AN IMPROVED THRESHOLDING FUNCTION AND SPARSE SUBSPACE DECOMPOSITION FOR SPEECH ENHANCEMENT AND ITS APPLICATION TO SPEECH RECOGNITION

Mohamed anouar Ben messaoud[1], Aïcha Bouzid[1]

[1]National Engineering School of Tunis, University of Tunis El Manar
anouar.benmessaoud@yahoo.fr

Kurzfassung: In this work, we propose an unsupervised monaural Arabic speech enhancement method that is based on two different techniques. The main idea is to determine an exact threshold value in the wavelet domain depending on the voicing state of the Arabic speech signal. Our proposed voiced/unvoiced decision algorithm based on the Multi-scale Product (MP) analysis is used. The MP is based on the multiplication of wavelet transform coefficients at three successive dyadic scales. Then, we apply a denoising technique based on the thresholding of the discrete wavelet transform coefficients. The threshold values change either when the signal is voiced or unvoiced. Further, a subspace decomposition-based post-processing technique is implemented. The Fast Fourier Transform (FFT) of the obtained frames is decomposed into three subspaces: sparse, low rank, and the remainder noise components. Experimental results show that the proposed approach outperforms the compared speech enhancement methods for noise-corrupted Arabic speech at low levels of SNR. Beside, we present the evaluation results for automatic recognition on enhanced Arabic speech signal. We reconstitute the clean Arabic speech from noisy observations based on a sparse imputation technique. It employs a non-parametric model and finding the sparsest combination of exemplars that jointly approximate the reliable features of a noisy Arabic utterance.

1 Introduction

Speech enhancement is a one of the most salient feature used in real-world applications, such as voice communication, hands-free accessories, voice-activated diallers for cellular phones, communicators in noisy cockpits, and automatic speech recognition (ASR) systems [1, 2]. Noise has a negative effect on both speech intelligibility and quality; a poor signal-to-noise ratio (SNR) may certainly result in a complete deficiency of speech intelligibility.

In order to reduce undesirable background noises corrupting Arabic speech and to improve the quality and the intelligibility of the speech in the presence of ambient noises, various algorithms have been adopted. Generally the algorithms can be ranged into two principal categories of single-channel and multi-channel speech enhancement techniques. Although multi-channel algorithms have significant efficiency in some applications, there are still many practical situations in which only single acquisition channel is applicable.

Fundamentally, there are four main types of approaches proposed in the field of single-channel speech enhancement: spectral subtraction, statistical model-based approaches, subspace and wavelet transform approaches.

The spectral subtraction is a single channel noise reduction from speech signals employs in the frequency domain by subtracting the noise spectral amplitude estimation from the noisy speech spectrum. This approach is effective in the case of stationary noise where the speech and the noise are independent [3, 4]. But it distorts the speech signal and introduces additional annoying residual noise in real environments.

In the statistical model-based approaches, the large variance of the spectrum coefficients is analysed as a Bayesian estimation problem [5]. Based on Gaussian statistics and a priori SNR estimation, Ephraim and Malah used the mean-square error (MMSE) to estimate the short-time spectral amplitude (STSA) [6], and the log spectral amplitude (LSA) [7]. Recently, Lu and Loizou proposed a method based on MMSE and maximum a posteriori (MAP) estimators derived using a Gaussian statistical model [8]. The main drawback of the statistical model is that it doesn't give a procedure to control the trade-off between the residual noise and the speech distortion.

Typically, a subspace approach for speech enhancement is used in the time domain when the speech estimation is formulated as a constrained optimization problem, where the speech distortions are minimized subject to the residual noise power level [9]. In [10], the authors propose a normalized least mean square (NLMS) to reduce the signal distortion. However, these adaptative methods are sensitive to the spectral flatness of the reference input.

In the last two decades, wavelet transforms (WT) have been applied on various research areas. The basic principle of speech denoising in wavelets family is based on the thresholding of the discrete wavelet transform coefficients (DWTC) to separate the target speech signal from those of noise. For noisy signal, applying a given threshold [11, 12, and 13] for all the DWTC irrespective of voiced or unvoiced speech regions may reduce undesirable background noises, but it can remove some unvoiced speech ranges with the additional noise. Thus has a negative effect on both speech quality and intelligibility. To solve this problem, the thresholding must be adapted over time. Therefore, some papers have proposed to design the threshold technique to ameliorate the quality of wavelet denoising approach and to remove the deficiency due to over thresholding of the speech signals by a simple time-invariant threshold [14, 15, 16, 17, and 18].

From this spectrum, we propose a wavelet thresholding approach for speech enhancement. Unlike traditional wavelet approaches, the threshold is adapted based on our voiced/unvoiced decision algorithm. The proposed approach is tested on noisy speech under various noise conditions including white noise, car noise, and babble noise. It is shown through experiment results that the use of the proposed approach can enhance the quality of noisy speech signal.

The present paper is organized as follows. Section 2 describes the block diagram of our proposed approach for speech enhancement. The subsection concerns the details of all the steps constituting our approach. Section 3 concerns the experimental results. Finally, section 4 concludes this paper and proposes the future work.

2 Proposed Approach

This section describes our proposed approach for Arabic speech enhancement. First, the signal is divided into 64 ms windows with 50% overlap between frames. Then, we calculate the discrete wavelet transform (DWT) of the noisy Arabic speech at four scales (3, 4, 5, and 6) using Daubechies wavelet. A thresholding process is applied on wavelet transform coefficients in each frame depending on the voicing nature of the frame. In fact, for voiced frames, we use a given threshold and for unvoiced frames, another threshold is fixed. For each scale, we apply the thresholding process. As a post-treatment, we apply an improved subspace decomposition technique.

The various stages are illustrated in Figure 1.

In the followed subsection, the voiced/unvoiced (V/UV) decision method will be closely analyzed.

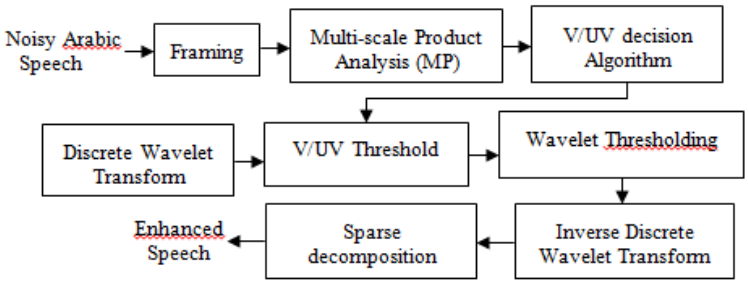

Figure 1 – Block Diagram for our Proposed Approach for Arabic Speech Enhancement

2.1 Voiced/Unvoiced Algorithm

The voiced/unvoiced algorithm is based essentially on the multi-scale product (MP) characteristics. In fact, the MP permits to suppress false peaks caused by noise, obtain a more simplified signal quasi-null signal in unvoiced frames and having a periodic structure in the voiced frames.

The proposed voiced/unvoiced algorithm uses the group classification of the multi-scale product in the frequency domain to classify the speech frames in voiced and unvoiced regions.

The system block diagram of our voiced/unvoiced classification algorithm is shown in the figure 2.

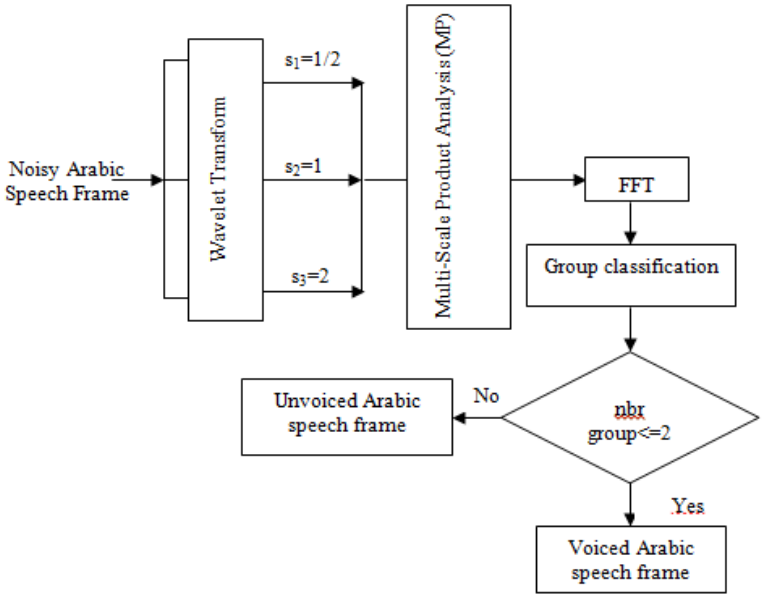

Figure 2 – Block Diagram for our Proposed Algorithm for Voiced/Unvoiced Decision

According to Mallat, the WT, has shown excellent capacities for the detection of signal singularities. When the wavelet function has specific selected properties, WT acts as a differential operator. The number of wavelet vanishing moments gives the order of the differentiation. This property of dyadic WT has been proven very useful for detecting pitch periods of speech signals [19].

The multi-scale product analysis consists of making the product of the wavelet transform coefficients of the speech at 3 dyadic scales. The wavelet used is the quadratic spline function at scales $s_1=2^{-1}$, $s_2=2^0$ and $s_3=2^1$.

The product $p(n)$ of wavelet transforms coefficients of the Arabic speech frame $x(n)$ at some successive dyadic scales is given as follows:

$$p(n)=\prod_{j=-1}^{j=1} W_{2^j}x(n) \quad (1)$$

Where $W_{2^j}x(n)$ is the wavelet transform of the Arabic speech frame $x(n)$ at scale 2^j.

A voiced frame of the FFT of MP gives a number of groups equal to 1 or to 2, whereas an unvoiced frame of the FFT of MP gives a number groups greater than 2.

The parameter for the determination of the V/UV decisions is the group classification. We apply the fast Fourier transform (FFT) function to the spectrum of the signal multi-scale product. We define the number of groups constituted by computing the distance separating 2 successive peak positions of the FFT of multi-scale product. Then, we rank this distance to compose a vector. The elements belong to the same group when the distance between two successive elements of the vector is less or equal to 10. In the case where the number of groups equal to 1 or to 2, the frame is declared as voiced, if not, the frame is considered un-voiced.

2.2 Wavelet Transform Thresholding

The proposed wavelet transform denoising technique is based on a modified wavelet thresholding (MWT). The DWT is applied on each frame of the input noisy Arabic speech, and the thresholding is applied on wavelet transform coefficients. However, we use a given threshold for voiced frames and another for unvoiced frames.

We apply a soft thresholding that is given by the following equation:

$$Th\_soft=\begin{cases} sgn(nc)(|nc|-th) & \text{for } |nc|\geq th \\ 0 & \text{for } |nc|<th \end{cases} \quad (2)$$

Where nc is the noisy wavelet coefficient and th is the threshold value proposed in [11]. This threshold is:

$$Th=\sigma\sqrt{2\log(N)} \quad (3)$$

Where σ and N are respectively the standard deviation of zero mean additive white Gaussian noise and the length of the noisy speech frame.

The standard deviation of noise must be estimated in order to determine the threshold value. The basic denoising method using WT considers that noise spectrum is white. Therefore we can estimate the standard deviation in [11]:

$$\sigma=\left(1/0.6745\right)\text{Median}(|c|) \quad (4)$$

Where c is the coefficient sequence of the noise WT.

As the real noise is colored, we use a level dependent threshold suggested in [20]:

$$Th_i=\sigma_i\sqrt{2\log(N_i)} \quad (5)$$

Where N_i is the number of the samples, and σ_i represents the noise variance estimated on the scale i given by Donoho:

$$\sigma_i = \left(1 / 0.6745\right) \text{Median}\left(|c_i|\right) (6)$$

Where c_i is the set of coefficients of the i^{th} wavelet band of noise.

If an analysis frame is classified as unvoiced, the threshold value used by equation (3) for white and by equation (5) for colored noise is multiplied respectively by a constant α and for a frame classified as voiced the threshold value is multiplied by the constant β.

2.3 Improved Subspace Decomposition

In the noisy Arabic speech frames, we decompose the noisy Arabic speech data matrix X into three subspaces:

$$X = S + Re + Lo (7)$$

Where S, Lo and Re represent sparse matrix, the low-rank components, and the remainder components respectively.

First, we divide the obtained enhanced Arabic speech into frames with length equal to 512 samples with half-length overlap. Second, we employ our subspace decomposition technique. We consider that the Lo lies on a low-rank sub-space, the speech signal is sparse, and Re is a random subspace. The assumption is based on the consideration that the speech signal is relatively sparse and has more variation whereas the noise spectrogram shows an iterative pattern.

Now, we solve the following convex optimization problem based on our work at [21]:

$$\min_{S,Lo} \left\| X - S - Lo \right\|_F^2 \text{ s.t. } card\left(S\right) \le s, rank\left(Lo\right) \le l (8)$$

Where $rank\left(Lo\right)$ is the rank of Lo, and $card\left(S\right)$ is the cardinality of S.

$\left\| . \right\|_F$ denotes the Frobenius norm of a matrix, s is the maximum number of entries in S, and l is low-rank constraint.

After the low-rank component and the sparse component are obtained, remainder noise Re is derived as $Re = X - S - Lo$.

Finally, we apply the inverse fast Fourier Transform function to obtain our enhanced Arabic speech signal.

3 Experiment and Results

In this section, the proposed approach for speech enhancement is evaluated using objective and Arabic speech recognition tests.

3.1 Simulation Conditions

To evaluate and compare the performance of our proposed method, we carried out simulations with the Arabic speech corpus, which is composed of phonetically balanced utterances. The Arabic speech signals are sampled at 8 kHz and it is composed of 1813 sentences belonging to three males and three female speakers. To test the robustness of our algorithm, we add three background noises (white, car, and babble) at three SNR levels (0, 5, and 10 dB) to the Arabic speech corpus [22]. The noise is taken from the Noisex92 [23].

3.2 Objective Evaluation Results

To measure the enhancement quality of the noisy Arabic speech, we calculate two parameters namely:

- The signal-to-noise ratio (SNR) measures the distortion of signal that reproduces the input signal.

- The segmental SNR (SegSNR) is formed by averaging frame level SNR estimates.

Our approach is compared to three recent speech enhancement methods named respectively the wavelet transform based on teager energy operator (WT-TEO) [14], geometric approach (GA) [4], and soft Masking using a priori SNR uncertainty on magnitude squared spectrum (MSS-SMPR) [8]. The results of the SNR and SegSNR measures are detailed in table 1.

For these two parameters, our proposed approach gives the best values except in two cases for the car noise at 10, and 5 dB where the MSS-SMPR method outperforms our approach. Moreover, the difference between the MSS-SMPR scores and those reached by our approach is not so large. Table 1 shows that the performance of the GA method remains acceptable but the WT-TEO keeps with the worst values.

Table 1 – Performance Comparison of Different Methods in Terms of SNR and SegSNR Measures

Type of Noise	SNR level	WT-TEO	Prop-Appr	GA	MSS-SMPR	WT-TEO	Prop-Appr	GA	MSS-SMPR
		SNR				SegSNR			
White	10 dB	12.31	14.34	12.51	13.70	1.49	2.08	1.34	1.81
	5 dB	8.11	10.12	8.39	9.45	-1.09	1.65	0.26	0.98
	0 dB	3.39	5.76	4.50	4.66	-1.08	1.02	-0.95	0.73
Car	10 dB	11.69	13.53	12.29	14.06	1.75	2.73	1.59	3.01
	5 dB	6.82	8.75	7.34	9.41	-0.34	1.06	0.01	1.35
	0 dB	2.97	4.55	3.18	3.56	-1.59	0.33	-1.43	-0.06
Babble	10 dB	11.03	13.70	11.42	11.78	1.26	2.38	1.11	1.74
	5 dB	5.94	8.36	6.45	7.59	-2.84	-0.90	-1.23	-1.25
	0 dB	3.55	6.07	3.81	4.24	-4.59	-1.83	-3.11	-2.69

3.3 Application to Rabic Speech Recognition

In this sub-section, we present the evaluation results for automatic recognition on enhanced Arabic speech signal. We reconstitute the clean speech from noisy observations based on the sparse imputation technique. It employs a non-parametric model and finding the sparsest combination of exemplars that jointly approximate the reliable features of a noisy utterance. That linear combination of clean speech exemplars is applied to replace the missing features. The advantage of the imputation approach is that the reconstructed clean speech features can be converted to cepstral features, which improves recognition accuracy at high SNR's. A clean speech frame is modeled by a mixture of Gaussians with diagonal covariance. This approach explains the imputation technique for a single Gaussian, but the results extend to a mixture of Gaussians.

For our experiments, we used a test set, which comprises 5 clean and 30 noisy Arabic speeches. The noisy speeches are composed of three noise types (white, babble, car) mixed at

the three SNR values. 23 frequency bands were used in all auditory filter-banks. The average recognition rate achieved by the sparse imputation technique is shown in table 2.

Table 2 – Arabic Speech Recognition Rate [%]

Type of noise	SNR level	Sparse imputation
White	10 dB	96.59
	5 dB	91.90
	0 dB	88.31
Car	10 dB	92.04
	5 dB	90.55
	0 dB	87.64
Babble	10 dB	94.78
	5 dB	89.24
	0 dB	83.86

The recognition accuracies exposed in table 2 indicate that sparse imputation technique can recover the missing data even at low SNR's.

4 Conclusion

The main idea of this paper is to apply convenient transformation on voiced sounds and other specific processing on unvoiced sounds to eliminate noise. It consists of thresholding the discrete wavelet transform coefficients calculated at four scales. But the threshold value depends on the voicing state of the analyzed frame.

The major contribution of this work consists to outperform the traditional thresholding based approaches. The proposed approach adapts the threshold value based on the V/UV decision and optimized subspace decomposition (SD) is applied.

Simulation results show that the proposed approach yields better results in terms of higher SNR, and SegSNR values than those of compared results in a better enhanced Arabic speech. Finally, the ASR results showed that the subspace decomposition (SD) is more advisable for speech enhancement applications than for ASR at low SNR.

5 References

[1] ARABI, P., and G. SHI: Phase-based dual- microphone robust speech enhancement. In IEEE Trans. Systems Man and Cybernetics Part B: Cybernetics, vol. 34, pp. 1763–1773, 2004.

[2] ORTEGA, A., E. LEIDA, and E. MASGRAU: Speech reinforcement system for car cabin communications. IEEE Trans. Speech Audio Process., vol. 13, pp. 917–929, 2005.

[3] BOLL, S.: Reduction of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process., vol. 2, pp. 112–120, 1979.

[4] YANG, L., and P. C. A LOIZOU: Geometric approach to spectral subtraction. Speech Comm., vol. 50, pp. 453–466, 2008.

[5] COHEN, I.: Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett., vol. 9, pp. 113–119, 2002.

[6] EPHRAIM, Y., and D. MALAH: Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process., vol. 32, pp. 1109–1130, 1984.

[7] EPHRAIM, Y., and D. MALAH: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process., vol. 33, pp. 443–448, 1985.
[8] LU, Y., and P. C. LOIZOU: Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. IEEE Trans. Audio Speech and Language Process., vol. 19, pp. 1123–1137, 2011.
[9] EPHRAIM, Y., and H. L. VAN TREES: A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process., vol. 3, (1995) pp. 251–266.
[10] WANG, J. F., and C. H. YANG, K. H. CHANG: Subspace tracking for speech enhancement in car noise environments. In Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing , 2004.
[11] DONOHO, D. L.: De-noising by soft-thresholding. IEEE Trans. Inform. Theory, vol. 41, pp. 613–627, 1995.
[12] MALLAT, S. G.: A Wavelet Tour of signal Processing. Burlington, 3rd ed. UK, Academic Press, Inc., 1999.
[13] PAN, Q., L. ZHANG, G. Z. DAI, and H. C. ZHANG: Two denoising methods by wavelet transform. IEEE Trans. Signal Process., vol. 47, pp. 3401–3406, 1999.
[14] BAHOURA, M., and J. ROUAT: Wavelet speech enhancement based on the teager energy operator. IEEE Signal Process. Lett., vol. 8, pp. 10–12, 2001.
[15] YU, G., E. BACRY, and S. MALLAT: Audio signal denoising with complex wavelets and adaptive block attenuation. In Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, 2007.
[16] SUMITHRA, A. M. G., and K. THANUSHKODI, Performance evaluation of different thresholding methods in time adaptive wavelet based speech enhancement. Internat. J. Engineering and Technology, vol. 1, pp. 440–447, 2009.
[17] HE, Z., and M. ZHANG: Detection and removal of impulsive white noise from noisy speech. In Proc. Wireless Comm. Networking and Mobile Computing, 2010.
[18] SAEED, A.: A new method for threshold selection in speech enhancement by wavelet thresholding. In Proc. Inter. Conf. on Computer Comm. and Management, 2011.
[19] KADAMBE, S., and G. FAYE BOUDREAUX-BARTELS: Application of the wavelet transform for pitch detection of speech signals. IEEE Trans. Info. Theory., vol. 38, pp. 917-924, 1992.
[20] JOHNSTON, I. M., and B. W. SILVERMAN: Wavelet threshold estimators for data with correlated noise. J. Roy. Statist.Soc. Ser. B., vol. 59, pp. 319–351, 1997.
[21] Ben Messaoud, M. A., and A. Bouzid: Sparse representations for single channel speech enhancement based on voiced/unvoiced classification. Circuits Syst. Signal Process., Published online 02 September 2016.
[22] HALABI, N.: Arabic speech corpus. Available at site http://en.arabicspeechcorpus.com/
[23] VARGA, A., and H. J. M. STEENEKEN: Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm., vol. 21, pp. 247–251, 1993.