A COMPUTATIONALLY-EFFICIENT SINGLE-CHANNEL SPEECH ENHANCEMENT ALGORITHM FOR MONAURAL HEARING AIDS

David Ayllón, Roberto Gil-Pita, Manuel Utrilla-Manso, Manuel Rosa-Zurera

Department of Signal Theory and Communications, University of Alcalá, Spain

ABSTRACT

A computationally-efficient single-channel speech enhancement algorithm to improve intelligibility in monaural hearing aids is presented in this paper. The algorithm combines a novel set of features with a simple supervised machine learning technique to estimate the frequency-domain Wiener filter for noise reduction, using extremely low computational resources. Results show a noticeable intelligibility improvement in terms of PESQ score and SNR_{ESI}, even for low input SNR, using only a 7% of the computational resources available in a state-of-the-art commercial hearing aid. The performance of the algorithm is comparable to the performance of current algorithms that use more computationally complex features and learning schemas.

Index Terms— Speech enhancement, Noise reduction, Time-frequency masking, Supervised learning.

1. INTRODUCTION

Speech enhancement in monaural hearing aids is an open and complex problem mainly due to two reasons. First, the improvement of speech intelligibility rather than speech quality is primordial for hearing-impaired people. Second, the computational resources and memory available in the digital signal processor (DSP) embedded in such devices is very low.

Traditional methods for single-channel noise reduction based on spectral subtraction [1, 2], the Wiener filter [3, 4], or the minimum mean-square error (MMSE) estimator [5, 6] have demonstrated their ability in reducing background noise, but they are not capable of improving speech intelligibility [7, 8]. The main reason is that these algorithms have been designed to improve speech quality, which can be easily improved by increasing the signal-to-noise ratio (SNR), rather than to improve speech intelligibility, which is only improved by suppressing the background noise without distorting the target speech signal. However, many traditional algorithms introduce speech distortions, usually as an annoying 'musical noise'.

Originated in the field of computational auditory scene analysis (CASA), the time-frequency (T-F) masking approach has grown in importance in recent years, due to its ability to improve the intelligibility of speech in noise. It is demonstrated in [8] that the ideal binary mask (IBM) defined in CASA [9] maximizes the articulation index (AI), a metric highly correlated with speech intelligibility. Unfortunately, the computation of the IBM needs to have access to the clean speech and noise signals and, in practice, it should be estimated from the corrupted speech signal. The CASA approach performs this estimation using features inspired in the human auditory system (pitch, amplitude and frequency modulation, onset/offset, etc.). However, it is conceptually and computationally simpler to use machine learning techniques to identify each T-F point as speech-dominated or noise-dominated.

Some prior works that use supervised learning to estimate a T-F mask are described in the following review. In [10], the IBM is estimated using an accurate binary Bayesian classifier that uses amplitude modulation spectrograms (AMS) as input features and trains Gaussian mixture models (GMM) to represent the distribution of each class. In [11], the previous classification schema is used to estimate a different binary mask based on magnitude spectrum constraints. The same approach has also been applied to estimate soft masks, which have proved to improve intelligibility better than binary masks. In [12], a soft mask is generated by estimating the local SNR using AMS as features and a multi-layer perceptron (MLP) as estimator. In [13], a smoothed ideal ratio mask (IRM) is estimated using deep neural networks (DNN) and features calculated in the Mel spectral domain.

In this paper, a novel machine learning algorithm to estimate a T-F soft mask is proposed. The main novelty of the algorithm resides in the proposed set of features and its extremely low computational cost, which makes its implementation in a commercial hearing aid easier than previous works in this field, which have used more computationally complex features (MFCCs, AMS, etc.) and learning schemas (GMM, MLP, DNN, etc.). In this work, both the proposed features and the estimator require low computational resources.

2. COMPUTATIONAL RESOURCES AVAILABLE FOR SIGNAL PROCESSING IN HEARING AIDS

In this section, the computational resources available for speech enhancement in a state-of-the-art commercial hearing

This work has been funded by the Spanish Ministry of Science and Innovation, under project TEC2012-38142-C04-02 and the scholarship AP2009-3932.

aid are quantitatively measured in terms of instructions to process each frequency band (IPF). The T-F analysis is based on a discrete Fourier transform (DFT) filterbank and usually implemented in a specific processor, hence it does not imply any consumption of computational resources from the main processor. Common DSPs embedded in hearing aids have a processor with a selective clock speed that usually goes from 1.28 MHz to 5.12 MHz. They have a Harvard architecture containing a multiplier-accumulator (MAC) with a set of instructions completed in a clock cycle. Then, the number of mega instructions per second (MIPS) is the clock speed value. The sampling rate (f_s) is usually adjustable (normally $f_s \leq 16$ kHz). Considering that the analysis and synthesis windows have a length of L_{WIN} samples working with 50% of overlap, and that the DFT-based frequency analysis contains K frequency bands, the IPF for each frame is calculated using the next expression:

$$IPF = \frac{MIPS}{K} \cdot \frac{L_{WIN}/2}{f_s}. \quad (1)$$

In the special case of a processor with a clock speed of 5.12 MHz (5 MIPS), f_s= 8 kHz, L_{WIN}= 64 samples, and $K = 32$, the IPF is 625. This IPF value will be considered as reference value in the remaining part of this paper. The computational resources are shared between the own speech enhancement algorithm, the multi-band compression-expansion algorithm (which is an indispensable algorithm), and other algorithms dedicated to feedback cancellation or automatic sound classification. Hence, the speech enhancement algorithm proposed in this paper will use only a part of the available IPF calculated in this section.

3. PROPOSED ALGORITHM

Let us consider $X(k,l) = S(k,l) + N(k,l)$ to be the short-time Fourier transform (STFT) of a speech signal $S(k,l)$ contaminated by noise $N(k,l)$, where k denotes frequency and l the time frame. The frequency-domain Wiener filter, which is inspired by the expression of the non-causal Wiener filter [3], is given by

$$M(k,l) := \frac{|S(k,l)|^2}{|S(k,l)|^2 + |N(k,l)|^2}. \quad (2)$$

The main goal in this work is the design of a computationally-efficient algorithm to estimate the Wiener mask from the mixture signal $X(k,l)$. The proposed solution is based on supervised learning, using a low-cost linear estimator whose weights are calculated during the training stage. The key point of the algorithm is the novel set of features proposed for estimation. The features require a small number of instructions to be calculated and allow the linear estimator to obtain low estimation errors.

Figure 1 shows a block diagram of the proposed enhancement algorithm. The diagram has been divided into two parts,

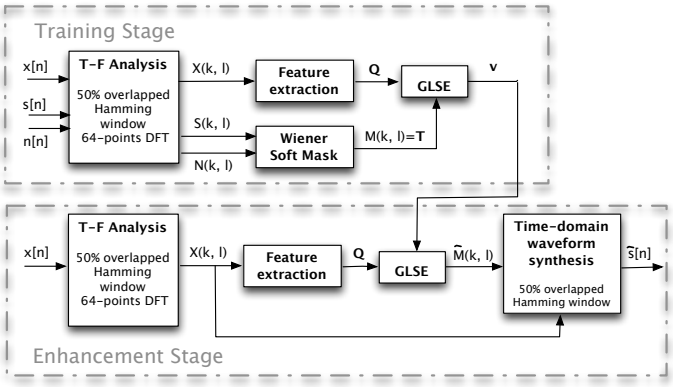


Fig. 1: Algorithm overview.

the training stage (top) and the enhancement stage (bottom). The first block (left) in the two stages represents the T-F analysis, which is performed by computing a 64-points DFT for each time frame, using a Hamming window with an overlap of 50%. In the training stage (top) the Wiener soft mask is calculated from the clean speech and noise signals, and it is used as target to train the estimator. The proposed estimator uses a set of features extracted from the STFT of the mixture to estimate the target mask. The weights calculated during the training stage are used during the enhancement stage to estimate the mask from the input noisy signal and to generate the enhanced speech signal. The different parts of the algorithm are explained in detail in the remaining of this section.

3.1. Generalized least squares estimator (GLSE)

Least squares estimation (LSE) is an approach that fits a parametrized mathematical model to the observed data by minimizing the mean square error (MSE) between the observed data and their expected values. In the case that the model combines linearly the unknown parameters, the method is known as linear least squares.

Let us define the pattern vector $\mathbf{x_i} = [1, x_{i1}, \ldots, x_{iP}]^T$, where $x_{i1}, \ldots, x_{iP}$ are P input features (i.e. the observations of the model). The pattern matrix $\mathbf{Q} = [\mathbf{x_1}, ..., \mathbf{x_L}]$ contains the patterns $\mathbf{x_i}$ of a set of L data samples. The output of the LSE is obtained as a weighted linear combination of the input features, according to $\mathbf{y} = \mathbf{v}^T\mathbf{Q}$, where the vector $\mathbf{v} = [v_0, v_1, v_2, \ldots, v_P]^T$ contains the bias v_0 and the weights applied to each of the P input features and $\mathbf{y}$ contains the output for the L input patterns. The vector $\mathbf{v}$ is calculated to minimize the error between the obtained output and its desired value. In the case of supervised learning, the desired output values are available and used for training. The MSE of the estimator is given by

$$MSE = \frac{1}{L}\|\mathbf{y} - \mathbf{t}\|^2 = \frac{1}{L}\left\|\mathbf{v}^T\mathbf{Q} - \mathbf{t}\right\|^2, \quad (3)$$

where $\mathbf{t} = [t_1, t_2, \cdots, t_L]^T$ is the target vector containing the desired output values for the L input patterns. The MSE is minimized by differentiating expression (3) with respect to

every weight in $\mathbf{v}$ and setting the result equal to zero, which yields the next expression

$$\mathbf{v} = \mathbf{t}\mathbf{Q}^T \left(\mathbf{Q}\mathbf{Q}^T\right)^{-1}. \quad (4)$$

In order to improve the performance of the linear LSE, it is proposed to introduce non-linear transformations of the input features, which are still linearly combined, unlike the non-linear least squares approach. The matrix $\mathbf{Q}$ is now defined as $\mathbf{Q} = [f_1(\mathbf{x_1}), \ldots, f_1(\mathbf{x_L}), \ldots, f_{N_T}(\mathbf{x_1}), \ldots, f_{N_T}(\mathbf{x_L})]$, where $f_1, \ldots, f_{N_T}$ are N_T linear or non-linear transformations performed over the original input features $\mathbf{x_i}$. The weight vector is then defined as $\mathbf{v} = [v_0, v_1, \ldots, v_{N_T \cdot P}]^T$, and it can still be obtained using expression (4). Henceforth, this is denoted generalized least squares estimator (GLSE).

In the problem at hand, a different GLSE is used to estimate the mask of each frequency band. The P input features are extracted from the mixture $X(k,l)$ for each of the L time frames (i.e. the L input patterns). The output of the GLSE for the k-th frequency band is $\mathbf{y}_k = [y(k,1), ..., y(k,l), ..., y(k,L)]^T$, and it is the estimation of the T-F mask (i.e. $\hat{M}(k,l) = y(k,l)$).

3.2. Proposed features for estimation

Let us assume that the output of the DFT-based analysis filterbank is $|X(k,l)|^2$. This information can be used as input feature by the estimator but, according to the aforementioned GLSE, further transformations of this feature can be also included. Specifically, the logarithm and the square logarithm, $\log(|X(k,l)|^2)$ and $\log^2(|X(k,l)|^2)$, have been experimentally found to provide the most meaningful information to the GLSE and they are included as input features.

Additionally, it is proposed to use the information of neighbor T-F points as input features. The logarithm and the square logarithm of N_F adjacent neighbor frequency bands of the current time frame (N_F upper and N_F lower bands) are also included as input features. The information regarding previous time frames is also included, but in a special way. An exponentially-weighted moving average (EWMA) of the logarithm and the square logarithm of the previous time frames is calculated for each frequency band, according to:

$$A(k,l) = (1-2^{-\alpha})A(k,l-1) + 2^{-\alpha} f(k,l), \quad \alpha \in \mathbb{Z}^+, \quad (5)$$

where $A(k,l)$ is the EWMA for the k-th frequency band in the l-th time frame, $f(k,l)$ represents the input value ($\log(|X(k,l)|^2)$ or $\log^2(|X(k,l)|^2)$), and α is a smoothing factor that controls the degree of weighting decrease. A lower value of α discounts older observations faster. The EWMA is calculated with $(D-1)$ different values of α, having $(D-1)$ different EWMA for each frequency band, which are included as input features. From the computational point of view, the use of exponential values ($2^{-\alpha}$) as filter coefficients is equivalent to shift a value α bits in memory, which reduces the computational cost associated to the computation of the EWMA.

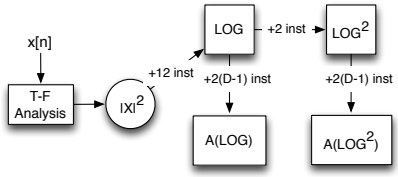


Fig. 2: Number of instructions associated to the computation of the proposed features.

In summary, according to the proposed feature schema, each T-F point has a total of $P = 2N_F + D$ input features.

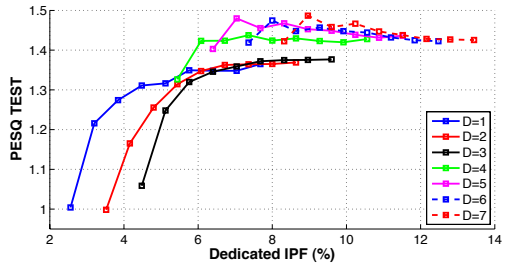
3.3. Computational cost of the proposed algorithm

In the enhancement stage, the estimation of the T-F mask only involves the operation $\mathbf{v}^T\mathbf{Q}$, using the fixed weights previously calculated in the training stage. The implementation of the proposed estimator is relatively simple, its computational cost being directly related to the number of features. Considering that the MAC operation is executed in a single instruction, the number of instructions required by the estimator can be reduced to $2P$, P being the number of input features included in $\mathbf{Q}$ (i.e. $P = 2N_F + D$). Assuming that the output of the T-F analysis filterbank is $|X(k,l)|^2$, and according to the standard assembler language used in this type of DSPs, the number of instructions required for the computation of the input features is $14+4(D-1)$, as shown in figure 2. According to this, the number of instructions necessary to process each frequency band is $IPF = 4N_F + 6D + 10$. Consequently, the values N_F and D should be selected to find a tradeoff between speech enhancement and computational cost.

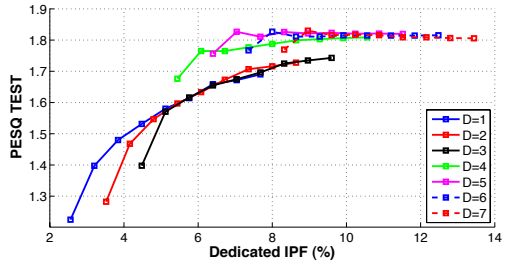
4. EXPERIMENTAL WORK AND RESULTS

4.1. Database setup

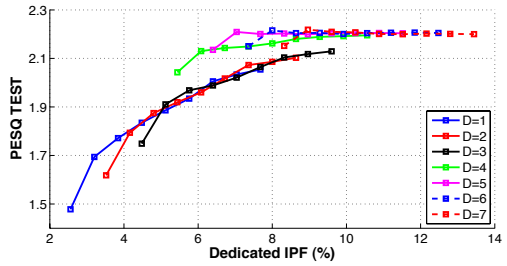
Speech and noise mixtures of different SNRs are generated, using the speech and noise signals contained in the NOIZEUS database described in [7]. The database contains 30 sentences produced by three male and three female speakers, corrupted by 8 different real-world noises at different SNRs, including suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The signals are sampled at 8 kHz. The 30 clean speech signals are normalized and linked together, one after the other, obtaining a speech segment of 80 seconds length. The 30 noise signals of each type of noise are also normalized and linked together, obtaining 8 different segments of 80 seconds length. The clean and noise signals are split into two different parts, one for training and another for testing. The training set consists of the 60% of the signals (56 seconds) and the test set consists of the 40% of the signals (24 seconds). Then, the training and test clean



(a) SNR = -5 dB



(b) SNR = 0 dB



(c) SNR = 5 dB

Fig. 3: PESQ obtained by the proposed algorithm in the test set, as a function of the percentage computational cost. D values are represented with lines of different colors, and N_F values are represented with squares over the lines (N_F increases from left to right).

speech segments are repeated 8 times, generating a signal of 448 seconds length in the case of training and 192 seconds length in the case of test. The 8 different noise segments are linked together, for the training and test sets separately. Finally, the clean and noise signals of both sets are normalized and mixed at the desired SNR.

4.2. Results

The computational cost of the proposed estimator directly depends on the values N_F and D. In order to find a tradeoff between speech enhancement and computational cost, the estimator is trained with different values of N_F and D. The value N_F has been varied from 0 to 8 and the value of D from 1 to 7, in both cases with steps of 1. Note that the specific case of $N_F = 0$ and $D = 1$ corresponds to the case of considering only the information of the current T-F point for estimation. The enhancement obtained over the test set is evaluated using an objective measure of speech quality and intelligibility. This measure is the PESQ score proposed in [14], which was first designed to evaluate the speech quality, but several works have reported high correlation between PESQ score and subjective listening tests [15, 16].

Table 1: PESQ scores corresponding to the unprocessed mixture (UN), the ideal Wiener mask (IWM), and the proposed algorithm in the test set (TEST), and SNR_{ESI} (dB) obtained by the proposed algorithm in the test set.

SNR	$PESQ$			$SNR_{ESI}(dB)$
	UN	IWM	TEST	
- 5 dB	1.32	2.46	1.48	5.41
0 dB	1.58	2.73	1.83	4.40
5 dB	1.88	3.02	2.21	3.41

Figure 3 represents the PESQ scores obtained by the proposed algorithm in the test set, as a function of the computational cost expressed in percentage of the total number of available IPF. The different values of D are represented with lines of different colors, and the different values of N_F are represented with squares over the lines (N_F increases from left to right). The SNR is - 5 dB in (a), 0 dB in (b), and 5 dB in (c). Analyzing the three graphs we find that a value of $D = 5$ and $N_F = 1$ represents a good tradeoff between the PESQ score and computational cost. This option represents only a 7% of the available IPF and the increment of D or N_F barely improves the PESQ score. Additionally, the importance of using the information of neighbor T-F points for the proposed estimator is clearly demonstrated.

Table 1 contains the PESQ score obtained by the unprocessed mixture (UN), the ideal Wiener mask (IWM), and the proposed algorithm in the test set with $D = 5$ and $N_F = 1$ (TEST), for the different SNRs. Although the PESQ scores obtained in the test set are still far from the ones obtained by the ideal (but unrealizable) Wiener mask, they represent an important increment in the PESQ score in comparison to the unprocessed mixture, and what is probably more important, they have been achieved using extremely low computational resources (7%). Additionally, the signal-to-residual spectrum measure SNR_{ESI} proposed in [8], whose correlation with speech intelligibility was found to be 0.81 [16], has been calculated for the output of the proposed algorithm in the test set (right column). The SNR_{ESI} is calculated to provide more meaning to the PESQ increments. In the worst case (SNR=-5 dB), a PESQ increment of 0.16 is obtained, which corresponds with a SNR_{ESI} of 5.41 dB, which is a good value.

Finally, the proposed algorithm is compared with the eMBM algorithm described in [11], in terms of PESQ score. The eMBM algorithm defines an ideal magnitude-constraints binary mask (IMBM) and trains two GMMs using AMSs features to estimate the IMBM with a two-class Bayesian classifier. The computational cost of this algorithm is clearly higher that the one of the algorithm proposed in this paper. Table 2 contains the PESQ increments achieved by the two algorithms for two types of noises (the ones evaluated

Table 2: Increments in the PESQ scores achieved by the eMBM algorithm and the proposed algorithm (TEST).

SNR	Airport babble		20-talker babble	
	eMBM	TEST	eMBM	TEST
- 5 dB	0.18	0.21	0.13	0.18
0 dB	0.32	0.28	0.14	0.24

in [11]), averaged over different time segments. In general, the proposed algorithm obtains slightly higher PESQ increments than the eMBM algorithm, and most importantly, this performance is obtained using very low computational resources.

5. CONCLUSIONS

In this paper, a computationally-efficient speech enhancement algorithm for monaural hearing aids has been proposed. The algorithm uses supervised learning to estimate a T-F soft mask based on the Wiener filter. Contrary to similar algorithms found in the literature, the computational complexity of the proposed estimator and the extraction of the input features is extremely low. The proposed solution only requires a 7% of the available computational resources in a state-of-the-art hearing aid and the obtained results support the ability of the algorithm to enhance noisy speech in terms of PESQ score and SNR_{ESI}. Additionally, it is noticeable the importance of using the information of neighbor T-F points to estimate the mask, especially the proposed EWMA related to previous time frames. The proposed estimator clearly fails without the information of the previous time frames.

To conclude, the proposed algorithm represents a computationally feasible solution for speech enhancement in commercial hearing aids. Although the optimal Wiener mask may be better estimated using more complex set of features and learning schemas, they are not probably realizable with such low computational resources.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[2] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, vol. 4, pp. 208-211.

[3] N. Wiener, *Smoothing of stationary time series*, Wiley, New York, US, 1949.

[4] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech", in *Proceedings of the IEEE*, vol. 67, no. 12, 1586-1604, 1979.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[6] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345-349, 1994.

[7] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Communications*, vol. 49, no. 7, pp. 588-601, 2007.

[8] P. 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.

[9] D. L. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press / Wiley-Interscience, 2006.

[10] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners", *The Journal of the Acoustical Society of America*, vol. 126, pp. 1486, 2009.

[11] G. Kim and P. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints", *IEEE Signal Processing Letters*, vol. 17, no. 12, pp. 1010-1013, 2010.

[12] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression", *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 11, no. 3, pp. 184-192, 2003.

[13] A. Narayanan and D. L. Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7092-7096.

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

[15] Y. Hu and P.C. Loizou. "Evaluation of objective quality measures for speech enhancement". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 229-238, 2008.

[16] J.Ma, Y. Hu and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387-3405, 2009.