BINAURAL MASK-INFORMED SPEECH ENHANCEMENT FOR HEARING AIDS WITH HEAD TRACKING

Alastair H. Moore, Leo Lightburn, Wei Xue, Patrick A. Naylor and Mike Brookes

Imperial College London
Electrical and Electronic Engineering, Exhibition Road, London

ABSTRACT

An end-to-end speech enhancement system for hearing aids is proposed which seeks to improve the intelligibility of binaural speech in noise during head movement. The system uses a reference beamformer whose look direction is informed by knowledge of the head orientation and the a priori known direction of the desired source. From this a time-frequency mask is estimated using a deep neural network. The binaural signals are obtained using bilateral beamformers followed by a classical minimum mean square error speech enhancer, modified to use the estimated mask as a speech presence probability prior. In simulated experiments, the improvement in a binaural intelligibility metric (DBSTOI) given by the proposed system relative to beamforming alone corresponds to an SNR improvement of 4 to 6 dB. Results also demonstrate the individual contributions of incorporating the mask and the head orientation-aware beam steering to the proposed system.

***Index Terms*—** Beamforming, Speech enhancement, Time-frequency mask, Assisted listening, Head rotation

1. INTRODUCTION

Whilst hearing aids are easily capable of improving audibility, making speech intelligible in the presence of high levels of background noise remains an important challenge. Two approaches to speech enhancement that can offer worthwhile intelligibility improvements are spatial selectivity and mask-informed enhancement. A hearing aid signal processing scheme is proposed which combines these approaches with dynamic head-tracking information.

Spatial selectivity can be obtained using beamforming whereby the signals from multiple microphones are filtered and combined to preserve signals arriving from one direction while suppressing those from other directions. It is typically assumed that the signal of interest, the desired source, is positioned in front of the listener and so the beam is steered to the front. If the signals from only one aid are used, the beam pattern is quite broad and so the suppression of interferers is limited. Combining signals from binaural aids allows narrower beams to be obtained at the expense of additional power consumption [1, 2]. An undesired consequence of combining signals from both ears is that the interaural differences associated with interfering signals can be reduced [3], degrading the ability of the auditory system to separate the speech and interference [4]. A number of approaches have been proposed which add constraints to the binaural beamformer in order to maintain the interaural cues of sources whose directions or signal statistics are known or can be estimated [5, 6, 7]. Estimation of source and interferer directions of arrival (DOAs) in low signal to noise ratio (SNR) conditions is a challenging problem in its own right. Moreover, natural head movements make online estimation of signal parameters difficult and so care must be taken to avoid instability in the beamforming filters.

In [8, 9] DOA estimation of two sources with rotating arrays was informed by an inertial measurement unit (IMU), with the resulting DOAs used to steer generalized-sidelobe-canceler beamformers for source separation. In our approach we similarly assume that the head orientation is available from an IMU. However, to ensure robustness at low SNRs, we propose a simple user calibration procedure to set the desired source DOA and signal-independent beamforming.

It has been shown that methods based on binary masks are able to improve the intelligibility of single-channel enhanced speech whereas conventional minimum mean squared error (MMSE)-based approaches are unable to do so [10]. Unfortunately, the direct application of a binary mask leads to poor perceived quality [11]. In this work, we (i) use a deep neural network (DNN) to estimate a binary mask that captures the modulations of the target source and (ii) use the mask to define the speech presence probability in a conventional speech enhancer [12, 13].

The signal model is described in Sec. 2. Our approach to combining signal independent, head-tracking-informed beamforming with mask-informed enhancement is outlined in Sec. 3. Experiments are presented in Sec. 4 and conclusions drawn in Sec. 5.

2. PROBLEM FORMULATION

The problem is formulated directly in the short term Fourier transform (STFT) domain with frequency index denoted ν and time frame denoted ℓ. The acoustic pressure at an arbitrary point in a free field, at which we place the origin of our coordinates system, is modeled as an infinite sum of plane waves. The DOA of a wavefront is denoted $\underline{\Omega} = (\underline{\vartheta}, \underline{\phi})$, where $\underline{\vartheta}$ is the inclination and $\underline{\phi}$ is the azimuth and the underline notation indicates angles expressed in world coordinates. Taking the coordinate origin to be the center of the listener's head, the signal at the origin due to the desired source is denoted $S_O(\nu, \ell)$ and has DOA $\underline{\Omega}_s$. The undesired signal, due to interfering sources and diffuse acoustic noise, is $V_O(\nu, \ell) = \int_{\underline{\Omega} \in \mathcal{S}^2} V(\nu, \ell, \underline{\Omega}) d\underline{\Omega}$, where $V(\nu, \ell, \underline{\Omega})$ is the acoustic noise signal arriving from direction $\underline{\Omega}$ and $d\underline{\Omega} = \sin(\underline{\vartheta}) d\underline{\vartheta} d\underline{\phi}$.

The signal received by the m^{th} microphone in the array is

$$Y_m(\nu, \ell) = X_m(\nu, \ell) + V_m(\nu, \ell) + \chi_m(\nu, \ell) \qquad (1)$$

where $X_m(\nu, \ell)$ and $V_m(\nu, \ell)$ represent the contributions due to the desired source and acoustic noise, respectively, and $\chi_m(\nu, \ell)$ is sensor noise, which is uncorrelated between the microphones.The transformations from $S_O(\nu, \ell)$ and $V_O(\nu, \ell)$ to $X_m(\nu, \ell)$ and $V_m(\nu, \ell)$, respectively, depend on the array's shape, which is assumed to be

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M026698/1].

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

fixed, and its orientation, which may vary with time. The array manifold, $H_m(\nu, \Omega)$, is defined as the relative transfer function (RTF) between the signal at the m^{th} microphone and the pressure that would have been observed at the origin for a plane wave with DOA with respect to the array, Ω. The apparent DOA with respect to the array at time frame ℓ, $\Omega(\ell)$, is a function of the array orientation, $\Lambda(\ell)$. Using the multiplicative transfer function (MTF) approximation [14], the dependence of the microphone signals on the free-field signals is

$$X_m(\nu,\ell) = H_m(\nu, \Omega_s(\ell))S_O(\nu,\ell) \tag{2a}$$

$$V_m(\nu,\ell) = \int_{\underline{\Omega}\in\mathcal{S}^2} H_m(\nu, \Omega(\ell))V(\nu,\ell,\underline{\Omega})d\underline{\Omega}. \tag{2b}$$

Our aim is to enhance the contribution of the desired source to the left and right reference channels, denoted $X_l(\nu,\ell)$ and $X_r(\nu,\ell)$, respectively, such that the predicted binaural intelligibility of the desired source is maximized.

It is assumed that the orientation, $\Lambda(\ell)$, of the head is available, for example using an IMU, and that the DOA of the desired source, $\underline{\Omega}_s$, is known, or can be obtained using a straightforward calibration step by the user.

3. PROPOSED SYSTEM

The proposed speech enhancement system for binaural hearing aids with head-tracking is shown in Fig. 1. The steering block determines the direction of the desired source with respect to the current orientation of the head. The reference beamformer is steered towards the desired source and estimates the desired free-field source signal at the origin. This signal is used to estimate a mask, $B(\nu,\ell)$, which aims to optimize monaural speech intelligibility. The bilateral beamformers are steered towards the listener's look direction (i.e. fixed with respect to the rotating head) and produce estimates of the desired source at the left and right reference channels. Finally, the estimated mask is used by the enhancement block to improve the intelligibility of the binaural signals.

3.1. Beamformers

The proposed system uses a total of three beamformers, identified by subscript $\alpha \in \{O, l, r\}$, which are described by the general equation

$$Z_\alpha(\nu,\ell) = \mathbf{w}_\alpha(\nu,\ell)^H \mathbf{y}_\alpha(\nu,\ell) \tag{3}$$

where $\mathbf{w}_\alpha(\nu,\ell)$ is the vector of filter weights, $\mathbf{y}_\alpha(\nu,\ell)$ is the vector of input signals and $(\cdot)^H$ is the conjugate transpose operator. To obtain optimal noise reduction under the constraint of no speech distortion, the filter weights are given by the well-known minimum variance distortionless response (MVDR) solution [15]

$$\mathbf{w}_\alpha(\nu,\ell) = \frac{\mathbf{R}_\alpha(\nu,\ell)^{-1}\mathbf{d}_\alpha(\nu,\ell)}{\mathbf{d}_\alpha(\nu,\ell)^H\mathbf{R}_\alpha(\nu,\ell)^{-1}\mathbf{d}_\alpha(\nu,\ell)} \tag{4}$$

where $\mathbf{R}_\alpha(\nu,\ell) = \mathbb{E}\{(\mathbf{y}_\alpha(\nu,\ell)-\mathbf{x}_\alpha(\nu,\ell))(\mathbf{y}_\alpha(\nu,\ell)-\mathbf{x}_\alpha(\nu,\ell))^H\}$ is the covariance matrix of the interference and $\mathbf{d}_\alpha(\nu,\ell)$ is the steering vector. For robustness, all three beamformers approximate $\mathbf{R}_\alpha(\nu,\ell)$ by assuming that the interference is due to a spherically isotropic noise field and no sensor noise. They differ in their input signals and steering vectors. Since the weights at each frequency are independent, the dependence on ν is omitted in the remainder of this section.

The reference beamformer, identified by $\alpha = O$, exploits all M input signals, $\mathbf{y}_O(\ell) = \begin{bmatrix}Y_1(\ell) \; Y_2(\ell) \; \cdots \; Y_M(\ell)\end{bmatrix}^T$, where $(\cdot)^T$ denotes the transpose operator, and is steered towards the desired source, $\Psi_O(\ell) = \Omega_s(\ell)$. Since the aim is to estimate S_O, i.e. the desired source signal at the origin, and the array manifold is defined in Sec. 2 as the RTF with respect to the origin, the required steering vector is

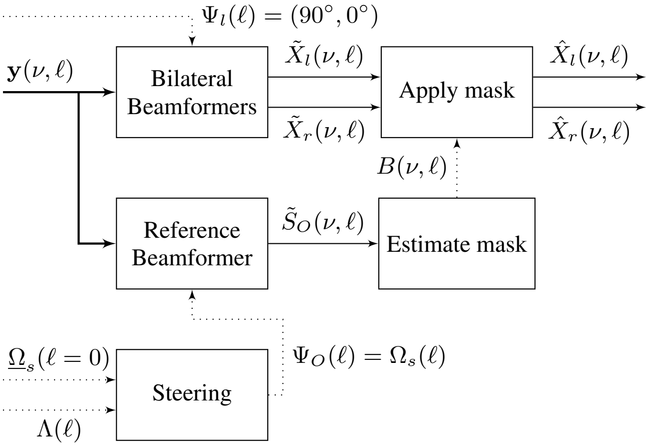


Fig. 1. Block diagram of proposed system. Solid lines represent acoustic signals with thick lines used for multichannel signals. Dashed lines represent parameter values.

$$\mathbf{d}_O(\ell) = \begin{bmatrix}H_1(\Psi_O(\ell)) \; H_2(\Psi_O(\ell)) \; \cdots \; H_M(\Psi_O(\ell))\end{bmatrix}^T. \tag{5}$$

The bilateral beamformers obtain, for each ear, a single channel estimate of the desired signal in the reference channel of that ear. The left and right beamformers, denoted by subscripts l and r respectively, each use microphones only from their respective ear and are, therefore, independent. Without loss of generality, the microphones are indexed such that those on the left (resp. right) side have odd (resp. even) values of m and the reference channel is $m = 1$ (resp. $m = 2$). Therefore $\mathbf{y}_l(\ell) = \begin{bmatrix}Y_1(\ell) \; Y_3(\ell) \; \cdots \; Y_{M-1}(\ell)\end{bmatrix}^T$ and $\mathbf{y}_r(\ell) = \begin{bmatrix}Y_2(\ell) \; Y_4(\ell) \; \cdots \; Y_M(\ell)\end{bmatrix}^T$.

In [4] it was observed that the beam patterns of bilateral beamformers steered towards the front of the array are similar to those of the reference channels over a wide range of angles, which may be expected to help natural binaural localization. The time-invariant steering vectors for the bilateral beamformers are therefore given by

$$\mathbf{d}_l = \begin{bmatrix}H_1(\Psi) \; H_3(\Psi) \; \cdots \; H_{M-1}(\Psi)\end{bmatrix}^T / H_1(\Psi) \tag{6}$$

$$\mathbf{d}_r = \begin{bmatrix}H_2(\Psi) \; H_4(\Psi) \; \cdots \; H_M(\Psi)\end{bmatrix}^T / H_2(\Psi) \tag{7}$$

with $\Psi_l(\ell) = \Psi_r(\ell) = \Psi = (90°, 0°)$. The loss in beamformer performance due to mismatch between the source DOA and the bilateral beamformers' look direction is not substantial (see Section 4.3).

3.2. Mask estimation

The target for the mask estimation algorithm is a modified version of an oracle mask presented in [16] which we term the high-resolution stochastic WSTOI-optimal binary mask (HSWOBM). The HSWOBM optimizes a version of the WSTOI intelligibility metric [17] with increased frequency resolution for a stochastic noise signal with a known power spectrum. This high resolution version of weighted STOI (WSTOI) is identical to WSTOI except that the correlation comparison is performed on modulation vectors computed in bands which use the full STFT frequency resolution.

The power compressed (15th root) cochleagram feature set from [18], modified to have 90 frequency channels with centers from 80 to 5000 Hz, is used as the input feature set for the DNN. This feature set is computed with 25.6 ms frames centered on the mask bins. The mask estimation procedure follows that of [18]: features within a sliding window of length 13 frames are concatenated to form the inputs of a DNN which simultaneously estimates all of the mask values within a 5 frame window centered on the same frame. The feature and estimation windows are incremented by 1 frame and the procedure repeated, so that each mask value is estimated 5 times. These estimates are averaged to obtain the overall mask estimate, $B(\nu, \ell)$, for frequency bin ν and time-frame ℓ.

The DNN has $90 \times 13 = 1170$ units in the input layer, 4 hidden layers each with 2000 rectified linear units, and $129 \times 5 = 645$ sigmoidal units in the output layer. Stochastic gradient descent with Nesterov momentum is used with a minibatch size of 1000, a learning rate of 0.1 and a momentum of 0.9. Dropout with a ratio of 0.2 is used for the input layer and 0.5 for all hidden layers. The loss function for training the DNN is the mean square error in the estimated mask with two time-frequency (TF) weightings applied. The first is the speech intelligibility index band importance [19]. For each TF cell, the second weighting is proportional to the reduction in high-resolution WSTOI [17] that results from inverting the mask in that cell of the oracle HSWOBM (i.e. the effect of a single error in the estimated mask).

3.3. Speech enhancement

In contrast to most mask-based enhancers, the estimated binary mask $B(\nu, \ell)$ is not applied directly to the noisy speech but is instead used to supply prior information to a classical speech enhancer [13] about the probability of speech presence in different TF regions using the approach from [12], which applies a gain function $G(\nu, \ell)$ to frequency bin ν of frame ℓ. The approach of [12] is modified to accommodate an estimated mask $B(\nu, \ell)$ that is continuous-valued by replacing the equations in Sec. 2.2 of [12] by $q(\nu, \ell) = Q^0 + \left(Q^1 - Q^0\right) B(\nu, \ell)$ and $G_{min} = G^0 + \left(G^1 - G^0\right) B(\nu, \ell)$.

To suppress artifacts during periods where no speech is detected a modified gain

$$G'(\nu, \ell) = \begin{cases} 0 & B(\nu, \ell) < 0.1 \\ G(\nu, \ell) & B(\nu, \ell) \geq 0.1 \end{cases}$$

is applied, rather than $G(\nu, \ell)$ directly.

4. SIMULATION EXPERIMENTS

The performance of the proposed system was evaluated in simulated experiments for a static sound source with continuously rotating and static array orientations. Using a free-field model of sound progation, as in (2), allows frame-based processing to be used even during rotation [20] while a diffuse noise model ensures that the noise covariance matrix is independent of the rotation angle. Our analysis is therefore concerned with the effectiveness of the mask-based enhancement and in particular the impact of steering the reference beamformer to match the DOA of the desired source with respect to the array.

4.1. Signal generation

In each trial the microphone signals consisted of 10 s of desired speech from a single target source, mixed with diffuse noise and spatially white sensor noise, as in (1). The desired speech source was located at $\underline{\Omega} = (90°, 30°)$, i.e. on the horizontal plane and offset towards the left of the listener's nominal position. The performance of the proposed system was evaluated under continuous sinusoidal rotation of the array between $\pm 30°$ with period 1 s. For comparison, three static orientations of the array, with yaw angle $30°$, $0°$ and $-30°$, were also evaluated. Note that these test conditions lead to DOAs of the desired source with respect to the rotated array in the range $0° - 60°$.

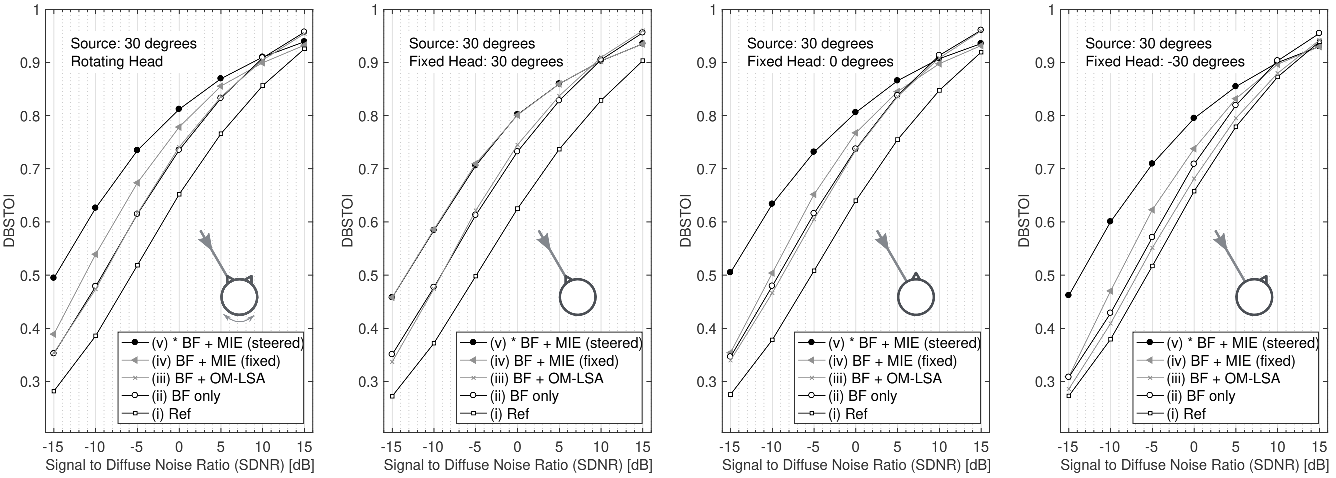
The desired speech propagation was modeled as a single plane wave as in (2a) while diffuse noise was modeled by discretizing the integral in (2b) as a summation over a 312-direction quadrature grid, with an independent identically distributed (iid) source impinging from each direction. Microphone signals for each individual plane-wave source were generated with fast convolution using Hamming-windowed signal frames of length 2 ms overlapping by 50%.

In each frame, $\mathbf{d}_O(\nu, \ell)$ was recalculated according to the current DOA with respect to the rotated array. To avoid the possibility of introducing interpolation errors, as would be the case for measured impulse responses, a binaural hearing aid array was simulated using an analytical model of the pressure field around a rigid sphere [21, 22] of radius 9 cm. The analytical model uses a spherical harmonic series whose truncation order was determined as in [23] to ensure the worst case error was less than -80 dB. The microphone array, consisting of 2 microphones at each ear with 2 cm separation, had a rectangular configuration. Placing the origin of our coordinates system at the centre of the sphere, with positive x, y and z-axes pointing forwards, left and up, respectively, the positions of the four microphones, prior to any rotations, were all on the horizontal plane at a radius of 10 cm, which corresponds to $\underline{\phi} \in \{\pm 84.3°, \pm 95.7°\}$. The number of frequency bins in $H_m(\nu, \underline{\Omega})$ was determined by the length of the filters in the time domain required given the array radius and the speed of sound.

Speech samples from the test partition of the TIMIT database [24] were concatenated to obtain a total duration of 10 s. For spherically isotropic diffuse noise, independent realizations of long term average speech spectrum (LTASS) noise [25] were generated for each direction and weighted according to the quadrature grid. Spatially white noise was generated as independent realizations of white Gaussian noise for each microphone.

The A-weighted level of the free-field desired speech at the origin was normalized according to [26, 27] while the power of the diffuse noise and spatially white noise were mixed to each obtain a particular A-weighted SNR, denoted signal to diffuse noise ratio (SDNR) and signal to spatially-white noise ratio (SWNR), respectively. The overall test set comprised 4 speech utterances (two males, two female) for each of 4 head rotation conditions (fixed at $30°$, $0°$ or $-30°$ or rotating sinusoidally between $\pm 30°$) and 7 SDNRs (-15 to 15 dB in 5 dB increments), giving a total of 112 trials. The SWNR was fixed at 30 dB.

To assess the contribution of the estimated mask to the proposed system we also evaluate the system performance when the 'Apply Mask' block in Fig. 1 is replaced with the unmodified optimally-modified log-spectral amplitude (OM-LSA) algorithm [28]. To assess the contribution of head-tracking to the proposed system we also evaluate the system performance with $\Psi_O = (90°, 0°)$.



(a) (b) (c) (d)

Fig. 2. Effect of processing on predicted binaural speech intelligibility for (a) rotating head and (b-d) fixed head orientations. The algorithms shown are (i) Ref: unprocessed reference microphones, (ii) BF only: $\tilde{X}_l(\cdot)$ and $\tilde{X}_r(\cdot)$, (iii) BF + OM-LSA: As (ii) with OM-LSA post process, (iv) BF + MAE (fixed): As (v) but reference beamformer is *not* steered towards the source, (v) Proposed system as shown in Fig. 1. Black lines with open symbols represent intermediate signals in the proposed system. Grey lines represent alternative processing strategies.

4.2. Implementation

The sampling frequency was 10 kHz. Beamforming was performed using overlap-add convolution with 20 ms Hamming-windowed frames overlapping by 50%. Beamformer weights were calculated as described in Sec. 3.1 with $\mathbf{R}_\alpha$ determined numerically using 1 s of spherically isotropic uncorrelated white Gaussian noise (WGN) from each direction on a 18×36 equiangular quadrature grid.

The DNN was trained with 3296 utterances from the training partition of the TIMIT database [24] combined with babble and speech shaped noise from the RSG.10 database [29]. The noisy utterances had average SNRs of $\{-2.7,\ -1.8,\ -0.6,\ 1.1\}$ dB for babble noise and $\{-4.0,\ -3.0,\ -1.7,\ 0.2\}$ dB for speech shaped noise. The target HSWOBM were optimized for WGN at -5 dB SNR. The parameters from [12] were used for the mask-informed enhancer.

4.3. Results and discussion

Fig. 2 shows the binaural STOI (DBSTOI) metric [30] for each test condition as a function of SDNR. A shift to the left of a performance curve represents an improvement in predicted intelligibility, which can be quantified in terms of the equivalent improvement in SDNR. This approach avoids the need to employ a non-linear mapping between DBSTOI values and predicted intelligibility, which may be a function of the specific listening conditions.

Fig. 2(a) shows performance of the compared methods with a sinusoidally rotating head. Using bilateral beamformers alone (indicated (ii) in the plot legends) gives an inprovement of 3-4 dB compared to the unprocessed case (indicated (i)). With our proposed mask-informed enhancement (v) performance is further improved over (ii) by 4–6 dB for SDNRs of -15–0 dB. At higher SDNRs the DBSTOI metric for the proposed method plateaus, whereas the beamformer alone continues to improve. In contrast, regardless of SDNR, OM-LSA without our proposed modifications (iii) offers no improvement over the bilateral beamformers alone (ii). It can therefore be concluded that the proposed mask is making a substantial improvement and is best suited to relatively poor SNR conditions.

Figs. 2(b-d) show the DBSTOI metric under static conditions for differerent head orientations. Consistent with pschoacoustic studies [31, 32], the unprocessed binaural reference signals (i) become more intelligible when the head is oriented away from the source. As for the rotating condition, (iii) offers no improvement over (ii). For the bilateral beamformer alone (ii) when the head is oriented away from the source (Figs. 2(c,d)) there is only a small reduction in peformance (1–2 dB) due to the mismatch of the beamformer look direction and the source DOA. However, after the proposed mask-informed enhancement the system performance is no worse for head rotation of -30° compared to 30°, justifying the use of fixed bilateral beamformers.

Comparing (iv) to (v) in Figs. 2(c,d) it can be seen that fixing $\Psi_O = (90^\circ, 0^\circ)$, i.e. not steering the reference beamformer towards the target, degrades intelligibility by 2.5–5 dB, depending on the input SDNR.

5. CONCLUSIONS

A novel signal processing approach for hearing aids has been proposed which exploits head-tracking information, target speaker mask estimation and mask-informed binaural speech enhancement. For an illustrative case with the listener's head oriented 30° away from the desired source, the benefit of the proposed mask-informed enhancement over bilateral beamformers alone is equivalent to a 6 dB improvement in SNR. The contribution of steering the reference beamformer towards the desired source accounts for between 2.5 and 5 dB. In contrast, the classical speech enhancement approach without our mask provides no measured intelligibility improvement.

6. 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. Ray Liu, Eds. John Wiley & Sons, Inc., 2008.

[2] S. M. Golan, S. Gannot, and I. Cohen, "A reduced bandwidth binaural MVDR beamformer," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2010.

[3] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2006.

[4] A. Kuklasinski and J. Jensen, "Multichannel Wiener Filters in binaural and bilateral hearing aids—Speech ntelligibility improvement and robustness to DoA errors," *J. Audio Eng. Soc. (AES)*, vol. 65, no. 1/2, pp. 8–16, 2017.

[5] T. J. Klasen, T. V. den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[6] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[7] E. Hadad, S. Doclo, and S. Gannot, "A generalized binaural MVDR beamformer with interferer relative transfer function preservation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2016, pp. 1643–1647.

[8] M. Zohourian, A. Archer-Boyd, and R. Martin, "Multi-channel speaker localization and separation using a model-based GSC and an inertial measurement unit," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5615–5619.

[9] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 430–434.

[10] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[11] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 892–902, Aug. 2014.

[12] L. Lightburn, E. De Sena, A. H. Moore, P. A. Naylor, and M. Brookes, "Improving the perceptual quality of ideal binary masked speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[13] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[14] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[15] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[16] L. Lightburn and M. Brookes, "SOBM - a binary mask for noisy speech that optimises an objective intelligibility metric," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5078–5082.

[17] ——, "A Weighted STOI Intelligibility Metric based on Mutual Information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[18] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, pp. 2604–2612, 2016.

[19] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute (ANSI)," ANSI Standard, 1997.

[20] V. Tourbabin and B. Rafaely, "Analysis of distortion in audio signals introduced by microphone motion," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2016, pp. 998–1002.

[21] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. San Diego, California, USA; London, UK: Academic Press, 1999.

[22] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing, 2017.

[23] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 193–204, 2014.

[24] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Nat. Inst. of Standards and Tech. (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[25] "Artificial voices," Intl. Telecommunications Union (ITU-T)," Standard, Sept. 1999.

[26] "Objective measurement of active speech level," Intl. Telecommunications Union (ITU-T)," Recommendation, Mar. 1993.

[27] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997–2016. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[28] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[29] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise data-base," TNO Institute for Perception, Tech. Rep., 1988.

[30] A. H. Andersen, J. M. d. Haan, Z. H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[31] W. E. Kock, "Binaural Localization and Masking," *J. Acoust. Soc. Am.*, vol. 22, no. 6, pp. 801–804, Nov. 1950.

[32] J. A. Grange and J. F. Culling, "The benefit of head orientation to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 703–712, Feb. 2016.