Robust Speech Recognition using Long Short-Term Memory Recurrent Neural Networks for Hybrid Acoustic Modelling

Jürgen T. Geiger, Zixing Zhang, Felix Weninger, Björn Schuller[2] and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Munich, Germany
[2]also with Department of Computing, Imperial College London, London, U.K.
geiger@tum.de

Abstract

One method to achieve robust speech recognition in adverse conditions including noise and reverberation is to employ acoustic modelling techniques involving neural networks. Using long short-term memory (LSTM) recurrent neural networks proved to be efficient for this task in a setup for phoneme prediction in a multi-stream GMM-HMM framework. These networks exploit a self-learnt amount of temporal context, which makes them especially suited for a noisy speech recognition task. One shortcoming of this approach is the necessity of a GMM acoustic model in the multi-stream framework. Furthermore, potential modelling power of the network is lost when predicting phonemes, compared to the classical hybrid setup where the network predicts HMM states. In this work, we propose to use LSTM networks in a hybrid HMM setup, in order to overcome these drawbacks. Experiments are performed using the medium-vocabulary recognition track of the 2nd CHiME challenge, containing speech utterances in a reverberant and noisy environment. A comparison of different network topologies for phoneme or state prediction used either in the hybrid or double-stream setup shows that state prediction networks perform better than networks predicting phonemes, leading to state-of-the-art results for this database.

Index Terms: acoustic modelling, robust speech recognition, neural networks, long short-term memory

1. Introduction

In recent years, neural networks (NNs) re-gained popularity for acoustic modelling in speech recognition [1], although the underlying methods had already been developed years ago [2]. Due to increased available computing power it is now possible to train large networks. Especially the utilisation of several hidden layers (making it a deep network) increases the modelling power of the system. For these reasons, deep NN (DNN) acoustic models were shown to outperform the conventional approach of Gaussian mixture models (GMMs). The GMM acoustic model in a hidden Markov model (HMM) framework is replaced by the network, which, instead of the GMM, creates the HMM state likelihoods. This approach is also referred to as the hybrid NN/HMM acoustic modelling approach. Such systems proved to be very robust in adverse conditions due to their increased modelling power [3, 4].

In addition, recurrent neural networks (RNNs) using the long short-term memory (LSTM) architecture [5] have been successful for acoustic modelling. Using the LSTM topology, these networks can exploit a self-learnt amount of long-range temporal context. This ability is helpful to improve noise robustness, e. g. in cases where a portion of frames within a longer window is spectrally masked by noise. Previously, in the context of robust speech recognition, LSTM networks were mostly used in a double-stream HMM architecture, where they are combined with the GMM acoustic model. This approach was first proposed in [6] and uses LSTM networks for phoneme prediction. The predicted phoneme probabilities are then used for decoding, jointly together with the GMM. Until now, LSTM networks have rarely been applied as an acoustic model on their own, predicting HMM states and using the hybrid acoustic modelling approach. A hybrid system that employs LSTMs for HMM state prediction could make use of the LSTM topology to exploit long-range temporal context and of the modelling power of a large network to be able to accurately predict HMM states.

1.1. Contribution

In this work, we propose to use LSTM RNNs for acoustic modelling in the hybrid NN/HMM system architecture. We employ LSTM networks to predict HMM states, and use the network predictions for acoustic modelling. Previous work using LSTMs for acoustic modeling has operated with the LSTM mapping feature vector sequences to context-independent phonemes. In the present work, we expand the size of the output space of the LSTM network to include the set of context-dependent states. Experiments are performed using the database of the medium-vocabulary recognition track of the 2nd CHiME Speech Separation and Recognition Challenge [7]. This database contains speech recordings in a reverberant domestic environment with non-stationary noise sources. In the experimental validation, we compare state prediction networks (in the hybrid setup) to the previous approach of predicting phonemes and using them in a double-stream architecture. In addition, the double-stream architecture can be used to combine the two different LSTM-based acoustic models. The experimental results show that with LSTMs, state prediction networks outperform networks predicting phonemes.

1.2. Related work

In [1], a broad overview of the application of DNNs for acoustic modelling in various speech recognition tasks is given. A DNN is created by using more than one hidden layer in a feed-forward neural network. By combining multiple restricted Boltzmann machines in a stack, a multilayer model called deep belief net can be created [8]. The application of DNNs for noise robust speech recognition was explored in [3]. In [9], deep LSTM RNNs were used for speech recognition on their own, without the need of an HMM framework, and in [10], the LSTM topology was employed in a hybrid HMM setup. Several studies employed NNs in the tandem system setup, predicting phones

and using the predictions as additional input features for an HMM: in [11], RNNs were compared to other NN architectures. In [12], an LSTM-HMM tandem system was succesfully applied for large-vocabulary continuous speech recognition. Using phoneme prediction LSTM networks in the double-stream approach was first proposed in [6]. This system was succesfully employed in the small-vocabulary recognition task of the 1st CHiME Challenge [13, 14]. In the 2nd CHiME challenge, we used that approach in combination with the provided baseline GMM acoustic model [15], and later together with an advanced GMM system [16]. Multi-stream HMM systems were originally proposed to combine independent feature streams [17]. For example, in this way, GMMs can be fused with NNs [18].

A short overview of the CHiME challenge is given in Section 2. The employed HMM-GMM system is described in Section 3. Section 4 introduces LSTMs and their application for acoustic modelling. The experimental setup and results are presented in Section 5. Some concluding remarks are given in Section 6.

2. CHiME medium-vocabulary track

The evaluation database of the 2nd CHiME medium-vocabulary recognition track was constructed from the Wall Street Journal (WSJ0) 5k vocabulary read speech corpus. Using recordings from a domestic environment, the clean speech utterances are convolved with impulse responses (simulating reverberation) and mixed with noise backgrounds. In order to obtain different signal-to-noise ratios (SNRs), the reverberated test utterances are temporally placed in the background noise, leading to SNR values ranging from -6 to 9 dB (in steps of 3 dB). The training set includes 7 138 utterances from 83 speakers (14.5 hours), in clean, reverberated and noisy form. 409 utterances (per SNR value) from 10 speakers form the development set (4.5 hours in total), while 330 utterances (per SNR value) from 8 other speakers are used as a test set (4 hours in total). Recognition systems are evaluated using word error rate (WER) in % as a measure.

3. HMM system

As HMM backend for the hybrid system and for experiments with the GMM acoustic model, we use a system similar to the one described in [19], implemented with the Kaldi speech recognition toolkit [20]. The system uses context-dependent triphone models with 40 phonemes (including silence). Each model has three HMM states and in total, there are 1 936 different HMM states, and 15 000 Gaussian components. Models are trained with the maximum likelihood principle. In addition, Linear Discriminant Analysis (LDA) [21] and maximum likelihood linear transform (MLLT) [22] are employed for feature decorrelation. LDA is applied on stacked MFCC vectors (13 coefficients over nine consecutive frames), reducing the resulting 117 dimensional vector to 40 dimensions. One after another, the clean, reverberated, and noisy training data are used for training. Then, LDA and MLLT are applied, before running another set of training iterations with the noisy training data. While the recordings in the CHiME database are stereo, features are extracted from mono signals, which are obtained by averaging over the two channels. Considering the recording setup of the CHiME database (fixed speaker position in front of the microphone), averaging over the two channels corresponds to delay-and-sum beamforming. Note that in contrast to the best system setup in [19], we do not use speaker adaptive training in our system, since it would require an additional decoding pass and furthermore assumes speaker identities to be known, which can generally not be expected.

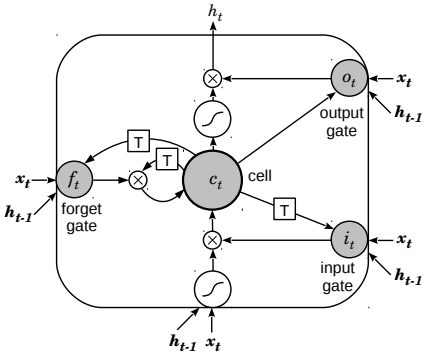

Figure 1: Long short-term memory block, containing a memory cell and the input, output and forget gates. T denotes a delay of one time step.

4. Acoustic modelling with neural networks

We compare two different methods for applying neural networks as an acoustic model within the HMM framework: predicting either phonemes or HMM states, both with LSTM networks.

4.1. Long Short-Term Memory recurrent neural networks

LSTM networks were first introduced in [5]. Compared to standard RNNs, LSTM RNNs are able to exploit a self-learnt amount of temporal context. The LSTM networks use so-called memory blocks instead of the conventional activation functions in the hidden layers. Each memory block consists of a memory cell and three gate units: the input gate, output gate, and forget gate, as depicted in Fig. 1. These gates control the behaviour of the memory block. The forget gate can reset the cell variable which leads to 'forgetting' the stored input $\boldsymbol{c}_t$, while the input and output gates are responsible for reading input from the feature vector $\boldsymbol{x}_t$ and writing output to $\boldsymbol{h}_t$, respectively. With this architecture, the network is capable of storing input over a longer period of time and thus exploiting a self-learnt amount of long-range temporal context. Furthermore, we use bidirectional RNNs [23]. Such networks process the input data in both directions with two separate hidden layers, exploiting context from both temporal directions [24]. The output of both hidden layers is then fed to the output layer. Additionally, the concept of using multiple hidden layers can also be applied here. Network training is performed using backpropagation through time, using the cross entropy as an error function. Our LSTM software is publicly available[1].

4.2. Acoustic modelling using phoneme predictions

In this approach, the network is trained to predict phonemes b, using a forced alignment of the training data (generated by the HMM system). The frame-wise phoneme predictions are converted into state likelihoods in the following way [6]: from the predicted phoneme probabilities $p(b_t|\boldsymbol{x}_t)$, frame-wise discrete phoneme predictions are obtained. These predictions are evaluated using the development set and the phoneme confusions are stored in a discrete probability table. Using a mapping from phonemes to HMM states leads to state likelihoods $p(\boldsymbol{x}_t|s_t)$. In

[1] https://sourceforge.net/p/currennt

Table 1: *Results for GMM acoustic modelling (WER [%] on the development set)*

SNR [dB]						
-6	-3	0	3	6	9	**Average**
64.3	55.6	47.3	40.0	36.0	29.7	45.5

this way, the HMMs do not directly model prediction probabilities of the LSTM, but instead, the confusions of the LSTM.

4.3. Acoustic modelling using state predictions

This method correponds to the classical hybrid system where the neural network is trained to predict HMM states s. The training targets are (as in the case of phoneme predictions) generated using a forced alignment of the HMM system. From the resulting posterior probabilities of the network $p(s_t|\boldsymbol{x}_t)$, the required state likelihoods are obtained using Bayes' rule.

4.4. Double-stream and hybrid decoding

The HMM state likelihoods are combined with the GMM acoustic model in a double-stream model topology. At every time step t, the likelihoods of the GMM and the NN acoustic model are joined multiplicatively,

$$p(\boldsymbol{x}_t|s_t) = p_G(\boldsymbol{x}_t|s_t)^{\lambda} \cdot p_N(\boldsymbol{x}_t|s_t)^{1-\lambda}, \qquad (1)$$

where p_G and p_N are the likelihoods of the GMM and the NN acoustic models, respectively, and λ is the stream weight of the GMM stream. Setting $\lambda = 0.0$ corresponds to hybrid NN/HMM acoustic modelling, where only the NN model is used. In addition, the double-stream setup can also be used to combine both LSTM acoustic models (phoneme and state prediction).

The biggest difference between the two methods of using neural network predictions for acoustic modelling is the number of training targets. For phoneme predictions, the network has 40 output units, whereas the network predicting HMM states has 1 936 output units (number of HMM states). The likelihoods derived from the phoneme predictions model the confusions the network makes.

5. Experiments

In our experiments, we compare the two different ways of acoustic modelling using LSTM networks (with different network topologies), performing experiments on the medium vocabulary track of the 2nd CHiME Challenge database as described in Section 2.

5.1. Parametrisation

The parametrisation of the HMM-GMM system has already been described in Section 3. This system is used to generate the forced alignments needed for setting the training targets for the NNs.

All evaluated NNs work with logarithmic Mel filterbank outputs as features. In other studies it was shown that with neural networks, those features perform better than MFCCs [1, 25]. We use 26 coefficients (plus root-mean-square energy) covering the frequency range from 20–8 000 Hz, together with delta and delta-delta coefficients, resulting in an 81-dimensional feature vector. Since the CHiME corpus contains noisy training data, all networks were trained in a multi-condition way, i. e., using noisy and reverberated-only training data together. The networks are trained through gradient descent with a learning rate of 10^{-5} and momentum of 0.9. During training, zero mean

Table 2: *Results for different phoneme prediction networks (WER [%] on the development set), either used as an acoustic model alone or combined with the GMM in the double-stream architecture.*

Network		GMM	
Function	**Layers**	-	✓
LSTM	81-128-90	59.1	39.8
BLSTM	81	59.7	39.7
BLSTM	81-128-90	49.0	36.5
BLSTM	100-100-100	45.3	34.9

Table 3: *Results for hybrid acoustic modelling with different state prediction networks (WER [%] on the development set)*

Network		
Function	**Layers**	**Average WER**
BLSTM	300	40.0
BLSTM	300-300	31.5
BLSTM	500-500	31.4
BLSTM	300-300-300	32.0

Gaussian noise with standard deviation 0.6 is added to the inputs in order to further improve generalisation. All weights were randomly initialised from a Gaussian distribution with mean 0 and standard deviation 0.1. After every training epoch, the average cross entropy error per sequence on the development set is evaluated. Training is aborted as soon as no improvement on the development set can be observed during 20 epochs.

Phoneme prediction networks work better in the double-stream setup where they are combined with the GMM. In this case, the stream weight is set to $\lambda = 0.5$ based on experience from previous work. In addition, for the phoneme prediction networks, we report results for the hybrid setup as well. The state prediction networks are capable of functioning as an acoustic model alone, without the double-stream setup. We tested different network configurations, to investigate how the topology influences the recognition performance, where we relied on our experience from previous works to determine the network size.

5.2. Development set results

Development set results for the employed GMM system are shown in Table 1, resulting in decreasing WER from 64.3 % to 29.7 % with increasing SNR, and an average WER of 45.5 %. These results demonstrate the difficulty of the recognition scenario.

Using phoneme prediction NNs as an acoustic model (cf. Section 4.2) leads to the results in Table 2. All NNs are evaluated in the hybrid and in the double-stream setup. The results show clearly that, in order to obtain reasonable results, it is required to combine the NN predictions with the GMM acoustic model. An LSTM network with three hidden layers improves the GMM result by 5.7 %, absolutely, and a BLSTM with one layer achieves similar results. Adding more layers to the BLSTM improves the result to 36.5 % average WER. Another improvement can be achieved by using a more straightforward network topology, compared to the other network topologies which were derived based on our previous experience. This system's performance (last row in Table 2), used as an acoustic model alone, is similar to the GMM; it will be used (in the multi-stream setup) in the experiments with the test set.

Results for the evaluated state prediction networks, used as an acoustic model in the hybrid setup, are shown in Table 3.

Table 4: *Test set results for selected systems (WER [%])*

System	SNR [dB] -6	-3	0	3	6	9	Avg.
GMM baseline [7]	70.4	63.1	58.4	51.1	45.3	41.7	55.0
evaluated GMM	60.2	50.6	44.9	37.0	31.0	27.6	41.9
GMM (discriminative learning+SAT) [19]	54.7	45.1	36.0	28.6	24.4	21.4	35.0
GMM (discriminative learning+SAT) + denoising [26]	44.1	35.5	28.1	21.2	17.4	14.8	26.9
GMM + BLSTM 78-128-90 [15]	58.6	50.1	43.9	37.1	32.7	28.3	41.8
DNN [4]	42.1	31.7	24.7	19.4	16.4	14.3	24.8
GMM + BLSTM (phonemes)	45.6	36.6	30.8	25.0	20.4	19.3	29.6
BLSTM (states)	40.3	32.2	25.0	19.8	16.8	15.8	25.0
BLSTM (phonemes) + BLSTM (states)	35.9	28.3	22.5	17.8	15.3	13.6	22.2

A BLSTM network with one layer achieves a similar performance (40.0 %) as a comparable phoneme prediction network (39.7 %, row three in Table 2). Adding a second layer leads to a substantial improvement (31.5 %). This network (which is later employed in the test set experiments) has 1.4 million weights in total. Increasing the layer size of this network (3.2 million weights) or adding a third layer (2 million weights) brings no benefit. These results show the limits of increasing the size of the network. Presumably, training methods known from deep learning are required to improve the performance of larger networks. The BLSTM network with two layers of size 300 (row three) was also evaluated in combination with the GMM, in the double-stream setup, resulting in a WER of 30.8 % (result not shown in the table). This is a further small improvement, which, however, comes at the computational cost of evaluating the GMM.

5.3. Test set results

In Table 4, we show experimental results of selected systems for the test set of the CHiME corpus. For comparison, we cite results of systems from the original CHiME challenge. The challenge baseline (55% WER) consists of a simple GMM system. With the GMM system evaluated in this paper, the WER is reduced to 41.9 %. This improvement can be attributed to a better system topology and training procedure, feature transformation in the form of LDA and MLLT, and the beam-forming approach employed in the front-end. The system presented in [19] additionally makes use of discriminative learning and speaker adaptive training (SAT), resulting in a WER of 35.0 %. Adding a denoising method as front-end processing to that system [26] reduced the WER to 26.9 %, which was the best entry to the 2nd CHiME challenge medium vocabulary track. In our original contribution to the challenge, we combined the double-stream BLSTM approach (though using MFCCs instead of Log-FB features) with the baseline GMM, decreasing the WER from 55.0 % to 41.8 %. Now, combining the BLSTM phoneme predictions with a better GMM system (41.9 %) results in a WER of 29.6 %. It is expected that the combination of a better GMM (as the one in [26] with a WER of 26.9 %) with the phoneme predictions of the LSTM can lead to even better results. Row seven in Table 4 (25.0 %) represents the result of the state prediction BLSTM in the hybrid system. Compared to the double-stream system (29.6 %), a relative WER reduction of 16 % is achieved. The last row shows the result of using the double-stream architecture to combine both different LSTM acoustic models. This combination leads to a further substantial improvement. In comparison to the DNN results (24.8 %) presented in [4], the BLSTM in the hybrid setup achieves similar results (25.0 %). However, there are two differences between these two systems, which complicates the comparison: first, while the DNN used clean GMM alignments for training as well as several iterations of alignment and re-training, the BLSTM in our work was trained using GMM alignments on noisy data, and only one iteration of training. On the other hand, our system used delay-and-sum beamforming as preprocessing.

6. Conclusions

We used LSTM RNNs as an acoustic model for a robust speech recognition system. Bidirectional LSTM networks were trained with HMM states as training targets, and the resulting predictions were converted into state likelihoods for decoding in the HMM framework in the hybrid setup. This method was compared to our previous approach of predicting phonemes with the LSTM, converting these phoneme predictions into state likelihoods and using them in a GMM-LSTM double-stream setup. The experimental results, obtained with the medium-vocabulary recognition track of the 2nd CHiME challenge, showed that the hybrid system (using state prediction networks) achieves a lower WER than the double-stream setup (using phoneme prediction networks). The BLSTM in the hybrid setup furthermore outperformed the best entry to the original CHiME challenge. In addition, a further improvement was obtained by combining both different LSTM acoustic models.

It was shown that the state prediction network profits more from a deep network topology, compared to the phoneme prediction network. Combining the state predictions with a GMM in the double-stream setup brought only a small improvement, because the GMM and LSTM acoustic models are probably strongly correlated.

Concerning future work, it is still unclear, how big the influence of front-end processing such as speech or feature denoising on NN systems for robust speech recognition is. In [3], it was shown that such techniques have a lower impact when applied as a front-end to the NN system, while in the study presented in [27], speech enhancement was able to improve a DNN recognition system. Therefore, it will be interesting to investigate the application of speech or feature enhancement as a front-end to the LSTM systems presented in this work. Our results showed that the employed method for phoneme prediction networks is complementary to GMM acoustic modelling and therefore, the double-stream system can profit from improvements in the GMM front-end.

Furthermore, methods such as generative pre-training could be applied to LSTM networks to improve their performance. Further investigations about the influence of network topology of LSTMs as well as better direct comparison to state-of-the-art DNN systems are neccesary to draw more general conclusions about the comparability of LSTMs and other DNNs.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Kluwer Academic Publishers, 1994.

[3] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7398–7402.

[4] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. ICASSP*. Florence, Italy: IEEE, 2014, pp. 5569–5573.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 4860–4863.

[7] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 126–130.

[8] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[9] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 6645–6649.

[10] A. Graves, N. Jaitly, and A.-R. Mohamed, "Speech recognition with deep recurrent neural networks," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, 2013, pp. 273–278.

[11] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4085–4088.

[12] C. Plahl, M. Kozielski, R. Schlüter, and H. Ney, "Feature combination and stacking of recurrent and non-recurrent neural networks for lvcsr," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 6714–6718.

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

[14] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 24–29.

[15] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the 2nd CHiME Challenge: Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF," in *Proc. of CHiME Workshop*, Vancouver, Canada, 2013, pp. 25–30.

[16] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and NMF for robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1037–1046, 2014.

[17] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 89–100, 2005.

[18] A. Hagen and A. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol. 19, no. 1, pp. 3–30, 2005.

[19] Y. Tachioka, S. Watanabe, and J. R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 6935–6939.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Honolulu, HI, USA, 2011.

[21] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. of ICASSP*, San Francisco, CA, USA, 1992, pp. 13–16.

[22] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. of ICASSP*, Istanbul, Turkey, 2000, pp. 1129–1132.

[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[25] 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.

[26] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proc. of CHiME Workshop*, Vancouver, Canada, 2013, pp. 19–24.

[27] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?" in *Proc. of Interspeech*, Lyon, France, 2013, pp. 2992–2996.