A MACHINE LEARNING APPROACH FOR COMPUTATIONALLY AND ENERGY EFFICIENT SPEECH ENHANCEMENT IN BINAURAL HEARING AIDS

David Ayllón[⋆†]*, Roberto Gil-Pita*[†] *and Manuel Rosa-Zurera*[†]

[⋆]R&D Department, Fonetic, Spain
[†]Department of Signal Theory and Communications, University of Alcala, Spain

ABSTRACT

A binaural speech enhancement algorithm that combines superdirective beamforming with time-frequency (TF) masking is proposed. Supervised machine learning is used to design a speech/noise classifier that estimates the ideal binary mask (IBM), which is further softened to reduce musical noise. The method is energy-efficient in two ways: the computational complexity is limited and the wireless data transmission optimized. The experimental work demonstrates the ability of the method to increase the intelligibility of speech corrupted by different types of noise in low SNR scenarios.

***Index Terms*—** Speech enhancement, Binaural hearing aids, Machine learning, Time-frequency masking.

1. INTRODUCTION

Binaural hearing aids improve the ability to localize and understand speech in noise, but with the ensuing increase in power consumption due to wireless data transmission. Roughly speaking, the current technology demands as much power to communicate both hearing aids as that required for the signal processing on a monaural device [1]. Binaural systems work with dual-channel input-output signals, although more than one microphone could be placed in each device. In the last years, binaural beamforming has been proposed for speech enhancement in binaural systems [2, 3, 4], but they only are able to preserve the spatial cues of the target source, which may cause some hearing discomfort.

Most works focused on binaural beamforming assume that the signals received at the right and left devices are available at both sides, which involves a high bandwidth communication. In practice, the signals are quantized before being transmitted, and the power consumption directly depends on the amount of exchanged information. This fact opens a new line of research: how to reduce the transmission bit rate without decreasing the performance of the enhancement system. Some of the first works in this direction are [5, 6, 7]. Unfortunately, the performance of these algorithms is notably affected when the bit rate decreases (e.g. lower than 16 kbps). Additionally, there is a problem associated to the use of binaural beamforming in hearing aids: the output of the beamformer (BF) is obtained by combining a weighted version of the input channels from both devices. If one or several input signals have been quantized and transmitted to the other device, the beamforming output is directly affected by quantization noise.

Recently, the work in [8] has proposed a novel schema for speech enhancement in binaural hearing aids. The algorithm is energy-efficient in two ways: the computational cost is limited and the data transmission optimized. Speech enhancement is obtained by (TF) masking. The ideal binary mask (IBM) [9] is estimated with a speech/noise linear classifier designed using supervised machine learning. Inspired in [8], the present work considers multiple input channels in each device. The new schema combines a fixed superdirective BF with TF masking. The fixed BF is able to reduce a high level of omnidirectional noise but it fails when rejecting directional noise [10]. The directional noise that remains at the output of the BF is removed by TF masking. A least squares linear discriminant analysis (LS-LDA) is designed to estimate the IBM, which is subsequently softened to reduce musical noise. The output speech intelligibility is evaluated with different types of noise.

2. PROPOSED ALGORITHM FOR AN EFFICIENT BINAURAL SPEECH ENHANCEMENT

Let us consider two wireless-connected hearing aids, each device containing N input channels. The signals impinging on the n-th microphone of the left (L) and right (R) devices are

$$x_{L/Rn}(t) = s_{L/Rn}(t) + \sum_{j=1}^{J} n^d_{L/Rnj}(t) + n^o_{L/Rn}(t) \quad (1)$$

where $s_{L/Rn}(t)$ are the contributions of the desired speech source to the L/R n-th microphone, $\sum_{j=1}^{J} n^d_{L/Rnj}(t)$ are the addition of J directional noise sources, and $n^o_{L/Rn}(t)$ are diffuse noise. The goal of the speech enhancement system is to produce an intelligible estimation of the original speech source, $s_{L/R}(t)$, from the corrupted input signals, $x_{L/Rn}(t)$. In addition, we assume that the target speaker is localized in

This work has been funded by the Spanish Ministry of Economy and Competitiveness, under project TEC2012-38142-C04-02

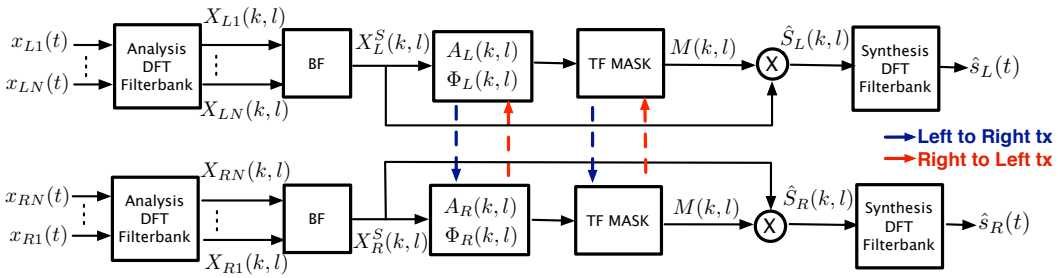

Fig. 1: Binaural speech enhancement system overview.

the straight ahead direction since, in a normal situation, the person is looking at the desired speaker.

Fig. 1 shows an overview of the binaural speech enhancement system proposed in this paper. The desired signal is enhanced in two steps: beamformation of the multichannel input signals in each device, and TF masking of the binaural steered signals. The second step requires the exchange of data between devices, and this wireless transmission is optimized to minimize power consumption and maximize speech enhancement at the same time.

2.1. Robust superdirective beamforming

As a first step to enhance the desired speech signal, each device includes a fixed superdirective BF steered to the straight-ahead direction (target source). A fixed superdirective beamforming is a computationally affordable solution to remove omnidirectional noise in hearing aids, since the filter coefficients can be pre-calculated and stored in the memory of the device.

The DFT of each time frame of the input signals is calculated by the analysis filterbank, obtaining $\mathbf{x}_{L/R}(k,l) = [X_{L/R1}(k,l), \cdots, X_{L/RN}(k,l)]^T$, where k represents frequency, $k = 1, \cdots, K$, and l the time frame, $l = 1, \cdots, L$. The steered signals are $X^S_{L/R}(k,l) = \mathbf{w}(k)^H \mathbf{x}_{L/R}(k,l)$, where $\mathbf{w}(k) = [W_1(k), \cdots, W_N(k)]^T$ is the frequency-domain weight vector, which is the same in both devices due to symmetry. In the proposed solution, a robust superdirective BF based on the minimum variance distortionless response (MVDR) filter [11] is implemented. The amplification of incoherent noise is avoided by establishing a lower limit on the white noise gain, as proposed in [12].

2.2. TF masking based on supervised machine learning

The second step is to calculate a TF mask to isolate the desired source from the directional and omnidirectional noise remaining at the output of the BF. A computationally affordable supervised machine learning algorithm is designed to estimate the IBM from the information contained in the left and right steered signals, $X^S_{L/R}(k,l)$, information that must be previously exchanged between devices. Particularly, the amplitudes (in dB) of the TF signals ($A_{L/R}(k,l)$) and the phases ($\Phi_{L/R}(k,l)$) are quantized and transmitted through the wireless link. Each device uses the information received from the other device and its own information to estimate the TF mask ($M(k,l)$). It is important to highlight that, in order to preserve the binaural cues, the TF mask applied in both devices must be the same. The output enhanced signals are obtained by applying the TF mask to the steered signals: $\hat{S}_{L/R}(k,l) = M(k,l) \cdot X^S_{L/R}(k,l)$. The synthesis filterbanks convert the enhanced TF signals into the time-domain ($\hat{s}_{L/R}(t)$).

According to the low computational resources available in hearing aids, the estimation of the IBM should be simple. The proposed method is based on a LS-LDA [13] designed to classify a TF point as speech or noise. A different classifier is designed for each frequency band k. Let us formulate the LS-LDA problem. The pattern matrix $\mathbf{Q}(k)$ of dimensions $((P+1)\text{x}L)$ contains the P input features of a set of L patterns (time frames) and a row of ones for the bias. The output of a LDA is obtained as a linear combination of the input features, $\mathbf{y}(k) = \mathbf{v}(k)^T \mathbf{Q}(k)$, where $\mathbf{y}(k) = [y(k,1), \ldots, y(k,L)]^T$ is a $(L\text{x}1)$ column-vector containing the output of the LDA and $\mathbf{v}(k) = [v(k,1), \ldots, v(k,P+1)]^T$ contains the bias and the weights applied to each of the P input features. For each of the patterns, the TF binary mask is generated according to

$$M(k,l) := \begin{cases} 1, & y(k,l) > y_0 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where y_0 is a threshold value set to $y_0 = 0.5$. In the case of least squares, the weights are adjusted to minimize the MSE of the classifier, $MSE(k) = \frac{1}{L} \|\mathbf{t}(k) - \mathbf{y}(k)\|^2$, where $\mathbf{t}(k) = [t(k,1), \cdots, t(k,L)]^T$ contains the target values that, in our problem, correspond with the IBM: '1' for speech and '0' for noise. The target IBM is calculated according to

$$t(k,l) := \begin{cases} 1, & P_S(k,l) > P_N(k,l) \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $P_S(k,l) = |S^S_L(k,l)|^2 + |S^S_R(k,l)|^2$ and $P_N = |\sum_{j=1}^{J} N^{dS}_{Lj}(k,l) + N^{oS}_L(k,l)|^2 + |\sum_{j=1}^{J} N^{dS}_{Rj}(k,l) + N^{oS}_R(k,l)|^2$, and $()^S$ means steered signal (i.e. BF output). To adjust the weights of the LS-LDA, the next optimization problem should be solved:

$$\hat{\mathbf{v}}(k) = \min_{\mathbf{v}(k)} \{ \left\| \mathbf{t}(k) - \mathbf{v}(k)^T \mathbf{Q}(k) \right\| \}. \tag{4}$$

Providing that the columns of matrix $\mathbf{Q}(k)$ are linearly independent, the minimization problem has a unique solution, and the weights are given by $\hat{\mathbf{v}}(k) = \mathbf{t}(k)\mathbf{Q}(k)^T \left(\mathbf{Q}(k)\mathbf{Q}(k)^T\right)^{-1}$. Finally, the binary mask is estimated with (2) and softened to reduce musical noise. The solution adopted in this work is very simple but effective: values of '1' are left unmodified, and values of '0' are replaced by an attenuation factor of 15 dB (different values have been tested).

The study carried out in [8] found that the most suitable set of features for the classification problem at hand, considering a tradeoff between the MSE of the classifier and computational cost, is $[A_L, abs(A_L - A_R), abs(\Phi_L - \Phi_R)]$. The study was performed with a system implemented asymmetrically (the mask was entirely calculated in one device). Hence, in the proposed symmetric implementation, the input features for the left device are $[A_L, abs(A_L - A_R), abs(\Phi_L - \Phi_R)]$ and for the right device are $[A_R, abs(A_L - A_R), abs(\Phi_L - \Phi_R)]$. Additionally, it was found that the information provided by the features calculated in neighbor time-frequency points is very valuable to the classifier. The use of 3 neighbor frequencies taken in each direction (upper frequencies and lower frequencies) and the use of 2 previous time frames represented a good tradeoff between signal enhancement and computational cost. According to this, the total number of features used by the classifier to classify each TF point is $P = 27$.

2.3. Transmission schema to optimize the power consumption

In order to limit the number of bits transmitted through the wireless link (and the power consumption), we propose to transmit a low bit rate version of $A_{L/R}(k,l)$ and $\Phi_{L/R}(k,l)$, where the number of bits used to code the amplitude and phase values may differ and they also may differ in each frequency band. Henceforth, the quantized values are denoted as $A_{L/R}^{B_{Ak}}(k,l)$ and $\Phi_{L/R}^{B_{Pk}}(k,l)$, where B_{Ak} is the number of bits used to code the amplitudes of the k-th band, and B_{Pk} the number of bits used to code the phases of the k-th band. $B_k = B_{Ak} + B_{Pk}$ represents the total number of bits transmitted per frequency band. If the total number of bits transmitted through the wireless channel is limited (i.e. the bit rate), they can be distributed among the different values of B_{Ak} and B_{Pk}, and this bit distribution can be optimized to maximize the output speech enhancement. According to this, the next optimization problem is formulated

$$\min_{B_{Ak}, B_{Pk}} MSE, \quad \text{s.t.:} \sum_{k=1}^{K} B_k \leq B_{LIMIT}, \qquad (5)$$

where $MSE = 1/K \sum_{k=1}^{K} MSE(k)$, and B_{LIMIT} the maximum number of transmitted bits. The values of B_{Ak} and B_{Pk} are limited between 0 and 8. Allowing to assign a value of 0 bits avoid the transmission of unnecessary information. Finding a closed solution for the optimization problem in (5) is quite complex, and its solution is approximated by a tailored evolutionary algorithm. The algorithm searches the best allocation of bits among frequency bands in order to minimize the average MSE (fitness function). Each candidate solution is a vector containing the number of bits (between 0 and 8) assigned to B_{Ak} and B_{Pk}. The details of the optimization algorithm can be found in [8].

The transmission schema is further optimized being implemented symmetrically: each device only computes the mask corresponding to half of the frequency bands and transmit it to the other device. This schema allows the devices to transmit only half of the quantized values of their amplitude and phase. If the left device computes the mask for the first half of bands, $M([1, \cdots, k/2], l)$, it should transmit $A_L^{B_{Ak}}([k/2 + 1 - N_{frecs}, \cdots, K], l)$ and $\Phi_L^{B_{Pk}}([k/2 + 1 - N_{frecs}, \cdots, K], l)$. The right device then computes the mask corresponding to the second half of bands, $M([k/2 + 1, \cdots, K], l)$ and transmits $A_R^{B_{Ak}}([1, \cdots, k/2 + N_{frecs}], l)$ and $\Phi_R^{B_{Pk}}([1, \cdots, k/2 + N_{frecs}], l)$.

2.4. Computational cost of the proposed system

The computational cost is measured in number of instructions per frequency band (IPF) required to process each time frame. The analysis and synthesis filterbanks are usually implemented in a specific processor, so these operations are not considered. The implementation of the spatial filters require N complex MAC operations for each band ($IPF = 2N$). The estimation of the TF mask involves the next steps: extraction of the input features ($IPF = 50$), LS-LDA ($IPF = 28$) and mask generation ($IPF = 4$), totalling $IPF = 82$. The application of the mask only requires 1 instruction. According to this, the total computational cost, with $N = 2$, is IPF=87. Considering a state-of-the-art commercial hearing aid, this represents only a 28% of the available IPF for signal processing [8].

3. EXPERIMENTAL WORK

3.1. Description of the experiments

A database of 3000 speech-in-noise binaural signals has been generated. It is split in two sets, one to design the speech/noise classifier (50 %) and other to test the algorithm (50 %). Speech signals are selected from the TIMIT database [14] and noise signals from an extensive database (1000 records) that contains both stationary and non-stationary noise. With the purpose of generalization, the speech and noise signals used to generate the test set are not included in the design set. Binaural mixtures are generated using the head-related impulse responses (HRIR) included in the CIPIC database [15]. Three different types of mixtures are generated: Type 1) 500 mixtures of speech with diffuse noise and two directional noise sources; Type 2) 500 mixtures of speech

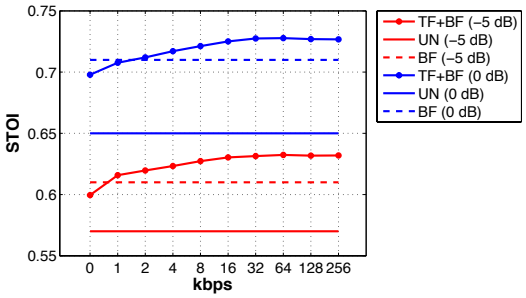

Fig. 2: Average STOI as a function of the transmission bit rate (kbps) for mixtures with SNR= -5 dB and SNR= 0 dB.

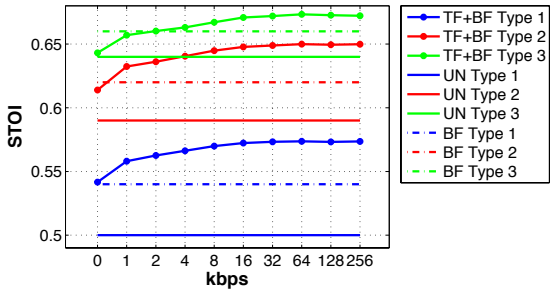

Fig. 3: Average STOI as a function of the transmission bit rate (kbps) and the type of noise. SNR= -5 dB.

with two directional noise sources; Type 3) 500 mixtures of speech with diffuse noise. Speech sources are placed in the front position, the two directional noise sources are placed at each side of the head at random positions, and diffuse noise is simulated by generating isotropic speech-shaped noise. The sampling rate is 16 kHz and the signals are transformed into the TF domain with a short-time Fourier transform (STFT) that uses a 128-points Hanning window with 50% of overlap (K = 64). Each hearing aid contains two microphones in endfire configuration, separated a distance of 0.7 cm.

The optimization problem formulated in (5) has been solved using different values of B_{LIMIT}, from 0 to 256 kbps. All the experiments have been repeated with SNR of 0 dB and -5 dB, which are low SNR values. The performance of the system is measured with the short-time objective intelligibility measure (STOI) proposed in [16], which shows high correlation with the intelligibility of TF weighted noisy speech. STOI values range from 0 to 1, higher values corresponding with higher intelligibility.

3.2. Results

Fig. 2 represents the obtained STOI values (averaged over the test set) as a function of the transmission bit rate (kbps) for mixtures with a SNR= -5 dB (red) and SNR= 0 dB (blue). It also shows the average STOI values of the unprocessed signals and the signals at the BF output (horizontal lines). The obtained STOI values demonstrate that the proposed system increases the output speech intelligibility. In the case of SNR=-5 dB, the initial average STOI has a value of 0.56, which is increased up to 0.61 at the output of the BF, which is an important increment. The application of the TF mask estimated with the proposed classifier obtains average STOI values around 0.64, and this value is kept practically constant for bit rates down to 8 kbps. Except in the case of 0 kbps, the STOI obtained by the estimated TF mask is higher than the one obtained at the output of the BF. The same relative behaviour is found in the case of SNR = 0 dB, but with higher STOI values.

Fig. 3 represents the average STOI values separated in different types of noise, for SNR=-5 dB. As it was expected, the lowest STOI values are obtained in the case of type 1, since speech is contaminated with the two types of noise. Comparing the results of type 2 and type 3, we can deduce that directional noise decreases more the output intelligibility than omnidirectional noise with the same power. However, the intelligibility improvement introduced by the proposed system is more noticeable in the case of type 1, followed by type 2, and finally in type 3. The differences between the beamforming output and the output of the TF mask are similar in the cases of type 1 and type 2, but they are smaller in the case of type 3. That means that most of the energy of the diffuse noise is already removed by the BF, and the TF mask does not introduce a noticeable improvement. Specifically, for bit rates lower than 4 kbps, the application of the TF mask is not beneficial if there is only diffuse noise.

4. CONCLUSIONS

From the results obtained in this work we can conclude that the proposed binaural speech enhancement system is able to increase the output speech intelligibility of speech corrupted with different types of noise in low SNRs, even with low transmission bit rates. In addition, the system is energy-efficient: it requires less than a 28% of the available computational resources and the transmission bit rate has been limited to reasonably affordable values that guarantee a minimum battery life, allowing to find a tradeoff between transmission bit rate and system performance.

Furthermore, the obtained results demonstrate that directional noise affects more the intelligibility than diffuse noise. Most of the diffuse noise power is removed by the BF, whereas most of the remaining directional noise power is removed by the TF mask. In an acoustic scenario when only omnidirectional noise is present, the application of the TF mask does not increase the output speech intelligibility as much as in cases where directional noise is also present, at least for low bit rates. From these results arose the idea of using an acoustic environment classifier, which is usually included in current hearing aids, to detect the presence of directional or diffuse noise and to decide whether to apply the TF mask or not. This problem should be further investigated in the future.

5. REFERENCES

[1] J.M. Kates, *Digital Hearing Aids*. Plural Pub, 2008.

[2] D.R. Campbell and P.W. Shields, "Speech enhancement using sub-band adaptive Griffiths-Jim signal processing," *Speech Commun.,* vol. 39, no. 1, pp. 97-110, 2003.

[3] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *J. Appl. Signal Process.,* vol. 2006, pp. 175-175, 2006.

[4] J. C. Rutledge, "A computational auditory scene analysis-enhanced beamforming approach for sound source separation," *J. Adv. Signal Process.,* vol. 2009, 2009.

[5] O. Roy and M. Vetterli, "Rate-constrained beamforming for collaborating hearing aids," *IEEE International Symposium on Information Theory,* pp. 2809-2813, 2006.

[6] S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Comparison of reduced-bandwidth MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Workshop Applications of Signal Processing to Audio and Acoustics,* pp. 223-226, 2007.

[7] S. Srinivasan and A. C. Den Brinker, "Rate-constrained beamforming in binaural hearing aids," *J.Adv. Signal Process.,* vol. 2009, no. 8, 2009.

[8] D. Ayllón, R. Gil-Pita and M. Rosa-Zurera, "Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids," *J. Adv. Signal Process.,* vol. 2013, no. 1, pp. 1-14, 2013.

[9] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* pp. 79-82, 2001.

[10] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. America,* vol. 99, no. 5, pp. 3138-3148, 1996.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of IEEE* , vol. 57, no. 8, pp. 1408-1418, 1969.

[12] H. Cox, R. Zeskind and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.,* vol. 35, pp. 1365-1376, 1987.

[13] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics* , vol. 7, no. 2, pp. 179-188, 1936.

[14] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," *DARPA Workshop on Speech Recognition,* pp. 93-99, 1986.

[15] V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, "The CIPIC HRTF database," *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* pp. 99-102, 2001.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Speech Audio Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, 2001.