A FLEXIBLE SPATIAL BLIND SOURCE EXTRACTION FRAMEWORK FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENTS

Francesco Nesta, Marco Matassoni*

Center of Information Technology, Fondazione Bruno Kessler
Trento, Italy
francesco.nesta@gmail.com,matasso@fbk.eu

Ramon Fernandez Astudillo†

Spoken Language Systems Laboratory, INESC-ID,
Lisbon
ramon@astudillo.com

ABSTRACT

Blind source extraction (BSE) is an attractive approach to enhance multichannel noisy speech data, as a preprocessing step for an automatic speech recognition system. BSE was successfully applied to the first Chime Pascal Challenge for improving the recognition rate of noisy commands in a small dictionary task. In this work we reviewed the BSE architecture and improved each system block in the framework in order to increase its flexibility and degree of blindness. Two different algorithms are finally implemented to address both Tracks of the 2nd Chime Challenge. To improve the overall performance, the output of the enhancement algorithm is then combined with robust ASR systems based on gammatone features analysis and on uncertainty decoding. Results obtained with different front-end and back-end configurations demonstrate the advantages of the proposed approaches.

Index Terms— multi-channel audio, source separation, robust speech recognition, speech enhancement, uncertainty decoding, gammatone features

1. INTRODUCTION

Voice-based human-machine interaction is attracting a lot of attention and significant results have been achieved in controlled conditions. Nevertheless speech acquisition, processing and recognition in non-ideal acoustic environments are still complex tasks [1][2].

The presence of environmental noise, reverberation and interfering speakers often causes a dramatic performance drop on automatic speech recognition (ASR) systems. To improve ASR robustness, different approaches have been widely investigated: spatial speech processing [3][4][5][6], alternative acoustic features [7], model compensation or adaptation [8, 9], uncertainty decoding [10] are popular approaches proposed to tackle this problem. However, in order to lead to effective solutions for real-world tasks, a careful combination of each single technique is necessary.

As such, in 2011 the first CHiME challenge considered the problem of recognizing speech in everyday noise situations. The 2nd CHiME challenge [11] extends the difficulty of the recognition task increasing the vocabulary size and introducing non-stationary mixing conditions for the target speaker, i.e. the speaker is allowed to do small movements.

In the former work [5] a complete system was proposed, which was able to sensibly improve the recognition performance for the 1st CHiME challenge task. The system is based on the effective combination of spatial processing based on the Blind Source Extraction (BSE) and robust ASR exploiting robust feature analysis. In this work [1] we revise and extend the original work in order to deal with both Track 1 and Track 2 of the 2nd CHiME Challenge. By combining different processing stages the described algorithm is then able to deal with both static and dynamic mixing conditions and is unsupervised, i.e. it does not require any specific knowledge on the target speaker other than the its direction. While the focus of this work is on the description of each stage required to build the full BSE processing, a detailed analysis is reported on the recognition performance obtained using robust ASR strategies, such as gammatone features or uncertainty decoding.

The article is organized as follows:

- Section 2 describes the general architecture of the BSE system which includes as main blocks the spatial dictionary learning, the constrained spatial filtering and the spectral filtering, described in Section 3, 4, 5;
- Section 6 describes the back-end systems used for the ASR.
- Section 7 discusses the recognition performance obtained combining different variants of the front-end and of the back-end systems and using different acoustic models and datasets in CHiME;
- concluding remarks end the discussion in Section 8.

2. BSE SYSTEM ARCHITECTURE

Although the BSE framework can be easily extended to a more general multichannel case, to simplify the discussion and be more consistent with the CHiME tasks we will explicitly refer to the case of two microphones.

An unknown number of source signals are recorded by an array of 2 microphones. We refer to the discrete time-frequency representation of signals, obtained for example through the Short-time Fourier Transform (STFT). Let $X_m(k,l)$ indicate the *l-th* STFT frame coefficients obtained for the k-*th* frequency bin for the m-*th* mixture signal. For convenience of notation we indicate the vector of signal mixtures as $\mathbf{X}(k,l) = [X_1(k,l)\ X_2(k,l)]^T$.

For each frequency bin k the time series $\mathbf{X}(k,l)$ obtained with all the incoming frames l represent the main input of a BSE system, whose general structure is depicted in Figure 1. The first block at the top of the chain has the goal to learn a dictionary of mixing systems

*now at Conexant Systems, Inc., Newport Beach, CA (USA)

†This work was partially funded by the FCT through the grant number SFRH/BPD/68428/2010 and the project PEst-OE/EEI/LA0021/2011.

[1] The research leading to these results has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement n 288121 - DIRHA.

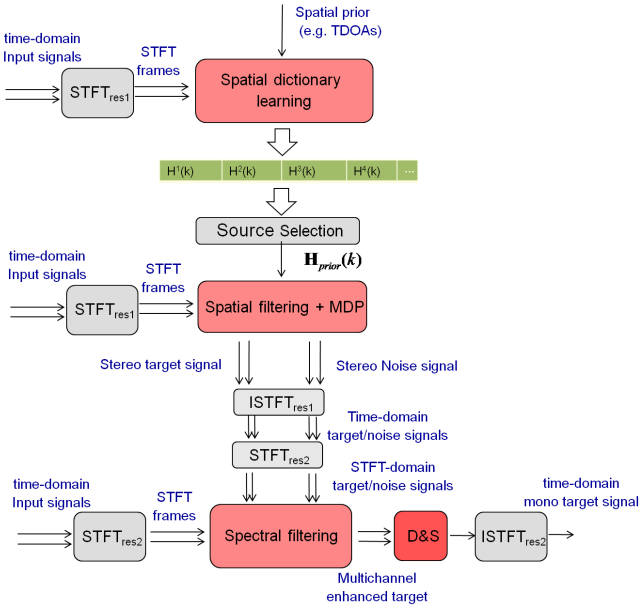

Fig. 1. Architecture of BSE system

describing the propagation of a source in a specific location. The mixing systems are related to the full echoic representation of the impulse responses between each spatial location and each microphone and represent a prior information used to constrain the subsequent spatial filtering. Among all the estimated mixing systems the target direction is used to select the one related to the wanted source. The selected mixing system, indicated as $H_{prior}(k)$, is used to constrain an ICA adaptation [5] in order to estimate a demixing system able to suppress the noise and preserve the target source signal. The stereo signals for both noise and target source are recovered by applying the Minimal Distortion Principle [12] to the estimated demixing matrix. Signals are then reconstructed back to time-domain and the noise and target signals are processed through a spectral filtering algorithm, such as an adaptive Wiener filter. The spectral filtering is applied to each channel separately using a time-frequency resolution better suited for the single channel spectral processing. Finally, the enhanced target signals are combined together into a single channel through a delay-and-sum beamformer and reconstructed back to time-domain.

3. SPATIAL DICTIONARY LEARNING

In a complex acoustic scene, multiple sources can be active and overlapping in some time instants. However, it is reasonable to assume that in a sufficient number of STFT frames only one source is dominant. Then, each frame is used to adapt the mixing system related to the spatial location of the dominating source. The learning is performed through a weighted Natural Gradient (wNG) similar to that proposed in [13]. First, we indicate with $H_m^o(k)$ the discrete frequency response between the location o and the microphone m, where k is the frequency bin index according to the given STFT analysis. A normalized vectorial representation of the response is obtained as

$$\mathbf{d}^o = \left[\frac{H_2^o(1)H_1^o(1)^*}{|H_2^o(1)H_1^o(1)^*|}, \cdots, \frac{H_2^o(N_{bins})H_1^o(N_{bins})^*}{|H_2^o(N_{bins})H_1^o(N_{bins})^*|} \right]^T \quad (1)$$

where N_{bins} indicates the total number of discrete frequency bins. The vector $\mathbf{d}^o$ gives a compact representation of the inter-channel phase difference, in the complex domain, which is different for each location o and reverberation conditions. If the room geometry and microphone array location is available, the spatial dictionary can be initialized using simulated frequency responses, e.g. through the image simulation method (ISM) [14][15]. In the simplest case where this information is not available the dictionary can be initialized using anechoic frequency response models describing the propagation of a source in a given direction as

$$H_1^o(k) = 1, \quad H_2^o(k) = e^{2\pi f_k \frac{d \times sin\theta^o}{c}} \quad (2)$$

where f_k indicates the frequency associated to the bin k, d is the microphone distance, c is the sound speed and θ^o is the angle of the source at the o-*th* location, with the respect to the broadside array direction. The mixing matrices associated to each location o, describing the acoustic propagation at the frequency bin k, are initialized as

$$\widehat{\mathbf{H}}^o(k) = \begin{bmatrix} 1 & 0 \\ H_2^o(k) & 1 \end{bmatrix}, \quad \forall o. \quad (3)$$

Similarly to the atom definition, each frame is represented as

$$\mathbf{R}^l = \left[\frac{X_2(1,l)X_1(1,l)^*}{|X_2(1,l)X_1(1,l)^*|}, \cdots, \frac{X_2(N_{bins},l)X_1(N_{bins},l)^*}{|X_2(N_{bins},l)X_1(N_{bins},l)^*|} \right]^T \quad (4)$$

For each frame we select the atom in the spatial dictionary best projecting with the observed frame l

$$\tilde{o} = \arg\max_o \Pr(o,l), \quad \Pr(o,l) = |(\mathbf{d}^o)^* \, \mathbf{R}^l|, \quad (5)$$

where $*$ indicates the complex transpose, and normalize the respective projection as

$$\overline{\Pr}(\tilde{o},l) = \frac{\Pr(\tilde{o},l) - \Pr_{\tilde{o}}^{min}}{\Pr_{\tilde{o}}^{max} - \Pr_{\tilde{o}}^{min}} \quad (6)$$

where $\Pr_{\tilde{o}}^{min}$ and $\Pr_{\tilde{o}}^{max}$ are the minimum and maximum projection of the atom $\tilde{o}$ with all the previously observed data frames. The normalized projection is then a weight with values ranging from 0 to 1, indicating the dominance of the source at the location $\tilde{o}$ and at the frame l.

A weighting matrix $\mathbf{D}^{\tilde{o}}(l)$ is defined as a diagonal matrix with the first element (i.e. p_{11}) equal to $\overline{\Pr}(\tilde{o},l)$ and the second element (i.e. p_{22}) set to $1-\overline{\Pr}(\tilde{o},l)$. Then, according to the weighted NG, for each frame l, the atom selected in (5) and its corresponding mixing system is updated as follows

$$\mathbf{Y}(k,l) = [\widehat{\mathbf{H}}^{\tilde{o}}(k)]^{-1}\mathbf{X}(k,l) \quad (7)$$

$$\Delta\mathbf{H}(k) = [\widehat{\mathbf{H}}^{\tilde{o}}(k)(\mathbf{I} - \Phi(\mathbf{Y}(k,l))\mathbf{Y}(k,l)^H)]\mathbf{D}^{\tilde{o}}(l) \quad (8)$$

$$\widehat{\mathbf{H}}^{\tilde{o}}(k) = \widehat{\mathbf{H}}^{\tilde{o}}(k) - \eta\Delta\mathbf{H}(k) \quad (9)$$

$$\mathbf{d}^{\tilde{o}} = \left[\frac{\widehat{H}_{21}^{\tilde{o}}(1)\widehat{H}_{11}^{\tilde{o}}(1)^*}{|\widehat{H}_{21}^{\tilde{o}}(1)\widehat{H}_{11}^{\tilde{o}}(1)^*|}, \cdots, \frac{\widehat{H}_{21}^{\tilde{o}}(N_{bins})\widehat{H}_{11}^{\tilde{o}}(N_{bins})^*}{|\widehat{H}_{21}^{\tilde{o}}(N_{bins})\widehat{H}_{11}^{\tilde{o}}(N_{bins})^*|} \right]^T \quad (10)$$

$\mathbf{Y}(k,l)$ is the vector of the demixed signal $[Y_1(k,l) \;\; Y_2(k,l)]^T$, where $\widehat{H}_{mn}^{\tilde{o}}(k)$ is the generic element of the matrix $\widehat{\mathbf{H}}^{\tilde{o}}(k)$, η is the adaptation step-size, $\mathbf{I}$ the identity matrix and $\Phi(\cdot)$ is a non-linearity. In practice, the weighting matrix induces the gradient to update the first column of $\widehat{\mathbf{H}}^{\tilde{o}}(k)$ when the source located in $\tilde{o}$ is dominant.

3.1. Source selection

The learned dictionary gives a spatial representation of the mixing parameters related to different source locations. However, it is also required a criterion to select the correct atom in order to selectively extract a given source of interest. In this work we select the atom having the maximum cumulative projection over all the frames of a sentence, but restricting the search only to atoms related to a given source direction range.

4. CONSTRAINED SPATIAL FILTERING

The spatial filtering can be applied with any kind of ICA adaptation, constrained with the mixing system selected from the dictionary. Here we use a constrained weighted NG where the adaptation is applied as follow

$$\widetilde{\mathbf{X}}(k,l) = [\widehat{\mathbf{H}}^{o^t}(k)]^{-1}\mathbf{X}(k,l) \tag{11}$$

$$\mathbf{Y}(k,l) = [\mathbf{H}(k)]^{-1}\widetilde{\mathbf{X}}(k,l) \tag{12}$$

$$\Delta\mathbf{H}(k) = [\mathbf{H}(k)(\mathbf{I} - \Phi(\mathbf{Y}(k,l))\mathbf{Y}(k,l)^H)]\mathbf{D}(l) \tag{13}$$

$$\mathbf{H}(k) = \mathbf{H}(k) - \eta\Delta\mathbf{H}(k) \tag{14}$$

$$\mathbf{W}(k) = [\mathbf{H}(k)]^{-1}[\widehat{\mathbf{H}}^{o^t}(k)]^{-1} \tag{15}$$

with $\widehat{\mathbf{H}}^{o^t}(k)$ indicating the matrix selected from the dictionary related to the location of the target source, $\mathbf{W}(k)$ is the full demixing matrix applied to the input data and $\mathbf{D}(l)$ is the weighting matrix. The matrix $\mathbf{D}(l)$ in this case defines the adaptation rate of the parameter related to the target source and to the noise sources and should be defined according to the expected scenario. In CHiME-like scenarios, the mixing systems of target and noise sources have different characteristics. The main speaker location can be considered relatively static, i.e. the speaker does not change location during the interaction. On the other hand, the noise sources can quickly move in the space and then the adaptation must quickly track variations of their mixing conditions. If the first element on the main diagonal of $\mathbf{D}(l)$ is set to zero (i.e. $p_{11} = 0$) the target mixing parameters will remain unaltered during the adaptation. This is acceptable if the target source is perfectly static since the mixing parameters estimated in the previous stage are enough accurate to describe the static part of the channel frequency response. However, if the target source is expected to do small movements, p_{11} should be > 0 in order to allow a certain degree of adaptation and compensate any mismatch. Similarly, the second diagonal element, p_{22} must be set to a higher value in order to better track fast variations in the noise mixing parameters. In general, p_{11} and p_{22} should be proportional to the probability of speech and noise presence, which in this work are approximated with $\overline{\Pr}(o^t, l)$ and $(1 - \overline{\Pr}(o^t, l))$. In order to be more robust to errors in $\overline{\Pr}(o^t, l)$, in this work we generate the weights through non-linear transformation of $\overline{\Pr}(o^t, l)$:

- **track1**: $p_{11} = 1 - \tanh[\alpha \times (1 - \overline{\Pr}(o^t, l))]$, $p_{22} = 1 - \tanh[\alpha \times \overline{\Pr}(o^t, l)]$
- **track2**: $p_{11} = 0, p_{22} = 1 - \tanh[\alpha \times \overline{\Pr}(o^t, l)]$

where α is a parameter defining the sensitivity of the weight to the target/noise presence. Note, while p_{22} is formulated in the same way for both Track1 and Track2, p_{11} is defined differently in order to better fit the characteristic of each track. In Track2 the speaker is static and the re-adaptation of the target mixing system is not necessary. On the other hand, in Track1 the target speaker is expected to do small movements and a certain degree of adaptation is required.

4.1. Forward-Backward on-line tracking

In common on-line adaptations filters are updated with the incoming data, starting from values estimated in the previous time instant. On the other hand, if multiple time frames are known beforehand, batch adaptations generally lead to more accurate results, on conditions that the mixing system remains stationary for the entire batch of analysis. However, in scenarios with highly non-stationary mixing conditions, such as those simulated in the CHiME challenge, batch adaptations are not appropriated since a continuous tracking of variations in the mixing parameters is necessary to recover the source signals with small distortion. An on-line adaptation would better adapt to local variations but it will not exploit future observations as done by batch processing.

In order to overcome this limitation, in this work we combine the advantages of on-line and batch adaptations using a backward-forward (BF) tracking on-line adaptation. For each noisy sentence of the CHiME datasets the adaptations in both "spatial dictionary learning" and "spatial filtering" stages, are iterated alternatively in the forward and backward direction over all the available data frames. This procedure implicitly constrains the overall adaptation to optimize the estimated filters over all the available data, without making any strong assumption of long-term stationarity as for traditional batch ICA optimizations.

An approximate pseudo-code description of the procedure is explained as follows:

Initialize l=0;d=1;
for N_{BF}
  while [($l < N_l$) and (d==1)] or [($l > 0$) and (d==-1)]
  set current frame to l=l+d;
  Apply the spatial dictionary learning as in (7)-(10)
  Apply the constrained spatial filtering as in (11)-(15)
  endwhile
if (d==-1)
  d=1 (set forward tracking)
else
  d=-1 (set backward tracking)
end
endfor

---

where N_{BF} is the number of BF tracking iterations and N_l is the total number of STFT frames. The BF procedure is run with a redundant number of iterations but with a small adaptation step-size. This approach improves the robustness of the parameters tracking against strong noise localized in some frames, and eventually converges to a globally optimized solution. For each frame, after the adaptation of the spatial filtering procedure the Minimal Distortion Principle (MDP) [12] is used to estimate the multichannel image of target source and noise signal (for more details see [5]). Finally, the STFT signals are reconstructed back to time-domain through a weighted Overlap-and-add (WOLA) using the Griffin and Lim's method [16]. It is worth noting that the reconstruction to time-domain is required because the spectral filtering operates at different time-frequency resolution, i.e. a fine temporal resolution is necessary to capture the non-stationarity of the source signals.

5. SPECTRAL FILTERING

For best performance, the enhanced target signals are not directly fed to the ASR but used to control a further spectral processing stage applied to each input channel. In this stage the recorded microphone signals are filtered with a spectral enhancement method operating in

a time-frequency resolution domain, different from that used for the constrained spatial filtering. In fact, the spectral filtering requires a higher temporal resolution domain in order to be consistent with the high non-stationarity of the recovered audio signals. On the other hand, the spatial filtering requires a high frequency resolution in order to approximate the mixing process from convolutive to linear.

In this work we adopted an adaptive Wiener filter applied separately to each channel of the recovered signals. Starting from the matrix $\mathbf{W}(k)$ estimated in (15) (in each frame l), we indicate with $Y^m_{\widetilde{m}}(k,l)$ the estimate of the *m-th* source signal recorded at the $\widetilde{m}$-*th* microphone, obtained applying the MDP [12]. Indicating with $P^t_{\widetilde{m}}(k,l)$ and $P^n_{\widetilde{m}}(k,l)$ the Power Spectral Density (PSD) of the target and noise at the $\widetilde{m}$-*th* microphone and frames l, the image of the target source signal at the $\widetilde{m}$-*th* microphones can be estimated as

$$S^1_{\widetilde{m}}(k,l) = \frac{P^t_{\widetilde{m}}(k,l)}{P^t_{\widetilde{m}}(k,l) + P^n_{\widetilde{m}}(k,l)} X_{\widetilde{m}}(k,l). \tag{16}$$

$P^n_{\widetilde{m}}(k,l)$ can be approximated with $|Y^2_{\widetilde{m}}(k,l)|^2$. Indeed, if the selected atom well represents the mixing parameters of the target source, the target signal is perfectly canceled from the noise output. In contrast, $P^t_{\widetilde{m}}(k,l)$ cannot be directly derived from $|Y^1_{\widetilde{m}}(k,l)|^2$ because in the general case of a number of sources $N > 2$, it is not possible to spatially suppress all the noise from the target outputs with a 2×2 demixing system. A possible estimation for $P^t_{\widetilde{m}}(k,l)$ is obtained as

$$e(k,l) = |Y^1_{\widetilde{m}}(k,l)| - g(k) \times |Y^2_{\widetilde{m}}(k,l)| \tag{17}$$

$$P^t_{\widetilde{m}}(k,l) = \max(|e(k,l)|^2, 0) \tag{18}$$

where $g(k)$ is a gain estimated in order to minimize the Mean Square Error (MSE) $E[|e(k,l)|^2]$. The gain $g(k)$ can be adapted on-line with a Normalized Least Mean Square algorithm (NLMS). The adaptation has the goal to remove the residual noise from the target signal, correlated to the estimated noise signal. However, if a residual component of the target is still present in $Y^2_{\widetilde{m}}(k,l)$ the filtering may result in an unwanted attenuation of the target signal. In order to reduce this distortion, filters are updated only when the target signal is sufficiently smaller than the output signal, e.g. $|Y^1_{\widetilde{m}}(k,l)| < \beta \times |Y^2_{\widetilde{m}}(k,l)|$.

As for the constrained spatial filtering, the enhanced frequency-domain signals are reconstructed back to time-domain through a weighted Overlap-and-add (WOLA) using the Griffin and Lim's method [16]. Since for the recognition task only a single signal is required, the channels are combined together with a delay&sum beamformer.

6. ASR BACK-END

6.1. Baseline ASR

In the experiments the short-term spectral analysis is performed with windows of $25ms$ and step-size of $10ms$. Mel Frequency Cepstral Coefficients (MFCCs) and log-energy plus the corresponding first and second order time derivatives are combined in a 39-size feature vector. Cepstral Mean Normalizations is also applied.

The recognition system for the *Track 1* task is based on whole-word HMMs with topology described in [17], trained with the reverberated Grid training data. 34 speaker-dependent (SD) models are derived. The sentences are sequences of the form:*[command] [color][preposition][letter][digit][adverb]*. Performance is measured as accuracy of two keywords for utterance (letter and digit).

The baseline ASR back-end for *Track2* is based on a popular setup [18]: the acoustic model comprises 39 phones plus two silence models (silence and short pause, tied together); the topology is 3-states left-to-right with no skips. Each phone HMM is represented by a GMM with 8 components while the silence uses 16 Gaussians. The provided training scripts starts from a clean acoustic model with some re-estimations steps. The language model is build from the standard 5K non-verbalized closed bigram provided in the original WSJ distribution.

A second set of experiments are carried out with an alternative feature set based on gammatone analysis in order to confirm the effectiveness shown in the first CHiME challenge.

The gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea; the filter center frequencies and bandwidths are derived from the filter's Equivalent Rectangular Bandwidth (ERB) as detailed in [19]. Additionally a Shifted Log is used as the (non-linear) compression function for the spectral representation:

$$Y = \log_{10}(X + \alpha_0) \tag{19}$$

where α_0 is a parameter that controls a threshold in the log function and simulates the human auditory rate-intensity curve. The 32-band energies are then decorrelated using the standard Discrete Cosine Transform to obtain a 13-dimensional observation vector, extended, as for the reference Mel feature, with first and second derivatives for a total of 39 components.

6.2. Sparsity based Acoustic Model Compensation

To further increase the robustness, the acoustic models of an ASR system can be dynamically compensated for the uncertainty of the incoming signal. In the context of multi-channel signal processing, the residual uncertainty after signal processing in STFT domain can be used for this purpose. An uncertain signal description in STFT domain can be related to an uncertain signal description in MFCC or other feature domains by using uncertainty propagation [10]. Once the uncertainty in MFCC domain has been estimated, observation uncertainty techniques such as uncertainty decoding and modified imputation can be used to compensate the acoustic models or the features. In [20] a measure of the residual uncertainty after beamforming and Wiener post-processing was attained from the residual MSE [20]. One disadvantage of this approach is that the MSE does not capture errors in the estimation of the target and noise PSDs. Furthermore, MSE propagation relies upon the assumption of additivity of target and noise, while the assumption of sparseness, either target or noise are active, often fits better source separation scenarios.

Here, we develop an uncertainty model in STFT domain based on the assumption of sparsity. We also briefly describe how to propagate such a model into STFT domain. Under the assumption of sparsity, each time-frequency bin of the observed signal $X_{\widetilde{m}}(k,l)$ contains either target or noise. In real world situations it is however impossible to determine which of the two signal is active with absolute certainty. In this work we model such uncertainty in following form. We consider each time-frequency bin of the target signal $S_{\widetilde{m}}(k,l)$ as an independent random variable with two possible outcomes. Either target is active and thus $|S_{\widetilde{m}}(k,l)| = |X_{\widetilde{m}}(k,l)|$ or noise is active and thus $|S_{\widetilde{m}}(k,l)| = 0$. The amplitude of the STFT of such an uncertain signal is therefore described as a matrix of independent scaled Bernoulli variables. This model is thus different from the complex Gaussian uncertain STFT model used in [20] and other works, and the conventional propagation formulas do not apply.

We can however make use of the transformations involved in the MFCC to approximate uncertainty propagation for this model. To attain this we approximate the large sum of scaled Bernoulli random variables at the Mel-filterbank

$$M_{jl} = \sum_{k=1}^{K} W_{jk} |S_{\tilde{m}}(k,l)| \tag{20}$$

by a continuous log-normal distribution as in [10, Sec. 6.3.3], where W_{jk} are the Mel-filterbank weights. As a result of this assumption, the distribution of the MFCC is Gaussian and only mean and variance of the amplitude at each time frequency bin $S_{\tilde{m}}(k,l)$ are needed to compute propagation. The only missing element is the probability of target activity $\hat{p}(k,l)$, which in this case is estimated from the Wiener gain of (16) as

$$\hat{p}(k,l) \approx \frac{\sqrt{P^t_{\tilde{m}}(k,l)}}{\sqrt{P^t_{\tilde{m}}(k,l)} + \sqrt{P^n_{\tilde{m}}(k,l)}} \tag{21}$$

Note that here an amplitude based gain, rather than a conventional Wiener gain, is used since it led to better results. In practice any binary mask could be used.

The Variance of each uncertain MFCC is then obtained from the log-normal assumption and the mean and variance of a scaled Bernulli distribution. This is given by

$$\Sigma^s_{iil} \approx \sum_{j=1}^{J} \sum_{j'=1}^{J} T_{ij} T_{ij'} \cdot \log\left(\frac{\sum_{k=1}^{K} W_{jk} W_{j'k} \Sigma^{|S|}_{\tilde{m}}(k,l)}{\bar{M}_{jl} \bar{M}_{jil}} + 1 \right) \tag{22}$$

where i is the DCT index, T_{ij} are the DCT coefficients and

$$\bar{M}^1_{jl} = \sum_{k=1}^{K} W_{jk} \hat{p}(k,l) |X_{\tilde{m}}(k,l)| \tag{23}$$

are the Mel-filterbank features of the spectrum filtered by (21). This is also equivalent to the expected Mel-filterbank features for the uncertainty model used. The variance of the amplitude of each uncertain Fourier coefficient can be determined from the variance of a Bernoulli random variable as

$$\Sigma^{|S|}_{\tilde{m}}(k,l) = \hat{p}(k,l)\Big(1 - \hat{p}(k,l)\Big)|X_{\tilde{m}}(k,l)|^2. \tag{24}$$

This measure of uncertainty is maximal when $\hat{p}(k,l) = 0.5$, that is when the PSDs of target and noise are equal. It is also minimal when one of the two PSDs is zero, thus penalizing the violation of the sparsity assumption with a higher uncertainty. Consequently this measure takes into consideration the errors in PSD estimation in a direct way, unlike MSE based estimation. It should be noted that the propagation of an uncertain spectrum using the log-normal assumption also implies a bias compensation of the mean of the MFCCs [10, Sec. 6.3.3]. This compensation led however to slightly worse results and was ignored. This could be due to the limited validity of the log-normal assumption for the scaled Bernoulli uncertainty model.

It should also be noted that measures of uncertainty based on the signal processing stage perform worse when used in noise matched conditions. The rationale for this is that these measures relate to how different is the estimated spectrum from the original clean signal. Since in noise matched conditions the models are trained with noisy data, distortions that would lead to low or medium uncertainties have been probably learned by the model and do not need to be compensated for. Uncertainty is therefore over-estimated in those cases.

In the case of sparsity based uncertainty, the artifacts caused by PSD estimation errors are learned from the training data and thus compensating them yields no improvement or a performance decrease. Consequently results are only provided for the *rever* test conditions, where the models were trained with reverberant speech.

General parameters
fs=16kHz, STFT window = Hamming
$\Phi(x) = \tanh(10 \cdot \lvert x \rvert) \exp(j\phi(x))$
BF_{iter}=2 (with-mem), BF_{iter}=10 (no-mem)
Spatial dictionary learning
STFT frame length/shift = 4096/256 samples
Dictionary size = 60 atoms, $\eta = 0.01$
Constrained spatial filtering
STFT frame length/shift = 4096/256 samples
$\eta = 0.05$, $\alpha = 2$
Spectral filtering
STFT frame length/shift = 1028/128 samples
NLMS step-size adaptation $\mu = 0.02$, $\beta = 1.2$

Table 1. Summary of parameters used in the BSE algorithm

7. PERFORMANCE EVALUATION

7.1. BSE parameter settings

The parameters of the BSE processing are summarized in Table 1 and were optimized only using the development dataset. For the evaluation of the ASR task we also considered two different operative modalities named as *with-mem* and *no-mem*. The first refers to the processing with memory over the entire dataset, i.e. the estimated spatial dictionary is sequentially propagated during the processing of the mixtures of each dataset. This is motivated by the fact that the spatial dictionary gives an average representation of the environmental acoustic, which should not change considerably over time. On the other hand, the second modality *no-mem* refers to the processing without memory, i.e. the spatial dictionary is re-initialized for each processed mixture.

7.2. Track 1

This section presents the results for *Track 1* reported as (keyword) recognition accuracies. Besides the reverberated (*rever*) and noisy (*noisy*) acoustic models (AM) already available in the CHiME datasets, a new set of models have been obtained by filtering the noisy training dataset with the same BSE processing. We refer to this set of models with *BSE-matched*, i.e. the ASR is directly matched with the BSE enhancement. Table 2 summarizes the setup adopted for acoustic model training.

id	**training dataset**	**BSE processing**
rever	reverberated	no
noisy	noisy	no
bse-matched	noisy	yes

Table 2. The sets used for AM training.

Table 3 reports the results obtained with the baseline ASR and the provided *rever* AM. Additional results are provided in Table 4, which shows the accuracies obtained when the AM training procedure is modified as described in [5] (we refer to it as a *modified*

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
dev	32.08	36.33	50.33	64.00	75.08	83.50	**56.9**
test	32.17	38.33	52.08	62.67	76.08	83.83	**57.5**

Table 3. Keyword recognition accuracies of unprocessed **test** and **dev** sets of Track 1, obtained with the ASR baseline and using *rever* AM.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
dev	46.92	49.00	60.00	73.33	81.33	89.42	**66.7**
test	43.58	49.42	63.42	72.00	82.75	89.67	**66.8**

Table 4. Keyword recognition accuracies of unprocessed **test** and **dev** sets of Track 1, obtained using *rever* data and the ASR baseline with the *modified training*.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	53.25	56.92	67.33	78.08	84.33	88.58	**71.4**
with-mem	55.67	61.25	69.75	81.25	86.33	91.58	**74.3**
no-mem	51.42	56.75	67.75	76.83	84.58	87.75	**70.9**
with-mem	53.75	60.25	74.08	81.92	87.67	90.33	**74.7**

Table 5. Keyword recognition accuracies of BSE processed **dev/test** sets, obtained using *rever* data and the ASR baseline with the *modified training*.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	68.67	71.08	78.50	84.42	87.25	89.50	**79.9**
with-mem	68.75	71.50	79.42	86.08	87.83	89.42	**80.5**
no-mem	66.58	74.08	80.42	85.00	86.17	89.33	**80.3**
with-mem	66.67	70.83	80.17	84.67	88.42	89.92	**80.1**

Table 6. Keyword recognition accuracies of BSE processed **dev/test** sets, obtained using *bse-matched* data and the ASR baseline with the *modified training*.

training). In both the cases no BSE processing was applied to the recorded mixtures.

While the BSE already produces a sensible improvement with the *rever* models (see Tables 5) , applying the processing also during training provides more robust models as demonstrated by Table 6. Indeed, the best performance is achieved in case of *bse-matched* training: the corresponding AM compensates the residual distortions introduced by the BSE processing and learned in the training phase.

Finally, Table 7 reports the performance for the test set, obtained with an alternative front-end based on gammatone analysis, proving the advantages of using robust features for further mitigate the effect of residual distortion not learned in the training. For the sake of completeness we also report in Table 8 the performance obtained using the original noisy AM models *noisy*, i.e. obtained from the noisy training dataset but not processed by the BSE.

7.3. Track 2

In this section a wide analysis of recognition performance is presented reporting the results on the two sets (**dev** and **test**). Note, to be compliant with the official metric used in CHIME for this track, the performance are evaluated through the Word Error Rates (%). As for Track 1, the two provided AMs (*rever* and *noisy*) result from two

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	69.17	74.08	81.25	87.08	89.67	90.50	**82.0**
with-mem	69.67	73.50	81.67	86.17	88.67	89.33	**81.5**
no-mem	70.00	75.42	83.75	87.00	90.08	91.33	**82.9**
with-mem	**69.00**	**77.50**	**83.50**	**87.42**	**90.42**	**91.33**	**83.2**

Table 7. Keyword recognition accuracies of BSE processed **dev/test** sets, obtained using *bse-matched* data and the ASR baseline with Gammatone front-end and the *modified training*.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	69.67	74.50	80.92	84.75	88.08	88.42	**81.1**
with-mem	69.58	74.67	81.17	87.33	89.17	88.92	**81.8**
no-mem	69.92	75.25	82.17	85.00	87.83	88.75	**81.5**
with-mem	68.67	75.17	83.67	86.50	88.42	89.67	**82.0**

Table 8. Keyword recognition accuracies of BSE processed **dev/test** sets, obtained using *noisy* data and the ASR baseline with Gammatone front-end and the *modified training*.

different training corpora: the *rever* AM derives from the purely reverberated signals while *noisy* AM exploits the corresponding set of signals with the additional superposition of the recorded background noise. Similarly to Track 1, the *bse-matched* models are derived by filtering the noisy training set with the corresponding BSE processing.

7.3.1. Baseline ASR

Tables 9 and 10 compare the baseline results (no processing) with WERs obtained when the BSE processing algorithms is applied, (for both *no-mem* and *with-mem* processing modalities)

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
baseline	86.3	82.8	76.1	71.4	63.0	55.9	**72.6**
no-mem	75.8	67.7	59.4	52.5	44.9	40.7	**56.8**
with-mem	64.0	57.0	50.4	45.8	39.7	36.3	**48.9**
baseline	88.0	83.2	78.1	71.9	65.2	55.9	**73.7**
no-mem	74.4	67.2	58.0	49.9	44.6	38.8	**55.5**
with-mem	64.2	56.6	50.6	44.6	40.5	35.7	**48.7**

Table 9. Word Error Rates on Track 2 CHiME **dev/test** sets with different processing configurations (MFCC front-end and *rever* AM).

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
baseline	73.2	67.4	59.9	55.7	49.1	44.3	**58.3**
no-mem	67.2	60.0	52.9	48.5	44.2	41.5	**52.4**
with-mem	57.5	51.0	46.9	43.3	39.8	38.2	**46.1**
baseline	70.4	63.1	58.4	51.1	45.3	41.7	**55.0**
no-mem	63.1	55.6	49.2	44.4	40.5	37.1	**48.3**
with-mem	55.0	49.3	43.9	40.7	37.7	36.3	**43.8**

Table 10. Word Error Rates on Track 2 CHiME **dev/test** sets with MFCC front-end, *noisy* AM and BSE processing.

As for Track 1, results shown in Tables 9-12 demonstrate the advantages of the BSE processing, especially when combined with noisy matched training. The introduction of the gammatone-based

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	60.6	51.9	44.7	41.1	36.0	33.5	**44.6**
with-mem	48.2	42.6	37.4	33.1	31.3	29.3	**37.0**
no-mem	56.4	48.3	41.1	35.5	32.7	29.6	**40.6**
with-mem	45.2	40.1	35.0	32.1	28.9	27.2	**34.8**

Table 11. Word Error Rates on Track 2 CHiME **dev/test** sets for *bse-matched* AM.

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	57.2	47.8	39.4	34.4	30.1	28.1	**39.5**
with-mem	43.3	37.5	32.3	28.8	26.0	25.4	**33.2**
no-mem	54.3	44.7	39.2	32.5	29.8	25.3	**37.6**
with-mem	**42.2**	**38.4**	**32.7**	**29.2**	**26.9**	**23.7**	**32.2**

Table 12. Word Error Rates on Track 2 CHiME **dev/test** sets with a different processing configuration (Gammatone front-end and *bse-matched* AM).

front-end provides an additional gain, confirming our past results [5]. It is worth noting that for this new front-end the corresponding AM is trained on the same signals used for the provided *rever* and *noise* AM but with a different procedure: indeed, a complete training step is directly applied to the reverberated or noisy signals (i.e. no re-estimation from a clean initial model).

7.3.2. Sparsity based Acoustic Model Compensation

Sparsity based acoustic model compensation with the two BSE variants, with and without memory, was tested when using the *rever* models. The setup differs slightly with respect to the setup explained in Section 6.1. These differences can be summarized as follows. First, the log-energy of the conventional MFCC front-end was changed to the 0^{th} cepstral coefficient and the filter gains where computed from amplitudes as indicated in (21). A new seed model was therefore trained for this new setup using the same WSJ0 recipe as the one provided in the challenge tool-set and the reverberated data. The only variation in the acoustic model was the use of word-internal rather than cross-word phonemes.

In addition to this, some additional changes apply to the setup but only when propagating uncertainty. First, the propagation through the final overlap and add step in Fig. 1 was ignored since it is computationally very expensive. For this purpose, the enhanced spectrum was directly fed to the feature extraction process and thus the number of frequency bins at the signal processing stage was reduced to 512. This led to a small mismatch between the non-propagated and the propagated features. The improvement attained by using dynamic compensation is in fact higher than the one reported here when compared to the same feature extraction configuration for non-propagated features.

Finally, it has to be taken into account that in the BSE, a different Wiener filter is applied to each channel by separate followed by a delay and sum. The random variable describing the uncertain spectrum is thus categorical and the approach explained in Section 7.3.2 is a simplification, although the same principle applies. Regarding dynamic compensation, conventional front-end uncertainty propagation was used [21]. It should be noted that modified imputation [22] did not yield any improvements.

Tables 13 and 14 display the results for the two BSE variants. As it can be observed, the use of sparsity based uncertainty compensation consistently improves BSE for all SNRs and for both BSE

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	70.9	61.5	53.3	45.0	40.6	34.7	**51.0**
no-mem+UD	67.3	57.5	49.1	42.4	38.5	32.7	**47.9**
with-mem	59.0	53.0	45.5	39.6	36.6	31.5	**44.2**
with-mem+UD	56.7	50.7	43.1	37.7	35.2	30.6	**42.3**

Table 13. Word Error Rates on task2 CHiME **test** set with second MFCC front-end, *rever* word-internal AM and BSE processing without and with sparsity based uncertainty decoding (+UD).

SNR	-6dB	-3dB	0 dB	3dB	6dB	9dB	**avg**
no-mem	71.5	62.6	52.9	46.6	40.1	35.6	**51.6**
no-mem+UD	69.2	60.4	50.3	45.1	39.0	34.4	**49.7**
with-mem	61.0	52.5	45.0	40.0	35.5	32.1	**44.4**
with-mem+UD	58.7	50.6	43.5	39.0	34.2	30.8	**42.8**

Table 14. Word Error Rates on CHiME **dev** set with second MFCC front-end, *rever* word-internal AM and BSE processing without and with sparsity based uncertainty decoding (+UD).

variants. Although not reported here, similar experiments using conventional MSE based uncertainty compensation produced little or no improvements in comparison. As explained in 7.3.2 this is mostly due to the fact that, unlike sparsity based uncertainty, MSE based uncertainty assumes no errors in the estimation of the spectral filtering parameters. Also, no improvements were attained in *bse-matched* conditions as the artifacts caused by residual noise were already learned by the model.

8. CONCLUDING REMARKS

In this article we revised and extended the multichannel Blind Source Extraction framework proposed in [5] in order to deal with both Track 1 and Track 2 of the 2nd CHiME challenge. A novel unsupervised spatial dictionary learning, combined with a backward-forward constrained on-line spatial filtering, allow an accurate enhancement of a localized (either static or moving) speech source in presence of real-world non-stationary noise and reverberation conditions. The capabilities of the BSE was demonstrated when used as pre-processing stage of a robust ASR system. It was shown that when BSE is combined with robust feature analysis and matched training, it produces a sensible improvement also for difficult medium vocabulary recognition tasks such as Track 2 of the 2nd CHiME Challenge. Furthermore, it was shown that when the training set cannot be matched with the BSE processing, the use of uncertainty decoding is able to sensibly improve the overall performance.

9. 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.

[3] 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.

[4] Z. Koldovsky, J. Malek, J. Nouza, and M. Balik, "Chime data separation based on target signal cancellation and noise masking," in *Proceedings of CHIME*, Florence, Italy, 2011.

[5] F. Nesta and M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments," *Computer Speech and Language*, 2012.

[6] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Proceedings of CHIME*, Florence, Italy, 2011.

[7] 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.

[8] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proceedings of ICASSP*, vol. 2, may 1996, pp. 733–736.

[9] J. Du and Q. Huo, "A feature compensation approach using high-order vector taylor series approximation of an explicit distortion model for noisy speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2285 –2293, nov. 2011.

[10] R. F. Astudillo, "Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, Technical University Berlin, 2010.

[11] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

[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] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal correlation," in *Proceedings LVA/ICA*, Mar 2012.

[14] M. Fakhry and F. Nesta, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pp. 1 –4, sept. 2012.

[15] F. Nesta and M. Fakhry, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," in *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, 1984.

[17] 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.

[18] K. Vertanen, "Baseline wsj acoustic models for htk and sphinx: Training recipes and recognition experiments," 2006. [Online]. Available: http://www.keithv.com/pub/baselinewsj/

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

[20] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da S. Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust asr in multi-source environments," *Computer Speech & Language*, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230812000575

[21] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with splice for noise robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 1, 2002, pp. I–57–I–60 vol.1.

[22] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.