Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?

Marc Delcroix, Yotaro Kubo, Tomohiro Nakatani, Atsushi Nakamura

NTT Communication Science Laboratories, NTT corporation,
2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan
{marc.delcroix,kubo.yotaro,nakatani.tomohiro,nakamura.atsushi}@lab.ntt.co.jp

Abstract

Using deep neural networks (DNNs) for automatic speech recognition (ASR) has recently attracted much attention due to the large performance improvement they provide for a variety of tasks. DNNs are known to be robust to overfitting and to be able to remove speaker variability. Another important cause of variability in speech is the presence of noise. A lot of research has been undertaken on noise robust ASR including front-end and back-end approaches. However most approaches have been developed or evaluated on traditional ASR systems based on Gaussian mixture models (GMMs). The question we try to address in this paper is whether conventional noise robust approaches can still be competitive when using recent DNN-based ASR systems. To this end, we compare experimentally the performance of DNN-based ASR systems in a distant speech recognition task, for DNNs trained with noise-free, noisy and enhanced speech. We confirm that DNNs are powerful when the training and testing conditions are well matched. However, the performance degrades in the presence of noise. The use of a speech enhancement pre-processor to reduce the noise variability significantly improves performance with performance improvement comparable to that observed with conventional GMM-based ASR systems.

Index Terms: robust speech recognition, deep neural network, speech enhancement, multi-condition training

1. Introduction

Recently there has been an increased interest in automatic speech recognition (ASR) systems that employ deep neural networks (DNNs) as acoustic models [1, 2]. Indeed, DNNs perform considerably better than conventional ASR systems based on Gaussian mixture models (GMMs) at various tasks [2]. The superiority of DNNs over GMMs can be explained by their structure that can model complex data in an efficient way, and a recently developed training procedure that prevents overfitting [3, 4]. Moreover, it was recently shown that DNNs can mitigate speaker variability, which has been a difficult issue for conventional GMM-based ASR systems [5, 6].

Another important cause of speech variability is the presence of noise or reverberation in the recordings. Noise is responsible for a mismatch between the observed speech and the acoustic model used for recognition. This mismatch severely degrades recognition performance, especially in the case of distant speech recognition [7]. There has been much research on noise robust ASR including acoustic model compensation/adaptation (back-end approaches) and speech/feature enhancement (front-end approaches) [8].

Back-end approaches try to transform the acoustic model to noisy speech using such techniques as parallel model combination (PMC) [9] or a vector Taylor series (VTS) [10, 11]. These approaches directly compensate the acoustic models and therefore are very successful in improving recognition accuracy. However, they largely rely on the use of conventional GMM-based acoustic models. Moreover, it is difficult to use such approaches when the noise characteristics change rapidly.

Front-end methods attempt to reduce noise/reverberation on speech or features before recognition. Many speech enhancement techniques have been designed for non-stationary distortions [12, 13, 14, 15, 16]. Moreover, as front-end approaches do not assume a particular ASR system, they are not limited to conventional GMM-based acoustic models. However, most speech enhancement approaches cannot completely remove noise and they introduce processing artifacts that cause inevitably some mismatch between the enhanced speech and the acoustic models. Several approaches have been investigated with a view to mitigating this mismatch with some success. These approaches include training the acoustic model on enhanced speech [17] (which can be considered as a simple form of feature-space noise adaptive training) or adapting it to the enhanced speech [18]. However, front-end approaches have been tested mostly with conventional ASR systems.

It is important to investigate how the knowledge accumulated in noise robust ASR for conventional GMM-based systems can be used with recent DNN-based systems. For example, it was shown in [5] that the effect of a conventional speaker compensation technique (vocal tract length normalization (VTLN)) vanishes when using DNNs. Therefore, in a similar way, *we should confirm whether front-end speech enhancement techniques can still provide a significant improvement when using DNNs for acoustic modeling*. It is known that neural networks can be made more robust by performing regularization during their training, which is achieved by the explicit [19] or implicit [2] introduction of noise. For example, the RBM training procedure adds noise implicitly by using stochastic binary hidden units that is known to act as a strong regularizer [2, 3, 4]. However, the characteristics of noise introduced during training may be very different from that of noise observed during testing, and therefore have a limited effect on noise robustness. In [19] it was argued that adding realistic noise during the training of a neural network could greatly improve performance by making the network insensitive to distortion and increasing the number of training samples. However, the results were obtained for relatively stationary noise that could be relatively well learned by the neural network.

In this paper we investigate the use of DNNs for noise robust speech recognition in the presence of rapidly changing noise. As in [19] we experiment with the effect of adding real noise during the training process. Moreover, we investigate the effect of reducing the noise variability using a speech enhance-

ment pre-processor. We confirm experimentally that DNN-based ASR systems provide a large improvement in recognition compared with conventional GMM-based systems when the training and testing conditions match relatively well. However, the performance deteriorates greatly in the presence of rapidly changing noise, even when we train DNNs on noisy speech. Reducing the noise variability with a speech enhancement pre-processor greatly improves the performance and tends to simplify training.

The organization of this paper is as follows. In Section 2 we review DNN-based ASR. In Section 3 we briefly describe the speech enhancement algorithm we used for our experiments. In Section 4 we present our experimental results. Finally Section 5 concludes the paper.

2. DNN-based acoustic modeling

First, let us briefly recall the principles of DNN-based ASR. A DNN is a type of neural network that has several hidden layers between its input and output layers. Such a network is known to be able to achieve an efficient representation of data that enables the creation of models with high representational power [4, 20]. Moreover, DNNs can be trained without overfitting using a recently developed training procedure [4, 3].

DNNs were recently employed to replace the GMMs conventionally used to model the hidden Markov model (HMM) state emission probabilities of acoustic models [1]. For speech recognition, a DNN is usually trained to output HMM state posterior probabilities given several frames of observed speech features as input. During decoding, the output HMM state posteriors are converted into a scaled likelihood to be used with conventional Viterbi or Forward-Backward algorithms.

The training procedure of the DNNs is a two stage process that is described in [1], and that consists of the unsupervised pre-training of a generative model, followed by a supervised discriminative fine tuning. For the pre-training, restricted Boltzmann machines (RBMs) are trained for each layer using the contrastive divergence algorithm [4, 3]. These RBMs are then stacked to form an initial value for the DNN. The initial DNN is then finely tuned using back-propagation. For the fine tuning step, the training data are force-aligned to the HMM states using a conventional GMM-based ASR system.

3. DOLPHIN speech enhancement

In this work, we use the recently proposed dominance based locational and power-spectral characteristics integration (DOLPHIN) algorithm as a speech enhancement pre-processor. The DOLPHIN algorithm is described in detail in [15, 21, 22, 23]. Here we only summarize the main characteristics of the algorithm.

DOLPHIN extends a conventional blind speech separation algorithm based on spatial information [24] to separate a target speech signal from noise by using both spatial and spectral cues in an integrated framework.

The spatial features used within DOLPHIN are extracted from multi-microphones (in the following we consider 2 microphones), and represent location information about the sound sources. These features are used with spatial models of speech and noise to determine a dominant source index, i.e. whether speech or noise is dominant for each time-frequency bin. In addition, DOLPHIN uses spectral features and spectral models of speech and noise to recover the entire spectral shape of speech, including both dominant spectral components and non-dominant spectral components that may be masked by the other sources. The spatial and spectral models are integrated within the DOLPHIN framework through the dominance source index, which can be estimated with the EM algorithm. After determining the dominant source index, speech enhancement is performed using MMSE.

DOLPHIN benefits from both spatial-based and spectral-based speech separation approaches, i.e. it can achieve high speech/noise separation with high-quality enhanced speech. Consequently, DOLPHIN was shown to provide a great improvement in terms of recognition accuracy. It was one of the components of the system we developed for the CHiME challenge, which achieved the best performance among the challenge participants [25]. We have also demonstrated the potential of DOLPHIN for more complex recognition tasks such as meeting recognition [22]. Consequently DOLPHIN can be considered a state-of-the art approach for dealing with rapidly changing noise conditions. Here we use the most recent and powerful version of DOLPHIN, which is based on spectral models using MFCCs [21].

So far, we have used DOLPHIN with conventional GMM-based ASR systems, and we observed a great improvement in performance. It is however important to confirm that a similar improvement could be obtained when using state-of-the-art DNN-based systems.

4. Experiments

We perform investigations using the the PASCAL 'CHiME' speech separation and recognition challenge task [25]. The CHiME challenge task involves command recognition in the presence of living room noise recorded with two distant microphones. The difficulty of the task comes from the challenging noise conditions that are highly non stationary and includes a large variety of noise sources including children's voices, vacuum cleaners, televisions or music. The variability of the noise makes it challenging to model properly. Consequently it is interesting to determine whether DNNs can cover such complex noise conditions.

4.1. Settings

The CHiME task consists of 6-word commands spoken by 34 English speakers. The commands are corrupted by background noise that was collected in a real living room. The recognition target comprises two key words consisting of a letter followed by a digit, which are included in the command. The training data consist of utterances spoken by the same 34 speakers and 6 hours of background noise data. The training utterances do not include noise (i.e. they are noise free) but they are corrupted by reverberation (the distance between the microphones and the speaker is 2 meters). The test data consist of a development set and an evaluation set that both include 600 reverberant utterances at 6 different SNRs ranging from -6 to 9 dB. The development set also includes the corresponding noise-free reverberant utterances. Note that the training data set and the test data sets all consist of reverberant speech for the same room (reverberation time of 300 msec.) but with different speaker positions and room configurations (doors open/closed ...) and therefore with different reverberant characteristics. A detailed description of the CHiME task can be found in [25]. In the following, all results are given in terms of keyword accuracy evaluated using the tools provided by the CHiME challenge organizers [26], and averaged over the 6 noise conditions.

We consider three types of training data sets used to train acoustic models,

- A noise-free data set, which consists of 17,000 training utterances (approximately 10 hours of speech).
- A noisy speech data set, which is artificially generated by adding randomly selected noise portions of the available background noise data scaled to achieve SNRs ranging from -6 to 9 dB. This data set is 42 times larger than the noise-free data set [23].
- An enhanced speech data set, which consists of the noisy speech data set processed with the DOLPHIN speech enhancement algorithm.

4.1.1. GMM-based system

For the baseline GMM-based ASR system we employed the speech recognizer platform SOLON [27], which was developed at NTT Communication Science Laboratories, to train the acoustic models and perform decoding. The acoustic models consisted of conventional left-to-right word HMMs with a total of 250 states each modeled by a GMM consisting of 7 Gaussian components for the noise-free model. For the noisy and enhanced models, we used 20 Gaussians per HMM state to cover the variability of the noisy and enhanced speech data. We trained speaker independent acoustic models using ML and the dMMI discriminative criterion [28] for each of the three training data sets (note that in this experiment we used dMMI settings so that it becomes equivalent to MPE [29]). The feature vectors consisted of 39 coefficients, 12 MFCCs, the log energy, and the delta and acceleration coefficients. For the feature extraction, we used a frame length and frame shift of 25 msec. and 10 msec., respectively. The features were processed with cepstral mean normalization (CMN) performed per utterance. These feature extraction settings correspond to those provided by the CHiME challenge baseline system [26].

4.1.2. DNN-based system

We trained the DNNs according to the recipe given in [1]. We used the same DNN structure that has been successful for many recognition tasks [2]. The DNN input features consist of 11 frames (5 left and 5 right context) of the MFFC based features. There were 2048 units per hidden layer and the output layer consisted of the 250 units corresponding to the 250 HMM states.

For pre-training, we generated two types of RBMs, one trained with noise-free data and one trained with noisy data. The latter was used for fine tuning both the noisy-speech and enhanced speech DNNs. Because the RBM training is complex in terms of computation, we could not generate an RBM trained directly on enhanced speech owing to time limitations.

In the following, we first discuss the results on the development set for the three training data sets and then briefly summarize the results we obtained for the evaluation set.

4.2. Development set

4.2.1. Noise-free speech model

Table 1 shows the performance we obtained when we trained the acoustic model with noise-free speech. We observe that performance is significantly better with the DNN-based system than with the discriminatively trained system (i.e. GMM-dMMI) for matched noise-free testing data, with a relative keyword error rate reduction of about 40%.

Interestingly, the performance obtained with the speaker independent DNN is very similar to that obtained with a speaker dependent discriminatively trained GMM-based system (96.8%). This observation seems to confirm that DNN can suppress speaker variability [6].

Table 1: Keyword accuracy for the development set for *acoustic models trained with noise-free speech*, and tested with noise-free, noisy and enhanced speech. The top two rows show the performance obtained with a GMM based acoustic model trained with ML and dMMI.

Model / Test	Noise-free / Noise-free	Noise-free / Noisy	Noise-free / Enhanced
GMM - ML	92.0	57.7	77.6
GMM - dMMI	94.7	59.6	80.4
DNN (1 layer)	96.3	59.8	82.9
DNN (2 layers)	96.8	57.5	81.8
DNN (3 layers)	96.5	56.3	80.9

As shown in the second column of Table 1 (i.e. Noisy), the performance degrades greatly when recognizing noisy speech. In this case, the performance of DNN and GMM based systems are equivalent. This confirms that DNNs are not robust as regards large mismatches between the training and testing conditions.

When using a speech enhancement pre-processor as shown in the third column of Table 1 (i.e. Enhanced), the performance can be improved by a 55% relative reduction in the keyword error rate, which is comparable to the improvement observed with a conventional GMM-based system. In this case, the DNN-based system provides a relative performance improvement of about 13 % compared with the GMM-based system when using only a single hidden layer, but the performance improvement vanishes when more layers are used suggesting overfitting to noise-free speech.

4.2.2. Noisy speech model

Table 2 shows the keyword accuracy for acoustic models trained with noisy and enhanced speech. The first column of Table 2 (i.e. Noisy/Noisy) shows the performance obtained for acoustic models trained and tested using noisy speech. For conciseness, we only show the performance when using 1 hidden layer and 6 hidden layers (i.e. Best) that was the best configuration. As expected the performance improves greatly compared with using the noise-free acoustic model. Moreover, the performance gap between DNN and GMM-based systems becomes much more significant (a relative keyword error rate reduction of about 30 %). Note that the DNN system trained and tested on noisy speech outperforms that DNN trained on noise-free speech and tested on enhanced speech. However, there remains a large performance gap compared with the noise-free testing conditions shown in Table 1.

One problem that may limit performance is the poor HMM state alignment of the noisy speech used for the DNN fine tuning. We therefore also tested the performance obtained when using noise-free speech for the HMM state alignment. Figure 1 compares the performance obtained on the development set as a function of the number of hidden layers when using noisy or noise-free speech for the HMM state alignment. We observed a consistent improvement when using noise-free speech for HMM state alignment (a relative keyword error rate reduction of about 5%). However, the remaining performance gap compared with the noise-free case in Table 1 indicates that it is difficult for DNNs to learn to discriminate speech from noise when noise has rapidly changing characteristics.

4.2.3. Enhanced speech model

We have seen that the performance improves greatly when we use a speech enhancement pre-processor. However, speech en-

Table 2: Keyword accuracy for the development set for *acoustic models trained and tested with noisy and enhanced speech*. The last two rows show the performance obtained with a DNN with 1 layer and with the best DNN configuration (i.e. 6 layers for Noisy and 3 layers for Enhanced).

Model/Test	Noisy/Noisy	Enhanced/Enhanced
GMM - ML	73.8	84.4
GMM - dMMI	78.3	88.5
DNN (1 layer)	82.3	89.7
DNN (Best)	85.7	90.9

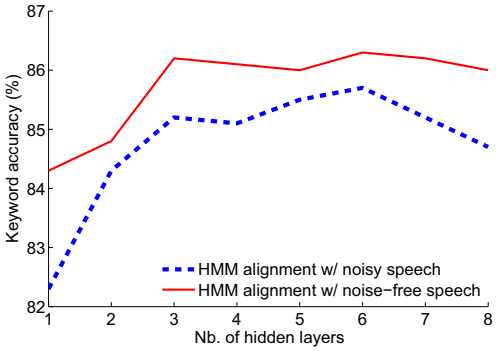

Figure 1: Keyword accuracy for the development set as a function of the number of hidden layers when using noisy speech for training and testing. The solid and dashed lines show the performance obtained when using noisy and noise-free speech for the HMM state alignment, respectively.

hancement pre-processing introduces a mismatch caused by the remaining noise and the processing artifacts. This may explain why the DNNs trained and tested on noisy speech outperformed those trained under noise-free conditions and tested with enhanced speech. Feature space noise adaptive training (i.e. training the acoustic models on enhanced speech) is a well known technique for mitigating the mismatch introduced by the speech enhancement pre-processor [11, 8] that we test here for DNN-based ASR systems. The second column of Table 2 (i.e. Enhanced/Enhanced) shows the performance obtained with models trained and tested with enhanced speech. In this case, the best performance was obtained with 3 hidden layers. Although the performance improvement introduced by training with enhanced speech is slightly less for the DNN-based system than for the GMM-based system, it is still large (33% relative keyword error rate reduction). Note that training the RBM was the most time consuming stage of the training process. Owing to time constraints, we used the RBMs trained with the noisy speech as the initial value for the fine-tuning. This may not be optimal, and we suspect that better performance could be obtained with RBMs directly trained with enhanced speech. As for the noisy speech experiment, we also undertook an investigation using noise-free speech for the HMM state alignment. However, here we did not observe the same improvement as in the noisy speech case.

Comparing the results in Table 2, we observe that reducing the variability introduced by noise using a speech enhancement pre-processor can outperform a system trained on noisy speech by a large margin. This is probably partly due to the fact that speech enhancement has access to spatial information, which was not available to the DNN. In future work we will investigate approaches designed to introduce spatial information directly into the DNN.

Table 3: Keyword accuracy for models trained with noisy speech for the evaluation set.

Model/Test	Noisy/Noisy	Enhanced/Enhanced
GMM - ML	73.2	85.5
GMM - dMMI	77.7	89.2
DNN	85.1	92.1
DNN (noise-free al.)	85.8	-

We also observed that using a speech enhancement pre-processor speeds up DNN training. Indeed, the speech enhancement pre-processor reduces the noise variability and consequently the model becomes easier to train. Accordingly, in our experiment, we observed that using enhanced speech required from about 12 to 47 % fewer iterations for the fine tuning to converge than when using noisy speech.

4.3. Evaluation set

Finally, Table 3 shows the results we obtained for the evaluation set of the CHiME challenge, for the models trained and tested with noisy and enhanced speech. We used the model configurations (i.e. number of hidden layers) that gave the best results for the development set. We observed the same tendency with the results as with the development set (Table 3 does not include results for noise-free speech because the CHiME challenge task does not include noise-free test data for the evaluation set). Note that these results are lower than the best results we obtained for the task [23], which was a 94.2% average keyword accuracy. There are several major differences that explain the poorer performance. The system proposed in [23] used speaker dependent acoustic models, performed unsupervised adaptation and combined the recognition results of several recognizer outputs obtained with different speech enhancement pre-processors. By integrating such techniques or their equivalent for DNNs, e.g. [5, 30], we expect the performance of the DNN-based system to be further improved.

5. Conclusions

We have discussed the robustness of DNN-based acoustic models to rapidly changing noise conditions. We confirmed that DNNs can provide significantly improved performance compared with conventional GMM-based ASR systems when the training and testing conditions match well. However, the performance degrades greatly in the presence of rapidly changing noise. We investigated two approaches for mitigating the mismatch, i.e. training the DNNs on realistic noisy data and using a speech enhancement pre-processor to reduce noise. We have shown experimentally that using a speech enhancement pre-processor can greatly improve performance in particular when the DNN can be trained with speech processed with the same speech enhancement pre-processor. Moreover, the improvement is comparable to that observed with conventional GMM-based systems. These results indicate that well designed speech/feature enhancement pre-processors can benefit DNN-based ASR systems.

Future work, will include investigations of approaches for performing denoising within a DNN framework such as [31, 32, 33] and to incorporate spatial information directly into the DNNs. Moreover, we will investigate the use of filter-bank features for noise robust DNNs as they have been shown to improve speech modeling ability [6].

6. References

[1] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 20, no. 1, pp. 14–22, 2012.

[2] 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: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] G. Hinton, "A practical guide to training restricted Boltzmann machines," Tech. Rep., 2010.

[4] Y. Bengio, *Learning deep architectures for AI*. Hanover, MA, USA: Now Publishers Inc., 2009.

[5] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2011, pp. 24–29.

[6] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, 2012, pp. 4273–4276.

[7] M. Woelfel and J. McDonough, *Distant speech recognition*. Chichester, UK: John Wiley and Sons Ltd, 2009.

[8] T. Virtanen, R. Singh, and B. Raj, *Technique for noise robustness in automatic speech recognition*. Chichester, UK: John Wiley and Sons Ltd, 2012.

[9] M. J. F. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[10] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, pp. 39–49, 1998.

[11] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 869–872.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing,*, vol. 9, no. 5, pp. 504–512, 2001.

[13] S. Rennie, J. Hershey, and P. Olsen, "Hierarchical variational loopy belief propagation for multi-talker speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2009, pp. 176–181.

[14] T. Nakatani, S. Araki, T. Yoshioka, and M. Fujimoto, "Multichannel source separation based on source location cue with log-spectral shaping by hidden Markov source model," in *Proc. Interspeech*, 2010, pp. 2766–2769.

[15] T. Nakatani, S. Araki, M. Delcroix, T. Yoshioka, and M. Fujimoto, "Reduction of highly nonstationary ambient noise by integrating spectral and locational characteristics of speech and noise for robust ASR," in *Proc. Interspeech*, 2011, pp. 1785–1788.

[16] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1692–1707, 2010.

[17] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 806–809.

[18] M. Delcroix, T. Nakatani, and S. Watanabe, "Combined static and dynamic variance adaptation for efficient interconnection of speech enhancement pre-processor with speech recognizer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4073–4076.

[19] P. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. International Conference on Document Analysis and Recognition,*, 2003, pp. 958–963.

[20] N. L. Roux and Y. Bengio, "Representational power of restricted Boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[21] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "LogMax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4029–4032.

[22] ——, "Dominance based integration of spatial and spectral features for speech enhancement," *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[23] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, no. 3, pp. 851 – 873, 2013.

[24] 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, no. 3, pp. 516–527, 2010.

[25] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621 – 633, 2013.

[26] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME peech separation and recognition challenge," *http://www.dcs.shef.ac.uk/spandh/chime/challenge.html*, cited April 24 2012.

[27] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2006.

[28] E. McDermott, S. Watanabe, and A. Nakamura, "Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training," in *Proc. Interspeech*, 2009, pp. 224–227.

[29] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. 105–108.

[30] 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 *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 366–369.

[31] Y. Tang, R. Salakhutdinov, and G. E. Hinton, "Robust Boltzmann machines for recognition and denoising," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2264–2271.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. international conference on Machine learning*, 2008, pp. 1096–1103.

[33] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, 2012.