

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>



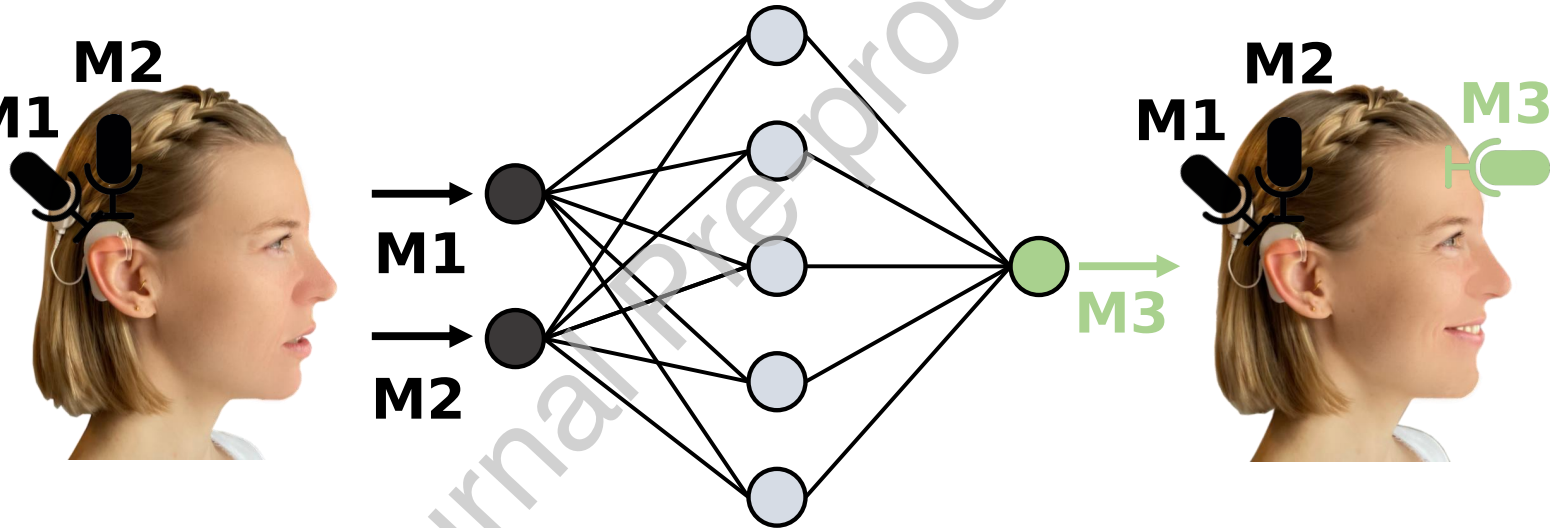
This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

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.

Journal Pre-proof



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

4 Tim Fischer^{a,b}, Marco Caversaccio^{a,b}, Wilhelm Wimmer^{a,b,c}

5 ^a*Hearing Research Laboratory, ARTORG Center for Biomedical Engineering Research,*
6 *University of Bern, Bern 3008, Switzerland*

7 ^b*Department of ENT, Head and Neck Surgery, Inselspital, Bern University Hospital,*
8 *University of Bern, Bern 3008, Switzerland*

9 ^c*corresponding author: Wilhelm Wimmer (wilhelm.wimmer@artorg.unibe.ch)*

10 **Abstract**

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

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

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

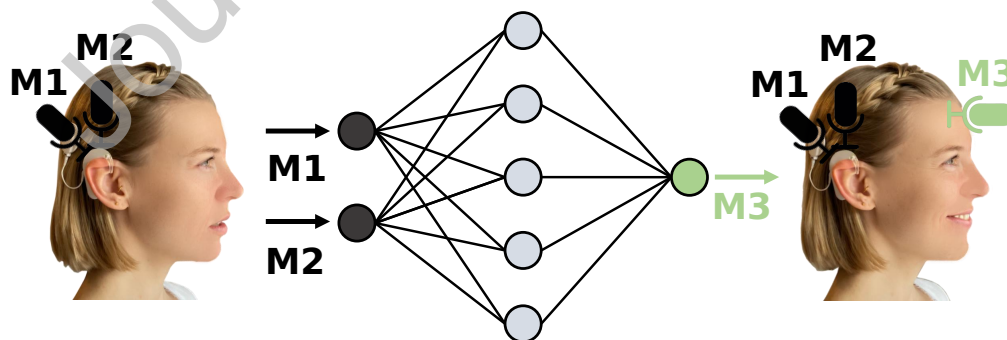
26 However, this microphone placement is impractical for hearing aid or
27 implant users, as it includes microphone positions such as on the forehead. To
28 overcome microphones' placement at impractical positions, we propose a deep
29 virtual sensing estimation of the corresponding audio signals. The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

36 Hearing aid and cochlear implant (CI) users might benefit from the pre-
37 sented approach using virtually sensed microphone signals, especially in noisy
38 environments.

39 *Keywords:* artificial intelligence, selective hearing, neural network,
40 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

Journal Pre-proof

41 1. Introduction

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

50 It is well known that the signal quality of beamforming algorithms in-
51 creases with the number of available input microphones and their position-
52 ing with respect to the target source [8, 9, 10, 11, 12, 13]. Using numerical
53 experiments, Feng et al. [8] showed that the microphone positions play an
54 essential role in the overall performance of beamforming algorithms. Jones
55 et al. [14] further showed for CI users that the microphone position at the
56 ear canal versus behind the ear led to more detailed interaural level differ-
57 ence (ILD) information due to the frequency transformations of the pinna
58 [15, 16]. In the specific case of unilateral CI users, it was demonstrated that
59 an additional microphone positioned at the contralateral ear led to increased
60 speech understanding in noise [17, 13, 18].

61 Since many conversations are held face to face [19], it is reasonable to as-
62 sume that additional microphones in positions other than the contralateral
63 ear canal, e.g., on the forehead, may further improve speech understanding.
64 However, the additional placement of microphones on the head is impractical
65 from the perspective of a hearing aid or CI user. One way of circumventing

66 this limitation may be to place the microphones virtually rather than phys-
67 ically. The results of several virtual microphone sensing approaches suggest
68 that estimating an additional microphone signal using information from the
69 available microphones may improve the speech quality in a cocktail party
70 scenario [20, 21, 22]. The microphone array used to record the reference sig-
71 nals was similar in the studies and consisted of 2 microphones positioned in a
72 straight line at a distance of 4 cm [20, 21] or 3 cm [22] from each other. To gen-
73 erate virtual microphone signals, the phase was linearly interpolated [20, 21]
74 or extrapolated [22] using measurements of the real microphone signals. In
75 Denk et al. [23], functions transformed the sound pressure at a microphone
76 positioned on a hearing aid to the pressure measured at the open eardrum.
77 The basis for the determination of these functions were the relative transfer
78 functions (RTFs) between the microphones, which in turn were determined
79 by head related transfer functions (HRTFs) measurements using frequency
80 sweeps in an anechoic chamber. Also using frequency sweeps, Corey et al.
81 [24] measured and evaluated impulse responses of 160 microphones spread
82 across the body and affixed to wearable accessories. Their results suggest
83 that microphone arrangements with large spatial distance spread across the
84 body provided the best signal-to-noise ratio (SNR) values. Unlike micro-
85 phones positioned on the head, the geometric arrangement of microphones
86 placed on clothing may change according to posture. Likely, the quality of
87 a beamforming algorithm defined for a specific microphone geometry suffers
88 from the continually changing microphone geometries in everyday life [25].

89 The tremendous progress in the field of machine learning leads to the
90 expectation that in the future, the RTFs between microphones can be de-

91 terminated purely data-driven, i.e., without prior knowledge of the specific
92 measurement setup. As a result, beamforming algorithms could be tuned
93 to individual array geometries by simply providing sufficient reference data
94 from the wearer without the need for anechoic chambers or knowledge of
95 the sound sources' positions. In the Mic2Mic publication [26] it was demon-
96 strated that even with unlabeled and unpaired data, audio signals between
97 different microphone domains could be translated. Based on the results, an
98 additional virtual microphone at the head of a hearing aid or CI user gen-
99 erated or learned solely by data-driven rules seems like a realistic scenario.
100 However, regardless of whether the microphones are placed virtually or phys-
101 ically on a subject's head, little is known about how their positioning affects
102 beamforming.

103 To continue the discussion, the first objective of this work was to system-
104 atically investigate the speech signal quality in complex acoustic scenarios
105 with varying head-mounted microphone arrangements and a minimum vari-
106 ance distortionless response (MVDR) beamformer as introduced by Souden
107 et al. [10]. Based on these measurements, virtual microphone signals at spe-
108 cific positions were estimated using a deep neural network. Finally, subjec-
109 tive listening tests were conducted to investigate to what extent the virtually
110 sensed microphone signals could improve the speech signal quality.

111 2. Methods

112 2.1. Linear observation model

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

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

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

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

126 2.2. MVDR beamforming

127 The MVDR beamformer minimizes the power of the beamformed signal
 128 while preserving the target signal, under the constraint of no distortion in the
 129 target signal [10]. The MVDR is a filter-and-sum beamformer and as such
 130 it applies different phase weights $h_i(f)$ to the i input microphone channels
 131 in order to steer the main lobe of the directivity pattern to the direction of

132 the target signal. The phase weights, or filters, are obtained in the frequency
 133 domain using [29]:

$$\mathbf{h}_{ref}(f) = [h_{1,ref}(f), \dots, h_{M,ref}(f)]^T = \frac{1}{\lambda(f)} (\mathbf{G}(f) - \mathbf{I}_{M \times M}) \mathbf{e}_{ref} \quad (1)$$

134 Where \mathbf{I} is the identity matrix and $\mathbf{G}(f)$ can be obtained by $\mathbf{G}(f) =$
 135 $\Phi_{noise}^{-1}(f) \Phi_{obs}(f)$ with $\lambda(f) = \text{trace}(\mathbf{G}(f)) - M$ [30, 10]. The spatial covari-
 136 ance matrices Φ can be computed by using time-frequency masks [29, 31, 32,
 137 33]. However, in this work we focus on the impact of additional microphone
 138 channels on the MVDR beamformers performance and extract $\Phi_{noise}^{-1}(f)$,
 139 $\Phi_{obs}(f)$ and $\Phi_{target}(f)$ from the noise, observation and target recordings.

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

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

Thus, the reference channel or microphone depends on $\mathbf{h}_r(f)$, which is
 the M -dimensional filter response (see Eq. 1) at the discrete frequency in-
 dex $f = 0, \dots, F - 1$, when \mathbf{e}_{ref} is set to \mathbf{e}_r . After the filters $\mathbf{h}_{ref}(f)$ 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 $y_i(t)$:

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

143 For the MVDR beamformer, the input signals were down-sampled to
 144 8 kHz and a Blackman window was applied [34]. Subsequently, an STFT
 145 (size = 256 and shift = 128) was performed. To reconstruct the signal, an

146 inverse short-time Fourier transform (ISTFFT) with the overlapadd strategy
 147 was applied. The herein used MVDR beamformer to evaluate the benefits
 148 of virtual microphone signals is just one application scenario. Theoretically,
 149 any multi-channel speech-enhancement algorithm could have been used to
 150 assess the benefits of virtually sensed microphone signals.

151 2.3. Data

152 The Bern cocktail party (BCP) dataset is tailored to this work, as it
 153 contains multi-microphone recordings of hearing aid or CI users in cocktail
 154 party scenarios [35]. For the recordings, 12 loudspeakers (Control 1 Pro,
 155 JBL, Northridge, USA) were aligned horizontally in a circle at the height
 156 of the ears (1.2 m) in an acoustic chamber [36, 37, 13]. For this work, we
 157 used the acoustic scenarios captured with 16 microphones (ICS-40619, TDK,
 158 Tokyo, Japan) attached to a head and torso simulator (Brel & Kjør, Type
 159 4128, Nærum, Denmark) (see Figures 1 and 2).

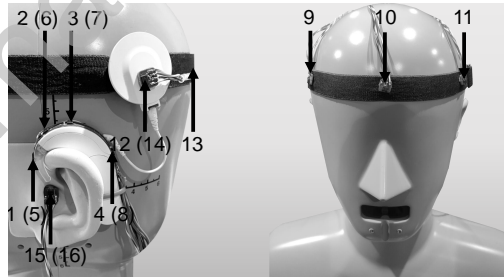


Figure 1: Placement of the 16 microphones used for cocktail party scenario recordings. The IDs refer to the microphone signals assignment in the multi-channel recording audio files [35]. Numbers in brackets refer to the contralateral (here: right side) assignment of the microphones. The sagittal plane 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. |

160 2.3.1. Test dataset

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

173 2.3.2. Training dataset

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

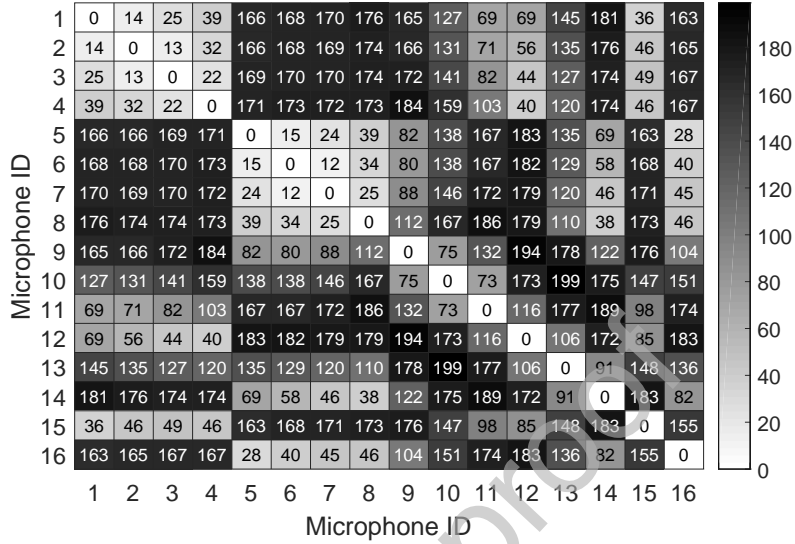


Figure 2: Euclidean distances in millimeters between the microphones for the head and torso simulator measurements [35].

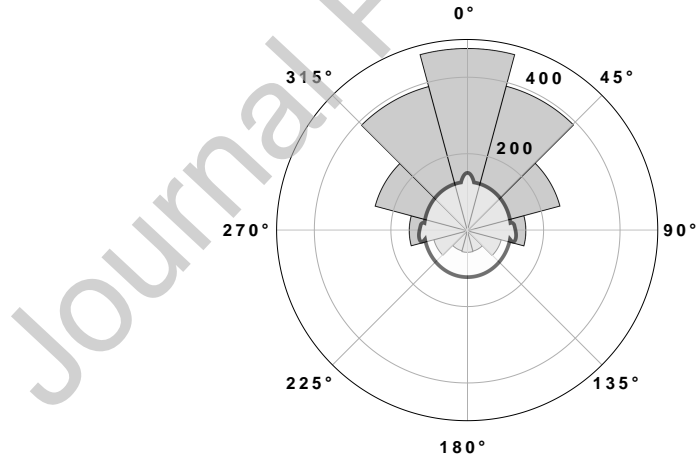


Figure 3: Circular histogram of the frequency of occurrence of spatial source directions in relation to the head and torso simulator azimuth. The audio files were selected such that the directional distribution assumes a von-Mises distribution with $\mu = 0.0$ and $\kappa = 1.1$ [35].

180 *2.4. Evaluation of microphone channel configurations*

181 Various microphone channel configurations were evaluated by adding or
 182 omitting microphone channels with respect to a reference microphone chan-
 183 nel configuration, as explained in detail later (Section 3, Tables 3-6). The
 184 results were computed by providing the MVDR beamformer [10] with the
 185 target and noise spatial covariance matrices Φ of the audio mixtures from
 186 the corresponding microphone configurations.

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

192 To cover realistic use cases regarding the benefits of different microphone
 193 configurations, the results were divided into 4 categories rather than pre-
 194 senting all possible microphone channel combinations: subsets of unilateral
 195 CI microphone configurations (see Table 3), unilateral CI microphone con-
 196 figurations with additional ipsilateral microphones (Table 4), unilateral CI
 197 microphone configurations with additional contralateral microphones (Table
 198 5), symmetric bilateral CI configurations with additional microphones (Table
 199 6). An overview of all measured microphone configurations can be found in
 200 Table 2.

201 For the evaluation of the microphone configurations (i.e., real recordings
 202 and virtually sensed microphone channels), the following objective speech
 203 quality metrics were assessed: perceptual evaluation of speech quality (PESQ)
 204 [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} | |

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

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

221 *2.5. Virtual sensing of a microphone channel*

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

227 Most applications of deep neural networks in the domain of audio signal
228 processing address the enhancement of speech signals by separating a target
229 source (speech) from a mixture of interfering noise sources [43]. In the work

230 presented here, however, no source separation was performed, but rather, in
 231 a transferred sense, a denoising of the reference signal, as explained in the
 232 following: Let the audio signal captured from a microphone inside the ear
 233 canal of the left ear be the reference signal and the audio signal inside the
 234 ear canal of the right ear the target signal. By trying to match the signal of
 235 the left ear to the right ear or denoise the left ear, we hypothesize that the
 236 network implicitly learns the RTF between the two microphone signals or, in
 237 other words, the "noise" to remove from the audio signal of the left ear. As
 238 a result, the network tries to virtually sense the right ear's audio input by
 239 using the signal of the left ear. To evaluate the quality of the virtually sensed
 240 microphones, spatial covariance matrices Φ with and without virtually sensed
 241 microphone signals were provided as input for the MVDR beamformer [10].
 242 The results were compared with the same metrics and statistics as with the
 243 real microphones measurements (see Section 2.4).

244 In this study, two microphone signals were used as reference signals, and
 245 three additional microphone signals were virtually sensed. The 2 reference
 246 signals consisted of the microphones $\{2, 3\}$ and were chosen because their
 247 spatial arrangement corresponds to that of a conventional CI audio proces-
 248 sor (see Figure 1 or Table 1). Motivated by the results of the head-mounted
 249 microphone measurements, the microphone on the forehead ($\{10\}$), the back
 250 ($\{13\}$) and inside the ear canal of the contralateral ear ($\{16\}$) were chosen
 251 as target signals for the virtual sensing approach. In the remainder of the
 252 manuscript, virtual channels are indicated by the subscript v . The resulting
 253 microphone configuration ($\{2, 3, 10_v, 13_v, 16_v\}$) provided the advantages
 254 as explained in the Discussion (Section 4.1): a high spatial spread of the

255 microphone signals [44], proximity to the target signal, and frequency trans-
 256 formations by the pinna and head shadow [15].

257 *2.5.1. Deep neural network architecture for the virtual sensing approach*

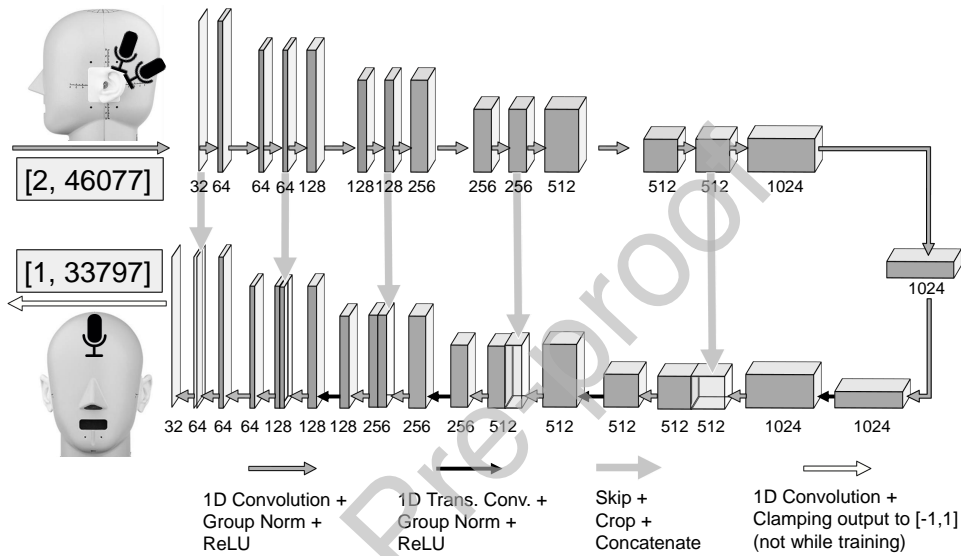


Figure 4: The proposed deep neural network architecture for the virtual sensing of additional microphone channels based on the work of Stoller et al. [42]. The numbers below the blocks describe the input channel size of the following convolution. Shown is an example for the estimation of the microphone 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.

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

262 [45, 46]. The well known U-Net structure is composed as a convolutional
263 autoencoder, and as such, consists of an encoder (contracting path), a bot-
264 tleneck, and a decoder (expanding path) [47]. A diagram of our network’s
265 architecture implementation is shown in Figure 4.

266 In the encoder, an increasing number of higher-level features on coarser
267 time scales were calculated, allowing the modeling of long-term dependen-
268 cies in the audio signal. Our implementation of the encoder consisted of
269 5 levels, with each level working on half the time resolution and twice the
270 number of feature maps as the previous one. In the bottleneck, the model
271 was forced to learn a compression of the input data, containing only the
272 relevant information (latent space) to construct the virtual microphone sig-
273 nal. The latent-space representation of the bottleneck layer was passed to
274 the decoder, which tried to learn a mapping of the input data to match the
275 desired virtual microphone signal. The decoder was the mirror image of the
276 encoder and also consisted of 5 levels. Each level worked on double the time
277 resolution and half the number of feature maps as the previous level. Based
278 on the results of initial tests, transposed convolutions were used for the up-
279 sampling process. Each convolution was followed by group normalization,
280 and a rectified linear unit (ReLU) activation function [48, 19]. By introduc-
281 ing the skip connections in the encoder-decoder architecture, the encoder’s
282 high-level features were concatenated with the local features computed dur-
283 ing the upsampling block of the decoding. The result of this concatenation
284 were multi-scale features that were fed in the output layer of the network
285 [47, 42]. The output of the last convolutional layer was the estimation of the
286 virtually sensed microphone signal.

287 The receptive field of the model was chosen to work with 2.1 s (46077
 288 samples), which provided an output vector with the desired test size of 1.5 s
 289 (33797 samples).

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

296 *2.5.2. Network training*

297 To train the deep virtual sensing network, we extracted measurement
 298 data from the two reference channels ($\{2, 3\}$) and the microphone channel to
 299 be estimated. Due to the large size of the BCP training dataset (see Section
 300 2.3.2), no data augmentation was necessary. In accordance with the original
 301 Wave-U-Net implementation [42], the audio data of the BCP dataset [35]
 302 was downsampled to 22.05 kHz. For evaluating the network’s performance,
 303 the absolute differences between the actual value and the predicted value (L_1
 304 loss) were used. To update the network weights iteratively based on training
 305 data, we applied the ADAM optimizer [49] with the default decay rates of
 306 $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 16 [42]. Instead of monotonically
 307 decreasing the learning rate, cyclical learning rates [50] were used with upper
 308 and lower boundaries of 0.0002 and 0.00001, respectively. Early stopping was
 309 performed after 10 epochs with only minimal improvement on the validation
 310 loss. Afterward, the best model was fine-tuned with lower learning rate limits
 311 (0.000001 to 0.00001) and a batch size of 8, again until 10 epochs without

312 improvement on the validation loss. The fine-tuned network was further used
313 to predict the virtual channels. The test dataset to evaluate the virtually
314 sensed microphone channels consisted of 2400 samples, which included the
315 audio files described in Section 2.3.1. Care was taken to ensure that none of
316 the test samples were used to validate or train the network.

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

322 *2.5.3. Subjective listening tests*

323 Twenty normal hearing participants (6 female, 14 male, mean age in years
324 = 29.8, SD = 3.6) performed a subjective listening test to evaluate the benefit
325 of the virtually sensed microphone signals on the speech quality. The test
326 was performed in a quiet environment, and stimuli were presented via high
327 definition insert earphones (Triple Driver, 1 More Inc. San Diego, CA) at
328 the most comfortable loudness levels as selected by the subjects.

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

336 To answer these questions, the participants were asked to listen to 3 audio

337 mixtures, all based on the same recording but either

- 338 • Beamformed based on the reference channels with additional virtual
339 channels ($\{2, 3\} + \{10_v\}, \{13_v\}, \{16_v\}$)
- 340 • Beamformed based on the reference channels only ($\{2, 3\}$)
- 341 • The non-beamformed recording of the channels $\{2, 3\}$

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

351 During the test and the trial runs, the participants were allowed to hear
352 the 4 audio files (1 target signal and 3 audio mixtures) as many times as de-
353 sired. The test stimuli consisted of 60 audio mixture quartets of 1.5 seconds
354 length per file, ensuring that each file contained the utterance of at least one
355 word. All audio mixtures were taken from the pool of the 2400 test files
356 described in Section 2.3 with distribution proportions as shown in Figure 3.
357 Evaluation of the presented audio files took about 20 minutes; no feedback
358 was given during or after the test. After evaluating 30 of the 60 audio files, a
359 pause of 3 minutes was taken during which the GUI was disabled. To mini-
360 mize order bias, the 2 stimuli blocks that were evaluated before and after the

361 pause were counter-balanced within the participants. The subjective listen-
362 ing evaluation was designed in accordance with the Declaration of Helsinki,
363 written informed consent was obtained from all participants.

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

372 3. Results

373 3.1. MVDR beamforming with unilateral channel configurations

374 Table 3 shows the PESQ, STOI and SI-SDR performances of unilateral
 375 single microphone configurations compared to the performance with the ref-
 376 erence configuration, i.e. a CI audio processor equipped with 4 microphones
 377 placed on top of the housing. For the PESQ and SI-SDR metric, the per-
 378 formances with single microphones were significantly worse than with the
 379 4-channel reference configuration (all $p = 0.001$). The same was observed
 380 for STOI ($p = 0.001$) except for the microphones $\{1, 4\}$ and $\{2, 4\}$ (both
 381 $p = 0.9$). In all 3 metrics, the microphones that were facing the front (front
 382 $\{10\}$, left temple $\{11\}$, forward facing (audio processor) $\{1\}$, see Figure 1
 383 or Table 1) achieved the best results, whereas the performance differences
 384 between channels $\{10\}$ and $\{11\}$ were not statistically significant in terms of
 385 PESQ and SI-SDR ($p = 0.608$, $p = 0.9$) but for STOI ($p = 0.001$). Between
 386 the microphones $\{1\}$ and $\{2\}$ the metrics PESQ, STOI and SI-SDR did not
 387 differ significantly ($p = 0.408$, $p = 0.9$, $p = 0.115$) (a significance-matrix
 388 showing the results of the post-hoc Nemenyi tests for Table 3 can be found
 389 in the Appendix (Figures A.1-A.3)).

390 When the same 4-channel reference configuration (microphones $\{1, 2, 3,$
 391 $4\}$) was extended by the aforementioned ipsilateral single microphone signals,
 392 again the front-facing microphones $\{10\}$ and $\{11\}$ (see Figure 3) provided the
 393 greatest benefit (see Table 4). The performance differences for all metrics
 394 when channel $\{10\}$ (front) was added did not differ significantly from the
 395 performance differences when channel $\{11\}$ (left temple) was added to the
 396 reference configuration (PESQ: $p = 0.792$, STOI: $p = 0.736$, SI-SDR: $p = 0.9$)

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

399 Since many CI audio processors record signals from 2 microphones posi-
 400 tioned on top of the housing, the performance of different spatial arrange-
 401 ments of 2 microphones placed on the audio processor compared to the 4-
 402 channel reference configuration (microphones {1, 2, 3, 4}) was investigated
 403 and is shown in Table 3. The arrangement with the largest spatial distance
 404 between the 2 microphones, namely the microphones on top of the audio pro-
 405 cessors facing the front and back ({1, 4}), achieved the best performance (see
 406 Figure 2 for a microphone distance matrix). The statistical analysis showed
 407 that the performance differences of the microphones {1, 4} did not differ sig-
 408 nificantly for PESQ and STOI from the results compared to the microphones
 409 on the audio processor facing the top and the back ({2, 4}) to the reference
 410 configuration ($p = 0.668$, $p = 0.9$). Both 2 channel microphone configura-
 411 tions did not differ significantly from the 4 channel reference configuration in
 412 terms of STOI (both $p = 0.9$). For the SI-SDR metric, the differences when
 413 adding {1, 4} did not differ statistically significantly from any of the tested
 414 2 channel configurations (all $p = 0.9$).

415 The arrangement with the smallest inter-microphone distance (micro-
 416 phones {2, 3}, see Figures 1 and 2), which is related to the conventional
 417 microphone positions of CI audio processors, achieved the lowest scores in
 418 2 (STOI and SI-SDR) of the 3 evaluated objective metrics, even though for
 419 SI-SDR the differences of this configuration did not differ significantly from
 420 any of the tested 2 channel configurations (all $p = 0.9$). For the metrics PESQ
 421 and STOI no significant differences in the performances between the micro-

422 phones $\{2, 3\}$, $\{1, 2\}$ or $\{1, 3\}$ were observed (PESQ: $p = 0.721$, $p = 0.601$,
423 STOI: $p = 0.884$, $p = 0.134$). Table 4 shows the impact on the PESQ, STOI
424 and SI-SDR metrics when additional ipsilateral, including those on the sagittal
425 plane, microphones were added to the the conventional microphone arrangement
426 ($\{2, 3\}$). The extension of the microphone arrangement ($\{2, 3\}$)
427 with forward facing microphones (front $\{10\}$ or left temple $\{11\}$) provided
428 the greatest benefit. For none of the 3 tested metrics did the performance
429 between adding the front ($\{10\}$) or left temple ($\{11\}$) microphone to the
430 conventional microphone arrangement differ significantly (PESQ: $p = 0.067$,
431 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 "†".

| Microphone IDs | Metric | | |
|---------------------------|--------------|----------------|--------------|
| | PESQ | STOI | SI-SDR |
| <i>Ref.: {1, 2, 3, 4}</i> | 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 "†".

| Microphone IDs | Metric | | |
|---------------------------|-------------|--------------------|--------------------|
| | PESQ | STOI | SI-SDR |
| <i>Ref.: {1, 2, 3, 4}</i> | <i>1.77</i> | <i>0.48</i> | <i>-29.07</i> |
| 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 [†] |
| <i>Ref.: {2, 3}</i> | <i>1.61</i> | <i>0.45</i> | <i>-30.38</i> |
| 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 |

432 *3.2. MVDR beamforming with bilateral channel configurations*

433 Table 5 shows the PESQ, STOI and SI-SDR performances when addi-
 434 tional bilateral microphones were added to the input signals of an unilateral
 435 CI audio processor equipped with 4 microphones placed on top of the housing
 436 (microphones {1, 2, 3, 4}, see Figure 1 or Table 1). When a single contralat-
 437 eral microphone was added, it was not the microphone closest to the target
 438 source (microphone {9}, temple) that provided the greatest benefit in terms
 439 of the human perception-related objective metrics PESQ and STOI, but the
 440 contralateral ear canal microphone {16}. Compared to adding channels {9}
 441 or {14} (temple or contralateral CI transmission coil), the improvement of
 442 the PESQ and STOI metrics were significantly better when adding the con-
 443 tralateral ear-canal microphone (all $p = 0.001$) (a significance-matrix show-
 444 ing all results of the post-hoc Nemenyi test for Table 5 can be found in the
 445 Appendix (Figures A.10-A.15)). However, in terms of SI-SDR, the input
 446 from the microphone on the contralateral CI transmission coil (microphone
 447 {14}) achieved the best SI-SDR values and even outperformed the micro-
 448 phone configuration compared to an additional contralateral 4-channel CI
 449 audio processor. All differences in SI-SDR with the contralateral transmis-
 450 sion coil microphone ({14}) compared to {9} (contralateral temple), {16}
 451 (contralateral ear canal) and Ref. ch. + {5, 6, 7, 8} were not statistically
 452 significant ($p = 0.362$, $p = 0.802$, $p = 0.409$). Since the cable connection
 453 between the CI transmission-coil and the audio processor could theoretically
 454 be exploited to transmit audio signals, a unilateral microphone configuration
 455 was also used as a reference, which included the coil signal ({12}) in addi-
 456 tion to the 4 microphones on the audio processors. The results showed in

457 Table 5 did differ only marginally and non significantly between the refer-
458 ence configuration with the CI transmission coil ($\{1, 2, 3, 4, 12\}$) and the
459 reference configuration without the CI transmission coil microphone ($\{1, 2,$
460 $3, 4\}$). The small benefit of adding microphone $\{12\}$ to the reference channel
461 configuration is also indicated by the results of Table 4.

462 An analysis of the results with a reference microphone configuration based
463 on the conventional spatial microphone arrangement in CI audio processors
464 (microphones $\{2, 3\}$, see Figure 1 or Table 1), lead to similar conclusions
465 as with the 4-channel microphone configuration described above (see Table
466 5). Again, the overall result of a single additional microphone positioned
467 at the contralateral ear-canal $\{16\}$ was best, but only with respect to STOI
468 and PESQ. For the PESQ metric, the performance with an additional mi-
469 cophone positioned in the contralateral ear canal differed non-significantly
470 compared to the performance with an additional microphone on the temple
471 ($\{9\}$) ($p = 0.763$). In terms of SI-SDR, the microphones on the contralateral
472 side which were close to the sagittal plane (temple $\{9\}$ and transmission coil
473 $\{14\}$) outperformed the contralateral ear-canal microphone $\{16\}$ when added
474 to the microphone configuration $\{2, 3\}$ ($p = 0.006$, $p = 0.9$). An additional,
475 identical, bilaterally connected processor with 2 microphones ($\{6, 7\}$) yielded
476 significantly better values in all metrics than adding the single microphones
477 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.

| Microphone IDs | Metric | | |
|---------------------------|-------------|-------------|---------------|
| | PESQ | STOI | SI-SDR |
| <i>Ref.: {1, 2, 3, 4}</i> | <i>1.77</i> | <i>0.48</i> | <i>-29.07</i> |
| 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 |
| <i>Ref.: {2, 3}</i> | <i>1.61</i> | <i>0.45</i> | <i>-30.38</i> |
| 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 |

478 When a bilateral CI processor microphone configuration was taken as a
 479 reference (microphones $\{1, 2, 3, 4, 5, 6, 7, 8\}$, see Table 6), adding a micro-
 480 phone to the front ($\{10\}$) provided more benefit than adding a microphone
 481 facing the back ($\{13\}$) (PESQ and STOI: $p = 0.001$), but for SI-SDR not
 482 statistically significant ($p = 0.515$) (a significance-matrix showing all results
 483 of the post-hoc Nemenyi test for Table 6 can be found in the Appendix (Fig-
 484 ures A.16-A.21)). The single front microphone achieved even similar and
 485 statistically not significantly differing STOI and SI-SDR values compared
 486 to the performance when adding 2 microphones at the left and right tem-
 487 ple ($\{9,11\}$) (both metrics $p = 0.9$). For PESQ however, the performance
 488 with the additional 2 temple microphones ($\{9,11\}$) differed statistically sig-
 489 nificant compared to the additional microphone facing to the front ($\{10\}$)
 490 ($p = 0.001$). Adding the signals of the two in-ear microphones ($\{15, 16\}$) to
 491 the bilateral CI processor microphone configuration (microphones $\{1, 2, 3, 4,$
 492 $5, 6, 7, 8\}$) did not provide any benefit, not even if only 2 bilateral ($\{2, 3, 6,$
 493 $7\}$) instead of 4 ($\{1, 2, 3, 4, 5, 6, 7, 8\}$) bilateral processor microphones were
 494 used as a reference microphone configuration. The full 16-channel micro-
 495 phone configuration achieved the statistically significant best PESQ scores
 496 (all $p = 0.001$). However, in terms of STOI and SI-SDR the performance did
 497 barely, and for SI-SDR non significantly, differ compared to the 8-channel ref-
 498 erence microphone configuration. Again, as with the unilateral 4-microphone
 499 CI audio processor configuration, adding the transmission-coil microphone
 500 signals ($\{12, 14\}$) to the bilateral microphone configurations ($\{1, 2, 3, 4, 5,$
 501 $6, 7, 8\}$ or $\{2, 3, 6, 7\}$) did barely and statistically not significant influence
 502 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 "†".

| Microphone IDs | Metric | | |
|--|-------------|-------------|---------------|
| | PESQ | STOI | SI-SDR |
| <i>Ref.: {1, 2, 3, 4, 5, 6, 7, 8}</i> | <i>2.07</i> | <i>0.54</i> | <i>-28.46</i> |
| 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† |
| <i>Ref.: {2, 3, 6, 7}</i> | <i>1.87</i> | <i>0.51</i> | <i>-28.94</i> |
| 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† |

503 *3.3. Virtual sensing of microphone channels*

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

513 The mean benefit in performance when additional virtual/measured mi-
 514 crophone signals were used for beamforming was 0.24/0.34 units for PESQ,
 515 0.06/0.07 units for STOI, and 1.17/1.25 dB for SI-SDR. For the PESQ and
 516 STOI metrics, the performance between the virtually sensed microphone sig-
 517 nals and the measured microphones signals differed significantly ($p < 0.001$).
 518 In terms of SI-SDR, no significant difference between the two configurations
 519 were observed ($p = 0.998$).

520 An analysis of the performance of the neural networks with respect to each
 521 of the estimated channels $\{16\}$, $\{13\}$ and $\{10\}$ showed that the mean benefit
 522 when an additional virtual/measured microphone signal was used for beam-
 523 forming was 0.154/0.211, 0.114/0.149, 0.178/0.219 for PESQ, 0.049/0.052,
 524 0.028/0.032, 0.042/0.048 for STOI, and 1.000/1.057, 0.938/0.877, 1.493/1.377
 525 for SI-SDR. For the metrics PESQ and STOI the differences in performance
 526 between the additional virtually estimated microphone and the measured mi-
 527 crophone were significant (all $p < 0.001$). For SI-SDR, the differences were

528 significant only with respect to microphone channel {10} ($p = 0.027$), but not
 529 for the channels {13} and {16} ($p = 0.244$, $p = 0.309$). The on average bad
 530 results for channel {16}, meaning the largest difference between the benefit
 531 of additional virtual/measured microphone signals, and the best results for
 532 channel {10} were also reflected in the validation losses of the trained net-
 533 works. For channel {16}, {13} and {10}, the best L_1 -losses on the validation
 534 set were 2.1×10^{-4} , 1.5×10^{-4} and 1.4×10^{-4} , respectively.

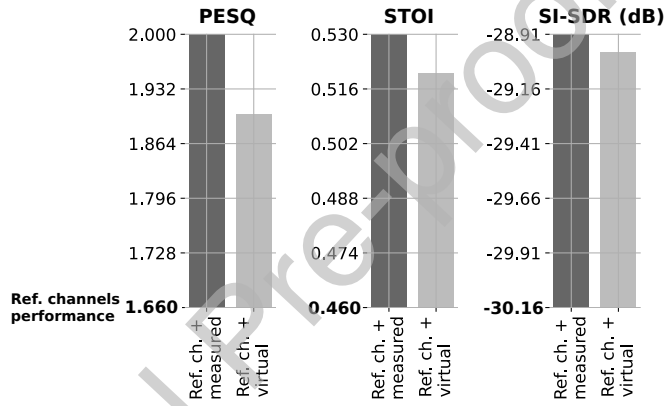


Figure 5: Comparison of overall perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) and scale-invariant speech to distortion ratio (SI-SDR) scores between 3 different microphone channel configurations used as input signals for the minimum variance distortionless 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.

535 *3.3.1. Subjective listening tests*

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

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

546 For all of the presented signal configurations, the distribution of the fre-
547 quency of choices differed significantly from the chance level of the test (all
548 $p < 0.001$).

549 To investigate if the subjects' choice of the signal most comfortable to
550 understand was dependent on the SNR of the original or raw audio mixture,
551 the SNRs of the corresponding raw audio mixtures were compared. It was
552 observed that the subjects preferred the beamformed signal with additional
553 virtual channels if the SNRs of the raw audio mixture were low (Mean=2.4,
554 SD=9.3) compared to the raw audio mixtures' SNRs when the beamformed
555 signal based on the reference channels only was selected (Mean=5.2, SD=8.0,
556 $p = 0.001$). The SNRs of the raw audio mixtures when the non-beamformed
557 signal was selected (Mean=2.1, SD=9.2) was not significantly different from
558 the SNRs of the raw audio mixtures when the beamformed signal with addi-
559 tional virtual channels was selected ($p = 0.987$). However, it was significantly

560 different from the SNRs of the raw audio mixtures when the beamformed sig-
 561 nal based on the reference channels was chosen ($p = 0.029$).

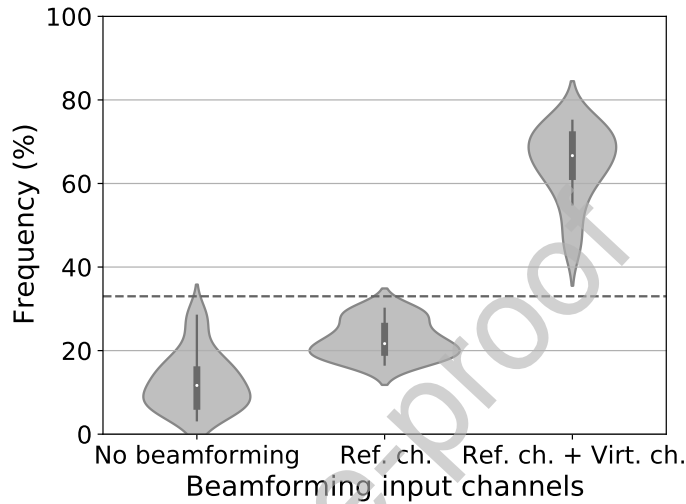


Figure 6: Violin plots [53] of the frequency of choices in the subjective listening test. The data represents the choices for the non-beamformed signal, the beamformed signal with the measured reference channel configuration as input channels (microphone IDs {2, 3}) and the beamformed signal with additional 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.

562 4. Discussion

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

569 4.1. Association of the speech quality and the microphone positioning

570 Our measurements of varying head-mounted microphone arrangements in
571 cocktail party scenarios confirmed that the performance of beamforming algo-
572 rithms and thus the speech quality improves with additional microphone sig-
573 nals [44]. Single-microphone speech-enhancement algorithms can only exploit
574 temporal and spectral information cues, whereas multi-microphone beam-
575 formers can additionally exploit the spatial information of the sound sources
576 [10, 44].

577 However, a high number of microphones alone does not necessarily lead
578 to a better speech quality [10]. In the case of bilaterally placed microphones
579 (Table 6), we observed saturation in terms of speech signal enhancement
580 with additional microphones that were placed close to the reference micro-
581 phones. In particular, the SI-SDR metric showed that noise from additional
582 microphone signals can dominate compared to the redundant information
583 in the audio signal used for speech enhancement. As also shown by Corey
584 et al. [24], the microphone arrangement's spatial diversity played a signifi-
585 cant role in the quality of the acoustic beamforming. The herein performed

586 measurements confirmed this finding since no improvements were observed
587 when additional microphones were placed at a distance of about 5 cm to the
588 reference microphones. It was assumed that even for low frequencies, these
589 microphones were too closely spaced to provide inter-microphone information
590 for the beamforming algorithm [24]. Besides, the microphones' distance was
591 too small for an effect of the acoustic head shadow [15]. With the same rea-
592 soning, the slightly worse result of the unilateral, conventional microphone
593 configuration ($\{2, 3\}$) and the good result of the arrangement with the largest
594 inter-microphone distance (front and back facing $\{1, 4\}$) compared to other
595 2-channel microphone arrangements on the audio processor can be argued.

596 Although adding a microphone with a high Euclidean distance to the ref-
597 erence microphone configuration is a good rule of thumb to improve acoustic
598 beamforming, other microphone positioning factors, such as exploiting the
599 acoustic head shadow [15], may be just as important. In the unilateral con-
600 figuration (see Table 4), we observed that the proximity to the most likely
601 target source with an additional microphone on the temple ($\{2, 3\} + \{11\}$)
602 was more important than the spatial diversity of the microphones with an
603 additional microphone placed on the back of the head ($\{2, 3\} + \{13\}$). In
604 addition to the proximity to the target signal and the microphone distance,
605 our measurements confirmed that the pinna's directional frequency trans-
606 formation provided relevant information for improving the quality of the
607 beamforming algorithm [15, 54, 16]. We observed that the most useful ad-
608 ditional contralateral microphone was neither the one closest to the target
609 signal ($\{11\}$, temple) nor the one with the highest Euclidean distance to the
610 reference microphone configuration ($\{14\}$, CI transmission coil). It was the

611 contralateral microphone placed in the ear canal facing away from the target
612 signal ($\{16\}$).

613 *4.2. Virtual sensing of head-mounted microphone signals*

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

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

634 In the measurements with the reference microphone configuration accord-
635 ing to conventional CI audio processors ($\{2, 3\}$), we observed that an addi-

636 tional microphone on the forehead produced similar improvements in speech
637 quality as an additional microphone placed at the entry of the contralateral
638 ear canal. However, due to the poor estimation of the contralateral ear sig-
639 nal by the neural network, a higher benefit was obtained with the virtual
640 microphone channel estimating the signal at the forehead. Therefore the
641 estimation of optimal microphone positions for neural network-based beam-
642 forming approaches requires further investigation.

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

652 *4.3. Limitations and outlook*

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

656 Due to the input data size of 2 seconds, the delay of the proposed net-
657 work architecture is too long to be applicable in a real hearing aid application.
658 However, this paper's main objective was to demonstrate a proof of concept
659 for purely data-driven virtual channel estimations in hearing aids or CIs.
660 Tackling the problem of latency and neural network complexity in online

661 speech enhancement is ongoing research [57, 58, 59, 60] with promising re-
662 sults and input frame lengths as little as 2 ms [60]. Future research should
663 investigate whether the significant reduction in network time delay required
664 for an application in hearing devices affects the performance of the presented
665 approach. In addition to progress in reducing the computational costs, sub-
666 stantial progress is continuously being made in other areas of speech signal
667 enhancement with artificial neural networks relevant for the methodology of
668 this work, such as in blind source separation (BSS) [61, 62, 63], acoustic
669 scene classification (ASC) [64, 65, 66], domain shift [26, 67] and the usage of
670 loss functions to optimize the parameters of the network based on the human
671 perception of speech [68, 59]. The results of Drude et al. [63] indicated, that
672 the benefit of the presented approach when using estimated coherence matrix-
673 ces may be different from the benefit achieved with the oracle matrices. For
674 computational time reasons, no sophisticated optimization of the presented
675 network’s architecture was performed. Further research may investigate the
676 optimal number and size of hidden layers for the presented approach.

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

683 One of the biggest challenges of the presented methodology to be ap-
684 plicable in a real-world application will be to ensure the robustness of the
685 network’s predictions in acoustic environments with high reverberation [69,

686 70, 71, 72]. In the context of this work, the first step in this direction would be
687 the use of more challenging acoustic training data, for example, by simulating
688 conditions with higher reverberation [73] or the use of dynamically moving
689 sound sources [36, 74]. Another possibility would be to record acoustic sce-
690 narios using a portable microphone array [75]. In a real-world application,
691 this data could be collected as part of an audiological fitting routine. In
692 both cases, whether the data was simulated or recorded in real environments
693 for each subject, the additional recordings and the personalization of the
694 network through transfer learning would most likely increase the robustness
695 of applied neural network solutions [76]. To account for the different head
696 geometries and thus varying inter-microphone features, the information of
697 3D head scans as provided in Fischer et al. [35] could be fed into a neural
698 network architecture that allows metadata injection.

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

707 5. Conclusions

708 In this work, real and virtual microphone signals were combined as in-
709 put for an MVDR beamformer to investigate the effects on speech quality

710 for hearing aid or CI users in cocktail party scenarios. The measurements
711 with respect to the number and spatial arrangement of real microphones in-
712 dicated that, optimally, microphones should be placed as close as possible
713 to the target source, encode monaural cues, and produce a large distance
714 spread by their spatial arrangement. In reality, however, it is inconvenient
715 to place the microphones according to these criteria. To overcome this prob-
716 lem, virtual microphone signals were estimated using a deep neural network
717 without explicit knowledge of the spatial microphone arrangement. The re-
718 sults of 3-alternative forced choice subjective listening tests and objective
719 speech quality metrics suggest that hearing aid or CI users might benefit
720 from virtually sensed microphone signals, especially in challenging cocktail
721 party scenarios.

722 **Appendix A. Additional Figures**

723 Please see appendix A.pdf for significance-matrices of the post-hoc Ne-
724 meny tests concerning the data in Tables 3-6.

725 **Funding**

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

728 **References**

- 729 [1] E. C. Cherry, Some Experiments on the Recognition
730 of Speech, with One and with Two Ears, The Journal
731 of the Acoustical Society of America 25 (1953) 975–979.

732 URL: <http://asa.scitation.org/doi/10.1121/1.1907229>.
733 doi:10.1121/1.1907229.

734 [2] J. Peissig, B. Kollmeier, Directivity of binaural noise reduction in spatial
735 multiple noise-source arrangements for normal and impaired listeners,
736 *The Journal of the Acoustical Society of America* 101 (1997) 1660–1670.

737 [3] N. Mesgarani, E. F. Chang, Selective cortical representation of attended
738 speaker in multi-talker speech perception, *Nature* 485 (2012) 233–236.

739 [4] P. Smaragdis, Blind separation of convolved mixtures in the frequency
740 domain, *Neurocomputing* 22 (1998) 21–34.

741 [5] S. Makino, T.-W. Lee, H. Sawada, Blind speech separation, volume 615,
742 Springer, 2007.

743 [6] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, D. Yu, Past review, current
744 progress, and challenges ahead on the cocktail party problem, *Frontiers*
745 *of Information Technology & Electronic Engineering* 19 (2018) 40–63.

746 [7] B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to
747 spatial filtering, *IEEE assp magazine* 5 (1988) 4–24.

748 [8] Z. G. Feng, K. F. C. Yiu, S. E. Nordholm, Placement design of micro-
749 phone arrays in near-field broadband beamformers, *IEEE transactions*
750 *on signal processing* 60 (2011) 1195–1204.

751 [9] J. Wouters, J. V. Berghe, Speech recognition in noise for cochlear im-
752 plantees with a two-microphone monaural adaptive noise reduction sys-
753 tem, *Ear and hearing* 22 (2001) 420–430.

- 754 [10] M. Souden, J. Benesty, S. Affes, On optimal frequency-domain multi-
755 channel linear filtering for noise reduction, *IEEE Transactions on audio,*
756 *speech, and language processing* 18 (2009) 260–276.
- 757 [11] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, J. Dmochowski, New
758 insights into the mvdr beamformer in room acoustics, *IEEE Transac-*
759 *tions on Audio, Speech, and Language Processing* 18 (2009) 158–170.
- 760 [12] W. Wimmer, M. Caversaccio, M. Kompis, Speech intelligibility in noise
761 with a single-unit cochlear implant audio processor, *Otology & neuro-*
762 *tology* 36 (2015) 1197–1202.
- 763 [13] W. Wimmer, M. Kompis, C. Stieger, M. Caversaccio, S. Weder, Direc-
764 tional microphone contralateral routing of signals in cochlear implant
765 users: A within-subjects comparison, *Ear and hearing* 38 (2017) 368–
766 373.
- 767 [14] H. G. Jones, A. Kan, R. Y. Litovsky, The effect of microphone place-
768 ment on interaural level differences and sound localization across the
769 horizontal plane in bilateral cochlear implant users, *Ear and hearing* 37
770 (2016) e341.
- 771 [15] J. Blauert, *Spatial hearing : the psychophysics of human sound local-*
772 *ization*, MIT Press, 1997.
- 773 [16] T. Fischer, C. Schmid, M. Kompis, G. Mantokoudis, M. Caversaccio,
774 W. Wimmer, Pinna-imitating microphone directionality improves sound
775 localization and discrimination in bilateral cochlear implant users, *Ear*
776 *and hearing* 42 (2021) 214.

- 777 [17] R. Arora, H. Amoodi, S. Stewart, L. Friesen, V. Lin, J. Nedzelski,
778 J. Chen, The addition of a contralateral routing of signals microphone
779 to a unilateral cochlear implant systema prospective study in speech
780 outcomes, *The Laryngoscope* 123 (2013) 746–751.
- 781 [18] M. F. Dorman, S. C. Natale, S. Agrawal, The value of unilateral cis,
782 ci-cros and bilateral cis, with and without beamformer microphones,
783 for speech understanding in a simulation of a restaurant environment,
784 *Audiology and Neurotology* 23 (2018) 270–276.
- 785 [19] Y. Wu, K. He, Group normalization, in: *Proceedings of the European*
786 *conference on computer vision (ECCV)*, 2018, pp. 3–19.
- 787 [20] H. Katahira, N. Ono, S. Miyabe, T. Yamada, S. Makino, Nonlinear
788 speech enhancement by virtual increase of channels and maximum snr
789 beamformer, *EURASIP Journal on Advances in Signal Processing* 2016
790 (2016) 1–8.
- 791 [21] K. Yamaoka, L. Li, N. Ono, S. Makino, T. Yamada, Cnn-based vir-
792 tual microphone signal estimation for mpdr beamforming in underdeter-
793 mined situations, in: *2019 27th European Signal Processing Conference*
794 *(EUSIPCO)*, 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8903040.
- 795 [22] R. Jinzai, K. Yamaoka, M. Matsumoto, S. Makino, T. Yamada, Wave-
796 length proportional arrangement of virtual microphones based on in-
797 terpolation/extrapolation for underdetermined speech enhancement, in:
798 *2019 27th European Signal Processing Conference (EUSIPCO)*, IEEE,
799 2019, pp. 1–5.

- 800 [23] F. Denk, S. M. Ernst, S. D. Ewert, B. Kollmeier, Adapting hearing
801 devices to the individual ear acoustics: Database and target response
802 correction functions for various device styles, *Trends in hearing* 22 (2018)
803 2331216518779313.
- 804 [24] R. M. Corey, N. Tsuda, A. C. Singer, Acoustic impulse responses for
805 wearable audio devices, in: *ICASSP 2019-2019 IEEE International Con-*
806 *ference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE,
807 2019, pp. 216–220.
- 808 [25] I. Himawan, S. Sridharan, I. McCowan, Dealing with uncertainty in
809 microphone placement in a microphone array speech recognition system,
810 in: *2008 IEEE International Conference on Acoustics, Speech and Signal*
811 *Processing*, IEEE, 2008, pp. 1565–1568.
- 812 [26] A. Mathur, A. Isopoussu, F. Kawsar, N. Berthouze, N. D. Lane,
813 *Mic2mic: Using cycle-consistent generative adversarial networks to over-*
814 *come microphone variability in speech systems*, in: *Proceedings of the*
815 *18th International Conference on Information Processing in Sensor Net-*
816 *works*, 2019, pp. 169–180.
- 817 [27] J. Capon, High-resolution frequency-wavenumber spectrum analysis,
818 *Proceedings of the IEEE* 57 (1969) 1408–1418.
- 819 [28] E. A. Habets, J. Benesty, S. Gannot, I. Cohen, The mvdr beamformer for
820 speech enhancement, in: *Speech Processing in Modern Communication*,
821 Springer, 2010, pp. 225–254.

- 822 [29] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, J. Le Roux,
823 Improved mvdr beamforming using single-channel mask prediction net-
824 works., in: Interspeech, 2016, pp. 1981–1985.
- 825 [30] J. Benesty, J. Chen, Y. Huang, Microphone array signal processing,
826 volume 1, Springer Science & Business Media, 2008.
- 827 [31] N. Ito, S. Araki, T. Nakatani, Complex angular central gaussian mix-
828 ture model for directional statistics in mask-based microphone array
829 signal processing, in: 2016 24th European Signal Processing Conference
830 (EUSIPCO), IEEE, 2016, pp. 1153–1157.
- 831 [32] T. Higuchi, K. Kinoshita, M. Delcroix, T. Nakatani, Adversarial train-
832 ing for data-driven speech enhancement without parallel corpus, in:
833 2017 IEEE Automatic Speech Recognition and Understanding Work-
834 shop (ASRU), IEEE, 2017, pp. 40–47.
- 835 [33] L. Drude, R. Haeb-Umbach, Tight integration of spatial and spectral
836 features for bss with deep clustering embeddings., in: Interspeech, 2017,
837 pp. 2650–2654.
- 838 [34] R. B. Blackman, J. W. Tukey, Particular pairs of windows, The mea-
839 surement of power spectra, from the point of view of communications
840 engineering (1959) 98–99.
- 841 [35] T. Fischer, M. Caversaccio, W. Wimmer, Multichannel acous-
842 tic source and image dataset for the cocktail party effect
843 in hearing aid and implant users, Scientific Data 7 (2020).

- 844 URL: <https://doi.org/10.1038/s41597-020-00777-8>.
845 doi:10.1038/s41597-020-00777-8.
- 846 [36] T. Fischer, M. Kompis, G. Mantokoudis, M. Caversaccio, W. Wimmer,
847 Dynamic sound field audiometry: Static and dynamic spatial hearing
848 tests in the full horizontal plane, *Applied Acoustics* 166 (2020) 107363.
849 doi:10.1016/j.apacoust.2020.107363.
- 850 [37] T. Fischer, M. Caversaccio, W. Wimmer, A front-back confusion metric
851 in horizontal sound localization: The fbc score, in: *ACM Symposium*
852 *on Applied Perception 2020*, 2020, pp. 1–5.
- 853 [38] Y.-H. Wu, E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, J. Oleson,
854 Characteristics of real-world signal-to-noise ratios and speech listening
855 situations of older adults with mild-to-moderate hearing loss, *Ear and*
856 *Hearing* 39 (2018) 293.
- 857 [39] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual
858 evaluation of speech quality (pesq)-a new method for speech quality as-
859 sessment of telephone networks and codecs, in: *2001 IEEE International*
860 *Conference on Acoustics, Speech, and Signal Processing. Proceedings*
861 *(Cat. No. 01CH37221)*, volume 2, IEEE, 2001, pp. 749–752.
- 862 [40] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for
863 intelligibility prediction of time–frequency weighted noisy speech, *IEEE*
864 *Transactions on Audio, Speech, and Language Processing* 19 (2011)
865 2125–2136.

- 866 [41] J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey, Sdr-half-baked
867 or well done?, in: ICASSP 2019-2019 IEEE International Conference
868 on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp.
869 626–630.
- 870 [42] D. Stoller, S. Ewert, S. Dixon, Wave-u-net: A multi-scale neural network
871 for end-to-end audio source separation, arXiv preprint arXiv:1806.03185
872 (2018).
- 873 [43] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath,
874 Deep learning for audio signal processing, IEEE Journal of Selected
875 Topics in Signal Processing 13 (2019) 206–219.
- 876 [44] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, K. R. Liu,
877 Acoustic beamforming for hearing aid applications, Handbook on array
878 processing and sensor networks (2010) 269–302.
- 879 [45] D. Griffin, J. Lim, Signal estimation from modified short-time fourier
880 transform, IEEE Transactions on acoustics, speech, and signal process-
881 ing 32 (1984) 236–243.
- 882 [46] K. Yatabe, Y. Masuyama, Y. Oikawa, Rectified linear unit can assist
883 griffin-lim phase recovery, in: 2018 16th International Workshop on
884 Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 555–559.
- 885 [47] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks
886 for biomedical image segmentation, in: International Conference on
887 Medical image computing and computer-assisted intervention, Springer,
888 2015, pp. 234–241.

- 889 [48] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, H. S.
890 Seung, Digital selection and analogue amplification coexist in a cortex-
891 inspired silicon circuit, *Nature* 405 (2000) 947–951.
- 892 [49] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization,
893 arXiv preprint arXiv:1412.6980 (2014).
- 894 [50] L. N. Smith, Cyclical learning rates for training neural networks, in:
895 2017 IEEE Winter Conference on Applications of Computer Vision
896 (WACV), IEEE, 2017, pp. 464–472.
- 897 [51] L. Torrey, J. Shavlik, Transfer learning, in: *Handbook of research*
898 *on machine learning applications and trends: algorithms, methods, and*
899 *techniques*, IGI global, 2010, pp. 242–264.
- 900 [52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan,
901 T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison,
902 A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chil-
903 amkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Py-
904 torch: An imperative style, high-performance deep learning
905 library, in: *Advances in Neural Information Processing Sys-*
906 *tems* 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL:
907 <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performar>
- 908 [53] J. L. Hintze, R. D. Nelson, Violin plots: a box plot-density trace syner-
909 gism, *The American Statistician* 52 (1998) 181–184.
- 910 [54] W. Wimmer, S. Weder, M. Caversaccio, M. Kompis, Speech intelli-

- 911 bility in noise with a pinna effect imitating cochlear implant processor,
912 *Otology & neurotology* 37 (2016) 19–23.
- 913 [55] S. Yu, J. C. Principe, Understanding autoencoders with information
914 theoretic concepts, *Neural Networks* 117 (2019) 104–123.
- 915 [56] F. Fan, J. Xiong, G. Wang, On interpretability of artificial neural net-
916 works, Preprint at <https://arxiv.org/abs/2001.02522> (2020).
- 917 [57] A. Pandey, D. Wang, Tcn: Temporal convolutional neural network for
918 real-time speech enhancement in the time domain, in: *ICASSP 2019-
919 2019 IEEE International Conference on Acoustics, Speech and Signal
920 Processing (ICASSP)*, IEEE, 2019, pp. 6875–6879.
- 921 [58] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, N. Harada, Real-
922 time speech enhancement using equilibrated rnn, in: *ICASSP 2020-
923 2020 IEEE International Conference on Acoustics, Speech and Signal
924 Processing (ICASSP)*, IEEE, 2020, pp. 851–855.
- 925 [59] Y. Koyama, T. Vuong, S. Uhlich, B. Raj, Exploring the best loss func-
926 tion for dnn-based low-latency speech enhancement with temporal con-
927 volutional networks, arXiv preprint arXiv:2005.11611 (2020).
- 928 [60] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency
929 magnitude masking for speech separation, *IEEE/ACM transactions on
930 audio, speech, and language processing* 27 (2019) 1256–1266.
- 931 [61] J. Lu, W. Cheng, D. He, Y. Zi, A novel underdetermined blind source
932 separation method with noise and unknown source number, *Journal of
933 Sound and Vibration* 457 (2019) 67–91.

- 934 [62] L. Drude, R. Haeb-Umbach, Integration of neural networks and proba-
935 bilistic spatial models for acoustic blind source separation, *IEEE Journal*
936 *of Selected Topics in Signal Processing* 13 (2019) 815–826.
- 937 [63] L. Drude, D. Hasenklever, R. Haeb-Umbach, Unsupervised training
938 of a deep clustering model for multichannel blind source separation, in:
939 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech*
940 *and Signal Processing (ICASSP)*, IEEE, 2019, pp. 695–699.
- 941 [64] K. Koutini, H. Eghbal-zadeh, M. Dorfer, G. Widmer, The receptive field
942 as a regularizer in deep convolutional neural networks for acoustic scene
943 classification, in: *2019 27th European Signal Processing Conference*
944 *(EUSIPCO)*, 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8902732.
- 945 [65] S. S. R. Phaye, E. Benetos, Y. Wang, Subspectralnet using sub-
946 spectrogram based convolutional neural networks for acoustic scene clas-
947 sification, in: *ICASSP 2019 - 2019 IEEE International Conference on*
948 *Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 825–829.
949 doi:10.1109/ICASSP.2019.8683288.
- 950 [66] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns:
951 Large-scale pretrained audio neural networks for audio pattern recogni-
952 tion, *IEEE/ACM Transactions on Audio, Speech, and Language Pro-*
953 *cessing* 28 (2020) 2880–2894.
- 954 [67] Z. Borsos, Y. Li, B. Gfeller, M. Tagliasacchi, Micaugment: One-shot
955 microphone style transfer, arXiv preprint arXiv:2010.09658 (2020).

- 956 [68] S.-W. Fu, C.-F. Liao, Y. Tsao, S.-D. Lin, Metricgan: Generative adver-
957 sarial networks based black-box metric scores optimization for speech
958 enhancement, arXiv preprint arXiv:1905.04874 (2019).
- 959 [69] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang,
960 Y. Gong, Cracking the cocktail party problem by multi-beam
961 deep attractor network, in: 2017 IEEE Automatic Speech Recog-
962 nition and Understanding Workshop (ASRU), 2017, pp. 437–444.
963 doi:10.1109/ASRU.2017.8268969.
- 964 [70] S. Inoue, H. Kameoka, L. Li, S. Seki, S. Makino, Joint separation and
965 dereverberation of reverberant mixtures with multichannel variational
966 autoencoder, in: ICASSP 2019 - 2019 IEEE International Conference
967 on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 96–100.
968 doi:10.1109/ICASSP.2019.8683497.
- 969 [71] F. Feng, M. Kowalski, Underdetermined reverberant blind
970 source separation: Sparse approaches for multiplicative and con-
971 volutive narrowband approximation, IEEE/ACM Transactions
972 on Audio, Speech, and Language Processing 27 (2019) 442–456.
973 doi:10.1109/TASLP.2018.2881925.
- 974 [72] Y. Xie, K. Xie, J. Yang, Z. Wu, S. Xie, Underdetermined reverberant
975 audio-source separation through improved expectation–maximization al-
976 gorithm, Circuits, Systems, and Signal Processing 38 (2019) 2877–2889.
- 977 [73] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre,
978 E. de la Rubia-Cuestas, L. Molina-Tanco, A. Reyes-Lecuona,

- 979 3D Tune-In Toolkit: An open-source library for real-time
980 binaural spatialisation, PLOS ONE 14 (2019) e0211899.
981 URL: <http://dx.plos.org/10.1371/journal.pone.0211899>.
982 doi:10.1371/journal.pone.0211899.
- 983 [74] T. Fischer, M. Caversaccio, W. Wimmer, System for combined hear-
984 ing and balance tests of a person with moving sound source devices,
985 2020. URL: <https://patents.google.com/patent/WO2020254462A1>,
986 WO Patent WO2020254462A1.
- 987 [75] I. Kiselev, E. Ceolini, D. Wong, A. De Cheveigne, S. Liu, Whisper:
988 Wirelessly synchronized distributed audio sensor platform, in: 2017
989 IEEE 42nd Conference on Local Computer Networks Workshops (LCN
990 Workshops), 2017, pp. 35–43. doi:10.1109/LCN.Workshops.2017.62.
- 991 [76] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He,
992 A comprehensive survey on transfer learning, Proceedings of the IEEE
993 109 (2021) 43–76. doi:10.1109/JPROC.2020.3004555.
- 994 [77] D. Marquardt, V. Hohmann, S. Doclo, Interaural coherence preserva-
995 tion in multi-channel wiener filtering-based noise reduction for binaural
996 hearing aids, IEEE/ACM Transactions on Audio, Speech, and Language
997 Processing 23 (2015) 2162–2176. doi:10.1109/TASLP.2015.2471096.
- 998 [78] D. Marquardt, S. Doclo, Interaural coherence preservation for binaural
999 noise reduction using partial noise estimation and spectral postfiltering,
1000 IEEE/ACM Transactions on Audio, Speech, and Language Processing
1001 26 (2018) 1261–1274. doi:10.1109/TASLP.2018.2823081.

- 1002 [79] N. Gößling, D. Marquardt, S. Doclo, Perceptual evaluation of binaural
1003 mvdr-based algorithms to preserve the interaural coherence of diffuse
1004 noise fields, Trends in Hearing 24 (2020) 2331216520919573.

Journal Pre-proof



^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

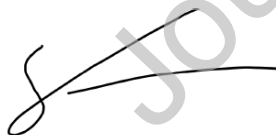
Author statement:

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

The logo of the University of Bern, featuring a stylized lowercase 'u' with a superscript 'b'.

^b
UNIVERSITÄT
BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

The logo for InselSpital, consisting of a green square icon with a white 'U' shape inside, followed by the text 'INSELSPITAL' in green.

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,

A handwritten signature in black ink, appearing to be 'W. Wimmer'.

Dr. Wilhelm Wimmer