THE MUNICH FEATURE ENHANCEMENT APPROACH TO THE 2ND CHIME CHALLENGE USING BLSTM RECURRENT NEURAL NETWORKS

Felix Weninger[1]*, Jürgen Geiger*[1]*, Martin Wöllmer*[2]*, Björn Schuller*[1]*, Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]BMW Group, Munich, Germany

felix.weninger@mytum.de

ABSTRACT

We present a highly efficient, data-based method for monaural feature enhancement targeted at automatic speech recognition (ASR) in reverberant environments with highly non-stationary noise. Our approach is based on bidirectional Long Short-Term Memory recurrent neural networks trained to map noise corrupted features to clean features. In extensive test runs, enhanced features are evaluated with gradually refined recognition back-ends, reaching from simple maximum likelihood (ML) trained recognisers to state-of-the-art ASR using discriminative training and model adaptation techniques. In the result, consistent improvements over the baseline ASR systems on both the small and medium vocabulary tasks of the 2nd CHiME Speech Separation and Recognition Challenge demonstrate the efficacy of the proposed method, achieving up to 52 % relative reduction of word error rate with respect to the multi-condition ML training baselines.

Index Terms— Long Short-Term Memory, recurrent neural networks, feature enhancement

1. INTRODUCTION

Decoding of speech in unfavourable acoustic conditions, especially in reverberated environments with interfering noise sources, is still a major challenge for today's automatic speech recognition (ASR) systems despite decades of research on this topic. Robustness of ASR systems can be addressed at different stages of the recognition process [1] – popular techniques comprise front-end speech enhancement, such as by microphone array processing or speech de-noising techniques, as well as improvements in the back-end by model adaptation or improved ASR architectures taking into account additional sources of information, such as neural networks. 'In between' one can also address noise-robust features – a popular expert crafted feature extraction scheme is RASTA-PLP [2] – or feature enhancement, defining a mapping from noisy to noise free speech features. An example for a data-based, non-parametric technique for feature enhancement is histogram equalisation [3]. Furthermore, feature enhancement by recurrent neural networks has been considered [4]. In particular, bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks (RNNs) have been employed in [5] for feature enhancement in highly non-stationary noise, by mapping noisy cepstral features to clean speech cepstral features, and have been shown to outperform traditional RNNs on this task.

In this contribution, we apply the methodology from [5] to the small vocabulary and the medium vocabulary ASR tasks of the 2nd CHiME Speech Separation and Recognition Challenge [6]. A major focus thereby is the interaction between feature enhancement and speech recognisers. In particular, we investigate the performance gain by using feature enhancement on top of refined speech recognition back-ends, comprising multi-condition training as well as more advanced adaptation techniques such as speaker adaptive transforms, as well as discriminative training with noisy data. Furthermore, we evaluate feature enhancement both in a 'plug-and-play' fashion, where enhanced cepstral features are used without modification of the back-end, and contrast this with the performance of using models re-trained with enhanced features. Both of these points have not been addressed in our earlier work [5] which only considered re-trained models and maximum likelihood (ML) recogniser training. In the following, we will first outline our feature enhancement methodology before describing the experimental setup and presenting the results on the CHiME Challenge data.

This research has been supported by the German Research Foundation (DFG) through grant no. SCHU 2508/4.

2. EVALUATION DATABASE

The small vocabulary task of the 2nd CHiME Challenge [6] consists of reverberated and noisy utterances from the Grid corpus resembling command-and-control utterances with a fixed grammar and a vocabulary size of 51. Utterances have been convolved with real room impulse responses measured in a domestic environment, and overlaid with realistic noise recorded from the same environment at signal-to-noise ratios (SNRs) from -6 to 9 dB, in steps of 3 dB. A closed set of 34 speakers is used for training, development, and testing in the small vocabulary task. The medium vocabulary task is created in a similar way, using the same noise corpus but the speaker independent development and evaluation test sets of the Wall Street Journal corpus (WSJ-0) with 5 k vocabulary size and disjoint sets of 84, 10, and 8 training, development, and test speakers. For both tasks, the same utterances are used at all SNRs in the development and test sets. The training sets comprise a randomly selected subset of utterances for each SNR. In the small vocabulary task, the training set has 17 000 utterances while the development and test sets consist of $6 \times 600 = 3\,600$ utterances. In the medium vocabulary task, there are 7 138 training, $6 \times 409 = 2\,454$ development, and $6 \times 330 = 1\,980$ test utterances. While the Challenge data is stereophonic, in our study we only consider monaural signal processing since we are especially interested in use cases such as multimedia information retrieval, where multi-channel audio with specified microphone placement is usually not available.

3. METHODOLOGY

Our feature enhancement approach is based on BLSTM recurrent neural networks that are trained to map cepstral features of noisy

speech to the corresponding features of noise free[1] speech, exploiting the context-sensitivity of the BLSTM technique. We have shown that the BLSTM architecture outperforms standard RNNs by a large margin on the feature enhancement task [5].

The basic architecture of Long Short-Term Memory (LSTM) networks was introduced in [7]. The underlying principle can be seen as an extension of conventional RNNs that enables the modelling of long-range temporal context for improved sequence labelling. LSTM networks are able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the regression or classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets (so-called *memory blocks*). Every memory block consists of self-connected *memory cells* and three multiplicative *gate* units (input, output, and forget gates). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer. Further details on the LSTM principle can be found in [8].

Standard RNNs have access to past but not to future context. To exploit both, past and future context, RNNs can be extended to *bidirectional* RNNs (BRNN), where two separate recurrent hidden layers scan the input sequences in opposite directions [9]. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. Bidirectional modelling can also be applied within an LSTM framework, which results in BLSTM. In the context of feature enhancement, BLSTM has been shown to outperform LSTM modelling at the expense of on-line capability [5].

4. EXPERIMENTS

4.1. Network Training

For each of the small and medium vocabulary tracks, a feature enhancement BLSTM network is trained on the task to map the official noisy training sets of the Challenge data to the corresponding reverberated training set. Only the isolated utterances are used. In our experiments, we use 39 cepstral mean normalised mel-frequency cepstral coefficients (MFCCs) exactly corresponding to the features employed by the Challenge baseline. In particular, the stereophonic signals are down-mixed to monophonic audio by averaging channels.

Our feature enhancement network has one input node for each noise corrupted input feature vector component and one output node for each regression target representing the noise free feature vector. In particular, the networks also predicts delta and acceleration coefficients, and can use the predicted deltas and accelerations as additional context information. Prior to network training, we compute the global means and variances of the reverberated and the noisy training set feature vectors and perform mean and variance normalisation of the network training targets and the network inputs accordingly. This normalisation was found necessary in order to ensure that cepstral, delta and acceleration coefficients are in similar order of magnitude for calculating the gradient of the error function in network training.

For the sake of consistency, the hyperparameters of the network and the training parameters are set exactly as in our previous study on conversational speech recognition in noise (yet without reverberation) [5]; no parameter tuning on the CHiME 2013 data is involved. The applied networks have three hidden layers consisting of 78, 128, and 78 memory blocks. Each memory block contains one memory cell. We train the networks through gradient descent with a learning rate of 10^{-5} and a momentum of 0.9. The gradient descent algorithm minimises the root mean squared error (RMSE) on the training data, across all six SNRs. Hence, the network is required to generalise to various levels of noise. Zero mean Gaussian noise with standard deviation 0.1 is added to the input activations in the training phase in order to further improve generalisation. Prior to training, all weights are randomly initialised in the range from -0.1 to 0.1. Input and output gates use hyperbolic tangent activation functions, while the forget gates have logistic activation functions. We use an early stopping strategy: In the training phase, we evaluate the overall error on the development set after every fifth epoch. More precisely, we compute the total RMSE of the enhanced noisy development features with respect to the ground truth reverberated development features. We abort training as soon as no improvement on the development set can be observed during 30 epochs. The network that achieved the best RMSE on the development set (with the mean taken across all six SNRs) is chosen as the final network. For the speaker-dependent small vocabulary recognition task, we also consider speaker-dependent feature enhancement networks. These are derived from the generic speaker-independent network, by running additional training epochs on only the data from only one specific speaker, and using early stopping as above, but evaluating the cost function only on the development data of this single speaker.

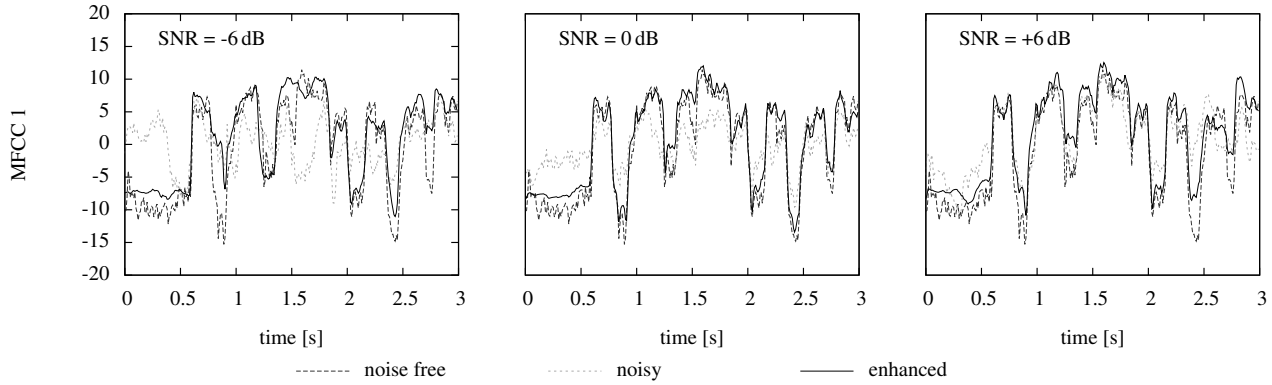
4.2. Feature Enhancement

Enhanced features are generated by simply presenting the frame-wise noisy MFCCs to the trained network and computing the output activations in a forward pass. Due to the normalisation of the training targets, the output activations are (approximately) mean and variance normalised, which does not match the features used to train the baseline models. Thus, to be able to use the enhanced features in a 'plug-and-play' fashion, i.e., without any recogniser modification, the global mean and variance normalisation is reverted after obtaining the enhanced MFCC features, to foster compatibility with the means and variances of the trained recognition models. More specifically, each enhanced feature vector is multiplied component-wisely with the corresponding variances of the reverberated training set, and the mean feature vector of the reverberated training set is added.

An example of the resulting features is presented in Figure 1. We depict the first MFCC of the utterance 050c0101 from the development set of the medium vocabulary task, at SNRs of -6, 0, and +6 dB: once as extracted from the noisy waveform, once as enhanced by the BLSTM, and once the 'ground truth' MFCC extracted from the corresponding reverberated, but noise free waveform. Only the interval [0 s, 3 s] is displayed for illustration. Note that the shapes of the noisy MFCC contours differ strongly among SNRs, since various noise segments have been used for mixing the noisy utterances. It can be seen that the BLSTM is able to reconstruct the noise free MFCC to some degree, even at -6 dB SNR. The RMSEs of the BLSTM enhanced MFCC contours, with respect to the noise free contour, are 3.58, 2.02, and 2.92 at -6, 0, and +6 dB in the displayed interval of the example utterances, while the noisy contours correspond to RMSEs of 7.17, 4.84, and 4.53. The higher RMSE of the enhanced MFCC at +6 dB can be explained by the presence of an interfering speaker (who is not present at lower SNRs – note that noise segments differ among SNRs in the CHiME corpus [6]). The behaviour of the BLSTM enhanced MFCC at the start of the utterance (interval [0 s, 0.6 s]) is interesting, as it is much smoother than the reverberated MFCC. In this interval, there is no speech, but rather well audible vocal noise (breathing) of the target speaker, which is apparently filtered out by the BLSTM as well.

[1] In this paper, we do not consider de-reverberation – hence, the enhanced features are not 'clean' in the sense of 'de-reverberated'.

Fig. 1: Results of BLSTM based enhancement of the first MFCC of utterance 050c0101, female speaker 050, WSJ-0 speaker independent 5 k vocabulary development set. Noisy MFCC 1 at SNR $\in \{-6, 0, +6\}$ dB in CHiME noise [6] vs. noise free and enhanced MFCC.



4.3. Speech Recognition Evaluation

For a more systematic (task-based) evaluation of feature enhancement, we conduct ASR experiments using a sequence of gradually refined recognisers. In line with the Challenge evaluation setup, we use word accuracy (WA, measured on the 35 letter and digit keywords) for the small vocabulary task, while we give word error rates (WER) for the medium vocabulary task. Note that in the small vocabulary task there are no insertion or deletion errors due to the fixed grammar decoding.

4.3.1. Baseline models

We evaluate the performance of the enhanced features using the baseline models provided by the Challenge organisers, as well as re-trained models using enhanced features. The baseline training and decoding setups provided by the Challenge organisers are used for easy reproducibility – only exchanging the set of feature files that is used. Both official baselines are implemented using HTK [10]. The baseline training procedures, however, differ between the small and medium vocabulary ASR tasks.

For the small vocabulary ASR task, the baseline setup performs HMM training 'from scratch', using either reverberated or noisy features. Thus, we evaluate our features in this setup by training on the reverberated and noisy training sets, respectively, using the training set features processed by the BLSTM.

For the medium vocabulary ASR task baseline, pre-trained clean models are used (based on the original WSJ corpus), which are then re-trained using Maximum Likelihood (ML) training. First, re-training is performed using reverberated features, yielding the baseline reverberated acoustic models. Then, another re-training step is done to provide the baseline reverberated and noisy acoustic models. For adapting these models to the enhanced features, the re-training steps are repeated using enhanced features. Manifold other re-training configurations (such as starting from the pre-trained clean model and training using enhanced features) can be thought of, and these were evaluated as well – they resulted in slightly different results (in the order of 1 % absolute average WA difference on the development set) and are not reported here for the sake of clarity.

4.3.2. Discriminatively trained / adapted models

For the medium vocabulary ASR task, we perform further experiments using the speech recognition system described in [11]. This system is based on the Kaldi speech recognition toolkit [12]. It uses state-of-the-art ASR techniques such as linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), speaker adaptive training (SAT) and discriminative training (DT) with boosted model-space or feature-space maximum mutual information (MMI) estimation for training with noisy data. Context-dependent triphone models are first trained with reverberated data using ML training. Then, ML training is continued with noisy training data, using LDA, MLLT and SAT. After that, boosted MMI discriminative training (boosting factor 0.1) is performed using noisy features, including feature-space and model-space components.

First, experiments are conducted using the reverberated models. Then, models are trained on enhanced reverberated features of the training set instead of unprocessed reverberated features, and their performance is evaluated on the enhanced features of the noisy development and test sets. Finally, the above mentioned ML and DT methods are applied to the reverberated models, using noisy features of the training set. Either unprocessed noisy features or enhanced noisy features are used for these steps. For evaluation on the noisy development and test sets, the corresponding features (unprocessed or enhanced) are used.

Language model weights employed during decoding are optimised by using the development set. All other training and decoding parameters exactly correspond to the ones used in [11] for the sake of transparency.

5. RESULTS

5.1. Small Vocabulary ASR

Table 1 shows the results obtained on the development and test sets of the 2nd CHiME Challenge, small vocabulary track. We report the mean accuracy across the six SNRs (-6 to 9 dB in 3 dB steps) on the development set, and the detailed accuracies per SNR on the test set. By simply 'plugging' the enhanced features into the reverberated acoustic models, we already obtain a large improvement (25.8 % absolute, 44 % relative increase in average WA on the test set to 83.3 %). Results can be significantly[2] improved by considering

[2]When we speak of significant differences, we mean statistical significance according to a simple z-test, using the significance level $\alpha = .05$. As a rule of thumb in the ranges of WA observed in our experiments on the small vocabulary task, results have to differ by 1.5 % absolute WA on average across

Table 1: *CHiME 2013 development and test set (small vocabulary track): (Key)word accuracies (% WA) using feature enhancement (FE), baseline reverberated and noisy recognisers, and matched condition recogniser trained with noisy features processed by feature enhancement. SI/SD: speaker-(in)dependent networks for feature enhancement.*

WA [%]	**Devel**	**Test**						**Test**
	Mean	**SNR** [dB]						**Mean**
		-6	-3	0	3	6	9	
Reverberated acoustic models								
Baseline	56.9	32.2	38.3	52.1	62.7	76.1	83.8	57.5
+ FE (SI)	83.1	71.6	78.3	83.1	86.7	89.3	91.1	83.3
+ FE (SD)	84.7	75.2	79.4	85.8	88.5	89.6	90.8	84.9
+ re-training	**84.8**	74.8	78.7	86.0	88.3	90.0	92.3	**85.0**
Reverberated + noisy acoustic models								
Baseline	68.8	49.3	58.7	67.5	75.1	78.8	82.9	68.7
+ FE (SI)	76.8	66.8	72.0	77.2	79.5	81.9	83.0	76.7
+ FE (SD)	77.2	69.5	72.7	78.2	80.3	82.7	83.9	77.9
+ re-training	**84.9**	74.9	78.6	86.2	88.3	89.8	92.3	**85.0**

speaker-dependent re-training of the networks (84.9 % average WA on test). Using enhanced reverberated features in model training does not significantly improve the results any further (85.0 % average WA on test). These figures are similar to those reported in [13] for BLSTM-based multi-stream ASR on the previous CHiME data set.

Considering noisy acoustic models, baseline accuracies are higher (68.7 % average WA), yet the improvement by using feature enhancement is smaller (up to 77.9 % average WA) – thus being significantly below the result using reverberated models. Especially for higher SNRs, feature enhancement does not significantly improve over the baseline. We believe that this can be attributed to training with only noisy data, resulting in larger model variances – note that the enhanced features have the variances of the reverberated training data, generating a mismatch. In fact, when training models from scratch using enhanced noisy features, the above-mentioned performance drop is avoided (85.0 % average WA, best result on the test set). Interestingly, these 'matched condition' models behave very similarly to the noise-free models coupled with feature enhancement.

5.2. Medium Vocabulary ASR

Results on the medium vocabulary (5 k) test set are shown in Table 2. In a 'plug-and-play' setup using the baseline reverberated acoustic models, average WER is halved from 73.70 % up to 46.01 %. Re-training using enhanced reverberated features significantly[3] decreases the error rate on the development set, but not on the test set. For the noisy baseline model, average WER decreases from 55.01 % to 48.16 % without re-training (similarly to the small vocabulary results, this stays below the performance with the reverberated model). Average WER is significantly reduced to 42.97 % with re-training using enhanced reverberated and noisy features.

Finally, we perform experiments with the Kaldi ASR system. With the acoustic models trained only on reverberated data, an average WER of 68.23 % is achieved, which is already better than the baseline reverberated models. Using enhanced reverberated features for training and evaluating on enhanced noisy features, the average WER is halved to 37.43 %. Notably, this result is better than the best result obtained with the official baseline recognition system and re-training (42.97 %, cf. above), probably owing to the system being tuned more towards less noisy conditions. Re-training models with noisy data and the described ML and DL techniques (whereby MMI including feature-space components was used) results in an average WER of 34.85 % without using feature enhancement, in accordance with the results reported by [11]. Feature enhancement leads to an additional relative WER reduction by 23.3 %, yielding 26.73 % average WER, which is our best achieved result. Note that when combining feature enhancement with DT, model-space adaptation (MMI) performed better than feature-space adaptation (fMMI). Overall, the relative improvements by feature enhancement are similar on development and test data.

6. CONCLUSIONS

We have demonstrated the efficacy of data-based cepstral domain feature enhancement for noise-robust ASR in challenging environments. Furthermore, improvements by advanced ASR model training have been shown to be complementary to the enhancement of ASR features by the proposed method. In the small vocabulary task, feature enhancement improves the test set WA (averaged over 6 SNRs) to 85.0 %, compared to the averaged WA of 68.7 % achieved with the baseline noisy acoustic models. For the medium vocabulary task, an average WER of 26.73 % is achieved, whereby the baseline system yields an average WER of 55.01 %. The improvements by the proposed method are all the more noticeable since it is fully monaural. Note, however, that this is not an official competition result in the Challenge, because learning a mapping between noisy and clean features was not allowed as per the Challenge guidelines. An advantage of the proposed method over monaural Mel or Fourier domain feature enhancement by non-negative matrix factorisation [14, 15] is that most of the computational complexity involved is shifted to a training phase, while evaluation can be done very efficiently – in contrast to typical NMF approaches involving little to no model pre-training but considerable effort in model evaluation.

The experiments in this study were focussed on the CHiME 2013 evaluation, involving non-stationary noise, yet from similar noise sources in training and test. Future work will hence concentrate on model generalisation, for example by joint speech and noise estimation in the given framework, as well as joint speech activity detection and context-sensitive de-noising in noisy acoustic streams.

SNRs, and by 4 % absolute WA per SNR to be significantly different.

[3] According to a z-test with $\alpha = .005$, treating the number of words as sample size.

Table 2: *CHiME 2013 development and test set (5 k medium vocabulary track): Word error rates (% WER) using feature enhancement (FE), baseline reverberated and noisy recognisers, and recognisers re-trained with enhanced features.*

WER [%]	**Devel Mean**	**Test SNR** [dB]						**Test Mean**
		-6	-3	0	3	6	9	
Reverberated acoustic models								
Baseline	72.56	87.97	83.19	78.05	71.87	65.23	55.91	73.70
+ FE	52.41	62.26	54.47	48.14	41.96	36.80	32.45	**46.01**
+ re-training	**50.97**	64.26	55.99	48.03	41.45	36.86	32.71	46.55
Reverberated + noisy acoustic models								
Baseline	58.27	70.43	63.09	58.42	51.06	45.32	41.73	55.01
+ FE	54.51	62.04	54.59	50.31	44.74	40.26	37.01	48.16
+ re-training	**47.62**	56.86	50.25	45.08	39.25	34.56	31.81	**42.97**
Kaldi ASR system (reverberated model, maximum likelihood training)								
Baseline	71.82	85.97	80.29	74.22	66.00	56.51	46.39	68.23
+ FE + re-training	**42.97**	56.45	46.25	38.13	32.32	27.83	23.63	**37.43**
Kaldi ASR system (reverberated + noisy model, discriminative training)								
Baseline	40.96	55.22	44.14	37.29	29.63	23.24	19.60	34.85
+ FE + re-training	**32.94**	42.67	33.92	27.50	21.78	18.38	16.16	**26.73**

Furthermore, other feature representations such as modulation spectrum based features (e.g., RASTA-PLP) will be investigated. Finally, we will also address feature space de-reverberation using context-sensitive BLSTM modelling.

7. REFERENCES

[1] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, ID 942617.

[2] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1992, vol. 1, pp. 121–124.

[3] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, and A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[4] S. Parveen and P. Green, "Speech enhancement with missing data techniques using recurrent neural networks," in *Proc. of ICASSP*, Montreal, Canada, 2004, pp. 733–736.

[5] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[6] 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.

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

[8] 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.

[9] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[10] S.J. Young, G. Evermann, M.J.F. Gales, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev, and P.C. Woodland, *The HTK book version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[11] 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.

[12] 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*, 2011.

[13] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J.F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-Negative Matrix Factorization for Highly Noise-Robust ASR: to Enhance or to Recognize?," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4681–4684.

[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] A. Hurmalainen, K. Mahkonen, J.F. Gemmeke, and T. Virtanen, "Exemplar-based recognition of speech in highly variable noise," in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 1–5.