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

Francesco Nesta, Marco Matassoni

Fondazione Bruno Kessler-Irst
via Sommarive 18, 38123 Trento, Italy
{nesta,matassoni}@fbk.eu

Abstract

This paper describes the system used to process the data of the CHiME Pascal 2011 competition, whose goal is to separate the desired speech and recognize the commands being spoken. The binaural recorded mixtures are processed by an on-line Semi-Blind Source Extraction algorithm. The algorithm is based on a multi-stage architecture combining the advantages of constrained Independent Component Analysis and Wiener-based processing, allowing the estimation of the target signal with limited distortion. The recovered target signal is then fed to the recognizer which uses noise robust features based on Gammatone Frequency Cepstral Coefficients. Moreover, model adaptation to actual processing is applied as a further stage to reduce the acoustic mismatch. Performance comparison between different model/algorithmic settings is reported for both development and test data sets.

Index Terms: blind source separation, speech enhancement, robust speech recognition

1. Introduction

Although speech processing technologies have being investigated actively since natural interaction is an appealing communication modality, speech acquisition, processing and recognition in a non-ideal acoustic environments are complex tasks due to presence of noise, reverberation and interfering speakers [1, 2] . CHiME is a speech corpus designed for investigating robust speech processing and recognition and for comparing achievements obtained in both speech enhancement and recognition communities [3]. The recorded data includes background recordings from a head simulator positioned in a domestic setting as well as binaural impulse responses collected in the same environment. By means of these genuine responses, utterances from the Grid corpus [4] have been added to this setting and mixed with the background noise to produce controlled and natural audio data. The task is to separate the speech and recognize the commands being spoken using systems that have been trained on noise-free commands and room noise recordings.

In general to improve automatic speech recognition (ASR) robustness, processing can be performed at signal, feature or model level [5, 2]. Speech enhancement techniques aim at improving the quality of speech signal captured through single microphone or microphone array. Robust acoustic features attempt to represent parameters less sensitive to noise or by enhancing the extracted features. Model adaptation approaches modify the acoustic model parameters to fit better the actual speech features. Another direction to tackle the problem of ASR in noisy environment is using techniques of speech enhancement to preprocess the noisy mixtures. Recent achievements in the field of the Blind Source Separation (BSS) context [6] have shown that binaural mixtures can be successfully processed by BSS methods in order to remove diffuse background noise from a given source of interest [7]. The array processing can then enhance the target speech in order to improve the performance of an ASR system.

In this work, a complete system designed to process the CHiME data is presented. A Semi-Blind Source Extraction (SBSE) system, based on the combination of Blind and Semi-Blind source separation [8][9] estimates the noise and target spectra which is later used to control the coefficients of a Wiener filter. This SBSE architecture allows a high suppression of the noise while maintaining the quality of the target source signal at an acceptable level. Results are provided in term of Keyword recognition Accuracy and various approaches are presented and discussed.

The paper is organized as follows: Section 2 presents the binaural signal processing module, detailing on the theoretical background and the adopted processing architecture. Section 3 introduces the recognition system, including the acoustic front end and training/adaptation procedures and reporting the partial recognition results. In Section 4 we summarize the experimental results and discuss the investigated techniques and, finally, Section 5 concludes the paper.

2. Semi-Blind Source Extraction

In this section a general formulation of the multichannel Blind Source Extraction problem is discussed. A specialization for the case of binaural recordings and the related system architecture is later described.

The problem is formulated considering the signals in the discrete time-frequency STFT domain. Indicate with k and l the frequency bin and STFT frame indices, respectively. Consider the case where M microphones observe at each frequency bin and frame, the image of the signals related to $N(k,l)$ sources. Note that the dependence of N with k and l indicates that the number of sources having non negligible energy may vary over time and frequency. In matrix notation one can write

$$\mathbf{x}(k,l) = \mathbf{H}(k,l)\mathbf{s}(k,l) \tag{1}$$

where $\mathbf{x}(k,l)$ is a column vector of the observed mixtures ($\mathbf{x}(k,l) = [x_1(k,l), \cdots, x_M(k,l)]^T$), $\mathbf{s}(k,l)$ is the column vector of the source signals ($\mathbf{s}(k,l) = [s_1(k,l), \cdots, s_{N(k,l)}(k,l)]^T$) and $\mathbf{H}(k,l)$ is a $M \times N(k,l)$ mixing matrix modeling the frequency responses between the microphones and the $N(k,l)$ sources.

In the case $N(k,l) = M$, the problem is determined and one can retrieve separated signal components $\mathbf{y}(k,l)$ by means of a set of demixing matrices $\mathbf{W}(k,l)$ as

$$\mathbf{y}(k,l) = \mathbf{W}(k,l)\mathbf{x}(k,l). \tag{2}$$

The time-varying demixing matrix $\mathbf{W}(k,l)$ can be determined, up to scaling and permutation ambiguities, by applying a complex-valued ICA. Note that, ICA requires a sufficient number of temporal observations for the mixtures $\mathbf{x}(k,l)$. However if the mixing conditions are quasi-stationary, i.e. do not change for a sufficient amount of time frames, an on-line ICA algorithm is sufficient to estimate the demixing matrices $\mathbf{W}(k,l)$, which can be modeled as

$$\mathbf{W}(k,l) = \mathbf{\Lambda}(k,l)\mathbf{\Pi}(k,l)\overline{\mathbf{H}}^{-1}(k,l), \tag{3}$$

where $\mathbf{\Lambda}(k,l)$ is an arbitrary diagonal scaling matrix, $\mathbf{\Pi}(k,l)$ is a permutation matrix and $\overline{\mathbf{H}}^{-1}(k,l)$ is an estimate of the inverse of the true time-varying mixing matrix $\mathbf{H}(k,l)$. For the case $N(k,l) > M$ the problem is underdetermined since there is no demixing matrix able to linearly separate the mixtures in the original signal components. When the number of the sources is known in advance, the mixing conditions are stationary and the reverberation time is not too high, clustering techniques based on the time-frequency disjointness have shown to be accurate enough for estimating STFT masks, which are used to isolate the spectral components of each source signal [10][11]. However, in realistic scenarios, the number of active sources is unpredictable and not constant over time. Moreover, clustering procedures are further complicated by the mixing conditions which are time-varying and can even change abruptly over time.

In general for the well-known determined case, i.e. $\mathbf{H}(k,l)$ is a square matrix, the solution to the separation problem is given by

$$\mathbf{y}(k,l) = \mathbf{W}(k,l)\mathbf{x}(k,l), \tag{4}$$
$$\mathbf{W}(k,l+1) = \mathbf{W}(k,l) + \eta[\Delta\mathbf{W}(k,l)] \tag{5}$$

where $\Delta\mathbf{W}(k,l)$ is the gradient which takes different form according to the cost function that is to be minimized. Considering the Natural Gradient adaptation based on the Kullback-Leibler divergence $\Delta\mathbf{W}(k,l)$ is determined as

$$\Delta\mathbf{W}(k,l) = \{\mathbf{I} - \Phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\}\mathbf{W}(k,l), \tag{6}$$

where $\Phi(\cdot)$ is a nonlinear function, $[\,\cdot\,]^H$ is the Hermitian (conjugate) transpose operator and $\mathbf{I}$ is the $M \times M$ identity matrix.

Although for the case $N(k,l) > M$ it is not possible to linearly separate all the signal sources, one may be interested in recovering only a single target source while filtering out the noise generated by the interfering sources. In fact it is still possible to estimate a demixing filter which would reduce the mutual dependence of the output components, and then potentially reducing the noise in the target signal. In general, since there are multiple sources the adaptation might converge to many equivalent solutions, i.e., the cost function that is to be minimized might have different equivalent minima. In other terms, without any constrain it is not guaranteed that the output signals would represent always the same source. Geometrically constrained adaptation were adopted for BSS, in order to prevent the permutation problem of frequency-domain implementation and improve the convergence stability of the overall adaptation [6]. Analogously, we define here a constrained ICA which is on the bases of the Semi-Blind Source Separation framework (SBSS), successfully applied to the Multichannel Acoustic Echo Cancellation (MCAEC) problem [9]. A demixing matrix constrain is imposed in order to force one of the outputs to give an estimation of a given target source, while the others give an estimation of the remaining interfering sources.

2.1. Constrained ICA adaptation

For the sake of simplicity we assume that the target source is always active and has stationary mixing parameters, i.e., it does not change location and modifications of the impulse responses (between the target source and the microphones) due to movements of other sources, are neglected. In order to guarantee that the first system output always corresponds to an estimation of the target source, the adaptation in (5) should be modified imposing the constraint

$$\mathbf{W}^{-1}(k,l) = [\mathbf{h}^1(k)|\cdots] \tag{7}$$

where $\mathbf{h}^1(k)$ indicates the column vector of the mixing parameters related to the target source, which is assumed to be known or approximatively estimated in advance. The constrain in (7) can be imposed in (4-6) as follows. First of all equation in (1) can be approximated as

$$\mathbf{x}(k,l) = \tilde{\mathbf{H}}(k,l)\mathbf{s}(k,l) \tag{8}$$

where $\tilde{\mathbf{H}}(k,l) = [\mathbf{h}^1(k)|\mathbf{h}^2(k)\cdots\mathbf{h}^M(k)]$ indicates the reduced mixing system of the target source and of the $M-1$ remaining most dominant sources. The matrix $\tilde{\mathbf{H}}(k,l)$ can be factorized as

$$\tilde{\mathbf{H}}(k,l) = [\mathbf{h}^1(k)|\mathbf{I}_{2\cdots M}] \times [\mathbf{c}|\mathbf{F}(k,l)] \tag{9}$$

where $\mathbf{I}_{2\cdots M}$ indicates the last $M-1$ columns of the $M \times M$ identity matrix, $\mathbf{c}$ is the $M \times 1$ column vector $[c,0,\cdots,0]^T$ with c an arbitrary constant and $\mathbf{F}(k)$ is an arbitrary $M \times M-1$ matrix resulting from the factorization. By inversion of (9) we obtain

$$\tilde{\mathbf{W}}(k,l) = \mathbf{W}_{constr}(k,l)\mathbf{W}_{prior}(k) \tag{10}$$

where $\mathbf{W}_{constr} = [\tilde{\mathbf{c}}|\tilde{\mathbf{F}}(k,l)]$ and $\mathbf{W}_{prior} = [\mathbf{h}^1(k)|\mathbf{I}_{2\cdots M}]^{-1}$. Here $\tilde{\mathbf{F}}(k,l)$ is an arbitrary $M \times M-1$ matrix resulting from the inversion and $\tilde{\mathbf{c}} = [1/c,0,\cdots,0]^T$. It follows that the first system output gives an estimation of the target source if the demixing matrix has the structure in (10). Substituting (10) in (4) we obtain

$$\mathbf{y}(k,l) = \mathbf{W}_{constr}(k,l)\mathbf{W}_{prior}(k)\mathbf{x}(k,l) \tag{11}$$
$$= \mathbf{W}_{constr}(k,l)\tilde{\mathbf{x}}(k,l) \tag{12}$$

where $\tilde{\mathbf{x}}(k,l)$ indicates the pre-processed mixtures according to the prior knowledge on the target mixing parameters. Therefore, a constrained adaptation is obtained by modifying (4-5) as

$$\tilde{\mathbf{x}}(k,l) = \mathbf{W}_{prior}(k)\mathbf{x}(k,l), \tag{13}$$
$$\mathbf{y}(k,l) = \mathbf{W}(k,l)\tilde{\mathbf{x}}(k,l), \tag{14}$$
$$\Delta\mathbf{W}_{constr}(k,l) = [\mu\Delta\mathbf{W}_1(k,l)|\Delta\mathbf{W}_{2\cdots M}(k,l)] \tag{15}$$
$$\mathbf{W}(k,l+1) = \mathbf{W}(k,l) + \eta[\Delta\mathbf{W}_{constr}(k,l)] \tag{16}$$

where μ is a scalar with values in the range between 0 and 1, $\Delta\mathbf{W}_1(k,l)$ is the $M \times 1$ matrix consisting of the first column of $\Delta\mathbf{W}(k,l)$ and $\Delta\mathbf{W}_{2\cdots M}(k,l)$ is the sub-matrix consisting in the last $M-1$ columns of $\Delta\mathbf{W}(k,l)$. The scalar μ defines the importance of the constraint imposed by $\mathbf{W}_{prior}(k)$. If $\mu = 1$ no constraint is imposed and the adaptation is equivalent to (4-6). On the other hand, if $\mu = 0$ the adaptation constrains the mixing parameters of the target to $\mathbf{h}^1(k)$, while it continuously adapts the parameters related to the interfering sources. Note that when $\mu = 0$ the constrained adaptation in (13-16) has the

same structure of the SBSS applied to the MCAEC problem [9]. In [9] a constrain in the matrix $\mathbf{W}$ was motivated by the fact that target signals are not present in the reference signals (used to estimate the echos) and consequently the full demixing system is to be constrained in order to have opportune zero entries in the corresponding mixing matrix. Similarly, in our case if the target mixing parameters are ideally known, the pre-processing is equivalent to null-steering in the direction of the target source, which means that the mixtures corresponding to the last $M-1$ elements of $\tilde{\mathbf{x}}(k,l)$ do not contain any contribute of the target source and can be considered as references of the remaining noise signals.

If $0 < \mu < 1$ the optimization is less constrained by the prior imposed by $\mathbf{W}_{prior}(k)$, that is, it is implicitly assumed a certain degree of uncertainty in the initial guess for $\mathbf{h}^1(k)$. For the sake of simplicity, in this work it is assumed that an "exact" knowledge of $\mathbf{h}^1(k)$ is available beforehand and μ is imposed to 0. In this case the signal of the target source is not affected by the permutation problem of the frequency-domain BSS since the order of the output is intrinsically forced by the constrain imposed in the mixing parameters. On the other hand, as discussed in the next section, an exact solution for the scaling ambiguity does not exist even when the mixing parameters are ideally known.

2.2. Scaling ambiguity

As shown in equation (3), neglecting the permutation ambiguity (which is assumed to be solved in the semi-blind case), the estimated demixing matrix $\mathbf{W}(k)$ is an estimation of the inverse mixing matrix up to a scaling ambiguity. A popular method used to reduce this ambiguity is the Minimal Distortion Principle (MDP) [12]. According to MDP, for estimating the multichannel image of the *m-th* source, the demixing matrix is normalized as

$$\mathbf{W}^m(k,l) = [\mathbf{O}_{M\times m-1}|[\mathbf{w}^m]^{-1}(k,l)|\mathbf{O}_{M\times M-m}]\mathbf{W}(k,l) \tag{17}$$

where $[\mathbf{w}^m]^{-1}(k,l)$ indicates the *m-th* column of $\mathbf{W}^{-1}(k,l)$. It may be shown that, for the determined case, the MDP leads to the estimation of the exact image of the source signals at each microphone. Assuming the permutation to be solved (i.e., $\mathbf{\Pi}(k,l) = \mathbf{I}$), substituting (3) in (17) and (4) we obtain

$$\begin{aligned}\mathbf{y}^m(k,l) &= [\mathbf{O}_{M\times m-1}|\mathbf{h}^m(k,l)|\mathbf{O}_{M\times M-m}]\mathbf{\Lambda}(k,l)^{-1}\mathbf{\Lambda}(k,l)\mathbf{s}(k,l)\\ &= \mathbf{h}^m(k,l)s_m(k,l) = \mathbf{s}^m(k,l)\end{aligned} \tag{18}$$

where $s_m(k,l)$ and $\mathbf{s}^m(k,l)$ are the *m-th* source signal and its multichannel image at the microphones, respectively, and $\mathbf{y}^m(k,l)$ indicates the normalized output signals when the scaling normalization is referred to the mixing parameters of the *m-th* column of $\mathbf{W}^{-1}(k,l)$. If $N(k,l) > M$ the source signal images cannot be exactly recovered and $\mathbf{y}^m(k,l)$ is not equivalent to $\mathbf{s}^m(k,l)$. For this case, indicating with $\mathbf{s}_L(k,l)$ and $\mathbf{s}_R(k,l)$ the vectors of the first M and the last $N(k,l)-M$ source signals, the mixing system in (1) can be factorized as

$$\mathbf{x}(k,l) = \mathbf{H}_L(k,l)\mathbf{s}_L(k,l) + \mathbf{H}_R(k,l)\mathbf{s}_R(k,l) \tag{19}$$

where $\mathbf{H}_L(k,l)$ and $\mathbf{H}_R(k,l)$ are the $M \times M$ and $M \times [N(k,l)-M]$ matrix partitions, respectively. We assume for simplicity that the estimated demixing matrix $\mathbf{W}(k,l)$ is a scaled version of the inverse of $\mathbf{H}_L(k,l)$,

$$\mathbf{W}(k,l) = \mathbf{\Lambda}(k,l)\mathbf{H}_L^{-1}(k,l) \tag{20}$$

which means that the first M signal components belong to the most dominant sources. Note that if the constrain in (7) is imposed, the first column of $\mathbf{H}_L(k,l)$ corresponds to the mixing parameters of the target source while the others columns would correspond to the remaining most dominant interfering sources. Applying the MDP normalization to $\mathbf{W}(k,l)$ as in (17) the image at the microphones of the *m-th* source signal is obtained as

$$\begin{aligned}\mathbf{y}^m(k,l) &= \mathbf{s}^m(k,l) + \\ &+ [\mathbf{O}_{M\times m-1}|\mathbf{h}^m(k,l)|\mathbf{O}_{M\times M-m}]\times \\ &\times \mathbf{H}_L^{-1}(k,l)\mathbf{H}_R(k,l)\mathbf{s}_R(k,l) \\ &= \mathbf{s}^m(k,l) + \mathbf{d}^m(k,l)\mathbf{e}(k,l),\end{aligned} \tag{21}$$

where $\mathbf{d}^m(k,l) = [\mathbf{O}_{M\times m-1}|\mathbf{h}^m(k,l)|\mathbf{O}_{M\times M-m}]$ and $\mathbf{e}(k,l) = \mathbf{H}_L^{-1}(k,l)\mathbf{H}_R(k,l)\mathbf{s}_R(k,l)$. As expected, further the desired target source image, the output contains a term related to the residual $N(k,l)-M$ source signals, which are not suppressed by the linear demixing. However, due to the STFT disjointness of acoustic signals it may be assumed that only M sources are not negligible in each k and l. Under this assumption $\mathbf{s}_R(k,l)$ may be neglected. However, note that the noise in the outputs due to the residual unsuppressed sources also depends on $\mathbf{H}_L(k,l)$. Therefore, even when $\mathbf{s}_R(k,l)$ is small the noise is not negligible if $\mathbf{H}_L(k,l)$ approaches the singularity and the output may result distorted. To reduce this drawback a straightforward normalization is to limit the magnitude of the outputs in order to force that the overall filtering is limited by unity gain. This is imposed as

$$\overline{y}_{\widetilde{m}}^m(k,l) = min(|y_{\widetilde{m}}^m(k,l)|, |x_{\widetilde{m}}(k,l)|)\frac{y_{\widetilde{m}}^m(k,l)}{|y_{\widetilde{m}}^m(k,l)|}, \tag{22}$$

Here $y_{\widetilde{m}}^m(k,l)$ indicates the image of the *m-th* source signal at the $\widetilde{m}$*-th* microphone (i.e. $\mathbf{y}^m(k,l) = [y_1^m(k,l),\cdots,y_M^m(k,l)]^T$).

2.3. Channel-wise Wiener filtering postprocessing

The proposed constrained ICA adaptation is able to enhance a given source of interest, as long as we have a partial knowledge on the target source mixing parameters. Since in general $N(k,l) > M$ the time-varying linear demixing is only partially able to suppress the noise, specifically the signals of the most $M-1$ dominant sources. In order to better extract the image at the microphones of the target source we apply to the input signal $\mathbf{x}(k,l)$ a Wiener filter, which gains are determined according to the target and noise source spectral power. Here we limit to the case of $M = 2$ but the reasoning can be easily extended to the general case $M \geq 2$.

Indicating with $P_{\widetilde{m}}^t(k,l)$ and $P_{\widetilde{m}}^r(k,l)$ the power density spectra of the target and residual noise at the $\widetilde{m}$*-th* microphone and frames l, the image of the target source signal at the $\widetilde{m}$*-th* microphones is recovered through a channel-wise Wiener filter as

$$s_{\widetilde{m}}^1(k,l) = \frac{P_{\widetilde{m}}^t(k,l)}{P_{\widetilde{m}}^t(k,l) + P_{\widetilde{m}}^r(k,l)}x_{\widetilde{m}}(k,l). \tag{23}$$

Note, the dependence of the filter gain with the STFT frame l indicates a time-varying filtering, which is required due to the non-stationarity of the sources. According to (22) $P_{\widetilde{m}}^r(k,l)$ can be approximated with $E[|\overline{y}_{\widetilde{m}}^2(k,l)|^2]$, where the expectation is determined as the smooth average of $|\overline{y}_{\widetilde{m}}^2(k,l)|^2$ over time (i.e., assuming local stationarity). An approximation of $P_{\widetilde{m}}^t(k,l)$ can be derived from $E[|s_{\widetilde{m}}^1(k,l)|^2]$, where the power spectra of the

target source is estimated as:

$$|s^1_{\tilde{m}}(k,l)|^2 = \begin{cases} |\hat{s}^1_{\tilde{m}}(k,l)|^2, & if \quad \hat{s}^1_{\tilde{m}}(k,l) > 0 \\ 0, & otherwise \end{cases}, \quad (24)$$

$$\hat{s}^1_{\tilde{m}}(k,l) = y^1_{\tilde{m}}(k,l) - \mathcal{C}_{\tilde{m}}(k,l) y^2_{\tilde{m}}(k,l) + o_{\tilde{m}}(k,l) \quad (25)$$

where $\mathcal{C}_{\tilde{m}}(k,l)$ is a normalized correlation coefficient and $o_{\tilde{m}}(k,l)$ is an offset, used to limit the over-substraction generated by the over-estimation of $\mathcal{C}_{\tilde{m}}(k,l)$. Equation (25) is motivated by the fact that, according to (21), both $y^1_{\tilde{m}}(k,l)$ and $y^2_{\tilde{m}}(k,l)$ contains the same residual error, up to different scaling factors. Due to the time-frequency disjointeness of acoustic signals in the STFT domain, $y^1_{\tilde{m}}(k,l)$ may be approximatively assumed to be equal to 0 in the points where $s_2(k,l)$ is dominant. Therefore the substraction in (25) has effect only in time-frequency points where $s_2(k,l)$ is not dominant. The coefficient $\mathcal{C}_{\tilde{m}}(k,l)$ is determined as

$$\mathcal{C}_{\tilde{m}}(k,l) = \frac{E[|y^1_{\tilde{m}}(k,l)||y^2_{\tilde{m}}(k,l)|}{E[|y^1_{\tilde{m}}(k,l)|^2]}, \quad (26)$$

where the expectation $E[\cdot]$ indicates a smooth average over l. In order to limit the statistical bias introduced by $s_1(k,l)$ and $s_2(k,l)$ the average is computed considering only points where $\frac{1}{\alpha} < \frac{|y^1_{\tilde{m}}(k,l)|}{|y^2_{\tilde{m}}(k,l)|} < \alpha$, with $\alpha > 1$. Finally, the offset $o_{\tilde{m}}(k,l)$ is determined from a smooth average of $|\hat{s}^1_{\tilde{m}}(k,l)|$, over values $\hat{s}^1_{\tilde{m}}(k,l) < 0$.

2.4. SBSE system architecture

The global architecture of the SBSE is based on a multiple stage processing as depicted in figure 1. The system has been coded in C++ and works in real-time on a laptop. The sampled time-domain signals are transformed in a discrete time-frequency representation applying a Short-Time Fourier Transform (STFT) with overlapped Hanning windows in order to obtain frames with a certain degree of continuity in time. In order to have an accurate estimation of the mixing parameters $\mathbf{h}^1(k)$, which is assumed to be stationary (i.e., independent on l), the Recursively-Regularized ICA in frequency-domain [8] is applied to a segment where the target source dominates the remaining noise. A constrained on-line ICA adaptation is adopted as in 2.1, using as constrain the estimated target mixing parameters and imposing $\mu = 0$. The estimated spectra is used to compute the gain of the Wiener filter for each channel in order to get the spatial images of the target source at microphones. The enhanced signals are then beamformed according to the target mixing parameters as

$$s_{beam}(k,l) = \begin{bmatrix} 1; \frac{\hat{h}^1_1(k)}{\hat{h}^1_2(k)} \end{bmatrix} \mathbf{s}^1(k,l) \quad (27)$$

where $\hat{h}^1_1(k)$ and $\hat{h}^1_2(k)$ are the estimated impulse responses of the target source, truncated to $64ms$ (in order to enhance only the direct-path and the early reflections). Since the overall adaptation is on-line, the system is applied directly to the unsegmented utterances in continuous audio. Finally, the resulting mono target signals are segmented and each utterance is fed to the ASR system. In this work the performance are evaluated with the ideal segmentation, since this information is assumed to be known beforehand. However, since the system has been coded for working in a real-time real-world application, automatic segmentation (based on the estimated source direction and spectral power) is currently under development.

Note that while the RR-ICA and SBSS subsystems use a high frequency resolution in order to better handle long impulse responses, the Wiener filter uses a lower frequency resolution. The RR-ICA and SBSS system are based on ICA, which requires that the observed mixtures are linear combination of the original signals. This is true in the STFT domain if the framing window is sufficiently larger than the impulse response length. However, due to the scaling ambiguity, the outputs $\mathbf{y}^m(k,l)$ contain also noise in the phase due to the residual components $\mathbf{d}^m(k,l)\mathbf{e}(k,l)$. As the STFT window gets larger this noise degrades the quality of the output, generating an "artificial reverberation" effect. For this reason, the output of the SBSS is not directly used to reconstruct the target signal but only to give an estimation of the target and residual noise power. Furthermore, the double STFT resolution of the overall system allows us to inherit the benefits of a high frequency resolution filtering and of an increased sparse source representation obtained with smaller STFT frames.

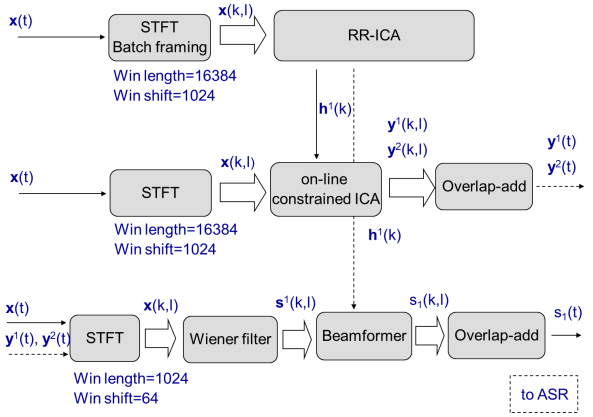

Figure 1: SBSE system architecture

3. Robust ASR

The task considers the problem of recognizing commands being spoken in a noisy living room from recordings made using a binaural manikin. The recognition system is based on the provided HTK setup: whole-word HMMs with topology described in [13] are trained with the reverberated Grid training data and speaker-dependent (SD) models are derived: the corpus consists of 34 speakers reading sentences which are simple sequences of the form:*[command] [color][preposition]*[**letter**]*[*[**digit**]*[adverb]*. As a result, the dictionary comprises 51 words and performance is measured as accuracy of two keywords for utterance (the letter and the digit tokens). The provided scripts perform the baseline training, the utterances decoding and related keyword accuracies computation.

3.1. Baseline

The baseline recognizer employs Mel Frequency Cepstral Coefficients (MFCC) and Cepstral Mean Normalisation.

Figure 2 reports Word Accuracy (%) for the development/test set before and after SBSE processing.

3.1.1. Acoustic features

To further improve performance of the ASR system we have introduced an alternative set of acoustic features, based on gam-

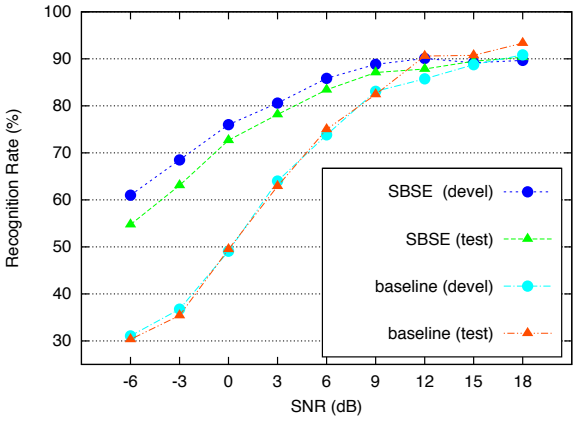

Figure 2: Keyword recognition accuracies for development and test sets, without and with SBSE processing for baseline setup.

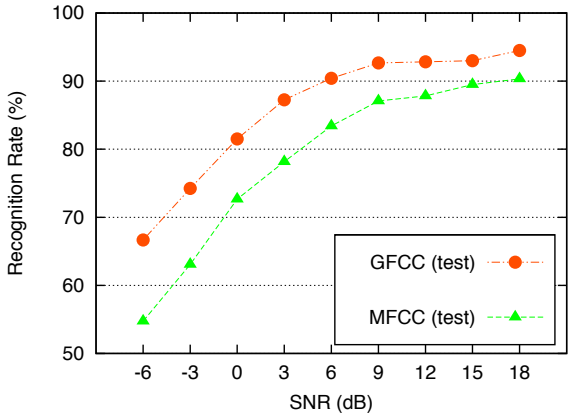

Figure 3: Comparison of recognition accuracies on test set using baseline MFCC and proposed GFCC.

matone analysis.

Gammatone filters (GF) are linear approximation of physiologically motivated processing performed by the cochlea, characterized by bandpass filters, whose impulse response is defined by:

$$g(t) = at^{n-1}cos(2\pi f_c t + \phi)e^{-2\pi bt} \quad (28)$$

where n is order of the filter, b is bandwidth of the filter, a is the amplitude, f_c is the filter center frequency and ϕ is the phase. The filter center frequencies and bandwidths are derived from the filter's Equivalent Rectangular Bandwidth (ERB) as detailed in [14]. The filter output of the m^{th} gammatone filter , X_m can be expressed by

$$X_m(t) = x(t) * h_m(t) \quad (29)$$

where $h_m(t)$ is the impulse response of the filter. These psychoacoustic inspired features prove to be robust against the residual noise and distortion induced by the SBSE processing.

In Figure 3 we present the comparison between the standard front-end (MFCC) and the proposed one (GFCC). For details about implementation and performance on other tasks see [15].

3.1.2. Model Training and Adaptation

To increase accuracy, we have then worked at model level, aiming at reducing the acoustic mismatch: typically it is possible to derive a suitable set of HMMs through retraining or adaptation. In this work two approaches have been tested: enlarged training and model adaptation.

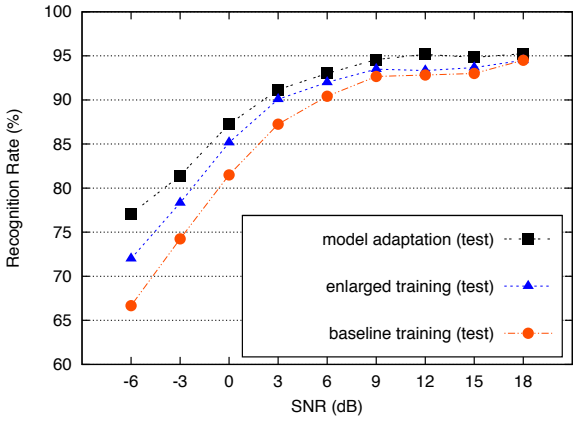

Figure 4: Keyword recognition accuracies on test set with different acoustic models: baseline, enlarged training (ET) and model adaptation (MA).

Enlarged Training (ET)

The standard training procedure uses the reverberated stereo signals; in the enlarged training different versions of the utterance are considered: besides the 17000 stereo files, we have added 17000 left and right monophonic channel signals and 17000 clean signals (the original waveform taken from the Grid corpus). The rationale behind this choice is that this redundancy can provide more generalization capability to the resulting acoustic model.

Model Adaptation (MA)

Starting from the Speaker Independent (SI) models, the baseline training procedure is modified applying a model adaptation based on a combination of MLLR and MAP. The development dataset and the test dataset are used to adapt the test and development models, respectively. MLLR is applied in the usual two-stage fashion: first a global adaptation is performed and the global transformation becomes the input transformation and a set of more specific transforms, using a regression class tree (with 128 nodes in our experiments), is estimated. After the MLLR step, an iteration of MAP adaption is performed. As a result, two sets of SD models are derived using the development and test material (i.e. all signals at different SNRs are pooled). Figure 4 shows the recognition curves for the two investigated approaches.

4. Summary and discussion

In the Tables 2 and 3 we report a summary of the recognition accuracies obtained with the illustrated techniques for the development and test sets, respectively. Different processing configurations are obtained combining the discussed methodologies, as described in Table 1.

The results show that the proposed processing is able to drastically reduce the impact of unwanted (and unknown) sources on the desired signal; the residual noise and distortion is effectively tackled by strategies at feature and model level. The introduction of parameters inspired by the human auditory system can provide additional robustness as well as some adaptation stages where the acoustic model benefits of audio examples of the SBSE chain.

The investigated methods maintain their benefits indepen-

C1	SBSE
C2	SBSE+GF
C3	SBSE+GF+ET
C4	SBSE+GF+MA
C5	SBSE+GF+ET+MA
C6	SBSE+MA
C7	SBSE+ET
C8	GF+ET+MA

Table 1: Processing configurations.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB
-	31.08	36.75	49.08	64.00	73.83	83.08
C1	61.08	68.67	76.00	80.67	85.83	88.83
C2	69.33	75.83	83.83	87.17	90.17	92.17
C3	76.08	81.67	87.33	89.92	92.17	93.67
C4	**79.25**	**84.25**	**88.25**	**91.00**	**93.08**	**93.92**
C5	**80.17**	**83.92**	**89.50**	**90.83**	**93.33**	**94.42**
C6	71.42	76.92	83.75	89.92	91.58	93.42
C7	66.33	73.50	79.17	83.83	86.50	90.83
C8	53.17	57.25	70.75	81.08	87.42	92.50

Table 2: Keyword recognition accuracies on CHiME development set according to the application of different processing configurations.

dently; we have observed improvements also with different combination of the strategies. For example SBSE provides tangible improvements also with MFCC and model adaptation. This may indicate also that GFCC are less sensitive to distortions introduced by the enhancement chain.

5. Conclusions

This paper presents and discuss recognition results on the PASCAL CHiME Speech Separation and Recognition Challenge. The binaural recorded mixtures are processed by a multistage Semi-Blind Source Extraction algorithm in order to obtain an estimation of the target signal.

The recovered target signal is processed by an automatic speech recognition (ASR) system which uses noise robust features based on Gammatone Frequency Cepstral Coefficients (GFCC). Standard MAP and MLLR adaptation techniques are then used to further mitigate the impact of the SBSE processing on the resulting models. Performance comparison between different model/algorithmic settings is reported for both development and test data sets. As discussed in Section 4 the proposed SBSE processing allows to tangibly reduce the error rate and demonstrate a good complementary with some standard approaches for robust speech recognition.

Future work is planned on a development of solutions in which SBSE techniques are tightly coupled with the ASR processing, for example developing the speech enhancement optimization in the acoustic features domain.

6. References

[1] W. Kellermann, "Some current challenges in multichannel acoustic signal processing," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3177–3178, 2006.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley and Sons, 2009.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB
-	30.33	35.42	49.50	62.92	75.00	82.42
C1	54.75	63.08	72.67	78.17	83.42	87.08
C2	66.67	74.25	81.50	87.25	90.42	92.67
C3	72.00	78.33	85.17	90.08	92.00	93.50
C4	**76.25**	**80.17**	**86.08**	**91.17**	**92.33**	**94.17**
C5	**77.08**	**81.42**	**87.25**	**91.17**	**93.00**	**94.58**
C6	71.17	76.42	82.58	86.50	88.83	91.67
C7	60.75	67.33	76.83	80.75	85.67	89.42
C8	51.58	57.25	70.67	79.67	85.92	92.67

Table 3: Keyword recognition accuracies on CHiME test set according to the application of different processing configurations.

[3] H. Christensen, J. Barker, N. Ma, and P. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Proceedings of Interspeech*, Makuhari, Japan, 2010.

[4] T. C. M. P., J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[5] J. Droppo and A. Acero, *Environmental Robustness*. Springer Handbook of Speech Processing, 2008.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech*, Nov. 2007.

[7] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 650 –664, May 2009.

[8] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 624 –639, march 2011.

[9] F. Nesta, T. Wada, and B.-H. Juang, "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 583 –599, 2011.

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[11] T. Melia, "Underdetermined blind source separation in echoic environments using linear arrays and sparse representations," Ph.D. dissertation, University College Dublin, 2007.

[12] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Symposium on ICA and Blind Signal Separation*, San Diego, CA, USA, Dec. 2001.

[13] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.

[14] M. Slaney, "An efficient implementation of the patterson holdsworth auditory filterbank," Apple Computers, Perception Group, Tech. Rep., 1993.

[15] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 570–573.