Improving the Separation of Concurrent Speech through Residual Echo Suppression

Christian Siegwart, Friedrich Faubel[1]*, Dietrich Klakow*

Spoken Language Systems, Saarland University, D-66123 Saarbrücken, Germany
Email: {christian.siegwart,friedrich.faubel}@lsv.uni-saarland.de
Web: www.lsv.uni-saarland.de

Abstract

This paper investigates the use of acoustic echo cancellation components in a speech separation system. The basic system uses a classical beamformer architecture, which separates the speech from different speakers based on spatial diversity. In order to get a better suppression of concurrent speech, we add a residual echo suppression stage, which has originally been developed in the area of acoustic echo cancellation. The speech separation performance of the proposed system is evaluated by means of automatic speech recognition experiments. The results show a clear improvement over standard beamforming and postfiltering approaches, with a word error rate of 44.2% compared to 68.1% for a superdirective beamformer (SDB) and 59.8% for an SDB with Zelinksi postfilter.

1 Introduction

Sound capturing systems for hands-free speech recognition [1] typically employ a spatial filter which extracts the signal from the desired speaker while suppressing noise and reverberation (under the assumption that these come from other directions). This approach works reasonably well in practice. But the PASCAL Speech Separation Challenge (SSC-II) [2, 3] has demonstrated that spatial filtering alone is insufficient for separating concurrent speech (i.e. speech from two simultaneous speakers, whose mixture is received at the microphones). Hence, the authors of [2, 3] proposed to use specialized cross-talk cancellation postfilters which exploit the spectral sparsity of speech through binary time-frequency masking. In this work, we present an alternative to these approaches. It consists of using the residual echo suppression technique from [4] in order to (1) estimate the residual at the output of two beamformers – each one steered at one of the speakers – and to then (2) suppress the estimated residual with a Wiener filter. This approach can be interpreted as removing that part which both beamformer outputs have in common. Hence, it should be expected to improve the separation.

Figure 1 again gives an overview of the proposed approach. It consists of a spacial filtering stage, which exploits the fact that speakers tend to reside at different locations in the room. This allows us to separate their speech by directing one beamformer at each of the speakers. Following [2], we here use a superdirective design [5], which is an appropriate choice for reverberant environments [6] such as the office room in which the SSC-II has been recorded [7]. This spatial separation stage is followed by a residual echo suppression stage, which further improves the separation quality by suppressing the residual at the beamformer outputs.

[1]This author has been supported by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction (MMCI).

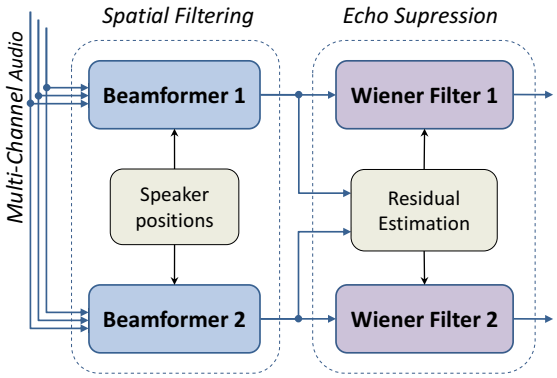

Figure 1: Proposed speech separation system with two beamformers and a residual echo suppression stage.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the beamforming techniques which are used in this work. This is followed by Section 3, which explains cross-talk cancellation based on time-frequency masking [2, 3, 8]. Section 4 describes the proposed approach of using residual echo suppression for this purpose. Section 5 finally presents the experimental results, including a discussion.

2 Beamforming

Beamforming is a way to influence the directivity of a microphone array. It is used to filter signals from a certain direction. Signals from other directions are attenuated. The filter characteristic of a microphone array is determined by both the array geometry, i.e. the positions $\mathbf{m}_1, \ldots, \mathbf{m}_M$ of the microphones, and the filter coefficients of the channels. For the following, let us denote the signal at the i-th microphone by $x_i(t)$ and let us further denote its Fourier coefficients by $X_i(\omega)$. Then beamforming (in the frequency domain) can be described as a multiplication of the Fourier coefficients $\mathbf{X}(\omega) = [X_1(\omega) \;\cdots\; X_M(\omega)]$ of the individual microphone signals with a weight vector $\mathbf{W}(\omega)$:

$$Y(\omega) = \underbrace{[W_1^*(\omega) \quad \cdots \quad W_M^*(\omega)]}_{\triangleq \mathbf{W}^H} \cdot \begin{bmatrix} X_1(\omega) \\ \vdots \\ X_M(\omega) \end{bmatrix}. \qquad (1)$$

In order to steer the array into a particular direction – denoted by the azimuth θ and elevation ϕ – we further define the array manifold vector

$$\mathbf{V}(\omega) = \begin{bmatrix} e^{-j\omega\tau_1} & \cdots & e^{-j\omega\tau_M} \end{bmatrix}^T \qquad (2)$$

where the time delays $\tau_i = -(\mathbf{a}^T \mathbf{m}_i / c)$ are calculated under the far-field assumption with $\mathbf{a}$ denoting the directional cosine $\mathbf{a} \triangleq [\cos\theta \sin\phi \;\; \sin\theta \sin\phi \;\; \cos\phi]^T$ and with c denoting the speed of sound.

2.1 MVDR Beamforming

In order to design an optimized beamformer, we may wish to minimize the noise power at the output of the beamformer. For that, let us express $\mathbf{X}(\omega)$ as a superposition of the desired signal $S(\omega)\mathbf{V}(\omega)$ with multi-channel noise $\mathbf{N}(\omega)$: $\mathbf{X}(\omega) = S(\omega)\mathbf{V}(\omega) + \mathbf{N}(\omega)$. Then writing the output of the beamformer as

$$Y(\omega) = \underbrace{\mathbf{W}^H(\omega)S(\omega)\mathbf{V}}_{Y_S(\omega)} + \underbrace{\mathbf{W}^H(\omega)\mathbf{N}(\omega)}_{Y_N(\omega)}, \qquad (3)$$

it becomes clear that the minimization of the noise output power is equivalent to minimizing the variance of $Y_N(\omega)$:

$$E\left\{|Y_n|^2\right\} = \mathbf{W}^H(\omega)\mathbf{\Sigma}_{nn}(\omega)\mathbf{W}(\omega). \qquad (4)$$

where $\Sigma_{nn}(\omega)$ is the power spectral density (PSD) matrix of the noise. As we do not wish the desired signal $S(\omega)$ to be either attenuated or amplified, we further use the distortionless constraint $\mathbf{W}^H(\omega)\mathbf{V}(\omega) = 1$. This leads to the minimum variance distortionless response (MVDR) beamformer whose weight vector is the solution of the following optimization problem:

$$\min_{\mathbf{W}} \ \mathbf{W}^H\Sigma_{nn}\mathbf{W} \quad \text{subject to} \quad \mathbf{W}^H\mathbf{V} = 1 \qquad (5)$$

where the dependency on ω has been dropped for the sake of readability. This optimization problem can now be solved with a Lagrange multiplier; and it yields:

$$\mathbf{W}_{mvdr}(\omega) = \frac{\Sigma_{nn}^{-1}(\omega)\mathbf{V}^H(\omega)}{\mathbf{V}^H(\omega)\Sigma_{nn}^{-1}(\omega)\mathbf{V}(\omega)}. \qquad (6)$$

2.2 Homogenous Noise Fields

In the particular case of a homogeneous noise field (i.e. the noise power is the same in each point) [5], the noise PSD matrix $\Sigma_{nn}(\omega)$ can be written $\Sigma_{nn}(\omega) = \Phi_{nn}(\omega)\Gamma_{nn}(\omega)$ where $\Phi_{nn}(\omega)$ denotes the noise power and where $\Gamma_{nn}(\omega)$ is the noise coherence matrix whose coefficients

$$(\Gamma_{nn})_{i,j}(\omega) = \frac{\Phi_{n_in_j}(\omega)}{\sqrt{\Phi_{n_in_i}(\omega)\Phi_{n_jn_j}(\omega)}} \qquad (7)$$

denote the coherence of the noise at the i-th and j-th microphone channels. Now plugging $\Sigma_{nn}(\omega) = \Phi_{nn}(\omega)\Gamma_{nn}(\omega)$ into the MVDR solution from above, we find that the weight vector $\mathbf{W}(\omega)$ becomes

$$\mathbf{W}_{\text{mvdr}}(\omega) = \frac{\Gamma_{nn}^{-1}(\omega)\mathbf{V}(\omega)}{\mathbf{V}^H(\omega)\Gamma_{nn}^{-1}(\omega)\mathbf{V}(\omega)} \qquad (8)$$

as the $\Phi_{nn}^{-1}(\omega)$ in the nominator and denominator cancel. Hence, in a homogenous noise field the beamformer is dependent on the noise coherence only.

2.3 Delay and Sum Beamforming

Choosing the identity matrix as a noise coherence matrix, i.e. $\Gamma_{nn} = I$, yields the well-known delay-and-sum (DSB) beamformer. Its weight vector is calculated according to:

$$\mathbf{W}_{\text{dsb}}(\omega) = \tfrac{1}{M}\mathbf{V}(\omega); \qquad (9)$$

and it is optimal for spatially uncorrelated (incoherent) noise such as white noise at the sensors.

2.4 Superdirective Beamforming

A popular alternative to the incoherent noise field assumption of the DSB is to use a spherically isotropic (or diffuse) noise field, whose coherence matrix $\Gamma_{nn}|_{\text{diff}}$ is calculated according to [5, 6]:

$$(\Gamma_{nn}|_{\text{diff}})_{i,j}(\omega) = \text{sinc}\left(\omega\frac{\|m_i - m_j\|}{c}\right). \qquad (10)$$

This is optimal for reverberant environments in which noise approaches the microphones from all directions [6]. Plugging $\Gamma_{nn}|_{\text{diff}}$ back into (8) yields the superdirective beamformer (SDB) [5]. As the name hints, it optimizes the directivity index (DI),

$$\text{DI}(\omega) = 10\log_{10}\left(\frac{|W^H(\omega)V(\omega)|^2}{W^H(\omega)\Gamma_{nn}|_{diff}(\omega)W(\omega)}\right), \qquad (11)$$

which describes the capability of an array to suppress a diffuse noise field.

2.5 MMI Beamforming

In order to compare our results to other methods that have been reported in the literature, we here also introduce the Minimum Mutual Information MMI) beamformer, which Kumatani et al. [3, 9] proposed as a solution to the speech separation problem. It uses two beamformers whose weights are jointly optimized to minimize the mutual information at the beamformer outputs. To maintain the distortionless constraint, the authors of [9] used the *Generalized Sidelobe Canceller* (GSC) configuration

$$\mathbf{Y}_i = (\mathbf{W}_{q,i} - \mathbf{B}_i\mathbf{W}_{a,i})^H\mathbf{X} \qquad (12)$$

with $\mathbf{W}_{q,i} = \mathbf{V}_i$ being the quiescent weight vector chosen for the i-th speaker (the dependency on ω has again been dropped for the sake of readability). The blocking matrix $\mathbf{B}_i$ by definition projects to the subspace which is orthogonal to $\mathbf{W}_{q,i}$. The active weight vectors $\mathbf{W}_{a,i}$ are optimized to individually minimize the mutual information

$$I(Y_1, Y_2) = \mathscr{E}\left\{\log\frac{p(Y_1,Y_2)}{p(Y_1)p(Y_2)}\right\} \qquad (13)$$

in each frequency bin. Due to the lack of an analytical solution, the $\mathbf{W}_{a,i}$ are refined in an iterative fashion, as explained in more detail in [9].

2.6 Zelinski Postfiltering

The MVDR beamformer in principle minimizes the noise power at the output of the array. But this minimization is restricted to linear solutions as it optimizes the scalar product with a weight vector. So, we might not find the best possible solution. In fact, it has been observed that the noise suppression can be further improved by nonlinear post-processing with a Zelinski postfilter [10, 11]. This approach was later shown to be the minimum mean square error (MMSE) solution to beamforming [12]:

$$\mathbf{w}_{\text{mmse}}(\omega) = \underbrace{\left(\frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{nn}(\omega)}\right)}_{H(\omega)}\mathbf{w}_{\text{mvdr}}(\omega). \qquad (14)$$

where $\mathbf{w}_{\text{mvdr}}(\omega)$ denotes the weight vector of an MVDR beamformer and where $H(\omega)$ is the frequency response of

a Wiener postfilter. The $\Phi_{ss}(\omega)$ and $\Phi_{nn}(\omega)$ denote the speech and noise power at the output of the array; and they may be estimated as follows [10, 11]:

$$\Phi_{ss} \approx \frac{2}{M(M-1)} \Re \left\{ \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} v_i^* \Phi_{x_i x_j} v_j \right\}, \qquad (15)$$

$$\Phi_{nn} \approx \frac{1}{M} \sum_{i=1}^{M} \Phi_{x_i x_i} - \Phi_{ss}. \qquad (16)$$

In these equations, $\Phi_{x_i x_j}$ and $\Phi_{x_i x_i}$ are the cross and power spectral densities of the microphone channels and v_i is the i-th coefficient of the array manifold vector (the dependency on ω was dropped for the sake of readability).

Although Zelinski postfiltering gives good results in practice, it also tends to overestimate the noise power. Hence, we here use a noise overestimation factor β which allows us to control the amount of speech distortion:

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \beta \Phi_{nn}(\omega)}. \qquad (17)$$

3 Binary Time/Frequency Masking

Unfortunately, the assumption of spatially uncorrelated noise (as it is used in the Zelinski postfilter) does not really hold in the presence of a strong directed interference such as a second speaker. Hence, Maganti et al. [8] proposed to replace the Zelinski postfilter by a specialized cross-talk cancellation postfilter. This filter uses the fact that different speakers tend to excite different frequency bands at a time and, consequently, extracts the clean speech spectrum $\hat{S}_i(\omega,t)$ of the i-th speaker at time t by applying a binary mask $M_i(\omega,t)$ to the beamformer output $Y_i(\omega,t)$:

$$\hat{S}_i(\omega,t) = M_i(\omega,t) \cdot Y_i(\omega,t), \quad i \in \{1,2\}. \qquad (18)$$

Optimally, the binary masks $M_i(\omega,t)$ would be set to 1 if the time-frequency unit (ω,t) belongs to the i-th speaker and it would set to 0 otherwise [13]. As, in practice, it is not known which time-frequency units are used by a speaker, Maganti et al. [8] estimated the masks based on the power ratio of the beamformer outputs:

$$\hat{M}_i(\omega,t) = \begin{cases} 1, & |Y_i(\omega,t)| \geq |Y_j(\omega,t)| \quad \forall j \\ 0, & otherwise \end{cases}. \qquad (19)$$

4 Residual Echo Suppression

As an alternative to binary masks [2, 3, 8], we here propose the use of residual echo suppression (RES) [4] as a speech separation postfilter. It is noteworthy that RES was originally developed for acoustic echo cancellation systems. It can be described as first estimating the residual based on the coherence function and then suppressing the estimated residual with a Wiener filter. For the first step, we need to estimate the cross-power spectral density (CSD) $\bar{\Phi}_{y_1 y_2}$ and the power spectral densities (PSD) $\bar{\Phi}_{y_1 y_1}$ and $\bar{\Phi}_{y_2 y_2}$. This is achieved here through Welsh averaging:

$$\begin{aligned} \bar{\Phi}_{y_1 y_2}(\omega) &= \alpha \cdot \bar{\Phi}_{y_1 y_2}(\omega) + (1-\alpha) \cdot \Phi_{y_1 y_2}(\omega) \\ \bar{\Phi}_{y_1 y_1}(\omega) &= \alpha \cdot \bar{\Phi}_{y_1 y_1}(\omega) + (1-\alpha) \cdot \Phi_{y_1 y_1}(\omega) \\ \bar{\Phi}_{y_2 y_2}(\omega) &= \alpha \cdot \bar{\Phi}_{y_2 y_2}(\omega) + (1-\alpha) \cdot \Phi_{y_2 y_2}(\omega) \end{aligned}$$

where $\Phi_{y_1 y_2}$, $\Phi_{y_1 y_1}$ and $\Phi_{y_2 y_2}$ denote the instantaneous CSD and PSD values and where α is a smoothing constant. After these exponentially decaying averages have been obtained, the coherence $\gamma_{y_1 y_2}$ between Y_1 and Y_2 can now be calculated according to:

$$\gamma_{y_1 y_2}(\omega) = \frac{\bar{\Phi}_{y_1 y_2}(\omega)}{\sqrt{\bar{\Phi}_{y_1 y_1}(\omega) \bar{\Phi}_{y_2 y_2}(\omega)}}. \qquad (20)$$

With this, we can now approximate the residual part of Y_j that is contained in Y_i as $\hat{R}_i(\omega) = \gamma_{y_1 y_2}(\omega) Y_j(\omega)$, $j \neq i$. The corresponding residual power $\hat{\Phi}_{r_i r_i}$ is obtained by taking the magnitude square[1] [4]:

$$\hat{\Phi}_{r_i r_i}(\omega) = \underbrace{\frac{|\bar{\Phi}_{y_1 y_2}(\omega)|^2}{\bar{\Phi}_{y_1 y_1}(\omega) \bar{\Phi}_{y_2 y_2}(\omega)}}_{=|\gamma_{y_1 y_2}(\omega)|^2} \bar{\Phi}_{y_j y_j}(\omega) \qquad (21)$$

In the second step, i.e. the construction of the Wiener filter, the clean speech power $\Phi_{s_i s_i}$ is estimated as $(\Phi_{y_i y_i} - \beta \hat{\Phi}_{r_i r_i})$ where β denotes a residual overestimation factor. This leads to the following Wiener filter transfer function:

$$H_i(\omega) = \frac{\max(\bar{\Phi}_{y_i y_i}(\omega) - \beta \cdot \Phi_{r_i r_i}(\omega), 0)}{\bar{\Phi}_{y_i y_i}(\omega)}, \qquad (22)$$

$i \in \{1,2\}$, which is to be multiplied to the corresponding beamformer output Y_i.

5 Experiments and Results

The performance of the proposed approach has been evaluated on the two speaker condition of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus[7]. This condition was used in the PASCAL Speech Separation Challenge II [2, 3] and it consists of two concurrent speakers which are simultaneously reading sentences from the Wall Street Journal. The total number of utterances is 356 (or 178, respectively, if we consider the fact that two sentences are read at a time [2]). The speech recognition system used in the experiments is identical to the one in [3], except that we use three passes only (instead of four): a first, unadapted pass; a second pass with unsupervised MLLR feature space adaptation; and a third pass with full MLLR adaptation. The estimated speaker positions are the same ones used in [3]. Hence, the results are absolutely comparable.

Table 5 shows the word error rates (WER) we obtained when separating the speech through beamforming only. This gives a direct comparison between delay-and-sum (DSB), superdirective (SDB) and minimum mutual information (MMI) beamforming [3], without the use of postfiltering. As expected, the MMI beamformer here performed best, with an absolute improvement of 20.2% compared to the DSB and 10.5% compared to the SDB.

	Used Beamformer		
	DSB	SDB	MMI
WER	78.87%	68.07%	57.58 %

Table 1: Word error rates for plain beamformers, without the use of a postfiltering techniques.

[1] Note that we have further replaced $|Y_j|^2$ by $\bar{\Phi}_{y_j y_j}$.

As a next step, we evaluated the combination of a superdirective beamformer with a Zelinski postfilter. The resulting word error rates are shown in Table 2, in dependency of the noise overestimation factor β, which controls the amount of speech distortion (see Section 2.6 for a more detailed explanation). Here, a value of $\beta = 0.5$ gave the best result, with a WER of 59.8% compared to 68.1% for a plain SDB. This result already comes quite close to an MMI beamformer, which achieves a WER of 57.0% when combined with a Zelinski postfilter.

	noise overestimation factor β				
	1.50	1.00	0.75	0.50	0.25
WER	64.7%	60.9%	62.3%	59.8%	60.9%

Table 2: Word error rate for an SDB with a Zelinski postfilter. The noise overestimation factor β is varied from a value of 1.5 to 0.25.

The results for the proposed residual echo suppression (RES) postfilter are finally shown in Table 3, again in combination with a superdirective beamformer. The best result is obtained with a residual overestimation factor β of 0.8, at a WER of 44.2%. This constitutes an absolute improvement of 24% over a plain SDB and still an improvement of 15% over an SDB with a Zelinski postfilter. At the same time, RES is computationally much more efficient than a Zelinski postfilter, as it does not require summing over $\mathcal{O}(M^2)$ channel combinations, see e.g. (15).

	residual overestimation factor β				
	1.0	0.8	0.6	0.4	0.2
WER	48.7%	44.2%	46.4%	47.0%	52.5%

Table 3: Word error rate for an SDB with an RES postfilter. The residual overestimation factor β is varied from a value of 1.0 to 0.2.

Table 4 again gives a comparison to other results that have been reported in the literature, namely the MMI beamformer from [3] as well as binary time-frequency masking (BM) [2, 3, 8, 14]. These results shows that the proposed RES postfilter still performs slightly better than the combination of a Zelsinki postfilter with binary masking (at least in the case of an SDB beamformer).

	Beamformer		
Postfilter	DSB	SDB	MMI
None	78.87%	68.07%	57.58%
Zelinski	69.07%	59.82%	56.98%
Zelinski + BM	51.03%	45.33%	49.99%
RES	N/A	44.20%	N/A

Table 4: Word error rates for the DSB, SDB and MMI beamformers with different postfiltering techniques.

6 Conclusions

We have shown how the residual echo suppression technique of an acoustic echo cancellation system can profitably be used for the speech separation task. The results show a significant improvement in word error rate, with an absolute improvement of 14% over a Zelinski postfilter and an improvement of 1% for a Zelinski postfilter with binary masking. The main strength of the proposed approach is its computational simplicity.

7 Acknowledgments

We would like to thank Kenichi Kumatani from Disney Research for kindly providing the MMI beamformed data of the speech separation challenge. Apart from him, we are also indebted to John McDonough who provided the speaker positions and the ASR framework, which has been used in the experiments.

References

[1] I. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons Inc, 2009.

[2] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, pp. 295–305, Springer-Verlag, June 2007.

[3] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Woelfel, and D. Klakow, "To separate speech: A system for recognizing simultaneous speech," in *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, pp. 283–294, Springer-Verlag, June 2007.

[4] G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1893–1896, IEEE, June 2002.

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

[6] 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.

[7] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357–362, IEEE, Dec. 2005.

[8] H. 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.

[9] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2527–2541, Nov. 2007.

[10] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proceedings of the 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2578–2581, IEEE, Apr. 1988.

[11] K. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, pp. 185–194, 1992.

[12] 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.

[13] S. Rickard and Z. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 529–532, IEEE, May 1993.

[14] R. Mahdian Toroghi, F. Faubel, and D. Klakow, "Multi-channel speech separation with soft time-frequency masking," in *SAPA-SCALE Conference*, Sept. 2012.