Large-scale, sequence-discriminative, joint adaptive training for masking-based robust ASR

Arun Narayanan, Ananya Misra, Kean Chin

Google Inc.

{arunnt,amisra,kkchin}@google.com

Abstract

Recently, it was shown that the performance of supervised time-frequency masking based robust automatic speech recognition techniques can be improved by training them jointly with the acoustic model [1]. The system in [1], termed deep neural network based joint adaptive training, used fully-connected feed-forward deep neural networks for estimating time-frequency masks and for acoustic modeling; stacked log mel spectra was used as features and training minimized cross entropy loss. In this work, we extend such jointly trained systems in several ways. First, we use recurrent neural networks based on long short-term memory (LSTM) units – this allows the use of unstacked features, simplifying joint optimization. Next, we use a sequence discriminative training criterion for optimizing parameters. Finally, we conduct experiments on large scale data and show that joint adaptive training can provide gains over a strong baseline. Systematic evaluations on noisy voice-search data show relative improvements ranging from 2% at 15 dB to 5.4% at -5 dB over a sequence discriminative, multi-condition trained LSTM acoustic model.

Index Terms: automatic speech recognition, noise robustness, joint adaptive training, deep neural network, LSTM

1. Introduction

With the adoption of artificial neural network (ANN) based acoustic models (AMs), automatic speech recognition (ASR) has made great strides and is gaining acceptance as a viable mode for communicating with our devices [2]. The current ASR systems work well in relatively clean conditions; the focus now is on reducing error rates in real noisy environments. The performance of even the best ASR systems in noisy conditions is much worse than in clean. The reasons are similar as with Gaussian mixture model (GMM) based AMs – the mismatch between training and test data. But contrary to GMM AMs, when trained using large-scale, multi-condition training (MTR) data, ANN AMs have shown much better performance in noise [3]. This paper focuses on improving performance of such MTR AMs in matched and unmatched noisy test conditions.

Supervised time-frequency (T-F) masking based feature enhancement algorithms have been shown to improve performance of MTR ANN AMs [1, 4]. Such algorithms use ANNs to estimate an ideal ratio mask (IRM) directly from noisy signals. The IRM, which is defined for a T-F representation of speech like a mel spectrogram, identifies the ratio of speech energy with respect to the mixture energy at each T-F bin [5]. Typically, the estimated ratio masks are used to remove noise via point-wise multiplication with the noisy spectrogram. The enhanced spectrogram is then used as input to the AM. While this shows improvement in extremely noisy conditions, performance can be improved using mask-based speech and noise estimates as additional features, and by training the mask estimator jointly with the acoustic model [1]. But improvements using such techniques have only been demonstrated on medium-large vocabulary tasks (CHiME-2 [6]) with around 15 hours of training data [1]. Moreover, typical systems use fully-connected feed-forward deep neural network-based (FFDNN) AMs. Small scale data and FFDNN AMs require the use of stacked features augmented with delta components for obtaining good performance, which in turn makes joint optimization structurally complicated. Furthermore, sequence discriminative training, which has now been shown to consistently improve performance of ANN-based AMs, has not been explored in the context of joint optimization. This study primarily addresses these shortcomings and extends these early works.

The rest of the paper is organized as follows. In Section 2, we provide additional context for this work. Section 3 provides system description. Experimental setup, and detailed results and analysis are presented in Section 4. We conclude with a discussion in Section 5.

2. Relationship to prior work

Time-frequency masking (TFM) has been widely studied in the context of speech enhancement [7, 8], speaker [9] and speech recognition [10]. While several alternatives exist for estimating TFMs, the current best systems estimate ratio or soft valued masks using ANNs trained with supervision. Most systems use FFDNNs, but recurrent neural networks (RNNs), specifically LSTM-RNNs, have also been shown to work well [4, 11]. Note that, although we use LSTMs, the mask estimator in our system is optimized using ASR criteria, unlike earlier work.

Although traditional enhancement techniques show large gains in performance when using ANN AMs trained in clean conditions, AMs trained using MTR data have been shown to be inherently noise robust [12, 10]. This is especially true when only a monaural signal is available as input. More importantly, it was shown in these works that retraining the AM using enhanced data, sometimes referred to as joint training, does not improve performance, and in some cases deteriorates performance. This was attributed to the the limited variability in training data when using enhanced features [12]. The distortions introduced by frontend processing are also a factor. Noise aware training (NAT), wherein a noise estimate obtained by averaging the first few frames of a noisy signal is used as additional features, was proposed as an alternative [12]. Subsequent studies show that such a static estimate of noise does not help when using a well-trained AM [1]. Instead, a dynamic estimate of noise obtained per frame using an estimated mask was shown to be better. We refer to such methods as mask-based noise aware training (mNAT) in this work. To the best of our knowledge,

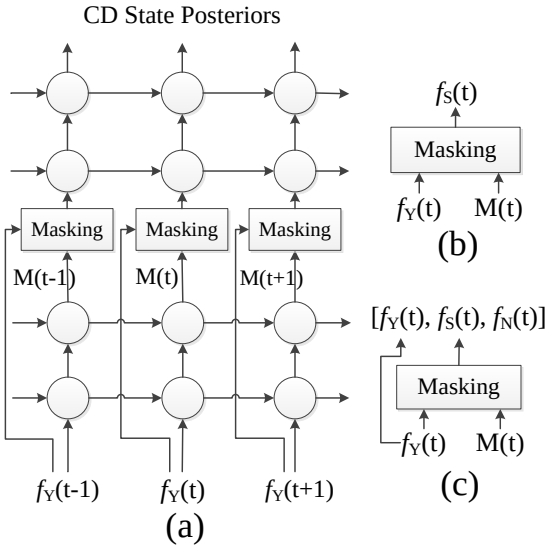

Figure 1: *Joint optimization using LSTM-RNNs: (a) The overall architecture. (b) Masking operation when using enhanced features as input to the AM. (c) Masking operation for mask-based noise aware training. The noise estimate is obtained using the inverted mask (1 - M(t)).*

none of these techniques have been evaluated using large-scale MTR-trained AMs.

For GMM AMs, joint (adaptive) training is used to jointly adapt parameters of the feature frontend and the acoustic models. Examples include speaker adaptive training [13], noise adaptive training [14], and VTS-based joint adaptive training [15]. In the context of ANN AMs, learning filterbank coefficients [16] and discriminative speaker codes [17] are examples of jointly optimizing the frontend. For noise robustness, FFDNN-based joint adaptive training (JAT) has been proposed for optimizing a masking-based enhancement frontend jointly with the AM using cross entropy loss [1]. Such a strategy can potentially be used for optimizing other enhancement techniques, like discriminative non-negative matrix factorization [18] and deep unfolding [19], as long as the frontend processing can be represented as a differentiable transformation of the noisy input features to enhanced clean features. As mentioned before, this study extends these early works.

3. System description

The overall architecture of the proposed system is shown in Fig. 1. 40-dimensional noisy log-mel spectrogram extracted from 25 msec windows with 10 msec hop size is used as input features. The same features are used as input to both the mask estimator and the (baseline) acoustic model. The features are normalized using the mean and the standard deviation of the training data. No utterance-level normalization is applied since it makes streaming recognition, which is one of the target applications of the proposed system, difficult.

Given log mel features, the mask estimation module uses LSTM-RNNs for mask estimation [20]. LSTMs are now widely used for modeling sequence data like speech [20] and language [21], and have been shown to consistently outperform FFDNNs. In addition to recurrent connections, LSTMs use input, output, and memory 'gates' to control the sequential flow of information, and cells that maintain internal states. This partly prevents vanishing gradients, making learning easier, and also allows them to learn long term dependencies better. LSTMs being recurrent allows us to use per-frame log mel features as input without any feature stacking or augmentation using delta components. This simplifies joint optimization, even though the use of LSTMs makes the individual models more complex. Further, it reduces delays in streaming recognition. As we will show in evaluations, for mask estimation LSTMs, perform at least as well as a well-trained FFDNN that uses stacked features. We use a 2-layer LSTM for mask estimation. Each layer uses 512 nodes and 256 projection nodes [20]. The input and output dimensionality are both 40 – the number of mel frequency bands.

The mask estimation module is trained to estimate the ideal ratio mask. Assuming that speech and noise are uncorrelated, the IRM is defined as the ratio of speech to mixture energy:

$$M(t,c) = \frac{\mathbf{X}(t,c)}{\mathbf{X}(t,c) + \mathbf{N}(t,c)}. \quad (1)$$

Here, $\mathbf{X}$ and $\mathbf{N}$ represent speech and noise mel spectrogram, respectively, t and c correspond to time and frequency indices. M represents the ideal ratio mask. Since we model real noisy signals that contain both background noise and reverberation, clean speech is no longer uncorrelated with the noise as it includes reverberation. Therefore, we treat reverberant speech as 'clean' speech, and everything else as noise. We do not directly address reverberation. With MTR AMs, we have found that mild reverberation does not severely affect performance.

Given an estimate of the IRM, $\widehat{M}$, and the original mean-variance normalized noisy features, $\mathbf{f}_Y$, the input to the acoustic model when using masked features directly is obtained as:

$$\widehat{\mathbf{f}}_X(t,c) = \mathbf{f}_Y(t,c) + (\alpha * \log(\max(\widehat{M}(t,c), \beta))/\sigma_c). \quad (2)$$

Here, $\widehat{\mathbf{f}}_X$ is an estimate of the normalized clean speech features. α is an exponential scaling factor for the mask to minimize speech distortion while allowing some residual noise, β is a spectral floor that prevents the log operation from misbehaving, and σ_c is the standard deviation of the c^{th} frequency channel of the noisy features. Masking directly in the normalized feature domain, as shown above, is made possible if we assume that the normalization parameters of the enhanced features are the same as the original noisy features. In practice, they are only similar; we have found that this does not affect performance since ANN parameters easily adapt to the values used for normalization as long as they are reasonably close to the true values. It should be pointed out that the masking operation is piecewise differentiable (piecewise because of the max operation), and can easily be incorporated into an ANN architecture.

For mask-based noise aware training, instead of using masked features, we stack the original noisy features, the speech estimate and a noise estimate ($\widehat{\mathbf{f}}_N$). $\widehat{\mathbf{f}}_N$ is obtained similar to Eq. 2, but using the inverted ratio mask, $1 - \widehat{M}$. mNAT has been shown to be better than standard NAT [22] for both Aurora4 [23] and CHiME2.

The acoustic model that we use is also based on LSTMs [20]. Specifically, it uses 2 LSTM layers, each with 832 nodes and 512 projection nodes. We use cross entropy followed by sequence discriminative training to optimize the model. For sequence training, we use state level minimum Bayes risk (sMBR), which optimizes the expected sequence error weighted by the state accuracy of the correct paths. sMBR has been shown to work well for ANN AMs [24, 25, 26] in prior work.

With the above formulation for obtaining features via masking, the full system can now be optimized to directly improve the ASR criterion of CE and sMBR. During joint optimization, the gradients of the ASR loss are used to update the weights of the mask estimator and the acoustic model. Initialization is necessary for joint optimization since it is harder for the joint model to learn an appropriate mask estimator from scratch using just the ASR losses. For our experiments, we first independently train the mask estimator. The acoustic model is then trained either using the noisy features or stacked features (noisy features, speech estimate, and noise estimate) by fixing the parameters of the mask estimator. The learned weights are used to initialize the corresponding joint models. During joint training, *only* the ASR losses are used for updating the weights of the full model. Joint optimization starts with CE training, and is followed by sequence training. Starting sequence training from the jointly optimized CE model was found to work better than either starting from the independently trained CE model, or the independently trained sMBR model.

4. Results

4.1. Experimental setup

All our acoustic models are trained using a 3 million utterance training set (~2000 hours) comprised of spoken English queries. The utterances are representative of Google search traffic, and are anonymized and hand-transcribed before use. The MTR set is created by mixing these relatively clean utterances with random noise segments collected from YouTube and daily life noisy environmental recordings. The signal-to-noise ratio (SNR) of the mixtures is set randomly to be between 5 dB and 25 dB. To model reverberant conditions, we use a room simulator based on the image model of reverberation [27]. We model moderate levels of reverberation, with T60 ranging form 0 to 400 msec. Speech and noise are assumed to originate from different locations. Note that since we only use monaural recordings, none of the presented systems use any spatial information implicitly or explicitly for enhancement. The evaluation set is similarly derived from a 30,000 utterance set (~30 hours). Noise is added artificially, similar to the training data, using the room simulator. For development, we mix this set with unseen segments of YouTube and daily life noises at similar SNR and reverberation settings. For the final evaluations, we mix these utterances with segments of unseen noise types taken from Aurora4 and NOISEX-92 [28]. SNRs are systematically varied from -5 dB to 15 dB at 5 dB intervals. Rooms modeled are similar, but contain more than one noise sources, unlike the training data. These differences ensure that our final evaluations focus on generalization properties of the systems considered. Only for pre-training the mask estimator, we use an alternative training set that consists of 800,000 separate utterances (~500 hours) created under similar noise and reverberation settings. All utterances in the training and evaluation sets are at 16 kHz.

The models are trained using asynchronous stochastic gradient descent (ASGD) using Google's distributed ANN training infrastructure [29]. For CE training we use ASGD, and for sMBR we use ASGD with AdaGrad. When using ASGD for joint optimization, we also clip gradients when their norm exceeds a preset maximum to prevent the log operation in Eq. 2 from back-propagating extremely large values. Note that AdaGrad automatically accounts for large gradients by scaling the learning rate down based on their moving average. The acoustic models use ~13k tied context dependent states.

Table 1: *Comparison of LSTM and FFDNN mask estimators.*

System	WER
CE baseline	18.52
+ FFDNN Masking	18.19
+ LSTM Masking	18.03

4.2. Development set results

We first compare performance obtained using LSTM and FFDNN mask estimators. The FFDNN mask estimator is trained using a context of 26 time frames, and has 4 layers, each with 512 nodes. Word error rates (WER) are shown in Tab. 1. α and β in Eq. 2 are set to 0.5 and 0.4, respectively. They are chosen via a grid search for optimal performance on this set. As shown, the FFDNN performs worse than the LSTM. In what follows, we only present results based on LSTM mask estimators.

WERs on the development set are shown in Tab. 2. The baselines include the independently-trained models: CE/sMBR baseline, independently-trained masking frontend (+ Masking), and mNAT. For CE training, denoising using a well-tuned spectral subtraction algorithm is also included as a baseline (CE w/ SS). The table also shows results after joint optimization of the direct masking (+ JAT) and the mNAT models (+ mJNAT). α and β are set to 0.5 and 0.01 for obtaining the speech estimate when performing joint optimization. A lower value for β is chosen since one of the goals of joint optimization is to achieve better noise suppression while preserving speech characteristics that are important for ASR. To obtain the noise estimate for mNAT and mJNAT, α is set to 1.0 and β to 0.01. Note that the mask is inverted before deriving the noise estimate; scaling and flooring, as in Eq. 2, is done after inversion. In mJNAT, the jointly estimated mask is used to recalculate both the speech and noise estimates. We also looked at variants where only the speech estimate is updated using the jointly estimated mask, with the noise estimate coming from the independently-trained mask as was suggested in earlier work [1]. But this performed slightly worse. It may be because with large training data, the ANN is able to adapt the mask appropriately for deriving both these estimates.

Comparing the CE results, masking and mNAT improves upon the noisy baseline by 0.5% absolute. JAT and mJNAT improve performance further by roughly 0.7 to 0.8% absolute, with mJNAT performing slightly better than JAT. In comparison, spectral subtraction fails to improve upon the CE baseline. sMBR improves the CE baseline by 3.4% absolute. Sequence training, although not directly addressing noise, seems to have a very significant effect as has already been observed in the CHiME-2 task [30]. Masking using independently-trained estimators fails to significantly improve this strong baseline. JAT improves performance by 0.3% absolute. mNAT performs comparably to JAT; mJNAT improves performance by another 0.2% absolute. Compared to the sMBR baseline, mJNAT improves WER by 0.6% absolute.

In Fig. 2 we show an example of the estimated masks before and after mJNAT. As shown, joint optimization uses the mask to identify segments of the spectrogram that are most speech-like and necessary for recognition, while aggressively suppressing noise. Interestingly, the jointly optimized masks are also able to correctly suppress noise in the original 'clean' recording (segments between 2.5 and 3 seconds). Comparing the jointly optimized mask before and after sMBR training, it

Table 2: *Comparison of various systems using CE- (left) and sMBR-trained (right) MTR acoustic models. SS stands for spectral subtraction.*

System	WER
CE baseline	18.52
+ Masking	18.03
+ JAT	17.30
mNAT	18.01
+ mJNAT	17.18
CE w/ SS	18.51

System	WER
sMBR baseline	15.15
+ Masking	15.13
+ JAT	14.81
mNAT	14.84
+ mJNAT	14.58

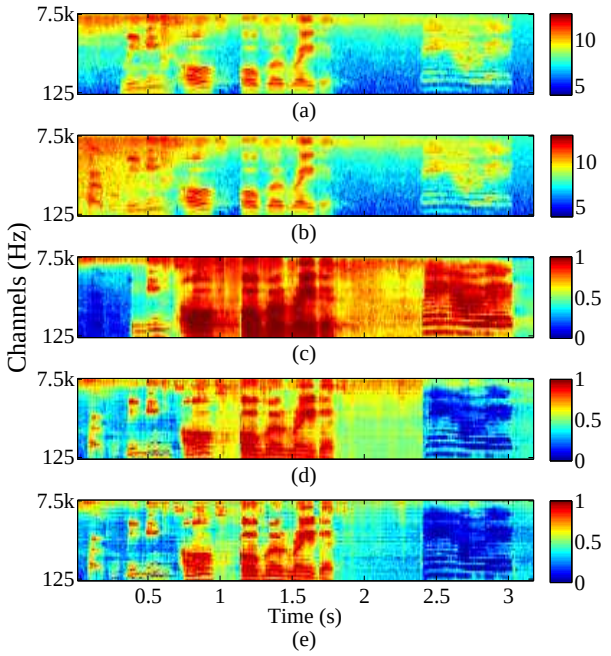

Figure 2: *Masks estimated before and after joint optimization: (a) The 'clean' log mel spectrogram. The clean signal being a real recording, already contains some noise. (b) The noisy mel spectrogram. (c) The independently estimated mask. (d) Estimated mask after mJNAT CE training. (e) Estimated mask after mJNAT sMBR.*

can be observed that sMBR training modifies the mask further for stronger noise suppression. It should be noted that the jointly optimized mask is not necessarily more similar to the IRM because the mask that is ideal for ASR is non-ideal for 'perfectly' resynthesizing the target. We believe that this makes joint optimization a more attractive strategy for training denoising frontends in conditions where the true clean speech is unavailable and only the correct transcription is given. This is quite common in large scale ASR tasks since obtaining real noisy signals with the correct transcription is easier than obtaining studio quality speech signals.

4.3. Evaluation set results

In this section we present results on the evaluation sets created with unseen noise types. The results in the previous section show that joint optimization typically performs better. Therefore, we only present the baseline and the results obtained after joint optimization. The results are shown in Tab. 3.

We obtain large improvements in performance over the baseline in low SNR conditions when using CE models. At -5 dB, JAT works better than mJNAT, and reduces WER by 6% absolute compared to the CE baseline. At 15 dB, mJNAT works better and reduces WER by 0.8% absolute. Compared to the stronger sequence-trained baseline, the overall relative improvements are lower. At -5 dB, JAT improves over the sMBR baseline by 2.2% absolute, and at 15 dB mJNAT improves by 0.3% absolute. As can be see, JAT works better than mJNAT at low SNRs. The training data for the models does not include a lot of low SNR signals, and it is likely that mJNAT is more sensitive to this due to the additional features it derives from the masks. Nonetheless, it is interesting to see that joint optimization continues to provide gains in unseen conditions over a strong baseline trained on large scale data.

Table 3: *Performance on the evaluation sets using CE- and sMBR-trained MTR AMs.*

System	WER				
	-5 dB	0 dB	5 dB	10 dB	15 dB
CE baseline	55.11	33.96	23.07	18.94	17.20
+ JAT	49.11	30.60	21.43	17.90	16.54
+ mJNAT	49.86	30.81	21.40	17.78	16.45
sMBR baseline	45.50	26.60	18.57	15.63	14.48
+ JAT	43.05	25.50	18.05	15.38	14.32
+ mJNAT	43.32	25.60	17.96	15.23	14.19

5. Discussions

In this paper we have shown that joint optimization continues to provide improvements in noisy conditions over a very strong sequence-trained acoustic model baseline, even when using large scale training data. Extending earlier work, we show that joint optimization can be simplified using LSTMs for mask estimation and acoustic modeling, which makes feature stacking and augmentation unnecessary. As with prior work, we have shown that mask-based noise aware training works better than direct masking in most cases, and that joint optimization provides further gains. The final system obtains relative improvements ranging from 5.4% at -5 dB to 2% at 15 dB.

In future work, we will explore the use of convolutional models in addition to LSTMs [31] in this framework. In initial experiments not reported in this work, we have observed that jointly optimized models learn characteristics that are complimentary to those learned by an acoustic model trained directly on noisy features: A simple averaging of posteriors from these two models seems to significantly improve WERs. Such model averaging techniques will be explored as part of future work. Finally, we will also explore ways of handling reverberation within this framework. While mild reverberation may not cause significant increase in WER given MTR-trained acoustic models, it is likely that a more systematic way of handling reverberation will prove useful in severely reverberant conditions.

6. Acknowledgements

The authors would like to thank Chanwoo Kim for the room simulator, Yuxuan Wang for the datasets used in initial experiments, and Richard Lyon, Andrew Senior, and Ron Weiss for useful discussions and feedback.

7. References

[1] A. Narayanan and D. L. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 92–101, 2015.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. W. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *Proceedings of the International Conference on Learning Representations*, 2013.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2015.

[5] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[6] E. Vincent, J. Barker, S. Watanabe, J. LeRoux, F. Nesta, and M. Matassoni, "The 2nd CHiME speech separation and recognition challenge," 2012. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task2.html

[7] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.

[8] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1581–1585.

[9] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[10] A. Narayanan and D. L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826–835, 2014.

[11] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings of IEEE GlobalSIP Symposium on Machine Learning Applications in Speech Processing*, 2014.

[12] M. L. Seltzer, D. Yu, and Y.-Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7398–7402.

[13] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings of the Fourth International Conference on Spoken Language*, vol. 2, 1996, pp. 1137–1140.

[14] O. Kalinli, M. L. Seltzer, and A. Acero, "Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3825–3828.

[15] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007, pp. 389–392.

[16] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 297–302.

[17] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7942–7946.

[18] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," *Proceedings of Interspeech*, 2014.

[19] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint arXiv:1409.2574*, 2014.

[20] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of Interspeech*, 2014.

[21] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling." in *Proceedings of Interspeech*, 2012.

[22] A. Narayanan and D. L. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 2523–2527.

[23] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evalutions," in *Proceedings of the European Conference on Speech Communication and Technology*, 2003, pp. 337–340.

[24] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3761–3764.

[25] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013.

[26] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," *Proceedings of Interspeech*, pp. 17–18, 2014.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[28] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[29] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[30] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," 2013, pp. 19–24.

[31] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2015.