Multichannel spatial clustering for robust far-field automatic speech recognition in mismatched conditions

Michael I Mandel[1]*, Jon P Barker*[2]

[1]Brooklyn College, CUNY, Department of Computer and Information Science
[2]University of Sheffield, Department of Computer Science

mim@sci.brooklyn.cuny.edu, j.barker@dcs.shef.ac.uk

Abstract

Recent automatic speech recognition (ASR) results are quite good when the training data is matched to the test data, but much worse when they differ in some important regard, like the number and arrangement of microphones or differences in reverberation and noise conditions. This paper proposes an unsupervised spatial clustering approach to microphone array processing that can overcome such train-test mismatches. This approach, known as Model-based EM Source Separation and Localization (MESSL), clusters spectrogram points based on the relative differences in phase and level between pairs of microphones. Here it is used for the first time to drive minimum variance distortionless response (MVDR) beamforming in several ways. We compare it to a standard delay-and-sum beamformer on the CHiME-3 noisy test set (real recordings), using each system as a pre-processor for the same recognizer trained on the AMI meeting corpus. We find that the spatial clustering front end reduces word error rates by between 9.9 and 17.1% relative to the baseline.

Index Terms: Spatial clustering, beamforming, minimum-variance distortionless response, microphone arrays

1. Introduction

While ASR systems using deep neural networks (DNN) as acoustic models have recently provided remarkable improvements in recognition performance [1], their discriminative nature makes them prone to over-fitting the conditions used to train them. For example, in the recent REVERB challenge [2], far-field multichannel automatic speech recognition (ASR) systems consistently performed more accurately in the simulated conditions that matched their training than in the real recording conditions that did not. In order to address generalization, DNN acoustic models can be trained using multi-condition data [3], for example to combine multichannel time-domain signals in noise [4]. DNN enhancement systems can similarly be trained explicitly to generalize to new source positions for a fixed array [5] or new non-speech noises [6].

While explicit generalization to new spatial configurations of microphones, sources, and rooms is expensive to include in discriminative training procedures, it can be naturally factored out of the data through beamforming. Traditional beamforming assumes a known array geometry, which hinders generalization to new conditions, but unsupervised localization-based clustering avoids this assumption. Successful systems of this type have been introduced for two-microphone separation [7, 8, 9], and in larger ad-hoc microphone arrays for localization [10], calibration [11], and construction of time-frequency (TF) masks [12]. While [13] performs unsupervised beamforming in distributed arrays for enhancement, the current paper performs unsupervised beamforming on compact arrays for ASR and is evaluated in much noisier conditions. Similarly [14] describe a clustering-based MVDR beamformer, but apply it to enhancement, not ASR. The current approach can be considered a fully probabilistic version of [15], which performs similar operations, but requires setting several heuristic parameters. The current approach is also able to better integrate information across frequencies, allowing it to be used with a standard large-vocabulary automatic speech recognizer, as opposed to the more limited missing data digit recognizer of [15]. A similar approach using complex Gaussian mixture model spatial clustering to drive MVDR beamforming was developed concurrently with ours [16].

At the core of the proposed approach is the spatial clustering algorithm known as Model-based Expectation Maximization Source Separation and Localization (MESSL). MESSL performs mask-based separation by clustering TF points in the joint space of interaural phase differences (IPD) and interaural level differences (ILD) [9]. Originally formulated for binaural (two-channel) recordings, it was recently extended to the multichannel case [17] by modeling all pairs of channels with the binaural model. While [17] applied the mask estimated by multichannel MESSL as a post-filter to the output of a standard beamformer [18], the current paper explores several approaches to driving MVDR beamforming with MESSL's outputs.

2. Binaural MESSL

In the absence of additive noise or multiple talkers (which will be considered shortly) a single source, $s(t)$, arriving at two ears, $\ell(t)$ and $r(t)$, through channels $h_\ell(t)$ and $h_r(t)$, can be written as

$$\ell(t) = s(t) * h_\ell(t) \qquad r(t) = s(t) * h_r(t). \tag{1}$$

The ratio of the short-time Fourier transforms, $\mathcal{F}\{\cdot\}$, of both equations is the interaural spectrogram, which provides the observations upon which MESSL operates

$$\frac{\mathcal{F}\{\ell(t)\}}{\mathcal{F}\{r(t)\}} = \frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}. \tag{2}$$

We parameterize this interaural spectrogram with $\alpha(\omega,t)$, the ILD (measured in dB) and $\phi(\omega,t)$, the IPD. Assuming that the channel to each ear includes a sufficient amount of energy from the direct-path, these observations are modeled as

$$\approx 10^{a(\omega)/20} e^{-j\omega\tau(\omega)} N(\omega,t) \tag{3}$$

where $a(\omega)$ models the ILD (measured in dB) of the direct-path signal, $\tau(\omega)$ models the ITD of the direct-path and early

echoes of the signal, and $N(\omega, t)$ is a complex valued, time- and frequency-dependent noise of relatively small magnitude.

In order to avoid spatial aliasing issues, the ITD is modeled as a discrete hidden variable and IPD observations are compared to IPDs predicted from each ITD. Another discrete random variable models the source that generated a particular observation, leading to hidden binary random variables $z_{k\tau}(\omega, t)$, which are 1 when TF point ω, t comes from source k and delay τ. At each combination of source and delay, the IPD and ILD observations are modeled by a Gaussian distribution with diagonal covariance, making this a Gaussian mixture model. We use the expectation maximization (EM) algorithm to estimate the maximum likelihood model parameters, Θ, from the observations, while simultaneously estimating the expected values of the $z_{k\tau}(\omega, t)$ indicator variables. Marginalizing those expected values over delay, τ, yields the probability that each TF point comes from each source, probabilistically separating them. The total log likelihood that MESSL maximizes is

$$\mathcal{L}(\Theta) = \sum_{\omega t} \log p(\phi(\omega, t), \alpha(\omega, t) \mid \Theta) \tag{4}$$

$$= \sum_{\omega t} \log \sum_{k\tau} \Big[p(z_{k\tau}(\omega, t) \mid \Theta) \cdot p(\phi(\omega, t), \alpha(\omega, t) \mid z_{k\tau}(\omega, t), \Theta) \Big]. \tag{5}$$

The E step computes the posterior of $z_{k\tau}(\omega, t)$ and the M step uses these posteriors to update the parameter estimates [9].

We also utilize a recently described extension to MESSL called MESSL-MRF [19], which uses the MESSL posterior over sources as the local potential in a grid-shaped pairwise Markov random field (MRF). This MRF penalizes the assignment of neighboring TF points to different sources, smoothing the masks and reducing musical noise. Exact inference in the grid-shaped MRF is intractable, so we use loopy belief propagation (LBP) to approximate a solution. In [19], the sum-product variant of LBP was used inside the EM algorithm to smooth the probabilistic masks. Here we additionally utilize the max-product variant of LBP to estimate a globally consistent hard mask after the source parameters have been estimated.

3. Multichannel MESSL

Multichannel MESSL [17] models every pair of microphones separately using the binaural model described in Section 2. These models are coordinated through a global TF mask for each source. While model parameters are difficult to translate between microphone pairs without calibration information, TF masks are much more consistent across pairs. Coordination through masks should therefore be agnostic to the microphone array geometry and allow adaptation to new microphone configurations without calibration. Multichannel MESSL on M microphones maximizes the following total log likelihood

$$\mathcal{L}(\Theta) = \frac{2}{M} \sum_{i<j=1}^{M} \mathcal{L}(\Theta_{ij}) \tag{6}$$

$$= \frac{2}{M} \sum_{i<j=1}^{M} \sum_{\omega t} \log \sum_{k\tau} \Big[p(z_{k\tau}(\omega, t) \mid \Theta_{ij}) \cdot p(\phi_{ij}(\omega, t), \alpha_{ij}(\omega, t) \mid z_{k\tau}(\omega, t), \Theta_{ij}) \Big]. \tag{7}$$

Averaging over all pairs in this way assumes that all microphone pairs are independent of one another, whereas in reality only $M - 1$ are. This false assumption leads to an over-confidence in the likelihoods that is compensated by the $\frac{2}{M}$ term. Preliminary experiments showed that using all pairs of microphones with this correction factor led to higher quality separations than designating a single microphone as reference and using $M - 1$ pairs. The E and M steps for the model then proceed almost as in the two-channel algorithm. In the E step, the likelihood of the observations for each microphone pair is calculated under each source model. These likelihoods are then multiplied across microphone pairs and normalized across sources to give the final global posterior masks. In the M step, these global masks are used to re-estimate the parameters of each pair-wise model.

Initializing the multichannel model requires initializing the pair-wise models and coordinating the source models across microphone pairs. We explored two different initializations. The first used the PHAT-histogram approach [20] to find the dominant peaks in cross-correlations between pairs of channels followed by several iterations of binaural MESSL to estimate a mask for each source. These masks were then used to align the sources across microphone pairs. This approach has the advantage of being self-contained. The second initialization used a TF mask derived from level differences between a beamformer output and a reference microphone. In the experiments below, this was between the output of BeamformIt [21] and a microphone facing away from the talker. The mask is constructed from the 30% of points where the beamformer output is maximally louder than the reference. This initialization has the advantage of automatically aligning the source models across microphone pairs, but can fail if the baseline beamformer fails in localization or separation.

4. MESSL-driven MVDR beamforming

Beamforming is the process of combining signals recorded from a microphone array into a single estimate of a target signal. This estimate is typically driven by an optimality criterion. One popular criterion for fixed (non-adaptive) filter-and-sum beamforming is that of minimum variance distortionless response (MVDR) [22], which aims to minimize the output power of the beamformer while preserving signals from a target "look" direction. In the short-time Fourier transform domain, if a target signal $S(\omega, t)$ is observed by a microphone array with M microphones, then the ith microphone signal is

$$X_i(\omega, t) = H_i(\omega) S(\omega, t) + N_i(\omega, t). \tag{8}$$

These microphone signals can be stacked into complex vectors at each TF point,

$$\mathbf{X}(\omega, t) = \mathbf{H}(\omega) S(\omega, t) + \mathbf{N}(\omega, t). \tag{9}$$

Then the linear MVDR beamformer at frequency ω is

$$\mathbf{w}^*(\omega) = \min_{\mathbf{w}} E\left\{ |\mathbf{w}^H \mathbf{X}(\omega, t)|^2 \right\} \text{ s.t. } \mathbf{w}^H \mathbf{d}(\omega) = 1 \tag{10}$$

where $\mathbf{d}(\omega)$ is a steering vector, indicating the direction that should be preserved with unity gain. Recently, [23] showed that this can be solved without the use of an explicit steering vector by

$$\mathbf{w}^*(\omega) = \frac{\Phi_{NN}^{-1}(\omega) \Phi_{HH}(\omega) e_{\text{ref}}}{\text{tr}\left(\Phi_{NN}^{-1}(\omega) \Phi_{HH}(\omega)\right)} \tag{11}$$

$$= \frac{(\Phi_{NN}^{-1}(\omega) \Phi_{XX}(\omega) - I) e_{\text{ref}}}{\text{tr}\left(\Phi_{NN}^{-1}(\omega) \Phi_{XX}(\omega)\right) - M} \tag{12}$$

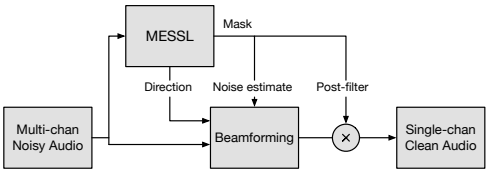

Figure 1: Three ways that MESSL outputs can drive minimum variance distortionless response beamforming: IPD parameters for look direction and masks for noise estimatation and/or non-linear post-filtering.

where M is the number of microphones, I is the $M \times M$ identity matrix, and e_{ref} is a vector of zeros with a single one selecting a reference microphone. This method allows the MVDR beamformer to be estimated without the use of an explicit steering vector, but still requires the estimation of $\Phi_{NN}(\omega)$, the noise spatial covariance, and either $\Phi_{XX}(\omega)$, the mixture spatial covariance, or $\Phi_{HH}(\omega)$, the target source spatial covariance. In our experiments, the denominator of these expressions was sometimes close to zero or even negative for a small set of frequencies, causing a large gain in the output at those frequencies and poor overall sound quality. We overcame this issue by enforcing that it be at least 1 (i.e., by replacing it with 1 if it was less than 1).

In the experiments discussed below, we explore the use of MESSL in driving MVDR beamforming in several ways, as illustrated in Figure 1. For estimating the noise spatial covariance matrices, $\Phi_{NN}(\omega)$, we compare using MESSL's masks to using the 400 to 800 ms of audio preceding the speech of each utterance, assumed to be noise-only, which is the approach taken by the baseline CHiME-3 system (see [24]). For estimating the steering vector, we compare an estimate of $\Phi_{HH}(\omega)$ based on MESSL's IPD parameters to a derivation from (12) using Φ_{XX}. For a non-linear post-filter, we compare the use of MESSL's masks to apply a gain to each TF point of the beamformed signal to the use of the unmodified output of the beamformer.

MESSL's mask for a single source, $M^{(k)}(\omega, t)$, can be used as a frequency-dependent voice activity detector to estimate $\Phi_{NN}(\omega)$ as

$$\Phi_{NN}(\omega) \approx \frac{\sum_{t=1}^{T} (1 - M(\omega, t))\, \mathbf{X}(\omega, t)\mathbf{X}^{H}(\omega, t)}{\sum_{t=1}^{T} (1 - M(\omega, t))}. \quad (13)$$

To avoid speech damage, we do not include observations at any frequency from frames when more than 40% of frequencies are predicted to be speech. To ensure that $\Phi_{NN}(\omega)$ is invertible, we always include the first M frames and the last $2M$ frames of an utterance in estimating it, numbers determined empirically.

Because MESSL models the IPD and ILD of each pair of microphones, it cannot reconstruct the magnitude of $\Phi_{HH}(\omega)$, but it can reconstruct its phase. Luckily, for compact arrays without a baffle between the microphones, the magnitude of the entries of this array should be close to unity. We thus compute the steering matrix entries

$$\Phi^{(k)}_{H_i H_j}(\omega) = \frac{\phi_{ijk\omega}}{|\phi_{ijk\omega}|} \text{ for } \phi_{ijk\omega} = E_\tau \left\{ e^{-j\omega\tau^{(k)}_{ij}(\omega)} \right\}. \quad (14)$$

This approach takes advantage of MESSL's frequency-varying IPD estimates and does not assume a pure delay between microphones, as standard steering vector formulations do.

Finally, MESSL's masks can be used as non-linear post-filters for the output of the MVDR beamformer. We found that using a maximum suppression of -9 dB $= 0.355$ at TF points where the mask was 0 reduced noise without causing noticeable artifacts. Preliminary ASR experiments also showed that -9 dB gave lower word error rates (WERs) than both -6 dB and -12 dB.

5. Experiments

We evaluate MESSL-MVDR in the context of ASR in mismatched far-field conditions. The baseline recognizer was trained on the AMI Meeting Corpus [25, 26], recorded on an 8-microphone circular array of diameter 10 cm. We use the multiple distant mic (MDM) condition processed by the BeamformIt tool [21], which performs delay-and-sum beamforming using time-varying source localization. We use the AMI Full-ASR partition training set (about 78 hours of speech) proposed in [27] and the corresponding Kaldi recipe with the provided automatic segmentations (version 1.6.1). The final acoustic model is a fully-connected deep neural network (DNN) that takes as input 40-dimensional log Mel filterbank features with first and second time derivatives [28]. This DNN is trained on labels aligned by a GMM-HMM model trained on MFCC features followed by linear discriminant analysis [29] and semi-tied covariance transforms [30], and discriminatively trained using the boosted maximum mutual information [31] criterion. The number of tied-states is roughly 4000.

This recognizer is tested on the live data portion of the CHiME-3 [24] dataset, which records speech input to a simulated tablet device in noisy environments. It uses a 6-microphone rectangular array around edge of the tablet, to which a talker whose mouth is 30–50 cm away reads sentences from the Wall Street Journal corpus (WSJ0). The recordings are made in four different noisy environments with an estimated signal to noise ratio of 0 dB. The acoustic model described above is used with the default CHiME-3 language model. Thus the training and test sets differ significantly in the number of microphones, array geometry, amount of reverberation, microphone array distance, amount and type of noise, speaking style, and vocabulary. MESSL is used only on the development and test sets, not in training.

5.1. Results

Table 1 shows the results of these experiments. The best system on the development set is shown in row 15 and used the MESSL noise estimate, MESSL post-filter, cross-correlation initialization for MESSL, and the mixture spatial covariance for (12). The columns of the table are ordered by the increase in word error rate on the development set caused by changing one of these parameters from this best setting. The rows of the table are ordered by the settings in each column. The parameter with the largest effect on this system is the noise estimate. Using the preceding 800 ms instead of the MESSL mask to estimate the noise results in a 2.75% absolute (14.0% relative) increase in development set word error rate (WER) (row 7 vs. 15). The second largest effect comes from the post-filter. Removing the post-filter results in a 2.4% absolute (12.2% relative) increase in WER (row 11 vs. 15). The last two parameters have smaller effects on the development set. Initializing MESSL from BeamformIt instead of using cross-correlations results in a 1.1% absolute (5.4% relative) increase in WER (row 13 vs. 15). Using the look direction from MESSL IPD instead of the mixture results in a 0.7% absolute (3.7% relative) increase in WER (row 14 vs. 15).

Baseline systems using BeamformIt are shown in the bottom two rows. The MESSL post-filter decreases WER for them

Table 1: Word error rates for recognizer trained on AMI data and tested on enhanced CHiME-3 real recordings. Noise estimates from the previous 400–800 ms (Prev) or MESSL mask. Post-filter not used (None) or MESSL mask. MESSL Initialized from BeamformIt or cross-correlation (Xcorr). Look direction from mixture (Mix) or from MESSL IPD. Bottom: BeamformIt baselines. Rows that are discussed in the text are colored.

	Noise est	Post-filt	MESSL Init	Look dir	WER (%) Dev	Test
1	Prev	None	—	Mix	29.2	48.6
2	Prev	None	BeamformIt	MESSL	26.1	39.7
3	Prev	None	Xcorr	MESSL	24.6	40.2
4	Prev	MESSL	BeamformIt	MESSL	22.8	35.4
5	Prev	MESSL	BeamformIt	Mix	23.2	39.5
6	Prev	MESSL	Xcorr	MESSL	20.8	35.6
7	Prev	MESSL	Xcorr	Mix	22.5	40.1
8	MESSL	None	BeamformIt	MESSL	26.7	43.9
9	MESSL	None	BeamformIt	Mix	22.4	32.4
10	MESSL	None	Xcorr	MESSL	23.1	41.3
11	MESSL	None	Xcorr	Mix	22.1	34.8
12	MESSL	MESSL	BeamformIt	MESSL	23.9	39.5
13	MESSL	MESSL	BeamformIt	Mix	20.8	*30.0*
14	MESSL	MESSL	Xcorr	MESSL	20.4	36.1
15	MESSL	MESSL	Xcorr	Mix	**19.7**	32.6
16	—	None	—	—	22.7	36.2
17	—	MESSL	—	—	20.6	31.0

by 2.1% absolute (9.3% relative) (row 16 vs. 17). Without a post-filter, two MESSL-MVDR systems (rows 9 and 11) achieve lower development and test WERs than the corresponding baseline (row 16), showing that MESSL can be used to effectively drive beamforming. With the post-filter, the same two systems (rows 13 and 15) perform comparably to the baseline (row 17). The MESSL-MVDR system that performs best on the development set (row 15) reduces WER on the test set by 3.6% absolute (9.9% relative) compared to the plain BeamformIt baseline. Consistent differences in performance have been seen on test and development sets for CHiME-3 [24], which might suggest looking directly for the best system on the test set, in which case, the best MESSL-MVDR system (row 13) reduces WER by 6.2% absolute (17.1% relative).

6. Conclusion

This paper has introduced the use of multichannel MESSL to drive minimum variance distortionless response beamforming. By clustering time-frequency points based on their spatial characteristics, this system is able to generalize to quite different recording conditions. Experiments recognizing data from CHiME-3 with a recognizer trained on AMI show that there are several ways of utilizing the outputs of MESSL with MVDR beamforming, including incorporating its mask into the noise spatial covariance estimate and as a post-filter and including its interaural phase differences into the target spatial covariance estimate. While using MESSL's outputs for spatial covariance estimates of the noise and for mask-based post-filtering improved ASR performance compared to a standard baseline, their use for estimating the target spatial covariance did not. In the future, combining speech models with spatial clustering could improve performance further. One possibility would be to generalize the speech models of [32] from binaural to multichannel recordings.

7. Acknowledgements

The work reported here was carried out during the 2015 Jelinek Memorial Summer Workshop on Speech and Language Technologies at the University of Washington, Seattle, and was supported by Johns Hopkins University via NSF Grant No IIS 1005411, and gifts from Google, Microsoft Research, Amazon, Mitsubishi Electric, and MERL. It is also based upon work supported by the NSF under Grant No IIS 1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Sig. Proc. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, New Paltz, NY, USA, Oct. 2013.

[3] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP*, vol. 12, 1987, pp. 705–708.

[4] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Proc. Interspeech*, 2015.

[5] Y. Jiang, D. Wang, and R. Liu, "Binaural deep neural network classification for reverberant speech segregation," in *Proc. Interspeech*, 2014, pp. 2400–2403.

[6] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE Tr. ASLP*, vol. 22, no. 4, pp. 826–835, Apr. 2014.

[7] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acous. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Tr. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Tr. ASLP*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[10] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *Proc. ICASSP*, vol. 2. IEEE, 2007, pp. II–761–II–764.

[11] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. ICASSP*. IEEE, May 2013, pp. 106–110.

[12] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Sig. Proc.*, vol. 87, pp. 1833–1847, 2007.

[13] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Tr. ASLP*, vol. 19, no. 4, pp. 661–676, May 2011.

[14] Y. Kawaguchi and M. Togami, "Soft masking based adaptation for time-frequency beamformers under reverberant and background noise environments," in *Proc. EUSIPCO*, 2010.

[15] M. Kühne, R. Togneri, and S. Nordholm, "Adaptive beamforming and soft missing data decoding for robust speech recognition in reverberant environments." in *Proc. Interspeech*, 2008, pp. 976–979.

[16] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline asr in noise," in *Proc. ICASSP*, Mar. 2016, pp. 5210–5214.

[17] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *Proc. ASRU*, 2015.

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Tr. ASLP*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[19] M. I. Mandel and N. Roman, "Enforcing consistency in spectral masks using markov random fields," in *Proc. EUSIPCO*, 2015.

[20] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Tr. Sys., Man, Cyb., C*, vol. 32, no. 4, pp. 474–484, Nov. 2002.

[21] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Tr. ASLP*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[22] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[23] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Tr. ASLP*, vol. 18, no. 2, pp. 260–276, 2010.

[24] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015.

[25] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus." *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[26] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Proc. ASRU*, Kyoto, 12 2007.

[27] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. ASRU*, Dec. 2013.

[28] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013, pp. 2345–2349.

[29] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1992, pp. 13–16.

[30] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Tr. ASLP*, vol. 7, no. 3, pp. 272–281, 1999.

[31] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008, pp. 4057–4060.

[32] R. Weiss, M. I. Mandel, and D. W. P. Ellis, "Combining localization cues and source model constraints for binaural source separation," *Sp. Comm.*, vol. 53, no. 5, pp. 606–621, 2011.