Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues

May, Tobias

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

Link to article, DOI:
10.1109/TASLP.2017.2765819

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues

Tobias May

***Abstract*—This study presents an algorithm for binaural speech dereverberation based on the supervised learning of short-term binaural cues. The proposed system combined a delay-and-sum beamformer (DSB) with a neural network-based post-filter that attenuated reverberant components in individual time-frequency (T-F) units. A multi-conditional training (MCT) procedure was used to simulate the uncertainties of short-term binaural cues in response to room reverberation by mixing the direct part of head related impulse responses (HRIRs) with diffuse noise. Despite being trained with only anechoic HRIRs, the proposed dereverberation algorithm was tested in a variety of reverberant environments and achieved considerable improvements relative to a coherence-based approach in terms of three objective metrics reflecting speech quality and speech intelligibility. Moreover, a systematic evaluation showed that the proposed system generalized very well to a wide range of acoustic conditions, including various measured binaural room impulse responses (BRIRs) reflecting different reverberation times, azimuth positions spanning the entire frontal hemifield, various source-receiver distances as well as different artificial heads.**

***Index Terms*—Dereverberation, binaural, coherence, neural networks, short-term direct-to-reverberant energy ratio, ideal ratio mask**

I. Introduction

In everyday listening situations, the sound mixture arriving at the listener's ears is a superposition of direct sound stemming from the target as well as early and late reflections from the walls or obstacles in the room, which are delayed and attenuated versions of the direct sound. While early reflections can be advantageous for speech intelligibility in rooms [1], [2], late reflections cause temporal and spectral smearing of the target signal characteristics, which reduces speech intelligibility for normal-hearing and hearing-impaired listeners [3]–[6]. In addition, the presence of reverberation deteriorates the performance of many technical applications, such as automatic speech recognition (ASR) [7] and speaker identification (SID) systems [8]. Thus, dereverberation algorithms can be assumed to be beneficial for a wide range of speech processing applications.

Microphone array processing techniques are fundamental building blocks in many speech processing applications and can enhance the target source by spatial filtering [9]. The most simple yet robust approach is the delay-and-sum beamformer (DSB), which can be steered towards a particular target source (look direction) by compensating for the time delay between the microphone signals. In this way, coherent signal components from the target source are constructively added, while diffuse signal components are attenuated. One limitation of such static beamformers is that the amount of attenuation of interfering sources from undesired directions is quite limited for a low number of microphones. In addition, beamformers can only partially suppress reverberation because reflections coming from the look direction are not attenuated.

Late reverberation can be modeled as an additive signal degradation [10] and can thus be removed by spectral enhancement strategies, where a real-valued gain function is applied to the short-time discrete Fourier transform (STFT) representation of the reverberant speech signal. In the context of single-channel dereverberation, this gain function is typically based on an estimate of the late reverberant power spectral density (PSD) [10], [11]. More sophisticated approaches aim at reducing both interfering noise and reverberation by either jointly estimating the noise and the late reverberant PSD [12] or by combining an autoregressive moving-average (ARMA) model of the late reverberant PSD with a hidden Markov model (HMM) of clean speech in a Bayesian filtering framework [13].

One frequently-used gain function for binaural dereverberation is based on the short-term interaural coherence (IC) function which measures the similarity between two ear signals. In this way, time-frequency (T-F) units dominated by the direct-sound can be distinguished from reverberation, for which the IC is typically lower. A linear coherence-to-gain mapping function was proposed in [14], whereas [15] presented a non-linear sigmoidal mapping based on coherence histograms which allowed for a much stronger attenuation of reverberant components. For a comprehensive overview of recent developments on speech dereverberation, the reader is referred to [16]. Apart from applications in the context of binaural dereverberation, such gain functions based on the short-term IC are also often employed as post-filters to increase the effectiveness of beamformers [17].

Instead of using a heuristic mapping as in [14] and [15], the advancement in the field of machine learning allows the use of supervised learning approaches, e.g. neural networks, to establish a mapping between a set of features and an explicit measure of direct-sound activity, as reflected by the short-term direct-to-reverberant energy ratio (DRR). In addition, instead of using one particular feature (e.g. the short-term IC), a

T. May is with the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK - 2800 Kgs. Lyngby, Denmark, e-mail: tobmay@elektro.dtu.dk.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

neural network can exploit the combination of complementary binaural features, such as IC, interaural phase differences (IPDs) and interaural level differences (ILDs). However, one of the main challenges of supervised learning approaches is the generalization to acoustic conditions not seen during training. The task of estimating a post-filter with a neural network is particularly challenging, because the network should ideally be independent of the room acoustic condition, the sound source direction and the underlying artificial head.

Recently, deep neural networks (DNNs) have been employed to segregate a target signal from a binaural mixture by combining spectral (monaural) and binaural features [18], [19]. Both studies used a set of reverberant binaural room impulse responses (BRIRs) for training. However, both studies assumed a fixed target direction of $0\,^{\circ}$, which is a significant limitation. Moreover, binaural features reflect relative differences between the ears and thus change systematically as a function of the sound source azimuth. In contrast, the influence of the head shadow on monaural features is changing less systematically with sound source azimuth and alters the absolute magnitude of monaural features. As a consequence, monaural features are less likely to generalize as well as binaural features when being used in a supervised learning framework, which may require the assumption of a fixed source position.

In the context of binaural sound source localization, a multi-conditional training (MCT) procedure (with Gaussian mixture models (GMMs) or DNNs) was shown to produce remarkably robust localization models with strong generalization capabilities [20]–[23]. The MCT was based on anechoic head related impulse responses (HRIRs) and the influence of room reverberation and competing sources was simulated by diffuse noise coming from all azimuth directions [21], [22]. The resulting localization model generalized very well to multiple competing sound sources, different artificial heads and a wide range of reverberant conditions. However, the applicability of such a training procedure for the task of speech dereverberation and its ability to generalize to arbitrary source directions has not yet been tested.

The current study presents a novel neural network-based post-filter for speech dereverberation. The network was trained with three types of short-term binaural features (IC, ILDs and IPDs), using a MCT strategy where binaural reverberant signals were simulated by mixing the direct part of HRIRs with diffuse noise. In addition, a time-alignment stage was incorporated to increase the effectiveness of the IPD feature. The performance of the proposed approach was compared to the coherence-based dereverberation algorithm developed by [15] and the single-channel Bayesian filtering approach presented in [13] using three objective metrics, namely the perceptual evaluation of speech quality (PESQ) [24], the short-time objective intelligibility (STOI) metric [25] and the normalized speech-to-reverberation modulation energy ratio (SRMR) [26]. Moreover, the generalization abilities of the neural network-based post-filter was evaluated in isolation and in combination with a DSB using a wide range of acoustic conditions, including various BRIRs representing different degrees of reverberation, several source-receiver distances, as well as different azimuth positions and artificial heads.

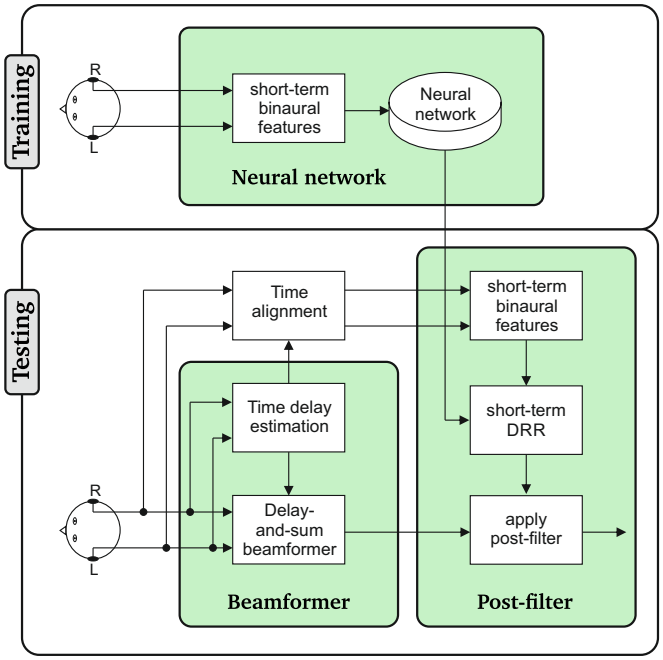


Fig. 1. Block diagram of the proposed dereverberation algorithm consisting of a fixed DSB beamformer and a neural network-based post-filter. The neural network was trained with short-term binaural features during an initial training stage.

II. System

The proposed dereverberation algorithm shown in Fig. 1 combined a fixed DSB with a neural network-based post-filter that estimated the short-term DRR in individual T-F units. The neural network was trained with short-term binaural features during an initial training stage. In the testing stage, the DSB was steered towards the estimated direction of the target source by performing broadband time-delay estimation (TDE). The estimated time delay was also used to time-align the left and right ear signals prior to feature extraction. The benefit of such a time alignment stage is discussed in Sect. IV-A. Afterwards, the monaural beamformer output was further processed by the network-based post-filter which attenuated reverberant components. Both building blocks, the DSB and the network-based post-filter, are described below.

A. Delay-and-sum beamformer (DSB)

The binaural signal was pre-processed by a static DSB that was steered towards the estimated direction of the target source (e.g. the most dominant source). The required time delay between both ear signals was estimated using the generalized cross-correlation (GCC) function [27], which was computed for the complete left- and right ear signals for time lags within the range of $[-1, 1]\,\mathrm{ms}$. To increase the resolution of the estimated time delay, the binaural signal was upsampled to a sampling frequency of $48\,\mathrm{kHz}$ prior to TDE. Afterwards, the binaural signal was time-aligned by delaying the leading ear signal and averaged. Finally, this monaural beamformer output was downsampled to $16\,\mathrm{kHz}$ and further processed by the network-based dereverberation stage.

B. Binaural features

Three different types of short-term binaural features were used to train the neural network-based dereverberation algorithm, namely the IC, ILDs and IPDs. All features were derived from the STFT representation of the binaural signals. The STFTs were computed by segmenting the left and right ear signals into overlapping frames of $32\,\mathrm{ms}$ duration with a time shift $\mathcal{S}_{\mathrm{STFT}}$ corresponding to $8\,\mathrm{ms}$. Each frame was Hamming-windowed and a 512-point discrete Fourier transform (DFT) was computed, producing the STFT representations of the left and the right ear signals, $X_{\mathrm{L}}(k,\ell)$ and $X_{\mathrm{R}}(k,\ell)$, where k and ℓ indicate frequency bin and time frame, respectively.

The spectral resolution of the binaural features was inspired by the human auditory system. Specifically, a set of 64 auditory filters was used that covered a frequency range between $65\,\mathrm{Hz}$ and $8\,\mathrm{kHz}$ according to the mel-frequency spacing [28], where $G(c,k)$ reflected the frequency-dependent response of auditory channel c.

1) Interaural coherence (IC): The IC in auditory channel c was calculated by

$$IC(c,\ell) = \sqrt{G(c,k)\left(\frac{|\Phi_{\mathrm{LR}}(k,\ell)|}{\sqrt{\Phi_{\mathrm{LL}}(k,\ell)\Phi_{\mathrm{RR}}(k,\ell)}}\right)^2}, \quad (1)$$

with $\Phi_{\mathrm{LL}}(k,\ell)$, $\Phi_{\mathrm{RR}}(k,\ell)$ and $\Phi_{\mathrm{LR}}(k,\ell)$ representing the exponentially weighted short-term auto- and cross-power spectral density functions [14]

$$\Phi_{\mathrm{LL}}(k,\ell) = \alpha\Phi_{\mathrm{LL}}(k,\ell-1) + (1-\alpha)X_{\mathrm{L}}(k,\ell)X_{\mathrm{L}}^*(k,\ell), \quad (2)$$

$$\Phi_{\mathrm{RR}}(k,\ell) = \alpha\Phi_{\mathrm{RR}}(k,\ell-1) + (1-\alpha)X_{\mathrm{R}}(k,\ell)X_{\mathrm{R}}^*(k,\ell), \quad (3)$$

$$\Phi_{\mathrm{LR}}(k,\ell) = \alpha\Phi_{\mathrm{LR}}(k,\ell-1) + (1-\alpha)X_{\mathrm{L}}(k,\ell)X_{\mathrm{R}}^*(k,\ell), \quad (4)$$

where the exponential weight α was controlled by the time constant τ according to $\alpha = \exp(-\mathcal{S}_{\mathrm{STFT}}/\tau)$ and * denotes the complex conjugate. The time constant was set to $\tau = 10\,\mathrm{ms}$.

2) Interaural level difference (ILD): The ILD was determined by the amplitude ratio between the STFTs of the two ear signals

$$ILD(c,\ell) = G(c,k)\left(20\log_{10}\left(\left|\frac{X_{\mathrm{R}}(k,\ell)}{X_{\mathrm{L}}(k,\ell)}\right|\right)\right). \quad (5)$$

3) Interaural phase difference (IPD): The IPD was derived by extracting the phase of the ratio of the left and right ear STFTs

$$IPD(c,\ell) = G(c,k)\arg\left(\frac{X_{\mathrm{R}}(k,\ell)}{X_{\mathrm{L}}(k,\ell)}\right). \quad (6)$$

C. Short-term DRR and IRM

The objective of the neural network was to estimate the short-term DRR based on a set of binaural features. Assuming that $d(t)$ and $r(t)$ represent binaural signals that were obtained by convolving a speech signal with the direct and reverberant part of a BRIR, separately, the corresponding STFT representations of the direct-sound and the reverberant signal components, $D(c,\ell)$ and $R(c,\ell)$, in the auditory domain were derived by

$$D(c,\ell) = \sum_k G(c,k)|X_{\mathrm{M}}^{\mathrm{D}}(k,\ell)|^2, \quad (7)$$

$$R(c,\ell) = \sum_k G(c,k)|X_{\mathrm{M}}^{\mathrm{R}}(k,\ell)|^2. \quad (8)$$

The binaural signals $d(t)$ and $r(t)$ were averaged across ears prior to computing the STFT, resulting in monaural spectrogram representations of the direct-sound and the reverberant signal components, $X_{\mathrm{M}}^{\mathrm{D}}$ and $X_{\mathrm{M}}^{\mathrm{R}}$, respectively. The required direct and reverberant components of the BRIR were identified by a time-windowing procedure outlined in [29]. Specifically, the first $1\,\mathrm{ms}$ after the maximum peak in the BRIR represented the direct-path component, whereas the remaining part of the BRIR was associated with reverberation.

Subsequently, the short-term DRR was computed by relating the short-term energy of the direct-sound signal components to the reverberant components

$$DRR(c,\ell) = 10\log_{10}\left(\frac{D(c,\ell)}{R(c,\ell)}\right). \quad (9)$$

Finally, the short-term DRR was expressed in terms of the ideal ratio mask (IRM), which is frequently used as a training target in supervised learning approaches [30]

$$IRM(c,\ell) = \left(\frac{DRR(c,\ell)}{DRR(c,\ell)+1}\right)^{\beta}, \quad (10)$$

$$= \left(\frac{D(c,\ell)}{D(c,\ell)+R(c,\ell)}\right)^{\beta}, \quad (11)$$

where the exponent β was set to 0.5.

D. Neural network architecture

A feedforward neural network was used to learn the mapping from the binaural features to IRM. The network consisted of an input layer, one hidden layer with rectified linear unit (ReLU) activation and 64 sigmoid output units representing the IRM in different auditory filters. The network was trained in full batch mode with the resilient back-propagation algorithm. To improve the generalization of the network, the mean squared error (MSE) performance function was modified with a weight decay regularization of 0.5 to reduce the risk of overfitting. Temporal context was incorporated by stacking the binaural feature vector across a predefined number of preceding time frames. In contrast to many other studies [19], [30], [31], future time frames were deliberately not considered here to limit the time delay of the algorithm and to ensure its applicability in real-time applications. To improve the generalization performance, an ensemble of 5 networks was separately trained and their output was averaged to predict the IRM.

III. Evaluation

A. Databases

Speech material from the TIMIT database [32] was used for the training and the testing stage. The training set contained

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
ANECHOIC HRIRS USED FOR TRAINING.

Database	Azimuth (°)		Distance (m)
	Range	Steps	
Berlin [36]	±90	5	3
Oldenburg [37]	±90	5	3
CATT [33]	±90	5	1.5

TABLE II
TWO SETS OF REVERBERANT BRIRS USED FOR EVALUATION.

Database	Azimuth (°)		Room	T_{60} (s)	Distance (m)
	Range	Steps			
Surrey [34]	±90	5	A	0.32	1.5
			B	0.47	1.5
			C	0.68	1.5
			D	0.89	1.5
Aachen [35]	±90	45	Stairway	1.1	2
			Aula	3.3	3

10 sentences from 630 speakers (438 males and 192 females), forming a total set of 6300 training sentences. The test set consisted of 8 sentences from 168 speakers (112 males and 56 females), forming a total set of 1344 testing sentences, from which none appeared in the training set.

Binaural signals were created by convolving monaural speech signals with HRIRs for anechoic conditions or BRIRs for reverberant conditions. As listed in Tab. I, a set of three anechoic HRIR databases (Berlin, Oldenburg and CATT) was used during the training stage. Whereas the Berlin database was recorded with a Knowles Electronic Manikin for Acoustic Research (KEMAR) of type 45BA, the Oldenburg database used a Brüel & Kjær head and torso simulator (HATS) of type 4128C. No details about the receiver used for the CATT-acoustics HRIRs were available [33]. The testing was based on two sets of reverberant BRIRs, namely the Surrey [34] and the Aachen database [35], which are summarized in Tab. II.

B. Model training

The neural network was trained with a set of 2000 binaural mixtures. To simulate the uncertainties of binaural cues in response to reverberant acoustic conditions, an MCT procedure according to [22] was employed. Specifically, each binaural mixture consisted of direct speech components mixed with diffuse noise at a specific signal-to-noise ratio (SNR). Direct speech components were created by convolving a randomly selected TIMIT sentence from the training set with the direct part (see Sect. II-C) of an HRIR. To ensure that the network generalized to different artificial heads and a range of source directions, three different HRIR databases were used during training that covered the full frontal azimuth range from -90 to $90\,^\circ$, as summarized in Tab. I. For each of the 2000 binaural mixtures, a randomly selected HRIR (one of three databases and one of 37 azimuth directions) was used. The presence of room reverberation was simulated by diffuse noise, which consisted of a mixture of 37 uncorrelated, white Gaussian noise sources that were placed across the frontal hemifield ranging from $-90\,^\circ$ to $90\,^\circ$ in steps of $5\,^\circ$ (using the same HRIR database as for the direct speech components). The long-term average spectrum (LTAS) of the diffuse noise was equalized to match the LTAS of the TIMIT speech corpus. The SNR between the direct speech components and the diffuse noise was randomly selected from a range between $0\,\mathrm{dB}$ and $15\,\mathrm{dB}$. During training, mean and variance normalization was performed using the entire feature space, while features used for testing were scaled using those normalization statistics measured during training.

C. Evaluation

To test the generalization abilities of the proposed approach to acoustic conditions not seen during training, two sets of measured BRIRs were considered as summarized in Tab. II. The first set was based on BRIRs from the Surrey database [34], which were measured with a Cortex Manikin Mk2 HATS at a distance of $1.5\,\mathrm{m}$ and ranged from -90 to $90\,^\circ$ azimuth in steps of $5\,^\circ$. For each of the four rooms (room A, B, C and D) and 37 azimuth angles, 5 TIMIT sentences were randomly selected from the test set, resulting in a set of $4 \times 37 \times 5 = 740$ reverberant binaural mixtures. The second set utilized BRIRs from the Aachen database [35], which were measured with a HMS2 artificial head by HEAD acoustics at distances of $2\,\mathrm{m}$ and $3\,\mathrm{m}$ and ranged from -90 to $90\,^\circ$ azimuth in steps of $45\,^\circ$. For each of the two rooms (stairway and aula carolina) and 5 azimuth angles, 5 TIMIT sentences were randomly selected from the test set, resulting in a set of $2 \times 5 \times 5 = 50$ reverberant binaural mixtures.

The network-based post-filter was compared to the coherence-based approach by [15], using the same block size of $32\,\mathrm{ms}$ with a time shift of $8\,\mathrm{ms}$. Similarly, the time constant involved in the IC calculation was set to $10\,\mathrm{ms}$, which gave better results than the original value of $100\,\mathrm{ms}$ suggested in [15]. The histogram-based coherence-to-gain mapping was performed using the magnitude-squared coherence function with a processing degree of 0.3, where the minimum gain value corresponded to $20\,\mathrm{dB}$ attenuation. In addition, the Bayesian filtering approach combining ARMA modeling with a HMM [13] was evaluated as a representative single-channel algorithm using the original implementation provided by the authors with the default set of parameters [38].

The dereverberation performance was assessed by three different objective metrics, namely PESQ [24] as provided by [39], STOI [25] and the normalized SRMR [26]. The SRMR metric has been shown to reflect the quality and intelligibility of reverberant speech [40] and therefore is commonly used to evaluate speech dereverberation algorithms [41]. In case of PESQ and STOI, the binaural signal consisting of the direct part only was used as a reference and the relative improvement compared to the unprocessed reverberant mixture was reported, producing ΔPESQ and ΔSTOI. The SRMR is a non-intrusive metric and, thus, the relative SRMR difference between the dereverberated speech signal and the unprocessed reverberant signal was computed, producing ΔSRMR. Binaural signals were averaged across ears prior to computing all objective metrics.

TABLE III
INFLUENCE OF DIFFERENT NEURAL NETWORK CONFIGURATIONS ON ΔPESQ AVERAGED ACROSS ALL FOUR ROOMS OF THE SURREY DATABASE. THE RELATIVE IMPROVEMENT IN PERCENT WITH RESPECT TO THE BASELINE CONFIGURATION IS SHOWN IN PARENTHESES.

Features	Temporal context	Network ensemble	# dim	Time alignment	# hidden units		
					128	256	512
IC	0	1	64	off	0.35 (baseline)	0.36 (+2.8 %)	0.36 (+2.9 %)
IC, ILD	0	1	128	off	0.39 (+11.8 %)	0.41 (+16.1 %)	0.42 (+19.1 %)
IC, ILD, IPD	0	1	192	off	0.42 (+20.4 %)	0.43 (+22.3 %)	0.44 (+25.5 %)
IC, ILD, IPD	4	1	960	off	0.44 (+26 %)	0.43 (+23.2 %)	0.46 (+30.7 %)
IC, ILD, IPD	4	5	960	off	0.45 (+29 %)	0.48 (+36.7 %)	0.51 (+44.0 %)
IC	0	1	64	on	0.35 (+0 %)	0.36 (+2.1 %)	0.36 (+2.0 %)
IC, ILD	0	1	128	on	0.40 (+12.4 %)	0.41 (+15.8 %)	0.42 (+18.6 %)
IC, ILD, IPD	0	1	192	on	0.44 (+25.6 %)	0.45 (+28.6 %)	0.46 (+29.1 %)
IC, ILD, IPD	4	1	960	on	0.47 (+33.5 %)	0.45 (+27.1 %)	0.48 (+35.9 %)
IC, ILD, IPD	4	5	960	on	0.49 (+37.6 %)	0.52 (+46.8 %)	0.55 (+55.7 %)

IV. EXPERIMENTS

Two experiments were conducted to assess the performance of the proposed dereverberation algorithm in a variety of reverberant acoustic conditions. The first experiment focused on the evaluation of the neural network-based post-filter and used the Surrey database to analyze the effectiveness of three types of short-term binaural features (IC, ILDs and IPDs) with and without time alignment. In addition, the impact of different neural network configurations, including the benefit of temporal context, ensemble averaging and the number of hidden units, was analyzed. The second experiment used both the Surrey and the Aachen database and compared the proposed neural network-based post-filter to the coherence-based dereverberation algorithm described in [15] and the single-channel Bayesian filtering approach [13]. Each of the algorithms was tested individually and in combination with a DSB. The single-channel approach was either applied to the reverberant speech averaged across both ears or to the monaural DSB output.

A. Experiment 1: Influence of short-term binaural features and neural network configuration

The dereverberation performance expressed in terms of ΔPESQ is shown in Tab. III, where each row represents a different configuration of the neural network. ΔPESQ scores were averaged across all 740 binaural mixtures using the Surrey database (see Sect. III-C). First, the results corresponding to the upper half of Tab. III are considered, which did not use the time alignment stage. When the neural network was trained only with the short-term IC using 128 hidden units (baseline configuration), a relative PESQ improvement ΔPESQ of 0.35 was achieved and performance did not increase further with increasing number of hidden units. When extending the feature space by ILDs and IPDs, an increase in ΔPESQ by 20.4 % compared to the baseline configuration was observed, which was even higher when a higher number of hidden units was used. The advantage of a broader network with more hidden units was apparent when including temporal context by stacking features from four preceding time frames. Finally, the ensemble averaging across five separately trained neural networks provided the overall largest improvement of 44 % in terms of ΔPESQ and was used in the second experiment.

One limitation of the IPD feature is it's inherent ambiguity due to spatial aliasing, whereby phase differences are wrapped to the interval $[-\pi, \pi]$. For a sound source located at $0\,^\circ$, the direct-sound signal components will produce uniform IPD values across frequency that are close to zero. This pattern, however, changes for lateral source positions, where the IPD changes linearly with frequency. Although T-F units dominated by reverberation will produce less systematic IPD values, IPDs associated with the direct-sound cover the full range between $[-\pi, \pi]$, making it more difficult for the network to distinguish between direct-sound and reverberant signal components.

Therefore, a time alignment stage prior to feature extraction was incorporated in the testing stage of the proposed algorithm (see block diagram in Fig. 1). In this way, coherent signal components associated with the direct-sound always corresponded to IPD values close to zero. Due to this consistency, the benefit of the IPD feature was substantially improved, as shown in the lower half of Tab. III. Importantly, the time alignment did not affect the contribution of the IC and the ILD feature. Thus, the neural network-based post-filter used in the second experiment included the time alignment stage.

B. Experiment 2: Comparison with coherence-based dereverberation and single-channel Bayesian filtering

Figure 2 shows the relative PESQ improvement of all tested algorithms for the Surrey database as a function of the reverberation time. Both the static beamformer ("DSB") and the coherence-based approach ("IC") of [15] provided similar improvements. Despite utilizing only one channel, the performance of Bayesian filtering ("BF") [13] was comparable to "IC", except for the condition with the lowest reverberation time ($T_{60} = 0.32$). The neural network-based post-filter ("NN") achieved substantially higher ΔPESQ scores and this performance benefit increased with increasing reverberation time. The combination of the DSB with either the coherence- ("DSB & IC") or the neural network-based post-filter ("DSB & NN") showed significant improvements, confirming that the processing of the beamformer, which mainly performed target enhancement, and the post-filters, which attenuated reverberant components, were complementary. Whereas the performance of "DSB & IC" was comparable to the network-based post-filter alone in two conditions ($T_{60} = 0.32\,\mathrm{s}$ and $T_{60} = 0.68\,\mathrm{s}$), the combination of the DSB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

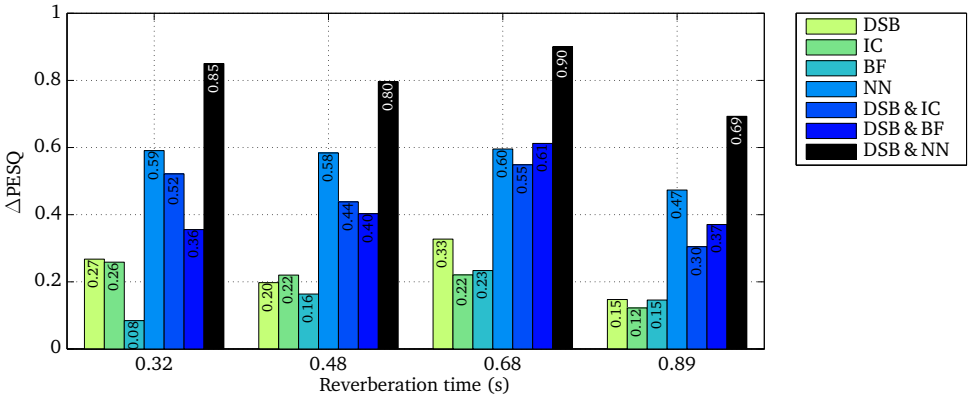


Fig. 2. ΔPESQ scores for the Surrey database as a function of the reverberation time. The results were averaged across all azimuth directions.

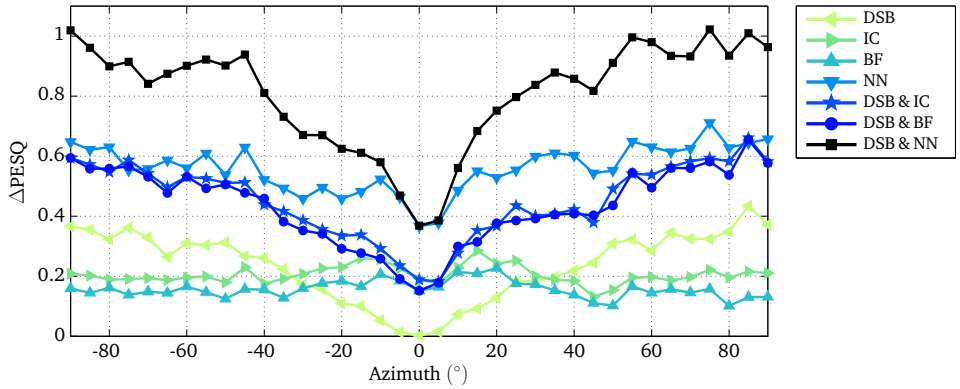


Fig. 3. ΔPESQ scores for the Surrey database as a function of the azimuth angle. The results were averaged across all rooms.

beamformer with the network-based post-filter was superior in all tested conditions.

Figure 3 shows the ΔPESQ scores of all tested algorithms as a function of the sound source azimuth. Results were averaged across all four rooms of the Surrey database. It can be seen that no benefit was achieved by the static beamformer ("DSB") for frontal source positions, which is due to the averaging of the reference signal across ears, that is already time-aligned for $0\,°$. However, performance increased with more lateral source positions, producing relative PESQ improvements of up to 0.4. In contrast, the benefit of the coherence-based approach ("IC") and single-channel Bayesian filtering ("BF") was largely independent of the source direction. The performance of the neural network-based post-filter ("NN") tended to be higher for source directions different from $0\,°$, which was presumably caused be the analysis of ILD and IPD features. Yet, the combination of the DSB with either the coherence- ("DSB & IC") or the neural network-based post-filter ("DSB & NN") showed a similar trend compared to the DSB alone, with substantially higher ΔPESQ improvements for lateral source positions, whereas the advantage of "DSB & NN" over "DSB & IC" and "DSB" was represented by a consistent offset across the full azimuth range.

Figure 4 presents ΔPESQ scores for the two rooms from the Aachen database as a function of the reverberation time. Despite the stronger amount of reverberation, the overall ranking of the different algorithms was fairly similar to the results obtained with the Surrey database (compare to Fig. 2). The improvement obtained with either the static beamformer ("DSB"), the coherence-based approach ("IC") or single-channel Bayesian filtering ("BF") was quite low, presumably due to the high amount of reverberation. Again, the combination of the DSB with the neural network-based

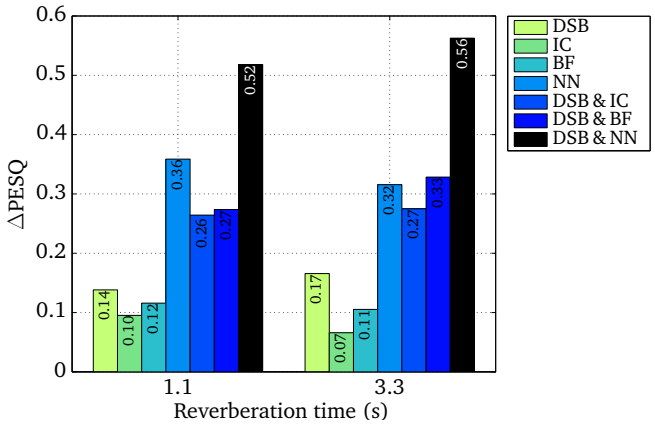


Fig. 4. ΔPESQ scores for the Aachen database as a function of the reverberation time. The results were averaged across all azimuth directions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE IV
OBJECTIVE PERFORMANCE METRICS OF ALL TESTED DEREVERBERATION ALGORITHMS AVERAGED ACROSS AZIMUTH DIRECTIONS. BOLDFACE INDICATES THE ALGORITHM THAT PRODUCED THE BEST RESULT FOR A GIVEN CONDITION.

Metric	Algorithm	Surrey [34]				Aachen [35]		Mean
		A	B	C	D	Stairway	Aula	
ΔPESQ	DSB	0.27	0.20	0.33	0.15	0.14	0.17	0.21
	IC	0.26	0.22	0.22	0.12	0.10	0.07	0.16
	BF	0.08	0.16	0.23	0.15	0.12	0.11	0.14
	NN	0.59	0.58	0.60	0.47	0.36	0.32	0.49
	DSB & IC	0.52	0.44	0.55	0.30	0.26	0.28	0.39
	DSB & BF	0.36	0.40	0.61	0.37	0.28	0.33	0.39
	DSB & NN	**0.85**	**0.80**	**0.90**	**0.69**	**0.52**	**0.56**	**0.72**
ΔSTOI (%)	DSB	3.01	3.67	2.95	5.36	4.27	5.68	4.15
	IC	0.39	1.16	-0.28	2.11	1.72	0.62	0.95
	BF	-2.83	-1.66	-2.35	-0.85	-1.43	0.82	-1.38
	NN	2.67	4.87	2.53	7.18	6.15	7.48	5.15
	DSB & IC	2.37	3.42	1.70	5.45	4.53	4.92	3.73
	DSB & BF	0.33	1.89	0.44	4.12	3.16	6.27	2.7
	DSB & NN	**4.06**	**6.08**	**3.85**	**9.02**	**7.49**	**9.74**	**6.71**
ΔSRMR	DSB	0.11	0.15	0.22	0.26	0.16	0.40	0.22
	IC	0.03	0.15	0.13	0.27	0.17	0.32	0.18
	BF	0.30	0.37	0.37	0.40	0.16	0.17	0.29
	NN	0.12	0.36	0.26	0.68	0.56	0.87	0.47
	DSB & IC	0.11	0.27	0.28	0.49	0.31	0.64	0.35
	DSB & BF	**0.40**	**0.50**	**0.58**	0.68	0.38	0.69	0.54
	DSB & NN	0.18	0.43	0.38	**0.82**	**0.64**	**1.10**	**0.59**

post-filter ("DSB & NN") achieved the best performance.

A summary of all three objective metrics is given in Tab. IV, where along with ΔPESQ, also ΔSTOI and ΔSRMR scores are shown for all tested algorithms. Results were averaged across all sound source directions for a particular room. It can be seen that the relative benefit of the network-based post-filter over the DSB and the coherence-based approach [15] is reflected in a consistent improvement in all three objective metrics. The single-channel Bayesian filtering approach performed well in terms of the ΔSRMR metric, but produced negative STOI scores. On average, the combination of the DSB with the neural network-based post-filter ("DSB & NN") performed best in all experimental conditions.

The effect of reverberation is illustrated in Fig. 5, where the spectrogram representation of a direct-sound signal (panel a) is compared to the reverberant binaural mixture (panel b) for a TIMIT sentence auralized at $-90\,^\circ$ in the Aachen stairway room ($T_{60} = 1.1\,\mathrm{s}$). The impact of reverberation on the speech signal in terms of temporal and spectral smearing is clearly visible. The remaining panels show the output of all tested dereverberation algorithms. Although the static beamformer ("DSB") shown in panel c) enhanced the target direction, which is reflected by the higher intensity of the speech harmonics, the amount of reverberation is hardly reduced. Both the coherence-based post-filter ("IC") presented in panel d) and single-channel Bayesian filtering ("BF") shown in panel e) were able to reduce the amount of temporal smearing to some extend, but the neural network-based post-filter ("NN"), shown in panel f), achieved a much stronger attenuation of the reverberant energy. This is particularly evident when comparing the gaps between words, and to some extend also the notches in between spectral harmonics, with the direct-sound signal shown in panel a). Finally, the combination of the DSB with the network-based post-filter ("DSB & NN") further enhanced the energy of speech harmonics, as illustrated in panel i).

V. DISCUSSION AND CONCLUSION

This study presented a novel neural network-based post-filter for speech dereverberation based on short-term binaural cues. The network was trained using a MCT procedure where binaural reverberant signals were simulated by mixing the direct part of HRIRs with diffuse noise. It was shown that dereverberation performance increased when three different types of short-term binaural features, namely IC, ILDs and IPDs, were jointly exploited. A systematic evaluation revealed that the proposed algorithm generalized very well to acoustic conditions not seen during training. Specifically, the network generalized to different artificial heads as well as a wide range of reverberant BRIRs, despite being trained with anechoic HRIRs.

A time alignment stage prior to feature extraction was shown to substantially increase the effectiveness of the IPD feature due to a more consistent distribution of IPD values associated with the direct-sound. Alternatively, the STFT-based feature extraction stage employed here could be replaced by an auditory-inspired front-end, where frequency-specific time differences between the two ear signals are typically analyzed by a cross-correlation function (CCF) in different subbands [20]. The CCF changes systematically with sound source azimuth and avoids the problem of estimating interaural time differences via peak picking. Consequently, using the entire CCF as a feature was shown to improve the performance of binaural sound source localization [23] and binaural speech segregation [18] systems and might as well be beneficial for the task of speech dereverberation.

The present study focused on the development of an effective post-filter for speech dereverberation. The sequential enhancement of the reverberant speech signal using a DSB and the neural network-based post-filter was shown to effectively

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2765819, IEEE/ACM Transactions on Audio, Speech, and Language Processing

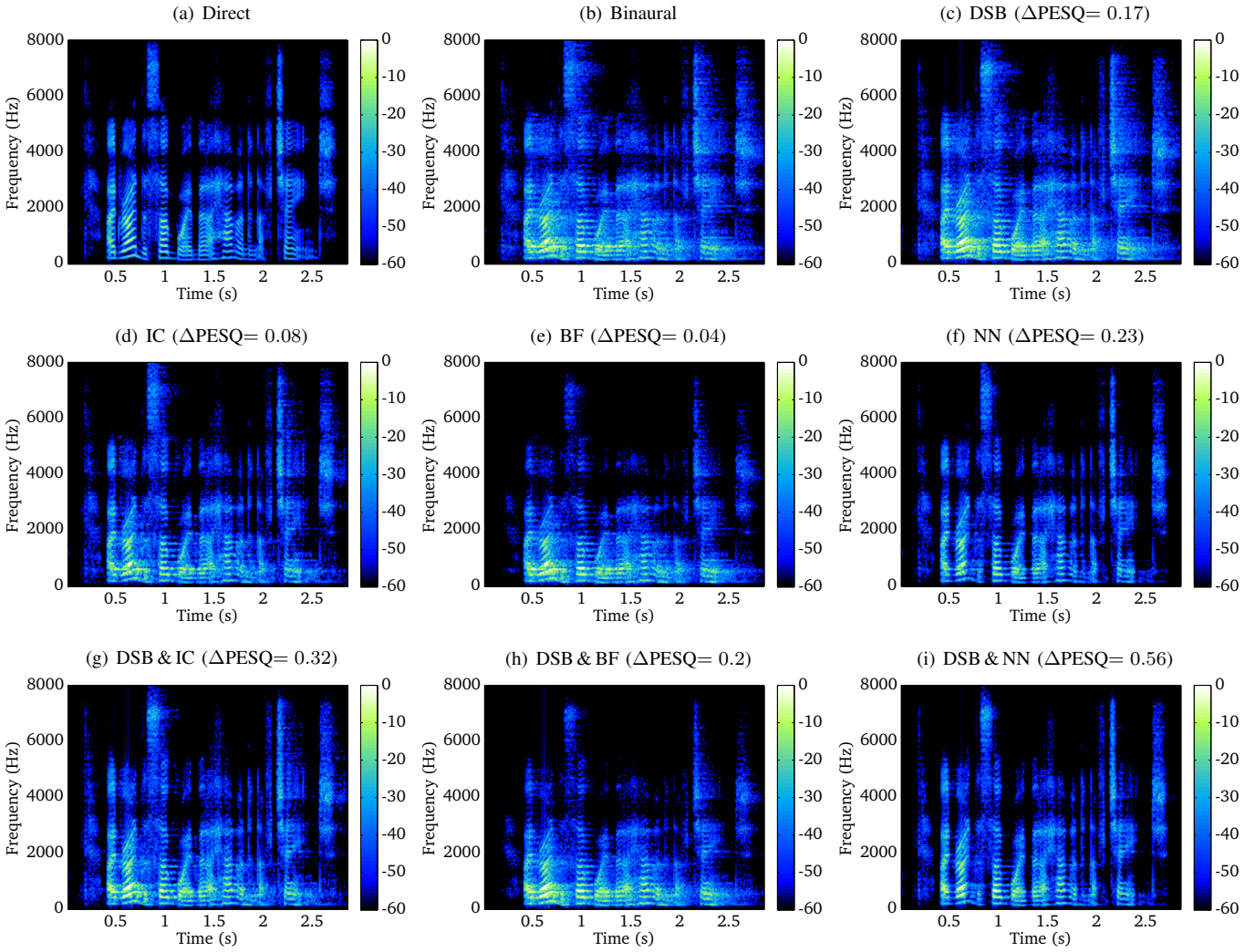


Fig. 5. Spectrogram representations for (a) the direct sound and (b) the unprocessed reverberant mixture consisting of a TIMIT sentence auralized at $-90\,^{\circ}$ in the Aachen stairway room ($T_{60} = 1.1\,\mathrm{s}$) along with all tested dereverberation algorithms: (c) DSB, (d) coherence-based post-filter (e) single-channel Bayesian filtering (f) neural network-based post-filter, (g) DSB with coherence-based post-filter, (h) DSB with single-channel Bayesian filtering and (i) DSB with neural network-based post-filter. Binaural signals were averaged across both channels prior to spectrogram analysis.

combine the complementary processing principles of target enhancement and reverberation attenuation, yielding an enhanced monaural output. For hearing aid applications, maintaining the interaural cues of the individual sound sources is required to preserve the spatial impression of the acoustic scene. Thus, the system presented here could be readily extended to a binaural cue-preserving dereverberation algorithm [42], where the neural network-based post-filter could be synchronously applied to both ear signals. In addition, more sophisticated beamformers, such as the minimum variance distortionless response (MVDR) beamformer, could be used, which can produce a true binaural output by estimating the speech components based on two reference signals (one for each ear) [43].

The proposed system was evaluated with a single target source in reverberant environments. Due to the robustness provided by the MCT, which has already been shown to enable accurate binaural localization in multi-source scenarios [22], the proposed approach has the potential to be applicable in reverberant conditions with multiple competing talkers, which has to be confirmed in future evaluations. Finally, future work will perform behavioral listeners tests to quantify the subjective benefit of the proposed dereverberation algorithm.

In summary, the neural network-based post-filter constitutes a powerful tool that allows to attenuate reverberant components independent of the room acoustic condition, the sound source direction, the source-receiver distance and the artificial head.

REFERENCES

[1] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[2] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 130, no. 2, pp. 996–1005, 2011.

[3] J. Moncur and D. Dirks, "Binaural and monaural speech intelligibility in reverberation," *J. Speech Hear. Res.*, vol. 10, no. 2, pp. 186–195, 1967.

[4] A. K. Nábělek, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. Speech Hear. Res.*, vol. 24, no. 3, pp. 375–383, 1981.

[5] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 3221–3232, 2011.

[6] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1607–1614, 2013.

[7] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1-2, pp. 123–142, 2004.

[8] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, 2012.

[9] M. Brandstein and D. Ward, Eds., *Microphone arrays: Signal processing techniques and applications*, Springer, Berlin, Germany, 2001.

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

[11] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, 2009.

[12] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *J. Appl. Signal Process.*, vol. 61, pp. 1–12, 2015.

[13] C. S. J. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 572–587, 2017.

[14] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, 1977.

[15] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2767–2777, 2013.

[16] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*, Springer, London, 1st edition, 2010.

[17] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays: Signal processing techniques and applications*, M. Brandstein and D. Ward, Eds., chapter 3, pp. 69–60. Springer, Berlin, Germany, 2001.

[18] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, 2014.

[19] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1075–1084, 2017.

[20] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, 2011.

[21] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[22] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proc. ICASSP*, 2015, pp. 2679–2683.

[23] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Interspeech*, 2016, pp. 3302–3306.

[24] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs p.862," *International Telecommunications Union (ITU-T) Recommendation*, 2001.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[26] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. IWAENC*, 2014, pp. 55–59.

[27] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[28] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[29] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2110–2117, 2002.

[30] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[31] D. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Technical report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1993.

[33] Institute of Sound Recording (IoSR), "Simulated room impulse responses," Software is available at https://www.surrey.ac.uk/department-music-media/research/institute-sound-recording-iosr/iosr-software-and-digital-resources, 2012.

[34] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modelling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, 2010.

[35] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digital Signal Process.*, 2009, pp. 1–5.

[36] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *Proc. 130th Conv. Audio Eng. Soc.*, 2011.

[37] H. Kayser, S. D. Ewert, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear hear-related and binaural room impulse responses," *J. Appl. Signal Process.*, 2009.

[38] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," Software is available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997-2017, last viewed August 2017.

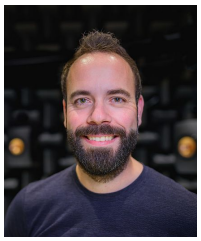
[39] P. C. Loizou, *Speech Enhancement: Theory and practice*, CRC Press, Hoboken, NJ, USA, 2nd edition, 2013.

[40] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[41] 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, pp. 1–4.

[42] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[43] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.



Tobias May studied hearing technology and audiology and received the M.Sc. degree from the University of Oldenburg, Oldenburg, Germany in 2007 and the binational Ph.D. degree from the University of Oldenburg in collaboration with the Eindhoven University of Technology, Eindhoven, The Netherlands. Since 2013, he has been with the Department of Electrical Engineering, Technical University of Denmark, first as a Postdoctoral Researcher (2013-2017), and since 2017 as an Assistant Professor. His research interests include computational auditory scene analysis, binaural signal processing, noise-robust speaker identification and hearing aid processing.