An Evaluation of Binaural Noise and Reverberation Reduction Algorithms

Masoumeh Azarpour1, Dominic Schmid1, Rainer Martin1, Timo Gerkmann2

1 Ruhr-Universität Bochum, Germany, Email: {masouneh.azarpour, dominic.schmid, rainer.martin}@rub.de
2 Universität Oldenburg, Germany, Email: timo.gerkmann@uni-oldenburg.de

Abstract
In this paper we evaluate several algorithms for binaural noise reduction and dereverberation and compare these algorithms in terms of their effectiveness in enhancing the quality and intelligibility of the speech signal. The algorithms that we consider exploit the coherence of the two microphone signals but also employ several enhancements to this basic and well-known scheme. These enhancements include smoothing techniques in the cepstral domain as well as the estimation and explicit use of the interaural impulse response in a matched-filter array approach. In the evaluation part of this paper we assess the noise reduction, the speech distortion, the strength of the dereverberation effect, the predicted intelligibility and the predicted overall quality of the processed signal. The comparison will highlight properties of the considered methods and performance trade-offs of these methods in general.

Introduction
In many speech communication scenarios the quality and the intelligibility of the acoustic signal is degraded by room reverberation and background noise. It is therefore desirable to mitigate these effects by means of a single or multi-channel speech enhancement technique. A particularly interesting class of methods, which is also highly relevant for hearing aids applications, uses the acoustic signals of binaural microphones. Binaural speech enhancement algorithms are more advantageous over bilateral or single-channel algorithms due to the use of the joint statistics of left and right microphone signals as well as due to the potential to preserve or improve the spatial perception of sounds [1, 2]. Towards this end, several binaural algorithms have been proposed and in most algorithms, especially those based on spectral coherence, the noise reduction and dereverberation effects are intrinsically coupled. However, noise reduction and dereverberation are difficult tasks on their own, therefore, many published works investigated either noise reduction or dereverberation.

The earliest and most prominent approach to binaural dereverberation [3] uses a spectral analysis-synthesis system and a measure of inter-microphone-correlation in frequency bands (such as the magnitude-squared coherence function) to suppress incoherent reverberation. This method has been also used for noise reduction, e.g., for hands-free speech communication in cars [4]. Single channel spectral subtraction, dual-channel noise power estimation, and the cross-correlation-based dereverberation methods have been successfully combined in [5]. Recently, variations of the cross-correlation method have been proposed which include a smoothing step in the cepstrum domain [6]. In other works such as in [7], single-channel dereverberation algorithms have been extended into a binaural framework and the performance of different bilateral filter gain computation methods has been investigated. The authors in [8] proposed a two-stage dereverberation algorithm to remove the early and late reverberation. The first stage is designed based on a statistical model of the room impulse responses and employs a spectral subtraction concept. The dual-channel Wiener filter in the second stage was utilized to reduced residual reverberation. In addition to achieving a significant dereverberation effect, the proposed algorithm is also able to keep the binaural cues unaffected.

In this paper we investigate the noise reduction and dereverberation performance of selected algorithms and address the arising difficulties. Two state-of-the-art algorithms which exploit the coherence of the two microphone signals along with several additional enhancements to the basic scheme are considered. These enhancements include smoothing techniques in the cepstral domain. Furthermore, a binaural noise reduction algorithm based on a matched filter array (MFA) using an estimated interaural impulse response [9] is included. The performance of the aforementioned algorithms is evaluated under different noise and reverberation scenarios in terms of noise reduction, speech distortion, the strength of the dereverberation effect, the predicted intelligibility and the predicted overall quality of the processed signal.

Binaural Signal Model
We consider simulated acoustic environments where the target speech $s(k)$ is convolved with acoustic room impulse responses $g_l$ and $g_r$ as captured by microphones positioned at the left and the right ear. The microphone signals are given in (1) where the subscripts $i \in \{l, r\}$ denote the signals corresponding to the left and right microphone, respectively,

$$y_i(k) = s(k) \ast g_i + v_i(k)$$

$$= s_{i,dr}(k) + s_{i,rv}(k) + v_i(k), \quad i \in \{l, r\},$$

where $\ast$ represents the linear convolution and $v_i(k)$ is the ambient noise. The direct path and reverberant speech components are represented by $s_{i,dr}(k)$ and $s_{i,rv}(k)$, respectively. The main objective of this paper is to recover the direct-path signal $s_{i,dr}(k)$.
Overview of the Algorithms

In this section the algorithms which are considered in the evaluation framework are briefly described. The required tuning parameters of the algorithms are chosen as they were suggested by their authors. In order to make the comparison consistent, the smoothing factor for estimating (cross-) power spectral densities which are used in almost all algorithms is set to the same fixed value.

Coherence-based dereverberation algorithm ABB [3]: The uncorrelated reverberation received at the two microphones is reduced based on the magnitude-squared coherence between microphones. Short-time Fourier spectra of input signals are then phase compensated and averaged in order to recover the target speech.

Dual-channel spectral subtraction-based noise reduction and dereverberation algorithm DBR [5]: This algorithm is a two-stage speech enhancement structure. In the first stage, the spectral amplitude of the microphone signals are estimated using a dual-channel noise power spectral density estimator. In the second stage an adaptive post-filter similar to the ABB approach [3] is employed to reduce residual noise and reverberation.

Binaural noise reduction using a binaural matched-filter array MFA [9]: In this work, MFA filters are employed as an approximation of the MVDR beamformer using the estimated interaural impulse response between left and right microphone signals. The estimation accuracy of the interaural impulse response affects the performance of the binaural noise reduction algorithm.

Most of the spectral noise reduction and dereverberation algorithms, which need an estimate of signal power spectra, suffer from fluctuations in the spectral gain functions. This artifact is called musical noise and is known to reduce the quality of the processed signal. In this paper, two techniques for suppressing musical noise are considered to remove such effects in conjunction with the aforementioned algorithms.

ABB with temporal cepstrum smoothing of spectral filter gains ABB-Cep: In [10] a technique has been proposed to smooth the cepstral representation of the filter gain function in order to reduce the musical noise. Due to the selective temporal smoothing of spectral functions (i.e. low-degree of smoothing for speech related cepstral coefficients; larger degree of smoothing for the remaining cepstral coefficients) the temporal characteristics of speech are not affected. The smoothing of the higher cepstral coefficients reduces outliers in the filter gain function.

ABB with cepstrum weighting ABB-CW: Since a temporal smoothing of the gain function as proposed in [10] may lead to a temporal smearing which decreases the dereverberation performance, an instantaneous cepstral weighting has been proposed [6]. The a posteriori probability that a cepstral coefficient represents the speech spectral structure was employed for the cepstral weighting function. Moreover, a priori knowledge about the speech was incorporated which is obtained by offline training of the parameters of the a posteriori probability. In contrast to [10], this technique does not need any voiced/unvoiced detector or a fundamental period estimator.

Objective Evaluation Measures

Many measures have been proposed to objectively evaluate the speech quality and intelligibility [11, 12]. For the most part these measures have been employed for the assessment of noise reduction algorithms. Recently, some works also explored these evaluation measures in the context of dereverberation algorithms [13]. However, as the evaluation of the dereverberation effects appears to be a difficult issue, no broad consensus on the most suitable measures has been reached so far. In this paper, we will use the following objective measures to assess various aspects of the considered algorithms: Perceptual Evaluation of Speech Quality (PESQ) [14] for the assessment of the overall signal quality, the Noise Reduction Ratio (NRR) for the noise reduction performance, the Cepstral Distance measure (CD) [15] for speech distortions and dereverberation, and the Short-Time Objective Intelligibility measure (STOI) [16]. In the following the NRR and CD measures will be briefly described. Note that both measures are computed on the noise and the speech signals, respectively. We employ a shadow-filter approach where the filter coefficients are adapted on the noisy signal and are copied into auxiliary filters for filtering the noise and speech signals individually.

The NRR is defined as the ratio of the power of the input noise signal to the power of the shadow-filtered noise signal in dB. A frequently used version of the CD [11], is expressed as

\[ d_{\text{CD}} = \frac{1}{M} \sum_{m=1}^{M} \min \left( \frac{10\sqrt{2}}{\ln(10)} \sum_{k=1}^{P} [d_m(k) - c_m(k)]^2, 10 \right) \]

where \( d_m \) and \( c_m \) are the cepstral coefficient vectors of the \( m \)th frame of reference and shadow-filtered reverberant speech signal, respectively. The cepstral coefficients are computed via the LPC spectrum and limited to a maximum value of 10 in each frame in order to remove outliers. For the average CD only the smallest 95% of the frame CD values are considered.

We note that the cepstral distance as defined above does not allow a meaningful comparison of short-time spectral envelopes in the reverberation tails. Furthermore, for strong reverberant conditions we observe a lot of large CD values on the frame level which are not taken into account by the standard measure. Also, we observe many outliers at the offsets of high-energy speech sounds. Therefore, instead of removing the outliers via the above limiting procedure we rather propose to combine the CD with a VAD and consider only CD values obtained during speech activity.
Experimental Setup and Results

All speech signals are taken from the TIMIT database [17] at a sampling frequency of 16 kHz. For the simulation of reverberant conditions we convolve the signals with binaural room impulse responses from the Aachen database [18], downsampled to 16 kHz. We select impulse responses from the meeting room ($T_{60} = 210$ ms), and the lecture room ($T_{60} = 700$ ms). The distance from the loud-speaker to the dummy head is about $d = 2$ m. We consider two types of additive observation noise signals: computer-generated diffuse, stationary white Gaussian noise and diffuse babble noise both of which have been generated using the algorithm proposed in [19]. In all algorithms we use a DFT length of 512 and the squared-root periodic Hanning window of length $M = 512$ for analysis and synthesis. The frame advance is set to half of frame length $M$. To investigate the performance of cepstral smoothing methods, the smoothing factor for the estimation of the (cross-) power spectral densities for post-filtering stages is set to $\alpha = 0.6$ to preserve transient speech components while tolerating some spectral outliers which are of less concern at this stage as they can be removed by cepstral smoothing techniques. Other required tuning parameters of algorithms are set as they were recommended by their authors. All algorithms produce binaural output signals. Since the target speaker was positioned in the frontal direction all evaluation measures are averaged over the left and right ears.

In Fig. 1 comparisons are shown with the aforementioned algorithms in terms of PESQ, NRR, CD, and STOI for different SNR and reverberant conditions. The NRR and the CD measures are calculated using the shadow-filtered noise and shadow-filtered reverberant speech. To compute a reference signal for all measures we convolve the first 20 ms of the impulse responses with the anechoic clean speech signal.

As it is shown in Fig. 1(a) the PESQ measure predicts a clear improvement for all algorithms. However, the MFA does not improve PESQ in low SNR conditions because the accurate estimation of the interaural impulse response degrades in low SNR. Furthermore, we observe in Fig. 1(b) that the modified ABB algorithms lead to a higher NRR than the original ABB and all of them are outperformed by the DBR algorithm. The high NRR of the DBR is explained by its two-stage design. Owing to its high NRR, the two-stage DBR algorithm also achieves a better overall speech quality than the other algorithms in low SNR conditions. The MFA results in the least amount of noise reduction and for this reason it achieves the lowest PESQ results.

Fig. 1(c) shows the CD-VAD results between the reference speech signal and the shadow-filtered reverberant speech signal. The speech distortions observed in this setup are due to (residual) reverberation and the distortions affected by the binaural processing schemes. For low reverberant condition the CD is clearly dominated by the distortions affected by the noise reduction filtering.

It can be seen that the large noise reduction introduced by the two-stage DBR comes with a large increase in spectral distortions as measured by CD. In contrast, the cepstral modifications ABB-Cep and ABB-CW increase the NRR and at the same time yield similar or even lower CD as compared to ABB in most conditions. With respect to the trade-off between NRR and CD, the cepstral weighting approach [6] outperforms both ABB as well as ABB-Cep except for high SNR conditions. It should be noted that the MFA algorithm, due to its less aggressive noise reduction, introduces the least amount of speech distortion. Without showing the corresponding results in a graph we furthermore note that all algorithms significantly improve the CD-VAD between the reference signal and the processed noisy and reverberant signal. All these observations have been confirmed through informal listening tests.

The STOI results shown in Fig. 1(d) indicate an improvement for both low and high reverberant conditions. However, in the high SNR (i.e., 20 dB) and low reverberation condition, DBR reduces the predicted speech intelligibility due to introduced speech distortion.

Conclusions

In this paper we compared several binaural noise and reverberation reduction algorithms in terms of their overall quality, the noise reduction performance, the speech distortions, and the predicted intelligibility. Four of them use the magnitude-squared coherence or variations of it to compute the spectral gain function. Furthermore, an algorithm based on a binaural matched-filter array (MFA) was considered.

The MFA achieves less speech distortion at the price of less noise reduction. The DBR approach significantly reduces the noise power and consequently improves the predicted overall speech quality. It suffers, however, from larger speech distortions. Among the two different cepstrum smoothing techniques the ABB-CW appears to outperform the ABB-Cep in all measures. In general the ABB-CW seems to give the best tradeoff between dereverberation, noise reduction, and speech distortion.

All in all the evaluation of the dereverberation performance remains a difficult task since the cepstral distance measure as used in this work indicates no improvements. Although the processed signals appear to be significantly less reverberant the distortions affected by the spectral gain outweighs the dereverberation effect.

References

Figure 1: Comparison of algorithms in terms of (a) PESQ, (b) NRR, (c) CD, and (d) STOI in different reverberant and SNR conditions.


