Method for Duration and Spectral Shape Transformations of Speech for the Diagnosis of Hearing Problems

F. Ykhlef

Division MultiMedia, Centre de Développement des Technologies Avancées, Algérie
fykhlef@cdta.dz

Abstract

In this paper, we present a method for duration and spectral shape transformations of speech based on the combination of pyramidal Discrete Wavelet Transform (DWT) and Synchronized Overlap and Addition (SOLA) technique. The developed scheme is devoted to the diagnosis of hearing problems. The SOLA technique is basically used for duration modifications. The pyramidal DWT gives a semi logarithmic frequency decomposition that represents the frequency scales of human hearing. In addition, it allows a separate manipulation of speech sub-bands. In the proposed scheme, we explore the frame synchronization provided by SOLA technique and the frequency reparation given by the DWT to develop simultaneous time-scale and frequency-magnitude transformations. We have compared the developed method to another scheme that uses the same concept in which the frequency decomposition is based on the Cosine Modulated Pseudo- Quadratic Mirror Filter Bank (CMP-QMF). A software system of speech transformations is implemented. Experimental results show that speech waveforms can be successfully modified to different time-scales and frequency-magnitudes with high quality.

Keywords: *SOLA, DWT, Hearing Problems, Diagnosis*

1. Introduction

The acoustic parameters that carry prosodic information (pitch, loudness, duration and timbre) are always characterized by sudden changes in continuous speech. For someone with hearing problems, these variations do not fit in his dynamic frequency range; therefore, there is lack of certain parts of the information. Listeners with hearing impairment, especially elderly people, often have difficulty in comprehending fast speech. In addition, they may have hearing loss at certain frequencies; generally, high frequencies. In this situation, hearing aids can offer certain improvement of speech intelligibility [1-3].

Amplifying sound through a hearing aid does not always help people with hearing loss better understand speech. However, compressing speech by decreasing its pitch, slowing down its speed, increasing its loudness through a new computer technology seems to help people with high frequency hearing loss better understand speech according to the university of Iowa research finding [4]. Although the latest hearing aids employ digital technology, they do not use it mainly for duration and spectral shape transformations simultaneously [5].

Several studies have been proposed in the literature to solve this problem [3, 5-11]. Slowing the speed of speech is a common technique for helping listeners with hearing impairment comprehended more easily. Nakamura *et al.*, [8] proposed a real time speech rate conversion especially designed for television sets. The portable digital speech rate converter for hearing impairment developed by Nejime *et al.*, [3] which slows speech without changing its pitch uses only a temporal Time-Scale Modification (TSM) algorithm.

One class of methods widely used in hearing research and applications is the sinusoidal representation of speech [6, 7]. Another family of methods uses a parametric concept of speech to reproduce the desired signal using a transformation model [9].

A. Montgomery and R. Edge [10] report the effects on speech intelligibility of two types of digital speech processing. The first type enhances the amplitude consonants to produce near-zero consonant/vowel intensity ratios. The second one increases the duration of consonants to provide an additional 30 ms of sound. It has been concluded that the development of the speech-processing techniques is necessary before their incorporation into a useful hearing aid.

G. Quagliaro *et al.*, [11] propose a method based on source-filter modeling to provide auditory correction of hearing-impaired individuals. The developed scheme includes several processing. Among them, we have pitch modification, voicing detection and spectrum estimation. Perceptual evaluation on a group of hearing impaired persons shows the efficiency of the proposed method.

F. Ykhlef *et al.*, [5] proposed a new concept for simultaneous time-scales and frequency-magnitudes transformation of speech. It is named Filter Bank Synchronous Overlap and Addition (FBSOLA). The developed concept explores the synchronization offered by Synchronized Overlap and Addition (SOLA) technique [12], which is basically used for duration modification, to accomplish simultaneous modifications of spectral shape and duration. The synchronization is carried out on the reconstructed signal and more precisely at the output of the filter bank for each frame. The spectral shape modification is performed in the desired sub-band. The Cosine Modulated Pseudo-Quadratic Mirror Filter bank (CMP-QMF) has been used to divide the speech signal into sixteen equally spaced channels.

In this paper, we present a modified scheme for duration and spectral shape transformation of speech based on the FBSOLA method. We have used in this application a semi logarithmic frequency spacing to generate the requested sub-bands. The decomposition is based on a pyramidal representation using the Discrete Wavelet Transform (DWT) to generate the frequency-scales used in audiometric tests [13]. The semi logarithmic repartition is appropriately approximating the real frequency scales of human perception. The proposed scheme is named pyramidal DWT-SOLA. It is basically devoted to speech repair, diagnosis and hearing disorders applications.

Our paper is organized as follows. Section 2 presents the main steps used to accomplish spectral shape and duration transformations. In Section 3, a software version of the proposed scheme is presented. Section 4 gives the main results from evaluating the signal processing part of the method. In addition, a comparative study between the proposed scheme and the one reported in [5] is given. Finally, a conclusion with the main perspectives to our work is laid in the last section.

2. Spectral Shape and Duration Transformations

2.1. Pyramidal DWT Representation

Wavelet transform has been intensively used in various fields of signal processing. It is a powerful tool for modeling non stationary signals such as speech that exhibit slow temporal variations at low frequencies and abrupt temporal changes at high frequencies.

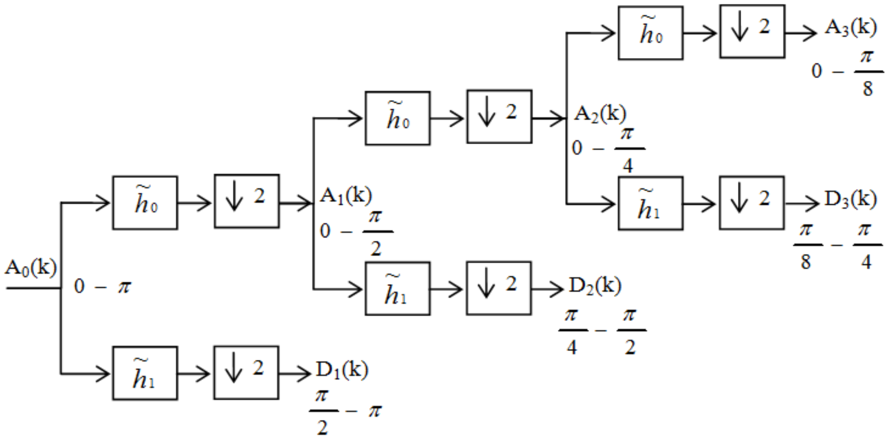


Figure 1. Three-Stage Pyramidal Decomposition

An efficient way to implement this scheme using filters was developed in 1989 by S. Mallat [14]. This practical filtering algorithm yields a pyramidal representation or what is called multi level analysis. A typical scheme for the DWT is depicted in Figure 1. The decomposition filters $\tilde{h}_0$ and $\tilde{h}_1$ were adopted to split the input signal into two overlapped and equally spaced frequency bands at the first level, where $\tilde{h}_0$ is a low pass filter and $\tilde{h}_1$ is a high pass one. Then, we decimate each band by a factor 2, such that the spectrum of each band is expanded to fill up the full frequency scale. The frequency bands are shown in Figure 1 as normalized form, where π denote the Nyquist frequency. The splitting, filtering and decimation can be repeated on the scaling coefficients to give the idea of multi level analysis.

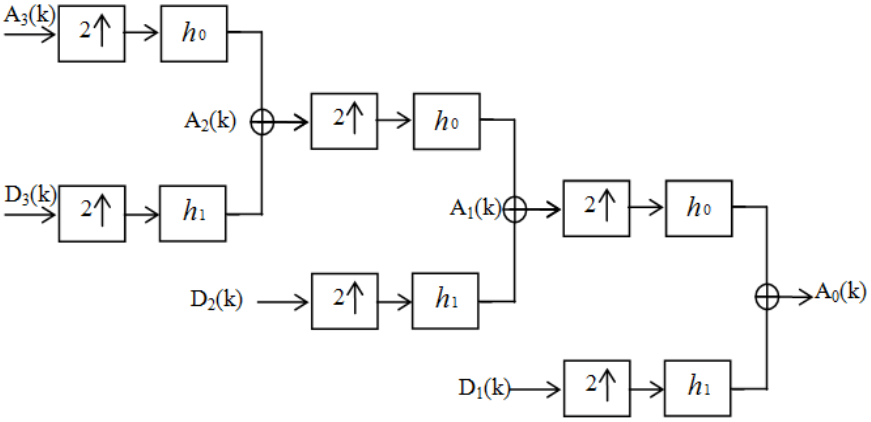


Figure 2. Three-Stage Pyramidal Reconstruction

The scaling coefficients are called respectively the approximation and detail coefficients. The approximation coefficients are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. The filter coefficients play a crucial role in a given DWT and have to satisfy orthonormalites and a certain degree of regularity.

Several set of filters can be found in the literature. As example, Daubechies, Haar, Symlets and Coiflets. In the reconstruction process, the approximation and detail coefficients at every level are upsampled by two, passed through the low pass and high pass synthesis filters and then added. This process is continued through the same number of levels as in the decomposition process to obtain the original signal (Figure 2).

The Mallat algorithm works equally well if the analysis filters, $\tilde{h}_0$ and $\tilde{h}_1$, are exchanged with the synthesis filters $\tilde{h}_0$ and $\tilde{h}_1$. To achieve perfect reconstruction without aliasing, the decomposition and reconstruction filters have to satisfy certain conditions according to the wavelet family [14].

2.2. Pyramidal DWT-SOLA

In the proposed scheme, the frequency-modified frames of the speech waveform are shifted and averaged according to a time scaling factor obtained at the highest cross-correlation points. Simple shifting and adding of the transformed frames would achieve the goal of modifying time-scales and amplifying the desired set of frequency but it would not conserve pitch periods, spectral magnitudes or phases. Therefore, it would be expected to produce poor speech quality. However, adding the frequency-modified frames in a synchronized fashion at point of highest cross-correlation serves to preserve time-dependent pitch, spectral magnitudes and phases to a large degree.

The speech waveform is frequency-decomposed by splitting each frame into several sub-bands using a pyramidal representation based on the DWT (Figures 1 and 2). The amplification is performed in the desired sub-band. The frequency-scales, used in the transformation structure, have to approximate the real frequency composition of human perception. Several representations have been reported in the literature. For example, we can mention Mel and Bark scales. In addition to these representations, it has been found that a pyramidal structure based on the DWT can highly approximate the frequency repartition of hearing system.

In this application, we use the pyramidal DWT to construct the frequency-scales adopted by the audiometric tests [13]. These scales correspond to the octave band representation of a set of frequencies which starts from 125 Hz and ends to 8 kHz {125, 250, 500, 1k, 2k, 4k, and 8k Hz}. The obtained sub-bands using the pyramidal DWT structure, which includes the decomposition and reconstruction steps, are mainly related to the sampling rate and the number of levels. For a fixed sampling frequency (Fs) of 16 kHz, the decomposition into six-stage (or seven sub-bands) gives the audiometric frequency-scales used in our application (Figure 3).

The corresponding gains for each sub-band are given as follows:

- G_6 for the sub band 4 to 8 kHz,
- G_5 for the sub band 2 to 4 kHz,
- G_4 for the sub band 1 to 2 kHz,
- G_3 for the sub band 0.5 to 1 kHz,
- G_2 for the sub band 0.25 to 0.5 kHz,
- G_1 for the sub band 0.125 to 0.25 kHz,
- G_0 for the sub band 0 to 0.125 kHz.

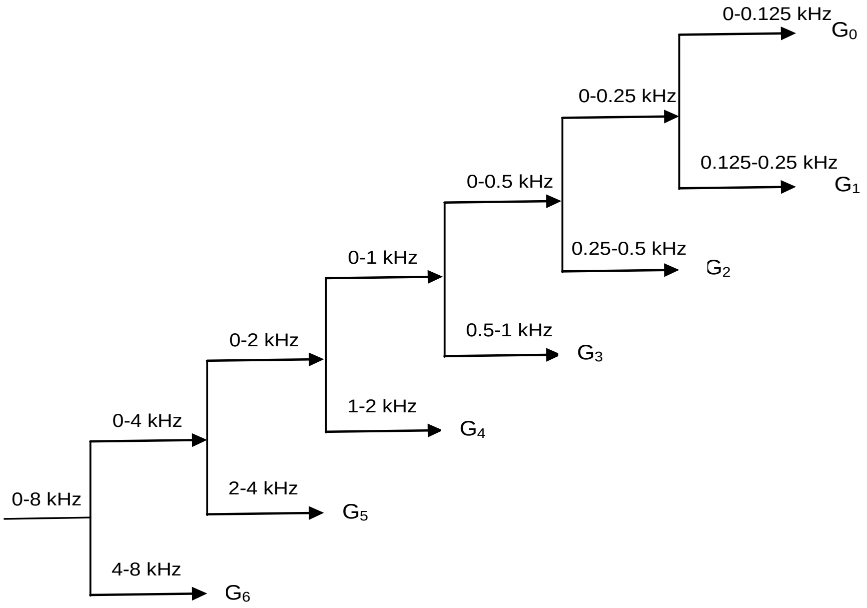


Figure 3. Six-Stage Pyramidal Decomposition of Single Frame for a Sampling Rate of 16 kHz

The reconstruction structure is obtained using the inverse scheme of the decomposition one. It is the same structure as shown in Figure 2 but the number of stages is equal to six. The amplification of the desired set of sub-bands is performed using the G_j (j=0 to 6) gain factors. In the pyramidal DWT-SOLA scheme, the speech signal x(n) is time-scaled using a factor α to give the output waveform y(n), where:

— α >1: corresponds to time expansion;

— α<1: corresponds to time compression.

Overlapping frames of size N are taken every S_a samples of x(n), then passed through the pyramidal structure in which the sub-bands are modified adequately. The reconstructed signal is named $\overleftrightarrow{x(n)}$. S_a denotes the analysis interframe interval. Each level in the analysis-synthesis structure must be adjusted to eliminate the lag introduced by the filtering operations. If S_s is the synthesis interframe, then S_s is related to S_a by:

$$S_s=\alpha S_a \qquad (1)$$

The synthesis is performed on a frame by frame basis, where each new analysis frame is added to the reconstructed signal. The algorithm is initialized using equation (2):

$$y(j)=x(j), \quad 0 \le j \le N\text{-}1 \qquad (2)$$

$\overleftrightarrow{x(mS_a + j)}$ denotes the m^{th} frame of the input signal that passes through the filter bank for 0≤j≤N-1. Each frame is synchronized and averaged with the neighborhood of $y(mS_s+j)$.

The alignment is obtained by computing the cross-correlation $R_m(k)$ at frame m between the two entities. It is represented as follows:

$$R_m(k) = \sum_{j=0}^{L-1} y(mS_s + k + j)\overleftrightarrow{x(mS_a + j)} \tag{3}$$

where $-N/2 \leq k \leq N/2$, and L is the number of points used to compute each cross-correlation.

k_m denotes the leg at which $R_m(k)$ is maximum. $\overleftrightarrow{x(mS_a + j)}$ is weighted and averaged by $y(mS_s+j+k_m)$ as described in [15]. According to the variable j, we have:

$0 \leq j \leq L_m-1$:

$$y(mS_s + k_m + j) = (1 - f(j))y(mS_s + k_m + j) + f(j)\overleftrightarrow{x(mS_a + j)} \tag{4.1}$$

$L_m \leq j \leq N-1$:

$$y(mS_s + k_m + j) = \overleftrightarrow{x(mS_a + j)} \tag{4.2}$$

L_m is the range of overlap of the two signals, and f(j) is a weighting function such that $0 \leq f(j) \leq 1$.

To reduce the high cross-correlation between y and $\overleftrightarrow{x}$ that can acquire for low value of L, this latter must be restricted to values greater than N/8 [15]. In this application, it is fixed to 1.2×N/8.

The choice of S_a and S_s depends on N and α. A smaller S_a will result in higher speech quality but increases the computation complexity. In practice, we search to maximize S_a without affecting the speech quality significantly. As a rule of thumb, we set:

$$\text{when}\, \alpha < 1:\ S_a = \frac{N}{2}, \tag{5.1}$$

$$\text{when}\, \alpha > 1:\ S_a = \frac{N}{2\alpha}, \tag{5.2}$$

The speech signal is divided into R frames, each one has a length of 128ms which is equivalent to N=2048 samples for an F_s of 16 kHz.

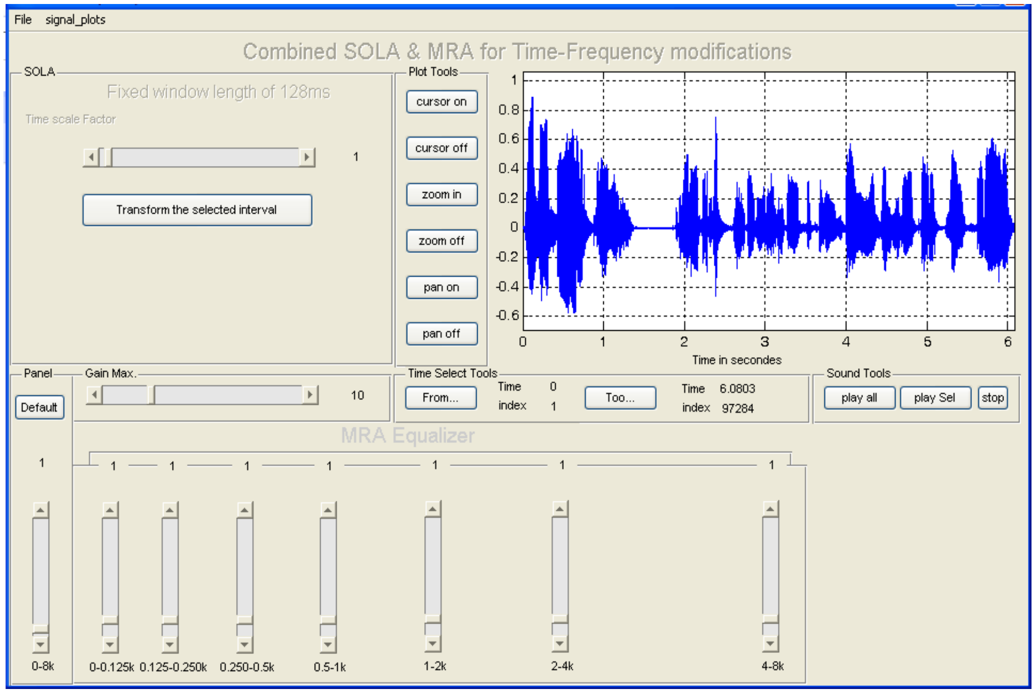
3. Software Description

We have designed a software system in order to perform complex transformation of speech waveforms (Figure 4). It can perform simultaneous duration and spectral shape transformations based on a time scale factor (α) and seven gain factors (seven sub-bands).The modifications can be done on the entire speech waveform or on selected intervals. The transformed speech can be saved as wave file. Furthermore, all the input parameters such as frame length, wavelet order or type and weighting function can be tuned. Different duration and spectral shape transformations can be performed on the same speech signal. The potential usefulness of the system for hearing research applications is illustrated by high frequency amplifications and speed slowing of the input waveform.

4. Results and Interpretations

We have assessed the proposed scheme by an informal subjective evaluation of Mean Opinion Score (MOS) by a group of three researchers with normal hearing capabilities.

The speech assessment is ranged from one to five which represent respectively the worst and best score.



Combined SOLA & MRA for Time-Frequency modifications

Fixed window length of 128ms

Time scale Factor 1

Gain Max. 10

Time 0
index 1

Time 6.0803
index 97284

Figure 4. Pyramidal DWT-SOLA

Each listener was asked to give his appreciation. The mean value of the obtained set of MOSs is considered in our evaluation.

Fifteen phonetically representative Arabic sentences, stored on a disk, were used as samples to transform the speech waveform using several duration and spectral shape combinations. These sentences cover the phonetic representation of Standard Arabic language which includes: occlusives (voiced and unvoiced), fricatives (voiced and unvoiced), nasals, liquids, vibrants, african phonemes, semi-vowels and vowels [5]. The speech corpus is sampled at 16 kHz.

We have compared the DWT-SOLA scheme to the CMP-QMF-SOLA reported in [5]. Technically, the main difference between both transformation schemes is related to frequency decomposition. The DWT-SOLA uses a semi logarithmic-scale which represents the set of frequencies defined in the audiometric tests. However, the CMP-QMF-SOLA scheme uses 16-channels equidistantly spaced with a bandwidth of 500 Hz. Our evaluation is twofold:

- Assessment of SOLA technique for TSM based on MOS test,
- Performance comparison between the two previous transformation schemes for simultaneous duration and spectral shape transformations.

The length of the analysis frame is equal to 128ms. It gives the best speech quality compared to other values. We have used this value for both schemes since they use a similar TSM technique.

For each sentence, we have modified the duration using the following set of factors: {0.4, 0.6, 0.9, 1.3, 1.6, 1.8, 2, 2.3, 2.6, 2.8 and 3.2}. MOS values for each factor are summarized in Table 1. The evaluation of TSM indicates that the setting parameters of the SOLA algorithm are well elaborated. Thus, the modified speech is obtained with good naturalness and intelligibility except for the highest and lowest factors where the transformed speech is slightly degraded. According to our perceptual evaluation, the duration modification has its best scores for time modification factors between 0.6 and 2.8.

Table 1. MOS Values for Duration Modifications

α	**MOS**
0.4	2.86
0.6	3.73
0.9	5
1.3	5
1.6	5
1.8	4.66
2	4.66
2.3	4.40
2.6	4.26
2.8	4
3.2	3.06

The time-frequency proprieties of both schemes are qualitatively summarized in Table 2. Technically, the proposed method for both schemes is based on an output similarity technique to achieve the synchronization for simultaneous duration and spectral shape transformations. Furthermore, this robust synchronization requires a window length greater than four times the local pitch period of a given voiced frame. For unvoiced frames, in the absence of periodicity, the same frame length is used in our application since we do not distinguish between these two regions.

The windowing technique is a linear fad-in/fad-out function. It is used in order to preserve the gain of the overlapped windows.

The computational efficiency of the CMP-QMF-SOLA structure, in terms of frequency decomposition, is medium compared to the DWT-SOLA structure.

The linear frequency repartition of 500 Hz bandwidth provided by the Cosine Modulated Pseudo-Quadratic Mirror Filter bank (CMP-QMF) requires an amount of 16-channels. However, the pyramidal DWT structure requires only six level of decomposition which is enough to approximate the human frequency repartition. Even if we use the pyramidal structure based on CMP-QMFs, the results will not be satisfactory due to the near perfect reconstruction allowed by this kind of filter bank.

Furthermore, the design problem of pseudo QMFs, more precisely the CMP-QMF, is constrained to the computation of the optimal coefficients of the prototype filter. Generally, designing such filters lead to a highly nonlinear optimization problem that cannot be easily solved.

Conversely, the use of the DWT offers perfect reconstruction of the output speech for any wavelets type or order. The spectral characteristics of the analysis-synthesis filters are extremely related to their orders. It is well known that high filter orders give perfect spectral characteristics of the analysis and synthesis filters. In addition, Daubechies filters offers an optimal time-frequency resolution compared to other orthogonal wavelets. This property cannot be achieved using the CMP-QMFs structure.

Table 2. DWT-SOLA and CMP-QMF-SOLA Proprieties

Prop. \ Scheme	CMP-QMF-SOLA	DWT-SOLA
Synchronization	output similarity	output similarity
Window length	fixed (> 4.pitch)	fixed (> 4.pitch)
Windowing	Linear weighting function f(j)	Linear weighting function f(j)
Computational efficiency	Medium	high
Approximation of hearing scales	No	Yes
Perfect reconstruction of the decomposed sub- bands	No	Yes
Optimization of the prototype Filter	Yes	No
Time-frequency resolution	low	High
Quality of transformations	Medium	High

The DWT-SOLA scheme uses Daubechies filters of order 16 which offers a good spectral representation of the decomposition and reconstruction filters. However, the optimal order of the prototype filter used by the CMP-QMF-SOLA scheme is equal to 51. The perceptual quality of simultaneous spectral shape and duration transformations depends on all these characteristics.

Moreover, the authors in [5] have found that the increase of the CMP-QMF sub-bands, for a number greater than 16 channels (in the CMP-QMF-SOLA scheme), will significantly reduce the quality of the modified speech and introduces some perceptible artefacts glitches. Conversely, The SOLA- DWT scheme offers a transformed speech with high quality even if the number of levels is increased.

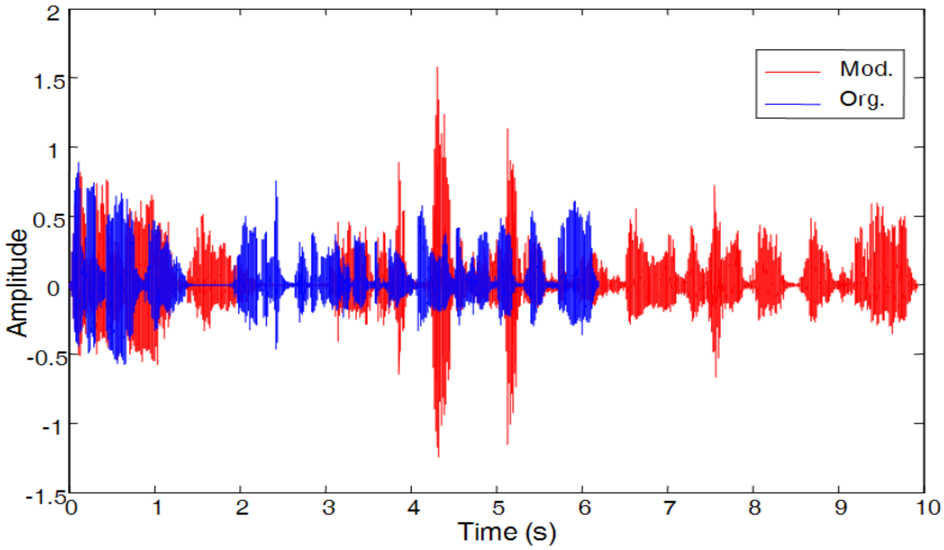


Figure 5. Simultaneous Spectral Shape and Duration Transformations via DWT-SOLA

Figure 5 shows duration dilation by a factor of 1.6 and an amplification of the higher frequency band (4 to 8 kHz) of the input speech via SOLA-DWT. The original signal is represented in blue and the transformed one is in red. The initial duration of the speech waveform was 6.2 seconds and has been transformed into 9.92 second simultaneously with the spectral component.

The spectral shape of the transformed speech is displayed in Figure 6 in which the desired frequency band is amplified and dilated according to the applied time-scale factor. The amplifications of the spectral components can be performed for any frequency band. An adequate perceptual evaluation need to be done by hearing impairment listeners to get their opinions.

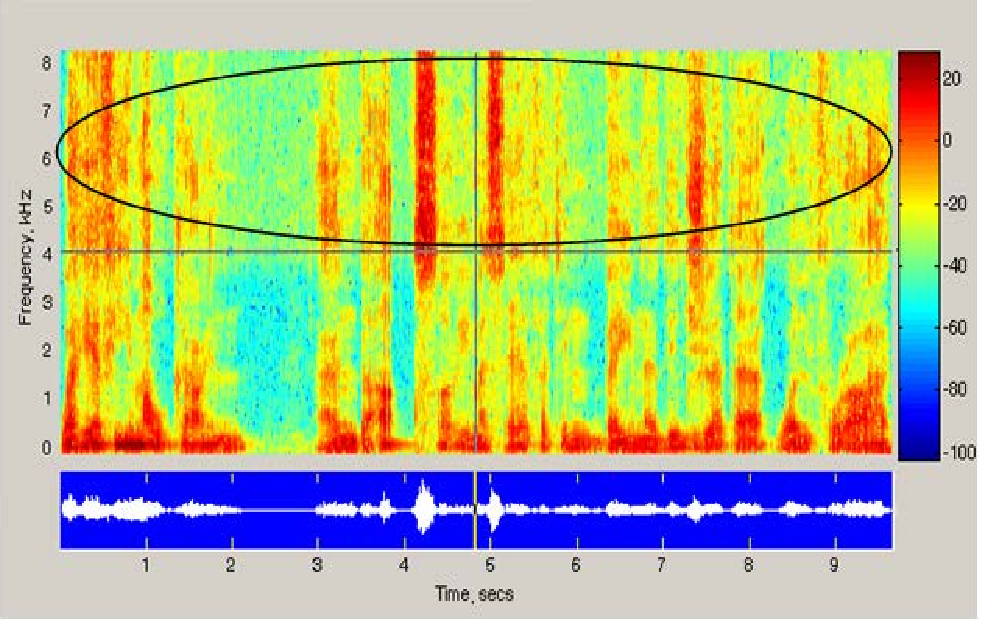


Figure 6. Spectrogram of the Transformed Speech via DWT-SOLA

5. Conclusion

We have proposed in this paper a transformation scheme of speech based on pyramidal DWT for frequency decomposition/amplification and SOLA algorithm for time-changing. The proposed scheme tackles the problem of simultaneous duration and spectral shape transformations of speech for listeners with hearing impairments. The semi logarithmic frequency scales provided by the pyramidal DWT approximate the real frequency repartition of human perception. Thus, the spectral shape is transformed based on a logarithmic frequency representation.

We have implemented a software version of the proposed speech transformation system using a pyramidal decomposition of six-stage. The proposed scheme is compared to another one based on pseudo-quadratic mirror filter bank. The resulting quality and versatility of the transformation make this system well suited to hearing related research. It can be used for the diagnosis of hearing disorders.

This paper reports on results from evaluating the signal processing part of the simultaneous spectral shape and duration transformations. Evaluation of the idea in field of tests for impaired listeners is still to be reported.

References

[1] A. M. Engebretson, “Benefits of Digital Hearing Aids”, IEEE Eng. Med. Biol. Mag., vol. 2, no, 13, **(1994)**.

[2] D. Van Tasell and T. Crain, “Hearing Loss, Speech, and Hearing Aids”, J. of Speech and Hear. Res., vol. 2, no. 36, **(1993).**

[3] Y. Nejime, T. Aritsuka and T. Imamura, “A Portable Digital Speech Rate Converter for Hearing Impairment”, IEEE Trans. Rehab. Eng., vol. 4, no. 2, **(1996)**.

[4] Web site of hearing aids lab, University of IOWA. Available: www.uiowa.edu., **(03.01.2013)**.

[5] F. Ykhlef, M. Bensebti, W. Benzaba, L. Bendaouia, R. Boutaleb and H. Meraoubi, “A New Method for Time-Frequency Modification of Speech Signal Using a Combining Cosine Modulated Pseudo-QMFBank & SOLA Algorithm for Listeners with Hearing Impairment”, Proceding of the 2nd International Conference on Advanced Computer Theory and Engineering, Cairo, Egypt, **(2009)** September 25-27.

[6] L. Cheng-Lung, Ch. Wen-Whei and Ch. Yuan-Chuan, “Spectral and Prosodic Transformations of Hearing-Impaired Mandarin Speech”, Speech Commun., vol. 2, no. 4, **(2006)**.

[7] J. Yvan and C. Howard Lee, “A Speech Analysis/Synthesis Software System for Hearing Research”, Proceding of the IEEE International Conference of Engineering in Medicine and Biology Society, Orlando, FL, USA, **(1991)** October 31-November 3.

[8] A. Nakamura, N. Seiyama, R. Ikezawa, T. Takagi and E. Miyasaka, “Real Time Speech Rate Converting System for Elderly People”, Procedding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Australia, **(1994)** April 19-22.

[9] K. Hermansen, F. K. Fink, U. Hartmann, “Hearing Aids for Profoundly Deaf People based on a New Parametric Concept”, Proceding of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, **(1993)** Octobre 17-20.

[10] A. Montgomery and R. Edge, “Evaluation of Two Speech Enhancement Techniques to Improve Intelligibility for Hearing-Impaired Adults. J. of Speech and Hear. Res. 3, vol. 31, **(1988).**

[11] G. Quagliaro, Ph. Gournay, F. Chartier and G. Guilmin, “Method and Device for the Processing of Sounds for Auditory Correction for Hearing Impaired Individuals”, U.S. Patent 6,408,273, **(2002)** June.

[12] S. Roucos and A. M. Wilgus, “High Quality Time-Modification for Speech”, Procedding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, USA, **(1985)** April 26-29.

[13] F. Ykhlef, W. Benzaba, L. Bendaouia, R. Boutaleb and A. Benia, “Computer Audiometer for Hearing Testing”, IEEE International Conference on Advances in Electronics and Micro-electronics, Valencia, Spain, **(2008)** September 29-October 04.

[14] S. Mallat, “Multiresolution Approximations and Wavelet Orthonormal Bases of L2(R)”, T. Am. Math. Soc. 315, **(1989)**.

[15] J. Makhoul and A. El-Jaroudi, “Time Scale Modification in Medium to Low Rate Speech Coding”, Proceding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, Japan, **(1986)** April 7-11.

Authors

Ykhlef Fayçal was born in Blida, Algeria, on July 30, 1979. He graduated from the Saad Dahlab University of Blida, Algeria, in 2002 in Electronic engineering. He received the Master of Science in speech and image processing from the same university in 2005. At present, he is a researcher at the Multimedia Laboratory of “Centre de Développement des Technologies Avancées”, Algiers, Algeria. His research interests include medical applications, signal and speech processing.