Binaural Noise-Reduction Method Based on Blind Source Separation and Perceptual Post Processing

Jorge I. Marin-Hurtado, Devangi N. Parikh, David V. Anderson

1School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA
2Department of Electronics Engineering, Universidad del Quindio, Armenia, Q., Colombia

Abstract

Binaural hearing aids include a wireless link to exchange the signals received at each side, allowing the implementation of more efficient noise-reduction algorithms for hostile environments such as babble noise. Although several binaural noise-reduction techniques have been proposed in the literature, only a few of them preserve localization cues of the target and interfering signals simultaneously without degrading the SNR improvement. This paper proposes a novel binaural noise-reduction method based on blind source separation (BSS) and a perceptual post-processing technique. Objective and subjective tests under four different scenarios were performed. The method showed good output sound quality, high SNR improvement at very low input SNR conditions, and preservation of localization cues for the signal and noise—outperforming both an existing BSS-based method and a multichannel Wiener filter (MWF).

Index Terms: Noise Reduction, Blind Source Separation, Perceptual Post-Processing, Binaural Hearing Aids

1. Introduction

A binaural hearing aid consists of a hearing aid placed on each ear and a wireless link to exchange information between both hearing aids. This arrangement provides significant advantages and user benefits with respect to independent monaural hearing aids, identified through psycho-acoustical studies of hearing perception in hostile environments (e.g., babble noise). These studies show that users prefer binaural noise-reduction strategies over monaural ones (i.e., single-channel noise reduction), and the relevance of the preservation of localization cues for target identification and speech intelligibility [1].

Binaural noise-reduction methods based on scene analysis, spectral subtraction, statistical methods, adaptive beam-forming [2], multichannel Wiener filtering (MWF) [3, 4], and blind source separation (BSS) [5, 6] have been proposed in the literature. The majority of these methods preserve the localization cues for the target signal, but only a few of them preserve the localization cues for the interfering signals [2, 3, 7, 6]. The methods that claim to preserve the localization cues for the interference signal are based on beamforming [2], MWF [3, 7], and BSS [6]. MWF methods provide better performance than beamforming methods [8]; however, the processing delay is high. The BSS method proposed in [6] demands similar or higher computational resources and processing delay than MWF methods, and their performance has not been compared to MWF. In this paper, a comparison between a BSS-based noise reduction method proposed in [6] (Aichner-07) and a MWF method [3] (MWF-N) shows that the BSS-based method provides higher SNR improvement than MWF-N at low input SNRs, but the quality of the output sound is lower than MWF-N. In addition, our subjective test shows that the Aichner-07 method is unable to preserve the localization cues for the interfering signals. These results motivate the exploration of alternative methods based on BSS.

We can improve the binaural BSS-based method by improving the post-processing state. When two-output BSS is used in binaural noise-reduction applications, the primary BSS output provides an estimate of the target signal, and the secondary BSS output, an estimate of the interfering signal. The processing is followed by post-processing to recover the localization cues. Aichner et al. [6] analyzed two different post-processing methods that were presented in [5, 6] and concluded that a post-processing method based on an adaptive filter for each ear (Aichner-07) provides the best performance. However, as mentioned before, in our experiments, the performance of Aichner-07 matched the performance of MWF-N. Thus, in this paper, we propose to use a BSS post processing inspired by the auditory perceptual model [9]. We selected this post-processing since it outperforms other BSS post-processing for speech-enhancement applications. This post-processing is modified so that it can be used for a binaural hearing aid. The proposed method outperforms the Aichner-07 and MWF-N methods, and it preserves the localization cues of both the target and interfering signals correctly.

2. Proposed method

The method proposed in [9] cannot be used directly for binaural hearing aids since it does not preserve localization cues. To recover the localization cues, the gains obtained by the BSS and perceptual post-processing algorithm are applied to the original unprocessed signals received at each side (Figure 1). If the original interaural time differences (ITD) and interaural level differences (ILD) remain unmodified in the enhanced signals, the localization cues for both target and interfering signals are preserved. Thus, applying the same set of gains to the original unprocessed signals ensures that the ITD and ILD are unmodified. Moreover, modifications have to be made to [9] so that the algorithm can be implemented in real time. To achieve low processing delay, we compute the gains and output signal on a sample-by-sample basis, and the parameters used to estimate these gains are updated on a frame-by-frame basis.

The block diagram of the proposed binaural noise-reduction technique is shown in Figure 1. Signals received at the left, $x_1$, and right, $x_2$, microphones are passed through a BSS algorithm to get the signals $u_1$ and $u_2$. An output selection algorithm...
identifies which BSS output contains the separated target signal \((y_1)\), or primary channel, and the separated interfering signal \((y_2)\), or secondary channel. These outputs, \(y_1\) and \(y_2\), are analyzed using a constant-Q filter bank, and then, the envelope in each sub-band is computed. These envelopes are used to estimate the signal-to-noise ratio (SNR) and to compute the noise-suppression gains. The SNR and gain are computed separately for each sub-band. These noise-suppression gains are computed to lower the noise floor, ensuring a perceptual removal of the noise in each sub-band. These gains are finally applied simultaneously to the original unprocessed signals by time-domain multiplication, and the output from each sub-band is summed together to produce the signals for the left and right ear.

The info-max BSS algorithm [10] was chosen because of low computational complexity and short processing delay. In the info-max method, unmixing filter coefficients are obtained by minimization of the cumulative density function (CDF) for the target signal. This CDF is approximated by a hyperbolic tangent function. The BSS block indicated in the Figure 1 is described by

\[
\begin{align*}
    u_1(n+1) &= x_1(n) + w_{12}^T(n)u_2(n) \\
    u_2(n+1) &= x_2(n) + w_{21}^T(n)u_1(n) \\
    w_{12}(n+1) &= w_{12}(n) - 2\mu \tanh(u_1(n+1))u_2(n) \\
    w_{21}(n+1) &= w_{21}(n) - 2\mu \tanh(u_2(n+1))u_1(n),
\end{align*}
\]

where \(x_1\) and \(x_2\) are the signals received at the left and right microphones, \(w_{12}\) and \(w_{21}\) are vectors of length \(N_u\) describing the unmixing filter coefficients, and \(u_1(n)\) and \(u_2(n)\) are vectors of length \(N_u\) whose elements are the previous outputs of the BSS algorithm, \(u_j(n) = [u_j(n) \ u_j(n-1) \ldots u_j(n-N_u+1)]^T\), \(j = 1, 2\), and \(n\) is the time index. To determine which BSS output contains the target signal, the energy of the envelopes of the signals \(u_1\) and \(u_2\) are compared. The output with the higher envelope energy is chosen as primary channel \(y_1\). To deal with moving sources, this update takes place every \(N\) samples.

The outputs of the BSS algorithm \(y_1\) and \(y_2\), as well as when the original time-domain unprocessed input signals at the left and right microphones, \(x_1\) and \(x_2\), are passed through a filter bank that resembles the auditory system. This filter bank was implemented using forth-order Butterworth filters. At 22 kHz sampling rate, each filter bank provides 24 sub-bands. At the output of the filter banks, the vectors \(x_j(l, k)\) and \(y_j(l, k)\) of length \(N\), \(j = 1, 2\), are obtained, where \(l\) corresponds to the frame index and \(k\) to the sub-band number. Although the signals \(x\) and \(y\) are obtained in the sample-by-sample basis, they are analyzed in non-overlapped frames of length \(N\) in order to compute the gain parameters as we will show next.

For each output \(y_j(l, k)\), the envelope is extracted using a full-wave rectifier followed by a low-pass filter. In particular, the primary envelope vector \(e_p(l, k)\) is extracted from \(y_1(l, k)\), and the secondary envelope vector \(e_s(l, k)\) from \(y_2(l, k)\). The low-pass filters are implemented using a first-order IIR filter whose cutoff frequency is selected to a fraction of the corresponding bandwidth of the band [9]. These cutoff frequencies were set to 1/5, 1/8 and 1/15 of the bandwidth of low, medium and high frequency bands, respectively. These fractions ensure that the envelope tracks the signal closely but at the same time does not change too rapidly to cause the gain to change rapidly.

The final outputs at the left \(z_1\) and the right \(z_2\) side are computed using the time-domain gains \(g_{1, k}\) produced by the perceptual post-processing stage:

\[
    z_j(l) = \sum_k g_{1, k} \circ z_j(l, k)
\]

where \(\circ\) denotes the element-wise product. The above equation is written in a vector form to state that the gains are computed using parameters updated on a frame-by-frame basis. However, the output values can be computed on a sample-by-sample, reducing the processing delay.

In [9], a method inspired by a perceptual modeling uses the entire knowledge of the signal to estimate these multiplicative gains. In this method, the gain modifies the envelope of each sub-band \(e_k\) such that \(e_k(l) = \beta e_k^\alpha(l)\). To provide noise reduction, the maximum envelope value is preserved (i.e. \(e_{\text{max}} = e_{\text{max}}\)) while the minimum envelope value is lowered (i.e. \(e_{	ext{min}} = K e_{\text{min}}\), where \(K\) is an expansion coefficient). Using the previous ideas, [9] developed a method to estimate \(\alpha\) and \(\beta\) for the entire signal. To provide a realistic implementation, equations in [9] are modified to a vector form to state the update of \(\alpha\) and \(\beta\) on a frame-by-frame basis every \(N\) samples:

\[
    g_{1, k} = \beta_{1, k} e_p(l, k)^{(\alpha_{1, k} - 1)}
\]

The factors \(\alpha\) and \(\beta\) are computed as

\[
    \beta_{1, k} = \max(e_{\text{max}}(k))^{1 - \alpha_{1, k}}
\]

\[
    \alpha_{1, k} = 1 - \log K / \log M_{1, k}
\]

where \(M_{1, k}\) is SNR estimated at \(k\)-th sub-band and frame \(l\), and \(e_{\text{max}}(k)\), a vector that holds the maximum values of the primary envelopes, at the sub-band \(k\), obtained from the previous \(N_{\text{max}}\) frames:

\[
    e_{\text{max}}(k) = \max(e_p(l, k) \ldots \max(e_p(l - N_{\text{max}}, k)))
\]

To avoid computational overflow, the value of \(\alpha\) was constrained to be in the range \([0, 5]\). To minimize artifacts and to achieve better quality outputs, the history stored in the vector \(e_{\text{max}}\) should hold at least one second. All experiments use a two-seconds memory, i.e. \(N_{\text{max}} = [2f_s/N]\). Since \(\alpha\) and \(\beta\) are fixed for a given frame, the gains can be also be computed in the sample-by-sample basis.

To estimate the SNR at the given sub-band and frame, the signal and noise power are obtained from the envelopes of the primary and secondary channel. Primary and secondary BSS outputs are used to reduce miss-classification errors in the SNR estimation when the input SNR is low. To obtain a reliable noise estimate, the noise power is updated using a rule derived from the noise PSD estimator proposed in [11]:

\[
    P_e = ||e_s(l, k)||^2
\]

\[
    \text{if}\ P_e - P_e(l - 1, k) < \epsilon \sqrt{\sigma_e(l - 1, k)}
\]

\[
    P_e(l, k) = \lambda_e P_e(l - 1, k) + (1 - \lambda_e) P_e
\]

\[
    \sigma_e(l, k) = \delta \sigma_e(l - 1, k) + (1 - \delta) [P_e - P_e(l - 1, k)]^2
\]

where \(P_e(l, k)\) is the noise power at the \(k\)-th sub-band and frame \(l\), \(\sigma_e(l, k)\) is an estimate of the variance of \(P_e\), \(\lambda\) and \(\delta\).
are time-constants to smooth the estimation, and $\epsilon$ is a threshold coefficient. Finally, the frame SNR is estimated by

$$M_{l,k} = \max \left( \frac{P_x(l,k)}{P_x(l,k)} - 1, 1 \right)$$  \hspace{1cm} (7)$$

where $P_x$ is the power of the primary channel estimated by

$$P_x(l,k) = \lambda_x^2 P_x(l-1,k) + (1 - \lambda_x) |e_x(l,k)|^2$$  \hspace{1cm} (8)$$

The values $\lambda_x = 0.95$, $\lambda_p = 0.9$, $\delta = 0.9$, and $\epsilon = 5$ were found to provide good performance in our experiments.

3. Experimental setup

The performance of the proposed method (BSS-PP) is compared to the existing methods described in [6] (Aichner-07) and [3] (MWF-N). Four different scenarios were tested: diffusive noise, babble noise, single interfering speech signal, and 4 distinguishable speakers placed at different locations (multi-talker scenario). Mixtures for these scenarios were created by filtering the target signal with the head related transfer functions (HRTF) measured for a KEMAR manikin [12]. The target signal was placed at eight different azimuth angles: 0°, 30°, 90°, 120°, 180°, 240°, 270° and 330°, where 0° corresponds to the front of the KEMAR, 90° corresponds to the right ear, and 270° to the left ear. Target signals were speech recordings of ten different speakers and sentences. For all scenarios, the interfering signals were added to the target signal at different SNR. For diffusive noise scenario, different uncorrelated pink noise sources were played simultaneously at 18 different spatial locations. For the babble noise scenario, the noise source was recorded in a cafeteria and added to the speech samples processed with the HRTFs. For the single interfering scenario, interference was located at 40°, and for the multi-talker scenario, interfering signals are speech samples placed at 40°, 80°, 200° and 260°. For all mixtures, a sampling frequency of $f_s = 22$ kHz was used.

The performance of the proposed method was analyzed using the broadband intelligibility weighted SNR improvement ($\Delta$SNR) [13]. The $\Delta$SNR values reported in this paper correspond to the average over all target speakers and angles. To assess the quality, we conducted a subjective test.

4. Results and discussion

The performance of the proposed method depends on the expansion parameter $K$ and the update frame-length $N$. Different experiments were carried out to determine the optimal values of $K$ and $N$. We established that the SNR improvement is very sensitive to the parameter $K$, and almost independent of the parameter $N$. In particular, values of $K \leq 0.01$ provide similar SNR improvement, and values of $K > 0.01$ produce a noticeable performance reduction. On the other hand, the sound quality increases with larger update frame-length $N$. However, no significant improvement in the sound quality was found for $N > 8192$ under all scenarios. Hence, the values of $K = 0.01$ and $N = 8192$ were used for all experiments.

For diffusive and babble noise scenarios, $\Delta$SNR for MWF-N and Aichner-07 decreases with the input SNR (Figs. 2 and 3). This is not the case for BSS-PP, which provides higher SNR improvement when the input SNR increases. For these scenarios, the proposed method (BSS-PP) outperforms the Aichner-07 and MWF-N methods.

For multi-talker (Fig. 4), the proposed method outperforms Aichner-07 and MWF-N for input SNR higher than 0 dB. In addition, the $\Delta$SNR for the proposed method increases with an increasing in the input SNR. These behaviors can be explained by the fact that the BSS algorithm is more efficient to unmix directional interfering signals than uniform background noises, and at very low SNR conditions, the interfering signal is confused with the target signal. In a very low SNR case, the proposed noise reduction algorithm considers the target signal as the interfering signal, and produces an enhancement of the interfering signal instead of the target signal.

For the above reasons, the proposed algorithm is useful only to deal with background noises that are distributed uniformly in the space (e.g. babble noise), or in multi-talker scenarios when the input SNR is higher than 0 dB.

A subjective test was conducted to verify the performance of the proposed method. In the test, only babble noise and single interfering signal scenarios were tested. For babble noise, the input SNR was fixed at 0 dB, and for the single interfering signal scenario the SNR was 5 dB. The babble-noise samples were used to determine the ability of the algorithm to reduce the background noise, while the single-interfering samples were used to identify the ability of the algorithm to preserve the localization cues of both target and interfering signals. The test
samples included clean, unprocessed mixtures, and enhanced signals with the Aichner-07, MWF-N, and BSS-PP methods. These samples were presented randomly to the subject, and the subjects were asked to grade the quality of the target speech and the noise level. They were also asked to identify the direction of arrival for the target and interfering signals. A total of 22 subjects participated in the experiment.

Subjective test results are shown in Table 1. Subjects rated the speech quality from 1 (poor quality) to 5 (excellent quality), and the noise level from 1 (very noisy) to 5 (clean). The percentages reported for the preservation of localization cues correspond to the number of trials in which subjects identified correctly the direction of arrival. A statistical analysis using analysis of variance (ANOVA), and a multiple comparison statistical test according to Tukey’s HSD were used to assess significant differences using a 95% of confidence. Significant differences were obtained for speech quality, noise reduction, and noise cues ratings. Since no significant differences were obtained for the identification of target cues, all techniques are able to preserve the localization cues of the target signal. A different situation is present for the identification of the noise cues. In this case, the BSS-PP and MWF-N methods showed no significance difference with respect to the original signal, but there is a significant difference with respect to the Aichner-07 method, which supports the assumption that Aichner-07 method is unable to preserve the noise cues. With respect to quality, MWF-N showed the highest quality score, and it was the only technique with no significant difference with respect to the original noisy signal. This result suggests that subjects preferred an output sound quality more close the original unprocessed signal. With respect to noise level, BSS-PP was rated with the highest score, and the three techniques showed significant difference among them, which supports the claim that the proposed method is able to provide better noise reduction.

Finally, the proposed method provides processing advantages over other binaural noise-reduction methods. To process an input sample, the proposed method requires $4L_w + 31S + 6$ multiplications, $4L_w + 29S + 1$ additions, and 1 division, 2 hyperbolic tangents, and 1 power raising. $S$ is the number of sub-bands. The total number of operations per input sample is similar to MWF and lower than Aichner-07. Moreover, the processing delay in the proposed method is limited by the FIR unmixing filter of the BSS algorithm. This processing delay is lower than $L_w/2$ samples, or equivalently, lower than 1 ms for a 22 kHz sampling rate. Processing delays in the MWF and Aichner methods are around 6 ms.

5. Conclusions

A method based on blind source separation and perceptual post-processing technique for noise reduction in binaural hearing aids is presented. An objective metric—SNR improvement ($\Delta$SNR), and subjective tests were conducted to verify the capacity of the proposed method to reduce the background noise, provide good sound quality, and to preserve the localization cues. Results show that the proposed method outperforms existing methods based on BSS (Aichner-07) [6] and MWF (MWF-N) [3]. The $\Delta$SNR for the proposed method is always positive and above 4 dB. The proposed method shows to preserve the localization cues for both target and interfering signals, which is not the case of Aichner-07, in which the localization cues for the noise are lost.

6. References