Neural Network Front-ends Based Speech Recognition In Reverberant Environments

Zhen Zhang
National Computer Network Emergency Response
Technical Team/Coordination Center of China,
Beijing, China
zhangzhen@cert.org.cn

Peng Li
National Computer Network Emergency Response
Technical Team/Coordination Center of China,
Beijing, China
lipeng@cert.org.cn

***Abstract*—This paper presents an investigation of reverberant speech recognition using frond-ends based methods. A 2-channel dereverberation method is adopted to achieve robust dereverberation under different reverberant conditions. Also a 2-channel spectral enhancement method is used where the gain of each frequency bin is controlled by acoustic scene, which is detected based on the analysis of full-band coherent property. Deep Neural Network (DNN) is also presented as a feature extractor. The DNN based front-end allows a very flexible integration of meta-information. Bottle neck features is extracted in place of MFCC features used in HMM-GMM system. We evaluated our methods on the data provided by REVERB challenge. On simulated data, the DNN front-end yields more than 33% relative reduction in Word Error Rate (WER).**

Keywords—REVERB Challenge, deep neural networks, Dereverberation, speech recognition

I. INTRODUCTION

Reducing the ASR performance gap between reveberant speech and close-talking recordings has been an important research topic for a long time. Lots of researches have proved that audio processing is helpful in improving the quality of the reverberant speech. Among the front-end signal processing technologies, three categories of dereverberation methods are generally applied: 1) beamforming using microphone arrays, 2) spectral enhancement, 3) blind system identification and inversion[1].Spectral enhancement based dereverberation shows superiority due to its robustness in both reverberant and noisy environment[2].Fractional time delay alignment filter is applied to the reverberant signal, and the acoustic scene is classified by analyzing the coherent component. Based on the acoustic scene, an appropriate spectral enhancing scheme is selected to eliminate the interference as much as possible while keeping the speech distortion always in a low level.

Recently, acoustic model based on the Deep Neural Networks (DNNs) has gained popularity with the consistent improvement in recognition performance over earlier Neural Network based front-ends (e.g. [3]). DNNs are either deployed as the front-end for standard Hidden Markov Model based on Gaussian Mixture Models (HMM-GMMs), or in a hybrid form to directly estimate state level posteriors. As noted in several publications [4,5,6,7], DNNs show general word error rate (WER) improvements on the order of 10-30% relative across a variety of small and large vocabulary tasks when compared with HMM-GMMs built on classic features (e.g. MFCC, PLP). Neural network based features have long been used successfully in speech recognition [8,9,10]. While the early research did not involve deep layers [4], the path towards deep learning was laid in the stacking of bottleneck networks [9]. A DNN is a conventional Multi-Layer Perceptron (MLP) with many internal or hidden layers. The BottleNeck (BN) features extracted from the internal layer with a relatively small amount of neurons have been shown to effectively improve the performance of ASR systems. It is possibly due to the limited number of units which creates a constriction in the network and further forces the information pertinent to classification into a low dimensional representation [11,5,12]. In many ASR systems, the neural network based features performs better than the cepstrum or spectrum based features (e.g. MFCC, PLP).

According to REVERB challenge, the reverberant data is simulated or recorded in various rooms with different distances between source and microphones[13,14,15], and three kinds of utterance are provided: 1-channel, 2-channel and 8-channel. We choose the 2-channel dereverberation method for both Speech Enhancement (SE) task and Automatic Speech Recognition (ASR) task. Work in this paper aims to extend and enhance the preliminary work in [16] with advanced DNN based feature extraction. We followed the standard strategies where the DNN includes one input layer, three hidden layers, another bottleneck layer and one output layer. The BN features are extracted from the bottleneck layer of DNN trained to predict the context-dependent clustered triphone states.

The rest of this paper is organized as follows. Section II introduces the algorithms of 2-channel dereverberation and DNN based feature extractor. Section III presents experiment results. Finally, section IV concludes the paper and discusses future work.

II. RECOGNITION OF REVERB SPEECH

A. *System Overview*

Figure 1 shows the overall system proposed for REVERB challenge. It contains two basic modules, front-end module and back-end evaluation module. In the front-end module,

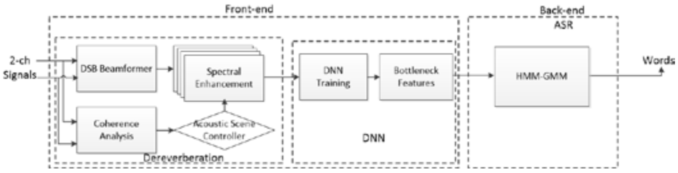

Fig.1. System structure

different from our previous system, we adopt a DNN based feature extractor. For consistent comparison, we use the same recognizer provided by REVERB challenge organizer.

B. Dereverberation and Noise Reduction

Spectral enhancement methods show their priority on the robustness in the condition of both noise and reverberation. In this section, an acoustic scene aware dereverberation method is utilized for the purpose of achieving environmental adaptation [16]. Two sensors are utilized since it is the basic topology of all microphone arrays and can be generalized to any other microphone arrays conveniently.

1) Signal model

Let *s(n)* represents the target clean signal, and $x_m(n)$ is the time-aligned signal at sensor $m(m=1,2)$. $n_m(n)$ is the noise of environment. $h_m(t)$ can be seen as the time-delayed RIR which is the convolution of real RIR from target signal to sensor $m(m=1,2)$ and alignment filter where the alignment filter is designed to steer the direction of target speech signal. The observed signal each channel can be expressed as follows.

$$x_m(n) = h_m(t) * s(n) + n_m(n) \tag{1}$$

Applying STFT to the 16kHz time-aligned signal, we have sinal expression at l th frame and k th frequency bin in time-frequency domain.

$$x_m(l,k) = H_m(l,k)S(l,k) + N_m(l,k) \tag{2}$$

2) Spectral enhancement

Spectral Enhancement method has a generalized form. The estimate of the amplitude spectrum of the target signal can be expressed as follows.

$$|\hat{S}(l,k)| = G(l,k)\,|\hat{X}(l,k)| \tag{3}$$

$G(l,k)$ is the gain estimated on each frequency bin and $|\hat{X}(l,k)|$ is the amplitude spectrum of signal to be enhanced. Before overlap-and-add scheme, regardless of leakage between frequency bins causing by STFT, both $G(l,k)$ and $|\hat{X}(l,k)|$ should chosen cautiously versus distortion to achieve robustness.

3) Environmental adaptation based processing

Reverberation, especially late reverberation, shows isotropic property as well as the environmental diffuse noise, while the direct sound shows strong coherent property [17]. There are two main acoustic scenes of the room we should blindly aware. One is the reflection condition which can be interpreted by reverberation time (T60) and the other one is the speake-mic distance. By estimating the proportion of coherent component, the effects of two are synthesized. All the diffuse part can be seen as noise to be filtered. We follow the Coherent-to-Diffuse energy Ratio (CDR) estimation in [18], which is expressed as follows.

$$\varepsilon(e^{j\Omega}) = \frac{|\sin c(\Omega f_S d_{mic}/c)|^2 - |\Gamma_{X_1X_2}(e^{j\Omega})|^2}{|\Gamma_{X_1X_2}(e^{j\Omega})|^2 - 1} \tag{4}$$

$\varepsilon(e^{j\Omega})$ is CDR estimator of each frequency bin and $\Gamma_{X_1X_2}$ is the expression of complex coherence function [19]. And a Wiener filter can be formed based on the estimation of CDR which can filter the non-coherent part. According to [18], it's more accurate when CDR is relatively large. Furthermore, We use global CDR, denoted by $\hat{\varepsilon}$, as an full-band acoustic scene aware controller indicating the level affected by reverberation and diffuse noise which can be interpreted as a direct sound activity detector to achieve environmental adaptation.

Up to now, lots of 1-channel and 2-channel dereverberation methods are proved efficiency under the framework of spectral enhancement and achieve robustness to noise compared with the methods using inverse filtering. Fixed beamforming, such as Delay-and-Sum Beamformer (DSB), helps to suppress the reverberation based on priori DOA information though its suppression ability is limited. Late reverberaion suppressing method using generalized statistic model of reverberation [20] shows outstanding performance especially when reverberation is strong. Based on the controller mentioned above, we compared $\hat{\varepsilon}$ with three constants σ_1 , σ_2 , σ_3 (from large to small, chosen 15dB, 10dB, 5dB).

Spectral enhancement strategy suggested above separates the acoustic scene into four cases. In the first case, the acoustic scene is ideal so that the speech signal recorded is very close to clean speech. In the second and third cases, a moderate tradeoff between dereverberation and signal distortion is achieved. In the last case, more reverberation reduction means better performance when reverberation is strong enough.

To avoid music noise, both time recursive and adjacent frequency gain smoothing are conducted. The recovered signal is obtained by inverse STFT and overlap-and-add scheme. The phase of recovered speech signal equals the noisy phase of beamformer output. All the processing is with windows of 512 points and step-size of 256 points which means the result of a DFT with length 512 (32 ms) at a shift of 16 ms.

C. DNN based features

Figure 2 shows the architecture. It is similar to those in [3,21,22]. All DNNs are trained feed-forward with the TNET[1] on GPUs. In a default TNET setup, 31 adjacent frame MFCC features are decorrelated and compressed with DCT into a dimension of 624（$31\times 39 \rightarrow 16\times 39$ .Global mean and vari-

[1] http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet

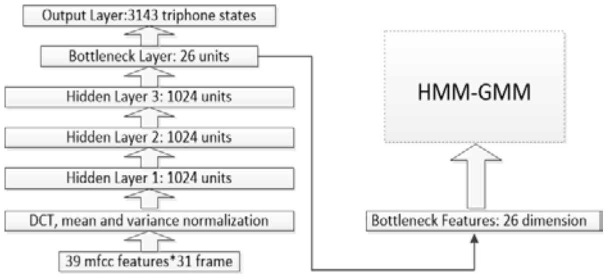
Output Layer:3143 triphone states
Bottleneck Layer: 26 units
Hidden Layer 3: 1024 units
Hidden Layer 2: 1024 units
Hidden Layer 1: 1024 units
DCT, mean and variance normalization
39 mfcc features*31 frame
HMM-GMM
Bottleneck Features: 26 dimension

Fig.2. DNN module structure

ance normalization are performed on each dimension before feeding as the DNN input. The 6 layered DNN structure is shown in Figure 2. On average 10% data in training set is reserved for cross validation in DNN training. The training stops automatically when the improvement of frame-based target classification accuracy on the cross validation set drops to below 0.1%.

The bottleneck layer is placed just before the output layer, as in our initial experiments this topology gives the best performance. We set to 26 dimension according to our experiments. DNNs are trained on classification targets of 3143 triphone states. In the bottleneck layer, linear BN features are extracted before the sigmoid activation.

III. Experiment

A. Data and system set

In the REVERB challenge, both simulated and really recorded data are provided. The simulated data (SimData) is convolved by clean utterance from WSJCAM0 corpus [13] with the recorded room impulse response (RIR) in different rooms. The reverberation time of the rooms are 250ms, 500ms and 700ms respectively. Recorded background noise is added to the reverberant data at a fixed signal-to-noise ratio (SNR) of 20dB. The really recorded data (RealData), utterances from the MC-WSJ-AV corpus [14], consists of utterances recorded in a noisy and reverberant room with reverberation time of 700ms. Both SimData and RealData include two types of distances between the speaker and microphone array (near=50cm and far=200cm). The test data are from the SimData and RealData databases under the following important assumptions. First, there is no drastic change in RIR within an utterance. Second, relative speaker-microphone position changes from utterance to utterance, which means the direction of arrival (DOA) of the target speech signal is uncertain, and this is essential to our dereverberation method. The recording 8-channel cicular array has diameter of 20cm and the 2-channel microphone distance, denoted by d_{mic} , can be calculated.

Under the framework of HTK based recognizer organizer provided, we re-train the acoustic model of "multi-condition" using HMMs structure and DNN respectively. The proper starting point is that the artificially distorted training signals are mismatch with the enhanced ones. Therefore, Substituting the 7861 reverberant noisy utterances by the enhanced signal and enlarging the re-training data with 7861*24 enhanced convolving utterances, we provide ASR result of this re-trained acoustic model. Then the seven possible cases are:

Clean+noEnh:"clean-condition" HMMs without dereverberation;

Multi+noEnh:"multi-condition" HMMs without dereverberation;

Clean+Enh:"clean-condition" HMMs with dereverberation;

Multi+Enh:"multi-condition" HMMs with dereverberation;

ReTrn+Enh:re-trained "multi-condition" HMMs with dereverberation;

DNN+Multi+Enh: "multi-condition" HMMs with features from DNN;

DNN+ReTrn+Enh: re-trained "multi-condition" HMMs with features from DNN;

B. Baseline experiments

For the ASR task, word error rate (WER) of test data is reported in I. Baseline models of ASR task as well as re-trained model are provided in this section. The "clean-condition" baseline system uses 39D mel-frequency cepstral coefficients (MFCCs) including Delta and Delta-Delta coefficients as features. As acoustic models, it employs tied-state HMMs with 10 Gaussian components per state trained according to the maximum-likelihood criterion [23]. All the training data for "clean-condition" HMMs is from WSJCAM0 corpus [13]. Further, the model is re-trained using the features of artificially distorted 7861 utterances to form the "multi-condition" HMMs. The utterances are in mixture with 24 kinds of RIRs and 6 kinds of noises. We test the enhanced signals on the two baseline systems both using and not using the unsupervised CMLLR model adaptation.

Tables I show the WER results on SimData and RealData datasets respectively. It includes the ASR results of "clean-condition" model, "multi-condition" model. WER of near, far data and their average are reported separately. As we can seen, "multi-condition" model performs better than "clean-condition" model. It achieves decrease on WER from 51.68% to 29.51% on average of SimData and from 88.53% to 56.94% on average of RealData. Consistent improvement across all recording conditions is achieved by using CMLLR which results in WER 25.25% on SimData and 48.85% on RealData.

C. 2-channel dereverberation

Performance of dereverberation is examined using both "clean-condition" and "multi-condition" acoustic model. According to table II, recognizing the test set with "clean-condition" model without CMLLR adaptation, the dereverberation method achieves decrease on WER from 51.68% to 37.06% on average of SimData and from 88.53% to 72.62% on average of RealData. Consistent improvement across all recording conditions is achieved by using CMLLR which results in WER 27.93% on SimData and 62.12% on RealData. However, recognizing with "multi-condition" model

TABLE I. Word error rate of baseline.

Test Data		**Word error rate(%)**									
		SimData							**RealData**		
		Room 1		**Room 2**		**Room 3**		**Ave.**	**Room 1**		**Av e.**
		Near	**Far**	**Near**	**Far**	**Near**	**Far**		**Near**	**Far**	
Clean+noEnh	NOCMLLR	18.06	25.38	42.98	82.20	53.54	88.04	51.68	89.72	87.34	88.53
	CMLLR	14.81	18.86	24.63	64.58	33.77	78.42	39.16	82.31	80.76	81.53
Multi+noEnh	NOCMLLR	20.60	21.15	23.70	38.72	28.08	44.86	29.51	58.45	55.44	56.94
	CMLLR	16.23	18.71	20.50	32.47	24.76	38.88	25.25	50.14	47.57	48.85

TABLE II. Word error rate of 2-channel dereverberation.

Test Data		**Word error rate(%)**									
		SimData							**RealData**		
		Room 1		**Room 2**		**Room 3**		**Ave.**	**Room 1**		**Ave.**
		Near	**Far**	**Near**	**Far**	**Near**	**Far**		**Near**	**Far**	
Clean+noEnh	NOCMLLR	17.43	25.25	27.85	49.48	36.51	65.94	37.06	73.91	71.34	72.62
	CMLLR	14.47	19.47	21.19	34.86	27.16	50.50	27.93	62.66	61.58	62.12
Multi+noEnh	NOCMLLR	23.64	36.46	27.72	37.69	34.00	45.85	34.22	59.95	59.49	59.72
	CMLLR	16.93	20.04	19.91	26.84	23.95	34.33	23.66	44.87	45.81	45.34
ReTrn+Enh	NOCMLLR	15.64	18.76	19.79	28.56	24.01	35.15	23.64	49.50	49.49	49.49
	CMLLR	**14.76**	**16.52**	**18.23**	**24.79**	**21.09**	**31.50**	**21.14**	**42.10**	**45.17**	**43.63**

without CMLLR adaptation, the performance (SimData: 34.22%, RealData: 59.72%) is worse than the "multi-condition" baseline (SimData: 29.51%, RealData: 56.94%) because the recognizing enhanced data is mismatch with the reverberant data used to train the "multi-condition" acoustic model, though using CMLLR gives a little improvement.

To overcome the mismatch, the re-trained "multi-condition" HMMs gives a better result. The optimized one has a WER 23.64% on average of SimData and 49.49% on average of RealData without CMLLR and finally 21.14% of SimData and 43.63% of RealData with CMLLR. The relative decreasing rates of WER are 59.09% and 50.7% each. What draws our attention is that the average WER of RealData is higher than that of SimData. Two reasons may cause the observation. One is that the utterances of RealData are not included in training set and another is that the simulated data can't imitate all the situations of real room environment.

D. *Bottleneck features from DNN*

TABLE III. WORD ERROR RATE OF BOTTLENECK FEATURES FROM DNN.

TEST DATA		WORD ERROR RATE(%)									
		SIMDATA							REALDATA		
		ROOM 1		ROOM 2		ROOM 3		AVE.	ROOM 1		AVE.
		NEAR	FAR	NEAR	FAR	NEAR	FAR		NEAR	FAR	
DNN+MULTI+ENH	NOCMLLR	17.67	23.36	23.56	29.39	30.35	39.36	27.28	59.60	58.71	59.16
	CMLLR	14.72	19.23	19.95	25.07	25.42	34.02	23.07	54.52	54.62	54.57
DNN+RETRN+ENH	NOCMLLR	16.89	18.25	18.68	23.80	21.64	27.97	21.21	51.36	51.13	51.25
	CMLLR	15.69	16.82	17.83	22.74	20.53	26.55	20.03	48.47	49.27	48.8

Table III shows the results for BN features based decoding. All the DNNs are trained with the triphone state targets force-aligned using training set. The decoding is performed on the test set. We achieved different performance without CMLLR and with CMLLR. On Multi+Enh" case, we got 33.43% relative decrease on SimData using BN features without CMLLR. But we only got 2.49% relative decrease with CMLLR. That is because we only use DNN as a front-end. During decoding process, we use the HMM-GMM structure of baseline system. CMLLR maybe not so work for the DNN features. Another reason lies the feature dimensions, we use 39 dimension MFCC features for our baseline system, but only 26 dimension BN features adopted in this paper. As for "ReTrn+Enh" case, we achieved 10.28% relate decrease compared with our results on 2-channel dereverberation. Similar situation for CMLLR, we only got 5.25% WER decrease on SimData. We can conclude that CMLLR works better on MFCC features than BN features. Another problem of BN features is we got worse performance on RealData, especially with CMLLR.That should be the mismatch between training data and testset and BN features are more sensitive than MFCC features. The further experiments will be investigated.

IV. CONCLUSION

We have presented out dereverberation approach to the REVERB challenge based on spectral enhancement. An acoustic scene aware technique is proposed to make dereverberation robust to different conditions. For ASR task, when it is combined with back-end ASR with matched training, it produces a significant decrease on WER.

The DNN based front-end was tested in the context of reverberant speech recognition. Experiments showed that on SimData, BN features gave an average 33.43% relative WER reduction over using MFCC features without CMLLR. Retraining did not give as much gain as we got on MFCC features. We also need to do further experiments on RealData about the performance difference between MFCC features and BN features.

References

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in Proc. Int. Workshop Acoust. Echo Noise Control, 2005.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 2, pp. 231–246, 2009.K. Elissa, "Title of paper if known," unpublished.

[3] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neckfeaturesforLVCSRofmeetings,"inAcoustics,Speechand SignalProcessing,2007.ICASSP2007.IEEEInternationalConference on, vol. 4, 2007, pp. IV–757–IV–760.

[4] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," J. Mach. Learn. Res., vol. 10, pp. 1–40, Jun. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1577069.1577070

[5] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in INTERSPEECH, 2011, pp. 437–440.

[6] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in AutomaticSpeechRecognitionandUnderstanding(ASRU),2013IEEE Workshop on, Dec 2013, pp. 285–290.

[7] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in ICASSP2014 - Speech and Language Processing (ICASSP2014 - SLTC), Florence, Italy, May 2014, pp. 5579–5583.

[8] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in Machine Learning for Multimodal Interaction, ser. LNCS. Springer Verlag, 2005, no. 3869, pp. 463–475.

[9] F. Grezl, M. Kafiat, and L. Burget, “Investigation into bottle -neck features for meeting speech recognition,” in Proc. Interspeech 2009, no. 9, 2009, pp. 2947–2950.

[10] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. el Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, “Transcribing meetings withtheAMIDA systems,” IEjeub2011blindEE Transactions on Audio, Speech and Language Processing, Aug. 2011.

[11] D.YuandM.L.Seltzer,“Improvedbottleneckfeaturesusingpretrained deep neural networks,” in INTERSPEECH, 2011, pp. 237–240.

[12] Z. Tüske, R. Schlüter, and H. Ney, “Deep hierarchical bottleneck MRASTA features for LVCSR,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, Canada, May 2013, pp. 6970–6974.

[13] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “Wsjcamo: A british english speech corpus for large vocabulary continuous speech recognition,” in Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, vol. 1. IEEE, 1995, pp. 81–84.

[14] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, “The multichannel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments,” in Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on. IEEE, 2005, pp. 357‑362.

[15] D.B.PaulandJ.M.Baker,“Thedesignforthewallstreetjournal-based csr corpus,” in Proceedings of the workshop on Speech and Natural Language. AssociationforComputationalLinguistics,1992,pp.357– 362.

[16] X. Wang, Y. Guo, X. Yang, Q. Fu, and Y. Yan, “Acoustic scene aware dereverberation using 2-channel spectral enhancement for reverb challenge,” REVERB(REverberant Voice Enhancement and Recognition Benchmark) Challenge, May 2014.

[17] H. Kuttruff, Room acoustics. Taylor & Francis, 2000.

[18] Jeub, Marco, et al. "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals." *Signal Processing Conference, 2011 19th European*. IEEE, 2011.

[19] I. A. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” Speech and Audio Processing, IEEE Transactions on, vol. 11, no. 6, pp. 709–716, 2003.

[20] E. A. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” Signal Processing Letters, IEEE, vol. 16, no. 9, pp. 770–773, 2009.

[21] F. Grezl and P. Fousek, “Optimizing bottle-neck features for LVCSR,” inAcoustics,SpeechandSignalProcessing,2008.ICASSP2008.IEEE International Conference on, 2008, pp. 4729–4732.

[22] K. Veselý, M. Karafiát, and F. Grézl, “Convolutive bottleneck network features for LVCSR.” in ASRU, D. Nahamoo and M. Picheny, Eds. IEEE, 2011, pp. 42–47. [Online]. Available: http://dblp.unitrier.de/db/conf/asru/asru2011.html#VeselyKG11

[23] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R.Haeb-Umbach,V.Leutnant,A.Sehr,W.Kellermann,R.Maasetal., “Proceedingsoftheieeeworkshoponapplicationsofsignalprocessing to audio and acoustics (waspaa-13),” THE REVERB CHALLENGE: A COMMONEVALUATIONFRAMEWORKFORDEREVERBERATION AND RECOGNITION OF REVERBERANT SPEECH, May 2013.