing interfering source signals.

In this work we approach robustness on the CHIME challenge data following two key directions

Source signal enhancement through statistical independence and multichannel data

Features better matching the human auditory perception

Semi Blind Source Extraction

Gammatone

BSS vs BSE

Blind Source Separation (BSS)

$\mathbf{s}(k,l)$ is a vector of N sources

$\mathbf{x}(k,l)$ is a vector of M mixtures (i.e. mic numbers)

$$\mathbf{x}(k,l) = \mathbf{H}(k)\mathbf{s}(k,l)$$

k = frequency bin index
l = frame index

- If N=M, the source signals are as:

$$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l) \approx \mathbf{s}(k,l)$$

$$\mathbf{W}(k)\mathbf{H}^{-1}(k) = \mathbf{I}$$ (up to order and scaling ambiguity)

In real-world N>M and may rapidly change over time!

Blind Source Extraction (BSE)

- The blind source extraction paradigm has been proposed to overcome those limitations[Takahashi, Saruwatari et all 2008].
- Mixtures are modeled as:

Image at microphones of the sum of the interfering sources

$$\mathbf{x}(k,l) = \mathbf{s}^t(k,l) + \mathbf{n}(k,l) = \mathbf{h}^t(k)s^t(k,l) + \mathbf{n}(k,l)$$

Image at microphones of the target source

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

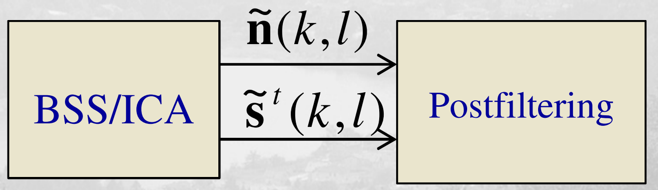
Limits of BSE

- We may estimate the noise in the mixture as:

$$\tilde{\mathbf{n}}(k,l) = \mathbf{w}(k)^T \mathbf{x}(k,l) = \mathbf{w}(k)^T [\mathbf{h}^t(k) s^t(k,l) + \mathbf{n}(k,l)]$$

where $\mathbf{w}(k)^T \mathbf{h}^t(k) = \mathbf{0}$

- If the target source is always active and dominant $\mathbf{w}(k)^T$ is one of the row of $\mathbf{W}(k)$, e.g. estimated through ICA.
- Once an estimation of $\mathbf{n}(k,l)$ and of the target source is obtained the signals are filtered through a non-linear time-varying filtering (e.g. Wiener filter, spectral subtraction,…), based on the estimation of the power spectral density of target source and noise signals.



Main issues:

1. Due to the scaling ambiguity, $\tilde{\mathbf{n}}(k,l)$ is a time-varying distorted approximation of $\mathbf{n}(k,l)$.
2. The target source cannot be estimated with a single linear demixing.

Incorrect estimation of the power spectral density of target and noise sources generates distortions in the recovered output signals!

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

On-line Semi-blind source extraction (SBSE)

The BSE is extended with a twofold modification:

Frequency mixtures are modeled as the sum of the signals of the target source and of the M-1 most dominant interfering sources. The intermittingly activity of the (unknown) interfering sources is modeled by a time-varying mixing matrix which leads to a better estimation of the noise components in each frequency and time frame.

$\mathbf{H}(k,l)$ is a M x N(k,l) mixing matrix $\Rightarrow$ $\mathbf{y}(k,l) = \mathbf{W}(k,l)\mathbf{x}(k,l)$

In order to better estimate the target source components a semi-blind source separation (SBSS) is realized. It nests a prior knowledge on **w**(k) directly the adaptation structure of ICA.

Assumption: the mixing matrix of the target is estimated beforehand in the signal chunks where it dominates the interfering sources.

Note: in a real-world application different strategies can be adopted to estimate the mixing matrix (e.g. as done in the demo presented at Interspeech 2011, a parallel batch off-line ICA can be applied on larger signals to supervise the on-line SBSS)

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

SBSS as a constrained ICA adaptation

- In order to guarantee that the first output is related to the target source, the ICA adaptation needs to be constrained, imposing

$$\mathbf{W}(k,l)^{-1} = [\mathbf{h}^t(k) \mid \ldots]$$

- It can be obtained as:

-If μ=0 an hard constraint is imposed
(e.g. equivalent to SBSS applied to MCAEC [Nesta et. all 2009/2011])

-If μ=1 no constraint is imposed

$$\begin{cases} \mathbf{W}_{prior}(k) = [\mathbf{h}^t(k) \mid \mathbf{I}_{2..M}]^{-1} \\ \tilde{\mathbf{x}}(k,l) = \mathbf{W}_{prior}(k)\mathbf{x}(k,l) \\ \mathbf{y}(k,l) = \mathbf{W}(k,l)\tilde{\mathbf{x}}(k,l) \\ \Delta\mathbf{W}(k,l) = \{\mathbf{I} - \phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\}\mathbf{W}(k,l) \\ \Delta\mathbf{W}_{constr}(k,l) = [\mu\Delta\mathbf{W}_1(k,l) \mid \Delta\mathbf{W}_{2..M}(k,l)] \\ \mathbf{W}(k,l+1) = \mathbf{W}(k,l) + \eta[\Delta\mathbf{W}_{constr}(k,l)] \end{cases}$$

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

Permutation and scaling ambiguity

Permutation

- If μ=0 the hard constraint avoids the permutation problem of frequency-domain BSS (on condition of an accurate mixing matrix prior).
- If the constraint is partially released permutation need to be fixed (e.g. through the GSCT)

Scaling

- Scaling ambiguity can be solved through the Minimal Distortion Principle (MDP) only if N(k,l)=M.
- If N(k,l)>M and **W**(k) approaches the singularity, the MDP may considerably overestimate the residual noise components not suppressed by the linear demixing.

A simple solution: non-linear clipping limiting the overall filtering by unit gain.

$$\bar{y}_{\tilde{m}}^{m}(k,l) = \min(|\,\bar{y}_{\tilde{m}}^{m}(k,l)\,|, |\,x_{\tilde{m}}(k,l)\,|)\frac{\bar{y}_{\tilde{m}}^{m}(k,l)}{|\,\bar{y}_{\tilde{m}}^{m}(k,l)\,|}$$

$x_{\tilde{m}}(k,l)$ Indicates the signal recorded at $\tilde{m}-th$ microphone

$\bar{y}_{\tilde{m}}^{m}(k,l)$ Indicates the projected back image of the $m-th$ source signal at the $\tilde{m}-th$ microphone

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

Channel-wise Wiener filter postfiltering

- Constrained SBSS can only enhance the target source signal by linear time-varying demixing:
- A post filtering is used to enhance the source of interest through a channel-wise adaptive Wiener filtering:

$$s^t_{\tilde{m}}(k,l) = \frac{P^t_{\tilde{m}}(k,l)}{P^t_{\tilde{m}}(k,l) + P^r_{\tilde{m}}(k,l)} x_{\tilde{m}}(k,l)$$

PSD of the target source $P^t_{\tilde{m}}(k,l) \approx E[|s^1_{\tilde{m}}(k,l)|^2]$

PSD of the noise $P^r_{\tilde{m}}(k,l) \approx E[|y^2_{\tilde{m}}(k,l)|^2]$

- For the 2-channel case:

$$|s^1_{\tilde{m}}(k,l)|^2 = \begin{cases} |\hat{s}^1_{\tilde{m}}(k,l)|^2 , & if\ \hat{s}^1_{\tilde{m}}(k,l) > 0 \\ 0 & otherwise \end{cases}$$

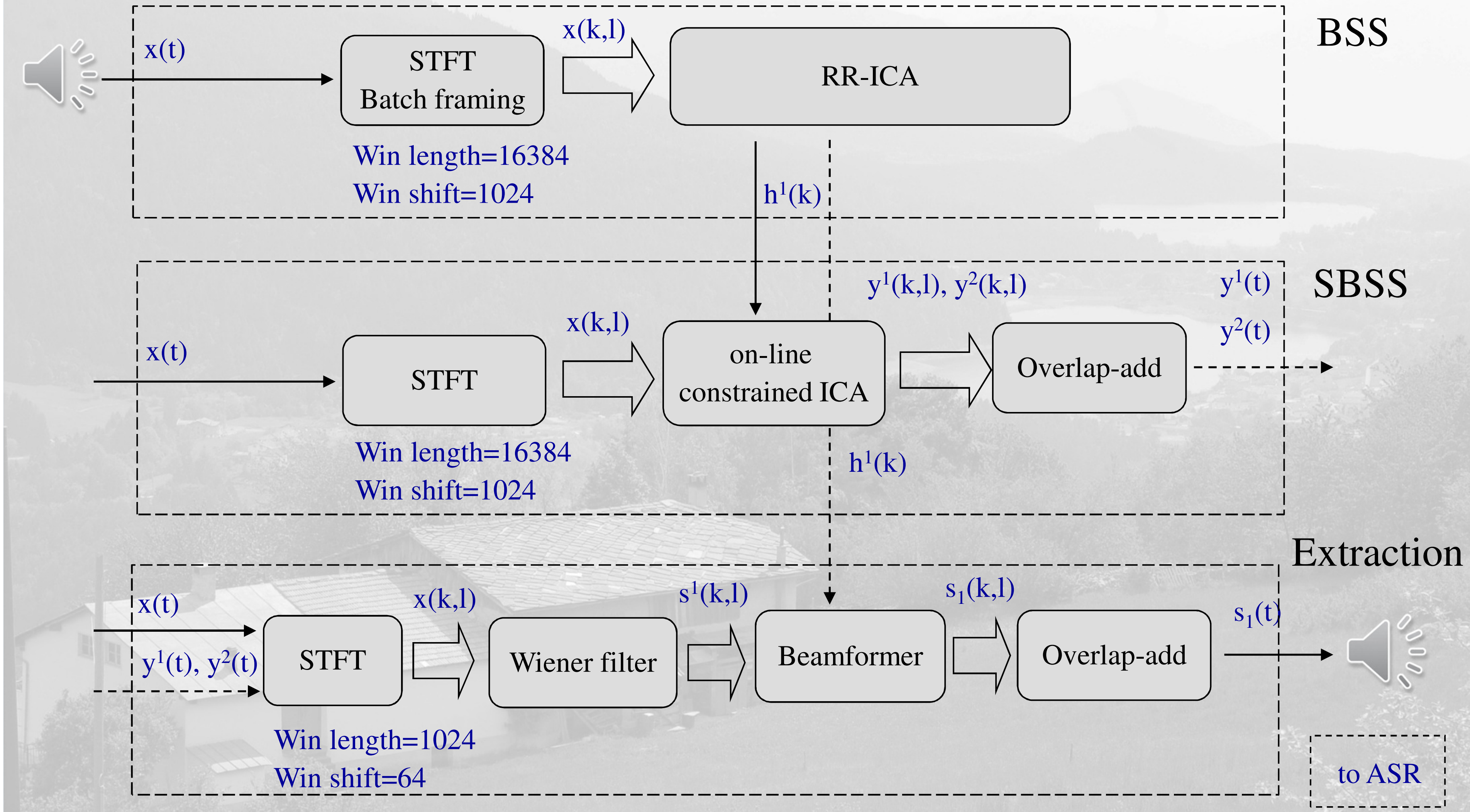
$$\hat{s}^1_{\tilde{m}}(k,l) = y^1_{\tilde{m}}(k,l) - C_{\tilde{m}}(k,l) y^2_{\tilde{m}}(k,l) + o_{\tilde{m}}(k,l)$$

Over-subtraction compensation

$$C_{\tilde{m}}(k,l) = \frac{E[|y^1_{\tilde{m}}(k,l)||y^2_{\tilde{m}}(k,l)|]}{E[|y^1_{\tilde{m}}(k,l)|^2]}$$

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

SBSE architecture



Robust Automatic Speech Recognition through on-line Semi Blind Source Extraction

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

Integration with Robust ASR (1/2)

Acoustic features based on Gammatone analysis:

- linear approximation of physiologically motivated processing performed by the cochlea
- bandpass filters, whose impulse response is defined by:

$$g_c(t) = at^{c-1}\cos(2\pi f_c t + \phi)e^{-2\pi b_c t}$$

- filter center frequencies and bandwidths are derived from the filter's Equivalent Rectangular Bandwidth
- output of the Gammatone filter:

$$x_c(t) = x(t) * g_c(t)$$

where $g_c(t)$ is the impulse response of the filter.

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

FONDAZIONE BRUNO KESSLER

Integration with Robust ASR (2/2)

Enlarged Training

- different versions of the utterance are considered:

Separate Right/Left channels, Right+Left, corresponding clean signals from Grid corpus

- Note: to guarantee the blindness with respect to the target signal contamination, the noisy signals are not used neither for the training nor for the adaptation.

Model Adaptation

- starting from the Speaker Independent models, model adaptation is applied,

based on a combination of MLLR and MAP:

1. MLLR is applied in two-stage fashion: global adaptation transform followed by specific transforms according to a 128 regression class tree
2. After the MLLR step, MAP adaption is performed.

- Two sets of SD models are derived using the development and test material

(i.e. all signals at different SNRs are pooled).

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

Experimental results (word accuracy %)

Development dataset

SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	AVG.
-	31.08	36.75	49.08	64.00	73.83	83.08	56.30
SBSE	61.08	68.67	76.00	80.67	85.83	88.83	76.84
SBSE+ET	66.33	73.50	79.17	83.83	86.50	90.83	80.02
SBSE+GF+ET	76.08	81.67	87.33	89.92	92.17	93.67	86.80
SBSE+GF+ET+MA	**80.17**	**83.92**	**89.50**	**90.83**	**93.33**	**94.42**	89.65

Test dataset

SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	AVG.
-	30.33	35.42	49.50	62.92	75.00	82.42	55.93
SBSE	54.75	63.08	72.67	78.17	83.42	87.08	73.19
SBSE+ET	60.75	67.33	76.83	80.75	85.67	89.42	76.79
SBSE+GF+ET	72.00	78.33	85.17	90.08	92.00	93.50	85.18
SBSE+GF+ET+MA	**77.08**	**81.42**	**87.25**	**91.17**	**93.00**	**94.58**	87.41

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu

Conclusions

Where we are…

- We proposed an advanced speech enhancement algorithm based on a Semi-blind source extraction.
- The enhancement chain introduces very low distortions in the recovered target signal even in presence of multiple real-world highly non-stationary noise sources.
- Promising results have been obtained in the CHIME challenge tasks, when combined with robust features derived by Gammatone analysis.

… and where we are going

- The target mixing parameters estimation is crucial: the more accurate it is, the more SNR improvement and the less distortions in the target signal.
- Spatial information (e.g. multiple TDOAs) can be used as a rough estimation for the mixing parameters→ source tracking is another key direction
- On going research activities concerns a better refinement of the estimated mixing parameters in a full blind fashion (e.g. exploiting other spatial cues, environmental awareness, …)
- Better combination of SBSE with Gammatone based features analysis

Francesco Nesta, Marco Matassoni {nesta,matassoni}@fbk.eu


Any questions? (not too many please!)