Real-Time Dereverberation for Deep Neural Network Speech Recognition

Andreas Schwarz, Christian Huemmer, Roland Maas, Walter Kellermann

Lehrstuhl für Multimediakommunikation und Signalverarbeitung, Friedrich-Alexander-Universität Erlangen-Nürnberg

Cauerstr. 7, 91052 Erlangen, Germany. E-Mail: {schwarz,huemmer,maas,wk}@lnt.de

Abstract

We evaluate a real-time multi-channel dereverberation method for the application to speech recognition with deep neural networks (DNN). The dereverberation method is based on modeling the reverberated signal as a mixture of a fully coherent direct path signal and a diffuse reverberation component, and estimating the coherent-to-diffuse power ratio (CDR) from the spatial coherence of the signals. The method can operate in real-time, i.e., without requiring processing of entire utterances. We compare CDR estimators which are "blind", i.e., do not require information about the direction of arrival (DOA) of the target signal, with estimators which make use of a DOA estimate. The impact of the dereverberation method on speech recognition accuracy with different DNN-based acoustic models is investigated with the REVERB challenge corpus and the Kaldi speech recognition toolkit.

Introduction

Most dereverberation methods require the collection of a considerable amount of data for the estimation of parameters before effective signal enhancement is possible. For example, approaches based on temporal decay models [1] require the estimation of the decay rate, approaches based on linear prediction require adaptation of the prediction filter [2], and methods based on channel inversion require estimates of the impulse responses from the source to the microphones [3]. In typical mobile speech recognition applications, voice is transmitted to a server to perform speech recognition. Due to the limited capacity of the transmission channel, and the desire to provide feedback while the user is still talking, transmission of voice should start as early as possible, without having to wait for several seconds or until the end of an utterance to perform signal processing. This requires a front-end which can operate with a low delay and does not require long-term estimation of signal or environmental characteristics. One such processing method is postfiltering based on the assumption of an uncorrelated [4, 5] or diffuse [6, 7] reverberation component.

In this paper, we evaluate the effect of coherence-based dereverberation on automatic speech recognition (ASR) with Deep Neural Networks (DNNs). First, we review the signal model and the concept of signal enhancement based on an estimate of the coherent-to-diffuse power ratio (CDR), of which a detailed description can be found in [7]. Then, we describe the direction of arrival (DOA)-independent and DOA-dependent CDR estimators which are used for the evaluation in this paper, as well as methods for DOA estimation. We describe the evaluated hybrid DNN-HMM ASR system, and compare the impact of dereverberation using the different CDR estimators on ASR word error rate (WER).

Signal Model

We consider a reverberated and noisy speech signal recorded by two omnidirectional microphones. The signal $x_i(t)$ recorded at the i-th microphone is composed of the coherent desired signal component $s_i(t)$ and the diffuse undesired component $n_i(t)$ comprising additive noise and late reverberation, i.e., $x_i(t) = s_i(t) + n_i(t),\ i = 1, 2$. The microphone, desired, and noise signals are represented in the short-time Fourier transform (STFT) domain by the corresponding uppercase letters, i.e., $X_i(k, f)$, $S_i(k, f)$ and $N_i(k, f)$, respectively, with the discrete frame index k and continuous frequency f, and the auto- and cross-power spectra $\Phi_{x_i x_j}(k, f)$, $\Phi_{s_i s_j}(k, f)$, $\Phi_{n_i n_j}(k, f)$. Note that the continuous frequency f is used here for generality; in practice, f denotes discrete values along the frequency axis. It is assumed that the auto-power spectra of all signal components are identical at both microphones, i.e., $\Phi_{s_i s_i}(k, f) = \Phi_s(k, f), \Phi_{n_i n_i}(k, f) = \Phi_n(k, f)$. The time- and frequency-dependent CDR can then be defined as

$$CDR(k, f) = \frac{\Phi_s(k, f)}{\Phi_n(k, f)}. \tag{1}$$

The complex spatial coherence functions of the desired signal and noise components are given by

$$\Gamma_s(f) = \frac{\Phi_{s_1 s_2}(k, f)}{\Phi_s(k, f)},\ \Gamma_n(f) = \frac{\Phi_{n_1 n_2}(k, f)}{\Phi_n(k, f)}, \tag{2}$$

and are assumed to be time-invariant, i.e., dependent only on the spatial characteristics of the signal components. It is furthermore assumed that signal and noise components are orthogonal, such that $\Phi_x(k, f) = \Phi_s(k, f) + \Phi_n(k, f)$. The complex spatial coherence of the mixed sound field can then be written as a function of the CDR and the signal and noise coherence functions:

$$\Gamma_x(k, f) = \frac{CDR(k, f)\Gamma_s(f) + \Gamma_n(f)}{CDR(k, f) + 1}. \tag{3}$$

The direct sound is now modeled as a coherent plane wave with a time difference of arrival Δt, while the undesired noise and late reverberation component is modeled as a diffuse (spherically isotropic) sound field. The corresponding spatial coherence functions for the direct and diffuse sound components are given by

$$\Gamma_s(f) = e^{j2\pi f \Delta t}, \tag{4}$$

$$\Gamma_n(f) = \Gamma_{\text{diffuse}}(f) = \text{sinc}(2\pi f d c^{-1}), \tag{5}$$

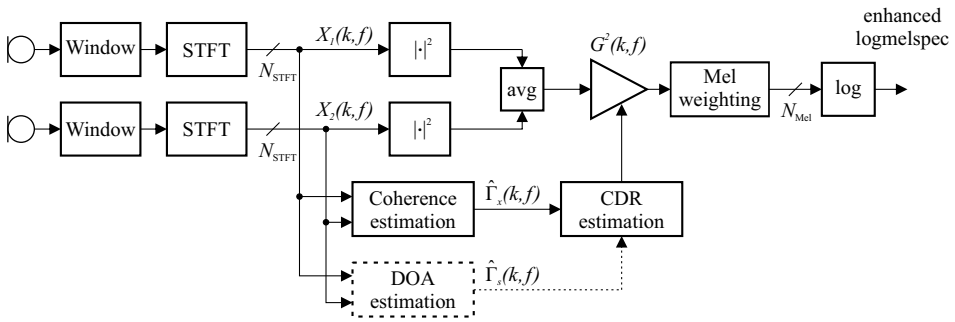

Figure 1: CDR-based dereverberation system [7] applied to logmelspec feature extraction

respectively, where c is the speed of sound. The direct signal coherence has a magnitude of one with a phase determined by the time difference of arrival (TDOA), while the diffuse noise coherence only depends on the known microphone spacing d.

For spectral enhancement it is desired to estimate the CDR from the short-time coherence of the mixed sound field $\Gamma_x(k,f)$. This coherence is first estimated as

$$\hat{\Gamma}_x(k,f) = \frac{\hat{\Phi}_{x_1x_2}(k,f)}{\sqrt{\hat{\Phi}_{x_1x_1}(k,f)\hat{\Phi}_{x_2x_2}(k,f)}}, \qquad (6)$$

where the spectral estimates $\hat{\Phi}_{x_ix_j}(k,f)$ are obtained by recursive averaging as follows:

$$\hat{\Phi}_{x_ix_j}(k,f) = \lambda\hat{\Phi}_{x_ix_j}(k-1,f) + (1-\lambda)X_i(k,f)X_j^*(k,f), \qquad (7)$$

with a constant smoothing factor λ between 0 and 1, chosen such that the time constant of the estimate is on the order of the stationarity interval of speech ($\ll$ 100 ms).

Coherence-based Spectral Enhancement

The STFT-domain spectral enhancement system evaluated in this paper is shown in Figure 1; a detailed description can be found in [7]. First, the squared magnitude spectra of both microphone signals are averaged, before the postfilter is applied to suppress diffuse signal components. The postfilter gain $G(k,f)$ is computed from an estimate of the CDR, which is obtained from an estimate of the spatial coherence $\hat{\Gamma}_x(k,f)$. The CDR can be estimated either with a "blind", i.e., DOA-independent estimator, which requires only the diffuse noise coherence model, or with a DOA-dependent estimator which additionally requires an estimate of the DOA or the target signal coherence $\hat{\Gamma}_s(k,f)$. The enhanced STFT-domain output spectrum after the postfilter is then transformed into logmelspec features using N_{Mel} Mel-spaced triangular weighting filters.

Blind (DOA-independent) CDR Estimation

In [7, 8] it is shown that knowledge or explicit estimation of the DOA is not required for the CDR estimation, since (3) can be solved for the CDR without requiring knowledge of $\Gamma_s(f)$. The corresponding blind (DOA-independent) estimator equation is given by [8, $\widehat{CDR}_{\text{prop3}}$]. Using this estimator, the proposed feature extraction system allows real-time operation, requiring only the quasi-instantaneous estimation of the short-time spectra according to (7).

DOA-dependent CDR Estimation

Several different DOA-dependent CDR estimators have been proposed [7], based on earlier optimum postfilter derivations for diffuse noise [6]. While they behave identically under ideal conditions, assuming the target DOA is exactly known, postfilters based on these estimators do not only suppress diffuse components, but also have an additional directional filtering effect. Also, they are more tolerant towards the estimation variance of the short-time coherence and deviation of the noise/reverberation coherence from the ideal diffuse model. For these reasons, they generally lead to stronger suppression of diffuse noise and better practical performance than the blind estimator [7]; on the other hand, they require the estimation of the target DOA.

The best-performing DOA-dependent estimator $\widehat{CDR}_{\text{prop2}}$ proposed in [7] is evaluated in this paper. We do not explicitly estimate a DOA, but estimate the direct signal coherence $\hat{\Gamma}_s(k,f)$ from the noisy speech signal by:

$$\hat{\Gamma}_s(k,f) = \exp\{j \arg \bar{\Gamma}_x(k,f)\}, \qquad (8)$$

where $\bar{\Gamma}_x(k,f)$ signifies a long-term estimate of the coherence between the microphone signals, in contrast to the short-time coherence estimate $\hat{\Gamma}_x(k,f)$. This long-term estimate is obtained analogously to (6) either from spectra which are averaged over an entire utterance, or, for real-time implementation, spectra which are recursively averaged as in (7), but using a higher smoothing factor.

ASR Engine

We employ the Kaldi toolkit [9] as ASR back-end system, configured in the same way as described in [10]. The WSJ0 trigram 5k language model of the REVERB challenge is used, and an acoustic model with 3551 context-dependent triphone states. We set up a GMM-HMM baseline system (see [11, 12] for a detailed description) trained on the clean WSJCAM0 Cambridge Read News REVERB corpus [13]. The alignment of the training data to the HMM states is then extracted from the clean training data and used for the later multi-condition training

Table 1: ASR Word Error Rate for the REVERB challenge evaluation and development test sets.

Acoustic Model	Enhancement	Evaluation Set										Development Set	
		SimData							RealData			SimData	RealData
		Room 1		Room 2		Room 3		**Avg**	Room 1		**Avg**	**Avg**	**Avg**
		near	far	near	far	near	far		near	far			
clean	none	7.91	**14.30**	32.53	79.75	44.90	87.15	44.42	85.79	85.52	85.66	46.27	85.13
	DOA-independent	**7.86**	17.74	16.25	47.50	24.74	67.55	30.27	72.31	66.27	69.29	31.70	65.25
	DOA-dep., utterance	8.32	18.21	**15.69**	**43.63**	23.18	66.19	29.20	**71.12**	66.58	68.85	30.49	**64.34**
	DOA-dep., realtime	8.55	17.70	15.74	43.74	**23.11**	**65.85**	**29.12**	71.57	**65.94**	**68.76**	**30.30**	64.83
multi-condition	none	**5.74**	**6.67**	7.65	13.92	8.65	14.62	9.54	28.45	29.14	28.80	9.68	24.93
	DOA-independent	6.61	7.12	7.65	12.18	8.32	14.57	9.41	28.46	29.07	28.77	**9.13**	25.29
	DOA-dep., utterance	6.98	7.33	**7.10**	**11.66**	8.13	14.30	**9.25**	**27.98**	**27.68**	**27.83**	9.36	**24.13**
	DOA-dep., realtime	6.79	7.53	7.16	11.82	**8.04**	**14.23**	9.26	29.03	29.25	29.14	9.51	25.08

of a hybrid DNN-HMM system. The DNN is a maxout network [14] with 2-norm nonlinearities/activation functions and 4 hidden layers, each with an input dimension of 2000 and an output dimension of 400.

We use features consisting of $N_{\text{Mel}} = 24$ static logmelspec coefficients, generated with or without applying coherence-based spectral subtraction in the STFT domain as described in the previous sections. Also, Delta (Δ) and acceleration ($\Delta\Delta$) coefficients are appended, and mean and variance normalization and ± 5 frame splicing is applied to the entire resulting feature vector. Note that, although the features are computed in real-time, the mean and variance normalization is performed per utterance here; for real-time decoding before the entire utterance is available, this would need to be modified.

The training is performed on the REVERB multi-condition training set [15], consisting of 7861 noisy and reverberated utterances from the WSJCAM0 corpus, using greedy layer-wise supervised training, preconditioned stochastic gradient descent, "mixing up" [14] as well as final model combination [14]. The multi-condition data is processed by the proposed spectral enhancement before training. For comparison, we also train an acoustic model only on clean speech without noise and reverberation.

Evaluation

We evaluate the proposed system on the two-channel task of the REVERB challenge [15]. The REVERB evaluation test set consists of ∼5000 reverberant and noisy utterances, partially created by convolution of clean WSJCAM0 utterances with impulse responses and mixing with recorded noise sequences ("SimData"), and partially consisting of multichannel recordings of speakers in a reverberant and noisy room from the MC-WSJ-AV corpus ("RealData"). For SimData, the reverberation times of the three rooms are approx. 0.25 s, 0.5 s and 0.7 s and the source-microphone spacing is 0.5 m (near) or 2 m (far). For RealData, the reverberation time is approx 0.7 s and the source-microphone distance is 1 m (near) or 2.5 m (far). In both cases, an 8-channel circular microphone array with a diameter of 20 cm was used, of which two microphones with a spacing of $d = 8$ cm are selected for the two-channel recognition task which is evaluated here. STFT frame length is 25 ms with 10 ms shift, the smoothing factor for the coherence estimation is $\lambda = 0.68$.

Figure 2 shows the logmelspec features generated from the noisy and reverberant, enhanced (DOA-independent, DOA-dependent with utterance-based direct signal coherence estimation or real-time estimation by recursive averaging) and clean signals of an utterance from the REVERB corpus. It is noticeable that the spectral enhancement achieves a significant reduction of both reverberation and background noise. A notable difference between the DOA-dependent and -independent enhancement methods is that the noise floor is suppressed to a higher degree for the DOA-dependent enhancement, otherwise the visual difference is not significant.

Results for the WER are given in Table 1. We compare results for logmelspec features with no spectral enhancement (corresponding to the presented scheme with a constant gain $G(k, f) = 1$), and spectral enhancement using the DOA-independent or DOA-dependent estimator, the latter either using utterance-based or real-time estimation of the direct signal coherence. Results are given both for the clean-speech acoustic model and for the multi-condition-trained acoustic model, where for multi-condition training, the signals are preprocessed by the same spectral enhancement method as used for the respective evaluation.

For the clean-speech acoustic model, all variants of the proposed signal enhancement method lead to a significant WER reduction; an exception is Room 1, which has a very low amount of reverberation, and where the benefit of suppressed reverberation is consequently outweighed by the inevitable distortion introduced by the spectral subtraction. Also, the DOA-dependent estimators have a small but consistent advantage over the DOA-independent estimator, which is in line with earlier results for a small-vocabulary HMM-GMM-based ASR system trained on clean speech [7].

For the multi-condition acoustic model, the overall WER is dramatically reduced w.r.t. the clean speech model, however the improvement by spectral enhancement is marginal. Apparently the temporal characteristics of reverberation can be exploited by the DNN very effectively, and while the spectral enhancement uses additional spatial information which is not available to the DNN, the use of this information by applying spectral

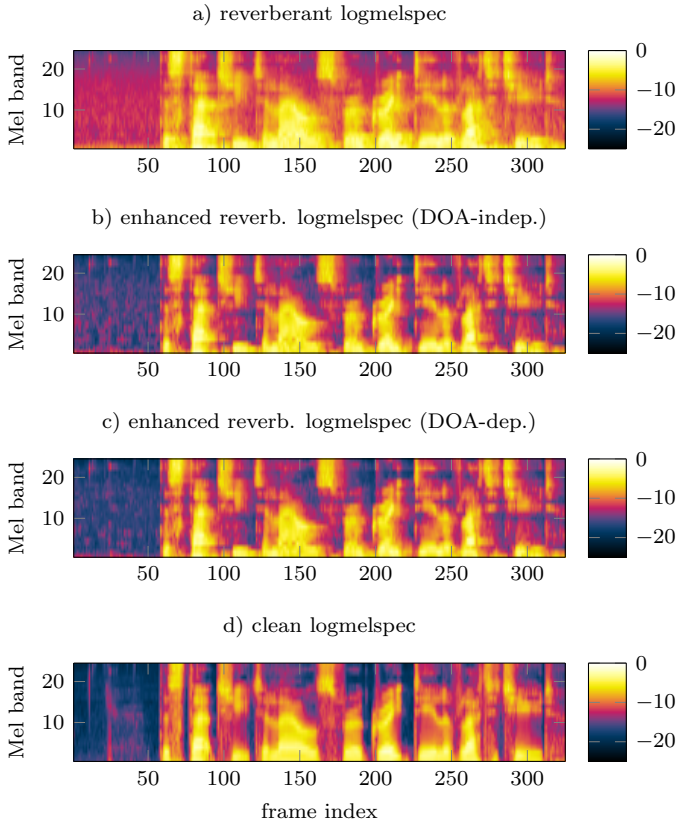

Figure 2: Features for the utterance "The statute allows for a great deal of latitude".

subtraction, along with the distortion it introduces, is apparently not beneficial. In [10], an alternative approach is described, where, instead of using spatial information to perform spectral enhancement, coherence-based *meldiffuseness* features features are extracted and appended to the noisy logmelspec feature vector (replacing the $\Delta\Delta$ coefficients of the logmelspec features). This approach was found to lead to a significant WER reduction.

Conclusion

We described a spectral enhancement system which works in real-time using instantaneous spatial coherence estimates. The system was shown to have a significant effect on recognition performance with a DNN-based acoustic model trained on clean speech, as similarly observed for a GMM-HMM-based recognizer and MFCC features in [7]. However, for a DNN acoustic model trained on multi-condition reverberated and noisy data, the advantage of the spectral enhancement was found to become insignificant. In [10], it is shown that, although the use of spatial coherence-based spectral subtraction does not lead to significant improvement, the same signal model can be used to extract coherence-based *meldiffuseness* features, which, when used as feature for the DNN, can significantly improve recognition performance.

References

[1] K. Lebart, J.-M. Boucher, and P. N. Denbigh. "A new method based on spectral subtraction for speech dereverberation". In: *Acta Acustica united with Acustica* 87.3 (2001), pp. 359–366.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation". In: *Proc. ICASSP*. 2008.

[3] M. Miyoshi and Y. Kaneda. "Inverse filtering of room acoustics". In: *IEEE Trans. Acoustics, Speech and Signal Processing* 36.2 (Feb. 1988), pp. 145–152.

[4] J. B. Allen, D. A. Berkley, and J. Blauert. "Multimicrophone signal-processing technique to remove room reverberation from speech signals". In: *J. Acoust. Soc. Am.* 62.4 (1977), pp. 912–915.

[5] R. Zelinski. "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms". In: *Proc. ICASSP*. 1988.

[6] I. A. McCowan and H. Bourlard. "Microphone array post-filter based on noise field coherence". In: *IEEE Transactions on Speech and Audio Processing* 11.6 (2003), pp. 709–716.

[7] A. Schwarz and W. Kellermann. "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2015). arXiv: 1502.03784.

[8] A. Schwarz and W. Kellermann. "Unbiased Coherent-to-Diffuse Ratio Estimation for Dereverberation". In: *Proc. IWAENC*. 2014.

[9] D. Povey et al. "The Kaldi speech recognition toolkit". In: *Proc. ASRU*. 2011.

[10] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann. "Spatial Diffuseness Features for DNN-Based Speech Recognition in Noisy and Reverberant Environments". In: *Proc. ICASSP*. 2015. arXiv: 1410.2479.

[11] F. Weninger, S. Watanabe, J. Le Roux, J. R. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll. "The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement". In: *Proc. REVERB Workshop*. 2014.

[12] S. P. Rath, D. Povey, K. Vesely, and J. Cernocky. "Improved feature processing for Deep Neural Networks". In: *Proc. Interspeech*. 2013.

[13] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, P. Woodland, and S. Young. *WSJCAM0 Cambridge Read News for REVERB LDC2013E109*. 2013.

[14] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur. "Improving deep neural network acoustic models using generalized maxout networks". In: *Proc. ICASSP*. 2014.

[15] K. Kinoshita, M. Delcroix, 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. WASPAA*. 2013.