INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving speech intelligibility in binaural hearing aids by estimating a time-frequency mask with a weighted least squares classifier

David Ayllón [1,2]*, Roberto Gil-Pita*[2]*, Manuel Rosa-Zurera*[2]

[1]R&D Department, Fonetic, Spain
[2]Signal Theory and Communications Department, University of Alcala, Spain
david.ayllon@fonetic.com, roberto.gil@uah.es, manuel.rosa@uah.es

Abstract

An efficient algorithm for speech enhancement in binaural hearing aids is proposed. The algorithm is based on the estimation of a time-frequency mask using supervised machine learning. The standard least-squares linear classifier is reformulated to optimize a metric related to speech/noise separation. The method is energy-efficient in two ways: the computational complexity is limited and the wireless data transmission optimized. The ability of the algorithm to enhance speech contaminated with different types of noise and low SNR has been evaluated. Objective measures of speech intelligibility and speech quality demonstrate that the algorithm increments both the hearing comfort and speech understanding of the user. These results are supported by subjective listening tests.
Index Terms: speech enhancement, machine learning, hearing aids

1. Introduction

Binaural hearing aids improve the ability to localize and understand speech in noise in comparison to monaural devices, but they require an increment in power consumption due to wireless data transmission. The power restriction in hearing aids also limits the computational cost of the embedded signal processing algorithms, so they should be designed to be both computationally and energy efficient [1].

Nowadays, there are two main approaches for binaural speech enhancement. One is binaural beamforming, which performs spatial filtering with the signals arriving at both devices. Some examples can be found in [2, 3, 4]. Unfortunately, the performance of these algorithms is notably affected when the bit rate is limited (e.g. lower than 16 kbps). Another drawback is that the beamforming output is directly affected by quantization noise.

The second approach is based on time-frequency (TF) masking. It has been demonstrated in [5, 6] that the application of the ideal binary mask (IBM) [7] to separate speech in noisy conditions entails an improvement in speech intelligibility. A recent approach to estimate the IBM from noisy speech is the use of supervised machine learning. Some examples are found in [8, 9, 10]. However, these methods are based on deep neural networks, which are computationally expensive to be implemented in hearing aids.

1.1. Previous work

In [11] the authors proposed a novel schema for speech enhancement in binaural hearing aids based on supervised machine learning. The algorithm is energy-efficient in two ways: the computational cost is limited and the data transmission optimized. The IBM is estimated with a speech/noise classifier. The proposed classification schema combines a simple least squares linear classifier (LSLC) with a novel set of features extracted from the spectrogram of the received signal. Features include information from neighbor TF points.

The work has been extended in [12], combining a fixed superdirective beamformer (BF) with TF masking. The fixed BF is able to reduce a high level of omnidirectional noise but it fails when rejecting directional noise. The directional noise that remains at the output of the BF is removed by the estimated TF mask, which is subsequently softened to reduce musical noise. In the proposed scenario, it is assumed that the target speaker is located in the straight ahead direction since, in a normal situation, the person is looking at the desired speaker. The target speech is contaminated by the addition of one or several directional sources and diffuse noise. The speaker wears two wireless-connected hearing aids, each one containing two microphones in endfire conguration, separated a distance of 0.7 cm. 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). The BF coefficients have been calculated to be robust against incoherent noise, according to [13].

The computational cost of the previous algorithm has been measured. Considering a state-of-the-art commercial hearing aid, it only requires a 28% of the total computational capabilities of the signal processor. The data transmission is optimized with a novel schema that optimizes the amount of bits used to quantify the signals exchanged between devices. The details about the transmission schema can be found in [12].

2. Least-squares linear classification

In this section we recall the standard formulation of the LSLC classifier and its application to estimate the IBM. The formulation of a weighted least squares problem is further included. These two descriptions will help to understand the proposal in section 3.

2.1. Least squares linear classifier

First it is important to highlight that a different classifier is designed for each frequency band k. Let us define the pattern matrix $\mathbf{Q}(k)$ of dimensions (PxL) containing P input features from a set of L patterns (time frames). The output of a linear classifier 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 (Lx1) column-vector that contains the output of the classifier and $\mathbf{v}(k) = [v(k,1),\ldots,v(k,P)]^T$ contains the weights applied to each of the P input features. For each of the patterns, the TF binary mask is generated according

http://dx.doi.org/10.21437/Interspeech.2017-771

to

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

where y_0 is a threshold value set to $y_0 = 0.5$. In the case of least squares (LS), 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 ordinary least squares (OLS) solution is obtained solving the next optimization problem:

$$\hat{\mathbf{v}}(k)^{LS} = \min_{\mathbf{v}(k)} \{ \left\| \mathbf{t}(k) - \mathbf{v}(k)^T \mathbf{Q}(k) \right\| \}, \quad (2)$$

and the OLS estimates of the model coefficients is given by

$$\hat{\mathbf{v}}(k)^{LS} = \mathbf{t}(k)\mathbf{Q}(k)^T \left(\mathbf{Q}(k)\mathbf{Q}(k)^T \right)^{-1}. \quad (3)$$

2.2. Weighted Least Squares

Let us consider now that the variances of the observations (features) are unequal and/or correlated. In this case, the OLS technique may be inefficient. The generalized least squares (GLS) method estimates the weights by minimizing the squared Mahalanobis length of the error [14]:

$$\hat{\mathbf{v}}(k)^{GLS} = \min_{\mathbf{v}(k)} \{ (\mathbf{t}(k) - \mathbf{v}(k)^T \mathbf{Q}(k))^T \boldsymbol{\Omega}(k)^{-1} (\mathbf{t}(k) - \mathbf{v}(k)^T \mathbf{Q}(k)) \}, \quad (4)$$

where the matrix $\boldsymbol{\Omega}(k)$ contains the conditional variance of the error term. In this case, the estimator of the weights has the next expression:

$$\hat{\mathbf{v}}(k)^{GLS} = \mathbf{t}(k)\boldsymbol{\Omega}(k)^{-1}\mathbf{Q}(k)^T \left(\mathbf{Q}(k)\boldsymbol{\Omega}^{-1}\mathbf{Q}(k)^T \right)^{-1}. \quad (5)$$

The weighted least squares (WLS) is a special case of GLS in which the matrix $\boldsymbol{\Omega}(k)$ is diagonal (off-diagonal entries are 0), and this occurs when the variances of the observations are unequal but where there are no correlations among them. In this case, the calculations can be simplified by defining a weighting term $\mathbf{w}(k) = [w(k,1), \cdots, w(k,L)]$ whose values are given by $w(k,l) = 1/\sqrt{\boldsymbol{\Omega}(k,l,l)}$ (diagonal terms). The weights estimates can be obtained as

$$\hat{\mathbf{v}}(k)^{WLS} = \mathbf{t}'(k)\mathbf{Q}'(k)^T \left(\mathbf{Q}'(k)\mathbf{Q}'(k)^T \right)^{-1}, \quad (6)$$

where $\mathbf{t}'(k) = [w(k,1)t(k,1), \cdots, w(k,L)t(k,L)]$ and $Q'(k,p,l) = w(k,l)Q(k,p,l)$.

3. Weighted LSLC for TF mask estimation

The success of the IBM improving speech intelligibility is thanks to its ability separating sound sources [7]. The W-Disjoint Orthogonality (WDO) factor proposed in [15] is a good indicator of the quality of the source separation achieved by a TF binary mask. This motivates the main proposal of this paper: the estimation of a TF mask that maximizes the WDO factor instead of minimizing the MSE with the IBM, as proposed in [11, 12]. In this section, first a new objective function called *Two-channel WDO factor* is defined. And second, the standard LSLC is reformulated to optimize this function.

3.1. Two-channel W-Disjoint Orthogonality (WDO) factor

Let us define the next signals in the STFT domain and filtered by the beamformer: $S_L^S(k,l)$ and $S_R^S(k,l)$ are the target speech signals at the left/right devices, $N_L^{dS}(k,l)$ and $N_R^{dS}(k,l)$ are the addition of directional noises at the left/right devices, $N_L^{oS}(k,l)$ and $N_R^{oS}(k,l)$ are the steered diffuse noise at the left/right devices. The superindex $()^S$ means steered signal.

In a two-channel problem, the IBM can be calculated according to

$$IBM(k,l) := \begin{cases} 1, & P_S(k,l) > P_N(k,l) \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where $P_S(k,l) = |S_L^S(k,l)|^2 + |S_R^S(k,l)|^2$ and $P_N = |N_L^{dS}(k,l) + N_L^{oS}(k,l)|^2 + |N_R^{dS}(k,l) + N_R^{oS}(k,l)|^2$. Considering the definition of the WDO factor in [15] the WDO associated to the separation of the target speech source from two channels can be expressed as

$$WDO = \frac{\sum_{(k,l)} M(k,l)(P_S(k,l) - P_N(k,l))}{\sum_{(k,l)} P_S(k,l)}, \quad (8)$$

where $M(k,l)$ is the applied TF mask. This expression can be rewritten as

$$WDO = \sum_{(k,l)} M(k,l)E(k,l), \quad (9)$$

where

$$E(k,l) = \frac{P_S(k,l) - P_N(k,l)}{\sum_{(k,l)} P_S(k,l)}. \quad (10)$$

Note that $E(k,l)$ is a constant value for a given mixture.

3.2. Weighted LSLC (WLSLC)

Let us focus now on the problem of finding the TF mask $M(k,l)$ that maximizes the WDO factor (i.e. source separation). Considering expression (9), the maximization problem is formulated according to

$$\max_{M(k,l)} \{ \sum_{(k,l)} M(k,l)E(k,l) \} \quad (11)$$

The value $E(k,l)$, defined in (10), can be decomposed in its modulus and sign, according to $E(k,l) = T(k,l)|E(k,l)|$, where $T(k,l)$ is the sign (+1, -1) and it is related to the target IBM defined in (7) through $T(k,l) = 2t(k,l) - 1$. Introducing this relationship into (11) yields

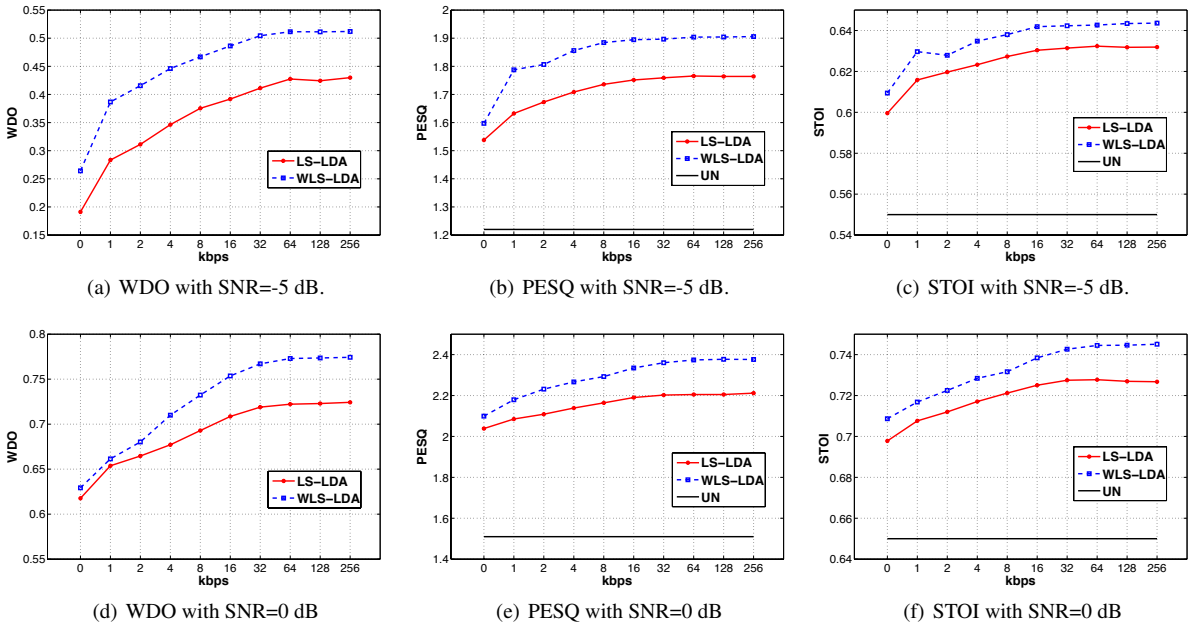
$$\max_{M(k,l)} \{ \sum_{(k,l)} M(k,l)(2t(k,l) - 1)|E(k,l)| \}. \quad (12)$$

Using the square values of $M(k,l)$ does not modify the values (i.e. '0' and '1'), which allows us to rewrite expression (12) as

$$\max_{M(k,l)} \{ \sum_{(k,l)} (2M(k,l)t(k,l) - M(k,l)^2)|E(k,l)| \}. \quad (13)$$

This maximization problem can be easily converted into the next minimization problem

$$\min_{M(k,l)} \{ \sum_{(k,l)} (M(k,l)^2 - 2M(k,l)t(k,l) + t(k,l)^2)|E(k,l)| \}, \quad (14)$$



(a) WDO with SNR=-5 dB. (b) PESQ with SNR=-5 dB. (c) STOI with SNR=-5 dB.

(d) WDO with SNR=0 dB (e) PESQ with SNR=0 dB (f) STOI with SNR=0 dB

Figure 1: *Two-channel WDO of speech, PESQ and STOI values, averaged over the test set, as a function of the transmission bit rate (kbps). The solid red line corresponds with the LS-LDA and the dashed blue line with the proposed WLS-LDA. The horizontal solid black lines represent the average values (PESQ or STOI) of the unprocessed signals (UN).*

where the addition of the term $t(k,l)^2$, which represents a constant value, allow us to rearrange the previous expression as

$$\min_{M(k,l)} \{ \sum_{(k,l)} (M(k,l) - t(k,l))^2 |E(k,l)| \}. \quad (15)$$

Since the values $M(k,l)$ are estimated from the output of the classifier $y(k,l)$, the previous expression is equivalent to expression (4). Hence, the maximization of the two-channel WDO is equivalent to the minimization of a weighted version of $MSE(k)$:

$$WMSE(k) = \frac{1}{L} \left\| (\mathbf{t}(k) - \mathbf{y}(k))\mathbf{w}(k) \right\|^2, \quad (16)$$

where the weighting terms are given by $\mathbf{w}(k) = \left[\sqrt{|E(k,1)|}, \ldots, \sqrt{|E(k,L|} \right]^T$. After computing $\mathbf{t}'(k)$ and $\mathbf{Q}'(k)$ as described in section 2.2, the weights of the WLSLC can be estimated using expression (6).

4. Objective evaluation

The comparison of the proposed WLSLC with the standard LSLC has been made with the same database used in [12]. It contains 3000 speech-in-noise binaural signals with three different types of mixtures: 1000 mixtures of speech with diffuse noise and two directional noise sources, 1000 mixtures of speech with two directional noise sources, and 1000 mixtures of speech with diffuse noise. The position of directional sources varies at random, and diffuse noise is simulated by generating isotropic speech-shaped noise. The speech signals are selected from the TIMIT database, and noise signals from a database that contains stationary and non-stationary noises. A 70% of the signals are used for training and the remaining 30% for testing. The data transmission has been limited to values that range from 0 to 256 kbps, and low SNRs of 0 dB and -5 dB have been used. The performance of the system is measured with the short-time objective intelligibility measure (STOI) [16], the two-channel WDO of the speech signal (8), and the PESQ score [17].

Figure 1 represents the two-channel WDO of speech (a, d), PESQ values (b-e) and STOI values (c-f), as a function of the transmission bit rate (kbps), for SNRs of -5 and 0 dBs. The solid red line corresponds with the LSLC and the dashed blue line with the proposed WLSLC. The horizontal solid black lines represent the PESQ and STOI values of the unprocessed signals (UN). All values are an average over the test set. The WDO values obtained by the WLSLC are notably higher than the ones obtained by the LSLC, particularly in the worst case (SNR=-5 dB). This is the expected behavior, since in the case of WLSLC the WDO is directly optimized.

Concerning speech quality (PESQ) and intelligibility (STOI), the scores obtained by the WLSLC are higher than the ones obtained by the LSLC in any case. The difference remains more or less constant with the transmission bit rate. In the worst case (SNR=-5 dB), the initial PESQ score (UN) of 1.22 is increased to a value of 1.9 applying the proposed TF mask (WLSLC estimation). In the case of SNR=0 dB, the initial PESQ score of 1.51 is increased up to 2.4 by the estimated TF mask. Regarding the STOI, in the case of SNR=-5 dB, the unprocessed STOI is 0.55, which is increased up to 0.64 by the proposed system. The initial STOI for SNR=0 dB is 0.65, and it is increased to 0.74. The previous values correspond with a transmission bit rate of 256 kbps. However, in all cases, the PESQ and STOI values are practically constant for bit rates down to 8 kbps. For lower transmission rates, the performance starts to decrease, but the improvement respect to the unprocessed signal is still noticeable in any case.

5. Intelligibility listening test

5.1. Description of the test

In order to validate the intelligibility of the proposed algorithm with real listeners, we have conducted listening tests processing speech signals from a different database that the one used to train the speech enhancement system. All the subjects that have participated in the experiments are native Spanish speakers, so we have used a database of speech signals in Spanish [18] (the use of sentences degraded with noise in a foreign language would be a disadvantage). The database consists of 300 sentences of 2 seconds each, grouped in six lists with equivalent predictability. The lists were also equivalent in length, phonetic content, syllabic structure and word stress. Only the first 200 sentences are used in our experiments (lists 1 to 4).

The 200 sentences were corrupted by a combination of isotropic white noise and two random directional noises (random noise and random position). The signals were mixed with -5 and 0 dB SNR. The unprocessed signals (denoted as 'UN') were processed by the proposed algorithm, generating two different binaural signals: the enhanced signals when the bit rate is limited to 16 kbps (denoted as 'TFM-16'), and the enhanced signals when the bit rate is limited to 256 kbps (denoted as 'TFM-256').

Twelve listeners were volunteer for the experiment. Half of the participants were male and the other half female, with ages ranged from 24 to 45 years (mean age of 30.6 years). All the participants were totally alien to the research conducted in this paper and none of them reported having any hearing or language problems. Six of the listeners participated in the experiment with a SNR of 0 dB and the other six with a SNR of -5 dB. Each of the subjects listened to a total of 200 sentences randomly selected from the three sets (UN, TFM-16 and TFM-256), selecting different combinations of sentences for each subject among the 200 available sentences of each condition. The experiments were performed in an isolated and quiet room and stimuli were played to the listeners binaurally through Sennheiser HD 202 stereo headphones at a comfortable listening level that was fixed throughout the tests for the different subjects. Before to start the test, each subject listened to a set of sentences from the different conditions to get familiar with the testing procedure. The order of the conditions was randomly selected across subjects. A GUI was developed for the tests. The subjects were asked to play each signal and type the words they understood. The software allowed the subjects to play each signal a single time. The intelligibility performance was evaluated by counting the number of words correctly identified. The duration of each test was approximately 40 minutes.

5.2. Results

The results of the listening test are summarized in figure 2. The graph represents the percentage of correct words in the three different conditions (UN, TFM-16 and TFM-256). The blue bars represent the values averaged over the six subjects in the case of -5 dB SNR, and the red bars represent the values averaged over the six subjects in the case of 0 dB SNR. The standard deviation is represented by a vertical black line over each bar. We can easily deduce a substantial improvement in intelligibility of the enhanced signals (TFM-16 and TFM-256) in comparison to that obtained from unprocessed speech (UN). In the case of 0 dB SNR, the initial 30% points (UN) are increased to a 73% with the 16 kbps mask, and to 81% with the 256 kbps mask. According to this, the designed system is able to increase the intelligibility from 30% to 81%, which is equivalent to an improvement factor of 2.7.

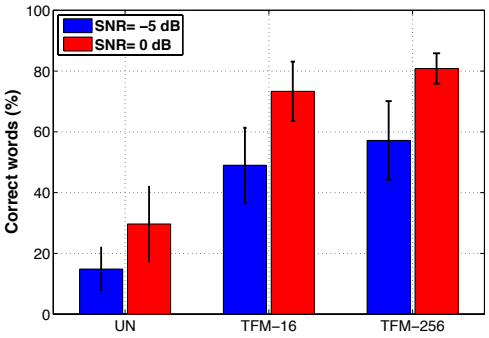


Figure 2: *Percentage of correct words in the three different conditions of the listening test.*

In the case of -5 dB SNR, the high level of noise causes that the initial intelligibility is very low (less than 15 % of the words are correctly identified). Nevertheless, the intelligibly obtained by the use of the 16 kbps mask increases the intelligibility to 49%, and the use of the 256 kbps mask increases the intelligibility to 57%. In this case, although the maximum output intelligibility of the system is not very high (57%), the increment respect to the original intelligibility (15%) is higher than in the case of 0 dB SNR, being equivalent to an improvement factor of 3.8.

6. Conclusions

This work presents a novel algorithm to estimate the TF mask for speech enhancement in binaural hearing aids. The paper introduces an update of a previous work presented by the authors. The experimental work has shown that the proposed method outperforms the results obtained by the previous algorithm in terms of speech quality and intelligibility.

The proposed solution has demonstrated to introduce important improvements in speech speech intelligibility (STOI) and speech quality (PESQ). In addition, these results are supported by subjective results obtained with a listening test. For instance, in the case of SNR= 0 dB, the percentage of correct words identified in the test is increased by a factor of 2.7, and in the case of -5 dB, by a factor of 3.8. These values represent a very important improvement in intelligibility for hearing aids users. Additionally, the performance of the system is practically unaltered with transmission bit rates that goes from 256 kbps down to 8 kbps, although the performance obtained with lower bit rates is also remarkable. This allows the reduction of the power required for data transmission and, together with the low computational cost of the enhancement algorithm, make the proposal efficient.

In summary, the proposed algorithm represents an affordable solution for speech enhancement in binaural hearing aids, being able to increment both the hearing comfort and speech understanding of the hearing impaired user.

7. Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitiveness, under project TEC2015-67387-C44-R.

8. References

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

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

[3] 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.

[4] S. Srinivasan and A. C. Den Brinker, "Rate-constrained beamforming in binaural hearing aids," *EURASIP Journal on Advances in Signal Processing* vol. 2009, no. 8, 2009.

[5] Y. Li and D.L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230-239, 2009.

[6] P.C. Loizou and G. Kim, "Reasons why current speech enhancement algorithms do not improve speech intelligibility and suggested solutions, " *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47-56, 2011.

[7] G. Hu and D.L. Wang. "Monaural speech segregation based on pitch tracking and amplitude modulation, " *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135-1150, 2004.

[8] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112-2121, 2014.

[9] Y. Xu, J. Du, L. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks", *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65-68, 2014

[10] Y. Zhao, D. Wang, I. Merks, T. Zhang, "DNN-based enhancement of noisy and reverberant speech", *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6525-6529, 2016

[11] D. Ayllón, R. Gil-Pita and M. Rosa-Zurera, "Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids," *EURASIP Journal on Advances in Signal Processing* vol. 2013, no. 1, pp. 1-14, 2013.

[12] D. Ayllón, R. Gil-Pita and M. Rosa-Zurera, "A machine learning approach for computationally and energy efficient speech enhancement in binaural hearing aids," *IEEE International Conference on Acoustics, Speech and Signal Processing*, no. 1, pp. 6515-6519, 2016.

[13] H. Cox, R. Zeskind and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing* vol. 35, pp. 1365-1376, 1987.

[14] T. Kariya and H. Kurata, *Generalized Least Squares.* Wiley, 2004.

[15] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing,* vol. 52, no. 7, pp. 1830-1847, 2004.

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

[17] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", *Recommendation P ITU-T.862*, 2001.

[18] T. Cervera and J. Gonzlez-Alvarez, "Test of Spanish sentences to measure speech intelligibility in noise conditions," *Behavior Research Methods* vol. 43, no. 2, pp. 459-467, 2001.