On the Impact of Localization Errors on HRTF-based Robust Least-Squares Beamforming

Hendrik Barfuss and Walter Kellermann*
Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg
{barfuss,wk}@lnt.de

Introduction

In a typical human-machine dialogue scenario, the target source and additional interfering sources, located at different positions from the target source, may be active at the same time. Clearly, these interfering sources have to be suppressed in order to establish a successful human-machine interaction. A common strategy is to apply spatial filtering techniques which are usually based on the free-field assumption of acoustic wave propagation. However, for scenarios where the microphones are mounted on a scatterer, the free-field assumption is not optimum, since the influence of the scatterer on the sound field is neglected. One example of such a scenario is a microphone array mounted on a robot head used for robot audition, which is also the focus of this article.

In order to design a beamformer which accounts for the influence of the scatterer, i.e., the robot head, on the sound field, the free-field steering vectors have to be replaced by Head-Related Transfer Functions (HRTFs)[1], see, e.g., [1].

In [2], we proposed an HRTF-based Robust Least-Squares Frequency-Invariant (RLSFI) beamformer design and verified experimentally that employing HRTFs instead of free-field steering vectors leads to a significantly improved beamforming performance and correspondingly better Automatic Speech Recognition (ASR), in a robot audition scenario. Since the proposed beamformer design depends on a set of HRTFs, the question arises how the beamformer performs if these HRTFs do not correspond to the true position of the target source, e.g., due to localization errors. Therefore, in this contribution, we investigate the impact of localization errors on the performance of the HRTF-based RLSFI beamformer.

The remainder of this article is organized as follows: In the next section, the HRTF-based beamformer design from [2] is briefly reviewed. After this, the results of our investigation of the HRTF robustness are presented, followed by a conclusion and an outlook to future work in the last section.

HRTF-based robust beamforming

Fig. 1 illustrates the block diagram of a Filter-and-Sum Beamformer (FSB), consisting of N microphones at positions $\mathbf{p}_n$, where $\mathbf{p}_n$ represents the position of the n-th microphone in Cartesian coordinates. In this article, vectors and matrices are denoted by lower- and upper-case boldface letters, respectively. The output signal $y[k]$ at time instant k is obtained by convolving the microphone signals $x_n[k]$ with Finite Impulse Response (FIR) filters $\mathbf{w}_n = [w_{n,0}, \ldots, w_{n,L-1}]^T$ of length L and a subsequent summation over all N channels. The beamformer response of an FSB is given as [3, 4]:

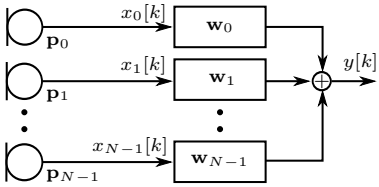


Figure 1: Block diagram of a filter-and-sum beamformer consisting of N microphones and FIR filters [2, 4].

$$B(\omega, \phi, \theta) = \sum_{n=0}^{N-1} W_n(\omega) g_n(\omega, \phi, \theta), \tag{1}$$

where $W_n(\omega) = \sum_{l=0}^{L-1} w_{n,l} e^{-j\omega l}$ is the Discrete-Time Fourier Transform (DTFT) representation of $\mathbf{w}_n$. Moreover, $g_n(\omega, \phi, \theta)$ is the response of the n-th microphone to a plane wave with frequency ω traveling in the direction (ϕ, θ), where ϕ and θ denote azimuth and elevation angle, respectively, and are defined as in [3].

In [4], the design of an RLSFI FSB was proposed, where a desired beamformer response $\hat{B}(\omega, \phi, \theta)$ is approximated in the Least-Squares (LS) sense at each frequency ω subject to a distortionless response constraint in the desired look direction and a constraint on the White Noise Gain (WNG). The LS approximation is performed for a discrete set of P frequencies ω_p and M look directions (ϕ_m, θ_m), and can be formulated in matrix notation as[2] [4]

$$\underset{\mathbf{w}_\mathrm{f}(\omega_p)}{\operatorname{argmin}} \|\mathbf{G}(\omega_p)\mathbf{w}_\mathrm{f}(\omega_p) - \hat{\mathbf{b}}\|_2^2 \tag{2}$$

subject to constraints on the WNG and the response in desired look direction, respectively:

$$\frac{|\mathbf{w}_\mathrm{f}^T(\omega_p)\mathbf{d}(\omega_p)|^2}{\mathbf{w}_\mathrm{f}^H(\omega_p)\mathbf{w}_\mathrm{f}(\omega_p)} \geq \gamma > 0, \quad \mathbf{w}_\mathrm{f}^T(\omega_p)\mathbf{d}(\omega_p) = 1, \tag{3}$$

*The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

[1] Note that in the context of this work, HRTFs only model the direct propagation path between a source and a microphone mounted on a robot head, but no reverberation components.

[2] A MATLAB design tool with a graphical user interface for the free-field-based design can be downloaded from http://goo.gl/obnZwY.

where $\mathbf{w}_\mathrm{f}(\omega_p) = [W_0(\omega_p), \ldots, W_{N-1}(\omega_p)]^T$, $[\mathbf{G}(\omega_p)]_{mn} = g_n(\omega_p, \phi_m, \theta_m)$, vector $\hat{\mathbf{b}} = [\hat{B}(\phi_0, \theta_0), \ldots, \hat{B}(\phi_{M-1}, \theta_{M-1})]^T$ contains the desired responses for all M discrete look directions, and $\mathbf{d}(\omega_p) = [g_0(\omega_p, \phi_\mathrm{d}, \theta_\mathrm{d}), \ldots, g_{N-1}(\omega_p, \phi_\mathrm{d}, \theta_\mathrm{d})]^T$ is the steering vector corresponding to the desired look direction $(\phi_\mathrm{d}, \theta_\mathrm{d})$. Operators $\|\cdot\|_2$, $(\cdot)^T$, and $(\cdot)^H$ denote the Euclidean norm, and the transpose and conjugate transpose of vectors or matrices, respectively. Note that the same desired response is chosen for all frequencies, as can be seen from the frequency-independent entries of $\hat{\mathbf{b}}$. Equations (2) and (3) can be interpreted as follows: The LS approximation of the desired beamformer response is given by (2). The first part of (3) represents the WNG constraint, with the lower bound γ on the WNG, which has to be defined by the user. The second part of (3) describes the distortionless response constraint which ensures that the target signal, coming from the desired look direction, passes the beamformer undistorted. The time-domain FIR filters $\mathbf{w}_n$ are obtained by solving (2), (3) for each frequency ω_p separately, followed by an FIR approximation of the optimum filter coefficients.

Assuming the microphones are located in the free field, the sensor response is given as

$$g_{n,\mathrm{FF}}(\omega_p, \phi_m, \theta_m) = e^{-j\mathbf{k}^T(\omega_p, \phi_m, \theta_m)\mathbf{p}_n}, \qquad (4)$$

where $\mathbf{k}(\omega_p, \phi_m, \theta_m)$ denotes the wave vector which depends on the current frequency and look direction, and the speed of sound [3]. Thus, matrix $\mathbf{G}(\omega_p)$ in (2) contains the well-known free-field-based steering vectors with respect to the M look directions and the N microphones, and vector $\mathbf{d}(\omega_p)$ in (3) is the free-field-based steering vector corresponding to the desired look direction.

The HRTF-based RLSFI beamformer design, as proposed in [2], is obtained by including measured or simulated HRTFs in (2) and (3) instead of free-field-based steering vectors. In this case, the sensor response is given as

$$g_{n,\mathrm{HRTF}}(\omega_p, \phi_m, \theta_m) = h_{mn}(\omega_p), \qquad (5)$$

where $h_{mn}(\omega_p)$ is the HRTF modeling the propagation between the m-th source position and n-th sensor at frequency ω_p. Consequently, $\mathbf{G}(\omega_p)$ now consists of all HRTFs between the M look directions and the N microphones, and $\mathbf{d}(\omega_p)$ contains the HRTFs corresponding to the desired look direction. Note that in contrast to the free-field-based design (4), the HRTFs-based design implicitly depends on the robot-source distance for which the HRTFs have been measured (see, e.g., [5]).

In Fig. 2, an example of the HRTF-based RLSFI beamformer according to (2), (3), and (5) is illustrated for a frequency range of 300 Hz $\leq f \leq$ 5000 Hz. The design was carried out for the 5-microphone robot head array illustrated in Fig 3(b). Beampatterns for two different WNG constraint values $\gamma_\mathrm{dB} = 10\log_{10}(\gamma) \in \{-10, -20\}$dB are shown to demonstrate the impact of the WNG constraint on the beamformer. It is important to note that the beampatterns were computed by evaluating (1) with (5). Thus, they effectively show the transfer

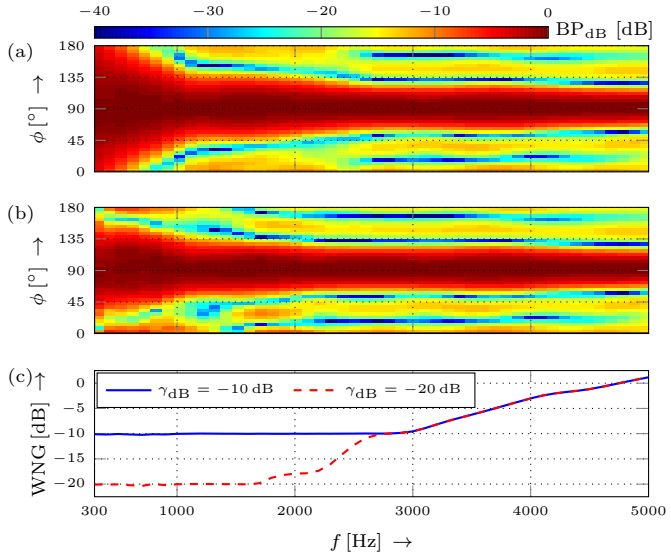


Figure 2: Design example of an HRTF-based RLSFI beamformer, designed for the 5-microphone robot head array in Fig. 3(b). Beampatterns for WNG constraints are illustrated in (a) $\gamma_\mathrm{dB} = -10$ dB and (b) $\gamma_\mathrm{dB} = -20$ dB. Subfigure (c) shows the resulting WNG.

function between source position and beamformer output with HRTFs modeling the acoustic system. We used a filter length of $L = 1024$ for the FIR approximation, and the HRTFs which were incorporated in the beamformer design were measured for a robot-source distance of 1.1m. The main beam was steered towards broadside. Figs. 2(a) and 2(b) illustrate the resulting beampatterns $\mathrm{BP}_\mathrm{dB}(\ldots) = 10\log_{10}(|B(\ldots)|^2)$ in dB and Fig. 2(c) shows the corresponding WNG in dB over frequency. It can be seen that both beamformers exhibit a good spatial selectivity above 1000Hz, and that a higher WNG constraint γ_dB leads to a broader beam at lower frequencies. Thus, the user can control the trade-off between robustness and spatial selectivity directly. Fig 2(c) confirms that both designs fulfill the required WNG with occasional slight deviations, which are due to the FIR approximation of the optimum filter coefficients. Note that a comparison of the beampatterns of the HRTF- and free-field-based beamformer with HRTFs modeling the acoustic system can be found in [2], illustrating the effect of the robot head as scatterer on the sound field.

Experimental results

In the following, we analyze the relative robustness of the HRTF-based beamformer design by comparing the impact of localization errors on the performance of the HRTF- and free-field-based RLSFI beamformer. More specifically, we investigate the impact of localization errors with respect to Direction-of-Arrival (DOA) and robot-source distance. At first, the experimental setup and performance measures are introduced, followed by a presentation of the experimental results.

Setup and parameters

We use Word Error Rates (WERs) of an automatic speech recognizer to evaluate the overall quality of the

enhanced signals at the beamformer outputs, since a high speech recognition accuracy is the main goal in robot audition. As ASR engine, we employed PocketSphinx [6] with a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM)-based acoustic model trained on clean speech from the GRID corpus [7], using MFCC+Δ+ΔΔ features and cepstral mean normalization. For the computation of the WER scores, only the letter and the number in the utterance were evaluated, as in the CHiME challenge [8]. Our test signal contained 200 utterances. In addition, the frequency-weighted segmental Signal-to-Noise Ratio (fwSegSNR) as defined in [9] was evaluated, where the target signal at the center microphone and at the beamformer output was used as reference signal for calculating the fwSegSNR at the input and output of the beamformer, respectively.

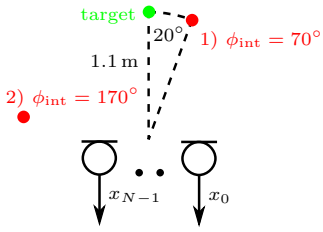
We created the microphone signals by convolving clean-speech source signals with Room Impulse Responses (RIRs), measured in a lab room with a reverberation time of $T_{60} = 190\,\text{ms}$ and a critical distance of approximately 1.2 m, using maximum-length sequences. The sampling rate of the speech signals and measured RIRs and HRTFs was 16kHz. The microphone positions at the robot head for which the RIRs were measured are illustrated in Fig. 3(b). The relative height of the source with respect to the robot head was 0.73 m, corresponding to an elevation angle $\theta = 56.4°$. This setup was chosen to simulate a taller human interacting with the NAO robot of height 0.57 m. The measurements were carried out for the robot looking towards broadside.

The set of HRTFs which is required for the HRTF-based beamformer design was measured for the same microphone configuration and robot-source distance as for the RIR measurements described above.

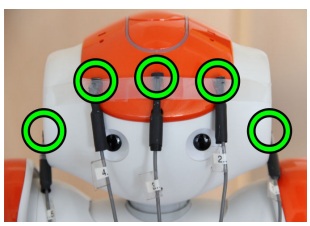
The HRTF- and free-field-based beamformers were designed for a filter length of $L = 1024$ taps and a WNG constraint with a lower bound of $\gamma_{\text{dB}} - 10$ dB.

Impact of localization errors with respect to direction of arrival

At first, the impact of DOA estimation errors on the beamforming performance is investigated. To this end, two two-speaker scenarios were evaluated, where the target source was always located at $\phi_{\text{d}} = 90°$ and the interfering source was located at 1) $\phi_{\text{int}} = 70°$ or 2) $\phi_{\text{int}} = 170°$, at a robot-source distance of $d = 1.1\,\text{m}$. The beamformer was steered towards $\phi_{\text{BF}} \in \{100°, 95°, 90°, 85°, 80°\}$, simulating localization errors of $\pm 5°$ and $\pm 10°$. The scenario was chosen to analyze the impact of localization errors on the beamformer performance in situations where an interfering source is 1) very close to or 2) relatively far away from the target source which is located directly in front of the robot. The evaluated two-speaker scenarios 1) and 2) are illustrated in Fig. 3(a), where target source and interfering source positions are illustrated by green and red filled circles, respectively.

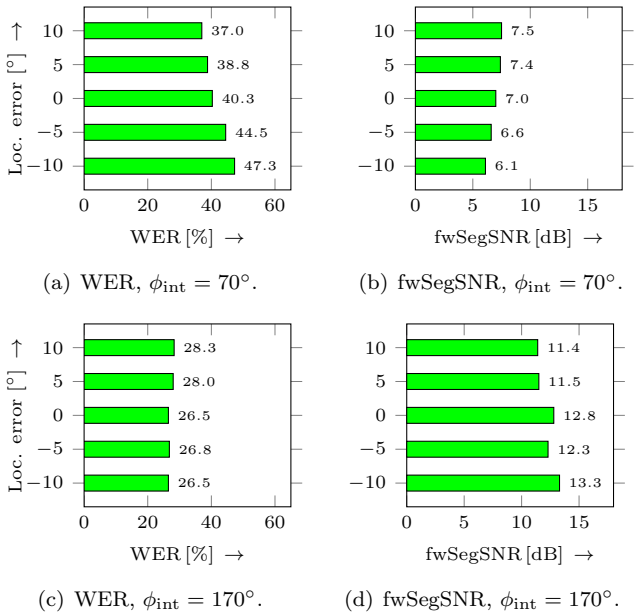


(a) Source positions.



(b) Microphone positions.

Figure 3: Illustration of the source positions of the two-speaker scenario and the employed microphone positions at the robot head.



(a) WER, $\phi_{\text{int}} = 70°$.

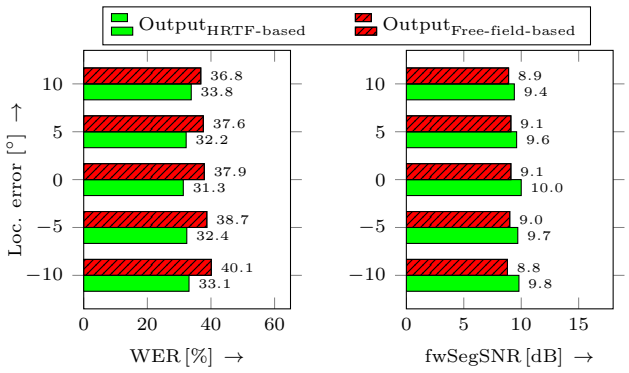
(b) fwSegSNR, $\phi_{\text{int}} = 70°$.

(c) WER, $\phi_{\text{int}} = 170°$.

(d) fwSegSNR, $\phi_{\text{int}} = 170°$.

Figure 4: Illustration of WERs in % and fwSegSNR levels in dB, obtained at the output of the HRTF-based beamformer for Scenario 1) $\phi_{\text{d}} = 90°$, $\phi_{\text{int}} = 70°$ and 2) $\phi_{\text{d}} = 90°$, $\phi_{\text{int}} = 170°$ for DOA estimation errors of $\pm 5°$ and $\pm 10°$. Measures at input: Scenario 1) $\text{WER}_{\text{in}} = 49.0\%$ and $\text{fwSegSNR}_{\text{in}} = 5.2\text{dB}$ and Scenario 2) $\text{WER}_{\text{in}} = 44.3\%$ and $\text{fwSegSNR}_{\text{in}} = 5.8\text{dB}$.

In Fig. 4, the results for the two scenarios are summarized. The subfigures on the left- and right-hand side show the WERs in % and fwSegSNR levels in dB obtained at the HRTF-based beamformer output, respectively. Each horizontal bar represents the results for one specific localization error of $\pm 5°$ or $\pm 10°$. From Figs. 4(a) and 4(b) it can be seen that when the interferer is very close to the target source, localization errors have a strong impact on the beamforming performance. When the beamformer is accidentally steered closer towards the interfering source (localization errors of $-5°$ and $-10°$), the beamforming performance decreases. This is because the beamformer's main beam is steered towards the interfering source, leading to a lower attenuation of the latter. If the localization error leads to the beamformer being steered away from the interfering source (localization errors of 5° and 10°), an increasing beamforming performance can be observed. This can be explained by the fact that in this particular scenario, a spatial null of the beampattern is getting closer to the interferer's direction the larger the localization error is. If the interferer is far away from the target source, as in Scenario 2), localization errors do not have a strong impact on the beam-



(a) Average WER. (b) Average fwSegSNR.

Figure 5: Illustration of average WERs in % and average fwSegSNR levels in dB, obtained at the output of the HRTF- and free-field-based beamformer for $\phi_{\text{d}} = 90°$, $\phi_{\text{int}} \in \{10° : (20°) : 70°, 110° : (20°) : 170°\}$, and localization errors of $\pm 5°$ and $\pm 10°$. Average input measures: $\text{WER}_{\text{in}} = 47.1\%$ and $\text{fwSegSNR}_{\text{in}} = 5.5\text{dB}$.

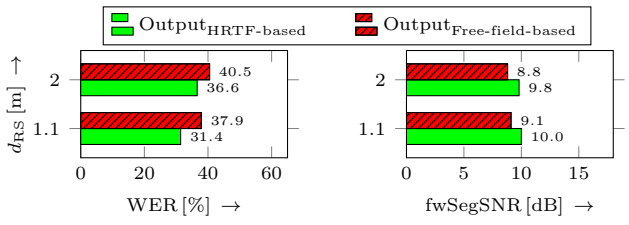
forming performance, which can be seen in Figs. 4(c) and 4(d), respectively.

In Figs. 5(a) and 5(b), the average WERs in % and fwSegSNR levels in dB, obtained at the output of the HRTF-based and free-field-based RLSFI beamformer, respectively, are illustrated. The presented results were averaged over eight different scenarios, where the target source was always located at $\phi_{\text{d}} = 90°$ and the interferer was located at positions between 10° and 70°, and 110° and 170°, in steps of 20°. It can be seen that the average performance of the HRTF-based beamformer decreases when there is a localization error. Furthermore, one can observe that the HRTF-based beamformer in general yields a better performance than the free-field-based beamformer, as was already shown in [2].

Impact of localization errors with respect to robot-source distance

In a second experiment, we evaluated the impact of localization errors with respect to the robot-source distance d_{RS}. Since in our experiment the robot head and the source are not at the same height, distance errors result in a mismatch between the elevation angle the beamformer is steered to and the elevation angle of the target source with respect to the robot head array. Here, we evaluated the beamformer performance for robot-source distances $d_{\text{RS}} \in \{1.1\text{m}, 2\text{m}\}$. The HRTF-based beamformer was designed using HRTFs measured for a robot-source distance of 1.1m. Thus, the elevation mismatch for $d_{\text{RS}} = 1.1\text{m}$, is 0°, whereas for $d_{\text{RS}} = 2\text{m}$, there is a mismatch of 13.5°, i.e., the beamformer is steered too high in elevation. To allow for a fair comparison, the same elevation angle was used for the free-field-based beamformer.

In Figs. 6(a) and 6(b), the average output WERs in % and fwSegSNR levels in dB of the HRTF- and free-field-based RLSFI beamformers are illustrated. The results were obtained for the same desired and interfering source positions as for Fig. 5. The results show that a mismatch with respect to the robot-source distance leads to a significant decrease in WER and to a slight decrease of the fwSegSNR. Apart from that, it is interesting to see that the HRTF-based beamformer still yields better results than the free-field-based beamformer.



(a) Average WER. (b) Average fwSegSNR.

Figure 6: Illustration of average WERs in % and fwSegSNR levels in dB, obtained at the output of the HRTF- and free-field-based beamformer for $\phi_{\text{d}} = 90°$, $\phi_{\text{int}} \in \{10° : (20°) : 70°, 110° : (20°) : 170°\}$, and robot-source distances of 1.1m and 2m. Average input measures: $\text{WER}_{\text{in}} = 47.3\%$ and $\text{fwSegSNR}_{\text{in}} = 5.5\text{dB}$.

Conclusion

In this work, we investigated the impact of localization errors on the performance of a recently proposed HRTF-based RLSFI beamformer. Localization errors with respect to the DOA of the target signal as well as the robot-source distance were evaluated. The results confirmed that both, erroneous DOA and robot-source distance estimates lead to a significant decrease in beamforming performance. Thus, it is of vital importance to use a set of HRTFs for the design of the HRTF-based RLSFI beamformer, which matches the position of the target source. Future work includes analysis of the effect of localization errors on the behaviour of the HRTF-based beamformer for sources in the near-field, and an extension of the RLSFI beamformer design to two-dimensional beam steering.

References

[1] Maazaoui, M., Abed-Meraim, K., and Grenier, Y.: Blind source separation for robot audition using fixed HRTF beamforming, EURASIP J. Advances Signal Processing, 2012 (58), Mar. 2012.

[2] Barfuss, H., Huemmer, C., Lamani, G., Schwarz, A., and Kellermann, W.: HRTF-based robust least-squares frequency-invariant beamforming, IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA), Oct. 2015.

[3] Van Trees, H.L.: Detection, estimation, and modulation theory: Optimum array processing, Wiley, 2004.

[4] Mabande, E., Schad, A., and Kellermann, W.: Design of robust superdirective beamformers as a convex optimization problem, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Apr. 2009.

[5] Blauert, J.: Spatial hearing: The psychophysics of human sound localization, The MIT Press, 1997.

[6] Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., and Rudnicky, A.I.: PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), May 2006.

[7] Cooke, M., Barker, J., Cunningham, S., and Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition, J. Acoustical Society America, 120 (5), Nov. 2006.

[8] Christensen, H., Barker, J., Ma, N., and Green, P.D.: The CHiME corpus: A resource and a challenge for computational hearing in multisource environments, INTERSPEECH, Sept. 2010.

[9] Hu, Y. and Loizou, P.C.: Evaluation of Objective Quality Measures for Speech Enhancement, Audio, Speech, Language Proc., IEEE Trans., 16 (1), Jan. 2008.