

IMPROVED NOISE ESTIMATION FOR THE BINAURAL MWF WITH INSTANTANEOUS ITF PRESERVATION

Daniel Marquardt¹, Lin Wang¹, Volker Hohmann², Simon Doclo¹

¹University of Oldenburg, Institute of Physics, Signal Processing Group, Oldenburg, Germany

²University of Oldenburg, Institute of Physics, Medical Physics, Oldenburg, Germany

{daniel.marquardt, simon.doclo}@uni-oldenburg.de

ABSTRACT

An important objective of binaural noise reduction algorithms in hearing aids is the preservation of the binaural cues. Recently a Multi-channel Wiener filter with instantaneous binaural cue preservation (MWF-ITFhc) has been presented, which relies on an accurate estimate of the noise signal vector. In this paper we propose a GSC-like structure for the MWF-ITFhc, comprising a blocking matrix and back projection. The perceptual difference between the original and the filtered signals is evaluated using an objective measure, based on a model of the binaural auditory processing. Experimental results show that the application of a blocking matrix with back projection increases the performance of the MWF-ITFhc in preserving the binaural cues of both the speech and the noise component.

Index Terms— Hearing aids, binaural cues, noise reduction

1. INTRODUCTION

Noise reduction algorithms in hearing aids are crucial to improve speech understanding in background noise for hearing impaired persons. For binaural hearing aids, algorithms that exploit multiple microphone signals from both the left and the right hearing aid are considered to be promising techniques for noise reduction, because in addition to spectral information spatial sound information can be exploited. In addition to reducing noise and limiting speech distortion, another important objective of binaural noise reduction algorithms is the preservation of the listeners impression of the auditory scene, in order to exploit the binaural hearing advantage and to avoid confusions due to a mismatch between the acoustical and the visual information. This can be achieved by preserving the Interaural Transfer Function (ITF) of the speech and the noise component, comprising both the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) binaural cues.

In [1] a binaural Multi-channel Wiener Filter (MWF) has been presented. It has been theoretically proven in [2] that this technique preserves the ITF of the speech component for a single speech source. On the contrary, the ITF of the noise

component is distorted, such that the ITF of the residual noise component is equal to the ITF of the speech component.

In addition, an extension of the MWF, namely the MWF-ITF, has been presented, by imposing a soft constraint on the preservation of the ITF of the noise component. Theoretical and experimental results in [2] have shown that a better preservation of the ITF of the noise component leads to a distortion of the ITF of the speech component, depending on the input SNR and a trade-off parameter. Hence, using the MWF-ITF it is not possible to preserve the speech and the noise ITF simultaneously.

To overcome this trade-off between preserving the binaural speech and noise cues, an instantaneous hard-constraint formulation of the noise ITF preservation term has been presented in [3], resulting in the MWF-ITFhc. This formulation allows perfect preservation of the noise ITF, given a perfect estimate of the noise signal vector. In this paper we propose a GSC-like structure for the MWF-ITFhc, comprising a blocking matrix and back projection to estimate the noise component in the microphone signals. To perceptually evaluate the binaural cue preservation, we also introduce an objective measure, based on the binaural localization model presented in [4] to ensure an objective evaluation that is related to the binaural perception of the human auditory system.

Experimental results show that incorporating a blocking matrix with back projection results in a better noise estimate for the MWF-ITFhc and less distortion of the binaural cues compared to other binaural cue preservation methods, quantified by the binaural auditory model based objective measure.

2. CONFIGURATION AND NOTATION

Consider the binaural hearing aid configuration in Figure 1, consisting of the left and the right microphone array with M microphones each. The frequency-domain representation of the m -th microphone signal in the left hearing aid $Y_{0,m}(k, l)$ can be written as

$$Y_{0,m}(k, l) = X_{0,m}(k, l) + V_{0,m}(k, l), \quad m = 0 \dots M - 1,$$

with $X_{0,m}(k, l)$ and $V_{0,m}(k, l)$ representing the speech and the noise component, k denoting the frequency index and l the frame index. The m -th microphone signal in the right hearing

This work was partly funded by the BMBF project "Modellbasierte Hörsysteme"

aid $Y_{1,m}(k, l)$ is defined similarly. For conciseness we will omit the variable k and l in the remainder of the paper, except where explicitly required.

We define the $2M$ -dimensional signal vector \mathbf{Y} as

$$\mathbf{Y} = [Y_{0,0} \dots Y_{0,M-1} Y_{1,0} \dots Y_{1,M-1}]^T. \quad (1)$$

The signal vector can be written as $\mathbf{Y} = \mathbf{X} + \mathbf{V}$, where \mathbf{X} and \mathbf{V} are defined similarly as \mathbf{Y} . Furthermore, we define the $4M$ -dimensional stacked weight vector \mathbf{W} as

$$\mathbf{W} = [\mathbf{W}_0 \quad \mathbf{W}_1]^T. \quad (2)$$

The output signal at the left hearing aid Z_0 is equal to

$$Z_0 = \mathbf{W}_0^H \mathbf{Y} = \mathbf{W}_0^H \mathbf{X} + \mathbf{W}_0^H \mathbf{V} = Z_{x,0} + Z_{v,0}, \quad (3)$$

where $Z_{x,0}$ represents the speech component and $Z_{v,0}$ represents the noise component in the output signal. Similarly, the output signal at the right hearing aid Z_1 can be defined by replacing \mathbf{W}_0 with \mathbf{W}_1 in (3).

The correlation matrices are defined as

$$\mathbf{R}_y = \mathcal{E} \{ \mathbf{Y} \mathbf{Y}^H \}, \quad \mathbf{R}_v = \mathcal{E} \{ \mathbf{V} \mathbf{V}^H \}, \quad \mathbf{R}_x = \mathbf{R}_y - \mathbf{R}_v, \quad (4)$$

where $\mathbf{R}_y(k)$ is estimated when speech is present and $\mathbf{R}_v(k)$ is estimated when speech is absent, depending on the decision of a Voice Activity Detector (VAD).

3. BINAURAL NOISE REDUCTION ALGORITHMS

In this section we briefly review the cost functions for the MWF [1], the MWF-ITF [2] and the MWF-ITFhc [3].

3.1. Binaural multi-channel Wiener filter (MWF)

The binaural MWF produces a minimum mean-square error (MMSE) estimate of the speech component in one of the microphone signals for both hearing aids. The MWF cost function estimating the speech components $X_{0,0}$ and $X_{1,0}$ in the left and the right hearing aid can be written as

$$J_{MWF}(\mathbf{W}) = \mathcal{E} \left\{ \left\| \begin{bmatrix} X_{0,0} - \mathbf{W}_0^H \mathbf{X} \\ X_{1,0} - \mathbf{W}_1^H \mathbf{X} \end{bmatrix} \right\|^2 + \mu \left\| \begin{bmatrix} \mathbf{W}_0^H \mathbf{V} \\ \mathbf{W}_1^H \mathbf{V} \end{bmatrix} \right\|^2 \right\},$$

where μ provides a trade-off between noise reduction and speech distortion and the first microphone is used as reference. The filter minimizing $J_{MWF}(\mathbf{W})$ is equal to

$$\mathbf{W}_{MWF} = \mathbf{R}^{-1} \mathbf{r}_x, \quad (5)$$

with

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_x + \mu \mathbf{R}_v & \mathbf{0}_{2M} \\ \mathbf{0}_{2M} & \mathbf{R}_x + \mu \mathbf{R}_v \end{bmatrix}, \quad \mathbf{r}_x = \begin{bmatrix} \mathbf{R}_x \mathbf{e}_0 \\ \mathbf{R}_x \mathbf{e}_1 \end{bmatrix}. \quad (6)$$

The vectors \mathbf{e}_0 and \mathbf{e}_1 are zero column vectors with $e_0(1) = 1$ and $e_1(M+1) = 1$. It has been shown in [2] that for a single speech source the ITF of the output speech and noise component are the same and equal to ITF_x^{in} , implying that all components are perceived as coming from the speech direction, which is generally undesired.

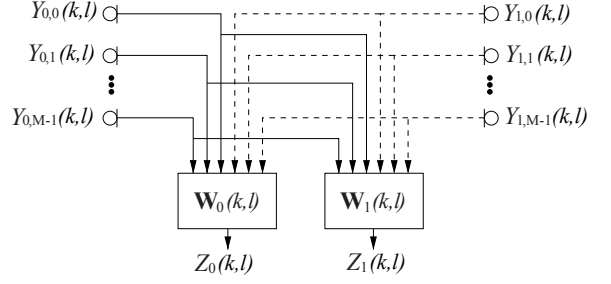


Fig. 1. Binaural hearing aid configuration

3.2. MWF with binaural cue preservation (MWF-ITF)

To reduce the distortion of the output noise ITF, an extension of the MWF cost function with a term related to the ITF of the noise component has been proposed and analyzed in [2]. The filter derived in [2] is equal to

$$\mathbf{W}_{MWF-ITF} = (\mathbf{R} + \delta \mathbf{R}_{vt})^{-1} \mathbf{r}_x, \quad (7)$$

with

$$\mathbf{R}_{vt} = \begin{bmatrix} \mathbf{R}_v & -ITF_v^{des,*} \mathbf{R}_v \\ -ITF_v^{des} \mathbf{R}_v & |ITF_v^{des}|^2 \mathbf{R}_v \end{bmatrix}, \quad (8)$$

where the desired ITF is calculated as

$$ITF_v^{des} = \frac{\mathbf{e}_0^T \mathbf{R}_v \mathbf{e}_1}{\mathbf{e}_1^T \mathbf{R}_v \mathbf{e}_1}, \quad (9)$$

which can be interpreted as an average input ITF. The parameter δ controls the emphasis on the noise ITF preservation term. It has been shown in [2] that the solution is always a trade-off between preserving the binaural cues of the speech component and preserving the binaural cues of the noise component, depending on the parameter δ and the output SNR.

3.3. MWF with instantaneous ITF preservation

To overcome the trade-off between preserving the binaural speech and noise cues, a modification of the MWF-ITF has been presented in [3] by adding a linear hard constraint on the instantaneous noise component ITF to the MWF cost function. The solution of the constrained optimization problem in [3] (MWF-ITFhc) is given by

$$\mathbf{W}_{MWF-ITFhc} = \mathbf{R}^{-1} \mathbf{r}_x - \frac{\mathbf{R}^{-1} \mathbf{C}^H \mathbf{C} \mathbf{R}^{-1} \mathbf{r}_x}{\mathbf{C} \mathbf{R}^{-1} \mathbf{C}^H} \quad (10)$$

with

$$\mathbf{C} = [\mathbf{V}^H \quad -ITF_v^{des,*} \mathbf{V}^H]. \quad (11)$$

With this formulation a perfect preservation of ITF_v^{des} in each frequency bin k and time frame l can be achieved. Note that the first part of (10) is fixed as in the MWF solution (5) and the second part contains the vector \mathbf{C} which is updated in each frame. Since (11) requires the noise signal vector \mathbf{V} to be available - which is not the case during speech segments - an estimate of the noise signal vector $\hat{\mathbf{V}}$ is required.

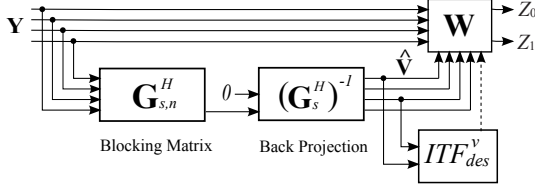


Fig. 2. GSC-like structure of the MWF-ITFhc

4. NOISE ESTIMATION

In [3] the noise signal vector $\hat{\mathbf{V}}$ was estimated using the binaural MWF, with $V_{0,m}$ and $V_{1,m}$ the desired signals, such that for each microphone m in both hearing aids an estimate of the noise component is computed. The performance of this noise estimation procedure relies on an accurate estimate of the second order signal statistics and produces leakage of the speech signal into the noise estimate $\hat{\mathbf{V}}$. To achieve a better estimate of the noise signal vector for the MWF-ITFhc, we propose a GSC-like structure [1], incorporating a blocking matrix and back projection (cf. Fig. 2). The blocking matrix creates a noise reference, by steering nulls towards the undesired source. This noise reference needs to be back projected in order to obtain an estimate of the noise signal vector $\hat{\mathbf{V}}$ in the microphones, which can then be used to compute an instantaneous estimate of the desired noise ITF.

Assuming N independent sources $Q_n(k, l)$ with $n = 1 \dots N$ and $N \leq 2M$, the blocking matrix can be designed exploiting the independency of the sources using blind source separation (BSS) [5], such that an estimate of the n -th source $\hat{Q}_n(k, l)$ is produced, i.e.

$$\hat{Q}_n(k, l) = \mathbf{G}_n^H(k) \mathbf{Y}(k, l) = \mathbf{D}_n^H(k) \mathbf{\Pi}(k) \mathbf{Q}(k, l), \quad (12)$$

with $\mathbf{G}_n(k)$ the n -th column of the $2M \times N$ -dimensional unmixing matrix $\mathbf{G}(k)$. Frequency-domain BSS algorithms inherently suffer from a scaling and permutation ambiguity [5] which is denoted in (12) by the permutation matrix $\mathbf{\Pi}(k)$ and $\mathbf{D}_n(k)$, the n -th column of the scaling matrix $\mathbf{D}(k)$.

For estimating the unmixing matrix $\mathbf{G}(k)$ and hence the blocking matrix $\mathbf{G}_n(k)$ we have used the frequency-domain BSS approach proposed in [6] which uses the Scaled Infomax algorithm [7] and aims to resolve the permutation ambiguities by exploiting the interfrequency dependence of the separated signals based on the power ratio measure of the separated signals. The scaling ambiguity is resolved by using the Minimal Distortion Principle i.e.

$$\mathbf{G}_s(k) = \text{diag}(\mathbf{G}_p^{-1}(k)) \mathbf{G}_p(k), \quad (13)$$

with $\mathbf{G}_p(k)$ being the demixing matrix after permutation alignment and $\mathbf{G}_s(k)$ the unmixing matrix after scaling correction. In (12) $\mathbf{G}_{s,n}(k)$ is now used instead of $\mathbf{G}_n(k)$ to estimate the independent sources. The contribution of the n -th source in the microphone signals can then be estimated using back projection, i.e.

$$\hat{\mathbf{V}}(k, l) = (\mathbf{G}_s^H)^{-1}(k) \begin{bmatrix} \mathbf{0}_{(n-1) \times 1} & \hat{Q}_n(k, l) & \mathbf{0}_{(N-n) \times 1} \end{bmatrix}^T \quad (14)$$

assuming $\hat{Q}_n(k, l)$ to be an estimate of the noise source. In case of perfect source separation in a scenario with only point sources the back projection perfectly recovers the noise signal vector and as such also the binaural cues of the noise signal. The noise signal vector $\hat{\mathbf{V}}$ can now be used for estimating a time-dependent desired ITF, contrary to (9) where ITF_v^{des} is constant over the considered signal. ITF_v^{des} is now estimated as

$$ITF_v^{des}(k, l) = \frac{\langle \hat{V}_{0,0}(k, l) \hat{V}_{1,0}^*(k, l) \rangle}{\langle \hat{V}_{1,0}(k, l) \hat{V}_{1,0}^*(k, l) \rangle}, \quad (15)$$

where $\langle \cdot \rangle$ denotes recursive smoothing.

5. OBJECTIVE PERFORMANCE MEASURES

In this section we briefly discuss the objective measures we have used for performance evaluation.

5.1. Binaural auditory model based ITD/ILD error

For the evaluation of the binaural cue preservation performance, we introduce an objective measure which is based on a model of binaural auditory processing that has been successfully applied for estimating the direction of arrival of speech sources [4]. This model incorporates the middle ear transfer characteristic, auditory band-pass filtering on the basilar membrane using a linear Gammatone filter bank, cochlear compression and half-wave rectification with additional low-pass filtering in the inner hair cells. A temporally smoothed ITD (≈ 5 ms time constant at 1 kHz) is calculated from the complex-valued output signals for each time sample t in the i -th gammatone filter. To discard segments that are not likely to originate from a point source, the interaural vector strength (IVS) has been proposed in [4] as a measure of psychoacoustic decorrelation sensitivity, i.e.

$$IVS_i(t) = \frac{|\int_0^\infty ITF_i(t - \tau) e^{-\tau/\tau_i^s} d\tau|}{\int_0^\infty |ITF_i(t - \tau)| e^{-\tau/\tau_i^s} d\tau}, \quad (16)$$

with $\tau_i^s = 2.5/f_i^c$ and f_i^c is the center frequency of the i -th gammatone filter. From the IVS a binary mask $w_i(t)$ is derived, i.e.

$$w_i(t) = \begin{cases} 1 & \text{if } IVS_i(t) \geq IVS_0 \ \& \ \frac{dIVS_i(t)}{dt} \geq 0 \\ 0 & \text{else} \end{cases} \quad (17)$$

where the threshold IVS_0 was set to 0.98 and the additional condition $\frac{dIVS_i(t)}{dt} \geq 0$ filters out misleading time segments caused by the sluggishness of the IVS due to the lowpass filtering of the ITF [4]. The mean difference of the reliable ITD values are calculated as

$$\Delta ITD = \left| \frac{\sum_i \sum_t w_i^y(t) ITD_i^y(t)}{\sum_i \sum_t w_i^y(t)} - \frac{\sum_i \sum_t w_i^z(t) ITD_i^z(t)}{\sum_i \sum_t w_i^z(t)} \right|, \quad (18)$$

where w_i^y is calculated from the input components and w_i^z is calculated from the output components. The ITD errors are evaluated up to the gammatone filter with a center frequency of 1.4 kHz and all ITDs above $700 \mu s$ are disregarded in the calculation of ΔITD . Ambiguities in the ITD between 700 Hz and 1400 Hz are resolved using the sign of the corresponding ILD values. The ILD errors are evaluated for all gammatone filter similarly to the ITD error by replacing ITD with ILD in (18).

5.2. Intelligibility Weighted SNR

To compare the performance of the algorithms in noise reduction we calculate the Intelligibility Weighted SNR [8] of the input and the output signals. The SNR gain of the left hearing aid is defined as

$$\Delta SNR_0 = \sum_k I(k) \frac{\mathbf{W}_0^H \mathbf{R}_x \mathbf{W}_0}{\mathbf{W}_0^H \mathbf{R}_v \mathbf{W}_0} - \sum_k I(k) \frac{\mathbf{e}_0^T \mathbf{R}_x \mathbf{e}_0}{\mathbf{e}_0^T \mathbf{R}_v \mathbf{e}_0}, \quad (19)$$

where $I(k)$ is a weighting function that takes the importance of different frequency bands for the speech intelligibility into account. ΔSNR_1 is defined in a similar way by replacing \mathbf{W}_0 with \mathbf{W}_1 and \mathbf{e}_0 with \mathbf{e}_1 .

6. EXPERIMENTAL RESULTS

In this section we perform simulations to investigate the binaural cue preservation and noise reduction performance of the MWF, MWF-ITF and the MWF-ITFhc with different noise estimation procedures, for a scenario consisting of one speech source and one noise source.

6.1. Setup

Binaural Behind-The-Ear Head-Related Impulse Responses measured in an office room ($T_{60} \approx 300$ ms) have been used to generate the speech and the noise signals. Each hearing aid was equipped with 2 microphones, therefore in total 4 microphone signals are available. The speech source (10 s taken from the OLSA speech material) was located in front of the listener at 0° and the interfering babble noise source was positioned at an azimuthal angle of 60° (right side of the head). The signals were processed at $f_s = 16$ kHz using a weighted overlap-add (WOLA) framework with a block size of 256 samples, an overlap of 75% between successive blocks and a Hann window.

\mathbf{R}_v was estimated using 5 seconds of a noise-only signal, preceding the noisy speech signal and \mathbf{R}_y was estimated using 10 seconds of noisy speech. The noise-only part was not taken into account during performance evaluation. The parameter μ in (6) was set to 1. The trade-off parameter δ in the MWF-ITF was set to $\delta = 4$, corresponding to a good trade-off between preservation of the speech cues and noise cues for this scenario. ITF_v^{des} was estimated using (9). We have used 2 noise estimation procedures for the MWF-ITFhc:

MWF-ITFhc (MWF) - The noise signal vector was estimated using the MWF, with $V_{0,m}$ and $V_{1,m}$ the desired signals as in

[3] and ITF_v^{des} was estimated using (9).

MWF-ITFhc (GSC) - The noise estimate was calculated using the GSC-like structure (cf. section 4) and ITF_v^{des} was calculated as in (15), where the smoothing corresponds to an averaging over 100 ms. For estimating the blocking matrix a WOLA framework with a block length of 4096 was used.

It is important to note that in the MWF-ITFhc, contrary to the MWF-ITF, no scenario dependent tuning of a trade-off parameter is required due to the hard-constraint formulation. The performance was evaluated for an intelligibility weighted input SNR in the first microphone of the left hearing aid ranging from -6 dB to 6 dB.

6.2. Binaural cue preservation and SNR gain

The results in preserving the *binaural cues of the speech component* are depicted in Figure 3 and 5.

Although in theory the MWF perfectly preserves the ITD/ILD of the speech component it can be noted that low ITD/ILD errors occur, which is due to estimation errors in the speech correlation matrix, especially at low SNRs. As expected the MWF-ITF shows the worst performance in preserving the speech ITD/ILD but the error significantly decreases with increasing input SNR. The MWF-ITFhc (MWF) introduces a slightly higher ITD error compared to the MWF, and the MWF-ITFhc (GSC) shows almost the same performance as the MWF. However, the MWF-ITF is clearly outperformed by the other algorithms in preserving the speech ITD/ILD.

The results in preserving the *binaural cues of the residual noise component* are depicted in Figure 4 and 6.

The performance of the MWF decreases with increasing input SNR, due to a better estimation of the speech correlation matrix in high SNR scenarios. The MWF-ITFhc (MWF) shows an increasing performance with increasing input SNR, however the MWF-ITFhc (MWF) is outperformed by the MWF-ITF. Below an input SNR of 2 dB the MWF-ITFhc (GSC) outperforms all algorithms in preserving the ITD of the residual noise cues. Above an input SNR of 2 dB all considered binaural cue preservation algorithms perform similarly and clearly outperform the MWF.

The *intelligibility weighted SNR gain* for the left and the right hearing aid is depicted in Figure 7 and 8. The MWF-ITFhc (GSC) shows the best performance for the left SNR gain. The MWF-ITF shows a clearly noticeable SNR gain improvement with increasing input SNR. The SNR gain in the right hearing aid is very similar for all algorithms.

7. CONCLUSION

In this paper we have shown that a better preservation of the binaural cues of both the speech and the noise component can be achieved by using a GSC-like structure, comprising a blocking matrix and back projection, for estimating the noise signal vector required in the MWF with instantaneous binaural cue preservation. In future work the more realistic scenario of multiple noise sources and diffuse noise needs to be addressed.

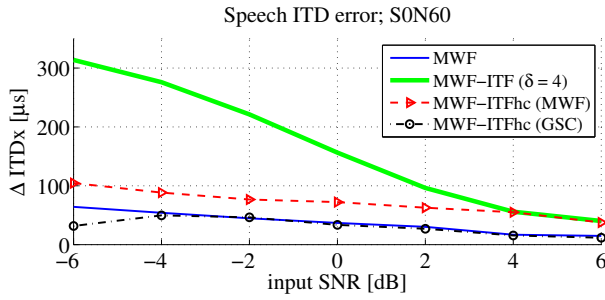


Fig. 3. ΔITD of the speech component

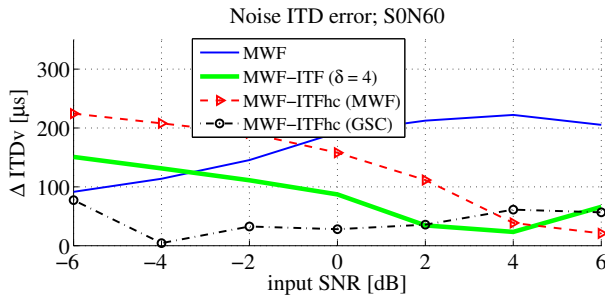


Fig. 4. ΔITD of the noise component

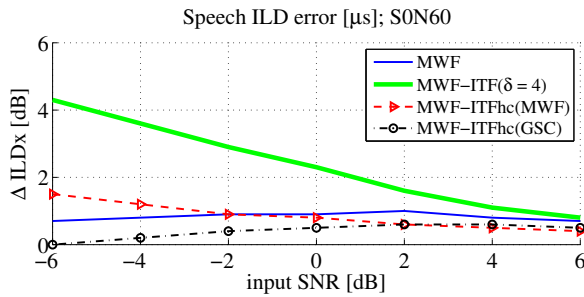


Fig. 5. ΔILD of the speech component

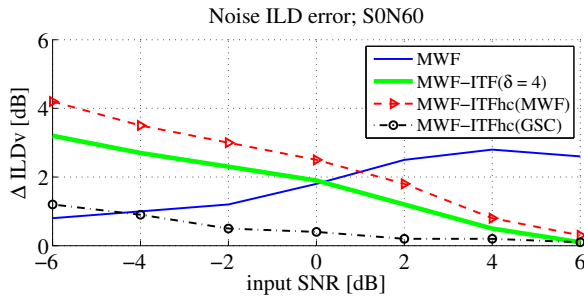


Fig. 6. ΔILD of the noise component

8. REFERENCES

- [1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, pp. 269–302. Wiley, 2010.
- [2] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Trans. on*

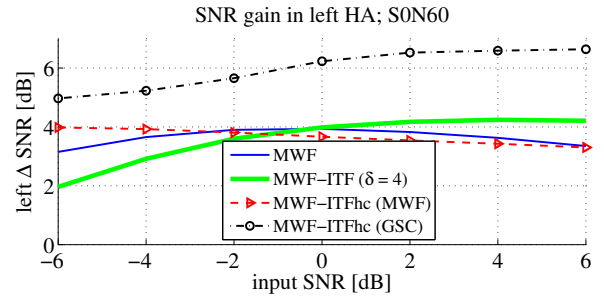


Fig. 7. ΔSNR for the left HA

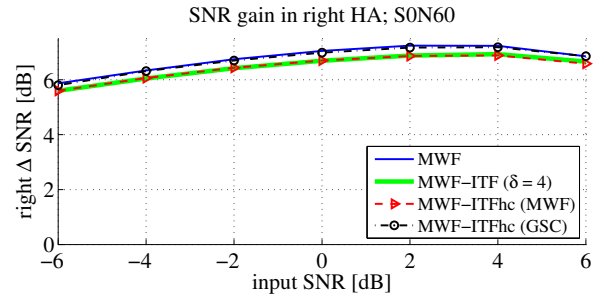


Fig. 8. ΔSNR for the right HA

Audio, Speech and Language Proc., vol. 18, no. 2, pp. 342–355, Feb. 2010.

- [3] D. Marquardt, V. Hohmann, and S. Doclo, "Binaural cue preservation for hearing aids using multi-channel wiener filter with instantaneous ITF preservation," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012.
- [4] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, pp. 592–605, 2011.
- [5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, pp. 299–327. Springer, 2005.
- [6] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 13, pp. 549–557, Mar. 2011.
- [7] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *Proc. ICASSP*, Honolulu HI, USA, Apr. 2007, pp. 637–640.
- [8] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.