A bit more on the ability of adaptation of speech signals

Un poco más de la habilidad de adaptación de las señales de voz

Dora M. Ballesteros L [1,2], Juan M. Moreno A [1]*

[1] Department of Electronic Engineering, Universitat Politecnica de Catalunya (UPC). Barcelona, Spain.

[2] Department of Telecommunications Engineering, Universidad Militar Nueva Granada (UMNG). Bogota, Colombia.

(Recibido el 31 de mayo de 2012. Aceptado el 18 de enero de 2013)

Abstract

Some traditional digital signal processing techniques encompass enhancement, filtering, coding, compression, detection and recognition. Recently, it has been presented a new hypothesis of signal processing known as the ability of adaptation of speech signals: *an original speech signal may sound similar to a target speech signal if a relocation process of its wavelet coefficients is applied*. This hypothesis is true under some conditions theoretically defined. In this paper we present the basic idea behind the hypothesis of adaptation and moreover, we test the hypothesis within four cases: speech signals with the same gender and language, speech signals with the same gender but different language, speech signals with the same language but different gender, and speech signals with different gender and language. It is found that the hypothesis is true if the requirements are satisfied, even if the gender or the language of the original and target signals are not the same.

---------- *Keywords:* Ability of adaptation, speech signals, wavelet coefficients, similarity

Resumen

Las técnicas tradicionales de procesamiento digital de señal incluyen mejoramiento, filtrado, codificación, compresión e identificación. Recientemente ha sido presentada una nueva hipótesis de procesamiento de señal conocida como la habilidad de adaptación de las señales de voz, en la que *una señal de voz (original) puede sonar similar a otra señal de voz (objetivo) si los coeficientes wavelet de la primera son re-ubicados*. Esta hipótesis es verdadera si se cumplen unas condiciones que han sido definidas

* Autor de correspondencia: teléfono: +34+93+401 56 91; fax: +34+93+401 67 56, correo electrónico: dora.maria.ballesteros@upc.edu (D. Ballesteros)

teóricamente. En este artículo presentamos la idea básica detrás de la hipótesis de adaptación y adicionalmente probamos la hipótesis en cuatro casos: señales de voz del mismo género e idioma, señales de voz del mismo género pero en diferente idioma, señales en el mismo idioma pero con diferente género, y finalmente, señales de voz que difieren tanto en el idioma como en el género. Una vez realizadas las pruebas, se estableció que la hipótesis de adaptación es válida incluso si el género (Femenino o Masculino) del hablante o el idioma del mensaje entre las dos señales (original y objetivo) no es el mismo.

----------*Palabras clave*: Habilidad de adaptación, señales de voz, coeficientes wavelet, índice de similitud

Introduction

In the area of speech processing there are a lot of techniques that manipulate the signal in order to enhance the quality of the signal [1 - 4], to identify patterns [5 - 7], to classify sounds [8, 9], to compress the signal [10 - 12] or to change the pitch of the voice [13]. But until now, to our knowledge, it has never been proposed a technique that modifies the signal so that it resembles (and sounds) like a target speech signal. This technique has been recently proposed by Ballesteros and Moreno [14] and this is the core of a speech-in-speech hiding scheme. In that scheme, the speech signal is camouflaged so that it resembles to the surrounding environment (the host speech signal) and then, the adapted-speech signal is hidden into the host speech signal [15].

The main step in the adaptation process is the relocation of the wavelet coefficients of the speech signal so that they resemble the behavior of the wavelet coefficients of the target speech signal. It is feasible because the distribution of the wavelet coefficients of the speech signal can be similar to the distribution of the wavelet coefficients of the target speech signal even if the speech signals have different behavior (plain-text, rhythm, gender of the speaker, among others). To obtain an adapted-speech signal similar to the target speech signal (the purpose of the adaptation process) it is necessary to satisfy the requirements of adaptation [14]. According to the previous study, it was found that the adaptation is feasible if and only if both signals have the same sampling frequency, time-scale and similar size of the non-silent time (or in other words, similar size of the non-zero wavelet coefficients). Additionally, it has been demonstrated that a vowel signal can be adapted to another one or to a voice signal and vice versa [14].

To enlarge the generalization of our proposed hypothesis of adaptation, in the current work we conduct several tests in which the gender or/and the language of the signals is not the same. The aim is to identify how the ability of adaptation runs in these conditions and in which cases the camouflage works best.

The rest of the paper is organized as follows. (Section 2) shows the background of Discrete Wavelet Transform (DWT) which plays an important role into the adaptation process. (Section 3) illustrates the idea behind the hypothesis of adaptation. (Section 4) shows the experimental tests related to the changes of the gender of the speaker and language of the plain-text. Finally, the paper is concluded in Section 4.

Background of the Discrete Wavelet Transform (DWT)

The DWT transforms a signal from time domain to time-frequency domain. It works in two steps, as follows [16]:

Firstly, the input signal (the speech signal in our case) is filtered with two filters, one half low-pass filter and one half high-pass filter;

Secondly, the above outputs are decimated by a factor of two.

It is illustrated in figure 1.

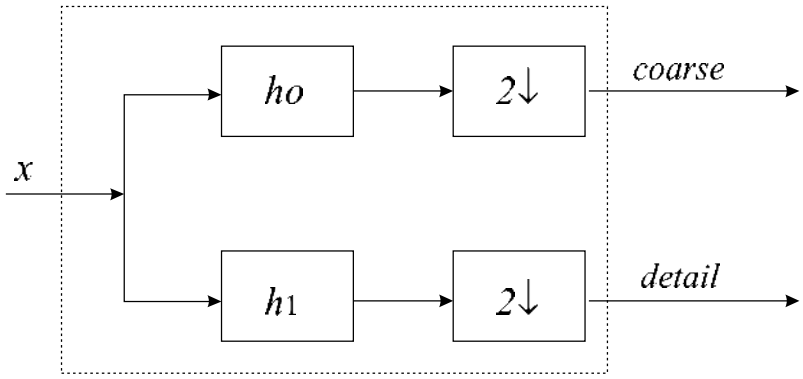


Figure 1 Scheme of the DWT

Since the detail of a signal is related to the high frequencies, the decimated-output of the high pass filter is known as the detail coefficients; in a similar way, since the approximation of a signal is related to the low frequencies, the decimated-output of the low pass filter is known as the coarse coefficients. Figure 1 illustrates the scheme in which x is the input signal, h_0 is the low pass filter, h_1 is the high pass filter and $2\downarrow$ is the decimation process by factor of two.

The basic idea behind the Ability of Adaptation

The speech signals can be considered as a signature of its owner because both the rhythm and tone are special characteristics that vary among people. For example, if the same plain-text is pronounced by two people, the time representation of their voices can be similar but their time-frequency representation (wavelet coefficients) is not. It is true even if the gender (and age) of the speaker is the same. Additionally, if the plain-text is modified, both the time and time-frequency representations of the speech signals will be completely different.

It can be easily illustrated with an example. Suppose there are two speech signals with different plaint-text, for example speech_1 with the plain-text *the purpose of this experimental test is to validate the hypothesis* and speech_2 with the plain-text *el propósito de esta prueba experimental es validar la hipótesis*. Both signals has the same sampling frequency (fs=8KHz), time-scale (t=5s), and are from a female speaker. Figure 2 shows the signals in time and wavelet domain.

As it is expected, both time and wavelet representations are different in each case. Now, two non-zero arrays are made from the non-zero wavelet coefficients of speech_1 and speech_2. Since there are a lot coefficients with magnitude close to zero, a threshold (*th*) is set which classifies the zero or the non-zero wavelet coefficients. If the magnitude of the wavelet coefficient is lower than *th*, then the thresholded coefficient is set to zero, but if this is higher than *th* (or equal), the amplitude of the coefficient is preserved. Once the two non-zero arrays have been obtained, their histograms are calculated. Figure 3 shows the histograms of the non-zero wavelet coefficients of the two speech signals.

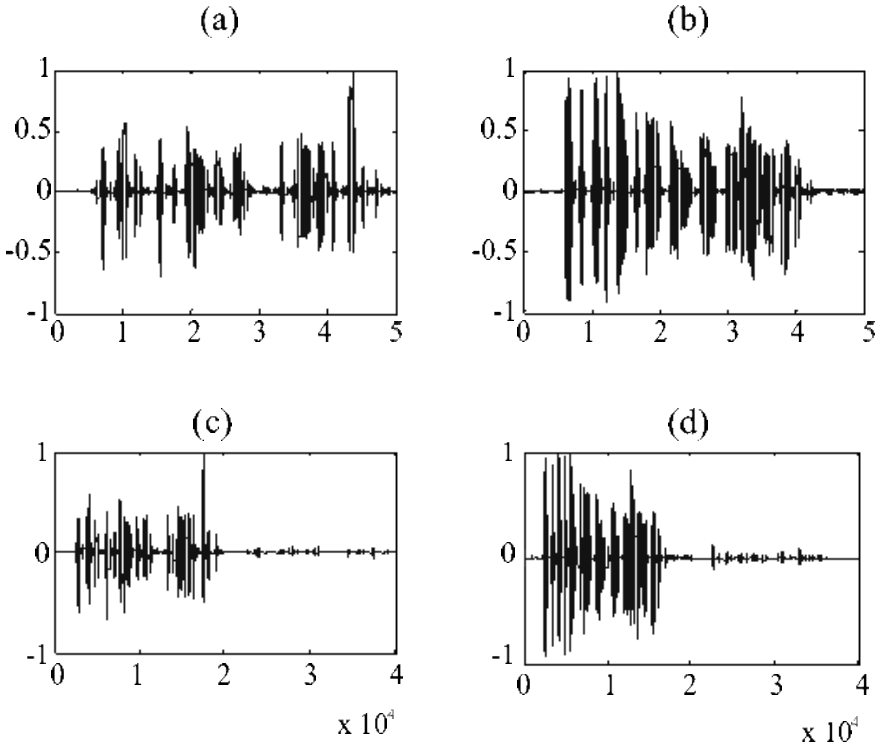


Figure 2 Time domain: a) speech$_1$; b) speech$_2$. Wavelet domain: c) speech$_1$; d) speech$_2$

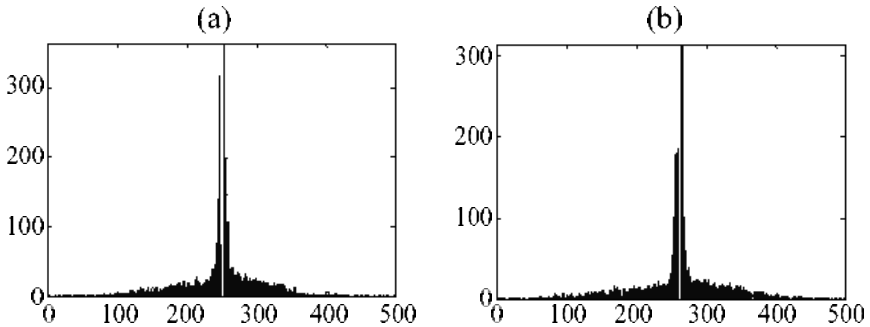


Figure 3 Histogram of the non-zero wavelet coefficients: a) from speech$_1$ signal; b) from speech$_2$ signal

According to figure 3, it is noticed that the histograms have similar shape. Since the *kurtosis* reflects the shape of a distribution, it is expected that the *kurtosis* of the above histograms is similar. The *kurtosis* is obtained as follows:

$$k = \frac{\sum_{i=1}^{N}(w_i - \mu)^4}{(N-1)\sigma^4} \qquad (1)$$

Where μ is the mean, σ^2 is the variance, k is the *kurtosis*, N is the total number of wavelet coefficients and w is the 1D-array of the non-zero wavelet coefficients of the speech signal.

In the current example, the *kurtosis* from the signals is 6.8 for speech$_1$ and 6.1 for speech$_2$. Since a similar shape of the histograms is related to a similar density of data, similar value of *kurtosis* means that the density distribution of the wavelet

coefficients is similar, too. In other words, although $speech_1$ and $speech_2$ sound different, the density distributions of their non-zero wavelet coefficients are similar. Therefore if the wavelet coefficients of $speech_2$ are relocated so that it resembles the wavelet coefficients of $speech_1$, the adapted-speech signal may sound similar to $speech_1$. This is the idea behind the ability of adaptation of speech signals.

In this context, $speech_2$ may sound similar to $speech_1$ (and vice versa) because their *kurtosis* of the non-zero wavelet coefficients is similar. Then, the adaptation is feasible if and only if the *kurtosis* and the number of the non-zero wavelet coefficients are similar between the speech signals.

Now, the steps of the adaptation process presented in [11] are carried out to adapt a speech signal to another one. Firstly, $speech_1$ is adapted so that it resembles $speech_2$. Secondly, $speech_2$ is adapted so that it resembles $speech_1$. Figure 4 shows the original speech signals and the adapted speech signals. To measure the similarity between the target speech signal and the adapted-speech signal, we use the Squared Pearson Correlation Coefficient (*SPCC*) because it can be seen as a speech distortion index [17, 18]. If the value is zero it implies that the adapted-speech signal is not similar to the target speech signal, but if the value is (close to) one, the adapted-speech signal is highly similar to the target speech signal and therefore the hypothesis is true. In the first case of adaptation, the index of similarity is 0.98 and the *ratio* of the non-zero wavelet coefficients is 1.23; in the second case the index of similarity is 0.98 and *ratio* is 0.81. It is worth noting that *ratio* of the second case of adaptation is *1/ratio* of the first case

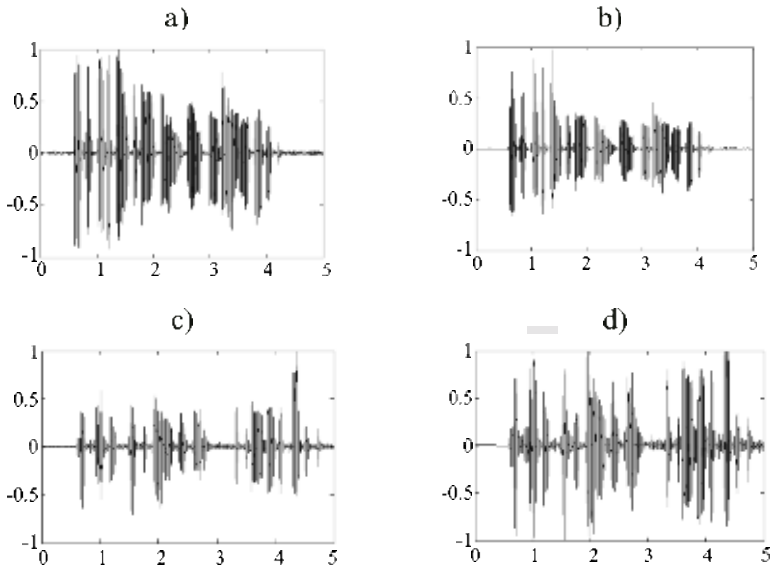


Figure 4 Time domain: a) $speech_2$; b) adapted-$speech_1$. c) $speech_1$; d) adapted-$speech_2$

Since the requirements of adaptation were satisfied (same *fs* as well as time-scale, and *ratio* close to 1), the adapted-speech signals are similar to the target speech signals. The high *SPCC* confirms the similarity. In the current example, a message in English language has been successfully adapted to a message in Spanish language and vice versa. The closer is the ratio to 1, the more similar are the *kurtosis* of the wavelet coefficients of the speech signals, and therefore the more similar is the adapted-speech signal to the target one. A broader validation of the hypothesis is presented in Section 4.

Experimental validation

The purpose of this experimental test is to validate the hypothesis of adaptation in relation to the gender and the language of the speech signals. Four scenarios are analyzed, as follows:

- First: the language of the messages and the gender of the speakers of both the speech and the target speech signals are the same.
- Second: the language of the messages is the same, but the gender of the speakers is different.
- Third: the gender of the speakers is the same, but the language of the messages is different.
- Fourth: both the gender of the speakers and the language of the messages are different.

The speech signals used in this work correspond to the following set:

- Group 1: five speech signals and five target speech signals from 10-female speakers in English language.
- Group 2: five speech signals and five target speech signals from 10-female speakers in Polish language.
- Group 3: five speech signals and five target speech signals from 10-male speakers in English language.
- Group 4: five speech signals and five target speech signals from 10-male speakers in Polish language.

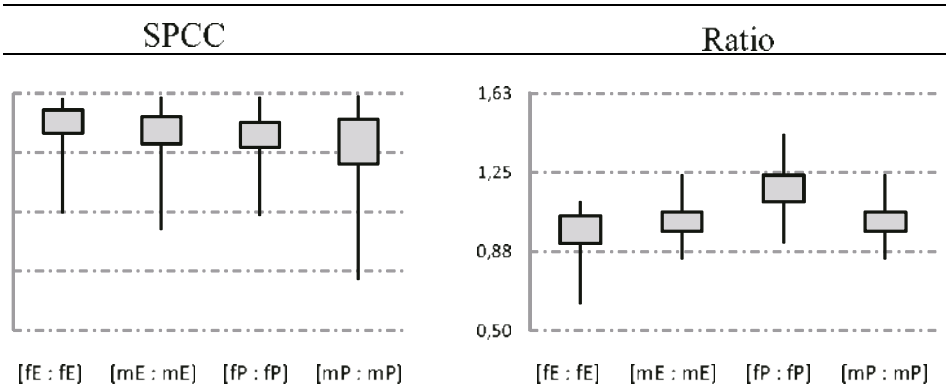
Every signal has been encoded with 16-bits, with a time-scale of ten seconds (it includes silent and non-silent time). In total there are 40 records belonging to 40 speakers. With the purpose to illustrate the generalization of the hypothesis; we select two languages phonetically different: English and Polish.

Now, we present the results by scenario. There are two measurements: the index of similarity, SPCC, and the *ratio* between the total non-zero wavelet coefficients of the speech signals (original and target). If *ratio* is 1, it means that the signals have the same number of non-zero wavelet coefficients, but if *ratio* is higher than 1 it means that the number of non-zero wavelet coefficients of the target speech signal is higher than the number of the non-zero wavelet coefficients of the original speech signal.

In tables 1-4 the results of the *SPCC* (first column) and the *ratio* of the non-zero wavelet coefficients (second column) are illustrated. There are 4 cases by scenario, each one with 25 tests, for a total by scenario of 100 tests. Every case is represented by its highest and lowest value, and the confidence interval of 95%. We use the following notation to name the scenario: the first and the second letter are related to the gender and the language of the original speech signal, respectively; while the third and the fourth letter are related to the gender and the language of the target speech signal, respectively. For example, the scenario [fP: mE] means that the original speech signal is from a female (f) speaker in Polish (P) language and the target speech signal is from a male (m) speaker in English (E) language.

According to table 1, most of the results have a similarity higher than 0.95 and all of them are higher than 0.9. It implies that the adapted-speech signals are highly similar to the target speech signals when simultaneously the gender of the speaker and the language of the message are unchanged. Since most of the ratios of the non-zero wavelet coefficients are in the interval [0.88, 1.25], it is concluded that if the number of the non-zero wavelet coefficients of the signals is similar, the adapted-speech signal will be highly similar to the target one.

Table 1 Results of the first scenario: same gender and same language



In table 2 and 3, the results of the second and third scenario are presented. Like the first scenario, in all cases the index of similarity is higher than 0.9. According to table 2, since the similarity between messages from female speakers is high even if the ratio of the non-zero wavelet coefficients is outside the interval [0.88 1.25], it can be suggested that the messages from female speakers are easier to camouflage than the messages from male speakers.

Table 2 Results of the second scenario: different gender but same language

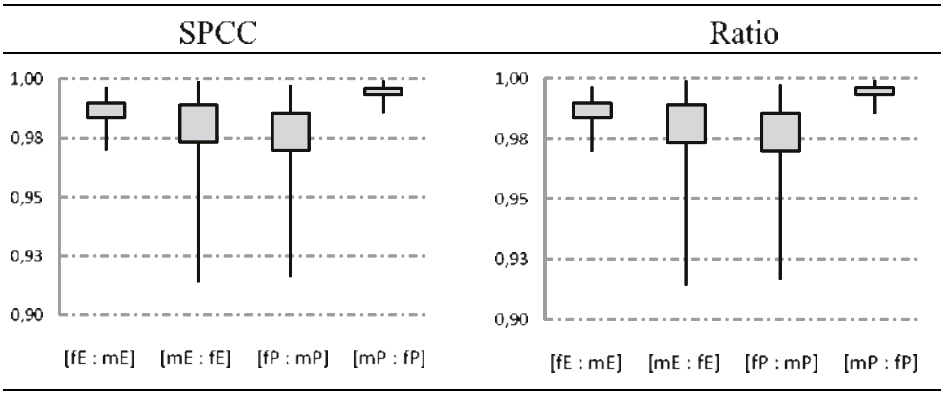
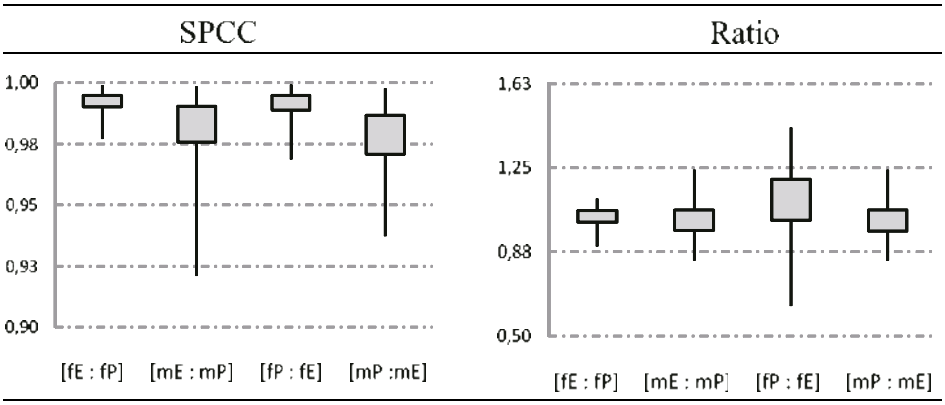


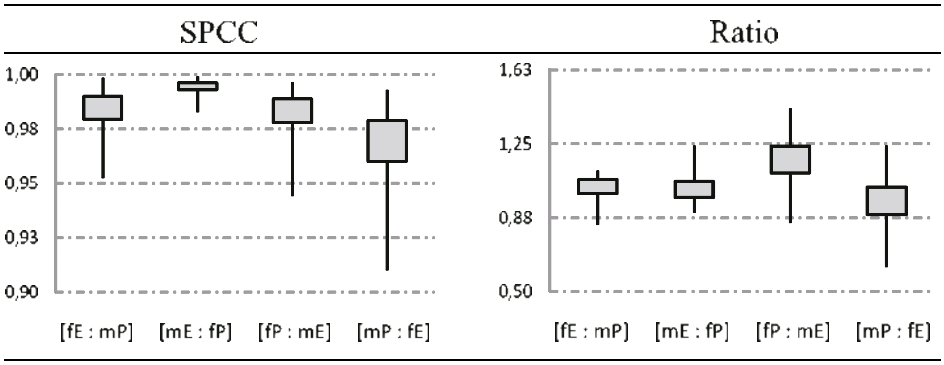
Table 3 Results of the third scenario: different language but same gender



Finally, the fourth scenario is presented in table 4. The first and second cases ([fE:mP] and [mE:fP]) have the *ratio* of non-zero wavelet coefficients into the interval [0.88 1.25] and the index of similarity higher than 0.95. The third and fourth cases ([fE:mP] and [mE:fP]) have the *ratio* outside the above interval and their index of similarity fell to 0.90. Nevertheless the index of similarity of the third case is better than of the fourth case. It can be concluded that the quality of the adapted-speech signal has a strong relationship with the *ratio* of the non-zero wavelet coefficients. If this value is into the range [0.88 1.25] it is expected that the similarity of the adapted-speech signal and the target speech signal will be high, but if the ratio is outside the above range, it is more desirable a ratio slightly higher than 1.25 instead of a ratio slightly lower than 0.88.

Summarizing, a speech signal can be adapted so that it resembles another speech signal if the theoretical requirements are satisfied even if the gender of the speaker and/or the language of the message is not the same.

Table 4. Results of the fourth scenario: different gender and different language



Conclusions

The basic idea behind the hypothesis of adaptation of speech signal was presented. It was illustrated that the histograms (and the kurtosis) of the non-zero wavelet coefficients between two speech signals will be similar if the theoretical conditions are satisfied. The closer are the histograms of the signals, the more similar is the adapted-speech signal to the target one. According to the results, it is confirmed that a speech signal may become similar to another speech signal, no matter the changes in gender of the speaker (female/male) or the language of the plain-text (English/Polish), if a requirement about the ratio of the non-zero wavelet coefficients is previously satisfied.

The importance of the results is that the adaptation in a useful tool to hide a speech signal into another speech signal without matter of the gender of the speaker or the language of the plain-text. The steganography model presented in [15] can be used with any kind of speech signals that satisfy the requirements of adaptation.

References

1 Y Hu, P. Loizou. "Speech enhancement based on wavelet thresholding the multitaper spectrum". *IEEE Transactions on Speech and Audio Processing*. Vol. 12. 2004. pp. 59- 67.

2 Y. Ghanbari, M. Karami. "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets". *Speech Communication*. Vol. 48. Issue 8. 2006. pp. 927-940.

3 Yu Shao, C. Hong. "A Generalized Time–Frequency Subtraction Method for Robust Speech Enhancement Based on Wavelet Filter Banks Modeling of Human Auditory System". *IEEE Transactions on Systems, Man, and Cybernetics.* Vol. 37. 2007. pp.877-889.

4 S. Senapati, S. Chakroborty, G. Saha. "Speech enhancement by joint statistical characterization in the Log Gabor Wavelet domain". *Speech Communication*. Vol. 50, 2008. pp. 504-518.

5 M. Eshaghi, M. Karami. "Voice activity detection based on using wavelet packet". *Digital Signal Processing*. Vol. 20. 2010. pp. 1102-1115.

6 C. Hsieh, E. Lai, Y. Wang. "Robust speech features based on wavelet transform with application to speaker identification". *IEEE Proceedings Vision, Image and Signal Processing*. Vol. 149. 2002. pp. 108- 114.

7 O. Farooq, S. Datta. "Wavelet-based denoising for robust feature extraction for speech recognition". Electronics Letters. Vol.39. 2003. pp. 163-165. DOI: 10.1049/el:20030068

8 E. Avci, Z. Hakan Akpolat. "Speech recognition using a wavelet packet adaptive network based fuzzy inference system". *Expert Systems with Applications*. Vol. 31. 2006. pp. 495-503.

9 J. Hung, H. Fan. "Subband Feature Statistics Normalization Techniques Based on a Discrete Wavelet Transform for Robust Speech Recognition". *IEEE Signal Processing Letters.* Vol. 16. 2009. pp. 806-809.

10 S. Joseph, P. Babu. *Speech compression using wavelet transform*. International Conference on Recent Trends in Information Technology (ICRTIT). 2011.

11 M. Osman, N. Al, H. Magboub, S. Alfandi. *Speech compression using LPC and wavelet*. Second International Conference on Computer Engineering and Technology (ICCET). 2010.

12 Z. Dan, M. Shengqian. *Speech Compression with Best Wavelet Packet Transform and SPIHT Algorithm*. Second International Conference on Computer Modeling and Simulation. 2010.

13 R. Veldhuis, H. He. "Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform". *Speech Communication,* Vol. 18. 1996. pp. 257-279.

14 D. Ballesteros, J. Moreno. "On the ability of adaptation of speech signals and data hiding". *Expert Systems with Applications*. Vol. 39. 2012. pp. 12574-12579.

15 D. Ballesteros, J. Moreno. "Highly transparent steganography model of speech signals using Efficient Wavelet Masking". *Expert Systems with Applications.* Vol. 39. 2012. pp. 9141-9149.

16 S. Mallat. *Wavelets and Filters Banks*. A wavelet tour of signal processing. 2nd Edition. Ed. Academic Press. Second Edition. 1999. pp. 255-264.

17 J. Benesty, C. Jingdong, H. Yiteng. "On the Importance of the Pearson Correlation Coefficient in Noise Reduction". *IEEE Transactions on Audio, Speech, and Language Processing*. 2008. pp. 757-765.

18 J. Benesty, C. Jingdong, H. Yiteng, I. Cohen. "Pearson Correlation Coefficient, in: Noise Reduction in Speech Processing". *Springer Topics in Signal Processing*. Vol. 2. 2009. pp. 1-4.