A Multi-Stage, Multi-Channel Processing System for Overlapping Speech Separation in a Real Scenario

Rahil Mahdian Toroghi, Youssef Oualil, Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany
Email: {rahil.mahdian, youssef.oualil, dietrich.klakow}@lsv.uni-saarland.de
Web: www.lsv.uni-saarland.de

Abstract

This paper addresses the problem of overlapping speech separation in a noisy room using a microphone array. The presented approach proposes a multistage processing framework to separate the desired sources and reduce the corruptive effects of noise, reverberation and interference. More specifically, 1) a beamformer separates the sources based on their location diversities, 2) a postfilter maximizes the output SNRs, and 3) a novel filter is derived to suppress the coherent terms at each output with respect to its contrasting one. Finally, 4) the clean signal is estimated using a modified masking filter. Exploiting the fact that a desired signal remains coherent within time frames, the mask is smoothed between frames to preserve this coherency and reduce the musical noise. Experiments on AMI-Wall Street Journal corpus show a significant improvement in speech quality, SNR, Source to Reverberation Ratio, and naturalness of the proposed method, compared to some methods in Blind Source Separation.

1 Introduction

Separation of speech sources which are recorded in closed areas is an essential requirement for several applications, such as meeting recognition, automatic classnotes transcribing, and so on. Speech separation is a hard problem, but can be facilitated to some degree by the use of an array of microphones, especially when the geometry of the array is also known a priori. Multiple recordings of the speech data enables us to denoise or dereverberate the signals of interest without distortion, at least theoretically [1]. Utilizing the fact that speakers are located at different positions in the room, spatial filtering (beamforming) can be used to exploit this spatial information of the sources and extract higher quality source signals out of the corrupted input array data.

In the presence of overlapping speakers, the conditions of the separation problem in a room environment get far more difficult to handle [2].

In this paper, following the line of thought of our previous work, [3], we present a multi-step processing system that is able to cope with the three corrupting effects found in every noisy echoic environment, namely noise, reverberation, and interference. The contribution of this paper is three-fold: 1) The system structure that can be used in any echoic environment along with the results that justify it, 2) Derivation of the model for a filter that suppresses the coherent terms from the signals, and 3) A modification on the binary mask that enables us to account for the signal correlations over the neighboring frames, especially when the signal contained in these frames is due to a voiced phoneme.

The remaining part of the paper is organized as follows. Next section reviews the background theory of the beamforming and postfiltering. Subsection 2.2 presents the problem formulation and justifies the processes used in the proposed structure. Subsection 2.3 reviews masking. Section 3 presents the experiments and is followed by the results and comparisons with some methods in Blind Source Separation (BSS).

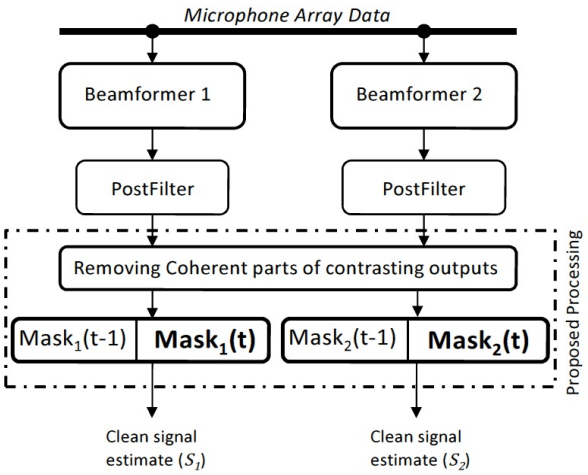

Figure 1: Block diagram of the multi-channel speech separation system of two sources in each frequency bin, used in this paper

2 Structure of the System

The overall system structure to separate two sources in a room is depicted in Fig. 1. This structure employs beamforming to extract the desired sources based on their unique geometrical position, a postfilter to increase the level of SNR, and two other stages: 1) A stage to suppress the portion of each output that is coherent with the contrasting output (that are emanated from the same sources), 2) A masking stage that accounts for source presence and temporal correlations in neighboring source frames. This structure can be utilized in realtime and only the masks calculated from the last frame need to be saved.

2.1 Background Theory

2.1.1 Beamforming

Beamforming (BF) aims at extracting the signal coming from the desired direction while suppressing the noise, reverberation, and interfering signals that are entering the array from other directions. The difference in the positions of the sources causes different Time Differences of Arrival (TDOA) with respect to the microphones in the array which is exploited in BF design. Let us consider a plane wave approaching the array aperture from a direction

$$\mathbf{a} = [\cos\theta\sin\phi \quad \sin\theta\sin\phi \quad \cos\phi]^T \qquad (1)$$

ISBN 978-3-8007-3640-9

with azimuth θ and elevation ϕ. Then, using the far field assumption, the delay which is introduced at the i-th microphone in position $\mathbf{m}_i$ $(i = 1 \dots L)$ in relation to the center of the array is $\tau_i = -\mathbf{a}^T \mathbf{m}_i / c$ where c denotes the speed of sound. Translating these delays to phase shifts in the frequency domain leads to the so-called *array manifold vector* $\mathbf{v}$, which is dependent on the sample rate and wave direction (through delays τ_i), as well as the angular frequency ω:

$$\mathbf{v}(\omega) = \begin{bmatrix} e^{-j\omega\tau_1} & \cdots & e^{-j\omega\tau_L} \end{bmatrix}^T \tag{2}$$

Now, denoting the frequency spectrum of the signals $x_i(t)$, $i = 1, \dots, L$, at the microphones by

$$\mathbf{X}(\omega) = \begin{bmatrix} X_1(\omega) & \cdots & X_L(\omega) \end{bmatrix}^T$$

the frequency spectrum $S(\omega)$ of the sound coming from the direction $\mathbf{a}$ can be extracted as the scalar product of $\mathbf{X}(\omega)$ with the weight vector $\mathbf{w}(\omega)$.

The weight vector $\mathbf{w}$ can be obtained through an optimization problem according to certain criteria (such as minimizing the beamformer output noise power) while, at the same time, maintaining the *distortionless constraint*, i.e. $\mathbf{w}^H \mathbf{v} = 1$. This leads to the following optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{\Sigma}_{nn} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{v} = 1 \tag{3}$$

where $\mathbf{\Sigma}_{nn}$ is the cross power spectral density (PSD) of the noise. Obtaining a reliable estimate of the cross PSD is a problem, that is typically overcome by making the homogeneous noise field assumption, for which $\mathbf{\Sigma}_{nn}$ can be written as [4]: $\mathbf{\Sigma}_{nn} = \mathbf{\Phi}_{nn} \mathbf{\Gamma}_{nn}$ where $\mathbf{\Phi}_{nn}$ denotes the noise power and $\mathbf{\Gamma}_{nn}$ is the noise coherence matrix. With this factorization, the MVDR solution devolves to:

$$\mathbf{w}_{\text{sdbf}} = \frac{P_n^{-1} \mathbf{\Gamma}_{nn}^{-1} \mathbf{v}}{P_n^{-1} \mathbf{v}^H \mathbf{\Gamma}_{nn}^{-1} \mathbf{v}} = \frac{\mathbf{\Gamma}_{nn}^{-1} \mathbf{v}}{\mathbf{v}^H \mathbf{\Gamma}_{nn}^{-1} \mathbf{v}} \tag{4}$$

By assuming a particular choice of $\mathbf{\Gamma}_{nn}$ for spherically isotropic noise field, which is optimal for reverberant environments [5], the result is called as Superdirective BF (SDBF) [4].

2.1.2 Postfiltering

The minimum mean square error solution to spatial filtering consists of an MVDR beamformer combined with a Wiener postfilter (PF) [6]. Following our work in [3], we extend the Wiener PF idea with the overestimation factor β, as follows:

$$H(\omega) = \frac{\mathbf{\Phi}_{ss}(\omega)}{\mathbf{\Phi}_{ss}(\omega) + \beta \mathbf{\Phi}_{nn}(\omega)} \tag{5}$$

where $\mathbf{\Phi}_{ss}(\omega)$ and $\mathbf{\Phi}_{nn}(\omega)$ denote the speech and noise power at the output of the array and can be estimated based on Zelinski method [7]. $\mathbf{\Phi}_{ss}(\omega)$ and $\mathbf{\Phi}_{nn}(\omega)$ in Zelinski estimation, are based on the incoherent noise assumption. But this may not be the case, in practice. Hence, we use a noise overestimation factor β in order to compensate for a possible systematic error. Our Automatic speech recognition experiments on MC-WSJ-AV corpus, [8], indicated that $\beta = 0.5$ gives the optimum result.

2.2 Proposed Coherent Removal Filters

The signals of the array microphones in Fourier domain can be writen as a linear combination of the atoms taken from a spatial basis dictionary, in which the base associated with the direct signal propagation is $\mathbf{v}_i, i = \{1,2\}$, and the rest of the bases which are concatenated in a matrix $\mathbf{\Lambda_{ri}}, i = \{1,2\}$ (column-wise), are associated with the reflections of the direct signal and other interferences. Formalization of this assumption for a two speaker scenario in a closed room while the microphone signals are corrupted with an additive ambient noise becomes, as:

$$\mathbf{X} = [\mathbf{v}_1 | \mathbf{\Lambda_{r1}}] \begin{bmatrix} S_1 \\ \mathbf{S_{1R}} \end{bmatrix} + [\mathbf{v}_2 | \mathbf{\Lambda_{r2}}] \begin{bmatrix} S_2 \\ \mathbf{S_{2R}} \end{bmatrix} + \mathbf{N} \tag{6}$$

Where $\mathbf{v}_i \in \mathbb{C}^M, (i = 1,2)$ are the array steering vectors toward the desired sources S_1 and S_2 with M as the number of microphones, $\mathbf{\Lambda_{ri}}, (i = 1,2)$ are the array steering matrices for all angles of the space but the desired sources, and $\mathbf{S_{iR}}, (i = 1,2)$ are the reverberated versions of the desired sources that contain the lagged version of the desired (current-time) sources with random lag times. $\mathbf{N}$ is multichannel ambient noise. All entities are transformed into short time frequency domain (STFT) and (ω, t) is dropped for simplicity.

Applying the weight vectors of the beamformers (corresponding to the sources) to (6), and assuming $\mathbf{w}_i^H \mathbf{v}_i = 1, \mathbf{w}_i^H \mathbf{v}_j \approx 0; i, j = \{1,2\}, i \neq j$, based on (3) and MVDR distortionless constraints, we get the following signals:

$$\begin{aligned} Z_1 &= \mathbf{w_1^H X} = S_1 + \mathbf{a}_1 \mathbf{S_{1R}} + \mathbf{a}_2 \mathbf{S_{2R}} + n_1 \\ Z_2 &= \mathbf{w_2^H X} = S_2 + \mathbf{b}_1 \acute{\mathbf{S}}_{\mathbf{1R}} + \mathbf{b}_2 \acute{\mathbf{S}}_{\mathbf{2R}} + n_2 \end{aligned} \tag{7}$$

where $\mathbf{a}_i, \mathbf{b}_i, i \in \{1,2\}$ are the gain vectors related to the reverberation terms of the desired sources (S_1, S_2) and interference parts ($\mathbf{S_{1R}}, \mathbf{S_{2R}}$, $\acute{\mathbf{S}}_{\mathbf{1R}}$, and $\acute{\mathbf{S}}_{\mathbf{2R}}$). Notice that, $\mathbf{S_{1R}}, \acute{\mathbf{S}}_{\mathbf{1R}}$ are not necessarily the same, since they are originated from S_1 but with different lags (randomly combined), and so does $\mathbf{S_{2R}}, \acute{\mathbf{S}}_{\mathbf{2R}}$. Consequently, there are subterms in Z_1 and Z_2 that are coherent and subterms that are incoherent. The coherent terms included in Z_1 and Z_2 outputs, make them dependent. n_1 and n_2 are the residual noise terms after the BFs. Since the noise terms are assumed independent of the signals and reverberation parts, we apply our modified postfiltering (PF) to the outputs of BFs (Z_1 and Z_2) to increase the SNR level, however, we assume that the structure of the equation (7) is preserved with a lower noise level.

Following the line of thought of our previous work [3], utilizing the mask directly after BF-PF stage could be erroneous. The reason comes from the deficiencies in beamforming which does not allow the interfering signals to be sufficiently suppressed. In addition, reverberation causes the signal spectral energy to smear in time and affect the mask estimation. Here we use our previously proposed logSigmoid mask and modify it to be applied as a coherency removal filter. The logSigmoid mask as in [3], is:

$$\widetilde{M}_{i,\sigma,\xi}(\omega, t) = \frac{1}{1 + \xi \left(\dfrac{|Z_j(\omega, t)|}{|Z_i(\omega, t)|} \right)^{\sigma}} \quad i \neq j \tag{8}$$

where ξ and σ are parameters to control the sharpness and scale matching of the mask, respectively. In our experiments with logSigmoid mask, we noticed that $\sigma = 2$ is

mostly the optimized value. Interestingly, this value corresponds to the power of the signals in spectral domain. Hence, we choose $\sigma = 2$, change the name of the mask to $G_z(\omega,t)$ filter, and approximate (8) with the binomial expansion with first two terms of the series that represent this function. Therefore, we have:

$$\begin{aligned}\widetilde{G}_{z_i,\xi}(\omega,t) &= 1-\xi\left(\frac{|Z_j(\omega,t)|}{|Z_i(\omega,t)|}\right)^2 \\ &= 1-\xi\left(\frac{P_{z_j}}{P_{z_i}}\right) \quad i\neq j \end{aligned} \tag{9}$$

where $P_{z_i}, P_{z_j}, (i,j) \in \{1,2\}$ denote the power spectrum of the contrasting signals. The coherence between two signals describes the strength of association between them, and is shown as:

$$\gamma_{z_i z_j} = (P_{z_i z_j})^2/(P_{z_i}P_{z_j}) \quad i\neq j \tag{10}$$

where $P_{z_i z_j}$ denotes the cross power spectral density of the two signals. logSigmoid mask (8) is a parameteric design that (roughly saying) improves the separation ability of the mask by choosing the parameters, so that the resulting signal in the output statistically resembles a clean speech signal (i.e., particularly Kurtosis maximization in subbands to follow the supergaussianity of the clean speech). These parameters we should learn in each frequency band, before employing the logSigmoid mask. Here, we only resort to one parameter, namely ξ, and we let it be proportional to the coherence of the signals Z_1 and Z_2 in power domain, since we intend to remove the coherency between Z_1 and Z_2 using these filters. Thus, in (9), we set $\xi = \lambda\gamma_{z_i z_j}$, and the filters are shown as:

$$G_{z_i}(\omega) = 1-\lambda\gamma_{z_i z_j}\frac{P_{z_j}}{P_{z_i}} \quad i\neq j \tag{11}$$

The value of the λ can also be viewed as the parameter that compensates for the approximation (expansion) error of (9). λ can be optimized for a measure that is related to the speech intelligibility [9], such as *maximum Kurtosis*. Notice that all the parameters introduced in this paper depend on frequency ω. The final filter equation to be applied to the outputs is:

$$G_{z_i}(\omega) = max\left\{(1-\lambda\gamma_{z_i z_j}\frac{P_{z_j}}{P_{z_i}}),0\right\} \quad i\neq j \tag{12}$$

2.3 Proposed Masking

After the coherent terms are excluded in previous filtering stage, there are still incoherent residual terms, from reverberation, interference and noise, that need to be suppressed. To contrast against these residual terms, we use masking. Using the fact that different speakers tend to excite different frequency bins at a time (also known as W-Disjoint Orthogonality, [10]) we use binary mask $M_i(\omega,t)$, with $i \in \{1,2\}$ for each output at every bin to extract the desired signal. Optimally, the value of $M_i(\omega,t)$ should be set to 1 if the TF bin (ω,t) belongs to the i-th speaker and should be set to 0 otherwise. However, since it is not known which TF bin belongs to which speaker, as in Maganti et al. [11], we use the absolute ratio of the contrasting outputs, as follows:

$$M_i(\omega,t) = \begin{cases}1, & |Z_i(\omega,t)| \geq |Z_j(\omega,t)| \quad \forall j\neq i \\ 0, & otherwise\end{cases} \tag{13}$$

We modify this mask to account for the correlation between adjacent speech frames. We know that if the signal is present in a frame, especially if the frame belongs to a voiced phoneme, there is a large correlation between the neighboring frames. However, attention to this correlation is missed in binary masking. Moreover, the speech signal related to the desired speaker remains coherent, during its activation period. We name the modified masking as smoothed Binary Mask (sBM) and our final stage of masking becomes, as follows:

$$\hat{M}_i(\omega,t) = \alpha\hat{M}_i(\omega,t-1) + (1-\alpha)M_i(\omega,t) \tag{14}$$

where α is the smoothing or correlation parameter in each frequency bin and can be optimized based on some measure being correlated with speech intelligibility, such as *maximum kurtosis*, as in Table 1. This table shows the efficiency of α, on the intelligiblity of the separated speech.

α	0.8	0.83	**0.85**	0.88	0.9	0.95
WER	51.13	41.62	**41**	41.52	42.10	48.15

Table 1: *Dependency of the WER (Word Error Rate %) to the mask smoothing factor α, when it is assumed constant.*

3 Experiments

Experiments are performed on AMI Wall-Street-Journal corpus [8]. This data set contains recordings of five pairs of speakers talking simultaneously to an array of 8 microphones planted symmetrically in a circle of 10cm radius. This is a challenging task for source separation given that the room is reverberant and includes significant background noise level. Position of the speakers have already been estimated as in [12]. This dataset has been used in the PASCAL Speech Separation Challenge II [13, 14]. The total number of utterances is 356 (or 178, respectively, if we consider the fact that two sentences are read at a time).

The result of the proposed method in Fig. 1 has been compared to some known methods in the field of Blind Source Separation (BSS) such as convolutive ICA (cICA) [15] of Ikeda et al., convolutive BSS (cBSS) method of L. Parra [16], as well as our previous method of logSigmoid masking (lgSigMsk) [3]. In our previous work, [3], we have already shown that SDBF outperforms other advanced bemaformers such as LCMV when the multistage structure, including BF combined with PF are followed by a masking stage. In addition, Table 2 shows that our new proposed method significantly outperforms the compared ones in case of Perceptual Speech Quality (PESQ), noise and reverberation enhancement measures shown by Segmental SNR, and Signal to Reverberation ratio, respectively. The cepstral and LPC based distances also show that the features of our method are closer to the natural speech. Evaluation measures are found online in **REVERB** Challenge workshop. Looking at Fig. 2 clearly shows that the spectrum of the proposed method is more enhanced than the other compared methods. Comparing the spectrograms with the clean one, we see that the remained interference is well removed in our methods, whereas in other BSS methods the interference components are still remained (a sample is marked in the figure). Noise is effectively removed however, it seems that there are parts of the clean signal that are also removed. This can be due to overestimation of the noise that has been removed in consecutive

Method	SegSNR	CD	LLD	SRMR	PESQ
BF/PF+cICA	-0.72	5.41	1.04	3.72	1.59
$\sim$ + cBSS	-0.31	5.31	0.96	5.56	1.53
$\sim$ + lgSigMsk	-0.30	5.96	1.11	7.17	1.45
Proposed	**2.05**	**4.89**	**0.95**	**8.30**	**1.96**

Table 2: *Comparison of our method with some known methods in BSS, applied to the outputs of BF/PF, based on the measures: Segmental-SNR (in dB), Cepstral Distance (CD), LPC based LogLikelihood Ratio Distance (LLD), Source to Reverberation Modulation Ratio (SRMR in dB) and Perceptual Evaluation Quality (*$1 \leq PESQ \leq 5$*).*

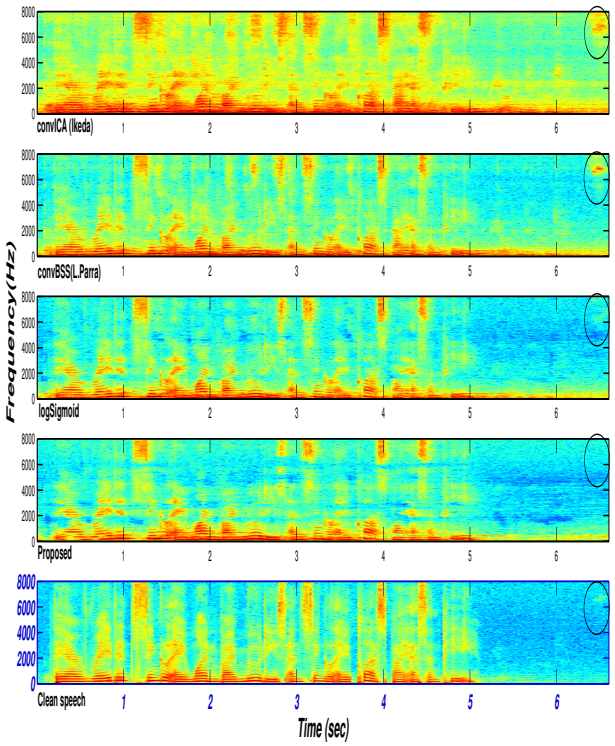

Figure 2: Spectrogram of a sample output from the compared methods

stages (postfilter and masking). Moreover, we are using a fixed smoothing value for the binary mask (α) than can make mistakes in frames that voiced phonemes are located between low energy frames. In these cases the mask is mostly dominated by the value of the low energy frames and the voiced information might be omitted. In general, noise, reverberation and interference are significantly removed. As a result, we see in Table 2 that in addition to the improvement in PESQ, the SNR and SRMR which are highly correlated with the intelligiblity are also highly improved. Moreover, the distances from the natural speech are also less than the compared BSS methods which again emphasize the higher intelligibility of the outcome of the proposed method.

4 Conclusion

We have investigated a speech separation system consisting of a beamforming, a postfiltering and two extra stages that perform as coherent removal filter and separation mask, respectively. We have derived the coherent removal filtering from our previously proposed logSigmoid mask by its expansion approximation. Moreover, we extended the binary mask based on the idea that, there is a high correlation between frames that contain speech. Totally, the proposed system structure showed its superior performance over some known techniques in BSS as well as our previously proposed logSigmoid masking system. Our future work, goes toward extending the idea of smoothed masking to be applied more sophisticatedly based on the features of the speakers.

References

[1] J. Benesty, S. Makino, and J. Chen in *Speech Enhancement*, pp. 27–29, Springer, 2005.

[2] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," *Proc. ICASSP*, vol. 5, pp. 497–500, Apr. 2003.

[3] R. M. Toroghi, F. Faubel, and D. Klakow, "Multichannel speech separation with soft time-frequency masking," *SAPA-SCALE conference*, Sept. 2012.

[4] M. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), pp. 19–38, Springer, 2001.

[5] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of the Acoustic Society of America*, vol. 27, pp. 1072–1077, Nov. 1955.

[6] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), pp. 39–62, Springer, 2001.

[7] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *Proc. ICASSP*, vol. 5, pp. 2578–2581, Apr. 1988.

[8] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," *ASRU*, pp. 357–362, Nov. 2005.

[9] G. Li and M. E. Lutman, "Sparseness and speech perception in noise," *Proc. SAPA*, pp. 7–11, Sept. 2006.

[10] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," *Proc. ICASSP*, vol. 1, pp. 529–532, May 2002.

[11] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2257–2269, Nov. 2007.

[12] Y. Oualil, F. Faubel, and D. Klakow, "A fast cumulative steered response power for multiple speaker detection and localization," *Proc. EUSIPCO*, Sept. 2013.

[13] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow, "To separate speech - a system for recognizing simultaneous speech," *Proc. MLMI*, pp. 283–294, June 2007.

[14] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," *Proc. MLMI*, pp. 295–305, June 2007.

[15] S. Ikeda, "A method of ica in time-frequency domain," in *in Proc. ICA*, pp. 365–371, 1999.

[16] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, pp. 320–327, May 2000.