Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST-Priority 2

DELIVERABLE NO: D1.5
Head-Geometry Beamformer with Binaural Output
and its Perceptual Quality Assessment

Date of deliverable: 31.12.2007
Actual submission date: 06.02.2008

Start date of project: 01.01.2006 Duration: 60
Organization name of lead contractor for this deliverable: **Carl von Ossietzky University Oldenburg**

Revision 1

Project co-funded by the European Commission within the Sixth Framework Program (2002 - 2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D1.5 Head-Geometry Beamformer with Binaural Output and its Perceptual Quality Assessment (OL)

Thomas Rohdenburg, Volker Hohmann, and Birger Kollmeier

February 6, 2008

Abstract

In this contribution different microphone array-based noise reduction schemes for hearing aids are suggested and compared in terms of their performance, signal quality and robustness against model errors. The algorithms all have binaural output and are evaluated using objective perceptual quality measures [17, 18, 21]. It has been shown earlier that these measures are able to predict subjective data that is relevant for the assessment of noise reduction algorithms. The quality measures showed clearly that fixed beamformers designed with head models were relatively robust against steering errors whereas for the adaptive beamformers tested in this study the robustness was limited and the benefit due to higher noise reduction depended on the noise scenario and the reliability of a direction of arrival estimation. Furthermore, binaural cue distortions introduced by the different binaural output strategies could be identified by the binaural speech intelligibility measure [21] even in case monaural quality values were similar. Thus, this perceptual quality measure seems to be suitable to discover the benefit that the listener might have from the effect of spatial unmasking.

Contents

1 Introduction

In modern hearing aids multi-channel noise reduction schemes based on small microphone arrays are used for speech enhancement. These algorithms exploit the spatial configuration of the interfering signals and therefore generally lead to less signal distortion and higher noise reduction than single-channel envelope filters. The human ability to separate sound sources in a complex situation, namely the cocktail-party effect, partly arises from the use of binaural localization cues. If binaural information is lost or distorted by the processing, the hearing impaired listener may not make use of the effect of spatial unmasking as efficiently as in the undistorted binaural condition. The intelligibility improvement introduced by a spatial filter is counteracted by the decrease due to the deteriorated efficiency of the spatial unmasking in this case. Although bilateral supply with hearing aids is motivated by a better directional-hearing ability, it has been shown in [1] that binaural cues are distorted if the hearing aids at the left and right ears work independently. Therefore, researchers have suggested microphone array based binaural spatial filtering techniques [2, 3, 4, 5] that assume a connection between the left and right hearing aid. In this study we analyzed fixed and adaptive beamformer algorithms, that exploit *a priori* knowledge about array position, wave propagation and direction of arrival as these seem to be slowly varying parameters that can be estimated and used for the adaptation of the algorithms. Information about the voice activity which might also be helpful for noise estimation was not used here. The beamformers that were calculated using the constrained minimum variance distortionless response (MVDR) design [6] had single channel outputs that were extended by a binaural stage. Three different strategies for generating a binaural output have been applied and evaluated by perceptual measures. Furthermore, the robustness of fixed and adaptive beamformers using different propagation models has been analyzed against steering error, array position and head-size mismatch by appropriate perceptual quality measures.

2 Acoustical Setup

Figure 1 shows schematically the acoustical setup and the coordinate system used for defining microphone positions and sound source directions. 6-channel signals ($M = 6$) have been recorded from two 3-channel behind-the-ear (BTE) hearing aid shells (Siemens Acuris) mounted on a Brüel & Kjær (B&K) head and torso simulator (HATS). The impulse responses (IRs) for all microphones have been measured with this setup in an anechoic room for azimuth directions 0-180° in 5° steps at an elevation of 0° (horizonal plane). In the following these are referred to as 6-channel head related transfer functions (HRTFs) in the frequency domain that include head-shadow and diffraction effects, and the characteristics of the microphones. Similarly, HRTFs have been measured in an office environment (reverberation time $\tau_{60} = 300$ ms). Directional target speech and interfering noise signals were calculated by filtering source signals with these HRTFs. In addition, real-world environmental noise has been recorded in a cafeteria and in an office room. Furthermore, an artificial diffuse noise has been generated by filtering a speech-colored random noise with the anechoic HRTFs from all directions and summing up all filtered noise signals. This signal simulates a cylindrical 2D-isotropic noise field. From the database of 6-channel directional speech and noise signals various mixtures have been calculated for different signal-to-noise ratios (SNRs). For condition 1) the input signal was composed from two directional signals filtered with HRTFs (target and interferer from 30° (front-left) and -135° (back-right) azimuth, respectively) and mixed with the recorded cafeteria noise to generate a near-to-realistic scenario. For condition 2) we used only one directional signal (speaker from 30° (left)) mixed with an artificial diffuse noise. The 30° direction was chosen because it is asymmetric to the array and offers a more general assessment of the beamformers properties than a fixed 0° look direction.

3 Algorithm

Figure 2 shows the block diagram of the noise reduction scheme which will be described in the following. Note that the algorithm is not limited to the 6-channel setup used here but applies to any M-channel

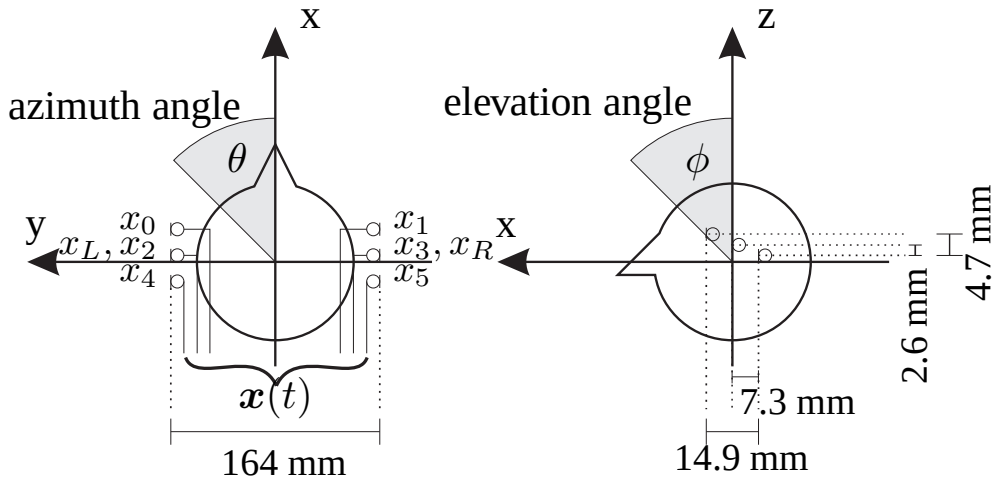


Figure 1: Acoustical setup: Two linear microphone arrays are mounted bilaterally on a B&K HATS. Each array consists of 3 hearing aid microphones mounted in a hearing aid shell with a distance of ca. 8 mm. The frontal direction is the x-axis which is equal to an azimuth angle $\theta = 0°$ and an elevation angle $\phi = 90°$.

microphone array mounted near to a head. Throughout the paper, vectors and matrices are printed in boldface, scalars in italics. t denotes the time, ω the radian frequency and k the block-index. The superscripts ^{T}, * and ^{H} denote the transposition, the complex conjugation and the Hermitian transposition, respectively.

3.1 Signal model

The multi-channel signal $\boldsymbol{x(t)} = [x_0(t), x_1(t), \ldots, x_{M-1}(t)]^T$ (Fig. 1,2) is assumed to be a mix of the directional signal $\boldsymbol{s}(t)$ and a noise signal $\boldsymbol{n}(t)$. In the frequency domain the signal model can be formulated as

$$\boldsymbol{X}(\omega, k) \quad = \quad \underbrace{\boldsymbol{d}_S(\omega) S(\omega, k)}_{\boldsymbol{S}(\omega,k)} + \boldsymbol{N}(\omega, k) \tag{1}$$

where the capital letters denote the time-frequency transformed signals of $\boldsymbol{x}, \boldsymbol{s}$, and $\boldsymbol{n}$ calculated by a short time Fourier transform (STFT). The propagation vector $\boldsymbol{d}_S(\omega) = \boldsymbol{d}(\omega, \theta_S, \phi_S)$ is the vector of transfer functions between the source signal $S(\omega)$ and the signal vector $\boldsymbol{S}(\omega)$ observed at the sensors. In general, the propagation vector for a signal source coming from the azimuth angle θ and the elevation angle ϕ is

$$\boldsymbol{d}(\omega, \theta, \phi) \quad = \quad [d_0(\omega, \theta, \phi), d_1(\omega, \theta, \phi), \ldots, d_{M-1}(\omega, \theta, \phi)]^T \tag{2}$$

where the transfer function to a microphone $i = 0 \ldots M-1$ is

$$d_i(\omega, \theta, \phi) \quad = \quad a_i(\omega, \theta, \phi) e^{-j\omega\tau_i(\omega,\theta,\phi)} \tag{3}$$

with the amplitude spectrum $a_i(\omega, \theta, \phi)$ and the group-delay $\tau_i(\omega, \theta, \phi)$.

3.2 Beamformer

A fixed filter-and-sum beamformer can be designed in the frequency domain to produce a monaural output that contains less noise energy than the multi-channel input signal $\boldsymbol{X}$ by

$$Y_f(\omega, k) \quad = \quad \sum_{i=0}^{M-1} W_i^*(\omega) X_i(\omega, k) = \boldsymbol{W}^H(\omega) \boldsymbol{X}(\omega, k). \tag{4}$$

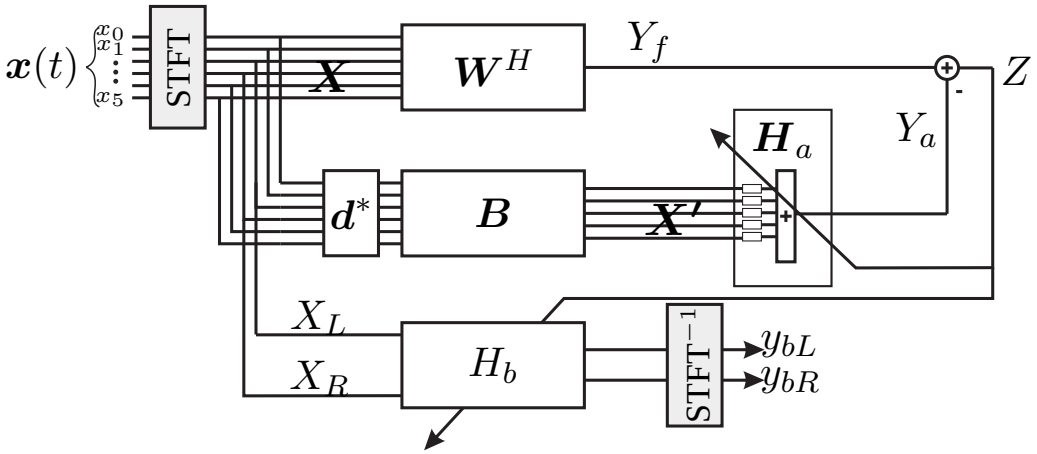


Figure 2: Multi-channel beamformer system with binaural output. $\boldsymbol{W}^H$ is the fixed beamformer filter, $\boldsymbol{B}$ denotes the blocking matrix, $\boldsymbol{H}_a$ is the adaptive filter, and H_b is the filter that generates a binaural output from the reference microphone signals $X_2(= X_L)$ and $X_3(= X_R)$ at the left and right ear.

The optimal filter $\boldsymbol{W}$ can be calculated by the well-known Minimum Variance Distortionless Response (MVDR) solution [6]:

$$\boldsymbol{W}(\omega,\theta,\phi) \quad = \quad \frac{\boldsymbol{\Phi_{NN}}^{-1}(\omega)\boldsymbol{d}(\omega,\theta,\phi)}{\boldsymbol{d}^H(\omega,\theta,\phi)\boldsymbol{\Phi_{NN}}^{-1}(\omega)\boldsymbol{d}(\omega,\theta,\phi)} \tag{5}$$

where $\boldsymbol{\Phi_{NN}}^{-1}$ denotes the inverse noise correlation matrix which is discussed in 3.4.2.

The fixed beamformer can be extended by an adaptive noise cancelation path which consists of a delay- (and amplitude-) compensation step, denoted by the delay compensation vector $\boldsymbol{p}$, followed by a blocking matrix $\boldsymbol{B}$ (producing the noise reference $\boldsymbol{X'}$) and an a multi-channel Wiener filter that is adapted to cancel out noise components that $\boldsymbol{X'}$ and Y_f have in common. The (element-wise) Hadamard product of the delay compensation vector $\boldsymbol{p}$ and the propagation vector $\boldsymbol{d}$ should result in a zero-delay vector with amplitude 1:

$$\boldsymbol{p} \bullet \boldsymbol{d} \quad = \quad \mathbf{1} = [1,\ldots,1]^T \tag{6}$$

Thus, $\boldsymbol{p}$ is defined by

$$\boldsymbol{p}(\omega,\theta,\phi) \quad = \quad \left[\frac{{d_0}^*(\omega,\theta,\phi)}{|d_0(\omega,\theta,\phi)|^2}, \frac{{d_1}^*(\omega,\theta,\phi)}{|d_1(\omega,\theta,\phi)|^2}, \ldots, \frac{{d_{M-1}}^*(\omega,\theta,\phi)}{|d_{M-1}(\omega,\theta,\phi)|^2}\right]^T \tag{7}$$

and the blocking matrix (which is a $[M-1 \times M]$ - subtraction matrix) is [6]

$$\boldsymbol{B} \quad = \quad \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & 1 & -1 \end{bmatrix} \tag{8}$$

the noise reference matrix $\boldsymbol{X'}$ at the output of the blocking matrix is:

$$\boldsymbol{X'}(\omega,k) \quad = \quad \boldsymbol{B}(\boldsymbol{p}(\omega,\theta,\phi) \bullet \boldsymbol{X}(\omega,k)) \tag{9}$$

The multi-channel Wiener filter is designed with

$$\boldsymbol{H}_a(\omega) \quad = \quad \boldsymbol{\Phi_{X'X'}}^{-1}(\omega)\boldsymbol{\Phi_{X'Y_f}}(\omega) \tag{10}$$

where the PSD-matrix $\boldsymbol{\Phi}_{\boldsymbol{X'X'}}$ and the cross-PSD row vector $\boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}$ denote expectation values defined by

$$\boldsymbol{\Phi}_{\boldsymbol{X'X'}}(\omega) = \mathrm{E}\left\{\boldsymbol{X'}(\omega)\boldsymbol{X'}^{H}(\omega)\right\} \quad (11)$$

$$\boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}(\omega) = \mathrm{E}\left\{\boldsymbol{X'}(\omega)Y_f^{*}(\omega)\right\} \quad (12)$$

In practice, $\boldsymbol{\Phi}_{\boldsymbol{X'X'}}$ and $\boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}$ are calculated by recursively averaging instantaneous short-time spectra:

$$\boldsymbol{\Phi}_{\boldsymbol{X'X'}}(\omega,k) = \alpha\boldsymbol{\Phi}_{\boldsymbol{X'X'}}(\omega,k-1) + (1-\alpha)\boldsymbol{X'}(\omega,k)\boldsymbol{X'}^{H}(\omega,k) \quad (13)$$

$$\boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}(\omega,k) = \alpha\boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}(\omega,k-1) + (1-\alpha)\boldsymbol{X'}(\omega,k)Y_f^{*}(\omega,k) \quad (14)$$

Therefore, also the filter $\boldsymbol{H}_a$ is slowly varying over time and the noise estimate of the adaptive path, Y_a, is calculated by

$$Y_a(\omega,k) = \boldsymbol{H}_a^{H}(\omega,k)\boldsymbol{X'}(\omega,k) \quad (15)$$

which then can be subtracted from the fixed beamformer output so that we get the monaural output of the Generalized Sidelobe Canceller (GSC):

$$Z(\omega,k) = Y_f(\omega,k) - Y_a(\omega,k) \quad (16)$$

In summary, we get the monaural outputs of the two beamformer types:

$$\text{fixed}: \quad Z(\omega,k) = Y_f(\omega,k) = \boldsymbol{W}^{H}(\omega)\boldsymbol{X}(\omega,k) \quad (17)$$

$$\text{adaptive}: \quad Z(\omega,k) = \boldsymbol{W}^{H}(\omega)\boldsymbol{X}(\omega,k) - \boldsymbol{H}_a^{H}(\omega)\boldsymbol{X'}(\omega,k) \quad (18)$$

Thus, the difference between fixed and adaptive beamformer consists of an additional noise subtraction path which can be added to the fixed beamformer. Note, that the original GSC [7] uses a standard delay-and-sum (D&S) beamformer in the fixed processing path, whereas we use an arbitrary superdirective design here, which is discussed below.

3.3 Binaural output

The output can be extended to a binaural signal with left and right output signal Y_{bL} and Y_{bR}

$$\boldsymbol{Y_b}(\omega,k) = [Y_{bL}(\omega,k), Y_{bR}(\omega,k)]^{T} \quad (19)$$

with different strategies.

3.3.1 Target signal phase reconstruction

The simplest solution might be to reconstruct the phase and amplitude response of the target signal by multiplying the monaural output with the propagation coefficients d_L, d_R that relate to the reference microphones (denoted as x_L and x_R in Fig. 1) at the left and right hearing aid array, respectively:

$$Y_{bL}(\omega,k) = d_L(\omega,\theta,\phi)Z(\omega,k) \quad (20)$$

$$Y_{bR}(\omega,k) = d_R(\omega,\theta,\phi)Z(\omega,k) \quad (21)$$

However, this can only reconstruct the gross magnitude and phase characteristic of the target signal that is included in the assumed propagation model whereas the binaural information of the interfering noise signal is lost.

3.3.2 Binaural post-filter

A method to preserve the phase of both, signal and noise, can be realized according to [2] by applying a real-valued time-varying post-filter to the reference microphone signals X_L, X_R:

$$H_b(\omega,k) = \frac{\left(|d_L(\omega,\theta,\phi)|^2 + |d_R(\omega,\theta,\phi)|^2\right)\Phi_{ZZ}(\omega,k)}{\Phi_{X_LX_L}(\omega,k)+\Phi_{X_RX_R}(\omega,k)} \quad (22)$$

$$Y_{bL}(\omega,k) = H_b(\omega,k)X_L(\omega) \quad (23)$$

$$Y_{bR}(\omega,k) = H_b(\omega,k)X_R(\omega) \quad (24)$$

$\Phi_{ZZ}, \Phi_{X_LX_L}$ and $\Phi_{X_RX_R}$ denote the power spectral density estimates for the signals Z, X_L, X_R, respectively. In practice, these can be estimated by recursively smoothing instantaneous signal powers. The binaural post-filter can be interpreted as a single-channel envelope Wiener filter applied to both reference channels X_L, X_R. Additional gain rules known from single channel noise reduction systems can be applied here.

3.3.3 Bilateral Beamformer

To investigate the behavior of two independently working unilateral beamformers $\boldsymbol{W_L}$ (left) and $\boldsymbol{W_R}$ (right), the system depicted in Figure 2 can be split into two subarrays where $\boldsymbol{X_L} = [X_0, X_2, \ldots X_{M-2}]$ denotes the signal matrix of the left subarray using the even-numbered microphones and $\boldsymbol{X_R} = [X_1, X_3, \ldots X_{M-1}]$ denotes the signal of the right subarray using the odd-numbered microphones. $\boldsymbol{X'_L}, \boldsymbol{X'_R}$ are defined according to (9) but for shorter blocking matrices and delay compensation vectors $\boldsymbol{p_L}, \boldsymbol{p_R}$, respectively.

$$Y_{bL}(\omega,k) = Z_L(\omega,k) = \boldsymbol{W_L}^H(\omega)\boldsymbol{X}_L(\omega,k) - \boldsymbol{H}^H_{a_L}(\omega)\boldsymbol{X'}_L(\omega,k) \quad (25)$$

$$Y_{bR}(\omega,k) = Z_R(\omega,k) = \boldsymbol{W_R}^H(\omega)\boldsymbol{X}_R(\omega,k) - \boldsymbol{H}^H_{a_R}(\omega)\boldsymbol{X'}_R(\omega,k) \quad (26)$$

The subarrays do not need to be restricted to one side but can use any combination of microphones from both sides if a connection between the bilateral arrays exists. In the case of a complete bilaterally connected system every filter gets the complete M-channel information. However, in this case additional constraints have to be included into the beamformer design to partially reconstruct the binaural information of the target and noise signal. A detailed analysis on such binaural systems for two microphones can be found in [5] and for six microphones in [8].

In summary, three different methods that produce a binaural output can be distinguished. In the following, the signal phase reconstruction method is denoted as (BIN_PR), the binaural post-filter as (BIN_PF), and the bilateral system using only the left (respectively, right) subarray is denoted as (BIN_BL).

3.4 Influence of different propagation models on the beamformer design

The fixed beamformer coefficients given by (5) ideally reduce a noise field[1] with the correlation matrix $\boldsymbol{\Phi}_{NN}$ under the constraint of an undistorted signal response in the desired look direction. The more exactly $\boldsymbol{\Phi}_{NN}$ is known, the higher is the noise reduction performance. The absence of distortion for the MVDR beamformer, however, is only given if the propagation model $\boldsymbol{d}$ used for the beamformer design and the true signal wave propagation vector $\boldsymbol{d}_S$ perfectly match. In general, the exact transfer functions $\boldsymbol{d}_S$ are unknown and several assumptions about the wave propagation must be made. The influences of the exactness of the propagation model on the beamformer performance are discussed below.

3.4.1 Propagation vector

All effects could be perfectly integrated into the beamformer design if the transfer functions $\boldsymbol{d}_S$ could be measured in the situation of interest, including the room response, the head-shadow and diffraction

[1] a superposition of many unknown noise signals

effects, and the microphone characteristics. However, as estimating the room response for a given target signal is not feasible under realistic conditions the second-best solution is measuring the anechoic transfer functions of the system including the head-influences and the microphone characteristics. It may be useful for the beamformer design to normalize the measured anechoic HRTFs to the transfer function of a left/right reference microphone for the target direction $\theta_S = 0°$ [9], because the aim is not to reconstruct the target S itself but its corresponding signals observed at the two reference microphones at the left and right ear. To establish a reference propagation model these normalized HRTFs are directly used as a propagation vector $\boldsymbol{d}$ in (2). This model will be referred to as HRTF in the following.

If the anechoic HRTF is not available, the gross head-shadow and diffraction effects can be modeled by the wave propagation observed on a rigid sphere [10, 11]. For head-models, both, a_i and τ_i in (3) are angle and frequency dependent. In general, it is assumed that the target source is approximately in the horizontal plane, i.e., $\phi_S \approx 90°$. Therefore, the elevation angle ϕ_S will be disregarded in the following for the head-related wave propagation models used in this study. The first head model (HM1) by [11] is a simple and effective parametric model that estimates the characteristics of a sphere. The interaural time difference (ITD) cues are modeled by Woodworth and Schlosberg's frequency independent (ray-tracing) formula. The gross magnitude characteristics of the HRTF spectrum, namely the interaural level difference (ILD) cues, are covered by a single-pole, single-zero head-shadow filter which also accounts for an additional frequency dependent delay at low frequencies. For each microphone of the array an angle of a ray from the center of the sphere to the microphone θ_i, $i = 0 \ldots M-1$, can be calculated. Choosing the angle to the desired sound source θ_S and some additional model parameters (e.g. sphere radius $r = 8.2$ cm, speed of sound, fitting parameters $\alpha_{min}, \theta_{min}$, see [11]), the transfer function is calculated by

$$\boldsymbol{d}(\omega, \theta_S) \quad = \quad [H_{HM1}(\omega, \theta_S, \theta_0, \text{params}), \ldots, H_{HM1}(\omega, \theta_S, \theta_{M-1}, \text{params})]^T \tag{27}$$

The second head model (HM2) [10] additionally incorporates the distance of the source for modeling near-field effects and interference effects that introduce ripples in the response that are quite prominent on the shadowed side. It is numerically calculated by a recursive algorithm given in [10]. The propagation vector is built similar to HM1 (27)

The far-field assumption implies that all microphones *see* the target sound wave arriving from the same angles (θ_S, ϕ_S) as a planar wave. Additionally assuming free-field (FF), i.e., no objects inside the sound wave path and a unity microphone response $a_i(\omega, \theta, \phi) = 1$, $\forall(\omega, \theta, \phi, i)$, the propagation coefficient (3) simplifies to

$$\boldsymbol{d}(\omega, \theta_S, \phi_S) \quad = \quad \left[e^{-j\omega\tau_{00}(\theta_S,\phi_S)}, \ldots, e^{-j\omega\tau_{0M-1}(\theta_S,\phi_S)}\right]^T \tag{28}$$

where τ_{0i} is a constant group delay measured between a reference microphone 0 and microphone i. The group delay can easily be calculated based on the microphone array geometry where $\boldsymbol{l}_{0i}$ is the vector difference between a reference microphone 0 and the microphone i, c is the speed of sound, and $\boldsymbol{e}_r(\theta_S, \phi_S) = [\sin(\theta_S)\cos(\phi_S), \sin(\theta_S)\sin(\phi_S), \cos(\phi_S)]^T$ is the unit vector in target direction:

$$\tau_{0i}(\theta_S, \phi_S) \quad = \quad \frac{\boldsymbol{l_{0i}}^T \boldsymbol{e}_r(\theta_S, \phi_S)}{c} \tag{29}$$

Thus, under the FF assumption the beamformer can be designed knowing the relative microphone positions and the direction of the target signal.

3.4.2 Noise correlation matrix

The *normalized* cross power spectral density matrix of the noise is defined as

$$\boldsymbol{\Phi_{NN}}(\omega) \quad = \quad \frac{1}{\overline{\Phi}_{NN}(\omega)} \begin{pmatrix} \Phi_{N_0N_0}(\omega) & \Phi_{N_0N_1}(\omega) & \ldots & \Phi_{N_0N_{M-1}}(\omega) \\ \Phi_{N_1N_0}(\omega) & \Phi_{N_1N_1}(\omega) & \ldots & \Phi_{N_1N_{M-1}}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{N_{M-1}N_0}(\omega) & \Phi_{N_{M-1}N_1}(\omega) & \ldots & \Phi_{N_{M-1}N_{M-1}}(\omega) \end{pmatrix} \tag{30}$$

where the normalization factor $\overline{\Phi}_{NN}(\omega)$ forces the trace of $\boldsymbol{\Phi_{NN}}$ to equal M. In the MVDR beamformer equation (5) the inverse of the noise correlation matrix, $\boldsymbol{\Phi_{NN}}^{-1}$, can be interpreted as a decorrelation of the noise components included in $\boldsymbol{X}$. The simplest noise model makes the assumption that the noise is already uncorrelated, i.e. no further decorrelation is needed, and therefore it has a correlation matrix

$$\boldsymbol{\Phi_{NN}}(\omega) = \boldsymbol{\Phi_{NN}} \quad = \quad \boldsymbol{I} = \boldsymbol{\Phi_{NN}}^{-1} \tag{31}$$

The optimal MVDR beamformer design for uncorrelated noise leads to a delay-and-sum (D&S)-beamformer (aka: *conventional* beamformer):

$$\boldsymbol{W}(\omega) \quad = \quad \frac{\boldsymbol{d}(\omega)}{\boldsymbol{d}^H(\omega)\boldsymbol{d}(\boldsymbol{\omega})} = \frac{\boldsymbol{d}(\omega)}{\sum_i a_i^2(\omega)} = \frac{1}{M}\boldsymbol{d}(\omega), \quad a_i(\omega) = 1\,\forall i \tag{32}$$

By summing up uncorrelated noise and correlated signal components the theoretical SNRE is $10\log_{10}(M)$ dB, i.e., ≈ 7.8dB for $M = 6$ Microphones. However, natural sound sources in general are spatially correlated and this knowledge can be used to design *superdirective* beamformers that have a higher directivity compared to conventional beamformers, especially for low frequencies. The correlation function of the noise depends on the frequency and the distance of the microphones. It can either be measured by long-term averaging the cross spectral densities $\Phi_{X_iX_k}$ between the microphones i, k during speech pauses, or estimated by using the same sound propagation model that is used for $\boldsymbol{d}$, which is shown in the following. The cross-spectral density of a signal Q arriving from azimuth angle θ as observed between the microphones i and k is

$$\begin{aligned}
\Phi_{X_iX_k}(\omega,\theta) &= \mathrm{E}\left\{Q(\omega)d_i(\omega,\theta)Q^*(\omega)d_k^*(\omega,\theta)\right\} && (33)\\
&= \Phi_{QQ}(\omega)d_i(\omega,\theta)d_k^*(\omega,\theta) && (34)\\
&= \Phi_{QQ}(\omega)a_i(\omega,\theta)a_k(\omega,\theta)e^{j\omega(\tau_i(\omega,\theta)-\tau_k(\omega,\theta))} && (35)
\end{aligned}$$

The noise cross-correlation matrix of all noise sources can be calculated as the sum of individual noise cross spectral densities arriving from different azimuth directions θ_v:

$$\Phi_{N_iN_k}(\omega) \quad = \quad \sum_{\theta_v} \Phi_{X_iX_k}(\omega,\theta_v) \tag{36}$$

If the directions of individual noise sources Q are unknown (which is mostly the case) the assumption of homogenously distributed sources is often made. Two typically used noise characteristics can be distinguished: 2D- or cylindrical isotropic noise which is a suitable model for rooms with comparatively high damping of ceiling and floor and 3D-isotropic or *diffuse* noise which is a model for noise sources homogenously distributed on a sphere, i.e., no preferred directivity. For free-field assumptions the noise models can be calculated analytically by solving the integral over an infinite number of noise sources from all directions. The characteristic of cylindrical isotropic noise is:

$$\Phi_{N_iN_k}(\omega) \quad = \quad J_0\left(\omega\frac{l_{ik}}{c}\right) \tag{37}$$

where J_0 is the zero-oder Bessel function of the first kind, l_{ik} the distance between the microphones i and k and c the speed of sound. Beamformers using this noise model can easily be modified for an optimal front-to-back ratio by adjusting the angle limits of the integral [6]. For spherically homogenous isotropic (diffuse) noise the integral over all azimuth and elevation angles leads to the well-known sinc-characteristic in free-field:

$$\Phi_{N_iN_k}(\omega) \quad = \quad \frac{\sin(\omega\frac{l_{ik}}{c})}{\omega\frac{l_{ik}}{c}} = \mathrm{sinc}\left(\omega\frac{l_{ik}}{c}\right) \tag{38}$$

However, for head-related systems these solutions of the integrals are not valid due to the more general definition of the propagation vector $\boldsymbol{d}$. In this case, the integrated HRTFs have been approximated by

summing over the propagation vectors from all direction using eq. (33)-(36). In summary, the different noise field models that were used are uncorrelated noise (uncorr), cylindrical isotropic noise (diff2D), spherical isotropic diffuse noise (diff3D), HRTF integrated noise (intHRTF) and long-term measured noise from real-world recordings (measured). For stability reasons of the beamformer design the noise correlations matrices have to be mixed with a certain amount of uncorrelated noise which is evaluated in section 5.

3.5 Algorithm combinations

The different propagation models, output types, and algorithm settings are summarized in Table 1. All combinations are possible and a subset of combinations was evaluated (see section 4).

Output Type	Wave Propagation Model $\boldsymbol{d}$	Noise field model $\boldsymbol{\Phi_{NN}}$	Beamformer type
BIN_PR	HRTF	uncorr	fixed
BIN_PF	HM2	diff2D	adaptive
BIN_BL	HM1	diff3D	
	FF	intHRTF	
		intHM2	
		measured	

Table 1: Algorithm combinations

4 Evaluation methods

For microphone arrays *signal- independent* measures exist to evaluate the theoretically performance to be expected for different noise field characteristics. These measures allow a rough estimate of the beamformer performance and are helpful for the numerical adjustment and optimization towards the desired system properties. In this study, modifications to existing measures that are suitable for head-worn systems are suggested and discussed below. For a more elaborate performance analysis, simulations with realistic signals, such as real-world recordings on a prototype array-system have to be done. The *signal-dependent* and *signal-independent* performance measures are described in sections 4.1 and 4.2.

4.1 Signal-independent performance measures and the influences of the head

4.1.1 Array gain

The array gain is a measure that shows the improvement of the SNR between the input signal of one sensor i and the output of the array. It is defined by

$$G_i(\omega) \quad = \quad \frac{\mathrm{SNR_{out}}(\omega)}{\mathrm{SNR_{in,i}}(\omega)} \tag{39}$$

If the input SNRs of all microphones ($\mathrm{SNR_{in,0}} \ldots \mathrm{SNR_{in,M-1}}$) are the same, then the array gain can be calculated for an arbitrary noise field $\boldsymbol{\Phi_{NN}}$ by [6]

$$G_i(\omega) = G(\omega) \quad = \quad \frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{NN}}(\omega)\boldsymbol{W}(\omega)} \tag{40}$$

If the beamformer coefficients $\boldsymbol{W}(\omega)$ are designed based on the true wave propagation vector $\boldsymbol{d}_S(\omega)$, the nominator in (40) equals 1, which means a distortionless response. However, if the propagation model $\boldsymbol{d}(\omega)$ in the beamformer design is changed or simplified compared to the true wave propagation $\boldsymbol{d}_S(\omega)$, the nominator shows the amount of signal distortion. The denominator that is to be minimized

by the beamformer shows the amount of noise reduction. For head-worn systems it is interesting to calculate the improvement compared to the source signal at the left and right ear position. Therefore, it is suggested to calculate the head-related array gain using the target signal power as observed at the reference microphones for the left and right head side:

$$G_L(\omega) = \frac{G(\omega)}{|d_L(\omega,\theta_S)|^2} \tag{41}$$

$$G_R(\omega) = \frac{G(\omega)}{|d_R(\omega,\theta_S)|^2} \tag{42}$$

Here d_L, d_R are the measured signal transfer functions to the left/right reference microphone, respectively. Note, that for free-field $G = G_L = G_R$.

4.1.2 White noise gain

The White Noise Gain (WNG) is a measure that shows the ability to reduce uncorrelated (i.e., spatially white) noise. Such noise can be associated to model errors, e.g., position, amplitude, phase errors, and self-noise of the microphones and is an important robustness measure for microphone arrays. If the WNG is small the beamformer is susceptible to uncorrelated noise (and model errors), i.e., such noise is increased rather than decreased. Thus, the WNG has to be limited to a minimum δ^2.

$$\mathrm{WNG}(\omega) = \frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{W}(\omega)} \geq \delta^2 \tag{43}$$

One of the most popular robust approaches to archive this is the diagonal loading algorithm [12, 13] :

$$\boldsymbol{W}(\omega,\theta,\phi) = \frac{(\boldsymbol{\Phi_{NN}}(\omega)+\mu(\omega)\boldsymbol{I})^{-1}\boldsymbol{d}(\omega,\theta,\phi)}{\boldsymbol{d}^H(\omega,\theta,\phi)(\boldsymbol{\Phi_{NN}}(\omega)+\mu(\omega)\boldsymbol{I})^{-1}\boldsymbol{d}(\omega,\theta,\phi)} \tag{44}$$

However, the choice of $\mu(\omega)$ that limits the WNG to a minimum of δ^2 is not simple. It can either be calculated in a multi-step iterative process [14] or via second order cone programming [12]. In this study, an iterative method is used and the importance of this constraint is studied based on the perceptual performance measures described in 4.2.

4.1.3 Directivity Index

The directivity index is a performance measure for directional microphones that shows the difference between target signal suppression and the suppression of noise coming from all directions, i.e., isotropic diffuse noise.

$$\mathrm{DI}(\omega) = 10\log_{10}\left(\frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{NN}}^{\mathrm{diffuse}}(\omega)\boldsymbol{W}(\omega)}\right) \tag{45}$$

To have a scalar performance value, the frequency dependent directivity index (DI) can be weighted by a band importance function a_k for speech perception taken from the articulation index [15, 16]. Thus, the sum over all bands k is

$$\mathrm{DI}_{\mathrm{AI}} = \sum_k a_k \mathrm{DI}(\omega_k) \tag{46}$$

4.1.4 Beampattern

The beampattern shows the array gain for noise signals arriving from different directions. Thus, in the denominator of (40) the noise correlation matrix $\boldsymbol{\Phi_{NN}}$ is replaced by the correlation matrix of a signal source in direction θ, ϕ with the assumed wave propagation $\boldsymbol{d}$:

$$\boldsymbol{\Phi_{DD}}(\omega,\theta,\phi) = \boldsymbol{d}(\omega,\theta,\phi)\boldsymbol{d}^H(\omega,\theta,\phi) \tag{47}$$

The beampattern is

$$|H(\omega,\theta,\phi)|^2 \quad = \quad -10\log_{10}\left(\frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_s(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi}_{\boldsymbol{DD}}(\omega,\theta,\phi)\boldsymbol{W}(\omega)}\right) \tag{48}$$

Note, that $|H(\omega,\theta,\phi)|^2 = 1$ only for the special case where $\boldsymbol{d}(\omega,\theta,\phi) = \boldsymbol{d}_S(\omega,\theta_S,\phi_S)$ accounting for a distortionless response.

Common visualizations of the beampattern are polar diagrams for specific frequencies ω or image plots (frequency over azimuth angle, color-coded intensity).

4.2 Signal-dependent performance measures

Signal-dependent performance measures allow for a more precise performance analysis especially if calculated on real-world recordings of typical acoustical scenes. For the performance measures used here, the separated desired signal and the noise signals have been processed with the same time-varying filters that have been calculated based on the mixture. This method, sometimes referred to as *shadow filtering*, is basically appropriate in simulation environments where the signal processing is disclosed. Given the target and the noise signals processed separately, different signal based performance measures such as the SNRE as well as perceptual quality measures can be calculated accurately.

4.2.1 Signal to Noise Ratio Enhancement (SNRE)

The SNR-Enhancement (SNRE) is the difference of the SNR at the output of the beamformer and a reference input-SNR, both measured in dB. For binaural systems the SNRE is calculated between the left (right) output of the binaural system and the left (right) input at the reference microphone, respectively. Although there exist many modifications to this measure, e.g., by using short-time (segmental) SNRE estimates or incorporating speech importance band weighting, the linear broadband SNRE is still an appropriate measure that had shown high correlations with subjective data on the assessment of background noise reduction [17]. Here, the SNR was calculated by taking the mean power of the broadband speech component on a dB-scale(excluding speech pauses, i.e. signal levels 60dB below peak level) minus the broadband noise power in dB. For head-worn systems bilateral performance evaluation is relevant because by simply taking the mean SNRE a better-ear effect would be ignored.

4.2.2 Perceptual Similarity Measure (PSM)

The quality measure PSM from PEMO-Q [18] estimates the perceptual similarity between the processed signal and the clean speech source signal. It has shown high correlations between objective and subjective data and has been used for quality assessment of noise reduction schemes in [19, 17, 20]. PSM increases with increasing (input) SNR. As we are interested in the quality enhancement introduced by the algorithm, we use the deduced measure ΔPSM that is calculated as the difference between the Perceptual Similarity Measure (PSM) of the output and of the unprocessed input signal.

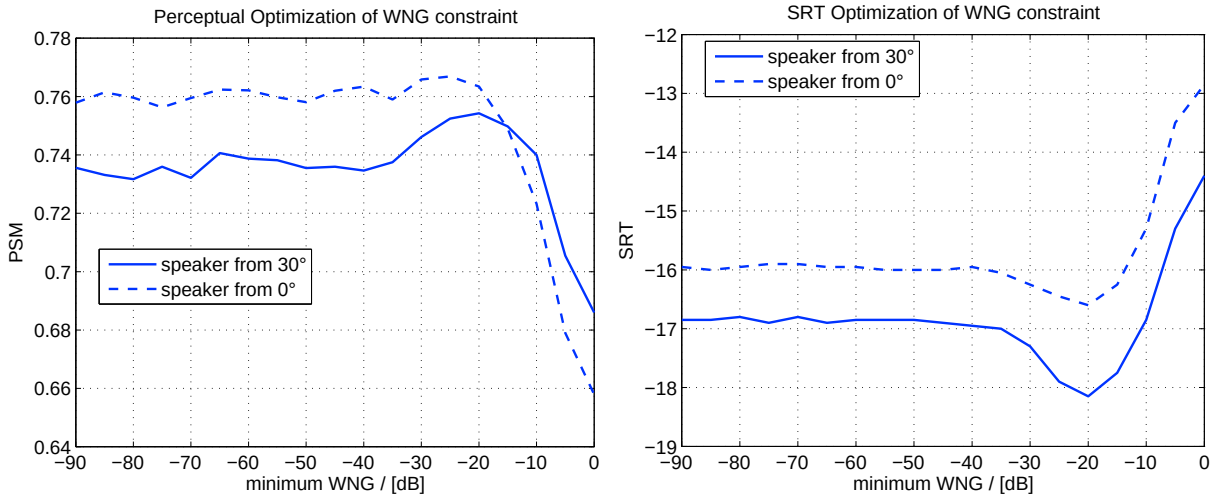
4.2.3 Binaural Speech Reception Threshold (SRT)

The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. In [21] a binaural model of speech intelligibility based on the equalization-cancelation (EC) processing by Durlach had been defined which is able to predict the SRT with high accuracy. If the estimated SRT for the output of a noise reduction scheme is lower than for the input signal this means that the speech intelligibility has increased due to the algorithm. However, as the speech intelligibility is a nonlinear function of the SNR and other signal features such as the preservation of binaural cues, we use the difference between output and input SRT, namely the ΔSRT, as an indirect measure for the increase of intelligibility. The binaural SRT measure as described in [21, 19] assumes a spatially stationary source configuration. To be applicable to moving sources it had to be extended to a block-wise measure with subsequent averaging across blocks.

5 Experiments and Results

5.1 Perceptual Optimization of the White Noise Gain Limitation

Spatially uncorrelated noise can be attributed to self-noise of the microphones as well as to statistical differences between the real-world acoustical scene and the assumed model. Thus, the attenuation of spatially white noise quantified by the WNG measure needs to be guaranteed by the beamformer coefficients to a certain amount. On the other hand, the directivity should be maximized for a maximum noise reduction. The trade-off between superdirectivity and white noise gain has been widely studied, e.g., in [6, 22]. In free-field, the limitation factor μ given in (44) should lie in the range between -10 dB to -30 dB to limit the white noise gain to a minimum of approximately $\delta^2 = -10$ dB which is equivalent to a maximum amplification of uncorrelated noise of 10 dB. However, as μ can be frequency dependent and the relation between $\mu(\omega)$ and δ^2 is none-linear the optimal white noise gain constraint can be found using perceptual quality measures for realistic microphones, model errors and typical realistic acoustical scenes. This perceptual optimization is shown in Figure 3. The x-axis shows the minimum δ^2 to which the white noise gain in (43) was limited. The beamformer coefficients were calculated iteratively due to (44) by increasing $\mu(\omega)$ so that the limit was reached. With these constrained beamformer coefficients real-world recordings have been processed and the perceptual similarity measure (PSM) Figure 3(a) and the speech reception threshold Figure 3(b) have been calculated. The results show that the maximum performance is reached at a white noise gain limit of $\delta^2 = -35$ dB.



(a) PSM (b) SRT

Figure 3: White noise gain constraint

5.2 Binaural output quality

Table 2 shows the performance results for the three binaural strategies (BIN_PF, BIN_PR, BIN_BL) which were evaluated for the fixed beamformers with different propagation models in signal condition 1). Although the mean SNRE values for BIN_PF and BIN_PR were in the same range, BIN_PF had a higher enhancement for the left channel and BIN_PR had a higher enhancement for the right channel. Interestingly, the SRT Gain of BIN_PF was significantly higher than for BIN_PR. This behavior can be explained as follows: As the beamformer output Z is monaural and the multiplication with the left and right propagation vectors only turns the output into the target direction, all signals are perceptually still coming from one direction. In other words: the localization cues for the background noise are lost. The binaural SRT measure can identify the difference as it considers the spatial arrangement of speech and noise signals to calculate the SRT. For this, it does not need explicit knowledge about the interaural time

Algorithm	SNRE L dB	SNRE R dB	mean SNRE dB	PSM L	PSM R	SRT Gain dB	SNR L dB	SNR R dB	SRT dB
FF_BIN_PF	8,1	9,9	9,0	0,66	0,54	7,6	8,0	4,5	-15,4
FF_BIN_PR	4,9	10,2	7,6	0,53	0,55	3,3	4,8	4,8	-11,1
FF_BIN_BL	4,0	4,0	4,0	0,55	0,29	4,6	3,9	-1,4	-12,4
HM1_BIN_PF	7,6	8,8	8,2	0,67	0,57	8,3	7,5	3,4	-16,1
HM1_BIN_PR	4,0	9,7	6,9	0,55	0,58	4,3	3,9	4,3	-12,1
HM1_BIN_BL	4,2	4,6	4,4	0,56	0,32	4,7	4,1	-0,8	-12,5
HM2_BIN_PF	9,0	10,9	10,0	0,69	0,61	8,4	8,9	5,5	-16,2
HM2_BIN_PR	6,5	13,0	9,8	0,59	0,62	5,1	6,4	7,6	-12,9
HM2_BIN_BL	4,4	4,6	4,5	0,56	0,31	4,8	4,3	-0,8	-12,6
HRTF_BIN_PF	9,2	11,4	10,3	0,71	0,64	8,5	9,1	6,0	-16,3
HRTF_BIN_PR	7,2	13,8	10,5	0,61	0,65	5,6	7,1	8,4	-13,4
HRTF_BIN_BL	5,0	6,4	5,7	0,57	0,36	5,1	4,9	1,0	-12,9
input	-	-	-	0,38	0,14	-	-0,1	-5,4	-7,8

Table 2: Binaural output quality

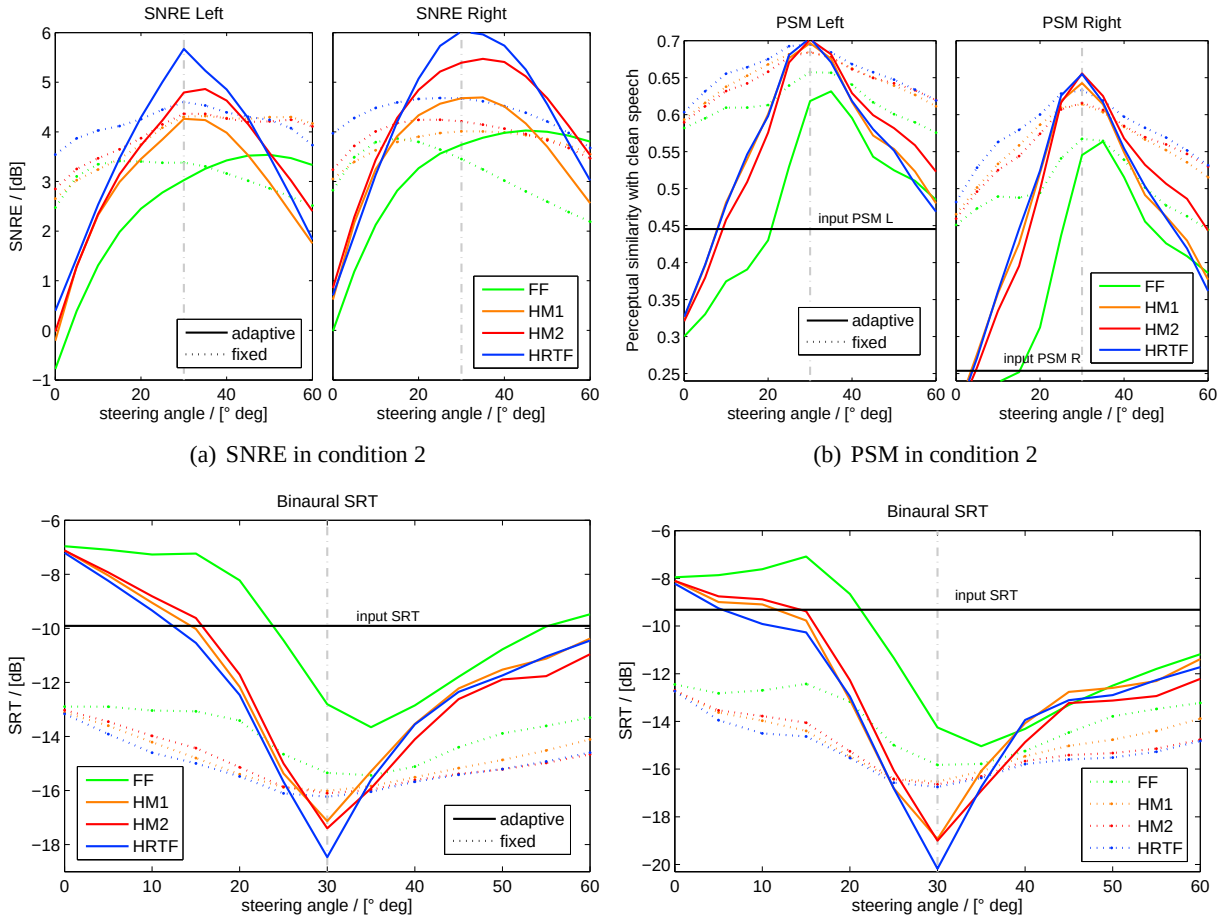
and level difference (ITD, ILD). For BIN_BL the noise reduction performance was reduced compared to BIN_PF and BIN_PR as the bilateral beamformer uses a subarray of only three microphones. However, as the distortion of the binaural cues for BIN_BL is lower than for BIN_PR, the values of the SRT are almost the same. In terms of the different propagation models, quality increases with the complexity and exactness of the model.

5.3 Robustness against steering errors

Figure 4 shows the three quality measures,(a) SNRE,(b) PSM and (c) SRT for different beamformers using the binaural post-filter (BIN_PF) in signal condition 2) over the steering angle of the beamformer. The dotted lines refer to the fixed beamformers, the solid lines to the (adaptive) GSCs and the black lines show the quality values for the unprocessed input signals. The target speech signal came from the 30° direction, so the best quality values should have been expected if the beamformer was steered in this direction. However, depending on the underlying model, algorithm and noise field, this might not always be the case. It can be seen that the free-field coefficients (green curves) are suboptimal for the head-mounted array because the maximum values are not aligned with the steering direction of the beamformer. Among all beamformers, the free-field propagation model leads to the lowest SNRE and the lowest perceptual quality values (PSM, SRT), because it does not incorporate any head-shadow and diffraction effects. The HRTF coefficients led to the highest noise reduction performance but the head models (HM1, HM2) showed comparable results in terms of the predicted overall quality and SRT. The fixed head model beamformers could enhance the SNR in diffuse isotropic noise by about 4 dB. The flatness of the dotted curves shows that they are relative robust against steering errors. The GSCs (solid lines) had approximately 1 dB higher SNREs than the fixed beamformers, but in terms of the estimated overall quality the advantages were small. The SRT estimate was 2 dB lower but these values were only stable within a steering mismatch of $\pm$ 5° degree which pointed out a lower robustness. However for condition 1) with a directional interfering noise source the adaptive beamformer could reduce the SRT by about 4dB more compared to the fixed beamformer that was optimized for suppressing isotropic noise (see Fig. 4 (d)). In summary it could be stated that the GSC was more susceptible to model errors and might only be beneficial in situations with directional interfering noise and small steering errors.

5.4 Robustness against model variation

The second head model (HM2) had shown a good performance that was comparable to the measured HRTFs. However, the robustness of the beamformer designed with HM2 against variations of head-size and position is important for practical applications. Figure 5(b) shows that the HM2 is relative robust against the mismatch between the position of the left and right hearing aid and the true array positions (during the recording of the signals). The same applies to the variation of the head-model's parameter "sphere-size" which is not shown here. This results motivate the use of the head-model for hearing aid



(a) SNRE in condition 2

(b) PSM in condition 2

(c) SRT in condition 2

(d) SRT in condition 1

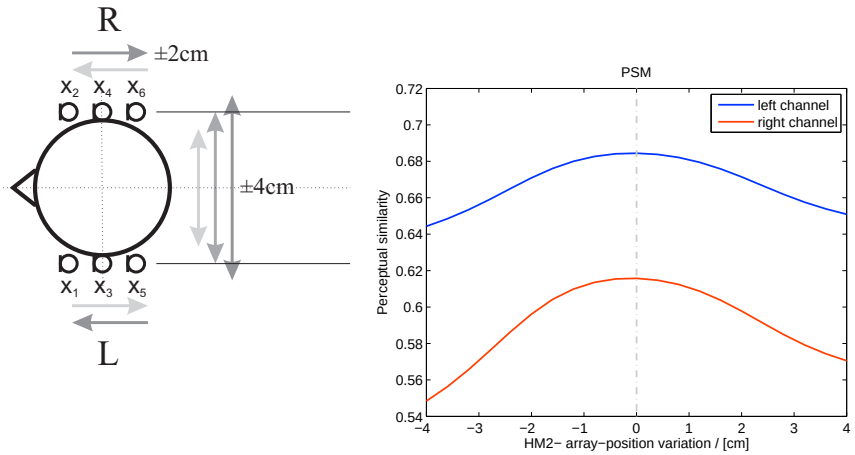
Figure 4: Robustness evaluation against steering mismatch

algorithms.

6 Conclusions

The robustness analysis has shown the importance of the incorporation of head-shadow and diffraction influences in the beamformer design for head-mounted arrays. The fixed beamformers designed with head models were relatively robust against steering errors whereas for adaptive beamformers the robustness was limited and a quality gain compared to fixed beamformers might only be reached in scenarios with directional noise sources and a reliable direction of arrival estimation. However, there are several approaches in literature to increase the robustness of the GSC [23] which have not been incorporated here.

The binaural speech intelligibility measure provides an integrative measure of binaural unmasking and could identify differences in the estimated speech-reception threshold (SRT) if binaural information was distorted. Therefore, it seems to be an appropriate measure to evaluate the perceptual quality of noise reduction schemes with binaural output. In combination with different near-to-realistic sound-scenarios the quality measures showed encouraging results towards a robustness testbench for multichannel-hearing aid algorithms with binaural output. Further work should concentrate on a further empirical validation of the objective perceptual measures.



(a) Model variation (b) Variation of subarray position

Figure 5: Robustness against variation of array position and model parameters for (HM2)

References

[1] T. Van de Bogaert, J. Wouters, T. Klasen, and M. Moonen, "Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 16-19 Oct. 2005, pp. 57–60.

[2] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *Journal on Advances in Signal Processing (EURASIP)*, vol. 2006, pp. Article ID 63 297, 14 pages, 2006.

[3] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[4] J. Desloge, W. Rabinowitz, and P. Zurek, "Microphone-array hearing aids with binaural output .i. fixed-processing systems," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, Nov 1997.

[5] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output. ii. a two-microphone adaptive system," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[6] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[7] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation*, vol. 30, pp. 27–34, 1982.

[8] 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. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[9] T. Lotter, "Single and multimicrophone speech enhancement for hearing aids," Ph.D. dissertation, RWTH Aachen, IND, NÃ$\frac{1}{4}$rnberg, Aug. 2004.

[10] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 5, pp. 3048–3058, 1998.

[11] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[12] S. Yan and Y. Ma, "Robust supergain beamforming for circular array via second-order cone programming," *Applied Acoustics*, vol. 66, no. 9, pp. 1018–1032, Sep. 2005.

[13] B. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 397–401, 1988.

[14] H. Cox, R. M. Zeskind, and T. Kodu, "Practical supergain," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[15] J. Greenberg and P. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *Journal of the Acoustical Society of America (JASA)*, vol. 91, no. 3, pp. 1662–1676, Mar. 1992.

[16] R. Stadler and W. Rabinowitz, "On the potential of fixed arrays for hearing aids," *Journal of the Acoustical Society of America (JASA)*, vol. 94, no. 3, pp. 1332–1342, Sep. 1993.

[17] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 169–172.

[18] R. Huber and B. Kollmeier, "Pemo-Q -A new Method for Objective Audio Quality Assessment using a Model of Auditory Perception." *IEEE Trans. on Audio, Speech and Language Processing*, 2006, special Issue on Objective Quality Assessment of Speech and Audio.

[19] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by means of Objective Perceptual Quality Measures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct. 2007, pp. 315–318.

[20] ——, "Subband-based parameter optimization in noise reduction schemes by means of objective perceptual quality measures," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*. Paris: Télécom Paris, Sep 12-14 2006.

[21] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[22] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[23] O. Hoshuyama and A. Sugiyama, "Robust adaptive beamforming," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 5, pp. 87–106.

Robustness analysis for multi-channel hearing aid algorithms with binaural output by means of objective perceptual quality measures

Thomas Rohdenburg, Volker Hohmann, Birger Kollmeier
Medizinische Physik, Universität Oldenburg, 26111 Oldenburg

Introduction

According to the ITU-T P.835 recommendation, subjective quality evaluation of noise reduction schemes involves (i) the perceived quality of the speech signal, (ii) the quality of the background signal and (iii) the overall quality. In [7] it has been shown that these subjective measures are predictable by objective measures in the case of monaural noise reduction schemes. In this study we extend the quality prediction to the case of multi-channel algorithms. These microphone array based algorithms have other influences on signal quality than single channel envelope filters as they exploit the spatial configuration of interfering signals and therefore in general lead to less signal distortion. For hearing aid applications, data from literature suggest that it is important that the beamformer preserves the binaural information so that the listener can make use of the effect of spatial unmasking. In order to generate a binaural output [6] was adopted.

Signal model and algorithms

The signals were generated using two 3-channel hearing aid headsets mounted on a dummy head. 6-channel HRTFs in an anechoic room and real-world environmental noise in a cafeteria have been recorded. The input signal was composed from two directional signals filtered with HRTFs (target and interferer from 30° and -135° azimuth, respectively) and mixed with the recorded cafeteria noise to generate a near-to realistic scenario. The multi-channel algorithms used here are fixed superdirective beamformers that are designed by the well-known constraint Minimum Variance Distortionless Response (MVDR) solution [2]. This solution allows to include different assumptions on the wave propagation of the target signal and the characteristics of the noise field as described by its cross power spectral density matrix. Three different beamformers were designed with the assumptions about wave propagation (i) in free-field (aka far-field assumption) (ff), (ii) in a simple spherical head model according to [3] (hm) and (iii) with measured 6-channel HRTFs in an anechoic room (hr). These beamformers had monaural outputs that were enhanced by a binaural post-filter according to [6]. The processing block-diagram is shown in Figure 1.

Signal independent quality measures

The beam-pattern is a well-established measure to evaluate the signal independent directional response of a beamformer. It is computed as the response of the array to a wavefront coming from a specific angle at a specific frequency [2]. In general, beam-patterns are only evaluated for far-field propagation. When beamformer coefficients that are designed for far-field are used in a near-field environment with head influences, the constraint of distortionless response may not be fulfilled and the far-field beampattern does not reflect the measured directional response. Therefore, in the near-field or if head-shadow and diffraction effects play a role, these effects also have to be incorporated in the beampattern calculation. Figure 2 shows the beampattern for farfield, beamformer coefficients steered to 30°, (a) evaluated in farfield and (b) evaluated in the nearfield (HRTF). As the beamformer should be designed for the head-mounted array, beampattern (b) shows the more realistic behavior. It can be seen that the target-signal will be distorted and the lateral noise reduction is poor, which is in line with the signal dependent perfomance measures (see below). Also, for other perfomance measures like the *directivity index* the head-shadow and diffraction effects need to be incorporated.

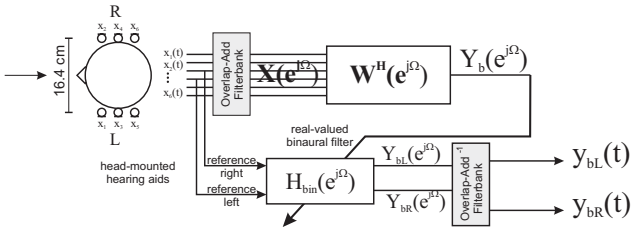


Figure 1: Signal model and beamformer setup

Signal dependent quality measures

SNRE

The SNR-Enhancement (SNRE) is the difference of the signal-to-noise ratio (SNR) at the output of the beamformer and a reference input-SNR, both measured in dB. For a comparison of multi-channel algorithms the choice of the reference is crucial. Here, the SNRE to different references (left, right, source, best microphone) are evaluated for a comparison with the perceptual measures (see below).

PSM

The quality measure PSM from PEMO-Q [4] estimates the perceptual similarity between the processed signal and the clean speech source signal. For monaural noise reduction schemes this measure has shown a high correlation with subjective overall quality ratings according to [5, 7]. Here, the PSM is measured between the clean speech source (before HRTF filtering) and the beamformer output (monaural) or the output of the binaural post-filter, respectively.

SRT

The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. In [1] a binaural model of speech intelligibility based on the equalization-cancelation (EC) processing by Durlach had been defined which is able to predict the SRT with high accuracy. For the objective quality assessment of binaural signals, we define a deduced measure here, namely the SRT Gain. The SRT Gain is calculated iteratively by reducing the SNR of the beamformer input signal until the predicted SRT has the same value as the original unprocessed reference signals. Thus, the SRT Gain is the amount of SNR reduction achieved by the algorithm as estimated by intelligibility estimates including spatial unmasking.

Results

The results in table 1 show that the beamformers with binaural outputs (D,E,F) in general have a higher SNRE. Although for the monaural A) a SNRE source of 5.8 dB was measured, the SNRE compared to the best microphone is almost zero. The same effect can be derived from the SRT Gain, it says that the SRT of the output is worse than the SRT of the unprocessed reference signal. This implies that this algorithm is not helpful to the listener, although the SNR is enhanced by 5.8 dB. Similar effects can also be seen for the other algorithms. The binaural algorithm F) has only a 1.5 dB higher mean SNRE than its monaural counterpart C), but the binaural output leads to an SRT Gain that is 4.3 dB higher. This means that the binaural algorithm can deal with an input signal that is 4.3 dB lower to gain the same speech intelligibility as the monaural algorithm.

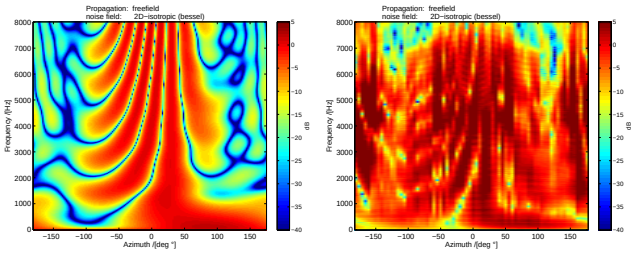
	Algorithm	SNRE Ref L	SNRE Ref R	SNRE source	SNRE best Mic	ΔPSM Ref L	ΔPSM Ref R	SRT Gain
A	mon-rec-ff-2d-fixed	1.6 dB	9.2 dB	5.8 dB	0.4 dB	0.06	0.30	-0.3 dB
B	mon-rec-hm-2d-fixed	2.3 dB	9.8 dB	6.4 dB	1.1 dB	0.12	0.36	1.7 dB
C	mon-rec-hr-hrtf-fixed	4.7 dB	12.2 dB	8.8 dB	3.5 dB	0.20	0.43	3.9 dB
D	bin-rec-ff-2d-fixed	5.3 dB	8.5 dB	7.3 dB	4.1 dB	0.20	0.23	4.5 dB
E	bin-rec-hm-2d-fixed	6.4 dB	7.9 dB	7.5 dB	5.2 dB	0.25	0.29	7.2 dB
F	bin-rec-hr-hrtf-fixed	9.0 dB	11.0 dB	10.3 dB	7.8 dB	0.29	0.40	8.2 dB

Table 1: Performance results for 3 beamformer designs with monaural and binaural outputs

Figures 3 (a-b) show preliminary robustness results for the three beamformers with binaural outputs. The performance measures are plotted over the steering mismatch. The results show that for all quality measures, the free-field and the head-model beamformers do not reach the optimal value at a steering mismatch of 0°. This is because the head diffraction is not (or not sufficiently) incorporated in the coefficients which leads to a steering to higher angles. On the other hand, the gradient of the performance curves is slightly steeper for the hrtf-beamformer which points out that it is more sensitive to steering errors.

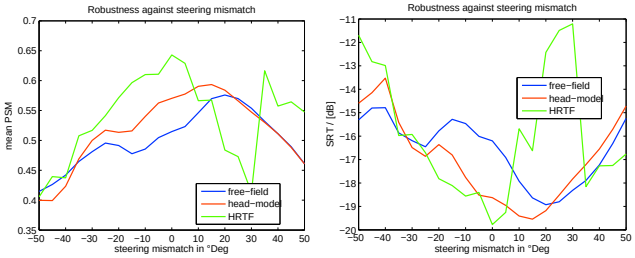
Outlook

Preliminary results have shown the importance of the incorporation of head-shadow and diffraction influences in both the beamformer designs and the performance measures. Furthermore, the performance measures showed a significantly higher quality if the beamformer was extended by a binaural post-filter. The new binaural quality measure showed encouraging results and is an important step towards a robustness testbench for multichannel hearing aid algorithms with binaural outputs.



(a) (b)

Figure 2: Beampatterns for far-field beamformer coefficients steered to 30° and used (a) in far-field and (b) in near-field environment



(a) PSM (b) SRT

Figure 3: Robustness against steering mismatch

Work supported by the EC (DIRAC project IST-027787) and BMBF

References

[1] Beutelmann, R. ; Brand, T. : Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. In: *Journal of the Acoustical Society of America* 120 (2006), Nr. 1, S. 331–342

[2] Kap. 2 In: Bitzer, J. ; Simmer, K. U.: *Superdirective Microphone Arrays*. Springer, 2001, S. 19–38

[3] Brown, P. C. ; Duda, R. O.: A Structural Model For Binaural Sound Synthesis. In: *IEEE Trans. on Speech and Audio Processing* 6 (1998), Sep, Nr. 5, S. 476–488

[4] Huber, R. ; Kollmeier, B. : PEMO-Q - A new Method for Objective Audio Quality Assessment using a Model of Auditory Perception. In: *IEEE Trans. on Audio, Speech and Language Processing* (2006). – Special Issue on Objective Quality Assessment of Speech and Audio

[5] ITU-T: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm / ITU. 2003 (Recommendation P.835). – Series P: Telephone Transmission Quality

[6] Lotter, T. ; Sauert, B. ; Vary, P. : A Stereo Input-Output Superdirective Beamformer for Dual Channel Noise Reduction. In: *Interspeech*. Lisbon, Portugal, sep 2005, S. 2285–2288

[7] Rohdenburg, T. ; Hohmann, V. ; Kollmeier, B. : Objective Perceptual Quality Measures for the Evaluation of Noise Reduction Schemes. In: Janse, C. (Hrsg.) ; Moonen, M. (Hrsg.) ; Sommen, P. (Hrsg.): *9th International Workshop on Acoustic Echo and Noise Control*. Eindhoven, 2005, S. 169–172

ROBUSTNESS ANALYSIS OF BINAURAL HEARING AID BEAMFORMER ALGORITHMS BY MEANS OF OBJECTIVE PERCEPTUAL QUALITY MEASURES

Thomas Rohdenburg, Volker Hohmann, Birger Kollmeier

Universität Oldenburg
Medizinische Physik
26111 Oldenburg, Germany
thomas.rohdenburg@uni-oldenburg.de

ABSTRACT

In this contribution different microphone array-based noise reduction schemes for hearing aids are suggested and compared in terms of their performance, signal quality and robustness against model errors. The algorithms all have binaural output and are evaluated using objective perceptual quality measures [1, 2, 3]. It has been shown earlier that these measures are able to predict subjective data that is relevant for the assessment of noise reduction algorithms. The quality measures showed clearly that fixed beamformers designed with head models were relatively robust against steering errors whereas for the adaptive beamformers tested in this study the robustness was limited and the benefit due to higher noise reduction depended on the noise scenario and the reliability of a direction of arrival estimation. Furthermore, binaural cue distortions introduced by the different binaural output strategies could be identified by the binaural speech intelligibility measure [3] even in case monaural quality values were similar. Thus, this perceptual quality measure seems to be suitable to discover the benefit that the listener might have from the effect of spatial unmasking.

1. INTRODUCTION

In modern hearing aids multi-channel noise reduction schemes based on small microphone arrays are used for speech enhancement. These algorithms exploit the spatial configuration of the interfering signals and therefore generally lead to less signal distortion and higher noise reduction than single-channel envelope filters. The human ability to separate sound sources in a complex situation, namely the cocktail-party effect, partly arises from the use of binaural localization cues. If binaural information is lost or distorted by the processing, the hearing impaired listener may not make use of the effect of spatial unmasking as efficiently as in the undistorted binaural condition. The intelligibility improvement introduced by a spatial filter is counteracted by the decrease due to the deteriorated efficiency of the spatial unmasking in this case. Although bilateral supply with hearing aids is motivated by a better directional-hearing ability, it has been shown in [4] that binaural cues are distorted if the hearing aids at the left and right ears work independently. Therefore, researchers have suggested microphone array based binaural spatial filtering techniques [5, 6, 7, 8] that assume a connection between the left and right hearing aid. In this study we analyzed fixed and adaptive beamformer algorithms, that exploit *a priori* knowledge about array position, wave propagation and direction of arrival as these seem to be slowly varying parameters that can be estimated and used for the adaptation of the algorithms. Information about the voice activity which might also be helpful for noise estimation was not used here. The beamformers that were calculated using the constraint minimum variance distortionless response (MVDR) design [9] had single channel outputs that were extended by a binaural stage. Three different strategies for generating a binaural output have been applied and evaluated by perceptual measures. Furthermore, the robustness of fixed and adaptive beamformers using different propagation models have been analyzed against steering error, array position and head-size mismatch by appropriate perceptual quality measures.

2. SIGNAL MODEL

The signals were recorded using two 3-channel behind-the-ear hearing aid shells mounted on a B&K dummy head. 6-channel head related transfer functions (HRTFs) in an anechoic room and real-world environmental noise in a cafeteria have been recorded. For condition 1) the input signal was composed from two directional signals filtered with HRTFs (target and interferer from 30° (front-left) and -135° (back-right) azimuth, respectively) and mixed with the recorded cafeteria noise to generate a near-to-realistic scenario. For condition 2) we used only one directional signal (speaker from 30° (left)) mixed with an artificial diffuse noise. The artificial noise was generated by summing up a speech-colored random noise that was filtered with HRTFs from all directions to simulate a 2D-isotropic noise field. This abated the influence of the noise field characteristic on the signal quality which was helpful for the analysis of the steering mismatch. The 30° direction was chosen because it is asymmetric to the array and offers a more general assessment of the beamformers properties than a fixed 0° look direction.

3. ALGORITHMS

The multi-channel algorithms used here are designed using the well-known constraint Minimum Variance Distortionless Response (MVDR) solution [9], Eq. (1),

$$\boldsymbol{W}(f) = \frac{\boldsymbol{\Phi_{NN}}^{-1}(f)\boldsymbol{d}(f)}{\boldsymbol{d^H}(f)\boldsymbol{\Phi_{NN}}^{-1}(f)\boldsymbol{d}(f)} \quad (1)$$

$$\boldsymbol{d}(f) = \left[a_0 e^{j2\pi f\tau_0}, a_1 e^{j2\pi f\tau_1}, \ldots, a_{M-1} e^{j2\pi f\tau_{M-1}}\right]^T \quad (2)$$

$$Y_f(f) = \boldsymbol{W^H}(f)\boldsymbol{X}(f) \quad (3)$$

where f denotes the frequency, $\boldsymbol{W}$ the beamformer coefficients, $\boldsymbol{d}$ the propagation vector, a_m and τ_m the amplitude and the group delay at microphone m, $\boldsymbol{X}$ the input vector, Y_f the output of the fixed beamformer (see Figure 1).

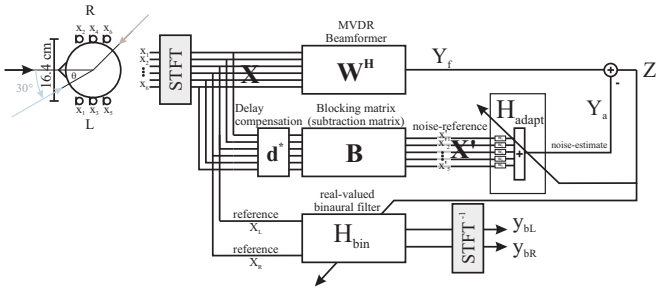


Figure 1: GSC beamformer and binaural post-filter

This solution allows to include different assumptions about the wave propagation of the target signal (included in the propagation vector $\boldsymbol{d}$, Eq. (2)), and the characteristics of the noise field as described by its cross power spectral density matrix $\boldsymbol{\Phi_{NN}}$.

The upper path of the signal diagram in Figure 1 shows the fixed beamformer which can be extended by an adaptive noise canceler path to form a Generalized Sidelobe Canceller (GSC) [10, 11]. Note that *fixed* beamformer means that a fixed noise field is assumed whereas a GSC can adapt to varying noise fields. However, for both beamformer types an adaptive steering to a moving target signal can be applied, e.g., if extended by a direction of arrival (DOA) estimation algorithm. Additionally, the beamformers are extended by a binaural stage with diverse methods to obtain a binaural output. All combinations of beamformer type (fixed/adaptive), binaural output method (bin1, bin2, bin3) and different assumptions about the wave propagation model (free-field (FF), head-model (HM1, HM2, HRTF)) are investigated in this study in terms of their performance and robustness.

Wave propagation models can be integrated into the beamformer design via the propagation vector $\boldsymbol{d}$ and the noise field cross power density matrix $\boldsymbol{\Phi_{NN}}$. For the free-field (FF), $\boldsymbol{d}$ has constant group-delay, τ_m, and unity amplitude, a_m, in the frequency domain. For head-models τ_m, a_m are frequency dependent accounting for head shadow and diffraction effects. The first head model (HM1) by [12] is a simple and effective parametric model that estimates the characteristic of a sphere. The interaural time difference (ITD) cues are modeled by Woodworth and Schlosberg's frequency independent (ray-tracing) formula. The gross magnitude characteristics of the HRTF spectrum, the interaural level difference (ILD) cues, are covered by a single-pole, single-zero head-shadow filter which also accounts for an additional frequency dependent delay for low frequencies [12]. The second head model (HM2) by [13] additionally incorporates the distance of the source for modeling near-field effects and interference effects that introduce ripples in the response that are quite prominent on the shadowed side. It is calculated by a recursive algorithm given in [13]. The third head model (HRTF) uses the measured HRTF of the respective microphones directly as the propagation vector. The noise field matrix $\boldsymbol{\Phi_{NN}}$ influences the amount of noise reduction achieved by the beamformer. In the free-field, a 3D-isotropic diffuse noise field matrix reduces to a coherence matrix with sinc-characteristic [9]. For the head-models the diffuse noise field is estimated by integrating the propagation vectors over all directions. Furthermore, $\boldsymbol{\Phi_{NN}}$ needs to be constrained to reduce super directivity for feasible designs [9, 10].

Binaural Outputs are calculated using three different methods:

(**i**) (bin1) The binaural output is generated by a real-valued time-varying post-filter based on [5] that is controlled by the monaural beamformer output Z:

$$H_{\mathrm{Bin}}(t,f) = \frac{\left(|d_L(f)|^2 + |d_R(f)|^2\right)\Phi_{ZZ}(t,f)}{\Phi_{X_LX_L}(t,f) + \Phi_{X_RX_R}(t,f)} \quad (4)$$

$$Y_{bL}(t,f) = H_{\mathrm{Bin}}(t,f)X_L(t,f) \quad (5)$$

$$Y_{bR}(t,f) = H_{\mathrm{Bin}}(t,f)X_R(t,f) \quad (6)$$

where X_L,X_R (see Fig. 1) denote the input signals and d_L, d_R the propagation vectors for the expected signal direction θ_S, at the left and right reference microphone, respectively. Φ_{ZZ}, $\Phi_{X_LX_L}$ and $\Phi_{X_RX_R}$ are the power spectral density estimates for the signals Z, X_L, X_R, respectively. As the filter is real-valued, the phase of signal and noise are kept and therefore also most of the binaural cues. However, the envelope filter might introduce additional signal distortions.

(**ii**) (bin2) The monaural beamformer output Z is multiplied by the propagation vectors of the reference microphones which reconstructs only the interaural phase of the signal and may degrade spatial unmasking effects:

$$Y_{bL}(t,f) = d_L(f)Z(t,f) \quad (7)$$

$$Y_{bR}(t,f) = d_R(f)Z(t,f) \quad (8)$$

(**iii**) (bin3) The array is split into a subarray of two parallel 3-channel beamformers $\boldsymbol{W_L}$, $\boldsymbol{W_R}$ which use common information about the target direction and the noise field. This simulates the behavior of independent bilateral hearing devices and binaural cues may be distorted as described in [4]:

$$Y_{bL}(t,f) = Z_L(t,f) = \boldsymbol{W_L^H}(f)\boldsymbol{X_{135}}(t,f) \quad (9)$$

$$Y_{bR}(t,f) = Z_R(t,f) = \boldsymbol{W_L^H}(f)\boldsymbol{X_{246}}(t,f) \quad (10)$$

where the numbers (1,3,5 and 2,4,6) refer to the microphones of the subarray, respectively.

4. QUALITY MEASURES

SNRE: The SNR-Enhancement (SNRE) is the difference of the signal-to-noise ratio (SNR) at the output of the beamformer and a reference input-SNR, both measured in dB. For a comparison of multi-channel algorithms the choice of the reference is crucial. Here, the SNRE is calculated between the left (right) output of the binaural stage and the left (right) input at the reference microphone, respectively.

PSM: The quality measure PSM from PEMO-Q [2] estimates the perceptual similarity between the processed signal and the clean speech source signal. For monaural noise reduction schemes this measure has shown a high correlation with subjective overall quality ratings according to [1, 14]. Here, the PSM is measured between the clean speech component at the left (right) reference microphone and the left (right) output of the binaural stage.

SRT: The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. In [3] a binaural model of speech intelligibility based on the equalization-cancelation (EC) processing by Durlach had been defined which is able to predict the SRT with high accuracy. For the objective quality assessment of binaural signals processed by noise reduction schemes, we are interested in the difference between the SRT of the input signal and the SRT of the output, namely the SRT Gain. Thus, the SRT Gain is the amount of SNR reduction achieved by

Algorithm	SNRE L dB	SNRE R dB	mean SNRE dB	PSM L	PSM R	SRT Gain dB	SNR L dB	SNR R dB	SRT dB
FF_bin1	8.1	9.9	9.0	0.66	0.54	7.6	8.0	4.5	-15.4
FF_bin2	4.9	10.2	7.6	0.53	0.55	3.3	4.8	4.8	-11.1
FF_bin3	4.0	4.0	4.0	0.55	0.29	4.6	3.9	-1.4	-12.4
HM1_bin1	7.6	8.8	8.2	0.67	0.57	8.3	7.5	3.4	-16.1
HM1_bin2	4.0	9.7	6.9	0.55	0.58	4.3	3.9	4.3	-12.1
HM1_bin3	4.2	4.6	4.4	0.56	0.32	4.7	4.1	-0.8	-12.5
HM2_bin1	9.0	10.9	10.0	0.69	0.61	8.4	8.9	5.5	-16.2
HM2_bin2	6.5	13.0	9.8	0.59	0.62	5.1	6.4	7.6	-12.9
HM2_bin3	4.4	4.6	4.5	0.56	0.31	4.8	4.3	-0.8	-12.6
HRTF_bin1	9.2	11.4	10.3	0.71	0.64	8.5	9.1	6.0	-16.3
HRTF_bin2	7.2	13.8	10.5	0.61	0.65	5.6	7.1	8.4	-13.4
HRTF_bin3	5.0	6.4	5.7	0.57	0.36	5.1	4.9	1.0	-12.9
input	-	-	-	0.38	0.14	-	-0.1	-5.4	-7.8

Table 1: Binaural output quality

the algorithm as estimated by intelligibility estimates including spatial unmasking. However, if the noise reduction algorithm is nonlinear the exact SRT Gain has to be calculated iteratively by reducing the SNR of the beamformer input signal until the predicted SRT has the same value as the original unprocessed reference signals.

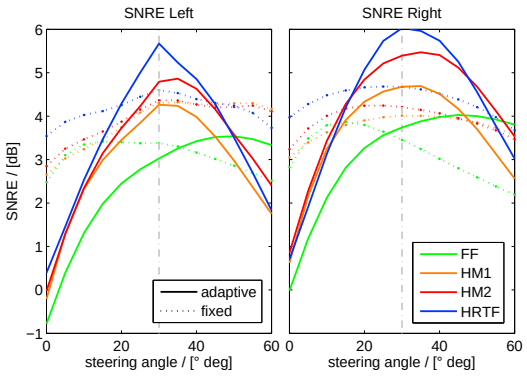
5. RESULTS

5.1. Binaural output quality

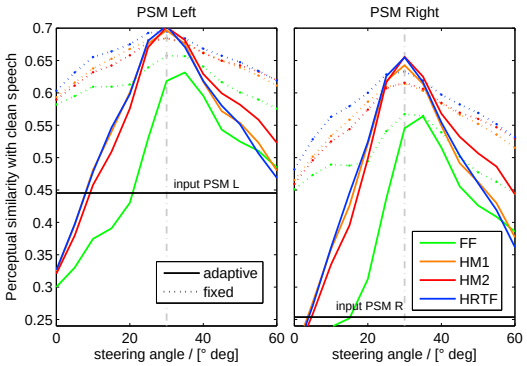
Table 1 shows the performance results for the three binaural strategies (bin1-3) which were evaluated for the fixed beamformers with different propagation models in signal condition 1). Although the mean SNRE values for bin1 and bin2 were in the same range, bin1 had a higher enhancement for the left channel and bin2 had a higher enhancement for the right channel. Interestingly, the SRT Gain of bin1 was significantly higher than for bin2. This behavior can be explained as follows: As the beamformer output Z is monaural and the multiplication with the left and right propagation vectors only turns the output into the target direction, all signals are perceptually still coming from one direction. In other words: the localization cues for the background noise are lost. The binaural SRT measure can identify the difference as it considers the spatial arrangement of speech and noise signals to calculate the SRT. For this, it does not need explicit knowledge about the interaural time and level difference (ITD, ILD). For bin3 the noise reduction performance was reduced compared to bin1 and bin2 as the bilateral beamformer uses a subarray of only three microphones. However, as the distortion of the binaural cues for bin3 is lower than for bin2, the values of the SRT are almost the same. In terms of the different propagation models, quality increases with the complexity and exactness of the model.

5.2. Robustness against steering errors

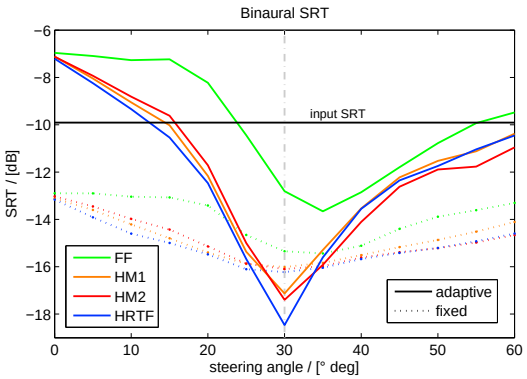
Figure 2 shows the three quality measures,(a) SNRE,(b) PSM and (c) SRT for different beamformers using the binaural post-filter (bin1) in signal condition 2) over the steering angle of the beamformer. The dotted lines refer to the fixed beamformers, the solid lines to the (adaptive) GSCs and the black lines show the quality values for the unprocessed input signals. The target speech signal came from the 30° direction, so the best quality values should have been expected if the beamformer was steered in this direction. However, depending on the underlying model, algorithm and noise field, this might not always be the case. It can be seen that the free-field coefficients (green curves) are suboptimal for the head-mounted array because the maximum values are not aligned with the steering direction of the beamformer. Among all beamformers, the free-field propagation model leads to the lowest SNRE and the lowest perceptual quality values (PSM, SRT), because it does not incorporate any head-shadow and diffraction effects. The HRTF



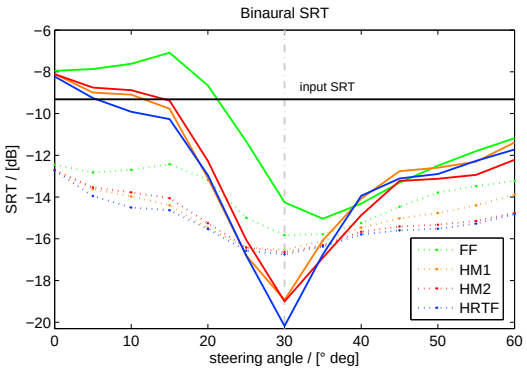
(a) SNRE in condition 2



(b) PSM in condition 2



(c) SRT in condition 2

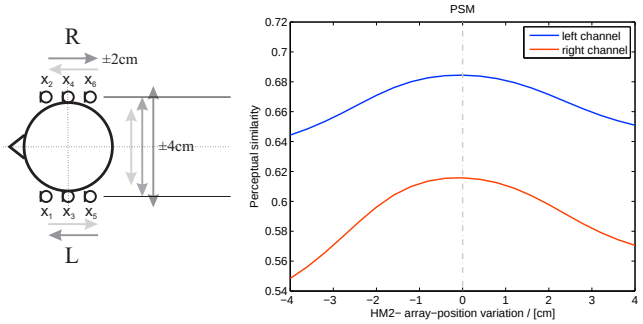


(d) SRT in condition 1

Figure 2: Robustness evaluation against steering mismatch

coefficients led to the highest noise reduction performance but the head models (HM1, HM2) showed comparable results in terms of the predicted overall quality and SRT. The fixed head model beamformers could enhance the SNR in diffuse isotropic noise by about 4 dB. The flatness of the dotted curves shows that they are relative robust against steering errors. The GSCs (solid lines) had approximately 1 dB higher SNREs than the fixed beamformers, but in terms of the estimated overall quality the advantages were small. The SRT estimate was 2 dB lower but these values were only stable within a steering mismatch of $\pm$ 5° degree which pointed out a lower robustness. However for condition 1) with a directional interfering noise source the adaptive beamformer could reduce the SRT by about 4dB more compared to the fixed beamformer that was optimized for suppressing isotropic noise (see Fig. 2 (d)). In summary it could be stated that the GSC was more susceptible to model errors and might only be beneficial in situations with directional interfering noise and small steering errors.

5.3. Robustness against model variation



(a) Model variation (b) Variation of subarray position

Figure 3: Robustness against variation of array position and model parameters for (HM2)

The second head model (HM2) had shown a good performance that was comparable to the measured HRTFs. However, the robustness of the beamformer designed with HM2 against variations of head-size and position is important for practical applications. Figure 3(b) shows that the HM2 is relative robust against the mismatch between the position of the left and right hearing aid and the true array positions (during the recording of the signals). The same applies to the variation of the head-model's parameter "sphere-size" which is not shown here. This results motivate the use of the head-model for hearing aid algorithms.

6. CONCLUSIONS

The robustness analysis has shown the importance of the incorporation of head-shadow and diffraction influences in the beamformer design for head-mounted arrays. The fixed beamformers designed with head models were relatively robust against steering errors whereas for adaptive beamformers the robustness was limited and a quality gain compared to fixed beamformers might only be reached in scenarios with directional noise sources and a reliable direction of arrival estimation. However, there are several approaches in literature to increase the robustness of the GSC [11] which have not been incorporated here.
The binaural speech intelligibility measure provides an integrative measure of binaural unmasking and could identify differences in the estimated speech-reception threshold (SRT) if binaural information was distorted. Therefore, it seems to be an appropriate measure to evaluate the perceptual quality of noise reduction schemes with binaural output. In combination with different near-to-realistic sound-scenarios the quality measures showed encouraging results towards a robustness testbench for multichannel-hearing aid algorithms with binaural output. Further work should concentrate on a further empirical validation of the objective perceptual measures.

Work supported by the EC (DIRAC project IST-027787), HearCom-Project (IST-004171) and BMBF

7. REFERENCES

[1] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 169–172.

[2] R. Huber and B. Kollmeier, "Pemo-q - a new method for objective audio quality assessment using a model of auditory perception." *IEEE Trans. on Audio, Speech and Language Processing*, 2006, special Issue on Objective Quality Assessment of Speech and Audio.

[3] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[4] 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," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005.

[5] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 63 297, 14 pages, 2006.

[6] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[7] J. Desloge, W. Rabinowitz, and P. Zurek, "Microphone-array hearing aids with binaural output .i. fixed-processing systems," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, Nov 1997.

[8] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output. ii. a two-microphone adaptive system," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[9] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[10] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation*, vol. 30, pp. 27–34, 1982.

[11] O. Hoshuyama and A. Sugiyama, "Robust adaptive beamforming," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 5, pp. 87–106.

[12] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[13] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 5, pp. 3048–3058, 1998.

[14] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU, Series P: Telephone Transmission Quality Recommendation P.835, Nov. 2003.

DIRECTION OF ARRIVAL ESTIMATION BASED ON THE DUAL DELAY LINE APPROACH FOR BINAURAL HEARING AID MICROPHONE ARRAYS

Stefan Goetze[1], Thomas Rohdenburg[2], Volker Hohmann[2], Birger Kollmeier[2], and Karl-Dirk Kammeyer[1]

[1]University of Bremen
Dept. of Communications Engineering
28334 Bremen, Germany
goetze@ant.uni-bremen.de

[2]University of Oldenburg
Medical Physics Group
26111 Oldenburg, Germany
thomas.rohdenburg@uni-oldenburg.de

ABSTRACT

Multi-channel beamformer algorithms are promising solutions for noise reduction in hearing aids as they exploit the spatial distribution of the interfering signals and therefore in general lead to less signal distortion than single channel algorithms. Beamformers need *a priori* information about the microphone array and the direction of arrival of the target speech source. For head-worn arrays it is usually assumed that the user physically steers the arrays' look direction toward the desired speech source. This may become unsatisfying for the hearing aid user for high directivity beamformers with a small main lobe and when the target signal source is moving. In this contribution an automatic steering (electronic control of the look direction) is applied based on the dual delay line approach after Liu et al. [1]. This approach is modified to be applicable for head-mounted hearing-aid arrays. We show that the original free-field approach does not work on a head-mounted array because of the inappropriate propagation model. If we apply the true HRTF or a spherical head propagation model, the estimate is reliable within $\pm 8^\circ$ degree mean estimation error for an input SNR of 10dB or higher. However, for lower SNR the method seems to be not robust enough.

Index Terms— Direction of Arrival (DOA), Head Related Transfer Function (HRTF), Noise Reduction, Beamforming

1. INTRODUCTION

In modern hearing aids multiple microphones are applied to reduce ambient noise by exploiting spatial information. Many contributions in the literature either assume a fixed look direction to zero degree or the Direction of Arrival (DOA) to be perfectly known. In the first case steering is accomplished by head movements to the desired source. However it has been shown by several authors that a steering mismatch due to a wrong estimation of the DOA severely degrades the beamformer performance [2, 3]. In this contribution the dual delay line approach after Liu et al. [1] is extended by the consideration of head shadowing effect to work with binaural beamforming algorithms for digital hearing aids. The performance of the system is analyzed in interaction with a binaural noise reduction scheme consisting of a fixed Minimum Variance Distortionless Response (MVDR) beamformer and a binaural post-filter.

The remainder of this paper is organized as follows: In Section 2 the proposed DOA estimation technique is reviewed for free-field assumptions of [1] and extended to work with Head Related Transfer Functions (HRTFs). In Section 3 the binaural noise reduction scheme is described. Simulation results for both, DOA estimation and noise reduction performance are presented in Section 4 and Section 5 gives some final conclusions.

Notation: Vectors and matrices are printed in boldface while scalars are printed in italic. k is the discrete time index, m the discrete frequency index and ℓ the discrete block index, respectively. The superscripts ^{T}, *, and ^{H} denote the transposition, the complex conjugation and the Hermitian transposition respectively.

2. ESTIMATION OF DIRECTION OF ARRIVAL

For noise reduction by microphone arrays a reliable estimate of the DOA of the desired sound source is a crucial point. The performance of beamforming noise reduction techniques is often heavily degraded if DOA estimation errors occur, especially if adaptive algorithms are applied [3].

2.1. Free-field assumptions

For the free-field assumption the dual delay line approach after Liu et al. [1] is promising because the spatial resolution can be directly influenced by choosing an appropriate number of sectors I. It will be briefly reviewed in the following with a somewhat modified notation and the specific problems caused by the shadow effects of the human head will be pointed out.

As depicted in Fig. 1 two microphones capture the sound signals $x_0[k]$ and $x_1[k]$ at two spatial positions $\mathbf{p}_0$ and $\mathbf{p}_1$. The time signals are multiplied by a Hann window $w[k]$ and transformed into the frequency domain

$$x^{(\ell)}[m] = \sum_{k=0}^{L_{\mathrm{DFT}}-1} x[\ell L_{\mathrm{Bl}} + k] w[k] e^{-j2\pi km/L_{\mathrm{DFT}}}. \quad (1)$$

Here L_{DFT} and L_{Bl} are the DFT-length and the block length, respectively. An appropriate zero-padding can be applied to reduce cyclic convolutions effects. For the reason of better readability the block index ℓ is omitted in the remainder if it is not necessary. Following [1] we divide the azimuth range of interest $\Phi = -90^\circ..90^\circ$ into I sectors as depicted in Figure 1.

For each sector i which corresponds to an angle Φ_i a propagation vector $\mathbf{d}[m, \Phi]$ for the left and the right channel can defined as

$$\mathbf{d}[m, \Phi] = \begin{bmatrix} |d_0[m, \Phi]| e^{-j2\pi m \frac{f_s}{M} \tau(\Phi_{i,0})} \\ |d_1[m, \Phi]| e^{-j2\pi m \frac{f_s}{M} \tau(\Phi_{i,1})} \end{bmatrix}. \quad (2)$$

For free field assumptions the absolute values of (2) equal one for all discrete frequencies ($|d_i[m, \Phi]| = 1, \forall m, \Phi$) and the differ-

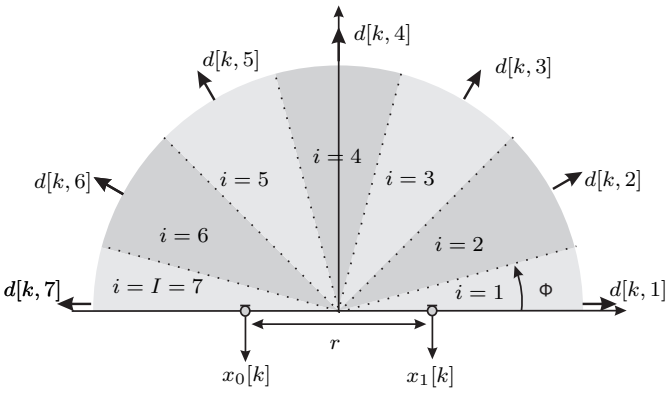


Fig. 1. Dual-microphone setup with $I = 7$ possible DOA sectors.

ence between the signal of the left channel $x_0[k]$ and the right channel $x_1[k]$ is just a time delay $\Delta\tau = \tau(\Phi_{i,0}) - \tau(\Phi_{i,1}) = \frac{r\cos\Phi_i}{c}$. Here r and $c = 344$ m/s are the inter-microphone distance and the speed of sound, respectively.

The microphone signals can be defined as a superposition of the desired signal $s[m]$ multiplied by the corresponding propagation vector $\mathbf{d}[m, \Phi]$ and some ambient noise $n[m]$:

$$x_0[m, \Phi] = s[m] \cdot d_0[m, \Phi] + n_0[m] \quad (3)$$

$$x_1[m, \Phi] = s[m] \cdot d_1[m, \Phi] + n_1[m] \quad (4)$$

Thus the desired direction of arrival can be obtained by

$$\Phi_{\text{opt}}[m] = \arg\min_{\Phi}[m] \{\Delta x[m, \Phi]\} \quad (5)$$

with

$$\Delta x[m, \Phi] = |x_0[m, \Phi]/d_0[m, \Phi] - x_1[m, \Phi]/d_1[m, \Phi]|\,. \quad (6)$$

Replacing $x_0[m, \Phi]$ (3) and $x_1[m, \Phi]$ (4) in (5) the minimization leads to a minimum of

$$v[m, \Phi] = |n_0[m]/d_0[m, \Phi] - n_1[m]/d_1[m, \Phi]| \quad (7)$$

at the angle $\Phi[m] = \Phi_{\text{opt}}[m]$. For free field assumptions the minimum of (7) gives a good estimate of the desired direction for a moderate noise level. Hence if head shadow effects have to be taken into account which results in a non-flat absolute value of the propagation factor ($|d_i[m, \Phi]| \neq 1$) the estimate fails completely.

2.2. Robustness improvements

For improving the robustness of the DOA estimation an averaging in time direction

$$\Delta\tilde{x}^{(\ell)}[m, \Phi] = \alpha \cdot \Delta x^{(\ell-1)}[m, \Phi] + (1 - \alpha) \cdot \Delta x^{(\ell)}[m, \Phi] \quad (8)$$

and in frequency direction

$$\hat{\Phi}_{\text{opt}} = \frac{1}{L_{\text{DFT}}} \sum_{m=0}^{L_{\text{DFT}}-1} \Phi_{\text{opt}}[m] \quad (9)$$

can be applied. Furthermore the maximum tracking speed of the DOA estimator should be limited to a certain threshold by

$$|\hat{\Phi}_{\text{opt}}^{(\ell-1)} - \hat{\Phi}_{\text{opt}}^{(\ell)}| < \xi \quad (10)$$

to avoid short but high estimation errors. This would lead to annoying artifacts if the beamformer steers to a completely wrong direction for a short period.

2.3. Head Shadowing Effects

If microphones are used which are mounted near the human head, e.g., on the frame of eyeglasses or in behind-the-ear (BTE) hearing-aids the free field assumption becomes invalid and the true Head Related Transfer Functions (HRTFs) have to be taken into account. For simulations 6-channel HRTFs were measured in an anechoic room using two three-channel BTE hearing aid shells mounted on a Brüel & Kjær (B&K) dummy head. Since in general HRTFs are unique for every human person they are not available for real-world DOA estimation. Thus head models have to be applied to estimate the HRTFs. In this contribution a head model by Duda [4, 5] is used which is a simple but effective parametric model that estimates the characteristics of a sphere. The interaural time difference (ITD) cues are modeled by Woodworth and Schlosberg's frequency independent (ray-tracing) formula. The gross magnitude characteristics of the HRTF spectrum, namely the interaural level difference (ILD) cues, are covered by a first order IIR head shadow filter which also accounts for an additional frequency dependent delay for low frequencies [5]. Near-field effects and interference effects that introduce ripples in the frequency response which are quite prominent on the shadowed side are incorporated and described in [4].

If a DOA estimator has to work near the human head shadowing effects have to be taken into account. As it is shown in Fig. 2 the HRTFs have strong level differences for different angles and thus the free-field assumption, where only the phase of the propagation factor is considered leads to wrong DOA estimates.

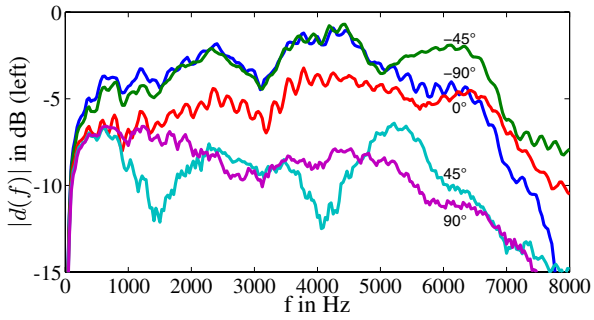


Fig. 2. Absolute values of Head Related Transfer Functions (HRTFs) of left channel.

3. MULTI-CHANNEL NOISE REDUCTION

Fig. 3 shows the system model of the multi-channel noise reduction scheme used in this paper. The discrete microphone signals $x_i[k], i = 1..6$ are transformed into the frequency domain by the Short Time Fourier Transform (STFT) (1). The DOA estimator feeds the MVDR beamformer with the propagation vector $\mathbf{d}[m, \hat{\Phi}_{\text{opt}}]$ corresponding to the estimated angle $\hat{\Phi}_{\text{opt}}$. The monaural beamformer output is further processed by the binaural post-filter $\mathbf{H}_{\text{Bin}}[m]$ to generate binaural output [3, 6] which is transformed back into time domain by the Inverse Short Time Fourier Transform (STFT^{-1}). The multi-channel algorithms used here are designed using the well-known constraint Minimum Variance Distortionless Response (MVDR) solution [7]:

$$\mathbf{W}[m] = \frac{\mathbf{\Gamma}_{NN}^{-1}[m]\mathbf{d}[m]}{\mathbf{d}^H[m]\mathbf{\Gamma}_{NN}^{-1}[m]\mathbf{d}[m]} \quad (11)$$

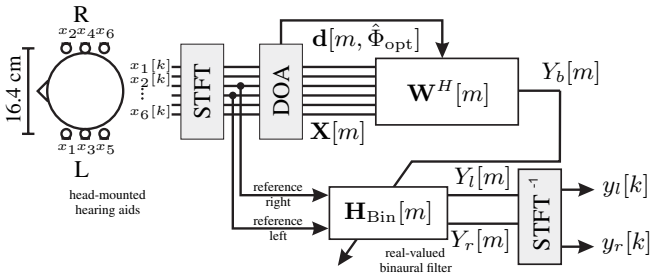


Fig. 3. Signal model and beamformer setup.

This solution allows to include different assumptions about the wave propagation of the target signal (included in the propagation vector $\mathbf{d}$), and the characteristics of the noise field as described by its cross power spectral density matrix $\mathbf{\Gamma_{NN}}[m]$. Although the beamformer is steered adaptively by the DOA estimator to variable directions, it is referred to as a *fixed* beamformer, as it is fixed in terms of the expected noise field. If the beamformer should optimally reduce noise from an arbitrary direction the beamformer coefficients can be designed with an isotropic noise field characteristic. For a diffuse noise field the cross power spectral density matrix $\mathbf{\Gamma_{NN}}[m]$ depends on the underlying propagation model and can be estimated by integrating the propagation vectors over all directions. For the free-field assumption the isotropic noise field $\mathbf{\Gamma_{NN}}[m]$ can be solved analytically: in 3-D the correlation can be described by a sinc-function [7], in cylindrical coordinates by a bessel-function. Due to the spatial filtering effect of the head the correlation between bilateral microphone signals is much lower than in free-field. Since the output of the beamformer is monaural we define a binaural post-filter according to [6]. The binaural post-filter $\mathbf{H}_{\text{Bin}}[m]$ controlled by the beamformer output is real-valued and therefore it preserves the interaural phase-difference between the two reference inputs from the left and right hearing-aid [3, 6].

4. SIMULATION RESULTS

The performance of the proposed algorithms for DOA estimation and for binaural noise reduction based on the imperfect real-world DOA estimates will be evaluated in the following. For simulations diffuse noise signals were generated by summing up speech-colored random noise filtered with measured HRTFs from all directions to simulate a 2D-isotropic noise field. A moving speaker was added for different input SNRs. The block length for all simulations was chosen to $L_{\text{Bl}} = 256$ with an overlap of 128 samples at a sampling frequency of $f_s = 16$kHz. The FFT-length was 512 samples, which means a zero padding factor of two. The number of possible angles was chosen to $I = 37$ which leads to a resolution of 5° for a range of $\Phi = -90^\circ..90^\circ$. The threshold for the maximum tracking speed of the algorithm was fixed to $\xi = 5^\circ$.

Fig. 4 shows the mean estimation error of the DOA estimator

$$\bar{e}_\Phi = \frac{1}{|\mathcal{A}|} \sum_{\mathcal{A}} \Phi - \hat{\Phi} \tag{12}$$

for different input SNRs. Here Φ and $\hat{\Phi}$ are the true and the estimated direction of arrival, respectively. $\mathcal{A}$ is the set of frames where speech is present and $|\mathcal{A}|$ its cardinality.

It can be seen from Fig. 4 that an estimation of the direction of arrival drastically fails if free-field assumptions are made (dash-dotted line). The use of the (in practice unknown) true HRTFs (solid line) lead to the best DOA estimates. The estimation using the head model according to eq. (5) only leads to a slight degradation and thus is a feasible approximation for the unknown true HRTF.

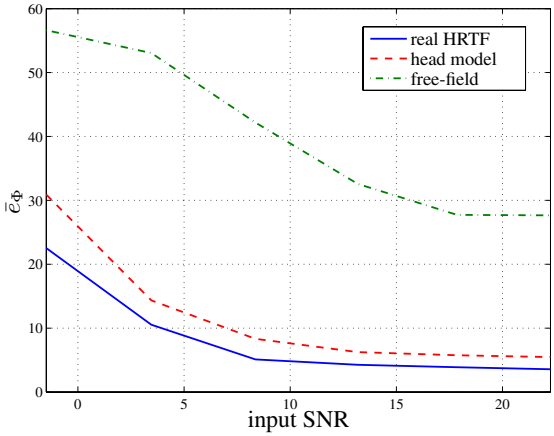


Fig. 4. Estimation error for a DOA estimator for different assumptions for the propagation vector over the input SNR.

For low input SNR (< 8dB) the estimation errors increase thus DOA estimation based on the dual delay line approach becomes unreliable. This is a general problem since the approach is based on looking for and comparing signal powers from different directions. For low SNR the signal power difference between clean speech + noisy speech from the desired direction and noisy speech from other directions is not sufficient for a reliable estimate. This result was also reported by other authors, e.g. [8]. Thus for low input SNR other DOA estimation methods should be applied, see e.g. [9] for an overview.

In Fig. 5 and 6 the performance of the binaural noise reduction scheme relying on real DOA estimates is evaluated by means of the Signal to Noise Ratio Enhancement (SNRE) and the Perceptual Similarity Measure (PSM) [10]. PSM is a speech quality measure from PEMO-Q [10] which estimates the perceptual similarity between the processed signal and a clean speech reference. This measure has shown a high correlation with subjective overall quality ratings [11]. Here the PSM is measured between the clean speech component at the left (right) reference microphone and the left (right) output of the binaural post-filter.

Fig. 5 shows the segmental SNRE between the left (right) output of the binaural post-filter and the left (right) reference channel. The SNRE is the difference of the Signal to Noise Ratio (SNR) at the output of the noise reduction scheme and a reference input SNR. It can be seen from Fig. 5 that if the binaural noise reduction scheme relies on DOA estimates based on free-field assumptions hardly any SNR enhancement is achieved (dash-dotted line). Although the use of true HRTFs leads to the best results (solid line), relying on the head model (dashed line) is capable of improving noisy speech when a head-mounted noise reduction device is applied. Fig. 5 gives the impression that the sound quality improvement increases for lower input SNRs. From Fig. 4 it is clear that this impression is misleading because mean DOA estimation errors at input SNRs lower than 5 dB are not satisfactory.

In Fig. 6 the PSM is shown which better reflects the perceived audio quality. Here it can be seen that the overall sound quality decreases drastically for lower input SNR. Again the results for the

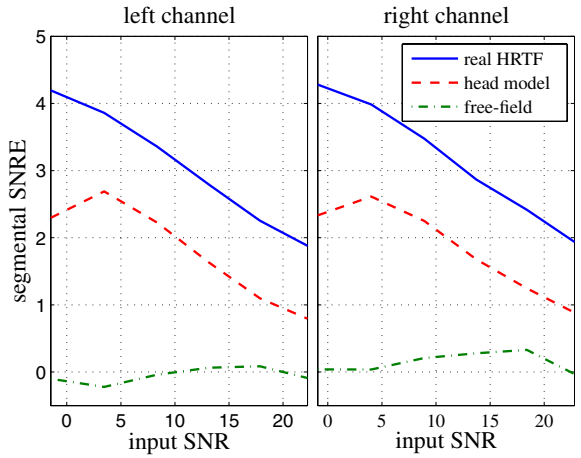


Fig. 5. SNRE of the beamformer steered by the DOA estimate for different input SNR.

head model give a good approximation for the real HRTFs, while free-field assumptions lead to a much lower sound quality.

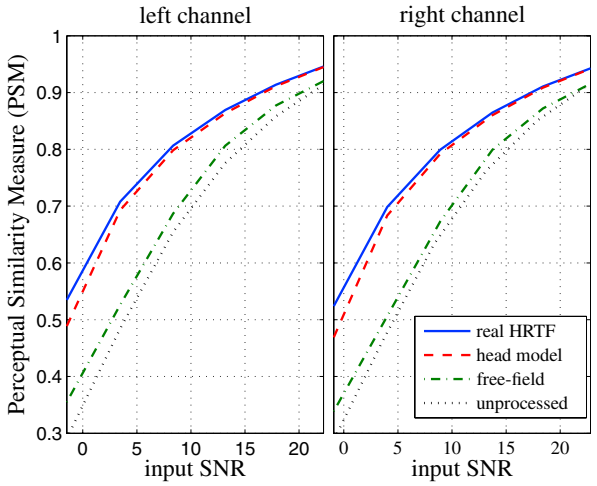


Fig. 6. PSM of the beamformer steered by the DOA estimate for different input SNR.

The so-called ΔPSM [11], which is the difference between the dotted line (unprocessed) and the particular PSM curve shows the quality improvement achieved by the processing. We see that for low input SNR the ΔPSM values are higher, which means that the improvement is better, but that the overall quality of the output signal is very poor. The ΔPSM values match with the SNRE curves from Fig. 5 but Fig. 6 additionally shows the overall quality and thus is more appropriate to compare the different methods.

5. CONCLUSIONS

In this work we analyzed the direction of arrival estimation method after Liu which is based on the delay line approach for the purpose of DOA estimation for hearing aid applications. It could be shown that the underlying free-field assumptions do not lead to satisfactory results and head related transfer functions have to be considered. Since in general it is impossible to estimate the true HRTFs, simulations based on a head model were performed, which showed good results for moderate input SNR. However, for low SNR environments the delay line approach is not capable to deliver reliable results and thus further methods need to be investigated for comparison.

6. REFERENCES

[1] C. Liu, B.C. Wheeler, D. O'Brian, R.C. Bilger, C.R.Lansing, and A.S. Feng, "Localization of Multiple Sound Sources with Two Microphones," *Journal of the Acoustical Society of America (JASA)*, vol. 108, no. 4, pp. 1888–1905, Oct. 2000.

[2] J. E. Greenberg and P. M. Zurek, "Microphone Array Hearing Aids," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 11, pp. 229–253. Springer, 2001.

[3] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by Means of Objective Perceptual Quality Measures," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA*, Oct. 2007.

[4] R.O. Duda and W.L. Martens, "Range Dependence of the Response of a Spherical Head Model," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

[5] P.C. Brown and R.O. Duda, "A Structural Model for Binaural Sound Synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sept. 1998.

[6] T. Lotter and P. Vary, "Dual Channel Speech Enhancement by Superdirective Beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. Article ID 63297, pp. 1–14, 2006.

[7] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 2, pp. 19–38. Springer, 2001.

[8] L. Calmes, G. Lakemeyer, and H. Wagner, "Azimuthal Sound Localization using Coincidence of Timing Across Frequency on a Robotic Platform," *Journal of the Acoustical Society of America (JASA)*, vol. 121, no. 4, pp. 2034–2048, Apr. 2007.

[9] G. Doblinger, "Localization and Tracking of Acoustical Sources," in *Topics in Acoustic Echo and Noise Control*, chapter 6, pp. 91 – 122. Springer, Berlin - Heidelberg, 2006.

[10] R. Huber and B. Kollmeier, "PEMO-Q - A New Method for Objective Audio Quality Assessment using a Model of Auditory Perception," *IEEE Trans. on Audio, Speech and Language Processing*, vol. Special Issue on Objectiv Quality Assessment of Speech and Audio, 2006.

[11] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Measures for the Evaluation of Noise Reduction Schemes," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005.

Work supported by the EC (DIRAC project IST-027787), BMBF and DFG

OBJECTIVE PERCEPTUAL QUALITY ASSESSMENT FOR SELF-STEERING BINAURAL HEARING AID MICROPHONE ARRAYS

Thomas Rohdenburg[1], *Stefan Goetze*[2], *Volker Hohmann*[1], *Karl-Dirk Kammeyer*[2] *and Birger Kollmeier*[1]

[1]University of Oldenburg
Medical Physics Group
26111 Oldenburg, Germany
thomas.rohdenburg@uni-oldenburg.de

[2]University of Bremen
Dept. of Communications Engineering
28334 Bremen, Germany
goetze@ant.uni-bremen.de

ABSTRACT

In this study a self-steering beamformer with binaural output for a head-worn microphone array is investigated in simulated and real-world conditions. The influence of the underlying sound propagation model on the estimation accuracy of the direction of arrival (DOA) estimation algorithm and the overall performance of the combined DOA-beamformer-system is evaluated. For this, technical performance measures as well as objective quality measures based on perceptual models of the auditory system are used. The self-steering beamformer showed better performance than a beamformer with fixed look-direction for SNR values above -2 dB if the propagation model includes at least a coarse head model.

Index Terms— Direction of arrival estimation, Array signal processing, Noise Reduction, Hearing aids, Perceptual audio quality estimation

1. INTRODUCTION

Multi-channel noise reduction schemes are promising solutions for hearing aids as they are capable to exploit the spatial distribution of the interfering signals. Thus, they lead generally to less signal distortion than single-channel noise reduction algorithms. For head-worn microphone arrays it is usually assumed that the look-direction is fixed at zero degrees, and that the user always turns his or her head towards the desired signal. This may become unsatisfying for the hearing aid user in particular for algorithms with a high spatial selectivity and if the signal of interest is moving. In this contribution a combination of a binaural beamformer [1, 2] and an automatic steering (electronic control of the look direction) based on the Generalized Cross Correlation (GCC) approach by Knapp and Carter [3] is applied. The importance of a proper model of wave propagation is investigated for a head-worn DOA-beamformer system. Furthermore, the performance of the system is evaluated in terms of estimation errors and signal-quality by means of objective perceptual measures that are based on models of the auditory system. With these measures the influences of inevitably occurring estimation errors can be quantified on a perceptual scale. Based on these results, the optimum compromise between algorithmic complexity and benefit can be derived.

Notation: Vectors and matrices are printed in boldface while scalars are printed in italic. k is the discrete time index and m the discrete frequency index. The superscripts ^{T}, *, and ^{H} denote the transposition, the complex conjugation and the Hermitian transposition, respectively.

Work supported by EC (DIRAC project IST-027787), HearCom-Project (IST-004171), BMBF and DFG

2. SIGNAL MODEL AND BINAURAL MULTI-CHANNEL NOISE REDUCTION

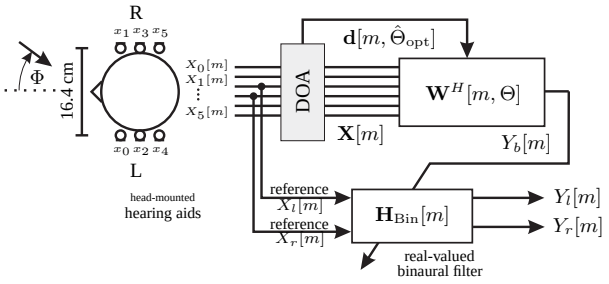


Fig. 1. Signal model and beamformer setup.

The noise reduction scheme used in this contribution is depicted in Fig. 1. With two 3-channel behind-the-ear (BTE) hearing aid shells mounted on a Brüel & Kjær (B&K) head and torso simulator (HATS), 6-channel head related transfer functions (HRTFs) were recorded in an anechoic room and in an office environment (reverberation time $\tau_{60} = 300$ ms) from different directions. A moving target signal was generated by filtering a speech signal with time-varying HRTFs that change due to a pre-defined virtual azimuth path (Fig. 2). Real-world environmental noise has also been recorded in a cafeteria and in an office room. Additionally, an artificial diffuse noise has been generated by summing up a speech-colored random noise that was filtered with HRTFs from all directions to simulate a cylindrical 2D-isotropic noise field. The moving speech signal was mixed with the noise signals at different signal-to-noise ratios (SNRs). In

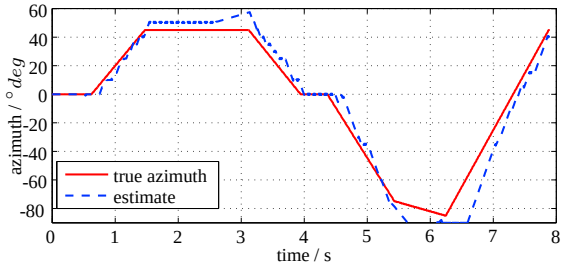


Fig. 2. Virtual azimuth path of moving speech source and its estimate for HM2 at 12 dB SNR.

Fig. 1, $X_i[m]$ denotes the audio-signal transformed into the fre-

quency domain by use of the short time Fourier transform (STFT), where $i = 0..5$ is the channel index. A DOA detection algorithm estimates the target signal's azimuth angle Θ which is used to steer the beamformer to this direction by means of the propagation vector $\mathbf{d}[m,\Theta]$. The beamformer $\mathbf{W}[m,\Theta]$ generates a single channel output $Y_b[m]$ via the well known Minimum Variance Distortionless Response (MVDR) approach [4]:

$$\mathbf{W}[m,\Theta] = \frac{\mathbf{\Gamma}_{NN}^{-1}[m]\mathbf{d}[m,\Theta]}{\mathbf{d}^H[m,\Theta]\mathbf{\Gamma}_{NN}^{-1}[m]\mathbf{d}[m,\Theta]}. \quad (1)$$

$$\mathbf{d}[m,\Theta] = [d_0[m,\Theta], d_1[m,\Theta], \ldots, d_{N-1}[m,\Theta]]^T \quad (2)$$

$$d_i[m,\Theta] = |d_i[m,\Theta]|e^{-j2\pi m\frac{f_s}{M}\tau_i[m,\Theta]},\ i = 0..N-1 \quad (3)$$

The fixed noise-field characteristic is coded in the coherence matrix $\mathbf{\Gamma_{NN}}[m]$ which additionally influences beamformer properties directivity and susceptibility to white noise, and therefore has to be constrained [4, 1]. Both, $\mathbf{d}[m,\Theta]$ and $\mathbf{\Gamma_{NN}}[m]$ depend on to the assumed wave propagation model which may differ from the true (and generally unknown) wave propagation from the source to the microphones. We distinguish four models, free-field (FF), two head models (HM1 [5], HM2 [6]) and the measured anechoic transfer functions from the source to the head-mounted hearing aid microphone array (HRTF). The simplest approach is to use a free-field / far-field assumption (FF), i.e., the sound propagation is modeled as a plane wave without interfering objects in the propagation path. For FF, $\mathbf{d}[m,\Theta]$ has unity magnitude, $|d_i[m,\Theta]| = 1\ \forall(i,m,\Theta)$ and constant group delay $\tau[m,\Theta] = \tau[\Theta]$ that can be calculated from the inter-microphone distance and the angle of incidence. For head-worn arrays it is beneficial to include knowledge about head shadow and diffraction effects [1, 11], especially for lateral target signal sources. Thus, head models by Duda et al. [5, 6] are applied which are effective parametric models that are based on the characteristics of a sphere. In HM1, the interaural time difference (ITD) cues are modeled by Woodworth and Schlosberg's frequency independent ray-tracing formula. The gross magnitude characteristics of the HRTF spectrum, namely the interaural level difference (ILD) cues, are covered by a first order IIR head shadow filter which also accounts for an additional frequency dependent delay at low frequencies [5]. In HM2, near-field effects and interference effects that introduce ripples in the frequency response which are quite prominent on the shadowed side are incorporated as described in [6]. For both head models (HM1, HM2) the frequency dependent group delay $\tau[m,\Theta]$ and magnitude have to be calculated for each microphone and angle of incidence due to [5, 6]. For HRTF, the propagation vector $\mathbf{d}[m,\Theta]$ equals the measured anechoic 6-channel HRTF for the angle of incidence Θ. $\Gamma_{NN}[m]$ can be estimated for a cylindrical isotropic diffuse noise field by integrating the propagation vectors over all directions Θ. For FF, this solution can be calculated via the Bessel function of the first kind of order zero. For the white noise gain constraints and further details see [4].

The binaural output is calculated by a real-valued time-varying post-filter based on [2] that is controlled by the monaural beamformer output Y_b:

$$H_{\mathrm{Bin}}[m] = \frac{\left(|d_l[m,\Theta]|^2 + |d_r[m,\Theta]|^2\right)\Phi_{Y_bY_b}[m]}{\Phi_{X_lX_l}[m] + \Phi_{X_rX_r}[m]} \quad (4)$$

$$Y_l[m] = H_{\mathrm{Bin}}[m]X_l[m] \quad (5)$$

$$Y_r[m] = H_{\mathrm{Bin}}[m]X_r[m] \quad (6)$$

Here $X_l[m], X_r[m]$ (see Fig. 1) denote the reference input signals and $d_l[m], d_r[m]$ the propagation coefficients for the estimated signal direction Θ_{opt}, at the left and right reference microphone, respectively. $\Phi_{Y_bY_b}[m]$, $\Phi_{X_lX_l}[m]$ and $\Phi_{X_rX_r}[m]$ are the power spectral density estimates for the signals $Y_b[m], X_l[m], X_r[m]$, respectively. As depicted in Fig. 1 we chose channel 3 and 4 as reference channels for the left and right site. For a detailed analysis of the binaural output see [1].

3. DIRECTION OF ARRIVAL ESTIMATION

Direction of arrival estimation is done by estimating the signal delay between microphone pair $x_l[k], x_r[k]$ via the PHAT-GCC (Phase Transform Generalized Cross Correlation) [3] which has been proven to give reliable estimates for various environments:

$$\tau_d = \arg\max_k R_{x_lx_r}[k] \quad (7)$$

with the (PHAT) generalized cross correlation [3]

$$R_{x_lx_r}[k] = \frac{1}{L_{\mathrm{DFT}}}\sum_{m=0}^{L_{\mathrm{DFT}}-1}\frac{\Phi_{x_lx_r}[m]}{|\Phi_{x_lx_r}[m]|}e^{j\frac{2\pi}{M}mk},\ k = 0..L_{\mathrm{DFT}}-1 \quad (8)$$

Typical signal delays that occur between the left and right microphones are about $8.3\mu s/1^\circ$deg in the range of $\pm30^\circ$ deg. For a sampling rate of 16 kHz these are 7.5° deg per sample. Thus, an appropriate oversampling of the generalized cross-correlation $R_{x_lx_r}[k]$ is suggested.

The time-delay of arrival due to diffraction is longer for lateral signals then expected in the free-field case. Therefore the time-delay corresponds to other angles of incidence for the head models than for the free-field. Fig. 3 depicts deviations that occur due to a wrong delay-to-azimuth mapping. Fig. 3(a) shows the time delay of arrival

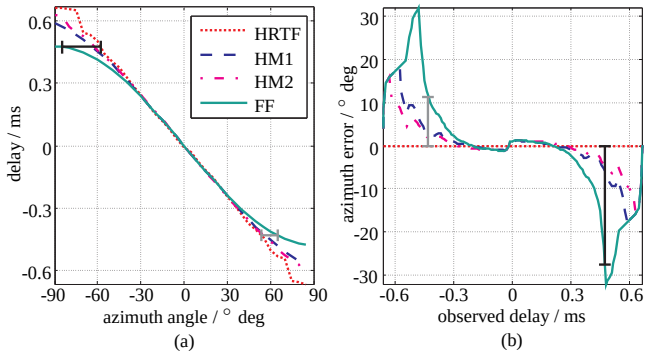


Fig. 3. Azimuth error for different time delays τ_d and propagation models.

between the microphones $x_l[k]$ and $x_r[k]$ against the azimuth angle for different propagation models. Between $\pm30^\circ$ the dependency is almost linear and only little deviations between the propagation models exist. For more lateral angles the differences increase due to the increased traveling time of the sound signals around the human head. In Fig. 3(b) the deviation of the estimated angle for the propagation model and true angle as determined from the measured HRTF is depicted. Note that for the free-field model (FF) delays beyond ±0.5 ms are assigned to $\pm90^\circ$. Therefore, the azimuth error decreases for values beyond these maximum delays. The gray and black bars show the corresponding values in (a) and (b). It can be seen that the head models give a better approximation of the true time delay than FF assumptions. Although the group delays for the

head models are frequency dependent [5], these effects are omitted here as they only apply for low frequencies (< 200 Hz). A maximum tracking speed of the DOA estimator is limited to $125^\circ/s$ as described in [11] to avoid sudden peaks in the DOA estimate that lead to severe disturbances of the subsequent beamformer. A simple speech activity detector based on the magnitude of $R_{x_l x_r}[k]$ is applied by updating the DOA estimate only if $R_{x_l x_r}[k]$ is greater than a threshold ξ. During speech pauses a tracking algorithm based on the last estimates continues the update of the azimuth estimate. However for the application in a hearing aid it might be useful to apply more sophisticated tracking algorithms that increase the robustness of the estimate while at the same time allowing for a quick change of direction due to a moving speaker. Here, our main focus lies on understanding the principle problems due to imperfect propagation models.

4. QUALITY ASSESSMENT

It has been shown in Fig. 3 that the assumption of an imperfect propagation model leads to systematic errors in the estimation of the signal-source direction. As we are interested in the influence of these estimation errors on the performance and signal quality for realistic scenarios we propose three performance measures.

SNRE: The SNR-Enhancement (SNRE) is the difference of the SNR at the output of the beamformer and a reference input-SNR, both measured in dB. For binaural systems the SNRE is calculated between the left (right) output of the binaural post-filter and the left (right) input at the reference microphone, respectively; by simply taking the mean SNRE a better-ear effect would be ignored.

PSM / ΔPSM: The quality measure PSM from PEMO-Q [7] estimates the perceptual similarity between the processed signal and the clean speech source signal. It has shown high correlations between objective and subjective data and has been used for quality assessment of noise reduction schemes in [1, 8, 9]. PSM increases with increasing (input) SNR. As we are interested in the quality enhancement introduced by the algorithm, we use the deduced measure ΔPSM that is calculated as the difference between the Perceptual Similarity Measure (PSM) of the output and of the unprocessed input signal.

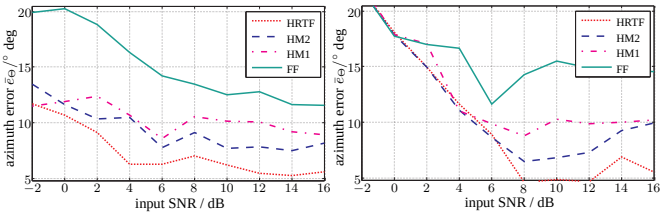
Binaural SRT / ΔSRT: The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. In [10] a binaural model of speech intelligibility based on the equalization-cancelation (EC) processing by Durlach had been defined which is able to predict the SRT with high accuracy. If the estimated SRT for the output of a noise reduction scheme is lower than for the input signal this means that the speech intelligibility has increased due to the algorithm. However, as the speech intelligibility is a nonlinear function of the SNR and other signal features such as the preservation of binaural cues, we use the difference between output and input SRT, namely the ΔSRT, as an indirect measure for the increase of intelligibility. The binaural SRT measure as described in [10, 1] assumes a spatially stationary source configuration. To be applicable to moving sources it had to be extended to a block-wise measure with subsequent averaging across blocks.

5. SIMULATION RESULTS

5.1. DOA Estimation Error

Fig. 4 shows the mean azimuth estimation error of the DOA algorithm $\bar{e}_\Theta = \frac{1}{|\mathcal{A}|}\sum_{\mathcal{A}} \Theta - \hat{\Theta}$ over the input SNR for the four propagation models. Here, Θ and $\hat{\Theta}$ are the true and the estimated direction of arrival, respectively. $\mathcal{A}$ is the set of frames where speech is present and $|\mathcal{A}|$ its cardinality. In artificial diffuse noise, Fig. 4(a), the mean azimuth error for the head models is below 15° degree at an SNR of -2 dB and falls below 10° for an SNR $> 2-4$ dB depending on the exactness of the model. The measured (in practice generally unknown) HRTF shows the best performance followed by HM2 which seems to be a feasible approximation. Assuming free-field, $\bar{e}_\Theta$ is persistently $3-7\,^\circ$ greater than for the head models.

The performance for this algorithm in a recorded real-world office environment with ambient noise, Fig.4(b), is worse at -2 dB SNR than for artificial diffuse noise, but $\bar{e}_\Theta$ also falls below 10° for an input SNR > 5 dB for the head models. Compared to the results gained in [11] where a DOA estimator based on the dual delay line approach was evaluated, it can be can be stated that the GCC-PHAT algorithm performs much better, particularly in noisy conditions.



(a) Articial diffuse noise (b) Office noise

Fig. 4. Mean DOA error in different noise conditions.

5.2. Objective Perceptual Quality of the whole system

Fig. 5 shows the performance measures described in Section 4 over the SNR of the input signal (SNR_{in}). If not indicated otherwise, results are shown for the diffuse noise. The Signal to Noise Ratio Enhancement (SNRE) in Fig. 5(a) slightly decreases with increasing SNR_{in} which is a fact common to all noise reduction systems as for infinite SNR_{in} the SNRE converges to zero. The ideal system (solid black line) has *a priori* information about the direction of arrival and uses the measured HRTF as a propagation model. Therefore, it should set the upper performance limit. Also, it would be expected that the systems with the most exact propagation model (HRTF and HM2, before HM1 and FF) have the highest SNRE. However, this is not seen in the right channel where FF (solid green) crosses HM2 (dashed blue). This is an artifact of the broadband SNRE measure that is suboptimal for quality assessment, as it does not incorporate signal distortions. For PSM in Fig. 5(b) the ranking behaves as expected: The ideal system sets the upper limit and the system with the fixed look direction to 0° shows the worst performance. The absolute PSM (not shown here) for the ideal system lies between 0.6 and 0.9 (where values close to 1 mean that the signal is perceptually undistinguishable from the clean speech [7]). A negative ΔPSM shows a signal degradation compared to the unprocessed signal, e.g., FF and 0° fixed at SNR>12 dB. For the head models ΔPSM is consistently higher than for the fixed system, whereas for FF the quality enhancement is marginal. Fig. 5(c) shows the decrease of the Speech Reception Threshold (SRT) due to the noise reduction that also incorporates the speech intelligibility benefit due to the preservation of binaural cues. Again, the ranking is consistent with the exactness of the propagation model. For input SNR values where the DOA estimation has low errors, HM2 and HRTF have less than 0.5 dB higher SRT than the ideal system. For FF, ΔSRT lies 1.5 dB higher than for the ideal system. All self-steered systems with head models have a lower SRT than the system fixed to 0° degree look-direction for all SNR_{in} whereas for FF this is the case at an SNR_{in} > 3dB. In those

cases steered systems are superior to fixed systems for the given input signals. Fig. 5(d) and 5(e) show the performance for real-world recordings in the office room mixed with (d) office ambient noise and (e) babble noise from a cafeteria. A ΔSRT close to the ideal system indicates a good performance which is given for the head models at a $\mathrm{SNR}_{in} > 4$ dB for the ambient noise (d) and a $\mathrm{SNR}_{in} > 9$ dB for babble noise (e). For FF, ΔSRT is significantly higher in (d) and it is close to the fixed system in (e). In summary it can be stated that for the difficult cafeteria noise condition where sudden correlated noise sources may occur, DOA estimation performance for a fast moving target signal source at low SNR is poor. However, for input SNR_{in} >9 dB automatic-steered systems are favorable, given an appropriate propagation model.

6. CONCLUSION

We presented a self-steering multi-channel noise reduction system with binaural output applicable to hearing aids. Estimation errors have been analyzed under the assumption of different wave propagation models. For a fast moving speech source under different simulated and real-world noise conditions, algorithm performance was evaluated using technically based measures and objective perceptual quality measures based on auditory models. The results show that for signal-to-noise ratios (SNRs) greater -2 dB self-steering systems are superior to fixed systems if a certain complexity of the propagation model is met. The DOA-beamformer system performs best in diffuse or ambient noise conditions. However, in difficult noise conditions such as cafeteria noise, the performance is lower than for a simulated system with a priori knowledge about the direction of arrival.

7. REFERENCES

[1] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by means of Objective Perceptual Quality Measures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct. 2007.

[2] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 63 297, 14 pages, 2006.

[3] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[4] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[5] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[6] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 5, pp. 3048–3058, 1998.

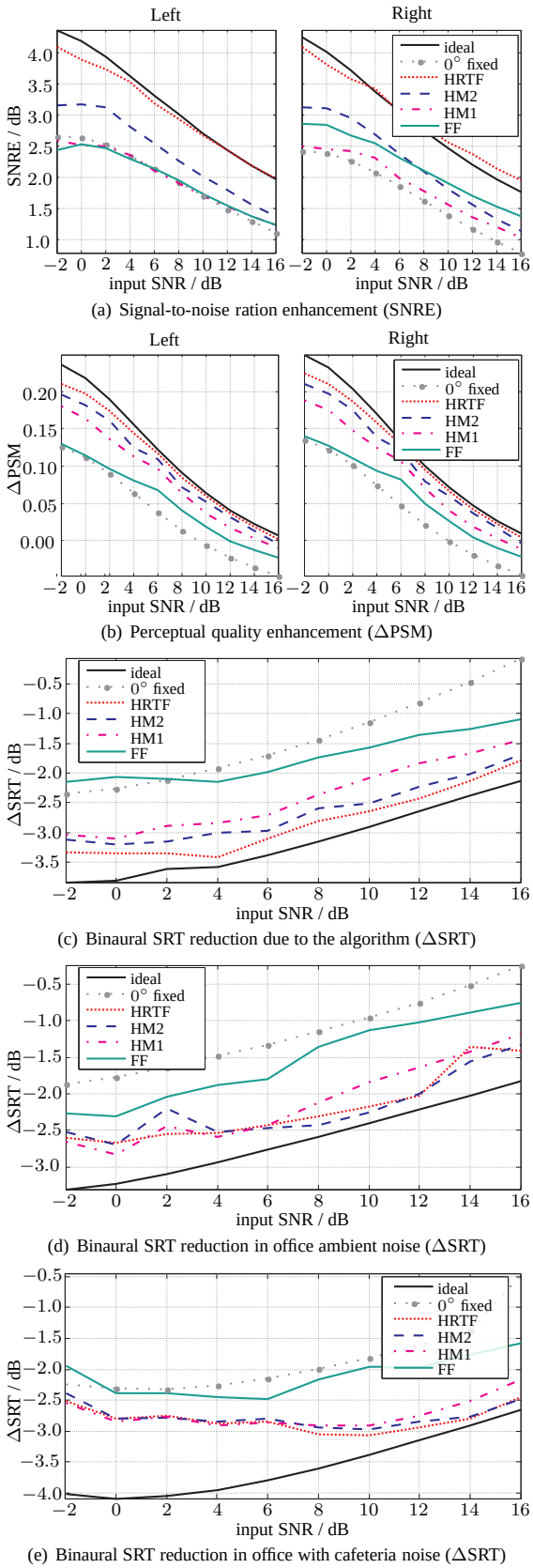
[7] R. Huber and B. Kollmeier, "Pemo-Q -A new Method for Objective Audio Quality Assessment using a Model of Auditory Perception." *IEEE Trans. on Audio, Speech and Language Processing*, 2006, special Issue on Objective Quality Assessment of Speech and Audio.

[8] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 169–172.

[9] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Subband-based parameter optimization in noise reduction schemes by means of objective perceptual quality measures," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, September 12-14 2006.

[10] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[11] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of Arrival Estimation based on the Dual Delay Line Approach for Binaural Hearing Aid Microphone Arrays," in *Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Xiamen, China, Nov. 2007.



(a) Signal-to-noise ration enhancement (SNRE)

(b) Perceptual quality enhancement (ΔPSM)

(c) Binaural SRT reduction due to the algorithm (ΔSRT)

(d) Binaural SRT reduction in office ambient noise (ΔSRT)

(e) Binaural SRT reduction in office with cafeteria noise (ΔSRT)

Fig. 5. Objective quality assessment of DOA plus beamformer system with different wave propagation models