MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Effectiveness of Discriminative Training and Feature Transformation for Reverberated and Noisy Speech

Tachioka, Y.; Watanabe, S.; Hershey, J.R.

TR2013-020 May 2013

Abstract

Automatic speech recognition in the presence of non-stationary interference and reverberation remains a challenging problem. The 2nd Annual Speech Separation and Recognition Challenge introduces a new and difficult task with time-varying reverberation and non-stationary interference including natural background speech, home noises, or music. This paper establishes baselines using state-of-the-art ASR techniques such as discriminative training and various feature transformation on the middle-vocabulary sub-task of this challenge. In addition, we propose an augmented discriminative feature transformation that introduces arbitrary features to a discriminative feature transformation. We present experimental results showing that discriminative training of model parameters and feature transforms is highly effective for this task, and that the augmented feature transformation provides some preliminary benefits. The training code will be released as an advanced ASR baseline.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

EFFECTIVENESS OF DISCRIMINATIVE TRAINING AND FEATURE TRANSFORMATION FOR REVERBERATED AND NOISY SPEECH

Yuuki Tachioka

Mitsubishi Electric
Information Technology R&D Center
5-1-1, Ofuna, Kamakura, Kanagawa, Japan

Shinji Watanabe, John R. Hershey

Mitsubishi Electric Research Laboratories
201, Broadway, Cambridge, MA, US

ABSTRACT

Automatic speech recognition in the presence of non-stationary interference and reverberation remains a challenging problem. The 2nd 'CHiME' Speech Separation and Recognition Challenge introduces a new and difficult task with time-varying reverberation and non-stationary interference including natural background speech, home noises, or music. This paper establishes baselines using state-of-the-art ASR techniques such as discriminative training and various feature transformation on the middle-vocabulary sub-task of this challenge. In addition, we propose an augmented discriminative feature transformation that introduces arbitrary features to a discriminative feature transformation. We present experimental results showing that discriminative training of model parameters and feature transforms is highly effective for this task, and that the augmented feature transformation provides some preliminary benefits. The training code will be released as an advanced ASR baseline.

Index Terms— Discriminative training, Feature transformation, Augmented discriminative feature transformation, CHiME challenge, Kaldi

1. INTRODUCTION

Recent advances in Automatic Speech Recognition (ASR) [1], have greatly improved the accuracy of speech recognition systems. Over the past ten years model training techniques have migrated from Maximum Likelihood (ML) estimation to discriminative training [2, 3, 4, 5, 6]. In addition, various types of feature transformations have been proposed and showed effectiveness [7, 8, 9, 10, 11, 12]. Although it is well known that the state-of-the-art ASR techniques are very effective in relatively clean speech conditions, we need further investigation of their effectiveness in challenging conditions such as environmental reverberation and noise. Both reverberation and noise can degrade recognition performance, and this has been a limiting factor in the expansion of speech recognition scenarios beyond close-talking microphones. It is well known that the acoustic scores become less reliable in noisy and reverberant conditions since acoustic feature patterns are corrupted. However, discriminative training generally optimizes the parameters to maximize the difference of scores between correct and incorrect word/phoneme sequences to reduce the confusability.

This paper aims to test these techniques in a challenging noisy speech recognition task. In particular, we focus on discriminative training and feature transformations for this problem. This paper also deals with several feature transformation approaches, which convert original features to new features based on linear transformations (Linear Discriminant Analysis (LDA) [7], Maximum Likelihood Linear Transformation (MLLT) [8, 9], Speaker Adaptive Training (SAT) [10], and discriminative non-linear feature transformation [11, 13, 14, 15, 16]).

In AURORA2 experiments, combination of LDA and MLLT improve the recognition accuracy [17]. LDA uses long context by context expansion (e.g., contiguous 9 frames) to exploit dynamic features, which reduces the influence of non-stationary noises. MLLT finds a linear transformation of features to reduce state-conditional feature correlations. SAT and Maximum Likelihood Linear Regression (MLLR) improve the recognition accuracy by adapting to unknown and changing noise conditions.

Discriminative non-linear feature transformations can provide yet further gains in performance, because the transformation is optimized to reduce the error rate in the context of the decoder (e.g., [18]). Some of the popular non-linear transforms provide an approximately piece-wise linear transform by the inclusion of "region-based" features based on Gaussian posterior probabilities. We propose to extend this basic approach by augmenting the set of region-based features to include additional non-linear features that may be relevant in noisy conditions.

There are three objectives in this paper. The first is to validate the effectiveness of the discriminative training and feature transformation for reverberated and noisy speech to answer how these techniques improve the recognition accuracy. We evaluate the performance improvement by these techniques using 2nd CHiME challenge Track 2, which is designed to evaluate the "word error rate" under reverberated and non-stationary noisy environments [19] and matches our interests. The 2nd is to build the CHiME challenge's baseline using public tools. We use a Kaldi toolkit [20] as an advanced ASR back end of CHiME baseline to a HTK [21] based ML baseline attached to CHiME. Participants in the CHiME challenge are not necessarily experts on the speech recognition, hence to make the baseline including various techniques helps their research. This baseline will be distributed for participants. The third is to experiment with alternative features in the discriminative feature transformation, which we call augmented discriminative feature transformation.

2. DISCRIMINATIVE TRAINING AND FEATURE TRANSFORMATION

2.1. Discriminative training

Discriminative training is a supervised learning principle that minimizes error in modeling labels and recognition results. Although there are several training methods available [2], this paper focuses primarily on the Maximum Mutual Information (MMI). In MMI, the

objective function is given as

$$\mathcal{F}_{\mathrm{MMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s)}, \quad (1)$$

where R is the number of training utterances and $\{\mathbf{x}_t\}_r$ is the r^{th} utterance features sequence. The acoustic model parameters λ are optimized by the extended Baum-Welch. r is the index of the training utterance. $\mathcal{H}_{s_r}$ and $\mathcal{H}_s$ are the HMM sequences of a correct label s_r and a recognition result s, respectively. p_λ is the likelihood of an acoustic model, κ is the acoustic scale, and p_L is the likelihood of a language model.

Utterances that contain many errors need to be considered intensively and evaluating phoneme accuracies (i.e., evaluating margins [6]) improves performance. In the boosted MMI (bMMI) [22], the standard MMI objective function is modified to include a term that "boosts" the effect of hypotheses with low phoneme accuracy:

$$\mathcal{F}_{\mathrm{bMMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s) e^{-bA(s,s_r)}}, \quad (2)$$

where $A(s, s_r)$ is the phoneme accuracy of s for a reference s_r, and $b\ (> 0)$ controls the strength of its effect. In this paper, we compare the performance of MMI and bMMI to that of ML.

2.2. Discriminative feature transformation

In addition to discriminative training, feature transformation based on the discriminative training criterion can be used [11]. This method estimates a matrix $\mathbf{M}$ that projects from high-dimensional non-linear features to low-dimensional transformed features, as shown in Eq. (3):

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{M}\mathbf{h}_t, \quad (3)$$

where $\mathbf{x}_t$, are the K-dimensional original features, $\mathbf{h}_t$ are L-dimensional nonlinear features, and $\mathbf{y}_t$ are the transformed features. Matrix $\mathbf{M}$'s dimension is $K \times L$. In this study, we validate the effectiveness of feature-space MMI (f-MMI) and its extension, feature-space boosted MMI (f-bMMI). In both of these, the features are constructed in the same way, but the objective function for training is different. After substituting $\mathbf{y}$ of Eq. (3) for $\mathbf{x}$ into Eqs. (1) and (2), we obtain the objective function for f-MMI:

$$\mathcal{F}_{\text{f-MMI}}(\mathbf{M}) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{y}_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{y}_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s)}. \quad (4)$$

Differentiating the objective function $\mathcal{F}$ by $\mathbf{M}$ as

$$\frac{\partial \mathcal{F}}{\partial \mathbf{M}} = \begin{bmatrix} \frac{\partial \mathcal{F}}{\mathbf{y}_1} & \cdots & \frac{\partial \mathcal{F}}{\mathbf{y}_{T_f}} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \cdots & \mathbf{h}_{T_f} \end{bmatrix}^{\mathrm{T}}, \quad (5)$$

where T denotes the transpose and T_f is the total number of frames of training data. The f-bMMI objective function is similarly constructed. The optimized matrix $\mathbf{M}$ is obtained by gradient descent. To form the features, N components of the Gaussian Mixture Models (GMM) are obtained by clustering the Gaussians in the initial triphone acoustic models into N components and re-estimating their parameters. The non-linear features $\mathbf{h}_t$ [13] are calculated as

$$\mathbf{h}_{t,n} = \left[p_{t,n}\left(\frac{x_{t,1} - \mu_{n,1}}{\sigma_{n,1}}\right), \cdots, p_{t,n}\left(\frac{x_{t,K} - \mu_{n,K}}{\sigma_{n,K}}\right), \alpha p_{t,n} \right]^{\mathrm{T}}, \quad (6)$$

where $\mu_{n,i}$ and $\sigma_{n,i}$ are the mean and variance in dimension i of the n^{th} Gaussian component. α is the scaling factor. $p_{t,n}$ are Gaussian component posteriors computed for each frame, approximated such that all but the Q-best posteriors are set to zero. This approximation is done in order to reduce computational cost by ensuring that $\mathbf{h}_t$ is sparse.

2.3. Augmented discriminative feature transformation

In f-MMI, the high-dimensional sparse features $\mathbf{h}_t$ in Eq. (3) represent the likelihood and posterior information. These features are similar to those of MFCC. It is most effective to use different types of features for noisy speech recognition, such as the tandem approach [23]. We propose a method that obtains new transformed features $\mathbf{y}'_t$ by adding features $\mathbf{h}'_t$ to $\mathbf{h}_t$ as

$$\mathbf{y}'_t = \mathbf{x}_t + \left[\mathbf{M}\mathbf{M}'\right] \begin{bmatrix} \mathbf{h}_t \\ \mathbf{h}'_t \end{bmatrix} \quad (7)$$

$$= \mathbf{x}_t + \mathbf{M}\mathbf{h}_t + \mathbf{M}'\mathbf{h}'_t. \quad (8)$$

Thus, the auxiliary feature is an additional bias term in the transformation. The objective function is given as

$$\mathcal{F}_{\text{af-MMI}}\left(\left[\mathbf{M}\mathbf{M}'\right]\right) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{y}'_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{y}'_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s)}. \quad (9)$$

The concatenated matrices $\mathbf{M}$ and $\mathbf{M}'$ are optimized through the above-described process. An advantage of this approach is that various features can be taken into account, such as the sparse vectors of dictionary learning [24], the posteriors of likelihood ratio test-based voice activity detection [25] or binary mask vectors [26]. By integrating features related to speech enhancement into the discriminative feature transformation framework, we hope to overcome some of the short-comings of feature-based noise compensation methods.

3. EXPERIMENTAL SETUP

We validated the effectiveness of discriminative training and feature transformation for reverberated and noisy speech on the 2^{nd} CHiME challenge. Additionally, our proposed "augmented discriminative feature transformation" was validated. CHiME consists of a small vocabulary task (Track 1) and a medium vocabulary task (Track 2). This paper focused on Track 2, whose utterances were taken from Wall Street Journal database (WSJ0). For natural speech recognition, medium vocabulary tasks are needed in addition to conventional small vocabulary (command-like) tasks. The training data set (si_tr_s) contains 7138 utterances from 83 speakers (si84), the evaluation data set (si_et_05) contains 330 utterances from 12 speakers (Nov'92), and the development set (si_dt_05) contains 409 utterances from 10 speakers. Acoustic models were trained using si_tr_s and tuned using si_dt_05. The language model size was 5 k (basic). These data simulate two types of realistic environments. There are two types of data: "reverberated," data made by convolving clean speech with impulse responses recorded at 2 m distant microphones, and "noisy" data made by adding noises to "reverberated" at SNR $= \{-6, -3, 0, 3, 6, 9\}$dB. Noises are non-stationary such as other speakers' utterances, home noises, or music. In this study the "isolated" noisy data was used, in which the noisy signal is about the same length as the speech, as opposed to the "embedded" noisy data, in which the noisy signal is longer than the speech. The "isolated" condition is similar to having good voice activity detection. Although the database provides two channels of data, in this paper, we only used the left channel data.

We describe the settings of acoustic feature and feature transformation. The baseline acoustic features are MFCC and PLP (1-13 order MFCCs (PLPs) + Δ + $\Delta\Delta$). It is well known that under certain assumptions linear discriminative analysis (LDA) transforms the features so that the classes are well separated. After concatenating the first 13 static MFCCs in nine contiguous frames, a total of 117 dimensional features are compressed into 40 dimensions by an LDA performed using classes identified by tri-phone HMM state alignments (2,500 states). Because the acoustic features are high dimensional, it is difficult to use full-covariance models (which consider correlations between dimensions), and, instead, diagonal-covariance models are widely used. This limitation can degrade the model's performance. Thus, there are several methods for transforming a feature space so the state-conditional covariances of the features. One such widely used transformation method is MLLT. Another challenge for modeling is the large variation among speakers. To address this problem, SAT is typically used. In SAT, training is conducted after adaptation, which transforms the input space into a canonical space so as to reduce the variance across speakers. In this study, we validated the effectiveness of LDA, MLLT, and SAT.

In discriminative feature transformation (section 2.2), $N = 400$ components in GMM were used and offset features were calculated for each MFCC dimension (a total 40 dimensions) with context expansion (9 frames). Feature $\mathbf{h}_t$'s dimension was $400 \times 40 \times 9$, and features at the top $Q = 2$ posteriors were selected and all other features were ignored. α was set to 5.

We summarize the experimental procedure based on the above setup as follows: First, a clean acoustic model was trained. The number of mono-phones was 40, including silence ("sil"). The tri-phone model had 2,500 states, and the total number of Gaussians was 15,000. Second, using their alignments and tri-phone tree structures, reverberated acoustic models were trained using the "reverberated" dataset. Third, noisy acoustic models were trained multi-conditionally using the "isolated" noisy dataset without any special pre-processing such as blind source separation. Finally, starting with this ML model, performed discriminative training and feature transformation for the "isolated" noisy dataset. The parameters used in our experiments were based on those in the WSJ tutorial included in the Kaldi toolkit.

4. RESULTS AND DISCUSSION

4.1. Clean speech (WSJ0)

We evaluated the Word Error Rate (WER) for clean speech (si_et_05), which is a baseline model for the following experiments. The WER of the tri-phone model is listed in Table 1. The bMMI discriminative training improves the WER relative to non-boosted training. The feature-based discriminative training further improves the result. LDA with MLLT improves the WER by 0.2% overall. The addition of SAT improves the WER by 1% overall.

4.2. Reverberated and noisy speech (CHiME)

4.2.1. Baseline

Starting from the initial tri-phone model trained on clean data, we retrained using reverberated and noisy data. Reverberation and noise causes errors in the alignment and reconstruction of tree structures. We consider whether alignment (A) and tree structures (T) are retrained on noisy data (y) versus the same to those of clean model (n). This generates three new conditions: (A=n,T=n), (A=y,T=n), and (A=y,T=y). The WER results in the rest of the paper are on the development set (si_dt_05). For the "reverberated" case, the WER of the tri-phone models (ML) are 12.69% (A=n,T=n), 12.05% (A=y,T=n), 12.35% (A=y,T=y). Using an alignment by the (A=y,T=n) model, which achieved the best performance, we retrained models on the "isolated" noisy dataset. The averages of these ML models are 56.29% (A=n,T=n), 56.37% (A=y,T=n), and 56.98% (A=y,T=y). The performance of the (A=y,T=y) model is inferior to that of the other models. The performance of (A=n,T=n) and (A=y,T=n) are almost equivalent; we use the (A=y,T=n) condition as a baseline model. Discriminative training and feature transformation were carried out starting with this model. The baseline ML model was also used to generate lattices used for the denominator of the discriminative objective function.

Table 1. WER[%] for clean speech (**si_et_05**) (tri-phone model, 2,500 states, 15,000 Gaussians).

	none	LDA+MLLT	LDA+MLLT+SAT
ML	5.34	5.10	4.15
MMI	4.91	4.58	3.44
bMMI	4.73	4.30	3.38
f-MMI	4.71	4.26	3.40
f-bMMI	4.35	4.02	2.90

4.2.2. PLP results

For the ML model, experiments using the PLP features were also performed. For clean speech, the WER of PLP is 5.38% , which is equivalent to that of MFCC (5.34%). For the "reverberated" case, the WER of PLP is 13.87% (A=y,T=n), which is worse than that of MFCC (12.05%). For the "isolated" noisy case, the average WER of PLP (A=y,T=n) is 57.35%, which is still worse than that of MFCC (56.37%). Although PLP is thought to be robust to noisy speech, these experiments show a possible weakness of PLP when handling reverberated speech.

4.2.3. Discriminative training

First, with regard to the MFCC features, the improvement of the WER by discriminative training from the ML baseline is shown in Table 2. Evidently discriminative training is better able to compensate for noise than ML training, even in scenarios such as this where the background noise is highly variable. The boosted model improves the WER by 1% relative to the non-boosted one, whereas the feature space technique improves the WER by 3% overall. We believe that the feature space is adapted for a target speaker to improve the WER and that this effect reduces the influence of other noises. In these tables, the boosting factor b is set to 0.1. In preliminary experiments, boosting factors were set to 0.05, 0.1, 0.2, and 0.5. Performance did not strongly depend on the boosting factors and that the optimized values of the boosting factor are approximately 0.1-0.2.

4.2.4. Feature transformation

In these experiments, the MFCC features were first transformed using LDA and MLLT. Table 3 shows the WER.As mentioned in the introduction, LDA and MLLT improve the model performance in ordinary noise conditions. These significant improvements may relate to the characteristics of the CHiME database. The CHiME database's

Table 2. WER[%] for isolated noisy speech (**si_dt_05**) (tri-phone model, discriminative training with MFCC features).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML	74.20	66.57	58.24	51.84	46.73	40.64	56.37
MMI	73.40	65.60	56.88	51.17	45.40	41.20	55.61
bMMI	72.78	64.71	55.69	50.83	44.00	40.27	54.71
f-MMI	69.94	62.50	54.51	48.74	42.73	38.34	52.79
f-bMMI	68.64	61.56	53.11	47.65	41.73	36.98	51.61

Table 3. WER[%] for isolated noisy speech (**si_dt_05**) (tri-phone model, discriminative training with MFCC+LDA+MLLT features).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML	70.95	62.62	53.98	47.37	40.27	34.84	51.67
MMI	68.55	61.12	53.41	46.32	39.52	34.30	50.54
bMMI	68.74	60.98	51.95	45.86	38.16	32.85	49.76
f-MMI	66.19	58.24	49.23	43.58	36.89	31.35	47.58
f-bMMI	66.65	57.46	48.25	42.99	35.71	31.07	47.02

noises includes interfering speech from other speakers such as children. LDA seems to be dealing effectively with this interference presumably by finding linear transforms that focus on the feature subspace least occupied by the interference.

It is also effective to use context to reduce the influence of non-stationary noises. Furthermore, although noises increase the correlations between MFCC coefficients in each dimension, MLLT reduces the correlations and improves the WER. Denominator lattices for discriminative training are re-generated using ML (MFCC+LDA+MLLT) model.

In another experiment we added SAT and MLLR to the LDA+MLLT model as shown in Table 4. In this case, because the amount of training data is not sufficient, transformation into a canonical space leads to an increase in the amount of training data effectively and the estimation accuracy of the acoustic models increases. Additionally, MLLR adaptation for a target speaker reduces the influence of noises. Denominator lattices for discriminative training were re-generated using the ML model.

4.2.5. Augmented discriminative feature transformation

Table 5 shows the WER of ML and f-MMI whose auxiliary features $\mathbf{h}'_t$ in Eq. (7) are static MFCC and PLP (13 dimensions each), respectively. In the ML model, as mentioned in section 4.2.2, the

Table 4. WER[%] for isolated noisy speech (**si_dt_05**) (tri-phone model, discriminative training with MFCC+LDA+MLLT+SAT features).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML	68.36	58.30	48.80	40.73	35.09	28.54	46.64
MMI	65.13	55.27	45.89	39.64	33.12	27.29	44.39
bMMI	64.60	55.10	45.82	39.05	32.72	26.86	44.03
f-MMI	63.09	52.62	42.44	36.29	31.01	25.52	41.83
f-bMMI	62.43	52.23	42.17	35.31	29.84	24.72	41.12

Table 5. WER[%] for isolated noisy speech (**si_dt_05**) (tri-phone model, discriminative feature transformation with MFCC (M) and PLP (P) features).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML(M)	74.20	66.57	58.24	51.84	46.73	40.64	56.37
ML(P)	74.57	67.50	59.76	53.02	47.00	42.23	57.35
f-MMI	69.94	62.50	54.51	48.74	42.73	38.34	52.79
(+M)	70.14	61.54	55.76	48.05	43.49	38.93	52.99
(+P)	69.52	62.31	54.48	48.59	42.94	37.90	52.62

performance of PLP is worse than that of MFCC. However, adding PLP to the discriminative feature transformation improves the WER, whereas adding MFCC does not improve the WER, presumably because of redundancy with the original features $\mathbf{x}_t$. On the other hand, since PLP features are different there can be some benefit to learning a transform based on them.

4.2.6. Evaluation set

Table 6 shows the WER of the evaluation set using the models tuned using the development set. The baseline is ML (MFCC), whereas on top of MFCC+LDA+MLLT+SAT, "Best 1" is ML and "Best 2" is f-bMMI. As a reference, the HTK based baseline was 55.01% [19]. Both discriminative training and feature transformation ("Best 2") achieve 33.22% error reductions relative to the baseline, and thus appear to be effective for reverberated and noisy speech.

Table 6. WER[%] for isolated noisy speech (**si_et_05**). The baseline is ML (MFCC), whereas on top of MFCC+LDA+MLLT+SAT, "Best 1" is ML and "Best 2" is feature-space boosted MMI.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
Baseline	69.79	62.71	55.86	46.89	42.07	37.49	52.47
Best 1	60.83	52.14	43.51	34.28	29.22	23.82	40.63
Best 2	54.70	45.11	35.98	28.64	24.38	21.39	35.04

5. CONCLUSION AND FUTURE WORK

We developed a state-of-the-art baseline for the 2$^{\text{nd}}$ 'CHiME' Speech Separation and Recognition Challenge. This baseline validated the effectiveness of both discriminative training and feature transformation on realistic reverberated and noisy environments. We proposed a framework to add auxiliary features to a discriminative feature transformation. Experiments show that these techniques, especially feature transformations, are effective for non-stationary interference and reverberation. The auxiliary feature approach provided some promising preliminary improvements. In future work, we plan to further investigate auxiliary features that reflect important characteristics of interference and reverberation, including auxiliary features derived from microphone array signal processing, for discriminative feature transformation.

6. REFERENCES

[1] J.M. Baker, L. Deng, J. Glass, S. Khudanpur, C.H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding part 1," *IEEE Signal Processing Magazine*, vol. 26, pp. 75–80, May 2009.

[2] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Magazine*, vol. 25, pp. 14–36, September 2008.

[3] L. Bahl, P. Brown, P. de Souza, and R. Mercer "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings ICASSP*. IEEE, 1986, pp. 49–52.

[4] D. Povey, and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings ICASSP*. IEEE, 2002, pp. 105–108.

[5] E. McDermott, T.J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 203–223, January 2007.

[6] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," in *Proceedings ICASSP*. IEEE, 2010, pp. 4894–4897.

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

[8] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proceedings ICASSP*. IEEE, 1998, pp. 661–664.

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

[10] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proceedings ICSLP*, ISCA, 1996, pp. 1137–1140.

[11] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proceedings ICASSP*. IEEE, 2005, pp. 961–964.

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, pp. 82-97, November 2012.

[13] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proceedings INTERSPEECH*. ISCA, 2005, pp. 2977–2980.

[14] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proceedings INTERSPEECH*. ISCA, 2005, pp. 989–992.

[15] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proceedings INTERSPEECH*. ISCA, 2006. pp. 1573–1576.

[16] M. Delcroix, A. Ogawa, S. Watanabe, T. Nakatani, and A. Nakamura, "Discriminative feature transforms using differenced maximum mutual information," in *Proceedings ICASSP*. IEEE, 2012, pp. 4753–4756.

[17] H. Erdoğan, "Regularizing linear discriminant analysis for speech recognition," in *Proceedings INTERSPEECH*. ISCA, 2005, pp. 3021–3024.

[18] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *Proceedings ASRU*. IEEE, 2007, pp. 238–247.

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings ICASSP*. IEEE, 2012, to appear.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings ASRU*. IEEE, 2011, pp. 1–4.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4.1)," *http://htk.eng.cam.ac.uk*, March 2009.

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

[23] D. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proceedings ICASSP*. IEEE, 2001, pp. 517–520.

[24] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, April 2006.

[25] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, January 1999.

[26] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 516–527, March 2011.