Speech selection and environmental adaptation for asynchronous speech recognition

Bo Ren*, Longbiao Wang*, Atsuhiko Kai† and Zhaofeng Zhang*
* Nagaoka University of Technology, Nagaoka 940-2188, Japan
E-mail:{s145011@stn, wang@vos, s147002@stn}.nagaokaut.ac.jp
† Shizuoka University, Hamamatsu 432-8561, Japan
E-mail: kai@sys.eng.shizuoka.ac.jp

***Abstract*—In this paper, we propose a robust distant-talking speech recognition system with asynchronous speech recording. This is implemented by combining automatic asynchronous speech (microphone or mobile terminal) selection and environmental adaptation with deep neural network based framework. Although applications using mobile terminals have attracted increasing attention, there are few studies that focus on distant-talking speech recognition with asynchronous mobile terminals. For the system proposed in this paper, by using bottleneck Features (BFs) from a Deep Neural Network (DNN) rather than the conventional Mel-Frequency Cesptral Coefficients (MFCCs), we adopted the state-of-the-art deep neural network acoustic model, environmental adaptation and automatic asynchronous speech selection. The proposed method was evaluated by using a reverberant WSJCAM0 corpus, which was emitted by a loudspeaker and recorded in a meeting room with multiple speakers by far-field multiple mobile terminals. By using the bottleneck features based DNN acoustic model with automatic asynchronous speech selection and environmental adaptation, the average Word Error Rate (WER) was reduced from 55.32% of the baseline system to 19.38%, i.e. the relative error reduction rate was 64.97%.**

Index Terms: distant-talking speech recognition, tandem DNN, hybrid DNN, model adaptation, asynchronous speech

I. INTRODUCTION

Many techniques have been proposed for robust automatic speech recognition in noise and reverberation, using multiple microphones such as a microphone array [1][2][3][4][5][6][7][8] [9]. These techniques require the synchronous signals of multiple microphones and the cost and preparation of the microphone array are considerable. The synchronous microphone array device is not available in many meeting rooms. In this paper, we present a robust hands-free speech recognition system using a ubiquitous asynchronous smart terminal such as a smart phone. A diagram of the proposed system is shown in Fig. 1. The proposed system consist of three components: (1) feature extraction and transformation, (2) DNN-based back-end processing, (3) automatic asynchronous speech selection and environmental adaptation. In this report, we are focusing on (1) feature transformation and (3) automatic asynchronous speech selection.

Many single-channel dereverberation methods have been proposed for robust distant-talking speech recognition [10][11][12][13]. Cepstral Mean Normalization (CMN) [14][15] may be considered the most general approach. It has been extensively examined and shown as a simple and effective way of reducing reverberation by normalizing cepstral features. However, the dereverberation of CMN is not completely effective in environments with late reverberation. Several studies have focused on mitigating the above problem [5][11][13][16][17]. A reverberation compensation method for speaker recognition using spectral subtraction, in which late reverberation is treated as additive noise, was proposed in [11]. A method based on Multi-Step Linear Prediction (MSLP) was proposed by [13], for both single and multiple microphones. The method first estimates late reverberations using long-term multi-step linear prediction, and then suppresses these with subsequent spectral subtraction. The drawback of this approach is that the optimal parameters for spectral subtraction are empirically estimated from a developing dataset, meaning that the late reverberation cannot be subtracted correctly as it is not precisely modeled. Previous work have shown that Denoising AutoEncoders (DAEs) [18], [19] is robust against the reverberation environment, because higher level representations and increased flexibility of the feature mapping function can be learned [20]. DAEs were treated as a flexible feature mapping method, which can generate robust features against the reverberation. Similar to DAEs, here we employ another feature mapping method, which generate bottleneck feature [21], [22] from a DNN with a bottleneck layer. This bottleneck layer creates a constriction in the network that force the information pertinent to classification into a low dimensional representation [23]. It has been shown that using a deeper neural network joint with Hidden Markov Models (HMMs) achieves state-of-the-art accuracy [24][25] and features derived from DNNs are strongly discriminated and invariant, i.e., less sensitive to perturbation in the input [26] . In this paper, by utilizing the property of DNN, reverberation robust bottleneck features are learned from a deep neural network.

DNN approaches have recently produced significant improvement in the accuracy of acoustic modelling for speech recognition, across a range of domains and evaluation datasets [24][25][27][28]. Compared to the Gaussian Mixture Model (GMM), DNN do not rely on the assumption that HMM states satisfy the gaussian distribution, and DNN typically employ a deep architecture to learn invariant and discriminative internal representations. In another words, DNNs may increase the

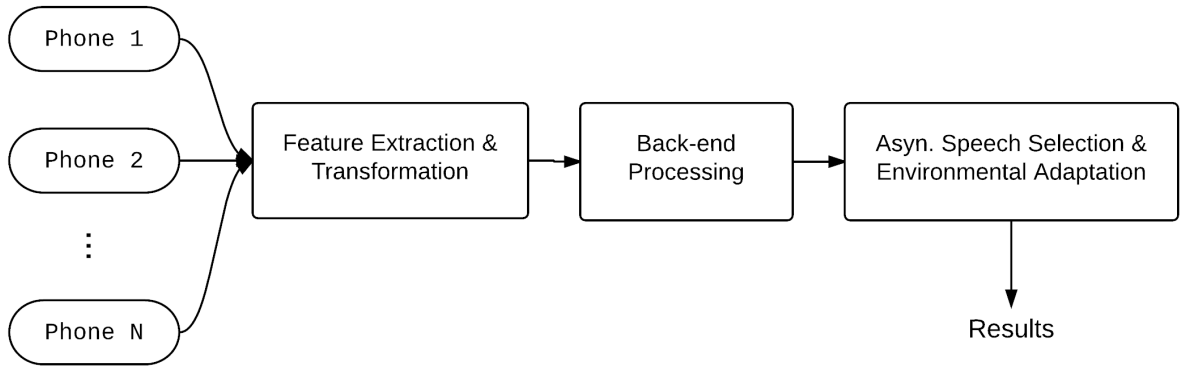

Fig. 1. System diagram of speech recognition with asynchronous speech recording

accuracy for speech recognition. Inspired by the improvement, instead of GMM, DNN was introduced into the proposed system.

Due to asynchronous speech recording are used in this situation, automatic mobile terminal selection and unsupervised environment adaptation are applied. For automatic mobile terminal selection, the optimal mobile terminal of a speech segment based on Voice Activity Detection (VAD) was automatically selected according to the maximum likelihood or power of the speech segment from each terminal. In this paper, an ideal VAD is applied. For environmental adaptation, we apply the Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation method [29][30], which is also called feature space Maximum Likelihood Linear Regression (fMLLR), to the features corresponding to the speech segment selected automatically.

II. BOTTLENECK FEATURE

Bottleneck features are generated from a deep neural network in which one of the internal layers has a small number of hidden units, relative to the size of the other layers. This small layer creates a constriction in the network that forces the information pertinent to classification into a low dimensional representation. Bottleneck features are most commonly used in an autoencoder which the neural network is trained to predict the input features themselves [31]. Because the activations at the bottleneck layer are a low-dimensional nonlinear function of the input features, an autoencoder can be viewed as a method of nonlinear dimensionality reduction. Bottleneck features for speech recognition are created from a DNN trained to predict phonemes or phoneme states. The inputs to the hidden units of the bottleneck layer are used as features for an HMM-based speech recognizer. These bottleneck features represent a nonlinear transformation and dimensionality reduction of the input features.

In the proposed system, we trained a bottleneck DNN with a bottleneck hidden layer of 42 hidden units rather than the other hidden layers with 1024 units to predict phoneme states. Including the bottleneck layers, there are totally 6 hidden layers in the deep architecture and the nolinear tanh function was used as the activation at all hidden layers. This bottleneck DNN was trained on the multi-condition training dataset, which contained around 16.5 hours speech sound, with around 8.5 million parameters using the well-known error back-propagation procedure [32]. The stochastic mini-batch gradient descent with a minibatch size of 512 samples was used to optimize cross entropy cost function. Input features used for bottleneck DNN is typically using variants of TRAPS features, in which long temporal windows of critical band energies are processed [33][34]. However, In most systems, the best performance is obtained by combining the bottleneck features derived from these inputs with traditional features, e.g. MFCCs or PLP. It is believed that these bottleneck features are complementary to the conventional features derived from the short-time spectra of input. Due to the existence of reverberation, we consider that the context frames contain some relative information which could extract discriminative and dereverberation robust bottleneck features. In our proposed system, the adjacent 9 frames (i.e. ± 4) of 12-dimensional MFCCs plus power (i.e. 117-dimensional in total) were used as input feature.

III. AUTOMATIC SPEECH SELECTION AND ENVIRONMENT ADAPTATION

In distant-talking speech recognition, recognition accuracy is significantly degraded by noise and reverberation. It depends on the distance and direction of the microphone and the speaker. In this situation where asynchronous mobile terminals

are used, it is important to select an optimal speech segment from an optimal mobile terminal. For the sake of simplicity, the ideal VAD was used in this paper. Thus, the challenge became how to automatically select an optimal mobile terminal. We selected the speech from the mobile terminal with the maximum likelihood criterion or maximum power criterion, on the assumption that this terminal is nearest to the speaker and is most appropriate for recognition. Speech selection were applied only in the evaluation steps, as training data were single channel speech rather than asynchronous speech. In a previous work [20], the maximum power criterion is tend to make an incorrect selection decision when the terminal is closed to some noise source emitting huge power. However the maximum likelihood criterion is less sensitive and robust against the complex real environments.

Given the speech segment of $n-th$ mobile terminal x_n, upon the maximum likelihood criterion, the optimal mobile terminal is selected as:

$$\hat{n} = \arg\max_{n} L_n, n = 1, \ldots, N \tag{1}$$

where L_n and N are the maximum likelihood of the speech segment x_n and the number of mobile terminals, respectively. As a result, the speech segment $x_{\hat{n}}$ was selected as the optimal speech.

For each speech segment $x_n, n \in \{1, \ldots, N\}$, it is split into frames $f_i, i \in \{1, \ldots, M\}$ and labeled by a context-dependent phoneme state l_i, $l_i \in \{1, 2, \ldots, J\}$. All the context-dependent phoneme states were generated by GMM-HMM acoustic models, the emission distribution of context-dependent phoneme state $j \in \{1, 2, \ldots, J\}$ is modeled by a gaussian distribution with mean μ_j and covariance Σ_j. The maximum likelihood L_n of the speech segment x_n, which is corresponding to the $n-th$ mobile terminal, would be as

$$L_n = \prod_{i=1}^{M} p(f_i, \mu_j, \Sigma_j) \tag{2}$$

where $p(f_i, \mu_j, \Sigma_j)$ is the posterior probability that the frame f_i is emitted from a context-dependent phoneme state $j \in \{1, 2, \ldots, J\}$.

fMLLR [29][30] is the method that modifies the featur values relative to gaussian distribution for each HMM state by using the regression matrix to reduce the mismatch between the adaptation data and models. This method is intended to obtain a transformation matrix for modifying the features, so that maximize the likelihood of the adaptation data. In this paper, we applied fMLLR for feature space adaptation, i.e. environment adaptation.

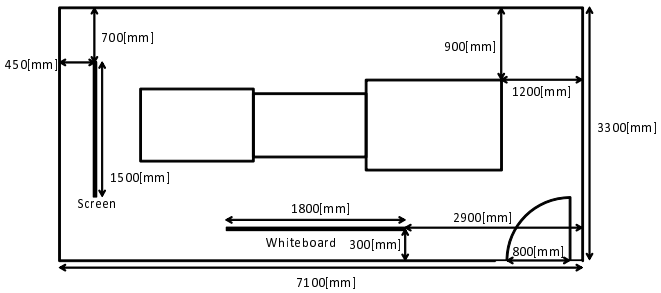

Fig. 2. Structure of the recording room (High: 250 cm, Reverberation time: about 0.6 second)

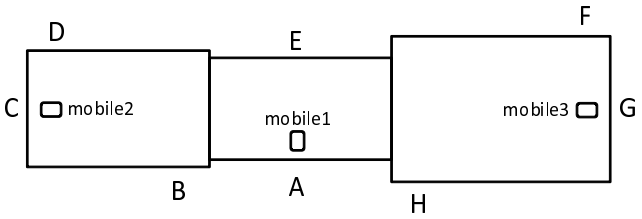

Fig. 3. Placement of the speakers and the mobile terminals (Height of the table: 70 cm, Height of the speaker: 85 cm)

IV. EXPERIMENT

A. Experiment Setup

1) training dataset: The training dataset provided by REVERB challenge (Reverberant Voice Enhancement anRecognition Benchmark) [35] was used. This dataset consists of the clean WSJCAM0 [36] training set and a multi-condition (MC) training set. Reverberant speech is generated from the clean WSJCAM0 training data by convolving the clean utterances with measured room impulse responses and adding recorded background noise. The reverberation times of the measured impulse responses range from approximately 0.1 to 0.8 sec. The number of speakers was 92 and the total number of utterances was 7861. This training dataset was used for both training of acoustic models and parameters of bottleneck DNN.

It should be noted that the recording rooms used for the multi-condition training data and test data were different.

2) evaluation dataset: To evaluate the proposed method, 100 utterances randomly selected from the evaluation test set of the WSJCAM0 corpus were emitted from a loudspeaker and recorded by three asynchronous mobile terminals (iPhone 4S) set in a seminar room. Fig. 2 shows the structure of the seminar (recording) room and Fig. 3 shows the positions of the loudspeakers and the placement of the mobile terminals. The speakers were fixed at eight positions from A to H shown

TABLE I
WERs (%) OF DNN SYSTEM WITH ENVIROMENTAL ADAPTATION, USING BFs

LDA_STC	Selection		No Selection (average)
	ML	MP	
No	20.02	**19.50**	31.23
Yes	**19.38**	19.91	29.83

in Fig. 3. We recorded 100 utterances in total, at positions A-H, using an iPhone 4S application called "PCM recording" for speech recording.

3) baseline systems: The well-known GMM and DNN were used to model output probabilities of the context-dependent HMM states in the baseline systems. This GMM were trained on the multi-condition training dataset with around 2000 tied-states and about 15000 gaussians. The 12-dimensional MFCCs puls power and their Δ and $\Delta\Delta$ coefficients, i.e. 39 dimension totally, were used as input features, which were also normalized by CMN for per-speaker. In the evaluation step, a feature space transform, which is obtained from the GMM-HMM model, was applied for unsupervised environment adaptation to alleviate the mismatch between training data and evaluation data. As the limit of training dataset, which is just around 16.5 hours, a deep neural network consisting of 2 hidden layers and 1204 units in each layer was trained with a initial learning rate of 0.015. Similar to bottleneck DNN, it was trained by using the stochastic mini-batch gradient descent with a minibatch size of 512 samples. At the decoding step, a standard 5k WSJ bigram language model was used. Both normalization and enviromental adaptation were used for both MFCC features and bottleneck features. we used the DNN implementation as reported in [37], which is part of the Kaldi toolkits [38].

B. Experiment Results

Table II shows the speech recognition results under multi-condition training dataset. "No Selection" shows the average recognition results of multiple mobile terminals when recognizing singly recorded for each mobile terminal without any selection. "ML" showes the speech recognition results of applying automatic mobile terminal selection with the Maximum Likelihood criterion, while "MP" employs the Maximum Power criterion. The best result in the same kind of system is **bold**. When bottleneck features were employed, a remarkable improvement was achieves in both GMM and DNN systems for all the evaluation datasets. The average Word Error Rates (WERs) of multiple mobile terminals were improved from 55.32% of MFCC features to 39.46% of bottleneck features in the conventional GMM system. We consider that one of most important reason that bottleneck DNN learns some invariant and discriminative representations with a complicated nonlinear transform, and this nonlinear transform is robust to environmental variations. Moreover, DNN system with bottleneck features gains more benefit, reduces the speech recognition WERs from 39.46% of GMM system to 34.50%. This reconfirm the conclusion from [39] that using both tandem DNN (i.e. bottleneck feature) and hybrid DNN achieves the best performance. As the environmental adaptation (i.e. fMLLR) was introduced into the systems, the speech recognition performance was improved. Because the environmental variation of speech is ubiquitous in reality, fMLLR applied here generalized the variation between training environment and adaptation environment.

By integrating the automatic mobile terminal selection, an additional reduction of WERs was obtained. In all the systems but DNN system with bottleneck features, the proposed maximum likelihood criterion automatic mobile terminal selection obtained a little more gain than the previous maximum power criterion. This confirms the idea that was introduced in Section 3. However, the best performance of this situation was showed in Table I. When an additional linear discriminant analysis (LDA) [40] with semi-tied covariance (STC) matrix [41] was applied, the maximum likelihood criterion achieves the best performance of 19.38% comparing to 55.32% of the baseline system, i.e. the reduction rate was 64.97% in the proposed system.

V. CONCLUSION

In this paper, we proposed a robust distant-talking speech recognition for asynchronous speech that was recorded using multiple mobile terminals. In this system, the reverberation and enviromental distortions robust bottleneck feature and the state-of-the-art DNN acoustic model were integrated. For alleviating the mismatch of the asynchronous speechs, we introduce a maximum likelihood criterion comparing to maximum power criterion in previous work to avoid the degradation of some high power noise. The best performance achieved in the DNN system using bottleneck features with a LDA_STC linear transform for de-corerelation was 19.38% comparing to 55.32% of the baseline system.

In the further, we will integrate some dereverberation methods into the proposed system to supress the degradation of dereverberation in a meeting room.

VI. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 15K16020 and a research grant from the Research Foundation for the Electrotechnology of Chubu (REFEC).

TABLE II
WERs (%) OF KINDS OF SYSTEMS

Feature	AM	fMLLR	Selection		No Selection (average)
			ML	MP	
MFCC	GMM	No	**36.43**	36.78	55.32
	GMM	Yes	**30.82**	32.40	48.98
	DNN	No	**28.20**	29.07	44.93
	DNN	Yes	**23.58**	25.10	37.42
BF	GMM	No	**25.39**	25.69	39.46
	GMM	Yes	**23.41**	24.34	38.12
	DNN	No	**21.42**	21.54	34.50
	DNN	Yes	20.02	**19.50**	31.23

REFERENCES

[1] AC Surendran and JL Flanagan, "Stable dereverberation using microphone arrays for speaker verification," *J. Acoust. Soc. Am.*, vol. 96, no. 5, pp. 3261–3262, 1994.

[2] Sharon Gannot and Marc Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Processing*, vol. 2003, no. 1, pp. 1074–1090, 2003.

[3] E. A P Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2005, vol. IV.

[4] Longbiao Wang, Norihide Kitaoka, and Seiichi Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN," *EURASIP J. Appl. Signal Processing*, vol. 2006, pp. 1–11, 2006.

[5] Marc Delcroix, Takafumi Hikichi, and Masato Miyoshi, "Precise dereverberation using multichannel linear prediction," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2007, vol. 15, pp. 430–440.

[6] Longbiao Wang, Norihide Kitaoka, and Seiichi Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Commun.*, vol. 49, pp. 501–513, 2007.

[7] Kinoshita Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Masato Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2009, vol. 17, pp. 534–545.

[8] Wang Longbiao, Norihide Kitaoka, and Seiichi Nakagawa, "Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm," *IEICE Trans. Inf. Syst.*, vol. 94, no. 3, pp. 659–667, 2011.

[9] Longbiao Wang, Zhaofeng Zhang, and Atsuhiko Kai, "Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2013, pp. 7224–7228.

[10] Mingyang Wu and DeLiang Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2006, vol. 14, pp. 774–784.

[11] Qin Jin, Tanja Schultz, and Alex Waibel, "Far-field speaker recognition," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2007, vol. 15, pp. 2023–2032.

[12] Seyed Omid Sadjadi and John H L Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2011, pp. 5448–5451.

[13] Keisuke Kinoshita, Tomohiro Nakatani, and Masato Miyoshi, "Spectral Subtraction Steered by Multi-Step Forward Linear Prediction For Single Channel Speech Dereverberation," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2006, vol. 1.

[14] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *Acoust. Speech Signal Process. IEEE Trans.*, vol. 29, 1981.

[15] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, and Alejandro Acero, "Efficient cepstral normalization for robust speech recognition," *Proc. Work. Hum. Lang. Technol.*, p. 69, 1993.

[16] Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[17] Zhaofeng Zhang, Longbiao Wang, and Atsuhiko Kai, "Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation," *EURASIP J. Audio, Speech, Music Process.*, vol. 2014, no. 1, pp. 15, 2014.

[18] Yuma Ueda, Longbiao Wang, Atsuhiko Kai, Xiong Xiao, EngSiong Chng, and Haizhou Li, "Single-channel dereverberation for distant-talking speech recognition by combining denoising autoencoder and temporal structure normalization," *Journal of Signal Processing Systems*, 2015.

[19] Zhaofeng Zhang, Longbiao Wang, Atsuhiko Kai, Kyohei Odani, Weifeng Li, and Masahiro Iwahashi, "Fusion of deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," *EURASIP Journal on Audio, Music and Speech Processing*, 2015.

[20] Longbiao Wang, Bo Ren, Yuma Ueda, Atsuhiko Kai, Shunta Teraoka, and Taku Fukushima, "Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording," in *APSIPA ASC*, 2014.

[21] Bo Ren, Longbiao Wang, Liang Lu, Yuma Ueda, and Atsuhiko Kai, "Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition," *Multimedia Tools and Applications*, 2015.

[22] Takanori Yamada, Longbiao Wang, and Kai Atsuhiko, "Improvement of distant-talking speaker identidication using bottleneck features of DNN," in *Interspeech 2013*, 2013, pp. 3661–3664.

[23] František Grézl, Martin Karafiát, Stanislav Kontár, and Jan Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.* IEEE, 2007, vol. 4, pp. 757–760.

[24] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[25] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton, "Acoustic Modeling Using Deep Belief Networks," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.* 2012, vol. 20, pp. 14–22, IEEE.

[26] Dong Yu, Michael L Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide, "Feature Learning in Deep Neural Networks - Study on Speech Recognition Tasks," *CoRR*, vol. abs/1301.3, pp. 1–9, 2013.

[27] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, "Deep neural networks

for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. November, pp. 82–97, 2012.

[28] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.* 2012, vol. 20, pp. 30–42, IEEE.

[29] Mark JF Gales and Philip C Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, no. 4, pp. 249–264, 1996.

[30] Mark JF Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[31] Geoffrey E. Hinton and RR R Salakhutdinov, "Reducing the dimensionality of data with neural networks.," *Science*, vol. 313, no. July, pp. 504–507, 2006.

[32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[33] Hynek Hermansky and Sangita Sharma, "TRAPS – Classifiers of Temporal Patterns," in *Proc. ICSLP*, 1998, vol. 3, pp. 1003–1006.

[34] František Grézl and Petr Fousek, "Optimizing bottle-neck features for LVCSR," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.*, 2008, pp. 4729–4732.

[35] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Appl. Signal Process. to Audio Acoust. (WASPAA), 2013 IEEE Work.* IEEE, 2013, pp. 1–4.

[36] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals, "Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.* 1995, pp. 81–84, IEEE.

[37] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf.* 2014, pp. 215–219, IEEE.

[38] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and Others, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[39] Martin Sundermeyer, Ralf Schl, and Hermann Ney, "Context-Dependent MLPs for LVCSR : TANDEM , Hybrid or Both ?," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2012.

[40] Suresh Balakrishnama and Aravind Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Inst. Signal Inf. Process.*, vol. 11, pp. 1–9, 1998.

[41] Mark JF Gales, "Semi-tied covariance matrices for hidden Markov models," *Speech Audio Process. IEEE Trans.*, vol. 7, no. 3, pp. 272–281, 1999.