INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Optimizing DNN Adaptation for Recognition of Enhanced Speech

Marco Matassoni, Alessio Brutti, Daniele Falavigna

Center for Information and Communication Technology
Fondazione Bruno Kessler, via Sommarive 18, Trento (Italy)
{matasso,brutti,falavi}@fbk.eu

Abstract

Speech enhancement directly using deep neural network (DNN) is of major interest due to the capability of DNN to tangibly reduce the impact of noisy conditions in speech recognition tasks. Similarly, DNN based acoustic model adaptation to new environmental conditions is another challenging topic. In this paper we present an analysis of acoustic model adaptation in presence of a disjoint speech enhancement component, identifying an optimal setting for improving the speech recognition performance. Adaptation is derived from a consolidated technique that introduces in the training process a regularization term to prevent overfitting. We propose to optimize the adaptation of the clean acoustic models towards the enhanced speech by tuning the regularization term based on the degree of enhancement. Experiments on a popular noisy dataset (e.g., AURORA-4) demonstrate the validity of the proposed approach.

Index Terms: robust speech recognition, DNN adaptation, speech enhancement

1. Introduction

Speech enhancement is a long studied topic by the scientific community dealing with automatic speech recognition (ASR), since it has been demonstrated for a variety of tasks, environments and applications that the removal/reduction of both noise and reverberation significantly improves the recognition performance. This has become particularly evident recently, when the introduction of deep learning-based speech enhancement has led to dramatic performance improvements.

When a speech enhancement component is available, a common approach consists in re-training acoustic models from scratch using an enhanced version of the training material. However, in many application scenarios this is not feasible (e.g., in case of highly variable acoustic conditions) and only a small set of data is available, both for learning the enhancement network and for DNN acoustic modeling. Therefore one has to resort to adapting previously trained acoustic models. For *unsupervised* DNN adaptation (i.e. with transcriptions obtained automatically with a preliminary recognition pass), we recently exploited [1] DNN regularization based on the Kullback-Leibler divergence (KLD) between original and actual output DNN distributions, as proposed in [2]. We also showed the effectiveness of the approach in combination with a module that estimates the quality of each sentence in the adaptation set in order to filter out the low quality ones.

In this work, coping with *supervised* DNN adaptation, we propose to modify the method described in [1, 3] by properly weighing the sentences available in an adaptation set. In this way, the acoustic observations in the adaptation set contribute differently to the creation of the regularization term and, consequently, to the definition of the total loss function to minimize.

Experimenting speech enhancement to improve the recognition of the test set of AURORA-4 noisy corpus [4, 5], we adopt an approach based on the joint usage of deep learning, to map the noisy features into clean ones, and of regularization to adapt the parameters of the original DNN acoustic models, trained on clean data, to the enhanced data. In this way we achieve significant word error rate reduction, with respect to the sole use of the enhanced features, even when the size of the adaptation set is small. Although the adopted enhancement is based on stereo data that partly limits its usefulness, the scope of the work is to investigate the effectiveness of acoustic model adaptation when a given DNN-based pre-processing strategy is employed. In summary, our contribution is two fold: 1) we show that the adaptation procedure must be tailored to the speech enhancement performance to achieve high word error rate (WER) reduction; 2) we propose a way to automatically control the adaptation, sentence by sentence, independent of the enhancement scheme.

This paper is organized as follows: Section 2 introduces the topic, discussing the main approaches while Section 3 presents the proposed technique; in Section 4 the recognition task and the system are presented, discussing then the experimental results in Section 5. Finally, Section 6 draws some conclusions and discusses directions for future work.

2. Related works

After some pioneeristic works of decades ago [6, 7, 8], recently deep learning based speech enhancement has gained large popularity becoming in practice a standard procedure and leading to an impressive amount of publications and a huge variety of solutions [9, 10, 11]. Speech enhancement methods can be grouped into two main categories: *i)* spectrogram enhancement optimizing the MSE as cost function [12, 13, 14, 15, 16]; *ii)* estimation of power spectral density (PSD) of speech and noise for spectral masking [17, 18, 19, 20, 21, 22]. In both cases a variety of deep networks are used, such as: feed-forward DNN, denoising autoencoder (DAE), convolutional neural network (CNN), bidirectional long-short term memory (B-LSTM).

DNN adaptation by regularization is usually employed to avoid overfitting the adaptation data. To this aim well known approaches make use of regularization terms based either on $L2$ norm [23] of the network weights or on Kullback-Leibler divergence between the target output distribution and the output distribution of the original network [2]. Other approaches can be implemented through the retraining of only subsets of the the network weights, e.g. the weights of the input layer, as in feature discriminative linear regression (fDLR) [24], or the weights of the output layer, as in output-features discriminative linear regression (oDLR) [25]. A variant of fDLR is described in [26], which proposes to adapt the DNN parameters within a maxi-

http://dx.doi.org/10.21437/Interspeech.2017-755

mum a posterior (MAP) framework. More generally, methods based on the application of linear transformations to input, output or hidden DNN layers can be found in [27, 28, 29, 30, 31]. The usage of dropout for DNN adaptation has been investigated in [32]. Finally, the application of a momentum term to update the DNN weights, the use of small values for the learning rate as well as of an early stopping criterion can be thought, at some extent, as ways to regularize a given original model.

3. Proposed Approach

As previously mentioned, the approach addressed in this work for acoustic model adaptation of an ASR system trained on clean data, consists in defining and optimizing a loss function that controls the adaptation depending on the distance between the distribution of the clean training material and the enhanced (or not) adaptation set. Figure 1 illustrates the proposed approach that introduces a suitable weighting function for controlling the regularization term for the DNN adaptation process, according to the actual enhancement level.

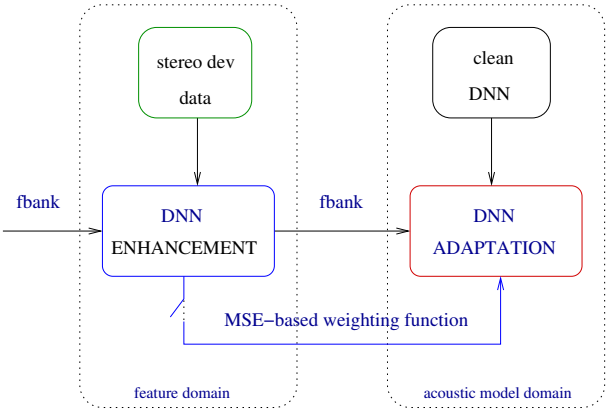

Figure 1: *System architecture: the DNN-based enhancement component generates enhanced filter-bank features as well as sentence-by-sentence weights used as regularization parameters for the forthcoming DNN adaptation procedure.*

3.1. DNN enhancement

Following the recent trend, we employ a DNN-based speech enhancement, implementing a rather simple scheme consisting of a feed-forward DNN which operates on the filter-banks. Given the noisy filter-bank input features $\mathbf{y}_t$ and the network output $\hat{\mathbf{o}}_t = f(\Theta, \mathbf{y}_t)$, the network parameters Θ are optimized to minimize the mean square error (MSE) with respect to the clean filter-banks $\mathbf{o}_t$:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{o}}_t - \mathbf{o}_t\|^2 + \lambda\|\mathbf{W}\|_2^2, \quad (1)$$

where λ is a regularization term, $\mathbf{W}$ are the network weights and K is the size of the mini-batch.

3.2. DNN adaptation

In a previous work [1] we demonstrated the benefits of KLD regularization both in cross and self DNN adaptation conditions. With *cross-adaptation* we mean that the adaptation set is disjointed from the test set, while in *self-adaptation* test and adaptation sets coincide. In this latter case the risk of overfitting the adaptation data is augmented by the fact that the target labels can be affected by recognition errors.

According to [2], the KLD regularization term is added defining a target output distribution P as linear interpolation between the original distribution $\overset{*}{p}$ and the distribution $\hat{p}$ computed via forced alignment with the adaptation data. In formulas:

$$P[s_i|\hat{\mathbf{o}}_t] = (1-\alpha)\hat{p}[s_i|\hat{\mathbf{o}}_t] + \alpha\overset{*}{p}[s_i|\hat{\mathbf{o}}_t] \quad 0 \leq \alpha \leq 1, \quad (2)$$

where $s_i, 1 \leq i \leq I$ represents the i^{th} output unit (i.e. the i^{th} HMM state) of the DNN to adapt, $\hat{\mathbf{o}}_t, 1 \leq t \leq T$ is the acoustic observation (enhanced filter-banks) at time t and α is the regularization coefficient. In [2] it is demonstrated that KLD regularization can be implemented by minimizing the cross-entropy of the distribution P, defined above, over the adaptation data.

In eq. 2, $\alpha = 0$ corresponds to a *pure* retraining over the adaptation data, (i.e. completely trusting them), while the value $\alpha = 1$ forces the output probability distribution of the adapted DNN to follow that of the original DNN (i.e. no adaptation). Hence, the expected optimal value of α will be close to 0 when the size of the adaptation set is large and the transcriptions are not affected by errors (i.e. in supervised conditions). On the contrary, when the size of the adaptation set is small and/or its transcription can be affected by errors (i.e. in the case of unsupervised adaptation), the optimal value of α should increase. However, even in supervised conditions the regularization term plays a crucial role, in particular when the adaptation set is small. As a matter of facts, it is rather intuitive to assume that, if the distribution of the features $\hat{\mathbf{o}}_t$ is very far from the original one, retraining the acoustic model is preferable. Conversely, as the mismatch is reduced, preserving the information in the original model may be beneficial.

The usage of a KLD based regularization term is due to its explicit dependence on the adaptation data, making the approach more suitable for investigating adaptation based on utterance weighing, as will be explained below.

3.3. Adaptive regularization

In [1] we experimented the usage of sentence dependent values of α for unsupervised adaptation, specifically we computed, for each adaptation sentence k, a value α_k as a function of an estimate of the WER of the k^{th} sentence itself.

In this work we adapt the acoustic models starting from those trained on the clean material; therefore, we propose estimating the regularization parameter as a function of a suitable *distance* between the enhanced signals and the clean ones. A reasonable metric can be derived (at sentence level) as function of the MSE between the features in the k^{th} clean sentence and those in the corresponding noisy one:

$$\hat{\alpha}_k = f(\text{MSE}_k), \quad (3)$$

where:

$$\text{MSE}_k = \frac{1}{K}\sum_{t=1}^{K} \|\hat{\mathbf{o}}_t - \mathbf{o}_t\|^2, \quad (4)$$

and K is the length of the k^{th} sentence. Analyzing some preliminary experiments, we choose a sigmoid function to map the MSE on the adaptive regularization coefficient $\hat{\alpha}_k$:

$$\hat{\alpha}_k = \frac{1}{1+\exp\left[-\sigma\left(\text{MSE}_k - \mu\right)\right]}. \quad (5)$$

Although the regularization term is heuristically derived, note that the MSE criterion employed in the equations above measures the distortion between the target DNN output distribution (representing the clean features) and the actual DNN output distribution (representing the enhanced/noisy features). Therefore the metric used to compare the two distributions could be defined through their KLD, similarly to what done for the HMM states, i.e. the same term used to derive eq. 2 [2]. This topic will be investigated in future work, as commented in the Section 6.

4. ASR system

The ASR system employed in the experiments is based on the KALDI open source software toolkit [33]. This integrates hybrid DNN-HMMs acoustic model in a static decoding network built by means of finite state transducers.

For this work we partly follow the original AURORA-4 recipe [1] for the clean condition (`train_si84_clean`) since we are interested in investigating adaptation strategies starting from clean acoustic models; the other option is to use multi-condition training (`train_si84_multi`) but in this case it would be more difficult to fairly evaluate the combination of enhancement and adaptation to noisy conditions.

The acoustic features are based on filter-bank energies consisting of 40 log Mel scaled filters. Feature vectors are computed every 10*ms* by using a Hamming window of 25*ms* length and are mean/variance normalized on a sentence-by-sentence basis. This differs from the original recipe but, again, the main goal is to investigate the impact of adaptation and enhancement in a very general setting without any assumption on speakers. The baseline DNN is trained using the Karel's setup [34] included in the KALDI toolkit. To this aim the 8,138 training utterances were aligned to their transcriptions by means of the baseline GMM-HMM models. An 11-frame context window (5 frames on each side) is used as input to form a 440 dimensional feature vector. The DNN has 7 hidden layers, each with 2,048 neurons. The DNN is trained in several stages including Restricted Boltzmann Machines (RBM) pre-training, and mini-batch stochastic gradient descent training. Initially, the learning rate is 0.008 and it is halved every time the relative difference in frame accuracy between two epochs on a cross-validation set falls below 0.5%. A frame accuracy improvement on the cross-validation set lower than 0.1% stops the optimization.

All experiments involving adaptation of the baseline DNN, aimed at minimizing the regularized cross-entropy of the adaptation data, defined by means of eq. 2, are performed according to the above recipe. The adaptation procedure is implemented as a re-training pass using a limited number of epochs (5) and a lower learning rate. The target labels of the adaptation set are derived from the alignment obtained on the *clean* data (i.e. exploiting the availability of stereo data). This choice is motivated by the fact that in this way the synergy between enhancement and adaptation does not incorporate the possible better alignment produced by enhanced features.

The language model (LM) used for decoding the standard evaluation set is based on the original pruned 5k closed-word trigram arpa model.

5. Experiments and results

Although the AURORA-4 dataset features simulated additive noise and reverberation effects and hence has known limits as recognition tasks, still represents an interesting experimental framework for evaluating robustness because of the variety of represented acoustic conditions.

Speech enhancement is based on a standard feed-forward DNN with 3 layers of 2048 units with ReLu activations; dropout with probability 0.1 and regularization term $\lambda = 10^{-4}$. The learning rate is 0.05 with 10% reduction at every epoch. Enhancement is performed on the 40 filter-bank feature vector used in the ASR, considering an 11-frame context which results in a dimensionality of the input feature vector of 440. The network is trained on the AURORA-4 development set, comprising 10 speakers uttering 330 sentences (overall 4620 total sentences). We consider 3 different networks (*enh0*, *enh1* and *enh2*) obtained stopping the network training after different epochs, to mimic different enhancement performance. In addition we consider, as upper-bound, a network trained on both the development and test sets (*oracle*). The DNN-based speech enhancement has been implemented using Theano [35, 36].

The parameters of the sigmoid mapping between the MSE and $\hat{\alpha}_k$ in eq. 5 have been empirically set to $\sigma = 3.5$ and $\mu = 1.8$ from an evaluation of the WER on the development set.

As reference, Table 1 reports the baseline performance in terms of WER on the AURORA-4 test set for the noisy signals (i.e., no enhancement) and the 3 enhancements using both the clean WSJ acoustic models and re-training the acoustic models on the noisy or enhanced development set. Note how the DNN-based enhancement improves the quality of the signals, considerably reducing the WER on the clean acoustic models. Nevertheless, training specific noisy acoustic models still provides better performance than training a new acoustic model from the enhanced material, unless a very powerful enhancement is available (i.e. enhancement trained on the test signals, referred to as *oracle* in the table.).

Table 1: *Baseline WERs for different combinations of enhancement (noisy means no enhancement) and acoustic models, clean or retrained on the development set ($\alpha = 0$). The 4 disaggregated AURORA-4 conditions are reported: A=clean close talk; B=noisy close talk; C=clean far talk; D=noisy far talk; oracle refers to enhancement trained on the test signals;* ada *indicates adapted models.*

Test/Model	Conditions				
	A	B	C	D	Total
noisy/clean	2.67	22.6	21.78	43.61	30.12
enh0/clean	3.64	14.92	9.79	28.99	19.78
enh1/clean	3.53	11.16	7.73	23.53	15.67
enh2/clean	3.29	10.48	7.23	23.29	15.22
oracle/clean	3.19	8.62	5.57	18.44	12.22
noisy/ada noisy	5.59	8.52	8.93	17.34	12.12
enh2/ada *enh2*	5.38	9.25	8.00	18.33	12.78
oracle/ada *orac.*	5.19	8.19	7.02	14.72	10.69

A considerable improvement is achieved when performing DNN adaptation with a proper KLD regularization. Figure 2 reports the WER on the entire AURORA-4 test set (averaging among all the acoustic conditions) as a function of the regularization coefficient α in eq. 2 when a constant value is used on the whole development set. Results for $\alpha = 0$ corresponds to retraining on the development set, while for $\alpha \rightarrow 1$ the performance would tend to those of the clean acoustic models (compare with baselines in Table 1). It can be observed how the lowest WER is obtained for different values of α which depend

[1] `https://github.com/kaldi-asr/kaldi...aurora4/s5`

on the quality of the enhancement available. When noisy adaptation material is used the best WER is obtained for $\alpha = 0.1$, meaning that less information is kept from the clean acoustic models. Conversely, as the quality of the enhanced signals increases, reducing the mismatch with respect to the clean models, larger values of α result to be optimal.

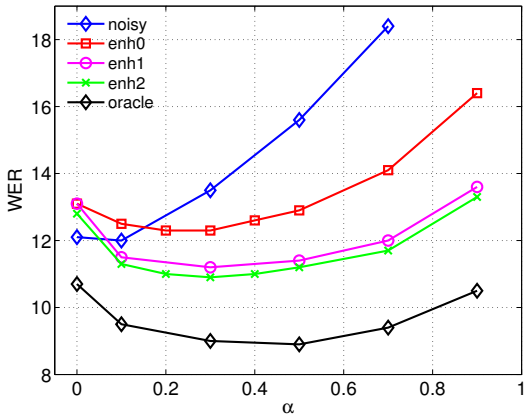

Figure 2: *WER on the full AURORA-4 test set as a function of the adaptation parameter α using the noisy signals, the 3 enhancements (enh0, enh1, enh2) and the upper-bound (oracle, trained on both development and test sets).*

Note also that for $\alpha = 0$ using the noisy signals (i.e. the parameters of the DNN are updated completely trusting the adaptation set) would results in better performance than performing filter-bank enhancement. However, properly setting the adaptation coefficient leads to a considerable performance improvement. This is in line with our assumption that the amount of adaptation to be used is related to the distance between the adaptation material and the original training signals.

Table 2: *WERs combining speech enhancement and model adaptation with constant and adaptive α on the full AURORA-4 test set and on the 4 disaggregated conditions: A=clean close talk; B=noisy close talk; C=clean far talk; D=noisy far talk.*

Enhancement	Conditions				
	A	B	C	D	Total
noisy (α=0.1)	4.73	8.09	9.06	17.51	11.96
enh0 (α=0.3)	4.24	9.02	7.72	17.74	12.32
enh1 (α=0.3)	3.98	7.93	6.67	16.41	11.19
enh2 (α=0.3)	3.92	7.72	6.48	15.94	10.88
enh2 ($\hat{\alpha}$)	3.87	7.64	6.07	15.28	**10.53**

Finally, Table 2 reports the results achieved when employing an adaptive (i.e. sentence dependent) regularization term in eq. 2, obtained using eq. 3, in comparison with the performance obtained with the best a posteriori constant α. The adaptive regularization further reduces the WER also considering the disaggregated results in all the 4 AURORA-4 noisy conditions.

To conclude this analysis, Figure 3 shows the distribution of the $\hat{\alpha}_k$ obtained when using *enh0* and *enh2*. Note that the distribution related to *enh2* is more flat than that of *enh0*, which performs more poorly. Actually, most of $\hat{\alpha}_k$ resulting by processing the adaptation set by means of *enh0* are close to 0, meaning that filter-bank have not been enhanced and the acoustic model has to be adapted more than if *enh2* was used.

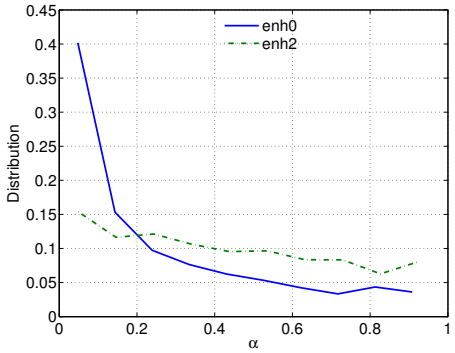

Figure 3: *$\hat{\alpha}_k$ distribution with two enhancement methods, enh0 and enh2, characterized by different de-noising performance.*

6. Conclusions

In this paper we have presented an approach to optimize the acoustic model adaptation process in presence of a speech enhancement component, based on a measure of the quality of the enhanced signals. The technique estimates specific regularization coefficients that are associated to each sentence of the adaptation set and are derived from the mean square error provided by the DNN-based enhancement component and learned simultaneously in the training process. In the forthcoming adaptation procedure these coefficients modulate the employed regularization term based on KLD and tend to optimize the resulting acoustic model.

The proposed method has been tested on AURORA-4, resulting in a considerable WER reduction in all conditions with respect to baselines. Clearly the enhancement procedure needs a stereo corpus that presents pairs of noisy and clean signals. This justifies the choice of AURORA-4 as experimental framework. Nevertheless, the obtained results show that it is possible to drive DNN adaptation exploiting information coming from the enhancement component, outperforming acoustic model retraining. A comparison with alternative approaches, such as joint training, is important to better asses the efficacy of the proposed approach and will be addressed in the future.

Future works will also address the use of more articulated distortion metrics, possibly based on the KLD between the original and enhanced speech distributions, introducing a theoretically-founded model. An alternative could be the use of a data-driven model for controlling the adaptation process as function of level of the enhancement, in place of the empirically derived sigmoid function, towards improving robustness against variable acoustic conditions, recognition quality and task complexity. Finally, another interesting topic is to study the impact of errors in the transcriptions since in this work we assume to perform adaptation on stereo data with exact supervision.

7. References

[1] D. Falavigna, M. Matassoni, S. Jalalvand, M. Negri, and M. Turchi, "DNN adaptation by automatic quality estimation of ASR hypotheses," *Computer Speech & Language*, 2016.

[2] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7893–7897.

[3] M. Matassoni, D. Falavigna, and D. Giuliani, "DNN adaptation for recognition of children speech through automatic utterance selection," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 644–651.

[4] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance analysis of the aurora large vocabulary baseline system," in *12th European Signal Processing Conference*, 2004, pp. 553–556.

[5] M. Delcroix, K. Kinoshita, C. Yu, A. Ogawa, T. Yoshioka, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5270–5274.

[6] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *in Handbook of Neural Networks for Speech Processing, Shigeru Katagiri, Ed.* Artech House, 1998, pp. 541–541.

[7] F. Xie and D. V. Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994.

[8] S. Tamura, "An analysis of a noise reduction neural network," in *IEEE nternational Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 2001–2004 vol.3.

[9] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: an overview of challenge systems and outcomes," in *IEEE Workshop on Automatic Speech R ecognition and Understanding*, 2013, pp. 162–167.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, 2016.

[11] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, Nov. 2016.

[12] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *Interspeech*, 2013, pp. 3512–3516.

[13] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[14] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1759–1763.

[15] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," *CoRR*, vol. abs/1605.02427, 2016.

[16] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Communication*, vol. 68, no. C, pp. 97–106, Apr. 2015.

[17] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.

[18] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, I. Illina, and A. Liutkus, "Robust ASR using neural network based speech enhancement and feature simulation," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 482–489.

[19] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3709–3713.

[20] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[21] A. Maas, Q. V. Le, T. M. ONeil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Interspeech*, 2012.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[23] X. Li and J. Bilmes, "Regularized Adaptation of Discriminative Classifiers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[24] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[25] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 366–369.

[26] Z. Huang, J. Li, M. Siniscalchi, I. Chen, C. Weng, and C. Lee, "Feature Space Maximum a Posteriori Linear Regression for Adaptation of Deep Neural Networks," in *Interspeech*, 2014, pp. 2992–2996.

[27] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Linear Hidden Transformations for Adaptation of Hybrid ANN/HMM Models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[28] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Interspeech*, 1995.

[29] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System," in *Interspeech*, 1995.

[30] B. Li and K. Sim, "Comparison of Discriminative Input and Output Transformation for Speaker Adaptation in the Hybrid NN/HMM Systems," in *Proc. of Interspeech*, 2010, pp. 526–529.

[31] S. Siniscalchi, J. Li, and C. Lee, "Hermitian Polynomial for Speaker Adaptation of Connectionist Speech Recognition Systems," *IEEE Trans. on Audio Speech and Language Processing*, vol. 21, no. 10, pp. 2152–2161, 2013.

[32] Y. Miao and F. Metze, "Improving low-resource CD-DNNHMM using dropout and multilingual DNN training," in *Interspeech*, 2013, pp. 2237–2241.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselỳ, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[34] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Interspeech*, 2011, pp. 2345–2349.

[35] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[36] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 285–290.