FAR-FIELD SPEECH RECOGNITION USING CNN-DNN-HMM WITH CONVOLUTION IN TIME

Takuya Yoshioka[1], *Shigeki Karita*[1,2], *and Tomohiro Nakatani*[1]

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2]Graduate School of Engineering, Osaka University, Osaka, Japan
{yoshioka.takuya,nakatani.tomohiro}@lab.ntt.co.jp, karita@nanase.comm.eng.osaka-u.ac.jp

ABSTRACT

Recent studies in speech recognition have shown that the performance of convolutional neural networks (CNNs) is superior to that of fully connected deep neural networks (DNNs). In this paper, we explore the use of CNNs in far-field speech recognition for dealing with reverberation, which blurs spectral energies along the time axis. Unlike most previous CNN applications to speech recognition, we consider convolution in time to examine whether it provides an improved reverberation modelling capability. Experimental results show that a CNN coupled with a fully connected DNN can model short time correlations in feature vectors with fewer parameters than a DNN and thus generalise better to unseen test environments. Combining this approach with signal-space dereverberation, which copes with long-term correlations, is shown to result in further improvement, where the gains from both approaches are almost additive. An initial investigation of the use of restricted convolution forms is also undertaken.

Index Terms— Far-field speech recognition, reverberation, convolutional neural network, deep neural network

1. INTRODUCTION

While speech recognition technology has matured to the point where it is utilised in a range of applications, far-field speech recognition remains a challenge. When a human voice is captured with a microphone at some distance from a talker in an enclosed space, such as a meeting room or a living room, the audio signal observed by the microphone is corrupted by reverberation and background noise, which impairs the speech quality. Even with Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid acoustic models [1–3] trained on corrupted data, the performance gap between far-field and close-talking set-ups is still large as demonstrated in the AMI meeting transcription task [4, 5] and the REVERB challenge task [6].

Since the effect of reverberation spans a number of consecutive time frames, far-field speech recognition systems have to account for the long-term statistical dependency inherent in the captured audio [7]. For conventional acoustic models based on Gaussian Mixture Model-Hidden Markov Models (GMM-HMMs), various approaches have been proposed for dealing with reverberation, ranging from signal enhancement techniques—such as Weighted Prediction Error (WPE) minimisation [8] and Bayesian filtering [9]—to dedicated modelling approaches—such as Direct CMLLR [10], REMOS [11], and Reverberant Vector Taylor Series (RVTS) compensation [12]. Of these techniques, signal-space convolution methods such as WPE were shown to be effective even for DNN-HMMs [6, 13].

DNN-HMM acoustic models already use a longer acoustic context (typically in the 100–200 ms range) than conventional GMM-HMMs by splicing feature vectors within a context window. The conventional DNN-HMMs model correlations between different frames within the window by using a fully connected network, which requires many parameters to be learned. On the other hand, the correlations that result from reverberation may be well organised since the reverberation effect can be (approximately) described with a small number of parameters in the power spectrum domain [14, 15]. If we could utilise a structured model to represent the reverberation-specific correlations with fewer parameters, the resultant model would generalise better to unseen environments than conventional DNN-HMMs. Compact models would also be useful when performing adaptation with a small quantity of data.

To model reverberation more efficiently, we consider using CNNs with convolution performed along the time axis. The convolution layer used in this paper performs multi-input multi-output convolution through an entire input sequence. The convolution layer shares weights among different connections by assuming that the relationship between neighouring frames does not vary over time. Such weight sharing significantly reduces the number of parameters to optimise. A causal linear filter is used to deal with reverberant noise, whose effect stems from the current and past frames. Although the impact of reverberation on acoustic features is highly nonlinear, we assume it to be approximated to some extent with a linear model [10, 16]. A pooling layer, which has often been combined with the convolution layer in previous studies, is not used in this work as pooling in time does not seem to improve the reverberation modelling capability. Fully connected layers are stacked on top of the convolution layer to compute HMM state posteriors with which Viterbi decoding is performed. In addition to using the CNNs, we also investigate the combined effect of the CNNs and WPE, which is a dereverberation method based on signal-space convolution.

In the remainder of this paper, Section 2 briefly discusses links between this work and previous studies. Section 3 describes the CNN-DNN-HMM framework used in this work. Section 4 shows experimental results and Section 5 concludes the paper. Appendix A provides a brief description of the WPE method. In the following, we use the term DNN to refer to a fully connected neural network.

2. RELATION TO PREVIOUS WORK

2.1. Acoustic Modelling with CNNs

In the speech recognition literature, there have been a few attempts to use CNNs along the time axis. Tóth performed phone recognition using a stacked DNN-HMM hybrid system [17] consisting of two DNNs that are organised hierarchically and trained jointly [18]. He called the lower level, or torso, DNN a CNN in light of the fact

that this network is applied to several different portions of an input feature sequence. However, since the torso DNN used in that work disregards the temporal order of feature vectors, it does not seem to serve as a reverberation model. Abdel-Hamid et al. [19] experimented with CNNs that were applied through time. They reported little performance improvement compared with standard DNN-HMMs while significant gains were obtained from convolution along the frequency axis. Sainath et al. [20] empoloyed two-directional filters, where convolution was performed through time and frequency.

In this paper, we apply a one-directional convolution operation with a causal filter through an entire feature vector sequence prior to context extension (see the diagram in Fig. 1). This is equivalent to extending the context window to the left. Our motivation for using this structure is to capture the sequential nature of reverberation more effectively than is possible with the previous approaches.

2.2. Far-Field Speech Recognition

There have been a limited number of studies on DNN-HMM acoustic modelling for far-field speech recognition. Swietojanski et al. applied a CNN-DNN-HMM approach to a meeting transcription task based on a table-top microphone array, where convolution was performed across microphones [21]. Yoshioka et al. examined the impact of dereverberation processing on the performance of DNN-HMMs in the meeting transcription task [13]. A few systems that participated in the REVERB challenge employed DNN-HMMs for acoustic modelling while none of them used CNNs [6, 22–24]. Supervised enhancement approaches using stereo corpora [25, 26] can possibly be applied to far-field speech recognition although few studies have been performed to evaluate these approaches in far-field tasks.

3. CNN-DNN-HMM WITH CONVOLUTION IN TIME

Figure 1 shows the processing flow of our proposed recognition system based on CNN-DNN-HMMs. A speech signal observed with a microphone is processed by a dereverberation method called WPE [8] to mitigate the reverberation effect in the signal space. Then, the dereverberated audio is converted into a sequence of feature vectors using a window of 25 ms shifted by 10 ms. We use 24-channel log mel-filter bank outputs plus their first and second-order delta coefficients as acoustic features. The resultant feature vectors are normalised so that they have a zero mean and unit variance in each dimension. We perform mean normalisation at an utterance level and variance normalisation at a corpus level. Normalisation processing is followed by a convolution layer, or a CNN, which applies a multi-input multi-output linear filter to the normalised 72-dimensional feature vector sequence. At each frame, the filter outputs are spliced with neighbouring frames within a context window and input to a fully connected DNN to produce HMM state posteriors. At training time, the CNN and DNN parameters are jointly optimised using state alignments generated with a baseline GMM-HMM system. Cross-entropy training is performed with Stochastic Gradient Descent (SGD). At test time, the HMM posteriors produced by the CNN-DNN are translated into pseudo likelihoods to perform Viterbi decoding.

3.1. Convolution Layer

The convolution layer in the left pipeline in Fig. 1 acts as follows. Let $\boldsymbol{x}_t$ denote a 72-dimensional feature vector, comprising static, delta, and delta-delta log mel-filter bank outputs. Convolution is performed on the feature vector sequence, $(\boldsymbol{x}_t)_{t=0,\cdots,T-1}$, as follows:

$$\boldsymbol{y}_t = \sigma\left(\sum_{k=0}^{K-1} \boldsymbol{A}_k \boldsymbol{x}_{t-k} + \boldsymbol{b}\right), \tag{1}$$

where T is an utterance length measured with the number of frames, K is a filter (i.e., convolution) order, and $\sigma(\cdot)$ denotes an activation function. Each $\boldsymbol{A}_k$ is an M-by-N matrix while $\boldsymbol{b}$ is an M-dimensional bias vector. $\boldsymbol{A}_0, \cdots, \boldsymbol{A}_{K-1}$ constitute the multi-input multi-output filter of the convolution layer. In our case, N corresponds to the feature vector dimensionality, i.e., 72. Following [27], we refer to M as the number of feature maps. In (1), the input sequence is extended to the left by $K-1$ frames with $\boldsymbol{x}_0$ to compute the outputs at the beginning of an utterance. After computing the convolution layer outputs, $\boldsymbol{y}_{t-L}, \cdots, \boldsymbol{y}_{t+L}$ are input into the DNN at each frame t.

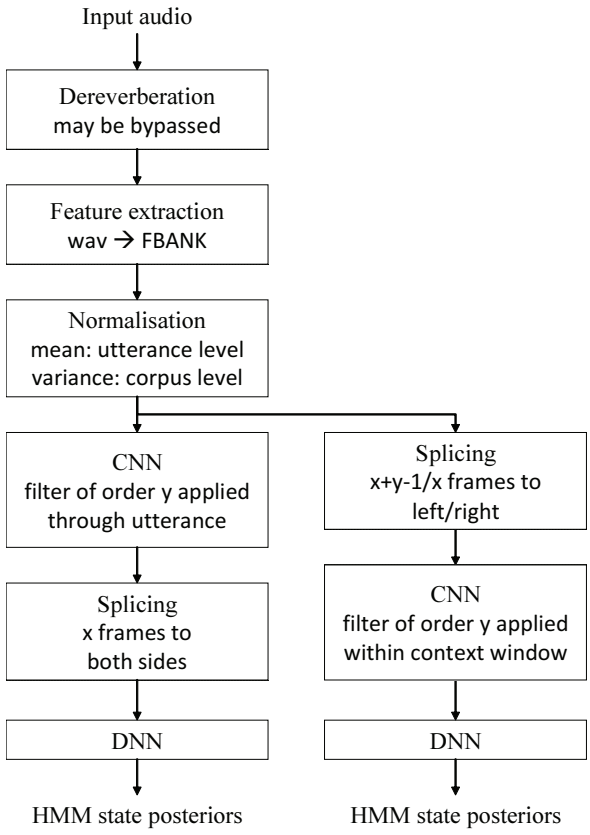

Fig. 1. Processing flow diagram of the proposed system: (left path) original form; (right path) equivalent form.

The CNN-DNN described above can be implemented using existing codes for CNN processing. Figure 2 illustrates the way in which the CNN-DNN translates feature vectors into state posteriors for frame t. The picture shows that the forward pass of our CNN-DNN can be realised by: 1) extending the context window by $K-1$ frames to the left; 2) applying the extended context window directly to raw feature vectors; 3) performing convolution within the extended context window without frame padding at the window edges, which yields the same number of output frames as the original context window size; and 4) forwarding the convolution outputs through the DNN. This means that our CNN-DNN can be imple-

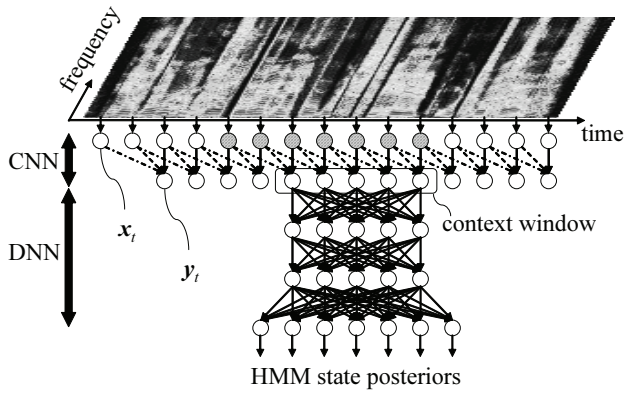

Fig. 2. Computation of HMM state posteriors. Arrows with the same line type (solid, dashed, or dash-dotted) share connection weights in the convolution layer.

mented by using the right pipeline in Fig. 1. With this approach, back-propagation can be performed in a conventional way.

3.2. Structured Forms of Convolution

It is possible to impose a structure on the convolution matrices in (1) to take advantage of the characterstics of input feature vectors. The simplest structure would be a diagonal convolution matrix, where $\boldsymbol{A}_k$ is assumed to be diagonal for all k values. The use of diagonal matrices makes sense when log mel-filter bank features are used since reverberation can be assumed to affect each filter bank separately in ordinary rooms. The use of diagonal matrices significantly reduces the number of parameters to optimise. An alternative structure that can be reasonably imposed on $\boldsymbol{A}_k$ is a block diagonal form. Using block diagonal matrices allows features belonging to the same class (static, delta, delta-delta, etc.) to be processed jointly. In our experiments, we consider full and diagonal matrices.

4. EXPERIMENTAL RESULTS

4.1. Set-ups

We examined the performance of our proposed CNN-DNN-HMM approach on the corpus provided by the REVERB challenge [28,29]. The challenge was held in 2014 to evaluate speech recognition systems in far-field set-ups. The test set comprises simulated data (SimData) and real recordings (RealData), where the SimData were generated by convolving room impulse responses with anechoic speech signals followed by the addition of a moderate amount of background noise. The training data consist of 15 hours of far-field speech signals that were generated by simulation. Since the training set contains no real recordings, there is a significant acoustic mismatch between the training data and the RealData. A single-microphone set-up was considered.

Our recognition systems used acoustic features consisting of 24 log mel-filter bank outputs plus their delta and delta-delta coefficients, which were context-extended with a symmetric window of nine frames. The DNNs had five hidden layers each with 1024 neurons and 3129 output neurons, or target HMM states. State alignments were generated for the training set by using a maximum likelihood GMM-HMM acoustic model. DNN training was performed with layerwise discriminative pre-training [30] followed by cross-entropy fine-tuning, where optimisation was performed with SGD using a minibatch of 128 frames. During fine-tuning, after each training epoch, cross entropy was evaluated on a 5% held-out set and the learning rate was halved if there was no cross entropy improvement over the previous iteration. Fine-tuning was stopped once the learning rate had been reduced six times. These configurations were tuned to our baseline DNN-HMM system. CNN-DNN-HMM systems had the same number of layers as the DNN baseline systems, i.e. one convolution layer plus four fully-connected hidden layers. 72 feature maps were used with a linear activation.

For decoding, we used both trigram and recurrent neural network (RNN) language models (LMs). When performing RNN decoding, LM scores were computed by interpolating trigram and RNN scores with an interpolation coefficient of 0.5.

4.2. Results

Table 1 shows the performance of CNN-DNN-HMM systems with raw (i.e., non-dereverberated) data. The use of CNNs with low orders (S2–S5) yielded gains over the baseline DNN-HMM system (P1) on RealData with both decoding schemes (i.e., trigram and RNN) while neither significant improvement nor degradation was observed on SimData. The larger acoustic mismatch between the training set and RealData than between the training set and SimData means that the CNN-DNN-HMM systems are more robust against environmental mismatch than the conventional DNN-HMM system while both systems exhibit a comparable modelling capability in far-field speech recognition. To our disappointment, increasing the filter order in the CNN-DNN-HMMs gradually degraded the performance, indicating that CNNs with temporal convolution are useful for modelling the short-term correlations that exist in feature vectors. This is probably because the CNNs start to capture irrelevant acoustic aspects that manifest themselves only in the training set when the convolution order is long.

Table 1. %WERs of proposed CNN-DNN-HMM systems using raw microphone signals.

System	Filter order	Trigram		RNN	
		Sim	Real	Sim	Real
P1	—	8.39	27.7	7.39	26.4
S2	2	8.49	25.5	7.60	23.8
S3	3	8.64	26.7	7.57	24.4
S4	4	8.50	26.8	7.46	26.2
S5	5	8.46	26.7	7.63	25.5
S6	6	8.64	27.2	7.75	25.9
S7	7	8.46	27.6	7.69	26.2
S8	8	8.39	27.2	7.20	25.9
S9	9	8.75	28.1	7.71	26.8

A simple approach to coping with both short- and long-term correlations resulting from reverberation is to preprocess input signals with a dereverberation technique prior to recognition with the CNN-DNN-HMMs. To validate this approach, the same investigation was conducted using dereverberated data as inputs. We performed dereverberation with the WPE method [8]. Table 2 shows the WER results. As with the results in Table 1, gains were obtained on RealData when low-order filters were used in CNNs although the gains were small with RNN decoding. The results show the gains from signal-space convolution (i.e., dereverberation) and feature-space convolution (i.e., CNN processing) to be almost additive. With the best configuration, we achieved overall relative improvements of 15.9%

Table 2. %WERs of proposed CNN-DNN-HMM systems using signals dereverberated with WPE.

System	Filter order	Trigram		RNN	
		Sim	Real	Sim	Real
Q1	—	7.30	25.3	6.42	23.5
T2	2	7.28	23.3	6.19	22.5
T3	3	7.19	24.6	6.36	23.0
T4	4	7.24	24.4	6.56	23.3
T5	5	7.35	24.1	6.48	23.5
T6	6	7.47	24.3	6.57	23.9
T7	7	7.62	25.2	6.68	23.8
T8	8	7.70	27.0	6.74	26.5
T9	9	7.55	26.8	6.71	25.6

and 14.8% with trigram and RNN decoding, respectively, compared with P1 system for RealData.

Table 3. %WERs of proposed CNN-DNN-HMM systems using the same temporal coverage as the baseline. Inputs to the systems were signals dereverberated with WPE.

System	Filter order (K)	Trigram		RNN	
		Sim	Real	Sim	Real
Q1	—	7.30	25.3	6.42	23.5
U2	2	7.25	23.9	6.43	22.4
U3	3	7.02	24.6	6.25	22.7

One possible concern is that the gains obtaind in the experiments described above resulted from the extended temporal coverage of the CNN-DNN-HMMs rather than the use of CNNs because a convolution operation effectively extends the period of time that an overall system can cover. To check that this was not the case, we performed experiments using a truncated context window in CNN-DNN-HMM systems. Here, we spliced convolution layer outputs by using a $(4-K+1)$-left 4-right context window where K is the filter order so that the time period covered by the CNN-DNN-HMM systems became the same as that of the baseline system. Table 3 shows the WER results. We can see that using the CNNs produced similar gains to those in Table 2, which implies that the convolutional structure is the primary cause of the performance gains. Although the results described above show that CNN-DNN-HMMs with temporal convolution are more robust against mismatches between training and test environments than conventional DNN-HMMs, it has yet to be clarified whether this improvement is the result of capturing the characteristics of far-field speech. Further investigation is required to clarify the impact of temporal convolution-based CNNs under various conditions.

Table 4. Comparison of full-matrix convolution and diagonal-matrix convolution. Dereverberated signals and trigram decoding was used.

Filter order (K)	Sim		Real	
	Full (Tx in Tab.2)	Diag	Full (Tx in Tab.2)	Diag
2	7.28	7.11	23.3	25.0
4	7.24	7.28	24.4	24.1
6	7.47	7.41	24.3	24.7
8	7.70	7.55	27.0	25.4

The performance with diagonal convolution matrices was also investigated. Table 4 compares the performance of a diagonal convolution scheme with that of a full convolution scheme. From these results, the diagonal convolution scheme appears to be slightly less sensitive to the choice of filter order probably because of its fewer adjustable parameters. Note that the diagonal convolution scheme can be used only for filter bank features since it assumes that reverberation affects each feature dimension separately.

5. CONCLUSION

In this paper, we examined the performance of CNN-DNN-HMM acoustic models with convolution in time for far-field speech recognition. While the CNN-DNN-HMM approach did not improve the recognition performance in seen test environments, meaningful gains were obtained for unseen real recordings. Since our results indicated the CNN-DNN-HMM approach to be useful for modelling short time correlations, the approach was further combined with a dereverberation technique based on signal-space convolution to cope with both short- and long-term correlations. The gains from the two approaches were found to be additive. From our results, it is still unclear whether the performance gains achieved in this paper are limited to far-field set-ups and this will be investigated in future work. Another interesting direction of future work would be to adapt or retrain the convolution layer to test environments. Since the number of convolution parameters is limited, especially with a diagonal convolution scheme, convolution layer retraining is expected to allow for rapid adaptation.

A. WPE METHOD FOR SIGNAL-SPACE DEREVERBERATION

WPE aims at removing the reverberation effect from an observed acoustic signal, thus generating a less reverberant signal. The method is based on linear prediction, which means that convolution is performed in the signal space.

Let $y_t[k]$ denote the STFT coefficient of the observed signal, where t and k are the time frame and frequency bin indices, respectively. WPE convolves the sequence of the STFT coefficients, $(y_k[t])_{0\le t<T}$, with a linear filter in each frequency bin as follows:

$$x_t[k] = y_t[k] - \sum_{\tau=T_\perp}^{T_\top} g_\tau^*[k] y_{t-\tau}[k], \tag{2}$$

where $*$ stands for complex conjugation and T denotes an utterance length measured with the number of frames. $T_\perp$ and $T_\top$ define the time period in which the filter has an effect. $T_\perp$ is normally set at 3 while $T_\top$ has a large value to deal with long-term reverberation. $G = (g_{T_\perp}, \cdots, g_{T_\top})$ defines the filter and needs to be optimised. The frequency bin index k is omitted below.

The filter G is optimised to minimise the following objective function:

$$F_{\mathrm{WPE}} = \sum_{t=0}^{T-1} \left(\frac{\left| y_t - \sum_{\tau=T_\perp}^{T_\top} g_\tau^* y_{t-\tau} \right|^2}{\theta_t} + \log \theta_t \right), \tag{3}$$

where $\Theta = (\theta_t)_{0\le t<T}$ is a set of auxiliary variables. These variables need to be optimised jointly with G, which leads to interleaved updates of G and Θ. After optimisation has been completed, the resultant filter is applied to $(y_t)_{0\le t<T}$ to generate dereverberated STFT coefficients $(x_t)_{0\le t<T}$, followed by waveform synthesis with an overlap add method.

B. REFERENCES

[1] N. Morgan and H. Bourlard, "Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Process. Mag.*, vol. 12, no. 3, pp. 24–42, 1995.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

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

[4] T. Yoshioka and M. J. F. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Comp. Speech, Language*, vol. 31, no. 1, pp. 65–86, 2015.

[5] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2014, pp. 172–176.

[6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Worksh.*, 2014.

[7] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[8] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

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

[10] M. J. F. Gales and Y.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2011, pp. 121–126.

[11] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1676–1691, 2010.

[12] Y.-Q. Wang and M. J. F. Gales, "Improving reverberant VTS for hands-free robust speech recognition," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2011, pp. 113–118.

[13] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5527–5531.

[14] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, pp. 359–366, 2001.

[15] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Eindhoven University of Technology, 2006.

[16] K. Kumar and R. Stern, "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4282–4285.

[17] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2013, pp. 138–143.

[18] L. Tóth, "Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 190–194.

[19] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proc. Interspeech*, 2013, pp. 3366–3370.

[20] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2013, pp. 315–320.

[21] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[22] X. Xiao, Z. Shengkui, D. H. H. Nguyen, Z. Xionghu, D. Jones, E.-S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB Worksh.*, 2014.

[23] F. J. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. T. Geiger, B. W. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement," in *Proc. REVERB Worksh.*, 2014.

[24] V. Mitra, W. Wang, Y. Lei, A. Kathol, G. Sivaraman, and C. Espy-Wilson, "Robust features and system fusion for reverberation-robust speech recognition," in *Proc. REVERB Worksh.*, 2014.

[25] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2523–2527.

[26] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4656–4660.

[27] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[28] K. Kinoshita, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Worksh. Appl. Signal Process. Audio, Acoust.*, 2013.

[29] "The REVERB challenge—evaluating de-reverberation and ASR techniques in reverberant environments," http://reverb2014.dereverberation.com/.

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