A BINAURAL HEARING AID SPEECH ENHANCEMENT METHOD MAINTAINING SPATIAL AWARENESS FOR THE USER

Joachim Thiemann, Menno Müller and Steven van de Par

Carl-von-Ossietzky University Oldenburg, Cluster of Excellence 'Hearing4All'
Oldenburg, Germany

ABSTRACT

Multi-channel hearing aids can use directional algorithms to enhance speech signals based on their spatial location. In the case where a hearing aid user is fitted with a binaural hearing aid, it is important that the binaural cues are kept intact, such that the user does not loose spatial awareness, the ability to localize sounds, or the benefits of spatial unmasking. Typically algorithms focus on rendering the source of interest in the correct spatial location, but degrade all other source positions in the auditory scene. In this paper, we present an algorithm that uses a binary mask such that the target signal is enhanced but the background noise remains unmodified except for an attenuation. We also present two variations of the algorithm, and in initial evaluations find that this type of mask-based processing has promising performance.

Index Terms— Hearing Aids, Spatial Rendering, Speech Enhancement, Beamforming

1. INTRODUCTION

Many modern hearing aids employ multi-channel noise reduction methods based on small microphone arrays to exploit the spatial separation of the sound sources in the environment. These multi-channel methods (such as beamforming [1, 2]) are in general capable of lower distortion and better noise suppression than single-channel enhancement techniques.

For hearing aid users requiring assistance on both ears, multi-channel hearing aids exist in various configurations. It has been shown that binaural cues can be distorted if the hearing aids work independently for each ear, reducing the overall intelligibility (due to reduced spatial unmasking in the auditory system) [3]. To alleviate this problem, the two hearing aids can be linked to form a single array with two outputs where the binaural cues can be controlled [4].

Using a speech enhancement algorithm can lead to distorting the binaural cues especially of the background noise. In many circumstances, this can be very disturbing to the user since important information about the user's surroundings is removed. One can imagine many scenarios where this can be not just disturbing, but even dangerous, such as in traffic or work situations where equipment indicators need to be heard. As a result, we aim to develop algorithms for multi-channel hearing aids that obtain good enhancement of the target signal, while preserving the spatial impression of both the target signal as well as the background noise.

This research was conducted within the Hearing4All cluster of excellence with funding from DFG grant 1077.

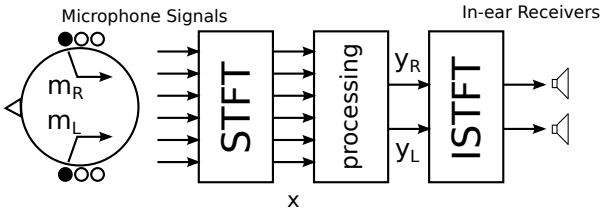


Fig. 1: Overview of array processing of sound in a multi-channel binaural hearing aid. Small circles represent the microphones, the filled circles showing the left and right reference microphones.

In this article, we present a method that uses a binary mask in the time-frequency (T-F) plane to create the signals presented to the hearing aid user. At the resolution of the T-F plane, the binary mask controls if the signal is taken from the enhancement algorithm or the reference microphones without processing. This means that in the absence of a highly localized target source, the user hears a completely unmodified (except for a possible gain factor) signal. This type of manipulation is already used in multi-microphone methods, and is similar to methods found in blind source separation [5].

The basics of multi-channel directional speech enhancement are described in the following section. Section 3 describes our proposed modification and some variations. In section 4, we describe our preliminary objective and subjective evaluation of the algorithm and its variations compared to some established multi-channel hearing aid speech enhancement algorithms.

2. BACKGROUND

We consider hearing aids with a small number of microphones that are closely spaced in the direct vicinity of the ear where all microphones of the hearing aids are processed in a sin-

gle device. Figure 1 shows an overview of such a system with 3 microphones on each ear. Note that for each ear, one of the microphones is designated as a reference microphone. We assume that the direction of the target signal is known. Working in the short-time fourier transform (STFT) domain, we write $\mathbf{x}(f,n) = [x_1(f,n)\, x_2(f,n) \dots x_M(f,n)]^T$ for the M-channel microphone signal, and $y_L(f,n)$ and $y_R(f,n)$ for the left and right ear signal respectively. We use f and n as the frequency and time indices of the T-F plane.

A well-known algorithm for directional enhancement of multi-channel microphone signals is the Minimum Variance Distortionless Response (MVDR) beamformer [6], where the filter coefficients are computed as

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_{\mathrm{NN}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathrm{NN}}^{-1}(f)\mathbf{d}(f)}, \tag{1}$$

and the single-channel output is computed as

$$y_{\mathrm{bf}}(f,n) = \mathbf{w}^H(f)\mathbf{x}(f,n). \tag{2}$$

The MVDR beamformer relies on the noise covariance matrix $\mathbf{\Phi}_{\mathrm{NN}}$ and the steering vector $\mathbf{d}$: note that we keep these quantities fixed w.r.t. the time index n, restricting ourselves to a fixed beamformer for simplicity.

The vector $\mathbf{d}(f) = [d_1(f)\, d_2(f) \dots d_M(f)]^T$ steers the beamformer, and depends on the position of the target source. It can be set in a variety of ways, for example from the array geometry under free field assumptions or from measurements using signals under controlled conditions. We assume here that $\mathbf{d}$ is normalised by setting one of the elements d_m to 1 for each frequency f thus making the mth microphone the reference microphone (that is, the microphone at the spatial location where the signal estimation is referenced).

2.1. Beamforming for two ears

Without much added computational effort, the input $\mathbf{x}$ can be used by multiple beamformers [1, 7]. As a result, one method of using the MVDR beamformer for a binaural hearing aid is to compute two steering vectors $\mathbf{d}_L(f)$ and $\mathbf{d}_R(f)$ for the left and right ears, respectively, which simply use different microphone channels as reference ($m = m_L$ or m_R). These two outputs differ only in terms of a complex scaling factor. We refer to this as the binaural MVDR.

Another method to build a beamformer with outputs for each ear is to restrict $\mathbf{d}_L(f)$ and $\mathbf{d}_R(f)$ to only use those microphone channels that are on the left and right side of the head respectively. This corresponds to a bilateral hearing aid where each side is independent of the other [3, 7], and can be used as a reference method.

3. PROPOSED ENHANCEMENT ALGORITHM

As described in the previous section, in the output of the binaural MVDR beamformer all frequency bins of one channel are simply frequency-dependent complex scaled copies of the other channel. The perceived effect is that the entire signal (both the target and the background noise) appear to originate from the direction of the target signal [2]. This means it is impossible to localize interfering signals, even if they are not completely cancelled out.

Some approaches have been proposed to address the rendering of the overall binaural scene. One example presented in [8] is used as a comparison in section 4. This algorithm restricts modification of the input signal to a real-valued gain factor to avoid destroying interaural cues.

In this paper, we propose an approach based on a binary allocation of T-F bins as either the target signal or background noise, where background noise may be diffuse or localizable interfering sources. The output signal in each ear is computed by selecting, on a T-F bin basis, either the attenuated output from the respective reference microphone or the output from the MVDR beamformer. In this way the binaural cues of the background noise are preserved, and the binaural cues of the target signal can be controlled independently. The selection is based on determining if the energy in the T-F bin is dominated by the target signal or background noise. Denoting $y_{\mathrm{SBB},L}$ and $y_{\mathrm{SBB},R}$ ("selective binaural beamformer", SBB) for the first variant of our algorithm (left and right channels), this can succinctly be written as

$$y_{\mathrm{SBB},L}(f,n) = \begin{cases} \mathbf{w}_L^H(f)\mathbf{x}(f,n), & t(f,n) = 1, \\ \gamma x_{m_L}(f,n), & \text{otherwise}, \end{cases} \tag{3}$$

where $t(f,n)$ is the decision of the bin (f,n) being dominated by the target signal ($t(f,n) = 1$) or not ($t(f,n) \neq 1$). The right ear signal is computed in the same manner, with the same mask. The attenuation γ is a simple real scalar that determines how much of the original signal is kept in the output, and may be changed based on user preference.

Generating the mask $t(f,n)$ is a crucial part of the algorithm, and will be further studied in the future. In the current implementation, we use a method that relies on the spatial gain properties of the beamformer. We base the classification on the fact that if in a given T-F bin the beamformer output is of lower energy than the inputs of the reference microphones, the energy in that bin is most likely dominated by the background noise. Specifically, we compute

$$t(f,n) = \begin{cases} 1, & |\mathbf{w}_{\mathrm{be}}^H(f)\mathbf{x}(f,n)| > E_{x\mathrm{av}}(f,n), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbf{w}_{\mathrm{be}}^H(f)$ is the beamformer referenced to the side closer to the target, that is eq. (1) using $\mathbf{d}_L$ or $\mathbf{d}_R$ depending on the target signal being on the left or right side. The average input energy is computed as $E_{x\mathrm{av}}(f,n) = \frac{1}{M}\sum_m |x_m(f,n)|$.

3.1. Additional algorithm variants

We now explore some variations of the basic binary allocation algorithm proposed above. We begin by noting that

in those T-F bins where the energy is dominated by the target signal, the background noise is by definition insignificant (within some allowable margin). Thus, enhancement of the target signal can be achieved by simply not attenuating the detected target signal bins, i.e.

$$y_{\mathrm{SA},L}(f,n) = \begin{cases} x_{m_L}(f,n), & t(f,n)=1, \\ \gamma x_{m_L}(f,n), & \text{otherwise}, \end{cases} \quad (5)$$

("selective attenuation", SA) and similarly for $y_{\mathrm{SA},R}(f,n)$ for the second variant algorithm. We note that in this variant of the algorithm the beamformer is used *only* for calculating the T-F mask. Note that this variant is similar to the algorithm in [8], however with a gain function restricted to the values $\{\gamma, 1\}$.

Another possibility is to consider a single-channel output (e.g. the left ear) that is used to compute the mask, and artificially render it at the original location by applying a phase-shift on the STFT coefficients. The phase shift is based on a geometric calculation of the time difference of arrival (TDOA), computing $\phi(f) = e^{-2\pi j \omega(f) d_{\mathrm{ear}} \sin(\alpha)/c}$, where $\omega(f)$ is the center frequency (in Hz) of the STFT bin f, d_{ear} is the interaural distance (in m), α the angle specifying the direction of the target, and c is the speed of sound in air (m/s). Assuming the target source is located to the left, we write the third variant ("TDOA simulation", TS) of the algorithm as

$$y_{\mathrm{TS},L}(f,n) = \begin{cases} \mathbf{w}_L^H(f)\mathbf{x}(f,n), & t(f,n)=1, \\ \gamma x_{m_L}(f,n), & \text{otherwise}, \end{cases} \quad (6)$$

$$y_{\mathrm{TS},R}(f,n) = \begin{cases} \phi(f)\mathbf{w}_L^H(f)\mathbf{x}(f,n), & t(f,n)=1, \\ \gamma x_{m_R}(f,n), & \text{otherwise}. \end{cases} \quad (7)$$

If the target is located to the right of the hearing aid user, the channels need to be swapped as appropriate. The assumption that phase modification is sufficient to render the sound at the correct spatial location is based on the idea that interaural time differences (ITDs) are a very strong directional cue for human listeners and in exchange for the loss of interaural level difference cues, we get a significant boost in the level of the target signal in the ear that faces away from the target source.

4. EVALUATION

In our preliminary evaluation of the proposed methods, we use a binaural hearing aid with three microphones per hearing aid, where the microphones are arranged above and behind the pinna. We consider a reverberant environment with associated ambient noise which is both typical and challenging for hearing aid users. For this device, the impulse responses from selected points in the room to the hearing aid model are available, as well as impulse responses measured in an anechoic chamber. The full description of the device and the recordings can be found in [9], and we specifically use the "cafeteria" environment and ambient noise recordings.

We consider two positions relative to the hearing aid: Position A, 102 cm directly in front of the dummy head, and position B, 30° to the left from the center, 117.5 cm away. The speech signals are simulated by convolving the anechoic recordings by the HRIRs corresponding to those positions. Speech items are of two male and two female speakers. The steering vector $\mathbf{d}(f)$ is taken from the anechoic HRIRs (depending on target location, 0° or -30°), and we generate $\mathbf{d}_L(f)$ and $\mathbf{d}_R(f)$ by normalising w.r.t. the front left or the front right microphone. The noise covariance matrix estimate $\mathbf{\Phi}_{\mathrm{NN}}$ is computed from the anechoic HRIRs as well, using the assumption of a cylindrically isotropic noise field. This means the algorithm has no knowledge of the particular spectral or spatial characteristics of the noise added to the signal and instead computes $\mathbf{\Phi}_{\mathrm{NN}}{}'(f)$ by summing the HRIR from all directions. We use a small frequency-dependent value $\mu(f)$ to regularize $\mathbf{\Phi}_{\mathrm{NN}}(f)$ towards low frequencies, by

$$\mathbf{\Phi}_{\mathrm{NN}}(f) = (1-\mu(f))\mathbf{\Phi}_{\mathrm{NN}}{}'(f) + \mu(f)\mathbf{I}, \quad (8)$$

where $\mu(f) = \frac{1}{f^8}$, found empirically. The effect of the regularization vanishes beyond the first few bins.

4.1. Comparisons to related algorithms

We compare the three proposed algorithm variants ("SBB", "SA", and "TS") to the simple bilateral enhancement, binaural MVDR ("bilat" and "binaural" respectively, see sec. 2.1) as well as the algorithm in [8] ("Lot06"), since it is conceptually very similar in design and purpose. However, since Lot06 is described for 2-channel inputs, the calculation of $Z(k)$ in [8] is modified for 6-channel input to remove any advantage that our proposed algorithms may have simply due to the increased number of microphones. All processing is done on 16 kHz sampled audio files, and the signals are transformed into frequency domain using a 1024 point STFT with full overlap. The attenuation factor γ is set to 0.3.

4.2. Objective Evaluation

The objective evaluation of our algorithms focuses on the amount of enhancement relative to the reference microphone signals (the front left and right microphones) alone. We consider a target at position A (0°) or B (-30°), mixed with ambient recorded noise at an input segmental SNR (iSNR) of -6, -3, 0, 3 and 6 dB. SegSNRs are averaged between the left and right channels, using segments of 1024 samples. To compute the output SegSNR, the unmixed target and background noise signals are processed in the same manner (that is, using the same mask) as the mixture.

Tables 1a shows the SegSNR enhancement (SNRE) w.r.t. the reference microphones for the target at position A. In terms of pure enhancement the traditional binaural MVDR provides the highest gain. In this algorithm, the background noise however is not rendered accurately and hence

Table 1: Comparion of SNR Enhancement, in dB

(a) Target at 0°

iSNR	SBB	SA	TS	Lot06	binaural	bilat
-6	2.68	2.23	2.58	2.94	5.22	3.36
-3	2.92	2.08	2.82	2.69	5.19	3.36
0	3.13	1.90	3.02	2.41	5.17	3.37
3	3.25	1.66	3.16	2.09	5.11	3.35
6	3.50	1.39	3.39	1.62	5.01	3.33

(b) Target at -30°

iSNR	SBB	SA	TS	Lot06	binaural	bilat
-6	3.43	2.64	4.48	2.55	5.37	2.58
-3	3.78	2.56	4.84	2.36	5.36	2.56
0	3.99	2.32	4.98	2.08	5.32	2.51
3	4.09	1.98	4.94	1.74	5.26	2.44
6	4.08	1.54	4.84	1.29	5.10	2.30

Table 2: SNRE per channel, Target at -30°

Channel	SBB	SA	TS	Lot06	binaural	bilat
Left	3.48	2.82	3.48	2.38	3.46	2.14
Right	4.26	1.60	6.15	1.63	7.10	2.81

can be greatly suppressed. Of the four algorithms designed to render the acoustic scene accurately, the two algorithms mixing the beamformer output with the input signal ("SBB" and "TS") outperform those that simply apply a gain to the input. However, only at large input SNRs, the performance approaches the performance of the bilateral beamformer.

The situation changes however when the target is not in the front center, as shown in Table 1b. Here, both SBB and TS show a considerably higher SNR enhancement, with the TS algorithm even approaching the binaural MVDR at high input SNR.

In Table 2, the SNRE is averaged for all iSNR conditions, but given for the left and right channels individually. Like the binaural MVDR beamformer, the TS algorithm (and, to a lesser degree, the SBB algorithm) has a drastic gain in the ear that is facing away from the source.

4.3. Subjective Evaluation

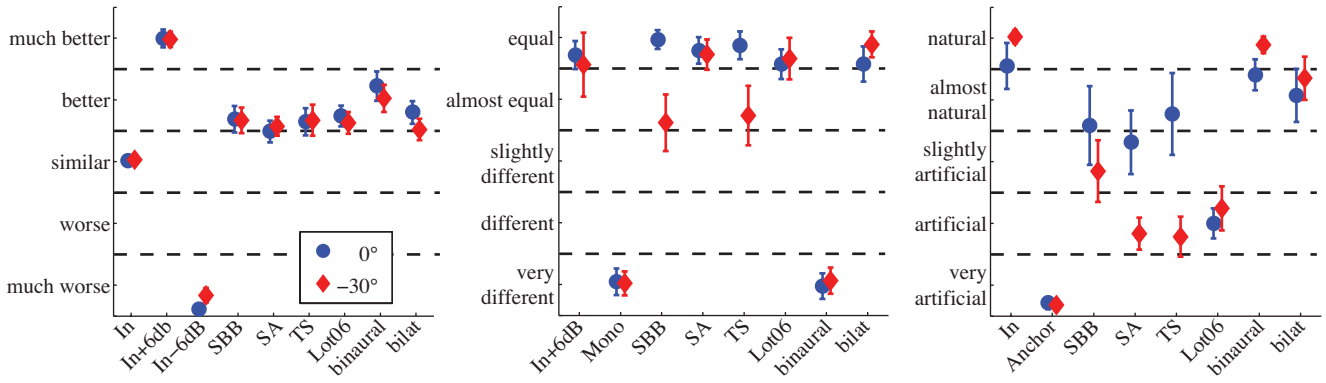
To obtain a subjective assessment of the proposed algorithms, we adapt the MUSHRA (ITU-R BS.1534) testing methodology [10]. MUSHRA as originally designed is not a suitable method since it assumes that all algorithms under test will degrade the subjective quality, relative to a known reference, of the signal to some degree. As we are assessing a speech enhancement algorithm with a focus on spatial rendering, we modify MUSHRA such that a) the user is not asked to locate a reference, b) we add a high quality and a low quality anchor as appropriate. The high quality anchor for the intelligibility and spatial rendering tests is a mixture where the target speech signal is boosted 6 dB compared to the input mixture processed by the algorithms under test, while for the naturalness test the input signal is used. The low quality anchor is different for each test run depending on the property of the algorithms the subjects are evaluating.

To give listeners a background source that is localizeable, in the subjective tests the target source is combined with a background signal that is a mix of the ambient noise and an interfering speaker. The spatial location of the target and interferer are such that if the target is at pos. A (see above), the interferer is at pos. B and vice versa. As an input signal, the target is mixed with an interferer with equal power (Segmental SNR 0 dB), and the ambient noise is added such that the target (only) to ambient noise has a segmental SNR of -6 dB. Listeners are given a visual (written) indication if the target speaker is supposed to be in front or at -30°. The results are from six normal hearing individuals, evenly split between male and female, with an average age of about 28 years.

In the first test, the listeners are asked to evaluate the speech intelligibility of the target speaker. As a low quality anchor we use a mixture similar to the signal being processed with the target in the mixture 6 dB lower than in the test signal. From initial test runs, we find that the differences are very difficult to judge; to ensure that we truly observe an enhancement we include the input signal in this test. Shown in Fig 2a, all algorithms under test show some apparent enhancement over the reference, but in this limited evaluation no algorithm shows a clear advantage over any other algorithm in terms of speech enhancement. A better measure to evaluate the enhancement is to measure the speech reception threshold (STR), which will be performed in future studies.

The reconstruction of the auditory scene in terms of spatial location is evaluated in the second test, where the results are shown in Fig. 2b. For this test, the anchor is the input signal presented transaurally, that is, as an identical mono signal in both ears. Here, we see the problem of the binaural MVDR: it is judged just as bad as the reference mono signal, since it is effectively a mono signal as well, even when the target is located off-center. The bilateral method performs surprisingly well, indicating that overall the binaural cues are left intact. Comparing the proposed algorithms with the reference Lotter algorithm, we see that the former appear to perform slightly better, though the sample size is too small to make a definitive statement. If the target is located off-center however, the SBB and TS algorithms show a distinct drop in performance.

Finally, Fig. 2c shows the results where listeners are asked to evaluate the signal in terms of "naturalness," where artefacts such as musical noise or speech distortion should be judged as artificial. Here, the anchor is a signal processed with a mask that causes a great deal of musical noise. This task was much harder for the listeners, as can be seen by the large variance that the analysis of the responses reveals. As in the spatial scene reconstruction test described above, the pro-



(a) Speech intelligibility (b) Spatial scene rendering (c) Naturalness (artefacts)

Fig. 2: Subjective evaluation results

posed algorithms show poor performance if the target signal is not in the center. Surprisingly though, Lotters algorithm is evaluated as having poor performance even if the target is in the center.

5. DISCUSSION AND CONCLUSION

The algorithms presented here attempt to balance the requirement of enhancing a speech signal that originates from a known direction in space yet preserve the spatial rendering of the background noise. The key idea is to create a T-F mask that distinguishes between target speech and background noise. Where the T-F mask indicates noise, the input signal is passed only through an attenuator, leaving all binaural cues unmodified. The target speech signal on the other hand can be rendered in a variety of ways, and we present three methods of doing so.

The methods we present show some promise, especially the SBB algorithm. Currently, it appears that the beamformer is a significant limitation of the enhancement quality, which also affects the mask that is computed. Ongoing research aims at improving the mask generation, including an extension to multi-target enhancement.

REFERENCES

[1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds., chapter 9, pp. 269–302. Wiley, 2010.

[2] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 18, no. 2, pp. 342–355, Feb 2010.

[3] T. Van den Bogaert, T. J. Klasen, M. Moonen, and J. Wouters, "Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids," in *Proc. IEEE Workshop on Applications of Signal Proc. to Audio and Acoust. (WASPAA)*, 2005, pp. 57–60.

[4] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 124, no. 1, Jul. 2008.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*. Springer Verlag, 2010.

[7] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-Array Hearing Aids with Binaural Output — Part I: Fixed-Processing Systems," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 5, no. 6, pp. 529–542, Nov. 1997.

[8] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. on Applied Sig. Proc.*, vol. 2006, pp. 1–14, 2006.

[9] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, 2009.

[10] ITU-R, "ITU-R Recommendation BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems," 2003.