JOINT CONSTRAINED MAXIMUM LIKELIHOOD REGRESSION FOR OVERLAPPING SPEECH RECOGNITION

Kenichi Kumatani [‡], *Rita Singh* [†], *Friedrich Faubel* [§], *John McDonough* [†], *Youssef Oualil* [§]

[‡] Spansion Inc., Sunnyvale, CA, USA

[†] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

[§] Spoken Language Systems, Saarland University, Saarbrücken, Germany

ABSTRACT

Adaptation techniques for speech recognition are very effective in single-speaker scenarios. However, when distant microphones capture overlapping speech from multiple speakers, conventional speaker adaptation methods are less effective. The putative signal for any speaker contains interference from other speakers. Consequently, any adaptation technique adapts the model to the interfering speakers as well, which leads to degradation of recognition performance for the desired speaker. In this work, we develop a new feature-space adaptation method for overlapping speech. We first build a beamformer to enhance speech from each active speaker. After that, we compute speech feature vectors from the output of each beamformer. We then *jointly* transform the feature vectors from all speakers to maximize the likelihood of their respective acoustic models. Experiments run on the speech separation challenge data collected under the AMI project demonstrate the effectiveness of our adaptation method. An absolute word error rate (WER) reduction up to 14 % was achieved in the case of delay-and-sum beamforming. With *minimum mutual information* (MMI) beamforming, our adaptation method achieved a WER of 31.5 %. To the best of our knowledge, this is the lowest WER reported on this task.

Index Terms— feature-space adaptation, overlapping speech, speech separation, distant speech recognition, microphone array

1. INTRODUCTION

Overlapping speech is often observed in natural conversation [1, 2]. Shriberg et al. reported in [1] that 30 % to 50 % of all utterances contain interfering speech from another speaker in telephone conversations and meetings. Speech recognition performance is degraded when multiple talkers are speaking simultaneously [1–6].

Microphone array techniques can effectively separate overlapping speech with a little distortion [3–9]. However, the separation performance is usually insufficient for distant speech recognition due to microphone errors, steering errors, spatial aliasing and limited directivity at lower frequencies [3–5]. Residual speech of the interfering speakers could limit the gains from speaker adaptation methods such as vocal tract length normalization (VTLN) [10, §6], constrained maximum likelihood linear regression (CMLLR) [11–13] and MLLR [12].

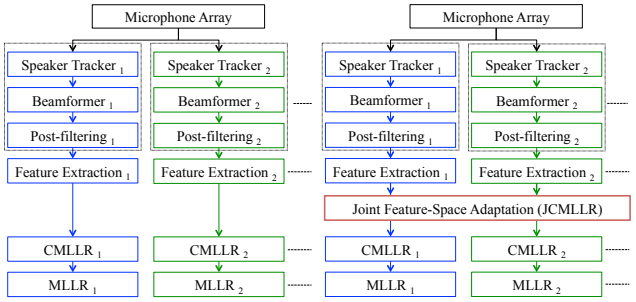
In this work, we develop a new microphone-array-based feature-space adaptation technique for recognition of overlapping speech. First, we build a beamformer to steer to each speaker, and compute a speech feature vector with the beamformer's output. The features for each speaker contain residual interference from the other speakers. We compute linear transforms that modify the features from each of the speakers to maximize the likelihood of recognition hypotheses. While conceptually similar to the well-known constrained MLLR (CMLLR) algorithm, the key difference is that we learn the transforms for all the speakers *jointly*. This has a dual effect – not only does it improve the statistical match of the features for any speaker to the recognizer, it simultaneously also *attenuates* the leakage from other speakers into the features.

Experiments on the Multi-Channel Wall Street Journal Audio Visual (MC-WSJ-AV) corpus, comprising recordings of overlapping speech uttered by human subjects captured with circular arrays [14], show that the proposed technique can greatly enhance speech recognition accuracy obtained even with the best beamforming algorithms. Moreover, the proposed method makes no assumptions about how the separated signals for the speakers were obtained, and can be used with *any* beamforming or signal separation algorithm, making it potentially a very useful tool for signal separation and distant speech recognition. Due to its conceptual similarity to the CMLLR algorithm, we dub our proposed algorithm *Joint* CMLLR, or JCMLLR.

The balance of this paper is organized as follows. In Section 2, we review prior work on speech separation for distant speech recognition. In Section 3, we formulate a problem; the linear transformation is described for a joint vector that consists of the feature vectors computed from the outputs of the beamformers. Section 4 describes an estimation algorithm for the joint linear transformation parameters based on the maximum likelihood (ML) criterion. In Section 5, we show recognition experiments on the MC-WSJ-AV corpus; overlapping speech uttered by real humans were captured with real circular arrays [14]. Finally, we conclude this work and describe possible future work in Section 6.

2. RELATION TO PRIOR WORK

Figure 1 shows a contrast between conventional speech recognition systems and our proposed system for overlapping speech. As illustrated in Figure 1 a), the conventional speech recognition system for overlapping speech typically consists of a microphone array processing module including a speaker tracker, beamformer and post-filter, a feature extraction module, and feature-space and model-space adaptation components. Given a position estimated by the speaker tracker, the beamformer is built to emphasize the sound wave coming from the direction of interest. The beamformed signal may be further enhanced by post-filtering. The front-end of the speech recognizer then computes the feature vector from the enhanced speech signals. The speech feature vector is adapted to a target speaker by an affine transformation by the CMLLR method [11–13]. In addition, MLLR [15] is performed to adapt Gaussian pa-



a) Conventional system b) Our system

Fig. 1. Conventional and proposed procedures for recognition of overlapping speech.

rameters of the HMM to the speaker. In our prior work [3, 8, 16], we found that the combination of CMLLR and MLLR significantly improves the recognition of overlapped speech.

In the case that overlapping speech is captured with distant microphones, the estimates of the speaker-dependent parameters for any speaker will be contaminated by speech from the other speakers due to the imperfect separation of signals from the target speaker. To mitigate this, we consider a pre-processing step where the concatenation of feature vectors for the individual speakers is decomposed into statistically independent components based on the maximum likelihood (ML) criterion. We refer to this pre-processing step as joint CMLLR (JCMLLR). On top of JCMLLR, we apply a cascade of conventional CMLLR and MLLR.

As it will be clear in Section 4, the formulation for JCMLLR is similar to estimation of an unmixing matrix in independent component analysis (ICA) [17, §9]. The main difference between our algorithm and typical ICA methods are:

- the use of an HMM [10, §7.1.1] for the probablity model, and
- estimation of the linear transformation in the feature domain; enabling us to use the HMMs in the speech recognizer.

We do not explicitly separate overlapping speech in the linear frequency domain such as in ML-based beamforming methods [18–20]. Instead, we assume that the observed feature vector can be approximately expressed as a linear combination of feature vectors of the individual speakers.

It is also worth mentioning that since we directly work on feature vectors derived from the speech, we do not need to address the permutation and scaling ambiguity problems encountered in the frequency-domain blind source separation (BSS) [21].

3. LINEAR TRANSFORM FOR SPEECH SEPARATION

We assume that $N_{\rm s}$ sound sources are detected by a speaker tracking system described in [22]. We enhance the signals coming from each of the directions of interest by beamforming [23]. Let us denote each speech feature vector computed from the signal associated with a sound source n as $\mathbf{o}_n$, where $n = 1, \cdots, N_{\rm s}$. Notice that $\mathbf{o}_n$ could also contain interference from the remaining sources due to imperfect separation performance by the beamformer.

Since each feature vector $\mathbf{o}_n$ combines information from all sources (with the dominant source being n), conversely, the features for any individual source can in turn be viewed as a function of all feature vectors $\mathbf{o}_n$, $n = 1, \cdots, N_{\rm s}$. We model this relationship by the following affine transformation

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_{N_{\rm s}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1N_{\rm s}} \\ \mathbf{W}_{21} & \cdots & \mathbf{W}_{2N_{\rm s}} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{N_{\rm s}1} & \cdots & \mathbf{W}_{N_{\rm s}N_{\rm s}} \end{bmatrix} \begin{bmatrix} \mathbf{o}_1 \\ \mathbf{o}_2 \\ \vdots \\ \mathbf{o}_{N_{\rm s}} \end{bmatrix} + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_{N_{\rm s}} \end{bmatrix}, \quad (1)$$

where $\mathbf{z}_m$ represents the estimated value of the feature vector isolated for the m-th sound source. $\mathbf{W}_{mn}$ is the transformation matrix for extracting the m-th sound source from the n-th observation vector and $\mathbf{d}_m$ is the linear shift for the m-th sound source.

For the sake of simplification, we rewrite (1) with concatenated matrices and vectors as

$$\mathbf{z} = \mathbf{W}\mathbf{o} + \mathbf{d}, \quad (2)$$

where

$$\mathbf{z} = [\mathbf{z}_1^T, \mathbf{z}_2^T, \cdots, \mathbf{z}_{N_{\rm s}}^T]^T, \quad (3)$$
$$\mathbf{o} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \cdots, \mathbf{o}_{N_{\rm s}}^T]^T, \quad (4)$$
$$\mathbf{d} = [\mathbf{d}_1^T, \mathbf{d}_2^T, \cdots, \mathbf{d}_{N_{\rm s}}^T]^T, \quad (5)$$

and $\mathbf{W}$ corresponds to the concatenated versions of the linear transformation matrices in (1), respectively.

Note that Equation 2 is very similar to the affine transform employed conventional CMLLR [11, 12] to adapt features for recognition:

$$\mathbf{y}_n = \mathbf{A}_n\mathbf{o}_n + \mathbf{b}_n. \quad (6)$$

Conventional CMLLR can also be employed with $\mathbf{z}_m$ values obtained from Equation 2:

$$\mathbf{y}_m = \mathbf{A}_m\mathbf{z}_m + \mathbf{b}_m. \quad (7)$$

During the estimation of $\mathbf{W}$ and $\mathbf{d}$, the CMLLR parameters, $\mathbf{A}_m$ and $\mathbf{b}_m$, are held fixed.

4. JOINT CMLLR FOR SPEECH SEPARATION

4.1. Distribution of transformed variables

Let us denote a probability density function (PDF) with model parameters $\mathcal{M}$ for the transformed feature vector $\mathbf{y}_m$ as $p_m(\mathbf{y}_m; \mathcal{M})$. Using the well-known relationship between the PDFs of variables related by an affine transform [17, §2.6], the log of the PDF of the original extended vector $\mathbf{o}$ can be written as

$$\log p(\mathbf{o}; \mathcal{M}) = \log|\mathbf{W}| + \sum_{m=1}^{N_{\rm s}} \log|\mathbf{A}_m| + \sum_{m=1}^{N_{\rm s}} \log p_m(\mathbf{y}_m; \mathcal{M}). \quad (8)$$

Notice that feature vectors of different speakers are assumed to be statistically independent with each other in (8).

Now let us assume that we observe $N_{\rm T}$ samples of $\mathbf{o}$, denoted by $\mathbf{o}(1), \mathbf{o}(2) \cdots, \mathbf{o}(N_{\rm T})$. Based on (7), we obtain the transformed features for the speaker m at a frame t as:

$$\mathbf{y}_m(t) = \mathbf{A}_m\mathbf{z}_m(t) + \mathbf{b}_m. \quad (9)$$

It is now straightforward to compute the log-likelihood based on (8):

$$\mathcal{L}(\mathbf{W},\mathbf{d}) = N_{\mathrm{T}} \log|\mathbf{W}| + N_{\mathrm{T}} \sum_{m=1}^{N_{\mathrm{s}}} \log|\mathbf{A}_m| + \sum_{m=1}^{N_{\mathrm{s}}} \sum_{t=1}^{N_{\mathrm{T}}} \log p_m\left(\mathbf{y}_m(t);\mathcal{M}\right). \quad (10)$$

To obtain a maximum-likelihood estimate of $\mathbf{W}$ and $\mathbf{d}$ so as to achieve speech feature separation as well as feature-space adaptation, we must maximize (10).

4.2. Cost function with HMM

We use an HMM with Gaussian mixture models (GMM) for state output distributions [10, §8] for computing the log-likelihood (10). Specifically, we use the HMM corresponding to the sentence (actual or recognized in a previous pass) for the recording. Transformation parameters are estimated using the expectation maximization (EM) algorithm. In the E-step of the EM algorithm, we compute the *a posteriori* probability of occupying the Gaussian component i_m for source m at a frame t, $\gamma_{i_m}(t)$, using the forward-backward algorithm. The linear transformation parameters are then updated in the M-step so as to maximize the following auxiliary function:

$$\mathcal{L}(\mathbf{W},\mathbf{d}) = \kappa + N_{\mathrm{T}} \log|\mathbf{W}| + N_{\mathrm{T}} \sum_{m=1}^{N_{\mathrm{s}}} \log|\mathbf{A}_m| - \frac{1}{2} \sum_{m=1}^{N_{\mathrm{s}}} \sum_{t=1}^{N_{\mathrm{T}}} \sum_{i_m=1}^{N_{\mathrm{w}}} \gamma_{i_m}(t) \left[N_{\mathrm{d}} \log 2\pi + \log|\mathbf{\Sigma}_{i_m}| + \left(\mathbf{y}_m(t) - \boldsymbol{\mu}_{i_m}\right)^T \mathbf{\Sigma}_{i_m}^{-1} \left(\mathbf{y}_m(t) - \boldsymbol{\mu}_{i_m}\right) \right], \quad (11)$$

where κ is a constant that depends on the state transition probabilities, N_{d} is the dimensionality of each feature vector $\mathbf{y}_m$, and $\boldsymbol{\mu}_{i_m}$ and $\mathbf{\Sigma}_{i_m}$ are the mean vector and covariance matrix for a Gaussian component i_m.

4.3. Estimation Algorithm

The natural gradient algorithm [17, §3.2] [24] is perhaps one of the most popular numerical optimization solvers in the field of ICA. Here, we derive necessary equations to implement the natural gradient algorithm. Upon taking the partial derivatives of (11) with respect to $\mathbf{W}$ and $\mathbf{d}$, we obtain the following:

$$\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{d})}{\partial \mathbf{W}} = N_{\mathrm{T}}[\mathbf{W}^T]^{-1} - \sum_{t=1}^{N_{\mathrm{T}}} \sum_{i_m=1}^{N_{\mathrm{w}}} \gamma_{i_m}(t) \mathbf{g}_{i_m}(t,\mathbf{W},\mathbf{d}) \left[\mathbf{x}(t)\right]^T, \quad (12)$$

$$\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{d})}{\partial \mathbf{d}} = -\sum_{t=1}^{N_{\mathrm{T}}} \sum_{i_m=1}^{N_{\mathrm{w}}} \gamma_{i_m}(t) \mathbf{g}_{i_m}(t,\mathbf{W},\mathbf{d}), \quad (13)$$

where

$$\mathbf{g}_{i_m}(t,\mathbf{W},\mathbf{d}) = \begin{bmatrix} \mathbf{\Sigma}_{i_m}^{-1}\left(\mathbf{y}_1(t) - \boldsymbol{\mu}_{i_m}\right) \\ \vdots \\ \mathbf{\Sigma}_{i_m}^{-1}\left(\mathbf{y}_j(t) - \boldsymbol{\mu}_{i_m}\right) \\ \vdots \\ \mathbf{\Sigma}_{i_m}^{-1}\left(\mathbf{y}_{N_{\mathrm{s}}}(t) - \boldsymbol{\mu}_{i_m}\right) \end{bmatrix}.$$

1: initialize the JCMLLR parameters: $\mathbf{W} \leftarrow \mathbf{I}$ and $\mathbf{d} \leftarrow \mathbf{0}$
2: **repeat**
3: **for** each beamformer's output **do**
4: estimate the VTLN warp factor as in [10, §6]
5: **end for**
6: **repeat**
7: update the linear shift $\mathbf{d}$ with (16)
8: **until** the expectation of (11) converges
9: **repeat**
10: update the joint transformation matrix $\mathbf{W}$ with (15)
11: **until** the expectation of (11) converges
12: **for** each speaker m **do**
13: update the CMLLR parameters with $\mathbf{z}_m$ as in [12]
14: **end for**
15: **for** each speaker m **do**
16: update the MLLR parameters with $\mathbf{y}_m$ as in [12]
17: **end for**
18: **until** the expectation of (11) converges

Fig. 2. Batch algorithm for the entire adaptation process

Eq. (13) indicates the steepest direction in the Euclidean orthogonal coordinate system. If the parameter space has a Riemannian metric structure, the steepest direction is the natural gradient. The direction of the natural gradient is obtained by post-multiplying (13) by $\mathbf{W}^T\mathbf{W}$ as

$$\mathbf{\Delta W} = \left[N_{\mathrm{T}}\mathbf{I} - \sum_{t=1}^{N_{\mathrm{T}}} \sum_{i_m=1}^{N_{\mathrm{w}}} \gamma_{i_m}(t) \mathbf{g}_{i_m}(t,\mathbf{W},\mathbf{d}) \left[\mathbf{x}(t)\right]^T \mathbf{W}^T \right] \mathbf{W} \quad (14)$$

It is clear from (13) and (14) that computation of the natural gradient does not require the inversion of $\mathbf{W}^T$ in every step, which leads to computational savings.

The transformation matrix $\mathbf{W}$ and linear shift $\mathbf{d}$ can be now updated as

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha_1 \mathcal{E}\left\{\mathbf{\Delta W}\right\}, \quad (15)$$

$$\mathbf{d} \leftarrow \mathbf{d} + \alpha_2 \mathcal{E}\left\{\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{d})}{\partial \mathbf{d}}\right\}, \quad (16)$$

where α_1 and α_2 are step parameters and $\mathcal{E}\{\cdot\}$ is the averaging operator over the samples. The linear shift is first estimated until the outputs of the auxiliary function (11) converge. Given the estimate of the linear shift, the JCMLLR matrix is updated.

Table 2 summarizes the batch algorithm for the entire adaptation process. We perform every adaptation step including JCMLLR in an unsupervised manner.

5. EXPERIMENTS

All the experiments reported here were conducted on the MC-WSJ-AV data collected under the Augmented Multi-party Interaction (AMI) project. The data set contains recordings of five pairs of speakers where each pair of speakers reads approximately 30 sentences taken from the 5000-word vocabulary of the Wall Street Journal (WSJ) task. The data from two simultaneously active speakers were recorded with two circular, eight-channel microphone arrays. The diameter of each array was 20 cm, and the sampling rate of the recordings was 16 kHz. As a reference, the close-talking data were also collected with a head-set microphone. The room size was 6.5 m $\times$ 4.9 m $\times$ 3.25 m and the reverberation time T_{60}

was approximately 380 milliseconds. In addition to being reverberant, the meeting room data collected include background noise from computers and the building ventilation. Some recordings also contain audible noise from outside the meeting room, such as that generated by passing cars and speakers in an adjacent rooms; see Lincoln et al. [14] for the details of the data collection apparatus. There are a total of 43.9 minutes of speech in the development set.

Prior to beamforming, we estimated the speaker's position with the Orion source tracking system [22]. Based on the average speaker position estimated for each utterance, we built the *minimum mutual information* (MMI) beamformer with the *generalized sidelobe canceler* (GSC) structure [25]. Subject to the distortionless constraint for the look direction, utterance-dependent active weight vectors were estimated based on the MMI criterion under the Gaussian probability density function. The MMI beamformer is capable of suppressing interfering sources without the signal cancellation effect encountered in conventional minimum variance distortionless response (MVDR) beamforming. Zelinski post-filtering [26] was further performed on the beamformed data to remove residual noise.

The feature extraction method used here was based on cepstral features estimated with a warped *minimum variance distortionless response* (MVDR) spectral envelope [27]. The MVDR models spectral peaks more accurately than spectral valleys, which leads to improved robustness in the presence of noise. Our speech feature analysis involved extracting 20 cepstral coefficients and global *cepstral mean subtraction* (CMS) with variance normalization. The frequency axis was warped so as to normalize the variations in vocal tract lengths by the VTLN method [10, §6]. After that, the warped cepstral coefficients were concatenated over 15 consecutive frames. The dimension of the concatenated vector was then reduced to 42 with *linear discriminant analysis* (LDA) [28]. Following the global CMS again in the LDA domain, we performed the global semi-tied covariance (STC) transformation [15], also known as the *maximum likelihood linear transformation* (MLLT).

The training data used for the experiments reported here was taken from the ICSI, NIST and CMU meeting corpora as well as the Transenglish Database (TED) corpus. The total amount of training data is approximately 100 hours. In addition to these corpora, approximately 12 hours of speech from the WSJCAM0 corpus [29] was used for HMM training in order to provide coverage of the British accents for the speakers in the MC-WSJ-AV development set. Acoustic models estimated with two different HMM training schemes were used for several decoding passes: conventional maximum likelihood (ML) HMM training [10, §8.1], and speaker-adapted training under the ML criterion (ML-SAT) [10, §8.1.3]. Our baseline system was fully continuous with 3,500 codebooks and a total of 180,656 Gaussian components. The full trigram language model for the 5,000 word WSJ task was used for decoding.

In addition to recognition with the unadapted models, we performed three adapted passes on the waveforms processed with each of the beamforming algorithms. Each adapted pass of decoding used a different acoustic model or speaker adaptation scheme. For each adapted pass, the adaptation parameters were estimated using the word lattices generated during the prior pass, as in [30]. A summary of the adapted decoding passes follows:

1. Estimate the VTLN, JCMLLR and conventional CMLLR parameters, then decode with the acoustic model.
2. Estimate MLLR parameters for each speaker on top of the adaptation method in the prior pass, then redecode with the acoustic model.
3. Estimate VTLN, JCMLLR, CMLLR, MLLR parameters, then redecode with the ML-SAT model.

Adapted pass		1st	2nd	3rd
Microphone array method	Joint adaptation			
D&S beamforming	without JCMLLR	73.6	59.2	56.8
	with JCMLLR	59.6	47.8	45.9
LCMV beamforming	without JCMLLR	58.6	47.4	45.8
	with JCMLLR	51.7	39.8	39.7
MMI beamforming	without JCMLLR	47.9	35.2	34.0
	with JCMLLR	45.5	33.2	31.5
Head-set microphone	without JCMLLR	26.5	23.4	23.0

Fig. 3. WERs with/without the JCMLLR method

Table 3 shows word error rates (WER) obtained with beamforming algorithms in each adapted pass. In the experiments, we used a delay-and-sum beamformer [31], a linearly-constrained minimum variance (LCMV) beamformer [31, §6.7] and the MMI beamformer [25]. For the LCMV and MMI beamformers, we place a null on the direction of the interfering source. Table 3 also shows the WERs obtained with the conventional adaptation methods and JCMLLR for each beamforming method. As a reference, the WERs for the close talking microphone (CTM) data are provided in Table 3. It is clear from Table 3 that JCMLLR can improve recognition performance for every beamforming method. We can observe from Table 3 that the poorer the speech separation performance of beamforming is, the more significant the improvement of JCMLLR becomes. These results imply that JCMLLR is also capable of unmixing overlapping speech although it does not directly separate a mixture of speech in the linear frequency domain. It is also clear from Table 3 that JCMLLR can consistently provide better recognition performance in the higher adapted passes where stronger speaker adaptation methods are used.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have described a new feature-space adaptation method for recognition of overlapping speech. The new method jointly estimates the constrained maximum likelihood regression (CMLLR) parameters for the feature vectors from multiple beamformers so as to obtain statistically independent components based on the ML criterion. We have confirmed the effectiveness of our method through a set of speech recognition experiments on real array data. We have also demonstrated that the cascade of joint CMLLR (JCMLLR) and conventional speaker adaptation methods further improves recognition performance.

We plan to mathematically investigate how the unmixing problem in the frequency domain can be approximated as a joint linear transformation in the feature domain. We also plan to evaluate different optimization algorithms for estimation of the JCMLLR parameters. Combinations of JCMLLR and other feature-space transformation methods such as discriminative features [32] are also possible. Furthermore, the framework of JCMLLR can be extended to model-space adaptation. It is also interesting to develop incremental update methods for this speaker adaptation scheme. These developments could be future work.

7. REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 1359–1362.

[2] Özgür Çetin and E. Shriberg, "Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap," in *Proc. ICASSP*, Toulouse, France, 2006.

[3] J.W. 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," in *Proc. MLMI*, Brno, Czech Republic, 2007.

[4] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proc. MLMI*, Brno, Czech Republic, 2007, pp. 295–305.

[5] R.M. Toroghi, F. Faubel, and D. Klakow, "Multi-channel speech separation with soft time-frequency masking," in *Proc. SAPA-SCALE Conference*, Portland, Oregon, USA, 2012.

[6] C. Siegwart, F. Faubel, and D. Klakow, "Improving the separation of concurrent speech through residual echo suppression," in *Proc. ITG Conference on Speech Communication*, Braunschweig, Germany, 2012, pp. 1–4.

[7] T. Virtanen, Rita Singh, and Bhiksha Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*, John Wiley & Sons, West Sussex, UK, 2012.

[8] K. Kumatani, J.W. McDonough, and Bhiksha Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[9] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, Chichester, UK, 2009.

[10] M. Wölfel and J.W. McDonough, *Distant Speech Recognition*, Wiley, London, 2009.

[11] V. Digalakis, D. Rtischev, and L. Neumeyer, "Fast speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 357–366, Sep. 1995.

[12] M. J. F. Gales, "Maximum likelihood linear transformations for HMM–based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[13] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech and Language*, vol. 26, no. 1, pp. 35–51, 2012.

[14] 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," in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 357–362.

[15] M.J.F. Gales, "Semi–tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[16] K. Kumatani, J.W. McDonough, D. Klakow, P. N. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," *IEEE Trans. on Audio, Speech and Language Processing*, August 2008.

[17] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," *Wiley Inter Science*, 2001.

[18] M. Reyes-Gomez, Bhiksha Raj, and D.P.W. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. ICASSP*, Hong Kong, April, 2003, pp. 664–667.

[19] M.L. Seltzer, Bhiksha Raj, and R.M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[20] B. Rauch, K. Kumatani, F. Faubel, J.W. McDonough, and D. Klakow, "On hidden markov model maximum negentropy beamforming," in *Proc. IWAENC*, Seattle, WA, USA, Sep. 2008.

[21] P. Comon and C Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press. Elsevier, 2010.

[22] T. Gehrig, U. Klee, J.W. McDonough, S. Ikbal, M. Wölfel, and C. Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, Pittsburgh, PA, USA, Sep. 2006.

[23] K. Kumatani, J.W. McDonough, S. Schact, D. Klakow, P. N. Garner, and W. Li, "Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming," in *Proc. ICASSP*, Las Vegas, NV, USA, April 2008.

[24] Shunichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

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

[26] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[27] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[28] R. Häb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1992, vol. 1, pp. 13–16.

[29] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Tech. Rep. CUED/F-INFENG/TR.192, Cambridge University Engineering Department (CUED) Speech Group, Sep. 1994.

[30] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, Salt Lake City, Utah, USA, 2001.

[31] H. L. Van Trees, *Optimum Array Processing*, Wiley–Interscience, New York, 2002.

[32] G. Saon and Jen-Tzung Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.