Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Downstream speech enhancement in a low directivity binaural hearing aid

Marinus M. Boone, Robertus C.G. Opdam and Anton Schlesinger

Laboratory of Acoustical Imaging and Sound Control, Delft University of Technology, Delft, The Netherlands

PACS: 43.60.Fg, 43.60.Jn, 43.66.Yw, 43.71.Gv

ABSTRACT

The combination of a binaural beamformer with an auditory model-based localizer and post-processor is presented. The intention of the design was to maintain a high degree of quality and listening ease while enhancing speech intelligibility. Therefore, a binaural and fixed minimum variance distortionless response beamformer was employed to establish a medium directivity at low self noise and high binaural naturalness. By applying models of binaural interaction at the two-channel output, sound sources are localized and separated. The process of binaural interaction was extended by an auditory model of across frequency interaction in the localizer and a model of modulation perception in the post-processor. To adapt the post-processor to the scene and to keep the introduced distortion at a low level, the parametric output of the localizer was used to steer the aperture of a spatial filter during post-processing. For that reason, Bayes theorem was applied to calculate the *a posteriori* probability of the target source in a complex acoustic scene and to use this probability for the formulation of a data-driven filter. The localizer and post-processor were assessed in a range of acoustic arrangements using an objective intelligibility and quality measure for nonlinearly processed speech. The results show a significant improvement in situations with one interferer and no decline of intelligibility and quality in more complex situations.

INTRODUCTION

There are generally two approaches in speech enhancement. One is realized through linear methods, the other through non-linear methods of sound scene segmentation. Linear methods are theoretically capable to fully reconstruct the desired speech. This however requires an impractical amount of processing time and impractical design solutions. Nevertheless, many successful linear approaches in speech enhancement exist with a lower separation power. Well known is the beamformer solution that combines a microphone array and a spatial filter. There are different beamforming methods that realize a compromise between target gain and self noise. The minimum variance distortionless response (MVDR) beamforming method allows for an adjustment of this compromise with a stabilization constant. The resulting gain of an MVDR beamformer is generally limited by the physical dimensions of the array as well as the required minimal processing robustness and the acoustical field for which it was optimized.

A non-linear method of speech enhancement can approximate an optimal filter with much less computational effort. It is often realized in a mask-based manner that acts in a time-frequency domain representation of the input signal. In spite of that, these non-linear mask-based approaches are difficult to handle. As such, they introduce signal distortion and their processing is subject to many design parameters. Moreover, the impact of non-linear signal distortion on speech intelligibility is difficult to evaluate. Subjective tests are laborious and most instrumental measures fail in the prediction of speech intelligibility for speech enhancement algorithms [1].

There are many local selection methods, i.e., whether a time-frequency bin belongs to the interferer or to the target source, for the generation of masks. The field of computational auditory scene analysis (CASA) developed algorithms of binaural interaction that have shown the ability to realize a benefit in speech intelligibility [2]. These approaches generally feature a robust processing in a wide range of acoustical situations, but also need a difficult and laborious parameter adjustment.

Localization algorithms can be used to detect the acoustic scene and to steer the post-processor. A range of auditory inspired localization algorithms have proven to imitate much of the human localization performance [3, 4]. Furthermore, they can operate in parallel to the online speech processing and can continuously update a data-driven post-processing filter.

In here, we combined a robust low directivity MVDR beamformer, which is implemented in a pair of spectacles, with the auditory model-based localization algorithm of Albani et al. [4] and the model-based post-processor of Kollmeier and Koch [5]. Both algorithms have each shown to be effective and robust realizations. In particular the combination of a two-channel beamformer and a binaural mask-based algorithm is considered to achieve a substantial improvement over single methods of speech enhancement [5]. Mainly because the pre- and post-processor add to a higher SNR gain in a complementary manner. I.e., the post-processor is put into its efficient SNR working range and the common front-back discrimination ambiguity is alleviated by the superposition of the directional pattern of the beamformer. Moreover, since the gains add, the requirements of speech intelligibility enhancement are lower for each processor, which is generally for the benefit of a lower target distortion and a natural listening experience.

In the following section we describe the algorithmic approach of the three constitutive parts. Thereafter the combined processing scheme is assessed with the three level intelligibility and quality index of Kates and Arehart [6]. These measures showed a fair degree of accuracy and were adapted to binaural listening. We conclude this contribution with a summary and an outlook.

ALGORITHM

The overall processing scheme is sketched in Figure 1. The combined processing scheme is realized in a sequential order

and is entirely performed in the digital frequency domain. The overall output demand is to enhance the signal in the frontal line of sight and to attenuate noise from all other directions. The two combined methods use different constraints to perform this task. The MVDR beamforming algorithm does not allow target distortion, while the binaural post-processor intrinsically performs target distortion and has therefore to be implemented with great care. A dichotic signal is maintained throughout the processing and presented to the ears. The following divisions give a synopsis of the applied MVDR beamforming method, the localizer and the post-processor.

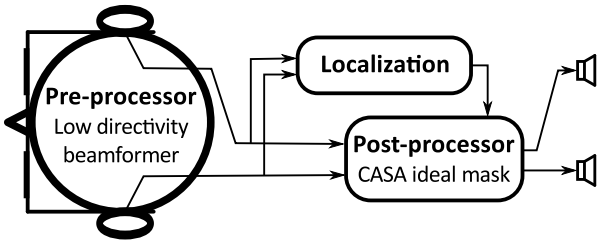

Figure 1: Combined processing scheme of a binaural MVDR beamformer, realized in the arms of a pair of spectacles, and an auditory model based localizer and post-processor.

Binaural Beamformer

The technique of MVDR beamforming facilitates maximal directivity gain for a given array configuration and assumptions of the noise field. The optimal filters are calculated in an optimization routine that maximizes the array's gain with the filters $\mathbf{F}(k)$ as parameter. The method generally suffers from a large noise sensitivity. The optimization procedure was therefore supplemented to achieve a maximum gain at a predefined noise sensitivity [7, 8]. The array of microphones is alined in an endfire configuration along the z-axis and resides in the arms of a pair of spectacles. The optimal filters are then calculated as follows: First, the propagation delay vector $\mathbf{W}(k,\theta,\phi)$ for $\theta = 0$, $\phi = 0$ for a plane wave can be formalized as:

$$\mathbf{W}(k) = [e^{jkz_1}\ e^{jkz_2} \dots\ e^{jkz_N}]^{\mathrm{T}}, \tag{1}$$

for omnidirectional transducers at positions $z_1 \dots, z_N$ (k is the wavenumber and $j = \sqrt{-1}$). When multiplying the propagation delay vector with the filter vector and summing them, the array frequency response to a plane wave from the target direction reads:

$$\Gamma_{\mathrm{T}}(k) = \mathbf{F}^{\mathrm{T}}(k)\mathbf{W}(k). \tag{2}$$

The array directivity factor, i.e., the averaged squared arrays target response $|\Gamma_{\mathrm{T}}(k,\theta,\phi)|^2$ divided by the average squared array response of sound incident from all directions $|\Gamma(k,\theta,\phi)|^2$, is employed as an objective function in the optimization process [8]. In matrix notation, the directivity factor is consequently given by:

$$Q(k) = \frac{\max_{\theta,\phi}\{\mathbf{F}^{\mathrm{H}}(k)\mathbf{W}^{*}(k)\mathbf{W}^{\mathrm{H}}(k)\mathbf{F}(k)\}}{\mathbf{F}^{\mathrm{H}}(k)\mathbf{S}^{\mathrm{T}}(k)\mathbf{F}(k)}, \tag{3}$$

where $(\circ)^{\mathrm{H}}$ is the Hermitian transpose, $(\circ)^{*}$ is the conjugate-complex operator and $\mathbf{S}$ is the cross-spectral density matrix of the microphones in the noise field. It is assumed that the noise field is isotropic and uniform, hence in the endfire configuration of the microphones the elements of $\mathbf{S}$ are given by:

$$S_{mn} = \frac{\sin(k(z_m - z_n))}{(k(z_m - z_n))}. \tag{4}$$

The optimization routine of the optimal beamforming method, which was developed in [9] and adapted to the presented application in [8], is defined as:

$$\min_{\mathbf{F}(k)}\{\mathbf{F}^{\mathrm{H}}(k)\mathbf{S}(k)^{\mathrm{T}}\mathbf{F}(k)\}, \tag{5}$$

while adhering to $\Gamma_{\mathrm{T}}(k) = 1$ and a predefined maximal allowed noise sensitivity $\mathbf{F}^{\mathrm{H}}(k)\mathbf{F}(k) < \Psi_{\max}$. Following [8], the solution to the problem is:

$$\mathbf{F}^{\mathrm{T}}_{\mathrm{opt},\beta}(k) = \frac{\mathbf{W}^{\mathrm{H}}(k)(\mathbf{S}(k) + \beta(k)\mathbf{I})^{-1}}{\mathbf{W}^{\mathrm{H}}(k)(\mathbf{S}(k) + \beta(k)\mathbf{I})^{-1}\mathbf{W}(k)}. \tag{6}$$

In this formula $(\circ)^{-1}$ indicates the inverse of a matrix, $\mathbf{I}$ the identity matrix and $\beta(k)$ is a stabilizing factor that determines the level of the sensor noise that is adapted during the optimization.

The filters were calculated for an array with 4 omnidirectional transducers at a length of 72 mm. Subjective listing tests yielded an advantage in SNR for normal hearing subjects of 7.5 dB and for hearing impaired subjects of 6.2 dB [8]. This notable enhancement of speech intelligibility attracted commercial attention and the development was further adapted for a casing of the entire processing of the highly directive hearing aid in the arms of a pair of spectacles. For daily use three programs of different directivity (omnidirectional, low and high directivity) were calculated. The filters are applied in a weighted overlap-add design in 31 subbands. A measurement of the directivity index ($DI(k) = 10\log_{10}(Q(k))$) of the hearing glasses worn by an artificial head showed an averaged gain in SNR of 7.2 dB in the high directivity mode and 4.4 dB for the low directivity mode [7]. The directivity index as a function of frequency for each program-mode and an artificial head are shown in Figure 2. The different modes show a similar progress of the directivity index. The array offers directivity from approx. 100 Hz to 5000 Hz in the low directivity mode as well as in the high directivity mode, and covers the spectrum of speech. In the omnidirectional mode, only the front microphones of the arrays are used which are at a prominent position in front of the head and therefore also result in some directivity.

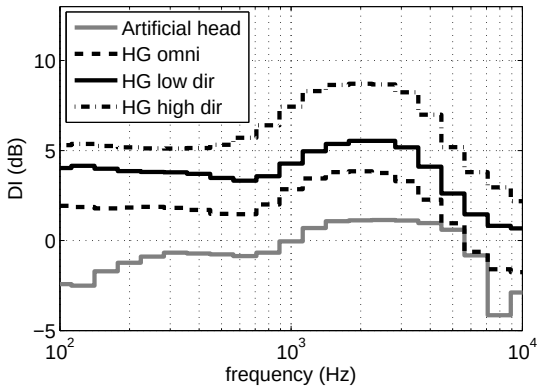

Figure 2: Directivity index in 1/3 octave-bands of three programs of the beamformer output and an artificial head at one ear.

The hearing glasses in the low directivity mode serve as the front-end in the here presented combined processing scheme and deliver a fixed gain in SNR to the subsequent units. The fixed beamformer is well suited for this task, as it provides a fixed binaural pattern, which is a prerequisite of the here applied binaural localizer and post-processor.

Localization

For acoustical localization purposes often the use of the independent component analysis (ICA) and the time delay of arrival (TDOA) method in combination with microphone arrays are suggested [10, 11]. These approaches can deliver perfect results, however the number of sources has to be known beforehand, large arrays of microphones are required, a free field placement of the receivers is needed or ambiguous (front-back) confusions occur due to symmetry. Generally the number of sources is not known in an acoustical environment and because a hearing aid is carried on the body, the use of large microphone arrays is not desirable. Furthermore, due to head and body shadow effects the free field placement of receivers is not possible. Therefore the choice for a CASA approach was made. CASA in fact utilizes the head shadow effect and binaural hearing to process the signal information. We adopted the basic principles from Albani et al. [4], which is based on neurophysiological studies of the Barn Owl [12, 13], and extended it with a preselection of time windows by the Internal Coherence (IC), multiple source detection in one time window and coincidence detection between the internal level- and time differences (ILD and ITD) cues. The details of this algorithm are given in the next section. The data processing within the algorithm is summarized in Figure 3. Each process block in this figure will be briefly discussed step by step hereafter. The first step is to transform the time

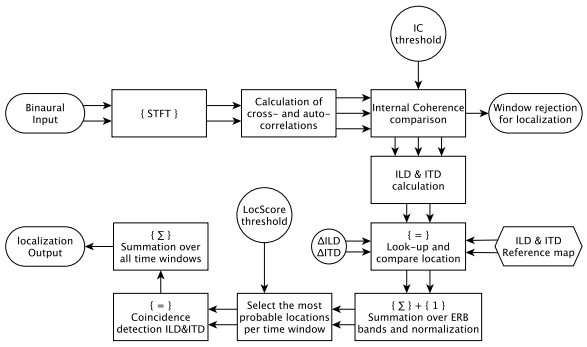

Figure 3: Flow chart of localization algorithm structure.

domain data (binaural input) into the frequency domain, by the use of an overlapping short time Fourier transform (STFT) representation [14]. A Hann window is used to probe the time data. A 50% overlap of the time windows assures the conservation of the total information, due to the shape of the Hann window. A fast Fourier transform (FFT) is carried out over each time window. The result is a collection of time-frequency (TF) bins, which represent the signal content over a region in time and frequency. In the next step, cross- and auto-spectra are calculated from these TF-bins:

$$\Phi_{ll}(f,n) \equiv \left\langle |L(f,n)|^2 \right\rangle, \quad (7)$$

$$\Phi_{rr}(f,n) \equiv \left\langle |R(f,n)|^2 \right\rangle, \quad (8)$$

$$\Phi_{lr}(f,n) \equiv \langle L(f,n)R^*(f,n) \rangle. \quad (9)$$

Respectively, Φ_{ll}, Φ_{rr} and Φ_{lr} are the auto-spectra of the left and right signals and is the cross-spectrum between them. L and R are the STFT representations of the left and right ear. f denotes the frequency bin index, where every bin has a bandwidth of 31.25 Hz. n is the number of the time window. The expectation operator $\langle \circ \rangle$ denotes the average over time and is defined as a first order low pass filter given in equation (10) and characterised by a coefficient $\gamma = exp\left(-\frac{1}{\tau f_s}\right)$, which is dependent on a time constant τ and the time window sample frequency f_s [15]:

$$\langle X(n) \rangle = (1-\gamma)X(n) + \gamma X(n-1). \quad (10)$$

These time averaged cross- and auto-spectra are used to compute the necessary three binaural cues for the localization process; Interaural Level Difference (ILD) ΔL, Interaural Time Difference (ITD) $\Delta\phi$ and the magnitude squared Interaural Coherence (IC) IC:

$$\Delta L(n,_{\text{ERB}}) = \frac{1}{N_b} \sum_{f=f_{start}(\text{ERB})}^{f_{end}(\text{ERB})} 10\log_{10}\left[\frac{\Phi_{ll}(f,n)}{\Phi_{rr}(f,n)}\right], \quad (11)$$

$$\Delta\phi(n,_{\text{ERB}}) = \frac{1}{N_b} \sum_{f=f_{start}(\text{ERB})}^{f_{end}(\text{ERB})} \frac{\arg\left(\Phi_{lr}(f,n)\right)}{2\pi f}, \quad (12)$$

$$IC(n,_{\text{ERB}}) = \frac{\sum_{f=f_{start}(\text{ERB})}^{f_{end}(\text{ERB})} |\Phi_{lr}(f,n)|^2}{\sum_{f=f_{start}(\text{ERB})}^{f_{end}(\text{ERB})} \Phi_{ll}(f,n) \sum_{f=f_{start}(\text{ERB})}^{f_{end}(\text{ERB})} \Phi_{rr}(f,n)}. \quad (13)$$

To apply the tonotopic decomposition of the human hearing, these cues are calculated in bands of equivalent rectangular bandwidths (ERB) [16]. Therefore, there is a summation over the frequency bins f, which belong within the selected ERB-band. For normalization purposes it is necessary to divide by the number of frequency bins used in the specific ERB-band, N_b. In the next step the IC cue is calculated to determine if there is sufficient data present in the current time window. When the IC value is above a certain threshold (between 0 (no coherence) and 1 (total coherence)) the information from that time window will be used for localization. Next, the ILD and ITD cues from the time windows that passed this test are calculated and compared to a map with reference ILD and ITD values per ERB-band. When, within a margin, a matching value of the map is found, the corresponding location gets a score point. The reference maps are built up using Plomp sentences [17] as sources from all possible directions. The output is a score matrix with the number of matches with the reference map values per location for each ERB-band, for both ILD and ITD cues. After that, for each location the scores of all or a selected range of ERB-bands are added up and normalized. Locations which have a higher normalized score than the threshold value will be considered as sources. This process is actually selecting the most probable sources per time window, separately based on the ILD and ITD cue. In this stage both ILD and ITD score vectors contain minor localization errors. Therefore in the next step a coincidence detection principle [18] is used to detect the concurring locations between ILD and ITD for each time window, providing a more robust localization result with a higher confidence interval. This is realized by minimizing the probability of a fault detection. The conditional probability of a fault detection by both ILD and ITD, $p(ILD \cap ITD) = p(ILD|ITD) \cdot p(ITD)$, is generally smaller than the separate probability of a fault detection, $p(ILD)$ or $p(ITD)$. As last step a summation of location scores over time windows is performed to give a certain location stability in time.

To show the performance of this localization algorithm a hypothetical situation with three speakers and a receiver is assumed. The three speakers are at 1 m distance of the receiver and have on average the same sound pressure level. The three speakers are located in front of the receiver at -40°, 0° and 40° azimuth and produce different sentences of the Plomp corpus [17]. Simulations with this situation are performed in three different environments; Anechoic, cafeteria background noise with a SNR of 5 dB and in a reverberant rectangular room of volume 1000 m^3 with $T_{60} = 1$s, which corresponds to a reverberation radius of 1.9 m. The results are given in Figure

4. From the total amount of time-frames, the algorithm used 37.3%, 31.7% and 27.4% of time-frames for the localization process in the anechoic condition, the cafeteria noise and the reverberant environment, respectively.

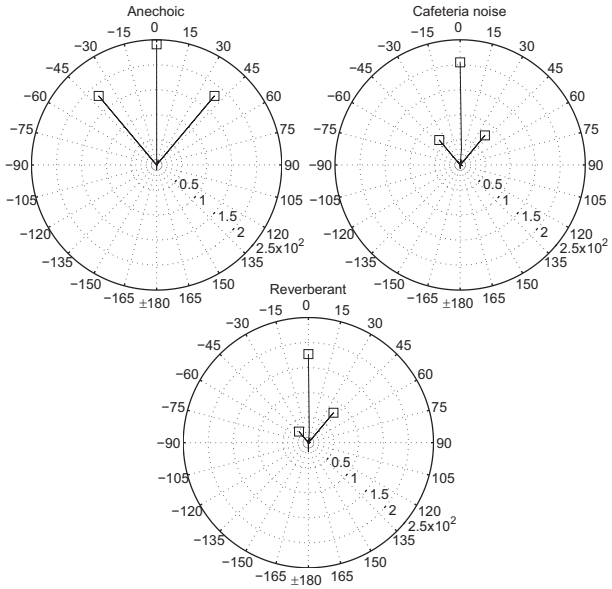

Figure 4: Localization performance of 3 speakers in an anechoic, cafeteria noise (5 dB SNR) and reverberant ($T_{60} = 1.0$ s) environment.

Mask-based Post-processor

The formation of masks is based on a local selection criterion that decides whether a certain time-frequency bin belongs to the target or to the noise. An example of the application based on a simple local SNR criterion is given in Figure 5. A class of successful binaural speech enhancement algorithms employs the deviation of acoustic sources from the mid-line as a selection criterion, i.e., their binaural differences. By defining a weighting function with a preferential listening direction, bins that correspond to this direction are weighted with one and bins that do not correspond to this direction are attenuated with a value smaller than one. Wittkop et al. [19, 2] give an overview of these algorithms that capture psychoacoustical principles of the spatial masking release.

Generally, these algorithms calculate the binaural cues from the comparison of the left and the right STFT signals. As an extension to the short-time spectra, the speech enhancement algorithm of Kollmeier and Koch [5] uses modulation spectra to analyze binaural disparity. The algorithm has been designed on the grounds of physiological and psychoacoustical results, where a decomposition of stimuli into different modulation frequencies has been found to be independent of center frequencies, i.e. the tonotopic representation of the cochlear frequency decomposition, as well as the spatial decomposition. Psychoacoustically, this particular processing of the auditory system is also known as co-modulation masking release. For speech intelligibility this generally means, that if two speakers differ in their fundamental frequency which modulate their respective spectra, they are each more intelligible than if their fundamentals collapse. The Kollmeier and Koch speech enhancement algorithm has been shown to be robust in different acoustic situations and to improve the SNR over a wide dynamic range by approximately 2 dB [5, 19].

In here, a variation of this algorithm is applied to the binaural output of the MVDR beamformer. The variation refers to the weighting function. First, a frequency independent direc-

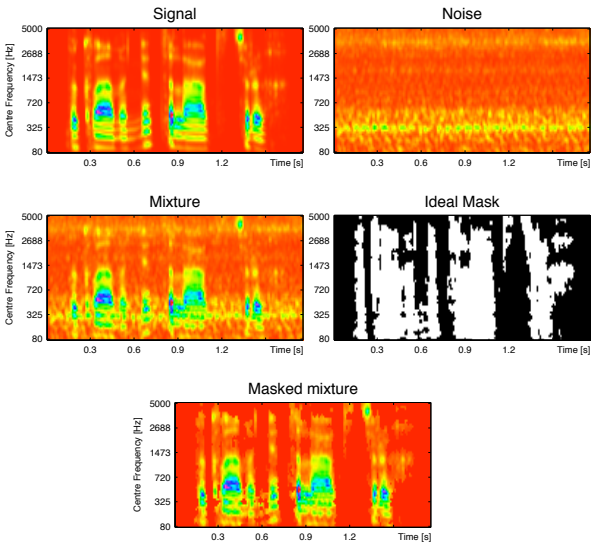

Figure 5: Illustration of the masking principle. The spectrogram of the sentence "We kwamen net te laat terug." said by male is shown in the top left panel (red denotes low energy and violet high energy). The sentence is mixed with noise and the aim is to filter the sentence by applying a mask (in the spectrotemporal domain). The noise is recorded in an autobus with an artificial dummy head. The spectrogram of the noise is shown in the top right panel. The spectrogram of the speech signal mixed with the noise is shown in the middle left panel (SNR=-7 dB). The middle right panel shows the mask. All bins with a SNR above -7 dB are white and the time-frequency bins with an SNR below -7 dB are black. The bottom panel shows the spectrogram of the mixture after multiplication with the mask.

tionality was applied, which is adopted from the lateral filter function of Wittkop and Hohmann [20]. Second, the reference values for the mask generation are data-driven. Therefore the parametric output of the localizer is used to calculate the *a posteriori* probability for the presence of the target source at a certain azimuthal position. By such means, the filter-function is constantly updated to the scene and the effects of distortion are reduced. The concept for source separation was introduced by Madhu [21, 22]. By way of example, he has shown that reverberation increases the variation of the binaural cues and that an *a posteriori* estimation corresponds to that change and opens the spatial aperture in mask-based filters.

For the most part, the here applied Kollmeier and Koch speech processor was implemented according to the algorithmic features given in [5]. We therefore restrict this report to the details of the variation from the initial implementation.

The parametric output of the localizer is first used to calculate an azimuthal non-symmetric aperture of the filter toward the target speaker. For that reason, the parametric output of the localizer is used as the input to an expectation-maximization (EM) routine in order to approximate the localization through a mixture of gaussians (MoG). This MoG fit $\hat{\theta}(n)$ is then employed to calculate the angular *a posteriori* probability of a target source $P_{t|\hat{\theta}}(n)$ in a time frame n. The approximation through a MoG is based on an *a priori* probability of P_i, the location θ_i with $\theta \in [-\pi, \pi]$ and a variance σ_i^2 of each source i with Q the total number of sources. The EM algorithm of Feder and Weinstein (see [23]) is applied to iteratively approximate the MoG estimation:

$$\hat{\theta}(n) = \sum_{i=1}^{Q} P_i exp\left(-\frac{(\theta(n)-\theta_i)^2}{2\sigma_i^2}\right). \qquad (14)$$

Thereafter, the *a posteriori* probability of the target source $P_{t|\hat{\theta}}$ is calculated with Bayes theorem:

$$P_{t|\hat{\theta}}(n) = \frac{\frac{1}{\sigma_t} P_t exp\left(-\frac{(\hat{\theta}(n)-\theta_t)^2}{2\sigma_t^2}\right)}{\sum_{i=1}^{Q} \frac{1}{\sigma_i} P_i exp\left(-\frac{(\hat{\theta}(n)-\theta_i)^2}{2\sigma_i^2}\right)}. \quad (15)$$

In this study, we assume the target source to be in front of the listener at zero degree. Figure 6 exemplifies three possible MoG fits of source combinations (dashed line) and the *a posteriori* target probability that adapts consequently to the scene. The

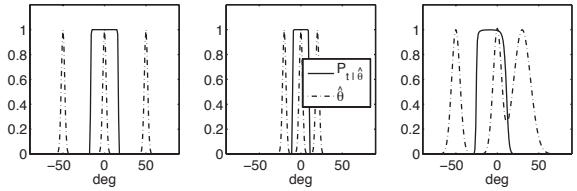

Figure 6: *A posteriori* probability of a target-source (solid line) in different arrangements (incident angles and variances) of sound sources (dashed line), where the target-source resides at zero degree.

distribution of the target probability is then used to construct a trapezoidal weighting function, where the reference values of the pass-range and the stop-range are calculated from $P_{t|\hat{\theta}} = 0.5$ and $P_{t|\hat{\theta}} = 0.05$, respectively.
Given that we yield the left and right short-time modulation spectra with $\Phi_l(n, f_c, f_m)$ and $\Phi_r(n, f_c, f_m)$, respectively, the level and phase differences can be calculated with:

$$\Delta L(n, f_c, f_m) = 10 \log_{10} \left| \frac{\Phi_{rr}(n, f_c, f_m)}{\Phi_{ll}(n, f_c, f_m)} \right| \quad (16)$$

and

$$\Delta\phi(n, f_c, f_m) = arg(\Phi_{lr}(n, f_c, f_m)), \quad (17)$$

where $\Phi_{rr}(n, f_c, f_m)$ as well as $\Phi_{ll}(n, f_c, f_m)$ are the time averaged auto-spectra of the right and left channel, respectively, and Φ_{lr} is the time-averaged cross-spectra. The time averaging is realized as in Eq. (10) with τ the time constant of a first order low pass filter. The binaural reference values ΔL_α and $\Delta\phi_\alpha$ with $\alpha \equiv [sccw; pccw; t; pcw; scw]$ were taken from a look-up table (*sccw* and *pccw* are the stop and pass angle counterclockwise, t is the target angle and *pcw* and *scw* denote the pass and stop angle clockwise). This database was built up from binaural recordings of transfer-functions in the horizontal plane at steps of 1 deg with an artificial head. ΔL_α and $\Delta\phi_\alpha$ were calculated from the convolution of the transfer-functions with a speech token of 30 s length. A voice activity detection algorithm was used to exclude pauses in the discourse of three male talkers that were subsequently summed to a mix. The trapezoidal weighting function is then calculated in analogy with the weighting function in [20]. A schematic sketch of a weighting function is given in Fig. 7.

Since the formalization is equal for the weighting function of ΔL and $\Delta\phi$, we only write down ΔL. We further reduce the notation of the weighting function to the clockwise side in Equations (18) and (19) and n, the time-index is dropped for brevity. First, the frequency-dependent reference values are calculated to establish a frequency-independent spatial filter:

$$\delta L_{pcw}(f_c, f_m) \equiv min\{|\Delta L_{pcw}(f_c, f_m) - \ldots \\ \Delta L_t(f_c, f_m)|\}, \{|\Delta L_{scw}(f_c, f_m) - \Delta L_t(f_c, f_m)|\}, \quad (18)$$

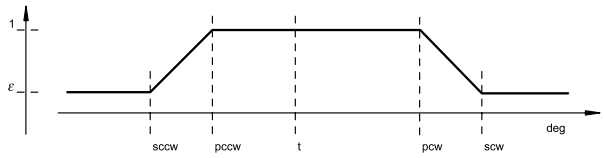

Figure 7: Schematic sketch of the mask-based spatial weighting function. *sccw* and *pccw* are the stop and pass angle counterclockwise, t is the target angle and *pcw* and *scw* denote the pass and stop angle clockwise of the binaural reference values.

and

$$\delta L_{scw}(f_c, f_m) \equiv max\{|\Delta L_{pcw}(f_c, f_m) - \ldots \\ \Delta L_t(f_c, f_m)|\}, \{|\Delta L_{scw}(f_c, f_m) - \Delta L_t(f_c, f_m)|\}. \quad (19)$$

$\Delta L(f_c, f_m)$ is corrected for asymmetries in the same way:

$$\delta L(f_c, f_m) \equiv \Delta L(f_c, f_m) - \Delta L_t(f_c, f_m). \quad (20)$$

The trapezoidal weighting function takes then the following form (f_c and f_m are dropped for brevity):

$$M_{t,\Delta L} = \begin{cases} \varepsilon & ; \text{ if } \delta L < \delta L_{sccw} \\ \Lambda_{ccw} & ; \text{ if } \delta L_{sccw} < \delta L < \delta L_{pccw} \\ 1 & ; \text{ if } \delta L_{pccw} < \delta L < \delta L_{pcw} \\ \Lambda_{cw} & ; \text{ if } \delta L_{pcw} < \delta L < \delta L_{scw} \\ \varepsilon & ; \text{ if } \delta L_{sccw} < \delta L \end{cases}, \quad (21)$$

herein Λ_{cw} is:

$$\Lambda_{cw} = \frac{(-1+\varepsilon)}{|\delta L_{scw} - \delta L_{pcw}|}(\delta L - \delta L_{pcw}) + 1, \quad (22)$$

and ε is the maximum attenuation.
$M_{t,\Delta L}(n, f_c, f_m)$ and $M_{t,\Delta\phi}(n, f_c, f_m)$ are then combined to a total weighting function for suppressing the interference to a target:

$$M_t(n, f_c, f_m) = (bM_{t,\Delta L}(n, f_c, f_m) + \ldots \\ (1-b)M_{t,\Delta\phi}(n, f_c, f_m))^e. \quad (23)$$

where e denotes an expansion exponent. The filtered modulation spectrum is inversely transformed to the STFT representation and corrected for the lost phase information as described in [5]. Finally, the left and right output of the MVDR beamformer are each ($Y_{t,MVDR}(n, f_c)$) weighted with the desired envelope of the target speaker in the STFT domain:

$$Y_t(n, f_c) = M_t(n, f_c) Y_{t,MVDR}(n, f_c). \quad (24)$$

In order to keep the processing as clear as possible, no smoothing of the short-time spectra or modulation spectra is applied.

EVALUATION

The combined processing scheme was tested in three setups. For the assessment of the speech intelligibility and quality, the appropriately weighted three level version of the Speech Intelligibility Index (SII) of Kates and Arehart [6] was applied. To include the dominating effect of the binaural advantage, the head shadow effect, a 'better ear' modeling was included in auditory critical bands for the three level index I3. A 'mean ear' modeling in auditory critical bands was calculated for the three level quality measure Q3. Therefore, the signal-to-distortion ratio (cf. Eq. (13) in [1]) is calculated in auditory critical bands and the maximum is taken in bands for the 'better ear' effect of the I3 measure, and the arithmetic mean in bands is calculated for the 'mean ear' for the Q3 measure.
As this contribution is mainly about the improvement of the

beamformer output, the analysis refers only to this improvement and not to the total improvement. Moreover, the analysis of the total improvement would be of limited validity, since the beamformer is optimized for an ideal diffuse noise field. However, to test the co-modulation masking release processing of the Kollmeier and Koch algorithm, pure speech samples for the target and the interferer were applied in the assessment. Preparatory, pauses in speech were excluded by a voice activity algorithm and sentences were faded into each other with an overlap add technique. Given a sufficient time-duration of the presentation, the short-time SII version, in here it is calculated in windows of about 30 ms length with an overlap of 50 %, is to a fair extend capable to predict speech intelligibility and quality for pure speech-maskers.
Three acoustical scenes at three different SNR levels were analyzed. For the speech material of different talkers, sentences of the Plomp corpus were used [17], prepared as above mentioned and combined to samples of 30 s length. The SNR levels were calculated at the ear level after the beamforming stage and comprised -5, 0 and 5 dB. Figure 8 gives the setups and results, where S stands for target-speaker and N stands for noise-speaker. The circles denote the initial unprocessed situation at the three SNRs. The Kollmeier Koch algorithm was adjusted

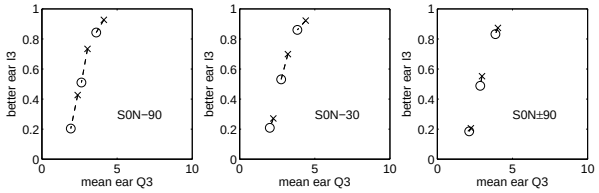

Figure 8: Results of the speech intelligibility improvement with the post-processor and the data-driven filter generation. ○ denote original situations and × denotes the processed situations.

by hand to achieve an optimal processing. The parameters of the algorithm are given in Table 1. As can be seen, the post-

Table 1: Parameters as applied in the simplified implementation of the Kollmeier Koch speech enhancement algorithm

e	b	ε	τ
2	0.95	0.2	0.2

processing improves both the speech intelligibility and quality throughout the range of SNRs and conditions. The improvement in speech intelligibility is the highest in the S0N-90 situation and almost no improvement is obtained in the two interferer situation S0N±90, where the sparsity and disjointness requirements of different sources in the time-frequency representation are violated. A smaller but clear decline in speech intelligibility improvement is also observed from the S0N-90 to the S0N-30 situation, when the binaural cues get less separable to build up a weighting function. In spite of that, an increase of 20 % speech intelligibility corresponds to an increase in SNR of approximately 1 to 2 dB and this marks an important advantage in the respective situations. Although the algorithm was not adjusted to yield a maximal quality improvement, no deterioration is observed in all conditions. This is an important outcome as listening ease and speech quality prevent fatigue and together form a requirement for a feasible speech enhancement algorithm.
A further improvement in signal quality might be achieved by smoothing the short-time spectrograms with a leaky integration in the cepstral domain [24], which is a technique that is generally used on binary masks. Together with the enhancement of the robustness of the Kollmeier Koch algorithm through a penalty measure of binaural variance (cf. weighting function in [5]), these technical advances are left to further research. The primary aim here was the combination of fundamental building blocks in a clear and straightforward design.

CONCLUSION

In order to achieve a high improvement in speech intelligibility while maintaining or improving speech quality, a combined speech enhancement scheme has been presented. It consists of a low-directivity MVDR beamformer, an auditory model-based localizer and an auditory model-based speech enhancement algorithm. The MVDR beamformer realizes a gain of 4.4 dB in a diffuse noise field at low self noise and features externalized binaural cues. These cues are exploited by the localizer and the post-processors. The localizer, which applies physiological principles of across frequency interaction, performs robust in a wide range of acoustical situations. Its parametric output is used to estimate a data-driven filter in the post-processing algorithm. This filter, realized as a time-frequency mask, acts on the total acoustic scene to enhance the intelligibility of the information coming from a target-speaker and maintains simultaneously a natural background and localization cues. Therewith the post-processor can potentially yield an audiological benefit. The here applied instrumental measures of speech intelligibility and quality reported an improvement of about 20 % at low to mid SNR levels in single-interferer situations. Also when the situations become more complex, the post-processing algorithm does not deteriorate speech intelligibility and quality.
The combination of a pre-processor and a post-processor as applied here is not new, e.g., [25]. The advantage of the here presented scheme is its robustness and its balance between distortion and speech intelligibility improvement. In spite of that, the entire processing chain holds a degree of complexity that is not easily manageable. We tried to simplify the processing where possible in this contribution to get a better insight into the crucial parts of the processing. However for a successful design, objective measures and standard tests of binaural speech intelligibility and quality for the nonlinear processors should be established to perform the difficult algorithmic parameter adjustment. Once these measures are available, a thorough refinement of speech enhancement approaches that are similar to the here presented work can follow.

ACKNOWLEDGMENTS

The authors like to thank Hendrik Elzinga for his help.

REFERENCES

[1] J. M. Kates and K. H. Arehart, “Coherence and the speech intelligibility index,” *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[2] D. Wang and G. Brown, *Computational Auditory Scene Analysis*. A John Wiley & Sohns, Inc., Publication, 2006.

[3] C. Liu, B. Wheeler, W. O’brien Jr., R. Bilger, C. Lansing, and A. Feng, “Localization of multiple sound sources with two microphones,” *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[4] S. Albani, J. Peissig, and B. Kollmeier, “Model of binaural localization resolving multiple sources and spatial ambiguities,” in *Psychoacoustics, Speech and Hearing Aids* (B. Kollmeier, ed.), pp. 227–232, World Scientific Publishing Co. Pte. Ltd., 1996.

[5] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *J. Acoust. Soc. Am.*, vol. 95, pp. 1593–1602, March 1994.

[6] J. Kates and K. Arehart, "A model of speech itelligibility and quality in hearing aids," in *IEEE Workshop on applications of signal processing to audio and acoustics*, (New Paltz), pp. 53–56, IEEE, 2005.

[7] M. M. Boone, "Directivity measurements on a highly directive hearing aid: the hearing glasses," *AES 120th Convention, Paris*, 2006.

[8] I. Merks, *Binaural Application of Microphone Arrays for Improvement Speech Intelligibility in a noisy Environment*. PhD thesis, University of Technology Delft, 2000.

[9] J. Capon, R. Greenfield, and R. Kolker, "Multidimensional maximum likelihood processing of a large aperture seismic array," *Proceedings of the IEEE*, vol. 55, no. 2, pp. 192–213, 1967.

[10] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[11] N. Madhu and R. Martin, *Advances in Digital Speech Transmission*, ch. Source localization with microphone arrays, pp. 156–170. John Wiley & Sons, Ltd, 2008.

[12] M. S. Brainard, E. I. Knudsen, and S. D. Esterly, "Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues," *J. Acoust. Soc. Am.*, vol. 91, no. 2, pp. 1015–1027, 1992.

[13] E. I. Knudsen and M. Konishi, "Mechanisms of sound localization in the barn owl (tyto alba)," *J. Comp. Physiol. A*, vol. 133, pp. 13–21, 1979.

[14] F. Hlawatsch and G. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Process. Mag.*, vol. 9, pp. 21–67, 1992.

[15] T. Wittkop, *Two-channel noise reduction algorithms motivated by models of binaural interaction*. PhD thesis, Carl-von-Ossietzky-Universität, Oldenburg, 2001.

[16] B. Moore, *An introduction to the psychology of hearing*, vol. 1, ch. 3, pp. 100–101. New York, USA: Academic press, 3rd ed., 1989.

[17] R. Plomp and A. Mimpen, "Improving the reliability of testing the speech reception threshold for sentences," *International Journal of Audiology*, vol. 18, no. 1, pp. 43–52, 1979.

[18] W. Bothe and W. Kolhörster, "Das Wesen der Höhenstrahlung," *Zeitschrift für Physik A*, vol. 56, no. 11, pp. 751–777, 1929.

[19] T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. Woods, and B. Kollmeier, "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *ACTA ACUSTICA united with ACUSTICA*, pp. 684–699, July 1997.

[20] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, vol. 39, pp. 111–138, 2003.

[21] N. Madhu, "Data-driven mask generation for source separation," (Denmark), ISAAR, 2009.

[22] N. Madhu and J. Wouters, "Localisation-based, situation-adaptive mask generation for source separation," (Cyprus), pp. 3–5, International Symposium on Communications, Control and Signal Processing (ISCCSP), 2010.

[23] M. Lázaro, I. Santamaría, and C. Pantaleón, "A new em-based training algorithm for RBF networks," *Neural Netw.*, vol. 16, no. 1, pp. 69–77, 2003.

[24] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *ICASSP*, 2008.

[25] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. J. O'Brien, B. Wheeler, and A. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, 2004.