Coherence-based Dereverberation for Automatic Speech Recognition

Andreas Schwarz, Andreas Brendel, Walter Kellermann

Lehrstuhl für Multimediakommunikation und Signalverarbeitung, Erlangen, Germany, E-Mail: {schwarz,brendel,wk}@lnt.de

Introduction

The idea of performing dereverberation using a short-time spatial coherence estimate dates back to 1977 [1], when it was proposed to essentially use the magnitude of the coherence as gain for reverberation suppression. Another heuristic method was recently proposed in [2], where a soft threshold function is used to compute a gain from the coherence magnitude, and the parameters of the threshold function are adapted depending on the histogram of the coherence magnitude in each frequency bin. Short-time coherence estimates have also been investigated in the context of beamforming as a so-called postfilter, and solutions for supression of uncorrelated and diffuse noise have been proposed [3]. In this contribution, we focus on methods where, first, the ratio between direct and reverberation signal components (coherent-to-diffuse ratio, CDR) is estimated from a short-time coherence estimate, and filter weights for reverberation suppression are computed from the CDR using, e.g., the Wiener filter or spectral subtraction rule. We compare and illustrate the behavior of a number of different CDR estimators that have been proposed over the past years, and propose a new variant. Finally, we compare the practical effect of the methods by processing reverberated speech and evaluating the recognition accuracy achieved by an automatic speech recognizer with the processed signals.

Signal model

We consider the recording of a reverberated speech signal by two omnidirectional microphones with spacing d. The auto- and cross-power spectral densities (PSD) of the microphone signals x_i are $\Phi_{x_i x_j}(k,f)$, $i,j = 1,2$, with the frame index k and frequency f. We assume that microphones are identical and closely spaced, therefore $\Phi_{x_i x_i} = \Phi_x$. The complex spatial coherence function is then given by

$$\Gamma_x(k,f) = \frac{\Phi_{x_1 x_2}(k,f)}{\Phi_x(k,f)}. \tag{1}$$

Furthermore, it is assumed that the direct path signal component with PSD Φ_s and the reverberation with PSD Φ_r are orthogonal, so that $\Phi_x = \Phi_d + \Phi_r$. We model the direct sound as a plane wave with the direction of arrival (DOA) θ with respect to the microphone axis, where $\theta = 0\,°$ corresponds to broadside direction, and the reverberation as a diffuse sound field [4]. The corresponding spatial coherence functions for the direct and reverberation components of the signals recorded at the microphones are

$$\Gamma_d(f) = e^{j2\pi f \Delta t}, \tag{2}$$

$$\Gamma_r(f) = \Gamma_{\text{diff}}(f) = \operatorname{sinc}(2\pi f \frac{d}{c}), \tag{3}$$

with the TDOA $\Delta t = \frac{d \sin(\theta)}{c}$. For hearing aids, where the head affects the coherence of the diffuse noise field, appropriate coherence models can be used instead [5].

Assuming that direct sound and reverberation are mutually orthogonal in the short-time spectral domain, the coherence of the mixture is given by

$$\Gamma_x(k,f) = \frac{\frac{\Phi_d(k,f)}{\Phi_r(k,f)}\Gamma_d(f) + \Gamma_r(f)}{\frac{\Phi_d(k,f)}{\Phi_r(k,f)} + 1}. \tag{4}$$

Solving for $\frac{\Phi_d}{\Phi_r}$ yields what we will denote as coherent to diffuse ratio (CDR) in the following:

$$CDR(k,f) = \frac{\Gamma_r(f) - \Gamma_x(k,f)}{\Gamma_x(k,f) - \Gamma_d(f)}. \tag{5}$$

Coherent-to-Diffuse Ratio Estimation

A short-time estimate $\hat{\Gamma}_x(k,f)$ for the spatial coherence can be obtained according to (1) from PSDs which have been estimated by, e.g., recursive averaging. However, (5) can in practice not be applied to obtain the CDR from an estimated coherence, since, due to mismatch between the coherence models and room acoustics, and the variance of the coherence estimate, the resulting CDR value will in general be complex-valued. A number of different practical estimator realizations have been proposed, which implicitly account for these errors, and which we will compare in the following. In order to illustrate their behavior, we visualize the output of different estimator realizations over the complex plane of possible coherence values $\hat{\Gamma}_x$ in Fig. 1. Results for a direct path TDOA $\Delta t = 0$ are shown in the first row, while in the second row, results are shown for $\Delta t = \frac{1}{3f}$. The black × marks the coherence of a fully coherent signal from the respective TDOA, while the white × marks the coherence of an ideal diffuse signal. The straight white line between these points marks the coherence values which would occur in theory under ideal conditions for different CDR values. In practice, the estimated coherence values $\hat{\Gamma}_x(k,f)$ will not lie exactly on this line. A good realization of an estimator should therefore be not only unbiased in the sense that the behaviour along the line matches (5) (which can be verified by inserting Γ_x according to (4) into the estimator equation), but also robust in the sense that small deviations of the coherence estimate from the model, e.g., a phase rotation caused by an inexact DOA estimate, do not lead to large deviations of the estimate.

The methods of Jeub et al. [6] and McCowan et al. [3] can be formulated as the CDR estimate

$$\widehat{CDR}_{\text{Jeub}}(k,f) = \frac{\Gamma_{\text{diff}}(f) - \operatorname{Re}\{e^{-j2\pi f \Delta t}\hat{\Gamma}_x(k,f)\}}{\operatorname{Re}\{e^{-j2\pi f \Delta t}\hat{\Gamma}_x(k,f)\} - 1}. \tag{6}$$

where the phase shift applied to the coherence estimate represents a time delay of one channel, which is proposed in the aforementioned publications in order to align the direct path component. However, this solution does not account for the fact that, for $\theta \neq 0\,°$, delaying one of the channels to achieve time alignment of the direct path also affects the phase of the coherence of the diffuse signal component. This leads to a biased estimate for low coherence values, as can bee seen in Fig. 1a (second row), where the CDR is underestimated for coherence values along the white line. This bias could be removed by correcting $\Gamma_{\text{diff}}(f)$ with the phase term $e^{-j2\pi f \Delta t}$.

Instead of time-aligning the signals, Thiergart et al. proposed to

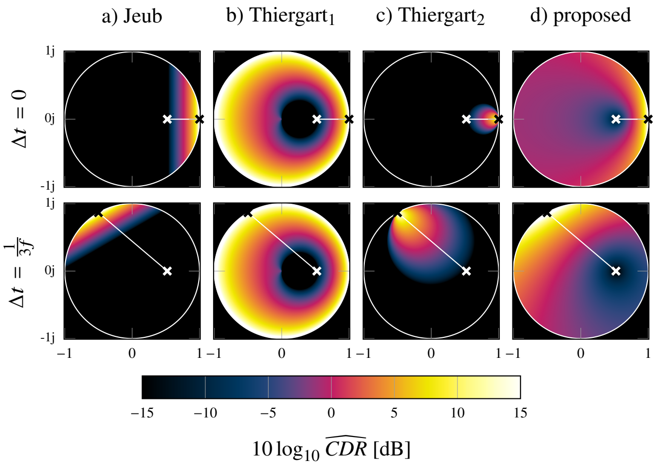

Figure 1: Coherent-to-diffuse ratio estimates as a function of complex spatial coherence $\hat{\Gamma}_x$, for $d = 8\,\text{cm}$, $f = 1.2\,\text{kHz}$. Different estimators (columns) and TDOA values (rows). Coherence of fully diffuse and fully coherent signals are highlighted.

use the phase of the cross-power spectral density (i.e., the phase of the estimated coherence $\arg\hat{\Gamma}_x$) as a phase estimate for the direct path model [7], which has the advantage of not requiring an explicit TDOA estimate:

$$\widehat{CDR}_{\text{Thiergart}}(k,f) = \text{Re}\left\{\frac{\Gamma_{\text{diff}}(f) - \hat{\Gamma}_x(k,f)}{\hat{\Gamma}_x(k,f) - e^{j\arg\hat{\Gamma}_x(k,f)}}\right\}. \quad (7)$$

However, the instantaneous phase of the coherence is not an exact estimate for the phase of the direct path, since, for small values of the CDR, the coherence estimate is mainly dominated by the coherence of the diffuse signal. For $\theta \neq 0\,°$, this causes a bias of the direct path phase estimate, which in turn leads to a bias in the CDR estimate (Fig. 1b, second row). If not the instantaneous phase, but a constant phase estimate is used for the coherence model of the direct path (e.g., from a TDOA estimate), the estimator is unbiased, but not robust towards phase deviations between the model and the coherence estimate, as can be seen in Fig. 1c: for values of $\hat{\Gamma}_x$ close to the unit circle, the CDR estimate drops sharply to zero when the phase of $\hat{\Gamma}_x$ exceeds the phase of Γ_d.

We propose another estimator that only has a slight bias for high CDR values, and is very robust towards model mismatch and estimation errors, which is illustrated in Fig. 1d:

$$\widehat{CDR}_{\text{prop.}}(k,f) = \left|\frac{e^{-j2\pi f\Delta t}\Gamma_{\text{diff}}(f) - e^{-j2\pi f\Delta t}\Gamma_x(k,f)}{\text{Re}\{e^{-j2\pi f\Delta t}\Gamma_x(k,f) - 1\}}\right|. \quad (8)$$

ASR Evaluation

We use 500 utterances from the GRID corpus which are reverberated by convolution with measured 2-channel impulse responses. The impulse responses were measured in two rooms: Room A, a lecture hall with a reverberation time $T_{60} \approx 1\,\text{s}$, and Room B, a large foyer with a reverberation time $T_{60} \approx 3.5\,\text{s}$. In each room, impulse responses were measured for 40 different source positions in $1\ldots4\,\text{m}$ distance from the microphones, spread over an angular range of $-90\ldots90\,°$. All processing is done at a sampling rate of 16 kHz using a filterbank with window length 1024, FFT size 512, and downsampling factor 128. The exponential forgetting factor for the PSD estimation is 0.68. For all methods, the reverberation suppression is applied to one microphone signal. In addition to the CDR-based methods, we evaluate a modified version of the method by Allen et al. [1] (where we directly use the magnitude of the coherence as the gain, and apply the enhancement to only one microphone instead of aligning), and the coherence-to-gain-mapping proposed by Westermann et al. [2] (parameter $k_p = 0.1$). We also evaluate the single-channel enhancement method based on an exponential decay model by Lebart et al. [8] (assuming perfect knowledge of the reverberation time). For the method of Lebart et al. and the CDR-based methods, spectral magnitude subtraction is applied to one microphone, with an oversubtraction factor that is optimized separately for each investigated method for maximum recognition performance. The lower limit for the filter gain is 0.1 in all cases. TDOA estimates obtained by cross-correlation are used for the CDR estimators which require the TDOA Δt.

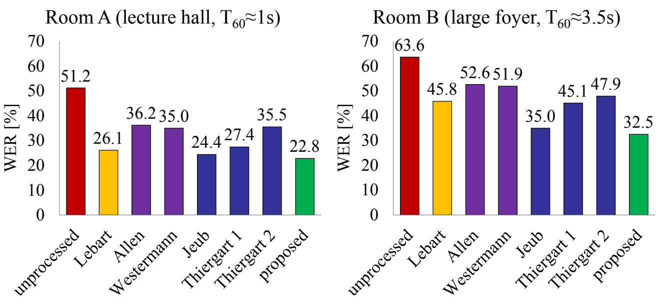

Figure 2: ASR Word Error Rate for reverberated unprocessed signals, and after processing with various dereverberation methods.

For ASR we use PocketSphinx with standard MFCC+Δ+ΔΔ features, trained on clean unreverberated speech. We evaluate the Word Error Rate (WER) for the letter and number in the GRID sentence; the WER for clean speech is 7.9%. Fig. 2 shows the WER for the unprocessed reverberated and dereverberated signals, with results averaged over all source positions. Since the methods by Allen et al. and Westermann et al. have a relatively weak dereverberating effect, and, unlike the methods based on spectral subtraction, offer no direct way of increasing the amount of suppression, WER improvements are low. Results for the CDR-based methods show that the implementation of the CDR estimator is crucial, and that the single-channel dereverberation performance can be significantly exceeded by the coherence-based dereverberation method using the proposed estimator.

References

[1] J. B. Allen, D. A. Berkley, and J. Blauert. “Multimicrophone signal-processing technique to remove room reverberation from speech signals”. In: *The Journal of the Acoustical Society of America* 62.4 (1977), pp. 912–915.

[2] A. Westermann, J. M. Buchholz, and T. Dau. “Binaural dereverberation based on interaural coherence histograms”. In: *The Journal of the Acoustical Society of America* 133.5 (2013), pp. 2767–2777.

[3] 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.

[4] R. K. Cook et al. “Measurement of Correlation Coefficients in Reverberant Sound Fields”. In: *The Journal of the Acoustical Society of America* 27.6 (1955), pp. 1072–1077.

[5] M. Jeub and P. Vary. “Binaural dereverberation based on a dual-channel Wiener filter with optimized noise field coherence”. In: *Proc. ICASSP*. 2010.

[6] M. Jeub et al. “Model-Based Dereverberation Preserving Binaural Cues”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010), pp. 1732–1745.

[7] O. Thiergart, G. Del Galdo, and E. A. P. Habets. “Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones”. In: *Proc. ICASSP*. 2012.

[8] 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.