Journal Pre-proof

Speech Signal Enhancement in Cocktail Party Scenarios by Deep Learning based Virtual Sensing of Head-Mounted Microphones

Tim Fischer, Marco Caversaccio, Wilhelm Wimmer

PII: S0378-5955(21)00128-3
DOI: https://doi.org/10.1016/j.heares.2021.108294
Reference: HEARES 108294

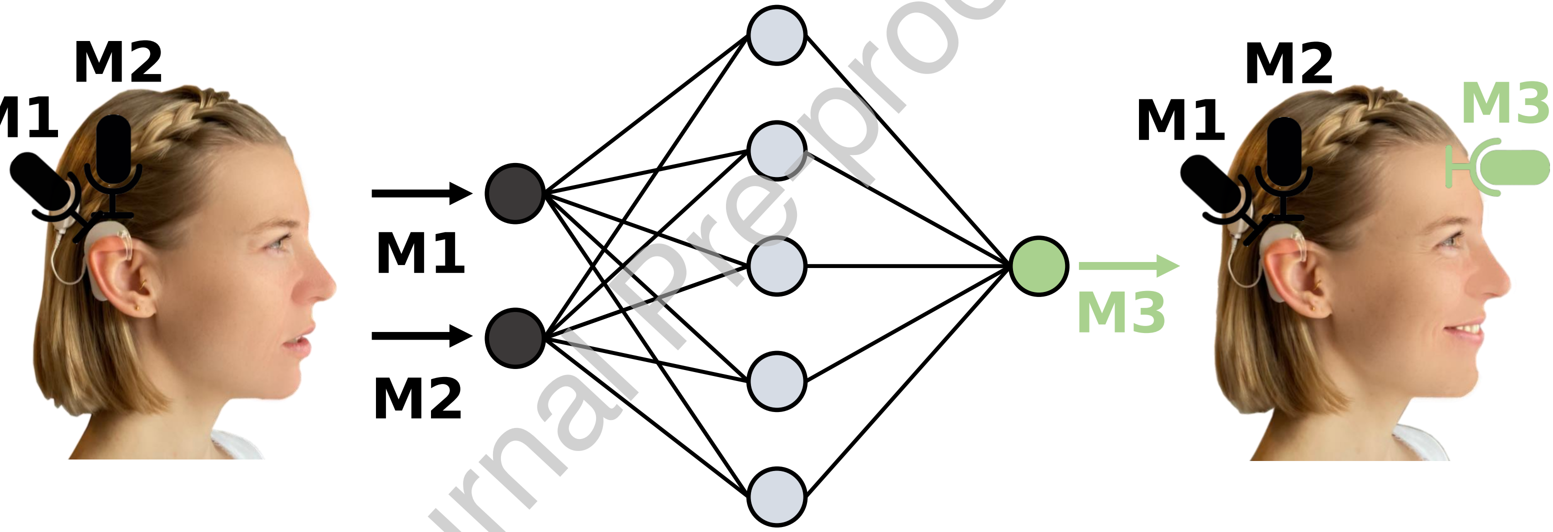
To appear in: *Hearing Research*

Received date: 8 February 2021
Revised date: 31 May 2021
Accepted date: 7 June 2021

Please cite this article as: Tim Fischer, Marco Caversaccio, Wilhelm Wimmer, Speech Signal Enhancement in Cocktail Party Scenarios by Deep Learning based Virtual Sensing of Head-Mounted Microphones, *Hearing Research* (2021), doi: https://doi.org/10.1016/j.heares.2021.108294

HIGHLIGHTS:

- Optimal positioning of the microphones is impractical.
- Deep learning can be used to virtually sense microphone signals.
- Virtual microphone signals can significantly improve the speech quality.



M1
M2
M1
M2
M3
M1
M2
M3

Speech Signal Enhancement in Cocktail Party Scenarios by Deep Learning based Virtual Sensing of Head-Mounted Microphones

Tim Fischer[a,b], Marco Caversaccio[a,b], Wilhelm Wimmer[a,b,c]

[a] *Hearing Research Laboratory, ARTORG Center for Biomedical Engineering Research, University of Bern, Bern 3008, Switzerland*
[b] *Department of ENT, Head and Neck Surgery, Inselspital, Bern University Hospital, University of Bern, Bern 3008, Switzerland*
[c] *corresponding author: Wilhelm Wimmer (wilhelm.wimmer@artorg.unibe.ch)*

Abstract

The cocktail party effect refers to the human sense of hearing's ability to pay attention to a single conversation while filtering out all other background noise. To mimic this human hearing ability for people with hearing loss, scientists integrate beamforming algorithms into the signal processing path of hearing aids or implants' audio processors.

Although these algorithms' performance strongly depends on the number and spatial arrangement of the microphones, most devices are equipped with a small number of microphones mounted close to each other on the audio processor housing.

We measured and evaluated the impact of the number and spatial arrangement of hearing aid or head-mounted microphones on the performance of the established Minimum Variance Distortionless Response beamformer in cocktail party scenarios. The measurements revealed that the optimal microphone placement exploits monaural cues (pinna-effect), is close to the target signal, and creates a large distance spread due to its spatial arrangement.

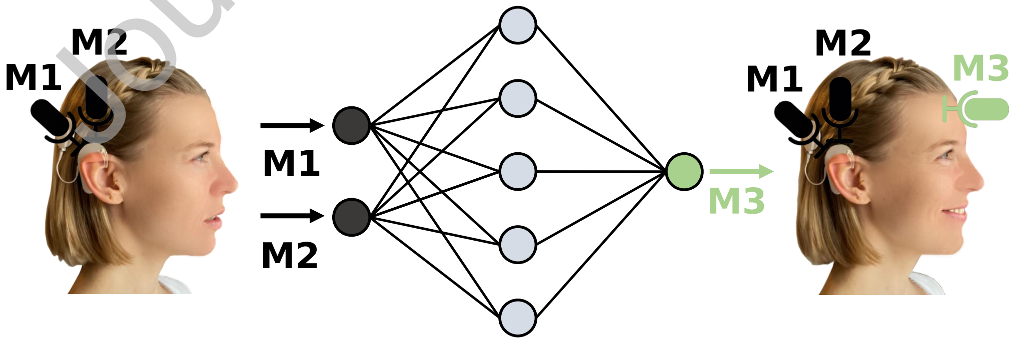
However, this microphone placement is impractical for hearing aid or implant users, as it includes microphone positions such as on the forehead. To overcome microphones' placement at impractical positions, we propose a deep virtual sensing estimation of the corresponding audio signals. The results of objective measures and a subjective listening test with 20 participants showed that the virtually sensed microphone signals significantly improved the speech quality, especially in cocktail party scenarios with low signal-to-noise ratios. Subjective speech quality was assessed using a 3-alternative forced choice procedure to determine which of the presented speech mixtures was most pleasant to understand.

Hearing aid and cochlear implant (CI) users might benefit from the presented approach using virtually sensed microphone signals, especially in noisy environments.

Keywords: artificial intelligence, selective hearing, neural network, beamformer, hearing aid, cochlear implant

Declarations of interest: none

Graphical Abstract



List of acronyms

SNR signal-to-noise ratio

BSS blind source separation

ASC acoustic scene classification

RTF relative transfer function

STFT short-time Fourier transform

ISTFT inverse short-time Fourier transform

SI-SDR scale-invariant speech to distortion ratio

SDR speech to distortion ratio

STOI short-time objective intelligibility

PESQ perceptual evaluation of speech quality

CI cochlear implant

MVDR minimum variance distortionless response

BCP Bern cocktail party

ILD interaural level difference

HRTF head related transfer function

ReLU rectified linear unit

GUI graphical user interface

dB	decibel

1. Introduction

Following a conversation in a noisy setting is difficult. In literature, this phenomenon is referred to as the cocktail-party problem. It describes an acoustic scenario, where multiple speech and noise sources with different intensities and directions of incidence overlap [1]. For normal-hearing persons, the auditory system can handle conflicting sounds and focus on a specific conversation [2, 3]. In hearing aids or CI audio processors, this separation of the conversational partner from a noise tangle is the goal of sophisticated beamforming algorithms [4, 5, 6, 7].

It is well known that the signal quality of beamforming algorithms increases with the number of available input microphones and their positioning with respect to the target source [8, 9, 10, 11, 12, 13]. Using numerical experiments, Feng et al. [8] showed that the microphone positions play an essential role in the overall performance of beamforming algorithms. Jones et al. [14] further showed for CI users that the microphone position at the ear canal versus behind the ear led to more detailed interaural level difference (ILD) information due to the frequency transformations of the pinna [15, 16]. In the specific case of unilateral CI users, it was demonstrated that an additional microphone positioned at the contralateral ear led to increased speech understanding in noise [17, 13, 18].

Since many conversations are held face to face [19], it is reasonable to assume that additional microphones in positions other than the contralateral ear canal, e.g., on the forehead, may further improve speech understanding. However, the additional placement of microphones on the head is impractical from the perspective of a hearing aid or CI user. One way of circumventing this limitation may be to place the microphones virtually rather than physically. The results of several virtual microphone sensing approaches suggest that estimating an additional microphone signal using information from the available microphones may improve the speech quality in a cocktail party scenario [20, 21, 22]. The microphone array used to record the reference signals was similar in the studies and consisted of 2 microphones positioned in a straight line at a distance of 4 cm [20, 21] or 3 cm [22] from each other. To generate virtual microphone signals, the phase was linearly interpolated [20, 21] or extrapolated [22] using measurements of the real microphone signals. In Denk et al. [23], functions transformed the sound pressure at a microphone positioned on a hearing aid to the pressure measured at the open eardrum. The basis for the determination of these functions were the relative transfer functions (RTFs) between the microphones, which in turn were determined by head related transfer functions (HRTFs) measurements using frequency sweeps in an anechoic chamber. Also using frequency sweeps, Corey et al. [24] measured and evaluated impulse responses of 160 microphones spread across the body and affixed to wearable accessories. Their results suggest that microphone arrangements with large spatial distance spread across the body provided the best signal-to-noise ratio (SNR) values. Unlike microphones positioned on the head, the geometric arrangement of microphones placed on clothing may change according to posture. Likely, the quality of a beamforming algorithm defined for a specific microphone geometry suffers from the continually changing microphone geometries in everyday life [25].

The tremendous progress in the field of machine learning leads to the expectation that in the future, the RTFs between microphones can be determined purely data-driven, i.e., without prior knowledge of the specific measurement setup. As a result, beamforming algorithms could be tuned to individual array geometries by simply providing sufficient reference data from the wearer without the need for anechoic chambers or knowledge of the sound sources' positions. In the Mic2Mic publication [26] it was demonstrated that even with unlabeled and unpaired data, audio signals between different microphone domains could be translated. Based on the results, an additional virtual microphone at the head of a hearing aid or CI user generated or learned solely by data-driven rules seems like a realistic scenario. However, regardless of whether the microphones are placed virtually or physically on a subject's head, little is known about how their positioning affects beamforming.

To continue the discussion, the first objective of this work was to systematically investigate the speech signal quality in complex acoustic scenarios with varying head-mounted microphone arrangements and a minimum variance distortionless response (MVDR) beamformer as introduced by Souden et al. [10]. Based on these measurements, virtual microphone signals at specific positions were estimated using a deep neural network. Finally, subjective listening tests were conducted to investigate to what extent the virtually sensed microphone signals could improve the speech signal quality.

2. Methods

2.1. Linear observation model

In this work, recordings from $M = 16$ microphones attached to a human head were used. Each of the $i = 1 \ldots M$ microphone signals $y_i(t)$ recorded varying acoustic cocktail party scenarios at time t. In the following, the cocktail party mixtures are described as the summation of the target speech source $s_i(t)$ and the noise $w_i(t)$ at microphone i:

$$y_i(t) = a_i s(t - \tau_i) + w_i(t)$$

where τ_i represents the time-delay of arrival and a_i is the amplitude modulation depending on the geometric arrangement of the microphones under the assumption of anechoic conditions. The noise $w_i(t)$ is assumed to be uncorrelated with the signal $s_i(t)$.

To enhance the perception of the target speech sources, the signals at each microphone can be combined using "beamforming" techniques. In this study, we used the widely studied MVDR beamformer [27, 28], which is introduced in the following section.

2.2. MVDR beamforming

The MVDR beamformer minimizes the power of the beamformed signal while preserving the target signal, under the constraint of no distortion in the target signal [10]. The MVDR is a filter-and-sum beamformer and as such it applies different phase weights $h_i(f)$ to the i input microphone channels in order to steer the main lobe of the directivity pattern to the direction of the target signal. The phase weights, or filters, are obtained in the frequency domain using [29]:

$$\boldsymbol{h}_{ref}(f) = [h_{1,ref}(f), \ldots, h_{M,ref}(f)]^T = \frac{1}{\lambda(f)} \left(\boldsymbol{G}(f) - \boldsymbol{I}_{M\times M}\right) \boldsymbol{e}_{\text{ref}} \quad (1)$$

Where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{G}(f)$ can be obtained by $\boldsymbol{G}(f) = \boldsymbol{\Phi}_{\text{noise}}^{-1}(f)\boldsymbol{\Phi}_{\text{obs}}(f)$ with $\lambda(f) = \text{trace}(\boldsymbol{G}(f)) - M$ [30, 10]. The spatial covariance matrices $\boldsymbol{\Phi}$ can be computed by using time-frequency masks [29, 31, 32, 33]. However, in this work we focus on the impact of additional microphone channels on the MVDR beamformers performance and extract $\boldsymbol{\Phi}_{\text{noise}}^{-1}(f)$, $\boldsymbol{\Phi}_{\text{obs}}(f)$ and $\boldsymbol{\Phi}_{\text{target}}(f)$ from the noise, observation and target recordings.

The standard unit vector of the reference microphone $\boldsymbol{e}_{\text{ref}}$, is selected by a maximum a posteriori expected SNR estimation. The reference microphone is chosen based on $\text{ref} = \underset{r}{\text{argmax}}\, \text{SNR}_{\text{post},r}$ [29] and:

$$\text{SNR}_{\text{post},r} = \frac{\sum_{f=0}^{F-1} \boldsymbol{h}_r^H(f)\boldsymbol{\Phi}_{\text{target}}(f)\boldsymbol{h}_r(f)}{\sum_{f=0}^{F-1} \boldsymbol{h}_r^H(f)\boldsymbol{\Phi}_{\text{noise}}(f)\boldsymbol{h}_r(f)}.$$

Thus, the reference channel or microphone depends on $\boldsymbol{h}_r(f)$, which is the M-dimensional filter response (see Eq. 1) at the discrete frequency index $f = 0, \ldots, F-1$, when $\boldsymbol{e}_{\text{ref}}$ is set to $\boldsymbol{e}_{\text{r}}$. After the filters $\boldsymbol{h}_{ref}(f)$ are computed, the beamformed output $z_{t,f}$ is obtained by using the short-time Fourier transforms (STFTs) $y_{i,t,f}$ of the microphone signals $y_i(t)$:

$$z_{t,f} = \sum_{i=1}^{M} h_{i,ref}(f) y_{i,t,f}$$

For the MVDR beamformer, the input signals were down-sampled to 8 kHz and a Blackman window was applied [34]. Subsequently, an STFT (size = 256 and shift = 128) was performed. To reconstruct the signal, an inverse short-time Fourier transform (ISTFT) with the overlapadd strategy was applied. The herein used MVDR beamformer to evaluate the benefits of virtual microphone signals is just one application scenario. Theoretically, any multi-channel speech-enhancement algorithm could have been used to assess the benefits of virtually sensed microphone signals.

2.3. Data

The Bern cocktail party (BCP) dataset is tailored to this work, as it contains multi-microphone recordings of hearing aid or CI users in cocktail party scenarios [35]. For the recordings, 12 loudspeakers (Control 1 Pro, JBL, Northridge, USA) were aligned horizontally in a circle at the height of the ears (1.2 m) in an acoustic chamber [36, 37, 13]. For this work, we used the acoustic scenarios captured with 16 microphones (ICS-40619, TDK, Tokyo, Japan) attached to a head and torso simulator (Brel & Kjær, Type 4128, Nærum, Denmark) (see Figures 1 and 2).

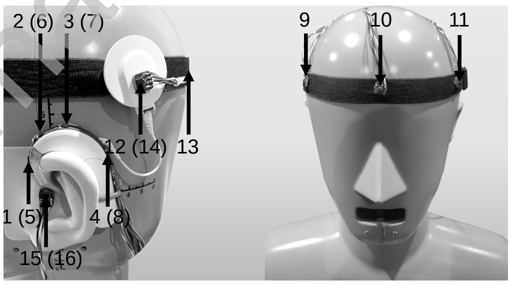


Figure 1: Placement of the 16 microphones used for cocktail party scenario recordings. The IDs refer to the microphone signals assignment in the multi-channel recording audio files [35]. Numbers in brackets refer to the contralateral (here: right side) assignment of the microphones. The sagittal plane is defined by a straight line between microphones 10 and 13 (front and back). A numeric description can be found in Table 1.

Table 1: Assignment of the 16 microphone positions to their respective IDs.

Microphone ID	Microphone position
{1}	Left audio processor. Facing forward.
{2}	Left audio processor. Facing to the top / forward.
{3}	Left audio processor. Facing to the top / backward.
{4}	Left audio processor. Facing back.
{5}	Right audio processor. Facing forward.
{6}	Right audio processor. Facing to the top / forward.
{7}	Right audio processor. Facing to the top / backward.
{8}	Right audio processor. Facing backward.
{9}	Right temple.
{10}	Front.
{11}	Left temple.
{12}	Left transmission coil.
{13}	Back.
{14}	Right transmission coil.
{15}	Left Ear. Entry of the ear canal.
{16}	Right Ear. Entry of the ear canal.

2.3.1. Test dataset

The results of this work were computed with an excerpt of 2400 samples from the BCP dataset [35]. The duration of each sample was 1.5 s, resulting in a total test dataset duration of 1 h. The samples were randomly chosen under the constraint, that a majority of the recordings contain a target source azimuth inside the field of view (i.e., $\pm 45°$), as this represents the most natural listening scenario [38] (see Figure 3). All samples were randomly selected from an SNR distribution which covered conversational speech levels with 1 to 3 competing speakers and varying background noise types and intensities. The distribution of the audio mixture on the 12 output channels covered scenarios of spatially separated and non-separated speech and noise sources. The samples or audio mixtures had a mean SNR value of 1.2 dB with a standard deviation of 10.9 dB.

2.3.2. Training dataset

For the training and validation of the deep neural network 65 h (78404 audio samples with 3 s duration each) were randomly selected from the head and torso simulator recordings of the BCP dataset [35], excluding the test dataset (see Section 2.3.1). Ninety percent of the samples were used for training and 10% for validation. Because of the large size of the training and validation dataset, no cross-validation was performed.

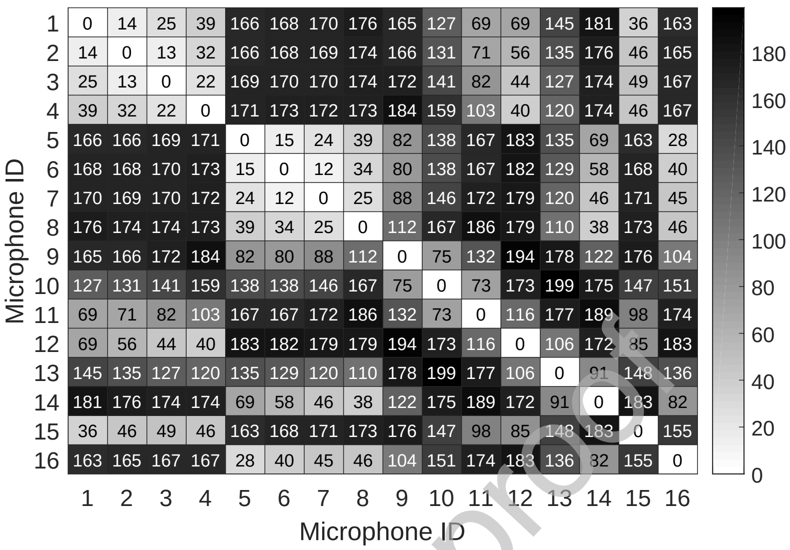


Figure 2: Euclidean distances in millimeters between the microphones for the head and torso simulator measurements [35].

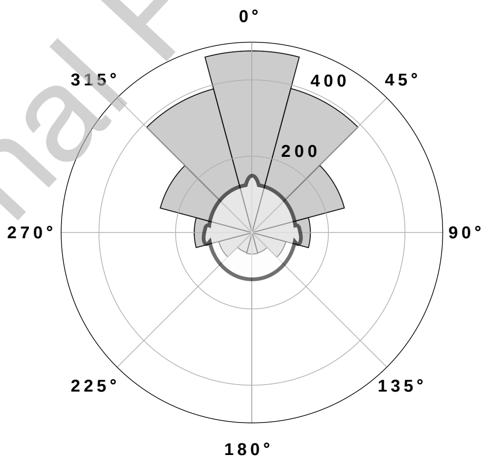


Figure 3: Circular histogram of the frequency of occurrence of spatial source directions in relation to the head and torso simulator azimuth. The audio files were were selected such that the directional distribution assumes a von-Mises distribution with $\mu = 0.0$ and $\kappa = 1.1$ [35].

2.4. Evaluation of microphone channel configurations

Various microphone channel configurations were evaluated by adding or omitting microphone channels with respect to a reference microphone channel configuration, as explained in detail later (Section 3, Tables 3-6). The results were computed by providing the MVDR beamformer [10] with the target and noise spatial covariance matrices $\mathbf{\Phi}$ of the audio mixtures from the corresponding microphone configurations.

The reference microphone configurations were selected to cover reasonable microphone inputs of hearing aid devices or audio processors. Care was also taken to ensure that all microphones in the unilateral reference microphone configurations could technically be connected to the audio processor using an existing cable such as from the CI transmission coil to the audio processor.

To cover realistic use cases regarding the benefits of different microphone configurations, the results were divided into 4 categories rather than presenting all possible microphone channel combinations: subsets of unilateral CI microphone configurations (see Table 3), unilateral CI microphone configurations with additional ipsilateral microphones (Table 4), unilateral CI microphone configurations with additional contralateral microphones (Table 5), symmetric bilateral CI configurations with additional microphones (Table 6). An overview of all measured microphone configurations can be found in Table 2.

For the evaluation of the microphone configurations (i.e., real recordings and virtually sensed microphone channels), the following objective speech quality metrics were assessed: perceptual evaluation of speech quality (PESQ) [39], short-time objective intelligibility (STOI) [40] and scale-invariant speech

Table 2: Overview of all measured microphone configurations.

Unilateral microphone configurations	Bilateral microphone configurations
{1}	{1, 2, 3, 4, 9}
{2}	{1, 2, 3, 4, 14}
{3}	{1, 2, 3, 4, 16}
{4}	{1, 2, 3, 4, 5, 6, 7, 8}
{10}	{1, 2, 3, 4, 5, 6, 7, 8, 10}
{11}	{1, 2, 3, 4, 5, 6, 7, 8, 13}
{12}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 11}
{13}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16}
{15}	{1, 2, 3, 4, 5, 6, 7, 8, 15, 16}
{1, 2}	{2, 3, 9}
{1, 2, 3, 4}	{2, 3, 14}
{1, 2, 3, 4, 10}	{2, 3, 16}
{1, 2, 3, 4, 11}	{2, 3, 6, 7}
{1, 2, 3, 4, 12}	{2, 3, 6, 7, 10}
{1, 2, 3, 4, 13}	{2, 3, 6, 7, 13}
{1, 2, 3, 4, 15}	{2, 3, 6, 7, 9, 11}
{1, 3}	{2, 3, 6, 7, 15, 16}
{1, 4}	{2, 3, 10, 13, 16}
{2, 3}	
{2, 3, 10}	
{2, 3, 11}	
{2, 3, 12}	
{2, 3, 13}	
{2, 3, 15}	
{2, 4}	
{3, 4}	

to distortion ratio (SI-SDR) [41]. The PESQ metric models the speech quality as perceived by human listeners. Analysis of speech-audio with the PESQ metric usually ranges from 1.0 (high distortion) to 4.5 (no distortion) [39]. The values of STOI range from 0.0 (no word correctly understood) to 1.0 (all words correctly understood) and highly correlate with the intelligibility of degraded speech signals [40]. The SI-SDR metric defines the energy ratio between the clean target signal and the acoustic distortions in decibel (dB). It is a slightly modified version of speech to distortion ratio (SDR), making it insensitive to power rescaling of the estimated signal [41].

For testing within a group of microphone configurations, the Friedman test was used (see Sections 3.1 and 3.2). To find the configurations that differed significantly after the Friedman test has rejected the null hypothesis, a post-hoc Nemenyi test was performed. In Section 3.3, two sets of paired samples were compared to each other with the two-sided Wilcoxon signed-rank test (no multiple testing). The significance level was chosen with $\alpha = 0.05$ for all statistical tests.

2.5. Virtual sensing of a microphone channel

The virtual sensing approach aimed to improve the speech quality in cocktail party scenarios by providing the beamformer with additional, virtually sensed, microphone signals. In this work, the estimation of the virtual microphone signals was realized by a purely data-driven deep learning approach on the raw-audio mixture without preprocessing [42].

Most applications of deep neural networks in the domain of audio signal processing address the enhancement of speech signals by separating a target source (speech) from a mixture of interfering noise sources [43]. In the work presented here, however, no source separation was performed, but rather, in a transferred sense, a denoising of the reference signal, as explained in the following: Let the audio signal captured from a microphone inside the ear canal of the left ear be the reference signal and the audio signal inside the ear canal of the right ear the target signal. By trying to match the signal of the left ear to the right ear or denoise the left ear, we hypothesize that the network implicitly learns the RTF between the two microphone signals or, in other words, the "noise" to remove from the audio signal of the left ear. As a result, the network tries to virtually sense the right ear's audio input by using the signal of the left ear. To evaluate the quality of the virtually sensed microphones, spatial covariance matrices $\mathbf{\Phi}$ with and without virtually sensed microphone signals were provided as input for the MVDR beamformer [10]. The results were compared with the same metrics and statistics as with the real microphones measurements (see Section 2.4).

In this study, two microphone signals were used as reference signals, and three additional microphone signals were virtually sensed. The 2 reference signals consisted of the microphones $\{2, 3\}$ and were chosen because their spatial arrangement corresponds to that of a conventional CI audio processor (see Figure 1 or Table 1). Motivated by the results of the head-mounted microphone measurements, the microphone on the forehead ($\{10\}$), the back ($\{13\}$) and inside the ear canal of the contralateral ear ($\{16\}$) were chosen as target signals for the virtual sensing approach. In the remainder of the manuscript, virtual channels are indicated by the subscript v. The resulting microphone configuration ($\{2,\ 3,\ 10_{\mathrm{v}},\ 13_{\mathrm{v}},\ 16_{\mathrm{v}}\}$) provided the advantages as explained in the Discussion (Section 4.1): a high spatial spread of the microphone signals [44], proximity to the target signal, and frequency transformations by the pinna and head shadow [15].

2.5.1. Deep neural network architecture for the virtual sensing approach

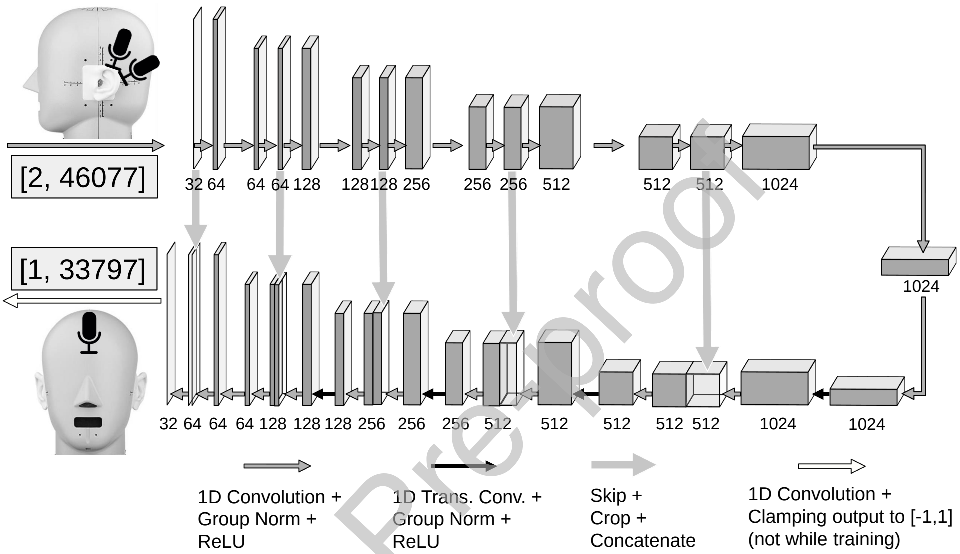


Figure 4: The proposed deep neural network architecture for the virtual sensing of additional microphone channels based on the work of Stoller et al. [42]. The numbers below the blocks describe the input channel size of the following convolution. Shown is an example for the estimation of the microphone signal on the forehead ({10}) with the measurement data of 2 microphones as positioned in conventional cochlear implant (CI) audio processors (microphones {2, 3}). The network's input and output data blocks denoted with "[A, B]" describe the number of channels (A) and the number of samples (B). For an illustration of the microphone placement, please see Figure 1.

The network architecture followed the U-Net adaption for end-to-end audio source separation in the time-domain [42]. The neural network operation on the raw-waveform in the time domain allowed to model the phase information of the audio signal, thus avoiding complex phase recovery algorithms

[45, 46]. The well known U-Net structure is composed as a convolutional autoencoder, and as such, consists of an encoder (contracting path), a bottleneck, and a decoder (expanding path) [47]. A diagram of our network's architecture implementation is shown in Figure 4.

In the encoder, an increasing number of higher-level features on coarser time scales were calculated, allowing the modeling of long-term dependencies in the audio signal. Our implementation of the encoder consisted of 5 levels, with each level working on half the time resolution and twice the number of feature maps as the previous one. In the bottleneck, the model was forced to learn a compression of the input data, containing only the relevant information (latent space) to construct the virtual microphone signal. The latent-space representation of the bottleneck layer was passed to the decoder, which tried to learn a mapping of the input data to match the desired virtual microphone signal. The decoder was the mirror image of the encoder and also consisted of 5 levels. Each level worked on double the time resolution and half the number of feature maps as the previous level. Based on the results of initial tests, transposed convolutions were used for the upsampling process. Each convolution was followed by group normalization, and a rectified linear unit (ReLU) activation function [48, 19]. By introducing the skip connections in the encoder-decoder architecture, the encoder's high-level features were concatenated with the local features computed during the upsampling block of the decoding. The result of this concatenation were multi-scale features that were fed in the output layer of the network [47, 42]. The output of the last convolutional layer was the estimation of the virtually sensed microphone signal.

The receptive field of the model was chosen to work with 2.1 s (46077 samples), which provided an output vector with the desired test size of 1.5 s (33797 samples).

Since no implicit zero padding was performed in the convolution operation, the neural network's output sample size was smaller than the input sample size. Avoiding zero-padding allowed the convolutions to be performed in the correct audio context. As a result, audio artifacts in the results could be minimized, and the temporal continuity of the audio signal was better preserved [42].

2.5.2. Network training

To train the deep virtual sensing network, we extracted measurement data from the two reference channels ($\{2, 3\}$) and the microphone channel to be estimated. Due to the large size of the BCP training dataset (see Section 2.3.2), no data augmentation was necessary. In accordance with the original Wave-U-Net implementation [42], the audio data of the BCP dataset [35] was downsampled to 22.05 kHz. For evaluating the network's performance, the absolute differences between the actual value and the predicted value (L_1 loss) were used. To update the network weights iteratively based on training data, we applied the ADAM optimizer [49] with the default decay rates of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 16 [42]. Instead of monotonically decreasing the learning rate, cyclical learning rates [50] were used with upper and lower boundaries of 0.0002 and 0.00001, respectively. Early stopping was performed after 10 epochs with only minimal improvement on the validation loss. Afterward, the best model was fine-tuned with lower learning rate limits (0.000001 to 0.00001) and a batch size of 8, again until 10 epochs without improvement on the validation loss. The fine-tuned network was further used to predict the virtual channels. The test dataset to evaluate the virtually sensed microphone channels consisted of 2400 samples, which included the audio files described in Section 2.3.1. Care was taken to ensure that none of the test samples were used to validate or train the network.

Since each virtual channel was estimated on a separate network, the networks were trained one after the other. The training time was reduced by successively using the previously trained network as a starting-point (transfer learning) [51]. All computations were performed with the open-source machine learning framework PyTorch version 1.6.0 [52].

2.5.3. Subjective listening tests

Twenty normal hearing participants (6 female, 14 male, mean age in years = 29.8, SD = 3.6) performed a subjective listening test to evaluate the benefit of the virtually sensed microphone signals on the speech quality. The test was performed in a quiet environment, and stimuli were presented via high definition insert earphones (Triple Driver, 1 More Inc. San Diego, CA) at the most comfortable loudness levels as selected by the subjects.

The questions of the subjective evaluation were twofold. First, we asked the subjects whether the signal processing applied by the MVDR beamformer lead to overall improved speech quality. Second, it was evaluated whether the beamformed signal based on the reference channels ($\{2, 3\}$) with additional virtual channels ($\{10_\mathrm{v}\}$, $\{13_\mathrm{v}\}$, $\{16_\mathrm{v}\}$) outperforms the beamformed signal without virtual channels available, i.e. only the measured channels $\{2, 3\}$ were used (see Figure 1 or Table 1 for a transcription of the channel IDs).

To answer these questions, the participants were asked to listen to 3 audio mixtures, all based on the same recording but either

- Beamformed based on the reference channels with additional virtual channels ($\{2, 3\} + \{10_\mathrm{v}\}, \{13_\mathrm{v}\}, \{16_\mathrm{v}\}$)
- Beamformed based on the reference channels only ($\{2, 3\}$)
- The non-beamformed recording of the channels $\{2, 3\}$

The 3 audio mixtures were randomly assigned to 3 buttons on a graphical user interface (GUI). Since the beamformer's task was to enhance the speech quality for a predefined target signal, a fourth button on the GUI labeled "Target Signal" played back a recording of the corresponding target speech signal without interfering background noise. Finally, the participants' task was to select from the 3 audio mixtures the one in which the target signal was most comfortable to understand. Before the test started, trial runs were conducted until the participants confirmed that they understood the test procedure.

During the test and the trial runs, the participants were allowed to hear the 4 audio files (1 target signal and 3 audio mixtures) as many times as desired. The test stimuli consisted of 60 audio mixture quartets of 1.5 seconds length per file, ensuring that each file contained the utterance of at least one word. All audio mixtures were taken from the pool of the 2400 test files described in Section 2.3 with distribution proportions as shown in Figure 3. Evaluation of the presented audio files took about 20 minutes; no feedback was given during or after the test. After evaluating 30 of the 60 audio files, a pause of 3 minutes was taken during which the GUI was disabled. To minimize order bias, the 2 stimuli blocks that were evaluated before and after the pause were counter-balanced within the participants. The subjective listening evaluation was designed in accordance with the Declaration of Helsinki, written informed consent was obtained from all participants.

A Kruskal-Wallis test was used to determine if the frequency of choices within the 3 response options differed significantly from each other. After the Kruskal-Wallis test has rejected the null hypothesis, a post-hoc Nemenyi test was performed to investigate which of the response distributions differed significantly from each other. To determine whether the response distributions differed significantly from the chance level of the test (33 %), a chi-square test was applied. The significance level was chosen with $\alpha = 0.05$ for all statistical tests.

3. Results

3.1. MVDR beamforming with unilateral channel configurations

Table 3 shows the PESQ, STOI and SI-SDR performances of unilateral single microphone configurations compared to the performance with the reference configuration, i.e. a CI audio processor equipped with 4 microphones placed on top of the housing. For the PESQ and SI-SDR metric, the performances with single microphones were significantly worse than with the 4-channel reference configuration (all $p = 0.001$). The same was observed for STOI ($p = 0.001$) except for the microphones {1, 4} and {2, 4} (both $p = 0.9$). In all 3 metrics, the microphones that were facing the front (front {10}, left temple {11}, forward facing (audio processor) {1}, see Figure 1 or Table 1) achieved the best results, whereas the performance differences between channels {10} and {11} were not statistically significant in terms of PESQ and SI-SDR ($p = 0.608$, $p = 0.9$) but for STOI ($p = 0.001$). Between the microphones {1} and {2} the metrics PESQ, STOI and SI-SDR did not differ significantly ($p = 0.408$, $p = 0.9$, $p = 0.115$) (a significance-matrix showing the results of the post-hoc Nemenyi tests for Table 3 can be found in the Appendix (Figures A.1-A.3)).

When the same 4-channel reference configuration (microphones {1, 2, 3, 4}) was extended by the aforementioned ipsilateral single microphone signals, again the front-facing microphones {10} and {11} (see Figure 3) provided the greatest benefit (see Table 4). The performance differences for all metrics when channel {10} (front) was added did not differ significantly from the performance differences when channel {11} (left temple) was added to the reference configuration (PESQ: $p = 0.792$, STOI: $p = 0.736$, SI-SDR: $p = 0.9$)

(a significance-matrix showing the results of the post-hoc Nemenyi tests for Table 4 can be found in the Appendix (Figures A.4-A.9)).

Since many CI audio processors record signals from 2 microphones positioned on top of the housing, the performance of different spatial arrangements of 2 microphones placed on the audio processor compared to the 4-channel reference configuration (microphones $\{1, 2, 3, 4\}$) was investigated and is shown in Table 3. The arrangement with the largest spatial distance between the 2 microphones, namely the microphones on top of the audio processors facing the front and back ($\{1, 4\}$), achieved the best performance (see Figure 2 for a microphone distance matrix). The statistical analysis showed that the performance differences of the microphones $\{1, 4\}$ did not differ significantly for PESQ and STOI from the results compared to the microphones on the audio processor facing the top and the back ($\{2, 4\}$) to the reference configuration ($p = 0.668$, $p = 0.9$). Both 2 channel microphone configurations did not differ significantly from the 4 channel reference configuration in terms of STOI (both $p = 0.9$). For the SI-SDR metric, the differences when adding $\{1, 4\}$ did not differ statistically significantly from any of the tested 2 channel configurations (all $p = 0.9$).

The arrangement with the smallest inter-microphone distance (microphones $\{2, 3\}$, see Figures 1 and 2), which is related to the conventional microphone positions of CI audio processors, achieved the lowest scores in 2 (STOI and SI-SDR) of the 3 evaluated objective metrics, even though for SI-SDR the differences of this configuration did not differ significantly from any of the tested 2 channel configurations (all $p = 0.9$). For the metrics PESQ and STOI no significant differences in the performances between the microphones {2, 3}, {1, 2} or {1, 3} were observed (PESQ: $p = 0.721$, $p = 0.601$, STOI: $p = 0.884$, $p = 0.134$). Table 4 shows the impact on the PESQ, STOI and SI-SDR metrics when additional ipsilateral, including those on the sagittal plane, microphones were added to the the conventional microphone arrangement ({2, 3}). The extension of the microphone arrangement ({2, 3}) with forward facing microphones (front {10} or left temple {11}) provided the greatest benefit. For none of the 3 tested metrics did the performance between adding the front ({10}) or left temple ({11}) microphone to the conventional microphone arrangement differ significantly (PESQ: $p = 0.067$, STOI: $p = 0.678$, SI-SDR: $p = 0.251$).

Table 3: Values represent the mean difference in the performance of the unilateral cochlear implant (CI) microphone configurations compared to the mean performance of the *reference channel configuration* including channels positioned on the sagittal plane (see Figure 1). The best result for each metric is marked in bold. All performance differences were statistically significant compared to the reference channel performance, except those marked with "†".

	Metric		
Microphone IDs	**PESQ**	**STOI**	**SI-SDR**
Ref.: {1, 2, 3, 4}	*1.77*	*0.48*	*-29.07*
{1}	-0.28	-0.06	-2.95
{2}	-0.28	-0.06	-3.13
{3}	-0.29	-0.06	-3.13
{4}	-0.31	-0.07	-3.32
{10}	**-0.24**	**-0.03**	-2.77
{11}	-0.25	-0.04	**-2.65**
{12}	-0.30	-0.07	-3.24
{13}	-0.35	-0.08	-3.52
{15}	-0.29	-0.06	-3.19
{1, 2}	-0.17	-0.03	-1.25
{3, 4}	-0.13	-0.02	-0.86
{1, 3}	-0.15	-0.03	-0.97
{1, 4}	**-0.08**	**-0.01**†	**-0.77**
{2, 3}	-0.16	-0.03	-1.32
{2, 4}	-0.09	-0.01†	-0.89

Table 4: Values represent the mean difference in the performance of unilateral cochlear implant (CI) microphone configurations when additional ipsilateral, including sagittal plane, microphones were added (see Figure 1). The performance difference is calculated in relation to the mean performance of the *reference channel configuration*. The best result for each metric is marked in bold. All performance differences were statistically significant compared to the reference channel performance, except those marked with "†".

	Metric		
Microphone IDs	**PESQ**	**STOI**	**SI-SDR**
Ref.: {1, 2, 3, 4}	*1.77*	*0.48*	*-29.07*
Ref. + {10}	0.18	**0.04**	**0.69**
Ref. + {11}	**0.20**	0.04	0.59
Ref. + {12}	0.02	<0.01	0.14†
Ref. + {13}	0.11	0.03	0.64
Ref. + {15}	0.01	<0.01†	-0,39†
Ref.: {2, 3}	*1.61*	*0.45*	*-30.38*
Ref. + {10}	0.22	**0.06**	**1.38**
Ref. + {11}	**0.23**	0.06	1.10
Ref. + {12}	0.12	0.03	0.81
Ref. + {13}	0.15	0.04	0.92
Ref. + {15}	0.03	<0.01	0.30

3.2. MVDR beamforming with bilateral channel configurations

Table 5 shows the PESQ, STOI and SI-SDR performances when additional bilateral microphones were added to the input signals of an unilateral CI audio processor equipped with 4 microphones placed on top of the housing (microphones {1, 2, 3, 4}, see Figure 1 or Table 1). When a single contralateral microphone was added, it was not the microphone closest to the target source (microphone {9}, temple) that provided the greatest benefit in terms of the human perception-related objective metrics PESQ and STOI, but the contralateral ear canal microphone {16}. Compared to adding channels {9} or {14} (temple or contralateral CI transmission coil), the improvement of the PESQ and STOI metrics were significantly better when adding the contralateral ear-canal microphone (all $p = 0.001$) (a significance-matrix showing all results of the post-hoc Nemenyi test for Table 5 can be found in the Appendix (Figures A.10-A.15)). However, in terms of SI-SDR, the input from the microphone on the contralateral CI transmission coil (microphone {14}) achieved the best SI-SDR values and even outperformed the microphone configuration compared to an additional contralateral 4-channel CI audio processor. All differences in SI-SDR with the contralateral transmission coil microphone ({14}) compared to {9} (contralateral temple), {16} (contralateral ear canal) and Ref. ch. + {5, 6, 7, 8} were not statistically significant ($p = 0.362$, $p = 0.802$, $p = 0.409$). Since the cable connection between the CI transmission-coil and the audio processor could theoretically be exploited to transmit audio signals, a unilateral microphone configuration was also used as a reference, which included the coil signal ({12}) in addition to the 4 microphones on the audio processors. The results showed in

Table 5 did differ only marginally and non significantly between the reference configuration with the CI transmission coil ({1, 2, 3, 4, 12}) and the reference configuration without the CI transmission coil microphone ({1, 2, 3, 4}). The small benefit of adding microphone {12} to the reference channel configuration is also indicated by the results of Table 4.

An analysis of the results with a reference microphone configuration based on the conventional spatial microphone arrangement in CI audio processors (microphones {2, 3}, see Figure 1 or Table 1), lead to similar conclusions as with the 4-channel microphone configuration described above (see Table 5). Again, the overall result of a single additional microphone positioned at the contralateral ear-canal {16} was best, but only with respect to STOI and PESQ. For the PESQ metric, the performance with an additional microphone positioned in the contralateral ear canal differed non-significantly compared to the performance with an additional microphone on the temple ({9}) ($p = 0.763$). In terms of SI-SDR, the microphones on the contralateral side which were close to the sagittal plane (temple {9} and transmission coil {14}) outperformed the contralateral ear-canal microphone {16} when added to the microphone configuration {2, 3} ($p = 0.006$, $p = 0.9$). An additional, identical, bilaterally connected processor with 2 microphones ({6, 7}) yielded significantly better values in all metrics than adding the single microphones shown in Table 5 (see Appendix Figure A.13-A.15 for p-values).

Table 5: Values represent the mean difference in the performance of unilateral cochlear implant (CI) microphone configurations when additional contralateral microphones were added (see Figure 1). The performance difference is calculated in relation to the mean performance of the *reference channel configuration*. The best result for each metric is marked in bold. All performance differences were statistically significant compared to the reference channel performance.

	Metric		
Microphone IDs	**PESQ**	**STOI**	**SI-SDR**
Ref.: {1, 2, 3, 4}	*1.77*	*0.48*	*-29.07*
Ref. + {9}	0.12	0.03	0.41
Ref. + {14}	0.16	0.03	**0.80**
Ref. + {16}	0.19	0.04	0.42
Ref. + {5, 6, 7, 8}	**0.30**	**0.05**	0.61
Ref.: {2, 3}	*1.61*	*0.45*	*-30.38*
Ref. + {9}	0.18	0.04	1.30
Ref. + {14}	0.19	0.04	1.28
Ref. + {16}	0.21	0.05	1.13
Ref. + {6, 7}	**0.26**	**0.06**	**1.44**

When a bilateral CI processor microphone configuration was taken as a reference (microphones {1, 2, 3, 4, 5, 6, 7, 8}, see Table 6), adding a microphone to the front ({10}) provided more benefit than adding a microphone facing the back ({13}) (PESQ and STOI: $p = 0.001$), but for SI-SDR not statistically significant ($p = 0.515$) (a significance-matrix showing all results of the post-hoc Nemenyi test for Table 6 can be found in the Appendix (Figures A.16-A.21)). The single front microphone achieved even similar and statistically not significantly differing STOI and SI-SDR values compared to the performance when adding 2 microphones at the left and right temple ({9,11}) (both metrics $p = 0.9$). For PESQ however, the performance with the additional 2 temple microphones ({9,11}) differed statistically significant compared to the additional microphone facing to the front ({10}) ($p = 0.001$). Adding the signals of the two in-ear microphones ({15, 16}) to the bilateral CI processor microphone configuration (microphones {1, 2, 3, 4, 5, 6, 7, 8}) did not provide any benefit, not even if only 2 bilateral ({2, 3, 6, 7}) instead of 4 ({1, 2, 3, 4, 5, 6, 7, 8}) bilateral processor microphones were used as a reference microphone configuration. The full 16-channel microphone configuration achieved the statistically significant best PESQ scores (all $p = 0.001$). However, in terms of STOI and SI-SDR the performance did barely, and for SI-SDR non significantly, differ compared to the 8-channel reference microphone configuration. Again, as with the unilateral 4-microphone CI audio processor configuration, adding the transmission-coil microphone signals ({12, 14}) to the bilateral microphone configurations ({1, 2, 3, 4, 5, 6, 7, 8} or {2, 3, 6, 7}) did barely and statistically not significant influence the performance metrics shown in Table 6.

Table 6: Values represent the mean difference in the performance of bilateral cochlear implant (CI) microphone configurations when additional microphones were added (see Figure 1). The performance difference is calculated in relation to the mean performance of the *reference channel configuration.* The best result for each metric is marked in bold. All performance differences were statistically significant compared to the reference channel performance, except those marked with "†".

	Metric		
Microphone IDs	**PESQ**	**STOI**	**SI-SDR**
Ref.: {1, 2, 3, 4, 5, 6, 7, 8}	*2.07*	*0.54*	*-28.46*
Ref. + {10}	0.11	0.01	0.02†
Ref. + {9, 11}	0.12	**0.02**	**0.11**
Ref.+ {15, 16}	-0.01	-0.01	-0.56
Ref.+ {13}	0.05	0.01	0.06†
Ref. + {9, 10, 11, 12, 13, 14, 15, 16}	**0.19**	0.01	-0.61†
Ref.: {2, 3, 6, 7}	*1.87*	*0.51*	*-28.94*
Ref. + {10}	0.16	0.03	0.49
Ref. + {13}	0.11	0.02	0.20
Ref. + {9, 11}	**0.22**	**0.04**	**0.81**
Ref. + {15, 16}	0.04	<0.01	-0.39†

3.3. *Virtual sensing of microphone channels*

The bar graphs in Figure 5 compare the performance in PESQ, STOI and SI-SDR (see Methods Section 2.4) between virtually sensed microphone signals and actually measured microphone signals placed at the same position on the head, i.e. the front ($\{10\}$), the back ($\{13\}$) and at the entry of the right external auditory canal ($\{16\}$) (see Figure 1 or Table 1). For all 3 objective speech quality metrics tested, adding virtually sensed microphone signals to the input signals of the MVDR beamformer resulted in a significant improvement compared to the performance with microphone signals as used in conventional CI audio processors ($\{2, 3\}$) ($p < 0.001$).

The mean benefit in performance when additional virtual/measured microphone signals were used for beamforming was 0.24/0.34 units for PESQ, 0.06/0.07 units for STOI, and 1.17/1.25 dB for SI-SDR. For the PESQ and STOI metrics, the performance between the virtually sensed microphone signals and the measured microphones signals differed significantly ($p < 0.001$). In terms of SI-SDR, no significant difference between the two configurations were observed ($p = 0.998$).

An analysis of the performance of the neural networks with respect to each of the estimated channels $\{16\}$, $\{13\}$ and $\{10\}$ showed that the mean benefit when an additional virtual/measured microphone signal was used for beamforming was 0.154/0.211, 0.114/0.149, 0.178/0.219 for PESQ, 0.049/0.052, 0.028/0.032, 0.042/0.048 for STOI, and 1.000/1.057, 0.938/0.877, 1.493/1.377 for SI-SDR. For the metrics PESQ and STOI the differences in performance between the additional virtually estimated microphone and the measured microphone were significant (all $p < 0.001$). For SI-SDR, the differences were significant only with respect to microphone channel $\{10\}$ ($p = 0.027$), but not for the channels $\{13\}$ and $\{16\}$ ($p = 0.244$, $p = 0.309$). The on average bad results for channel $\{16\}$, meaning the largest difference between the benefit of additional virtual/measured microphone signals, and the best results for channel $\{10\}$ were also reflected in the validation losses of the trained networks. For channel $\{16\}$, $\{13\}$ and $\{10\}$, the best L_1-losses on the validation set were 2.1×10^{-4}, 1.5×10^{-4} and 1.4×10^{-4}, respectively.

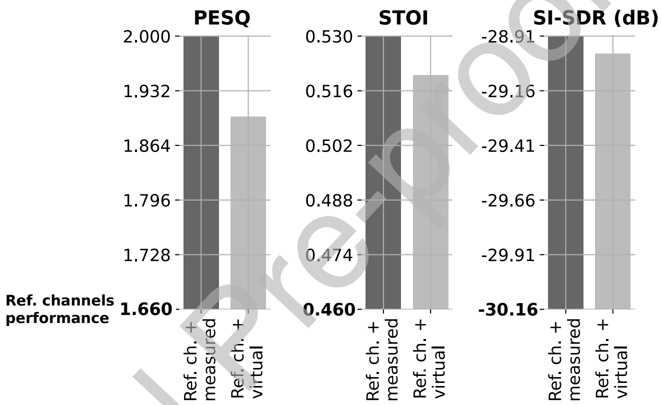


Figure 5: Comparison of overall perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) and scale-invariant speech to distortion ratio (SI-SDR) scores between 3 different microphone channel configurations used as input signals for the minimum variance distortionless response (MVDR) beamforming algorithm [10]: 1) Reference channel configuration according to the conventional microphone placement on CI audio processors (microphone IDs $\{2, 3\}$) (bold letters); 2) Reference channel configuration with additional measured (real) microphones (microphone IDs $\{2, 3\} + \{10, 13, 16\}$) (dark grey bar); 3) Reference channel configuration with additional virtually sensed microphones (microphone IDs $\{2, 3\} + \{10_v, 13_v, 16_v\}$) (light grey bar). The dataset used to evaluate the microphone channel configurations consisted of 2400 cocktail party audio samples, as described in Section 2.3. Please see Figure 1 or Table 1 for a description of the microphone IDs.

3.3.1. Subjective listening tests

Figure 6 shows that the participants preferred the audio mixture that was beamformed using the additional virtual channels (Mean=65%, SD=8%) compared to a beamformed signal generated using only the microphones as placed in CI audio processors (Mean=23%, SD=4%). This difference in selection frequency was statistically significant with $p < 0.001$.

The non-beamformed signal was rarely selected as the signal that was easiest to understand (Mean=13%, SD=7%). The beamformed signal based on the reference channel only and the beamformed signal based on additional virtual channels differed significantly to the non-beamformed audio mixture selection frequency ($p = 0.002$, $p < 0.001$).

For all of the presented signal configurations, the distribution of the frequency of choices differed significantly from the chance level of the test (all $p < 0.001$).

To investigate if the subjects' choice of the signal most comfortable to understand was dependent on the SNR of the original or raw audio mixture, the SNRs of the corresponding raw audio mixtures were compared. It was observed that the subjects preferred the beamformed signal with additional virtual channels if the SNRs of the raw audio mixture were low (Mean=2.4, SD=9.3) compared to the raw audio mixtures' SNRs when the beamformed signal based on the reference channels only was selected (Mean=5.2, SD=8.0, $p = 0.001$). The SNRs of the raw audio mixtures when the non-beamformed signal was selected (Mean=2.1, SD=9.2) was not significantly different from the SNRs of the raw audio mixtures when the beamformed signal with additional virtual channels was selected ($p = 0.987$). However, it was significantly different from the SNRs of the raw audio mixtures when the beamformed signal based on the reference channels was chosen ($p = 0.029$).

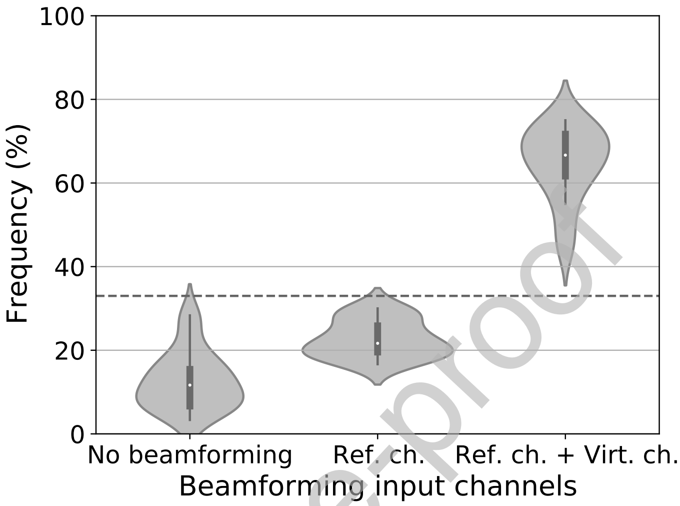


Figure 6: Violin plots [53] of the frequency of choices in the subjective listening test. The data represents the choices for the non-beamformed signal, the beamformed signal with the measured reference channel configuration as input channels (microphone IDs {2, 3}) and the beamformed signal with additional virtually sensed microphone signals as input channels (microphone IDs {2, 3} + {10_v, 13_v, 16_v}) (see Figure 1 or Table 1). The dashed horizontal line indicates the chance level of the test. The probability of observations taking a given value (Frequency (%)) is indicated by the violin's width, while each violin is normalized to have the same area. The thick black bar in the center of the violin represents the interquartile range. The thin black line extended from it represents the 95% confidence intervals, and the white dot represents the median.

4. Discussion

Herein, we presented a comprehensive comparison of different head-mounted microphone configurations and their effect on the output of an MVDR beamforming algorithm. The results showed that microphone positions, such as placing a microphone on the forehead, would be desirable for better speech understanding. Since these microphone positions are not practicable in reality, we proposed and evaluated a purely data-driven virtual sensing technique.

4.1. Association of the speech quality and the microphone positioning

Our measurements of varying head-mounted microphone arrangements in cocktail party scenarios confirmed that the performance of beamforming algorithms and thus the speech quality improves with additional microphone signals [44]. Single-microphone speech-enhancement algorithms can only exploit temporal and spectral information cues, whereas multi-microphone beamformers can additionally exploit the spatial information of the sound sources [10, 44].

However, a high number of microphones alone does not necessarily lead to a better speech quality [10]. In the case of bilaterally placed microphones (Table 6), we observed saturation in terms of speech signal enhancement with additional microphones that were placed close to the reference microphones. In particular, the SI-SDR metric showed that noise from additional microphone signals can dominate compared to the redundant information in the audio signal used for speech enhancement. As also shown by Corey et al. [24], the microphone arrangement's spatial diversity played a significant role in the quality of the acoustic beamforming. The herein performed measurements confirmed this finding since no improvements were observed when additional microphones were placed at a distance of about 5 cm to the reference microphones. It was assumed that even for low frequencies, these microphones were too closely spaced to provide inter-microphone information for the beamforming algorithm [24]. Besides, the microphones' distance was too small for an effect of the acoustic head shadow [15]. With the same reasoning, the slightly worse result of the unilateral, conventional microphone configuration ($\{2, 3\}$) and the good result of the arrangement with the largest inter-microphone distance (front and back facing $\{1, 4\}$) compared to other 2-channel microphone arrangements on the audio processor can be argued.

Although adding a microphone with a high Euclidean distance to the reference microphone configuration is a good rule of thumb to improve acoustic beamforming, other microphone positioning factors, such as exploiting the acoustic head shadow [15], may be just as important. In the unilateral configuration (see Table 4), we observed that the proximity to the most likely target source with an additional microphone on the temple ($\{2, 3\}+\{11\}$) was more important than the spatial diversity of the microphones with an additional microphone placed on the back of the head ($\{2, 3\}+\{13\}$). In addition to the proximity to the target signal and the microphone distance, our measurements confirmed that the pinna's directional frequency transformation provided relevant information for improving the quality of the beamforming algorithm [15, 54, 16]. We observed that the most useful additional contralateral microphone was neither the one closest to the target signal ($\{11\}$, temple) nor the one with the highest Euclidean distance to the reference microphone configuration ($\{14\}$, CI transmission coil). It was the contralateral microphone placed in the ear canal facing away from the target signal ({16}).

4.2. Virtual sensing of head-mounted microphone signals

In this work, we presented and evaluated a method for virtual sensing of microphone signals to improve the speech quality of hearing aid and CI users in noisy environments. The proposed methodology enabled to capture microphone signals at positions on the head, including but not limited to the forehead, where a physical placement of microphones is impractical. Our objective measurements showed, that adding strategically positioned virtual microphones on the head significantly improved the speech quality compared to the speech quality as obtained with a microphone arrangement found in conventional CI audio processors. This result was also confirmed in human listening tests using a 3-alternative forced-choice procedure with the task of selecting the speech mixture that was most comfortable to understand.

In addition to the general assumption that adding microphone signals to hearing aid applications can increase the performance of beamforming algorithms [44], we hypothesized and confirmed that replacing real microphone signals with virtual microphone signals can also increase beamformer performance. In contrast to the work presented in [22, 21, 20], our entirely data-driven approach showed that explicit knowledge of the real microphone's positioning might not be necessary to enhance the speech quality with virtual microphone channels. The mathematical reasoning for the success of our deep learning-based approach is the subject of ongoing research [55, 56].

In the measurements with the reference microphone configuration according to conventional CI audio processors ({2, 3}), we observed that an additional microphone on the forehead produced similar improvements in speech quality as an additional microphone placed at the entry of the contralateral ear canal. However, due to the poor estimation of the contralateral ear signal by the neural network, a higher benefit was obtained with the virtual microphone channel estimating the signal at the forehead. Therefore the estimation of optimal microphone positions for neural network-based beamforming approaches requires further investigation.

The subjective feedback of the 20 participants significantly showed that the additional virtual microphone signals were preferred, especially in cocktail party scenarios with low SNRs. On the other hand, the participants' choices also showed that in low SNRs scenarios, the MVDR beamforming, either with real or real and additional virtual channels, might degrade the subjective speech signal quality instead of enhancing it. This finding confirmed that although MVDR beamformers aim to keep the target signal undistorted [7], there was a trade-off between noise reduction and speech signal distortion [10].

4.3. Limitations and outlook

Although the virtually sensed microphones significantly improved the speech quality within this study, further research is needed before the methodology can be used in hearing aids or CI audio processors.

Due to the input data size of 2 seconds, the delay of the proposed network architecture is too long to be applicable in a real hearing aid application. However, this paper's main objective was to demonstrate a proof of concept for purely data-driven virtual channel estimations in hearing aids or CIs. Tackling the problem of latency and neural network complexity in online speech enhancement is ongoing research [57, 58, 59, 60] with promising results and input frame lengths as little as 2 ms [60]. Future research should investigate whether the significant reduction in network time delay required for an application in hearing devices affects the performance of the presented approach. In addition to progress in reducing the computational costs, substantial progress is continuously being made in other areas of speech signal enhancement with artificial neural networks relevant for the methodology of this work, such as in blind source separation (BSS) [61, 62, 63], acoustic scene classification (ASC) [64, 65, 66], domain shift [26, 67] and the usage of loss functions to optimize the parameters of the network based on the human perception of speech [68, 59]. The results of Drude et al. [63] indicated, that the benefit of the presented approach when using estimated coherence matrices may be different from the benefit achieved with the oracle matrices. For computational time reasons, no sophisticated optimization of the presented network's architecture was performed. Further research may investigate the optimal number and size of hidden layers for the presented approach.

Our approach follows a two-step procedure to estimate a virtual microphone channel that is used as an additional input to the beamformer. We chose this procedure to improve the compatibility with existing beamforming technology in current devices. However, the entire approach could be replaced by an end-to-end single-network artificial intelligence solution for hearing devices.

One of the biggest challenges of the presented methodology to be applicable in a real-world application will be to ensure the robustness of the network's predictions in acoustic environments with high reverberation [69, 70, 71, 72]. In the context of this work, the first step in this direction would be the use of more challenging acoustic training data, for example, by simulating conditions with higher reverberation [73] or the use of dynamically moving sound sources [36, 74]. Another possibility would be to record acoustic scenarios using a portable microphone array [75]. In a real-world application, this data could be collected as part of an audiological fitting routine. In both cases, whether the data was simulated or recorded in real environments for each subject, the additional recordings and the personalization of the network through transfer learning would most likely increase the robustness of applied neural network solutions [76]. To account for the different head geometries and thus varying inter-microphone features, the information of 3D head scans as provided in Fischer et al. [35] could be fed into a neural network architecture that allows metadata injection.

Although the speech quality may improve by applying the proposed measures, binaural cues would still be discarded, resulting in a low spatial quality of the perceived sounds [15]. It remains unclear whether the findings of this study will also hold for current state-of-the-art beamformers with binaural output. To preserve the binaural cues and thus improve the spatial quality of the MVDR beamforming algorithm [10], adaptations such as those proposed by Marquardt et al. [77] or Marquardt and Doclo [78] could yield improvements in this regard while still enhancing the speech quality [79].

5. Conclusions

In this work, real and virtual microphone signals were combined as input for an MVDR beamformer to investigate the effects on speech quality for hearing aid or CI users in cocktail party scenarios. The measurements with respect to the number and spatial arrangement of real microphones indicated that, optimally, microphones should be placed as close as possible to the target source, encode monaural cues, and produce a large distance spread by their spatial arrangement. In reality, however, it is inconvenient to place the microphones according to these criteria. To overcome this problem, virtual microphone signals were estimated using a deep neural network without explicit knowledge of the spatial microphone arrangement. The results of 3-alternative forced choice subjective listening tests and objective speech quality metrics suggest that hearing aid or CI users might benefit from virtually sensed microphone signals, especially in challenging cocktail party scenarios.

Appendix A. Additional Figures

Please see appendix_A.pdf for significance-matrices of the post-hoc Nemenyi tests concerning the data in Tables 3-6.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] E. C. Cherry, Some Experiments on the Recognition of Speech, with One and with Two Ears, The Journal of the Acoustical Society of America 25 (1953) 975–979.

URL: http://asa.scitation.org/doi/10.1121/1.1907229. doi:10.1121/1.1907229.

[2] J. Peissig, B. Kollmeier, Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners, The Journal of the Acoustical Society of America 101 (1997) 1660–1670.

[3] N. Mesgarani, E. F. Chang, Selective cortical representation of attended speaker in multi-talker speech perception, Nature 485 (2012) 233–236.

[4] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1998) 21–34.

[5] S. Makino, T.-W. Lee, H. Sawada, Blind speech separation, volume 615, Springer, 2007.

[6] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, D. Yu, Past review, current progress, and challenges ahead on the cocktail party problem, Frontiers of Information Technology & Electronic Engineering 19 (2018) 40–63.

[7] B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering, IEEE assp magazine 5 (1988) 4–24.

[8] Z. G. Feng, K. F. C. Yiu, S. E. Nordholm, Placement design of microphone arrays in near-field broadband beamformers, IEEE transactions on signal processing 60 (2011) 1195–1204.

[9] J. Wouters, J. V. Berghe, Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system, Ear and hearing 22 (2001) 420–430.

[10] M. Souden, J. Benesty, S. Affes, On optimal frequency-domain multichannel linear filtering for noise reduction, IEEE Transactions on audio, speech, and language processing 18 (2009) 260–276.

[11] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, J. Dmochowski, New insights into the mvdr beamformer in room acoustics, IEEE Transactions on Audio, Speech, and Language Processing 18 (2009) 158–170.

[12] W. Wimmer, M. Caversaccio, M. Kompis, Speech intelligibility in noise with a single-unit cochlear implant audio processor, Otology & neurotology 36 (2015) 1197–1202.

[13] W. Wimmer, M. Kompis, C. Stieger, M. Caversaccio, S. Weder, Directional microphone contralateral routing of signals in cochlear implant users: A within-subjects comparison, Ear and hearing 38 (2017) 368–373.

[14] H. G. Jones, A. Kan, R. Y. Litovsky, The effect of microphone placement on interaural level differences and sound localization across the horizontal plane in bilateral cochlear implant users, Ear and hearing 37 (2016) e341.

[15] J. Blauert, Spatial hearing : the psychophysics of human sound localization, MIT Press, 1997.

[16] T. Fischer, C. Schmid, M. Kompis, G. Mantokoudis, M. Caversaccio, W. Wimmer, Pinna-imitating microphone directionality improves sound localization and discrimination in bilateral cochlear implant users, Ear and hearing 42 (2021) 214.

[17] R. Arora, H. Amoodi, S. Stewart, L. Friesen, V. Lin, J. Nedzelski, J. Chen, The addition of a contralateral routing of signals microphone to a unilateral cochlear implant systema prospective study in speech outcomes, The Laryngoscope 123 (2013) 746–751.

[18] M. F. Dorman, S. C. Natale, S. Agrawal, The value of unilateral cis, ci-cros and bilateral cis, with and without beamformer microphones, for speech understanding in a simulation of a restaurant environment, Audiology and Neurotology 23 (2018) 270–276.

[19] Y. Wu, K. He, Group normalization, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[20] H. Katahira, N. Ono, S. Miyabe, T. Yamada, S. Makino, Nonlinear speech enhancement by virtual increase of channels and maximum snr beamformer, EURASIP Journal on Advances in Signal Processing 2016 (2016) 1–8.

[21] K. Yamaoka, L. Li, N. Ono, S. Makino, T. Yamada, Cnn-based virtual microphone signal estimation for mpdr beamforming in underdetermined situations, in: 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8903040.

[22] R. Jinzai, K. Yamaoka, M. Matsumoto, S. Makino, T. Yamada, Wavelength proportional arrangement of virtual microphones based on interpolation/extrapolation for underdetermined speech enhancement, in: 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 2019, pp. 1–5.

[23] F. Denk, S. M. Ernst, S. D. Ewert, B. Kollmeier, Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles, Trends in hearing 22 (2018) 2331216518779313.

[24] R. M. Corey, N. Tsuda, A. C. Singer, Acoustic impulse responses for wearable audio devices, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 216–220.

[25] I. Himawan, S. Sridharan, I. McCowan, Dealing with uncertainty in microphone placement in a microphone array speech recognition system, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 1565–1568.

[26] A. Mathur, A. Isopoussu, F. Kawsar, N. Berthouze, N. D. Lane, Mic2mic: Using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems, in: Proceedings of the 18th International Conference on Information Processing in Sensor Networks, 2019, pp. 169–180.

[27] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proceedings of the IEEE 57 (1969) 1408–1418.

[28] E. A. Habets, J. Benesty, S. Gannot, I. Cohen, The mvdr beamformer for speech enhancement, in: Speech Processing in Modern Communication, Springer, 2010, pp. 225–254.

[29] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, J. Le Roux, Improved mvdr beamforming using single-channel mask prediction networks., in: Interspeech, 2016, pp. 1981–1985.

[30] J. Benesty, J. Chen, Y. Huang, Microphone array signal processing, volume 1, Springer Science & Business Media, 2008.

[31] N. Ito, S. Araki, T. Nakatani, Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing, in: 2016 24th European Signal Processing Conference (EUSIPCO), IEEE, 2016, pp. 1153–1157.

[32] T. Higuchi, K. Kinoshita, M. Delcroix, T. Nakatani, Adversarial training for data-driven speech enhancement without parallel corpus, in: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2017, pp. 40–47.

[33] L. Drude, R. Haeb-Umbach, Tight integration of spatial and spectral features for bss with deep clustering embeddings., in: Interspeech, 2017, pp. 2650–2654.

[34] R. B. Blackman, J. W. Tukey, Particular pairs of windows, The measurement of power spectra, from the point of view of communications engineering (1959) 98–99.

[35] T. Fischer, M. Caversaccio, W. Wimmer, Multichannel acoustic source and image dataset for the cocktail party effect in hearing aid and implant users, Scientific Data 7 (2020).

URL: https://doi.org/10.1038%2Fs41597-020-00777-8. doi:10.1038/s41597-020-00777-8.

[36] T. Fischer, M. Kompis, G. Mantokoudis, M. Caversaccio, W. Wimmer, Dynamic sound field audiometry: Static and dynamic spatial hearing tests in the full horizontal plane, Applied Acoustics 166 (2020) 107363. doi:10.1016/j.apacoust.2020.107363.

[37] T. Fischer, M. Caversaccio, W. Wimmer, A front-back confusion metric in horizontal sound localization: The fbc score, in: ACM Symposium on Applied Perception 2020, 2020, pp. 1–5.

[38] Y.-H. Wu, E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, J. Oleson, Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild-to-moderate hearing loss, Ear and Hearing 39 (2018) 293.

[39] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, in: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), volume 2, IEEE, 2001, pp. 749–752.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Transactions on Audio, Speech, and Language Processing 19 (2011) 2125–2136.

[41] J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey, Sdr–half-baked or well done?, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 626–630.

[42] D. Stoller, S. Ewert, S. Dixon, Wave-u-net: A multi-scale neural network for end-to-end audio source separation, arXiv preprint arXiv:1806.03185 (2018).

[43] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE Journal of Selected Topics in Signal Processing 13 (2019) 206–219.

[44] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, K. R. Liu, Acoustic beamforming for hearing aid applications, Handbook on array processing and sensor networks (2010) 269–302.

[45] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Transactions on acoustics, speech, and signal processing 32 (1984) 236–243.

[46] K. Yatabe, Y. Masuyama, Y. Oikawa, Rectified linear unit can assist griffin-lim phase recovery, in: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 555–559.

[47] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[48] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, H. S. Seung, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, Nature 405 (2000) 947–951.

[49] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[50] L. N. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 464–472.

[51] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242–264.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performan

[53] J. L. Hintze, R. D. Nelson, Violin plots: a box plot-density trace synergism, The American Statistician 52 (1998) 181–184.

[54] W. Wimmer, S. Weder, M. Caversaccio, M. Kompis, Speech intelligi-

bility in noise with a pinna effect imitating cochlear implant processor, Otology & neurotology 37 (2016) 19–23.

[55] S. Yu, J. C. Principe, Understanding autoencoders with information theoretic concepts, Neural Networks 117 (2019) 104–123.

[56] F. Fan, J. Xiong, G. Wang, On interpretability of artificial neural networks, Preprint at https://arxiv. org/abs/2001.02522 (2020).

[57] A. Pandey, D. Wang, Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6875–6879.

[58] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, N. Harada, Real-time speech enhancement using equilibriated rnn, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 851–855.

[59] Y. Koyama, T. Vuong, S. Uhlich, B. Raj, Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks, arXiv preprint arXiv:2005.11611 (2020).

[60] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation, IEEE/ACM transactions on audio, speech, and language processing 27 (2019) 1256–1266.

[61] J. Lu, W. Cheng, D. He, Y. Zi, A novel underdetermined blind source separation method with noise and unknown source number, Journal of Sound and Vibration 457 (2019) 67–91.

[62] L. Drude, R. Haeb-Umbach, Integration of neural networks and probabilistic spatial models for acoustic blind source separation, IEEE Journal of Selected Topics in Signal Processing 13 (2019) 815–826.

[63] L. Drude, D. Hasenklever, R. Haeb-Umbach, Unsupervised training of a deep clustering model for multichannel blind source separation, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 695–699.

[64] K. Koutini, H. Eghbal-zadeh, M. Dorfer, G. Widmer, The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification, in: 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8902732.

[65] S. S. R. Phaye, E. Benetos, Y. Wang, Subspectralnet using sub-spectrogram based convolutional neural networks for acoustic scene classification, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 825–829. doi:10.1109/ICASSP.2019.8683288.

[66] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894.

[67] Z. Borsos, Y. Li, B. Gfeller, M. Tagliasacchi, Micaugment: One-shot microphone style transfer, arXiv preprint arXiv:2010.09658 (2020).

[68] S.-W. Fu, C.-F. Liao, Y. Tsao, S.-D. Lin, Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement, arXiv preprint arXiv:1905.04874 (2019).

[69] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, Y. Gong, Cracking the cocktail party problem by multi-beam deep attractor network, in: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 437–444. doi:10.1109/ASRU.2017.8268969.

[70] S. Inoue, H. Kameoka, L. Li, S. Seki, S. Makino, Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 96–100. doi:10.1109/ICASSP.2019.8683497.

[71] F. Feng, M. Kowalski, Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 442–456. doi:10.1109/TASLP.2018.2881925.

[72] Y. Xie, K. Xie, J. Yang, Z. Wu, S. Xie, Underdetermined reverberant audio-source separation through improved expectation–maximization algorithm, Circuits, Systems, and Signal Processing 38 (2019) 2877–2889.

[73] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, A. Reyes-Lecuona,

3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation, PLOS ONE 14 (2019) e0211899. URL: http://dx.plos.org/10.1371/journal.pone.0211899. doi:10.1371/journal.pone.0211899.

[74] T. Fischer, M. Caversaccio, W. Wimmer, System for combined hearing and balance tests of a person with moving sound source devices, 2020. URL: https://patents.google.com/patent/WO2020254462A1, WO Patent WO2020254462A1.

[75] I. Kiselev, E. Ceolini, D. Wong, A. De Cheveigne, S. Liu, Whisper: Wirelessly synchronized distributed audio sensor platform, in: 2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops), 2017, pp. 35–43. doi:10.1109/LCN.Workshops.2017.62.

[76] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (2021) 43–76. doi:10.1109/JPROC.2020.3004555.

[77] D. Marquardt, V. Hohmann, S. Doclo, Interaural coherence preservation in multi-channel wiener filtering-based noise reduction for binaural hearing aids, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2015) 2162–2176. doi:10.1109/TASLP.2015.2471096.

[78] D. Marquardt, S. Doclo, Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018) 1261–1274. doi:10.1109/TASLP.2018.2823081.

[79] N. Gößling, D. Marquardt, S. Doclo, Perceptual evaluation of binaural mvdr-based algorithms to preserve the interaural coherence of diffuse noise fields, Trends in Hearing 24 (2020) 2331216520919573.

UNIVERSITÄT BERN
ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH
INSELSPITAL

ARTORG Center for Biomedical Engineering Research, Murtenstrasse 50, CH-3008 Bern

Barbara Canlon, PhD
Editor-in-Chief
Hearing Research

Bern, February 04th 2021

Author statement:

Tim Fischer: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Investigation, Data Curation, Visualization

Marco Caversaccio: Resources, Supervision

Wilhelm Wimmer: Conceptualization , Validation, Writing - Review & Editing, Supervision, Methodology, Project administration

Kind regards,

Dr. Wilhelm Wimmer

Corresponding author

Tel: +41 (0)31 632 8789
Fax: +41 (0)31 632 7624
tim.fischer@artorg.unibe.ch
www.artorg.unibe.ch

b
UNIVERSITÄT
BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

INSELSPITAL

ARTORG Center for Biomedical Engineering Research, Murtenstrasse 50, CH-3008 Bern

Barbara Canlon, PhD
Editor-in-Chief
Hearing Research

Bern, February 04th 2021

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Kind regards,

Dr. Wilhelm Wimmer

Tel: +41 (0)31 632 8789
Fax: +41 (0)31 632 7624
tim.fischer@artorg.unibe.ch
www.artorg.unibe.ch