Journal of Computer Science 6 (4): 392-405, 2010
ISSN 1549-3636


# Prosody Modification of Standard Arabic Speech Using Combining Synchronous Overlap and Add With Fixed-Synthesis Algorithm and Multi Level Discrete Wavelet Transform

[1]Ykhlef Faycal, [2]Bensebti Mesaoud and [1]Bendaouia Lotfi
[1]Multimedia and System Architecture,
Center of Development of Advanced Technologies, Algiers, Algeria
[2]Department of Electronics, University of Saad Dahleb, Blida, Algeria

**Abstract: Problem statement:** The objective of prosody modification is to change the amplitude, duration and pitch ($F_0$) of speech segments without altering their spectral envelop. Applications are numerous, including, Text-To-Speech synthesis, transformation of voice characteristics and foreign language learning. Several approaches have been developed in the literature to achieve this goal. The main restrictions of these latter are in the modification range, the synthesized speech quality and naturalness of spoken language. The latest research studies provide evidence that the first Formant ($F_1$) and $F_0$ are dependent; suggesting that in order to preserve high quality and naturalness of the speech signal, any change to one of these parameters must be accompanied by a suitable modification of the other. **Approach:** This study introduced a prosody modification method using combining Synchronous Overlap and Add with Fixed-Synthesis (SOLAFS) algorithm and a multi level decomposition based on Discrete Wavelet Transform (DWT) to overcome the limitations cited above. It used Standard Arabic (SA) sounds. For a purpose of comparison, two techniques based on frame by frame processing were proposed. The first one consists in a pitch synchronous processing of the $m^{th}$ approximation level time segments used in SOLAFS algorithm. It was aimed to modify the prosody of the input speech without affecting the spectral envelop. The second one explores the correlation between $F_1$ and $F_0$ in the corresponding approximation level of SA sounded and modified duration and both $F_0$ and $F_1$ scales. It was based on a re-sampling method using FFT interpolation. The use of multi level analysis was aimed to provide independent control over the spectral envelope. In both techniques, the decomposition level depends on the chosen sampling Frequency ($F_S$). $F_0$ marking was based on multi level peaks comparison. Both techniques use an automatic speech classification algorithm based on modified version of the Johnson algorithm. **Results:** The performances of the proposed techniques were evaluated by listening tests using sentences in SA language sampled at an $F_S$ of 16 kHz. It was found that manipulation in the third approximation level of $F_0$ in conjunction with the local $F_1$ improved significantly the naturalness of the modified speech compared to the classical prosody modification. **Conclusion:** This improvement was most suitable for high $F_0$ scales from the fact that speaker generally increases $F_1$ as they increase their $F_0$. Further, the technique can be used in the manipulation of the remained formant structure.

**Key words:** Prosody modification, SOLAFS, PSOLA, intelligibility, naturalness, distortion

## INTRODUCTION

The purpose of prosody modification is to change the amplitude, duration and pitch ($F_0$) of a speech segment without affecting the timbre of the speaker voice. Amplitude modification can be easily accomplished by direct multiplication, but duration and $F_0$ changes are not so straightforward (Ykhlef *et al*., 2008). Applications are numerous, including, Text-To-Speech Synthesis (TTS), transformation of voice characteristics, foreign language learning but also audio monitoring or film/soundtrack post-synchronization (Moulines and Laroche, 1995). For instance, in a TTS system, it is necessary to modify the durations and $F_0$ contours of the basic units in order to incorporate the relevant suprasegmental knowledge in the utterance corresponding to the sequence of these units (Yegnanarayana *et al*., 1994).

**Corresponding Author:** Ykhlef Faycal, Multimedia and System Architecture, Center of Development of Advanced Technologies, Algeria

For this purpose, a number of algorithms have been proposed in the literature. It appears that most of these algorithms can be viewed as slight variations of a small number of basic schemes which are parametric and non-parametric methods (Moulines and Laroche, 1995). In parametric methods, parameters of the speech production model are estimated and used in the modification and the synthesis stages. As an example, Source Filter Modeling (Ykhlef *et al*., 2008) and Harmonic + Noise Model (HNS) (Laroche *et al*., 1993). Non-parametric methods are closely related to the speech production models, but the parameters of these models need not be estimated directly. These methods operate directly on the waveform to incorporate the desired prosody information. As an example, there are Synchronous Overlap and Add (SOLA) and pitch synchronous overlap and add (TD-PSOLA for time domain PSOLA and FDPSOLA frequency domain PSOLA) (Moulines and Laroche, 1995). Other variants of the previous methods have been developed. For example, Linear Prediction-PSOLA (LP-PSOLA) (Edgington and Lowry, 1996) and SOLA with Fixed-Synthesis (SOLAFS) (Hejna *et al.*, 1992). Basically, SOLA and SOLAFS are limited to Time Scale Modification (TSM), however, the TD-PSOLA and LP-PSOLA can modify both time and $F_0$-scales (Jun *et al*., 2009). The FD-PSOLA is used only for $F_0$-scale modification.

The standard implementation of SOLA method consists of dividing the input signal into overlapping segments of fixed length starting at an analysis instant. In the second step, these overlapping segments are shifted according to a given time scaling factor. Then the similarities in the area of the overlap intervals are searched for a discrete-time lag of maximum similarity using a cross correlation function. At this point, the overlapping blocks are weighted by a fade-in and fade-out functions and summed sample by sample (Roucos and Wilgus, 1985). The main drawback of the method is that the amount of overlap of the $m^{th}$ segment between the analysis segments and the output signal varies with the point of highest cross correlations, where more than two segments may overlap in certain regions (Hejna *et al*., 1992). As results, the quality of the output speech will be seriously affected.

To avoid problems related to the use of SOLA, the SOLAFS algorithm was proposed. Since the use of a dynamic synthesis interval is the major problem encountered with SOLA, SOLAFS constrained a fixed synthesis interval. The calculation of the cross-correlation is adapted so that the point of maximum similarity is applied to the analysis signal, rather than on the synthesis one. The modified speech signal using SOLAFS is obtained with high quality, but it depends mainly on the proper choice of the algorithm parameters (Jun *et al*., 2009). The algorithm is well suitable for real time implementations using large time scaling factors (Hejna *et al*., 1992).

The basic PSOLA method consists of deriving pitch synchronous analysis windows, using $F_0$ marks (Rao and Yegnanarayana, 2006; Hung *et al*., 2001). Analysis windows are obtained by direct multiplication of the speech signal with hanning window typically of length two or four pitch periods that are centered around the $F_0$ marks. The $F_0$ modification is achieved by changing the time intervals between the $F_0$ marks. Duration modification is achieved by either repeating or omitting the speech windows.

The PSOLA method suffers from amplitude, spectral and phase distortions due to direct manipulation of the speech signal. The position of $F_0$ marks determines the quality of the modified speech. However, the determination of these positions is not a trivial problem and could be difficult to implement robustly in real-time (Moulines and Laroche, 1995; Schnell *et al*., 2000). In addition, the $F_0$ and time manipulation differs from one region to another (silence, voiced, unvoiced, transient and mixed regions) (Schnell *et al*., 2000), which means that segments classification is a crucial task in this method.

PSOLA approach does not handle well voiced fricatives that are stretched considerably because of added buzziness (repeating segments induce periodicity at the high frequency that was not present in the original signal) or attenuation of the aspirated component (Hung *et al*., 2001). Further, transient signals (stops sounds) should not be modified because their dilatation is perceptively disagreeable (Peeters, 1998). The marks in the unvoiced segments are generally equidistantly placed (Peeters, 1998; Hung *et al*., 2001), they are used in duration modifications by repeating or omitting the corresponding segments according to a time scaling factor. Buzziness appears in unvoiced sounds when larger time scaling factors values are applied, due to the regular repetition of identical input segment. It is preferable to use other approaches in these regions such as granular synthesis technique (Schnell *et al*., 2000). LP-PSOLA is theoretically more suitable for $F_0$-scale modification, because it provide independent control over the spectral envelope for synthesis. It is based on the principle of residual excited vocoders (Moulines and Laroche, 1995; Rao and Yegnanarayana, 2006).

Several approaches are available in the literature for time-scale, $F_0$-scale or both time and $F_0$-scales

modification (Rao and Yegnanarayana, 2006; Muralishankar *et al*., 2004; Kumar and Jain, 2006). They differ in the amount of computation needed, modification range and the modified speech quality. Muralishankar *et al*. (2004) and Rao and Yegnanarayana (2006) respectively, perform $F_0$-scale and $F_0$ time-scale modification in the residual domain. The residual manipulation is likely to induce less distortion in the speech signal. The $F_0$ marking in (Rao and Yegnanarayana, 2006) was based on the instants of significant excitations using the property of average group-delay of minimum phase signals. Muralishankar *et al*. (2004), use the autocorrelation function for $F_0$ marking. The $F_0$-scale modification is done in both approaches by interpolating the residual signal using the Discrete Cosine Transform (DCT) (Rao and Yegnanarayana, 2006). Duration modification in (Rao and Yegnanarayana, 2006) is performed by repeating or omitting the speech segments. Kumar and Jain (2006) perform $F_0$-scale modification using the maximum modulus of Continuous Wavelet Transform (CWT). The results of $F_0$ shifting seem to be more promising in the residual domain by providing independent control over the spectral envelope.

Although the large existing prosody manipulation methods have achieved high level of intelligibility, this problem still attracts the attention of researchers, especially because of the limitations on the modification range and the naturalness of the resulting speech. Current methods promise reasonable voice quality and naturalness for not more than 1 octave modification rate, remarking that the quality degradation is noticeable for large scales. One of the reasons for this degradation is the assumption of a constant magnitude spectrum. However, the opposite has been shown to be true (Kain and Stylianou, 2000). An increase of $F_0$ was observed to cause a vowel boundary shift or a vowel height change (Hirahara, 1988). Additionally, an analytical study showed that the overlap of natural vowels in formant scatter plots can be normalized using $F_0$ information (Kain and Stylianou, 2000).

Syrdal and Steele (1985) provide evidence that the $F_1$ and $F_0$ are dependent; suggesting that in order to preserve a high quality and naturalness of the speech signal, any change to one of these parameters must be accompanied by a suitable modification of the other. In addition, they found that speakers generally increase $F_1$ as they increase their $F_0$. Such an approach is adopted in (Tanaka and Abe, 1997), where the spectral envelope is modified according to $F_0$ modification by a Vector Quantization (VQ) codebook mapping technique.

Based on these assumptions, in this paper, two techniques for prosody modification of Standard Arabic (SA) speech are proposed to overcome the limitations on the modification range, the synthesized speech quality, intelligibility and naturalness. The proposed techniques are based on combined version between the well known SOLAFS algorithm used for TSM of speech and a multi level decomposition based on Discrete Wavelet Transform (DWT). They are able to manipulate both time and $F_0$-scales for large values. The use of multi level analysis is aimed to provide independent control over the spectral envelope.

The first technique consists of dividing pitch synchronous analysis windows from the $m^{th}$ approximation level ($A_{Lm}$) time segments used in SOLAFS algorithm. It is named $A_{Lm}$-PSOLAFS. The $F_0$ of the input speech is shifted using overlap and add principal according to $F_0$ marks without affecting $F_1$ and the remainder spectral envelop. Duration modification is performed by SOLAFS algorithm in order to reduce buzziness that could appear in unvoiced sounds using the overlap and add principal.

The second one aims at modifying duration and both local $F_0$ and $F_1$ of the instantaneous analysis segments used in SOLAFS algorithm at a multi level decomposition, and exactly at their frequency band existence namely from 0 to 1 kHz for the SA sounds. It explores the correlation between these two entities in their frequency bandwidth. The $F_0$ and $F_1$ manipulations are done by re-sampling the time portions of the $m^{th}$ level approximation according to $F_0$ marks. Thus, $F_0$ and $F_1$ are only scaled on the desired frequency band. This technique is named $m^{th}$ level approximation pitch re-sampling SOLAFS, $A_{Lm}$PR-SOLAFS**.**

Both techniques do not need markers in the unvoiced region for time scaling operation since it is performed using the SOLAFS algorithm. However, segments classifications are needed and are based on a modified Johnson algorithm to detect silence (suku:n), voiced, unvoiced, and mixed SA sounds.

The decomposition level depends on the chosen sampling frequency ($F_S$). In this study, it is fixed to 16 kHz, thus the desired frequency bandwidth between 0 to 1 kHz, corresponds to the third approximation level and it is retained for both techniques.

The instants of excitation (marks) are extracted at local instants based on the $m^{th}$ approximation level peaks comparison and allocations, according to the local pitch period.

## MATERIALS AND METHODS

This text begins by presenting the main characteristics of the SA sounds especially its formantic structure. Then it presents the technique used for sound classification. In the next point, it explains the main steps of the SOLAFS algorithm used in TSM and focus for the most part on the practical implementation. Finally, the automatic methods of $F_0$ marking and modification of SA sounds proposed in this study and used for both techniques $AL_3$-PSOLAFS and $AL_3$PR-SOLAFS are described.

**Standard Arabic sounds description:** Arabic is a Semitic language and, is one of the oldest languages in the world. Currently, it is one of the most spoken languages in term of number of speakers (Alotaibi and Hussain, 2009). There are several Arabic dialects over the world of Arabic countries. This study focuses on the characteristics of the SA sounds.

The SA is composed of twenty eight phonetically distinct consonants phonemes, 3 short vowels (a, u, i) which contrast phonetically with their long counterparts (a:, u:, i:) and six corresponding variants of the short and the long vowels in emphatic context (Aissiou and Guerti, 2009). The silence in SA is called (suku:n). All the Arabic vowels are oral and fully voiced. They can be nasalized in the nasal consonants context. The difference between short and long ones is approximately double. The relative duration of the short vowels is from 100-150 ms. With the long ones, it is from 225-350 ms (Al-Ani and Salman, 1970). The characteristics that form a vowel are relatively more prominent and stable than consonants. Generally, consonants have less energy than vowels.

The SA phonemes classification is based on physiological speech parameters that consist on both horizontal and vertical places and various manners of articulations.

The twenty eight consonants are classified physiologically as:

- Thirteen fricatives: unvoiced ([θ], [ʕ], [X], [s], [ʃ], [S]*, [f]). voiced ([z], [D]*,[ε], [γ], [h])
- Eight stops: unvoiced ([?], [k], [q], [t]). voiced ([d], [b], [t*], [d*])
- Two nasals: voiced ([m], [n])
- Three sonorants: voiced ([y], [w], [r])
- One thrill: voiced [l]
- One affricate: voiced [dZ]

Where, (*) denotes an emphatic consonant.

The phonetic frame work in this study is that of the International Phonetic Alphabet (IPA) as used within the California University Phonological Segment Inventory Database (UPSID) (Aissiou and Guerti, 2009).

The SA is characterized by three phonetic phenomena which are the presence of the emphatic consonants, the geminate ones and by the presence of the glottal, pharyngeal, velar and uvular ones called back consonants. It possesses eight back phonemes. On the neighboring emphasis consonants, all Arabic vowels are strongly influenced. The germination corresponds to the consonants production with intensive energy concentration. At a phonological level, it is important to note that in Arabic the geminated segments can never occur at the beginning of the word. The entire Arabic consonants can be geminated except the glottal stop consonant [?] (Aissiou and Guerti, 2009).

It is well known that the spectral maxima and the fundamental frequency of voiced speech, i.e., formants and $F_0$ respectively play an important role in the identification of speech sounds and gender. The spectral information is used in analysis, synthesis and recognition of speech signal (Hung *et al*., 2001). In general, $F_0$ varies from 70-250 Hz for men, 150-400 Hz for women and from 200-600 Hz for children which is a large bandwidth of variation for the majority of the spoken language (Ykhlef *et al*., 2008). On the other hand, there is no reference study in the literature about formants evaluation of SA sounds in continuous speech. However, several studies have been reported for evaluating the formantic assessment of voiced SA sounds (Al-Ani and Salman 1970; Newman and Verhoeven, 2002; Mansour, 1998; Ghazeli, 1979; Belkaid, 1984; Abou, 1994).

Newman and Verhoeven (2002) present a comparative study of several formant assessments of SA vowels where they estimate the average values of the first and second formant of the six vowels. For our purpose, we have estimated the two first average formantic frequencies of the SA sounds for male and female speakers using a corpus composed of several sentences in SA continues speech. The corpus is spoken by native Arabic speakers in low noisy environments. The phonetic phenomena of SA are taken into account in this corpus. The Fs is fixed to 16 kHz. The formants trajectory is estimated using Linear Prediction Coding (LPC) spectral envelop based on Consonant-Vowel (CV) context of the six vowels of SA with conjunction to all consonants. The estimated average values are in agreement with the results found in (Newman and Verhoeven, 2002). The mean values of all studies are reported in Table 1. The average value of $F_1$ (and the expected modified ones for frequency scales less or equal to 2) is always included in the bandwidth 0 to 1 kHz.

Table 1: Mean values of SA vowel frequencies (Hz)

| i: | | i | | u: | | u | | a: | | a | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ | $F_1$ | $F_2$ |
| 332 | 2115 | 395.8 | 1872 | 338 | 992.8 | 393.3 | 1053.9 | 473.1 | 1180 | 460.5 | 1300 |

This result will be used in $F_0$-scale manipulation in order to shift the spectral envelop in the frequency band existence of $F_1$ and $F_0$ of voiced SA sounds for high naturalness $F_0$-scale modification.

**Sounds classification by DWT:** Sound classification is defined as the process of finding the boundaries in a natural language between words, syllables or phonemes. The sound classification is an important problem for number of fields in speech processing. However, in order to perform classification, one must consider the acoustic characteristics of the spoken language under study (Al-Manie *et al*., 2009). Consequently, an efficient, accurate, automatic and simple technique is needed to accomplish this objective. The traditional approach to tackle this problem is manual segmentation of speech, usually performed by specialized phoneticians. This method is based on listening and visual judgment on required boundaries (Lee *et al*., 2003). However, in this application, an automatic technique able to classify the main acoustic regions of SA speech, namely/suku:n/, unvoiced (unvoiced fricatives and stops), mixed (voiced fricatives) and fully voiced sounds (vowels, nasals, sonorants and affricate) is needed in order to modify only $F_0$ at the voiced parts. Furthermore, in continues speech, it is very difficult to detect with accuracy all the acoustic phenomenon of a given spoken language.

Many techniques were reported in the literature for automatic sound classification (Aissiou and Guerti, 2009; Johnson, 1996). The voiced/unvoiced classifications have proved to be well suitable for real time application. In essence, these techniques are based on energy and zeros crossing rate or autocorrelation functions. The difficulties in using these techniques are in setting the appropriate thresholds. A practical solution was proposed in (Ykhlef *et al*., 2008) for automatic determination of thresholds for both zeros crossing versus energy computation and autocorrelation function and has given good results. Further, it has been proved to be suitable for source filter modeling applications (Ykhlef *et al*., 2008). However, for high quality prosody modification, a technique for clearly detecting silence, voiced, unvoiced and especially mixed sounds is needed to classify the segments of the speech signal. The previous two techniques show some difficulties in automatic detecting of mixed sounds.

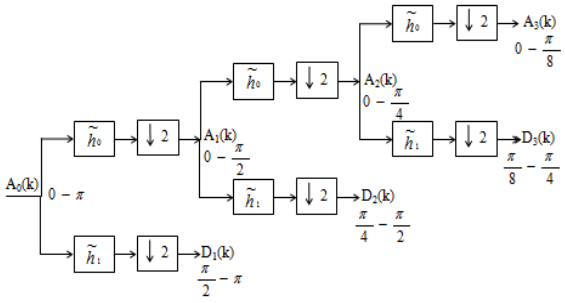


Fig. 1: Three stage multi level decomposition

Since in this study $F_0$ shifting is based on DWT, the use of Johnson (1996) algorithm which is able to detect clearly voiced, unvoiced and mixed sounds with some modification in the algorithm parameters according to the SA sounds is preferred. Further, the use of the algorithm with conjunction to an automatic /suku:n/ detector based on entropy computation of the time segments is also proposed.

Wavelet Transform (WT) has been intensively used in various fields of signal processing. Unlike Fourier, cosine and sine transforms, the WT has the advantage of using variable size time-windows for different frequency bands. This results in localization for both time and frequency simultaneously. Consequently, WT is a powerful tool for modeling non-stationary signals as speech that exhibit slow temporal variations at low frequencies and abrupt temporal changes at high frequencies. The use of WT in speech classification and pitch marking was guided by these proprieties. An efficient way to implement this scheme using filters was developed in 1988 by Mallat (1989).

The Mallat algorithm is in fact a classical scheme known by the signal processing community as a two-channel subband coder (Strang and Nguyen, 1996). This practical filtering algorithm yields a Fast Wavelet Transform (FWT) or what is called multi level analysis. A typical multi-level analysis for the DWT is depicted in Fig. 1. The decomposition filters, $\tilde{h}_0$ and $\tilde{h}_1$ were adopted to split the input speech signal into two overlapped and equally spaced frequency bands at first level, where $\tilde{h}_0$ is low-passed and $\tilde{h}_1$ is high-passed. Then, decimate each band by a factor 2, such that spectrum of each band is expanded to fill up the full frequency scale (Mallat, 1989). The frequency bands are

shown in Fig. 1 as normalized form, where π denote the Nyquist frequency. Consequently, the signal length of the low pass and high pass output filter is only half of the original one. The splitting, filtering and decimation can be repeated on the scaling coefficients to give the idea of multi level analysis. This results in a high frequency resolution in low bands and low frequency resolution in high bands. The scaling coefficients are given as follows (Popescu, 2001):

$$A_j(k) = \sum_{\ell} A_{j-1}(\ell)\tilde{h}_0(2k-\ell) \quad (1)$$

$$D_j(k) = \sum_{\ell} A_{j-1}(\ell)\tilde{h}_1(2k-\ell) \quad (2)$$

and are called respectively the $j^{th}$ approximations and details coefficients, where $k$ and $\ell \in Z$.

The approximations coefficients are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. The decomposition filter coefficients play a crucial role in a given DWT and have to satisfy orthonormalites and a certain degree of regularity. Several set of filter coefficients can be found in (Strang and Nguyen, 1996). This application uses Daubechies filters which give best results compared with other wavelets.

In the classification process, the SA speech signal is fragmented into m overlapping segments, of fixed length N; each segment being $S_a$ samples apart, where $S_a$ represents the step size. The algorithm starts by computing the nonnormalized Shannon entropy (En) of the $m^{th}$ segment (xan) as shown in Eq. 3:

$$En(xan) = \sum_{i=0}^{N-1} xan(i)^2 \log(xan(i)^2) \quad (3)$$

where, $xan_i$ denote the $i^{th}$ sample of the xan segment.

Assuming that, the discrete samples of the speech signal are normalized between -1 and 1. The $m^{th}$ segment is classified as /suku:n/ if its corresponding entropy does not exceed 0.1, which means 1% of the normalized amplitude. In the second step, we compute the energy (En) of the level one approximation and detail coefficients as shown in Eq. 4:

$$En(A_j) = \sum_{i=0}^{M-1} A_j(i)^2 \quad (4)$$

Where:

$A_j$ = The $j^{th}$ approximation coefficients of the xan segment

M = Its length which is supposed to be half of N

The details energy is obtained by replacing $A_j$ by $D_j$ in Eq. 4.

Assuming that, the whole energy of the $m^{th}$ segment is equal to the sum of the approximation and detail coefficients energies, which represents 100%. If the percentage of the energy concentrated in level one approximation coefficients is less than 30%, the segment is classified as unvoiced. If the percentage of the energy in level one approximation coefficients is between 30 and 97%, the segment is classified as mixed one. Finally, above 97% the segment is classified as voiced. The thresholds are obtained with practical tests using the SA corpus. The processing was based on overlapped segments as explained above. The choice of the Daubechies filters order is fixed in this study at 15, which represents an average value suitable for DWT classification. Figure 2 represents the classification results obtained by using the SA context /ahi/ extracted from the word /wahia/ (she is). The speech signal (speech) is represented in red color and the continues classification (class) is in black. The first segment (1) which is equal to one represents the interval of the first vowel [a].

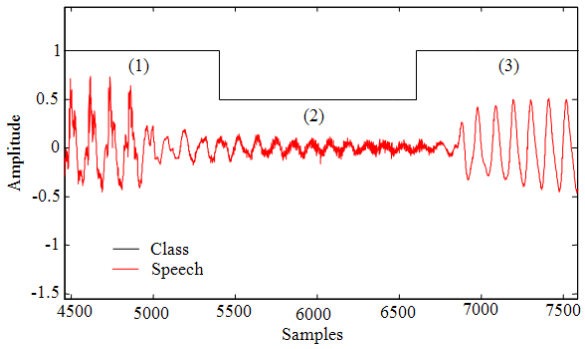


Fig. 2: Classification of the context /ahi/ from the word /wahia/

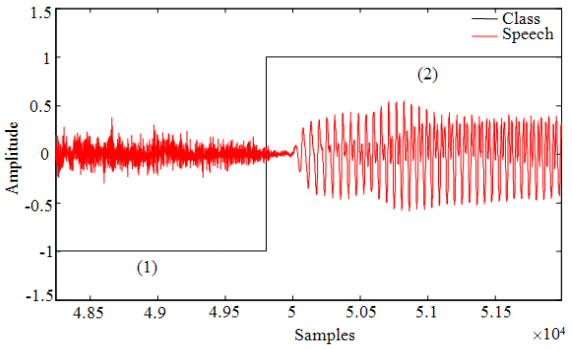


Fig. 3: Classification of the context /sa:/ from the word /tusa: ɛi du/

The second one (2), which is equal in this case to 0.5, represents the interval of the voiced fricative [h]. Acoustically, it denotes a mixed portion between voiced and unvoiced sounds. The latter segment represents the second vowel [i] from the previews context. Figure 3 represents the classification of the unvoiced fricative [s] (1) and the vowel [a:] (2) from the context /sa:/, extracted from the word /tusa: ɛidu/ (she help). In Fig. 2 and 3, the segment length and step size are fixed respectively to 560 and 160 samples for an $F_s$ equals to 16 kHz.

As a conclusion, it is verified practically that the classification of Arabic sounds using the SA corpus is well established with minimum errors classification.

**Time scale modification:** The TSM of speech signal is performed using the SOLAFS algorithm. Thus, the main steps of the algorithm are explained in details. For the most part, this study focuses on the practical implementation and the main parameters used for time and $F_0$-scales modification. More details about the algorithm can be found in (Hejna *et al*., 1992).

SOLAFS is an improvement of the SOLA method described in (Roucos and Wilgus, 1985). It uses a fixed synthesis interval to constraint the problem of segments overlap related to the use of a dynamic synthesis interval by the traditional method. It is more efficient then SOLA and provides greater flexibility in the choice of parameters. Further, the computational requirements and costs are simplified. It provides a robust TSM compared to the SOLA method.

In accordance with SOLAFS algorithm described in (Hejna *et al*., 1992), blocks of the input speech signal referred as analysis segments (Eq. 4), are taken at an average rate of $S_a$ with each starting allowed to vary within limits and an output signal is reconstructed using a fixed inter block offset $S_s$, i.e., the duration of overlap with the existing signal with each segment to be added is fixed. This is done by searching for segments of the input signal near the target starting position $mS_a$ which are similar to the portion of the output signal that will overlap when constructing the output signal (Fig. 4).

The analysis segments are chosen as follow:

$$x_m(n) = \begin{cases} x(mS_a + k_m + n) & n = 0..., W-1 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Where:
W = The segment length
$k_m$ = The number of samples of shift
x(n) = The sampled speech signal

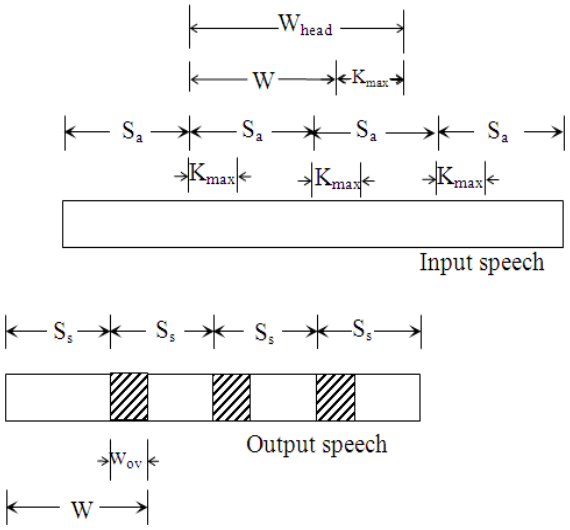


Fig. 4: Illustration of SOLAFS (time compression)

The output speech y(n) is reconstructed recursively with the following:

$$y(mS_s + n) \leftarrow \begin{cases} b(n)y(mS_s + n) + (1 - b(n))x_m(n) & n = 0..., W_{ov} - 1 \\ x_m(n) & n = W_{ov}..., W - 1 \end{cases} \quad (6)$$

Where:
$W_{ov} = W\text{-}S_s$ = The number of points in the overlap region
b(n) = A weighting function (fade in) such that $0 \le b(n) \le 1$

As it can be seen in (Eq. 5), $k_m$ affects the starting position of the input segments. It is determined by similarity measure using the normalized cross correlation between the input and the output in the region of overlap defined in (Eq. 7a and b):

$$k_m \leftarrow \begin{cases} t_m = k_{m-1} + (S_s - S_a) & \text{if } 0 \le t_m \le K_{max} \\ \max\limits_{0 \le k \le K_{max}} R^m{}_{xy}(k) & \text{otherwise} \end{cases} \quad (7a)$$

Where:

$$R^m{}_{xy}(k) = \frac{\sum_{i=0}^{W_{ov}-1} x(mS_a + k + i)y(mS_s + i)}{[\sum_{i=0}^{W_{ov}-1} x^2(mS_a + k + i) \sum_{i=0}^{W_{ov}-1} y^2(mS_s + i)]^{0.5}} \quad (7b)$$

$0 \le k \le K_{max}$

and $K_{max}$ is the maximum allowable shift from the initial starting position of the analysis segment. The variable $k_m$ is initialized by zero ($k_0 = 0$).

Basically, the implementation of SOLAFS algorithm starts by extracting the $m^{th}$ head vector $x_{head}$ of length $W_{head} = W + K_{max}$ each $S_a$ instant. This vector is obtained from the input speech signal in order to extract the $m^{th}$ analysis segments $x_m(n)$ of fixed length W (Fig. 4).

According to $k_m$, only W samples from the analysis input head vector $x_{head}$ will be allocated to the analysis $x_m(n)$ segment and will be used in the output reconstruction. The head vector is very important in our technique and both $F_0$ marking and shifting rely on it. The time scale factor is defined as:

$$F = \frac{S_a}{S_S} \quad (8)$$

When F is unity, the speech is unchanged, when F is greater than one, the speech is time compressed and when it is less than one, the speech is expanded. $K_{max}$ must be chosen to be larger than the largest expected pitch period in the input speech to avoid $F_0$ fracturing. In this application and for fixed $F_s$ of 16 kHz, the preferred choice of $K_{max}$ is set to 200 samples. The segments length and overlap respectively W and $W_{ov}$ are optimized by listening tests as 400 and 200 samples.

**Pitch manipulation:** At this level, the automatic $F_0$ marking and modification of SA sounds proposed in this study and used for both techniques $AL_3$-PSOLAFS and $AL_3$PR-SOLAFS will be described. In fact, the proposed techniques rely on the reconstruction of the $j^{th}$ approximation level of the $m^{th}$ head vector $x_{head}$. Thus, at first, the choice of multi level decomposition based on DWT in $F_0$-scale modification will be justified. Then, in the second point, the main steps for reconstructing the approximations and details themselves from their coefficient vectors are given, followed by a technical description of $F_0$ marking. Finally, a detailed description of the proposed $F_0$ shifting techniques is given.

**Pitch marking based on multi level decomposition:** $F_0$ marking is an important task in speech modification. The quality of the modified speech depends mainly on the positioning of the corresponding marks. These latter must be positioned pitch synchronously with the same place within the period (Laprie and Colotte, 1998). Glottal Closure Instants (GCI), i.e., the point where the vocal track system is excited, is generally chosen as the place of marks. Although several algorithms for determining GCI exist (Cheng and Shaughnessy, 1989; Moulines and Francesco, 1989), results are not sufficiently precise to perform automatic pitch marking. Some human assistance is thus required to correct errors. As an example, for vowels GCI are supposed to fall on the first negative peak of each period and pitch marking should select among the minima those which correspond to GCIs. However, considering the nature of speech signal, it may be wiser to select maxima instead of minima when it is found that $F_0$ marks are more reliable on that way (Laprie and Colotte, 1998). Furthermore, $F_0$ marking and scaling seem to be more promising by providing independent control over the spectral envelope. For example, marking and modification of local pitches in the LPC residual domain.

In this application, the independent control over the spectral envelope (second, third, fourth … formants) is performed in the Wavelet domain. $F_0$ marking and shifting are manipulated in the frequency band existence of the original $F_0$ and $F_1$ of voiced SA sounds. As mentioned earlier, these two latter are included in the bandwidth between 0 and 1 kHz. Since speech signals are sampled at an $F_s$ of 16 kHz, this frequency band relies within the $A_3$ approximation level of the original signal. Thus, the remained details ($D_3$ (1-2 kHz), $D_2$ (2-4 kHz) and $D_1$ (4-8 kHz)) will be unchanged.

The proposed $F_0$ marking technique relies on the reconstruction of the $j^{th}$ approximation level of the head vector $x_{head}$. The approximations at the $j^{th}$ level are characterized by the same length as the input speech signal. However, the approximation coefficients length obtained using the structure in Fig. 1 has half length of the original one. In fact, it is possible to reconstruct the approximations and details themselves from their coefficient vectors using the reverse process of decomposition as described in Fig. 5.

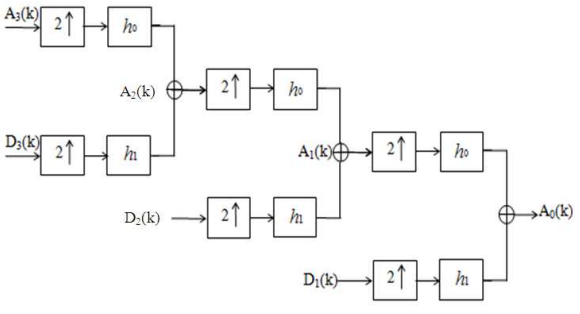


Fig. 5: Three stage multi level reconstruction

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. 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 $\tilde{h}_1$. As given by (Popescu, 2001), $(\tilde{h}_0(\ell))_{\ell\in Z} = (h^*_0(-\ell))_{\ell\in Z}$ and $(\tilde{h}_1(\ell))_{\ell\in Z} = (h^*_1(-\ell))_{\ell\in Z}$ correspond respectively to the impulse responses of the analysis and synthesis filters. The reconstruction formula is given as follows:

$$A_{j-1}(k) = \sum_{\ell} h_0(k-2\ell)A_j(\ell) + \sum_{\ell} h_1(k-2\ell)D_j(\ell) \quad (9)$$

To achieve perfect reconstruction without aliasing, the decomposition and reconstruction filters have to satisfy certain conditions according to the wavelet family filters. A technical discussion of how to design these filters is available on (Strang and Nguyen, 1996). The approximation and detail signals which have the same length as the input speech segment are named respectively $AL_j(k)$ and $DL_j(k)$ in order to defer them from their coefficients $A_j(k)$ and $D_j(k)$.

The $j^{th}$ level approximations are reconstructed from their coefficients by replacing the detail coefficients at each level by a zeros vector with the same length as the $j^{th}$ detail coefficients. The reconstructed signal corresponds to the desired $j^{th}$ level approximations. Depending on the impulse response length of the decomposition–reconstruction filters, the reconstructed signal must be left shifted in order to compensate the lag introduced by Finite Impulse Responses (FIR) filtering of the structures in Fig. 1 and 5. Similarly, we can reconstruct the $j^{th}$ level detail using the identical process. In this way, the reconstructed details and approximations are true constituents of the original signal. We find when we combine them by simple sample to sample adding that it does correspond to the original speech signal (Eq. 10):

$$xan(n) = A_{LR}(n) + \sum_{j=1}^{R} D_{Lj}(n) \quad (10)$$

Where:
xan(n) = The $n^{th}$ sample of the speech segment
R = The number of decomposition level

The $F_0$ marking and shifting method used in both techniques ($AL_3$-PSOLAFS and $AL_3$PR-SOLAFS) follow the same steps as in SOLAFS algorithm. It starts by the fragmentation of the input signal into m overlapping segments of fixed length $W_{head}$, each segment being $S_a$ samples apart, where the $m^{th}$ segment corresponds to the input head vector $x_{head}(n)$ used in SOLAFS algorithm.

The marking is based on the comparison of the $AL_3$ and ($AL_4$ or $AL_5$ or $AL_6$, according to the local pitch period) approximations peaks of the $m^{th}$ head vector $x_{head}$. The optimal choice of $F_0$ marks is obtained by affecting all the peaks instants found in ($AL_4$ or $AL_5$ or $AL_6$) to the nearest $AL_3$ ones which correspond finally to the local head vector marks. The $AL_4$, $AL_5$ and $AL_6$ frequency bandwidths are defined respectively for an $F_S$ of 16 kHz from 0-500, 0-250 and from 0-125 Hz. The difference between two successive peaks in the highest level corresponds to the local pitch period. The choice of one AL from the set of approximations ($AL_4$ or $AL_5$ or $AL_6$) in the comparison will be done according to the local pitch period, where the nearest left bandwidth of the approximations under study to the local $F_0$ will be taken as optimal approximation. The selected approximation in the comparison gives the best $F_0$ marking instants. Thus, an initial estimation of an average $F_0$ value should be performed on the $AL_3$ approximation by using for example the autocorrelation function. This comparison aims at reducing the erroneous marks that could appear in certain cases when an acoustic irregularity is occurring. To understand the process, an illustrative example is given as follows. Assuming a mal speaker with an average local pitch of 192 Hz, $AL_5$ is the best choice from the remainder approximations ($AL_6$ and $AL_4$) and will be used in the comparison because 250 Hz is the nearest side to the local $F_0$. Further, peaks bellow a threshold of 1% from the maximum amplitude of the input speech signal are not taken in consideration because it may leads to erroneous marks. Practically, highest Wavelet orders lead to optimal marks positions. Nevertheless, it has been found that an average order of 15 gives good results.

Figure 6 represents $F_0$ marking of a voiced speech segment chosen arbitrary. The signal in blue color denote the $AL_3$ approximation and in red the corresponding $AL_5$. Marks are reported as black impulses in the local maximum of the $AL_3$ approximation. As it is shown, the peaks instants in $AL_5$ are automatically allocated to the nearest ones in the $AL_3$ approximation.

The process is repeated for all voiced speech segments until all the speech duration is reached.

**Pitch modifications of $AL_3$ approximation:** After obtaining the $AL_3$ analytic $F_0$ marks of the input head vector $x_{head}$ under study, the next step is to derive the synthesis ones for a desired pitch period modification factor $\beta$.

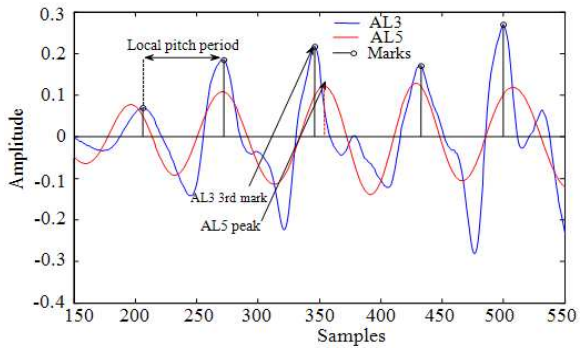


Fig. 6: Pitch marking of voiced segment

The process is performed for each segment using Eq. 11:

$$Ms(k+1) = Ms(k) + \beta \times P(Ma(k)) \quad (11)$$

where, Ma(k) and Ms(k) denote respectively the $k^{th}$ analytic and synthetic $F_0$ marks and P(Ma(k)) the $k^{th}$ pitch period obtained between two successive analytic marks.

In a given voiced region, the initial value of synthetic mark is equal to the initial analytic one. The choice of the corresponding analysis portions or windows (portions of length P(Ma(k)) for $AL_3$PR-SOLAFS or window of length 2*P(Ma(k) for $AL_3$-PSOLAFS) used in the modification process is performed by minimizing the time distance between the $k^{th}$ synthetic mark and all the analytic marks of the segment under study as shown in Eq. 12:

$$\min_k \left| Ms(k) - Ma \right| \quad (12)$$

where, k depends on the number of synthetic marks in the segment under study.

Thus, according to this minimization, some portions will be repeated or deleted. The $F_0$ modification in the proposed $AL_3$PR-SOLAFS technique is scaled by re-sampling the $AL_3$ approximation portions obtained between two successive analytic marks in order to fit the new desired pitch period duration for the obtained synthetic marks. The Fast Interpolation Algorithm based on FFT described in (Prasad and Satyanarayana, 1986) is used. The $F_0$ shift is accompanied by $F_1$ shift, from the fact that we are re-sampling the frequency bandwidth of both $F_0$ and $F_1$.

Practically, the re-sampling of the $AL_3$ portions introduces high frequency discontinuities in the points of connection of the reconstructed input head vector $x_{head}$. The problem is solved by passing each $AL_3$ approximation after processing through a fifth order median filter. After the modification of $F_0$ and $F_1$ in the $AL_3$ approximation level, the modified head vector will be obtained by simply adding the three details and the modified approximation using Eq. 10. Then, according to the obtained shift $k_m$ (as explained previously), only W sample which correspond to the modified $m^{th}$ analysis segment $x_m(n)$ will be used in the reconstruction of the modified speech. The process is repeated for all input speech segments ($x_{head}$) using the same steps as in TSM except fixing F equal to one for no duration modification (it can be changed in the case of modification of both pitch and duration).

For the second technique $AL_3$-PSOLAFS, $F_0$-scale modification is performed in the $AL_3$ approximation level based on pitch synchronous overlap and adds using hanning window of length two pitch periods. We have used the formula given bellow:

$$\tilde{A}_{L3}(n) = \frac{\sum_i A_{L3i}(n) h_i(n - Ms(i))}{\sum_i h_i(n - Ms(j))} \quad (13)$$

where, $\tilde{A}_{L3}$ correspond to the reconstructed third approximation level of the head vector and h the hanning window used.

The choice of the corresponding analysis window in the modification process is performed by minimizing the time distance as given by Eq. 12.

The modified head vector will be obtained by simply adding the remained three details with the modified approximation.

The $AL_3$–PSOLAFS technique is performed using $F_0$ markers as explained above. Then, the same as in the first technique, depending on $k_m$, only W sample of the $m^{th}$ analysis segment $x_m(n)$ will used in the reconstruction of the modified speech. The process is repeated for all input speech segments using SOLAFS algorithm by fixing F equals to one for no duration modification. This latter can be changed in the case of modification of $F_0$ and duration together.

## RESULTS

Experiments were carried out to investigate the potential advantages of using the proposed prosody modification techniques to enhance the naturalness of the modified speech. First, a set of five continues sentences from our SA speech corpus that contain all acoustic representation and characteristics of the

language under study are selected. These selected sentences contain male and female speakers. The $F_s$ frequency is still fixed to 16 kHz. The performance evaluation of prosody modification is performed using an informal subjective evaluation of Mean Opinion Score (MOS) by a group of six researchers. The tests were carried out in the laboratory by playing the modified speech through loudspeakers. The speech quality assessments include intelligibility, distortion and naturalness. All are ranged from one to five with five meanings the best. Speech intelligibility is related to the amount of speech items that is recognized correctly, while speech distortion is related to the quality of a reproduce speech signal with respect to the amount of audible distortions. Speech naturalness is related to the degree of reproduced speech that could be normally spoken by ordinary human being.

For duration modification, the proper choice of SOLAFS algorithm parameters was fixed as follows:

- The maximum allowable shift $K_{max}$ is fixed to 200 samples
- The segments length $x_m(n)$ is 400 samples
- The overlap region $W_{ov}$ is fixed to 200 samples

For each sentence, the duration was modified by factors 0.5, 0.7, 0.9, 1.2, 1.5, 1.7 and 2. We have been interested in evaluating the speech duration to intelligibility and naturalness of the sentences under study. In the whole, TSM using SOLAFS algorithm, generates a modified speech with high quality and there is no need to evaluate the speech distortion since this criterion is taken in consideration in the choice of the algorithm parameters.

The MOSs for each duration modification factors are given in Table 2.

The evaluation of the pitch period modification includes distortion tests from the fact that high scales can degrade the quality of the modified speech in these kinds of algorithms. The pitch periods was modified by factors 0.5, 0.7, 0.9, 1.2, 1.5, 1.7, 1.9 and 2.2. The last factor is used only for the $AL_3$PR-SOLAFS technique in order to estimate the quality assessments for large modifications. The performances of the proposed techniques $AL_3$PR-SOLAFS (1) and $AL_3$-PSOLAFS (2) for each pitch period modification factors are given in Table 3.

Table 2: MOSs values for time scale modification

| F | Intelligibility | Naturalness |
|---|---|---|
| 0.5 | 4.67 | 4.33 |
| 0.7 | 4.67 | 4.50 |
| 0.9 | 5.00 | 5.00 |
| 1.2 | 4.83 | 4.66 |
| 1.5 | 4.66 | 4.50 |
| 1.7 | 4.33 | 4.16 |
| 2 | 3.83 | 3.66 |

A random stationary interval from the original (org, in blue) and modified (mod, in red) speeches scaled by $AL_3$PR-SOLAFS technique are plotted in Fig. 7.

Table 3: MOSs values for pitch period modification

| | Intelligibility | | Naturalness | | Distortion | |
|---|---|---|---|---|---|---|
| β | (1) | (2) | (1) | (2) | (1) | (2) |
| 0.5 | 3.83 | 3.83 | 3.83 | 2.50 | 3.83 | 3.66 |
| 0.7 | 4.83 | 4.83 | 4.43 | 4.00 | 4.16 | 4.00 |
| 0.9 | 5.00 | 4.83 | 4.83 | 4.50 | 5.00 | 4.83 |
| 1.2 | 4.66 | 4.66 | 4.33 | 4.50 | 4.16 | 4.00 |
| 1.5 | 4.16 | 4.31 | 3.83 | 4.00 | 4.50 | 4.16 |
| 1.7 | 4.00 | 4.17 | 3.50 | 3.83 | 4.16 | 3.83 |
| 1.9 | 3.50 | 3.97 | 3.16 | 3.16 | 4.43 | 1.50 |
| 2.2 | 3.33 | - | 2.50 | - | 2.33 | - |

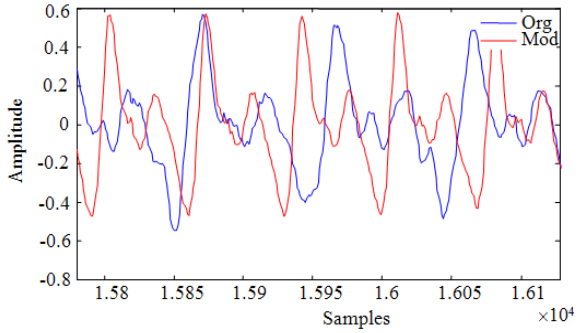


Fig. 7: Temporal representation of a modified and original portion of speech by $A_{L3}$PR-SOLAFS for pitch period modification factor β = 0.7

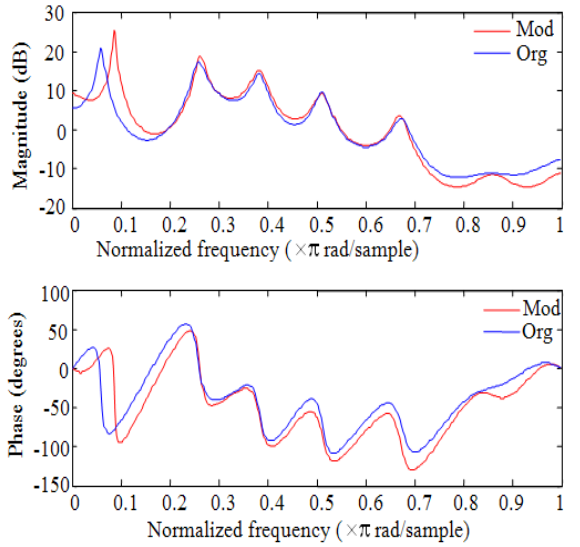


Fig. 8: Spectral envelop (magnitude and phase) of a modified and original portion by $A_{L3}$PR-SOLAFS for pitch period modification factor β = 0.7

The pitch period modification factor used is equal to 0.7. The corresponding spectral envelop (amplitude (or magnitude in dB) and phase) obtained using LPC analysis of order 19 are plotted in Fig. 8.

## DISCUSSION

The evaluation results for TSM (Table 2) indicate that the setting parameters of the SOLAFS algorithm are well fixed. They give good naturalness and intelligibility except for high compression ratios were the speech intelligibility is slightly degraded compared to the whole cases due to the fast replay of the utterances under study.

The $F_0$-scales modification is presented for high and low pitch period modification as reported in Table 3. For pitch period modification factors between (0.9 and 1.5), both techniques seem to have good performances in intelligibility, distortion and naturalness assessments. Beyond 1.5 both techniques are degraded in naturalness and distortion assessments by incrementing the pitch period modification factors which correspond to decrementing in $F_0$ range. The naturalness of the output speech is not degraded similarly for both techniques. For $AL_3$PR-SOLAFS technique, high pitch period modification factors leads to acceptable degree of naturalness but in some ways, it sounds like metallic effects with the increment of the modification factors (greater then 1.5). Pitch period modification factors greater than 2 are used only by the first technique, but the quality and naturalness are reversely proportional to these latter. It could be used in the case of $AL_3$-PSOLAFS technique by changing the length of the window to four pitches. On the other hand, in the $AL_3$-PSOLAFS technique, the increment in the pitch period modification factor beyond 1.5 leads to a certain sensation of husky voices which degraded significantly the naturalness of the modified speech. Further, the modified speech quality is seriously affected when large values factors are applied.

For low modification pitch period factors (lower than one), both techniques have an acceptable assessment in intelligibility and distortion but the naturalness performances of the modified speech using $AL_3$PR-SOLAFS technique is clearly improved compared to $AL_3$-PSOLAFS which still be perceived as husky voice when low modifications factors are applied. This is because in natural speech, low pitch periods (high $F_0$ values) also correspond to some extent increased in formants frequencies (especially $F_1$). Furthermore, the $AL_3$-PSOLAFS technique suffers from amplitude and phase distortion when large scales are applied. However, the amplitude and phase of the modified speech using $AL_3$PR-SOLAFS is not affected at all. The use of small modification factors around unity (β and F equal to 0.9) in the experiment aims to evaluate the accuracy of the proposed techniques in the reconstruction of the modified speeches without perceptual and acoustical degradation, which is clearly proved in MOSs tests.

Based on subjective hearing tests, it can be noticed that $F_0$-scale modification of mixed SA sounds is perceptually disagreeable and it should be copied with no modifications. On the whole, there is no need for $F_0$ marking procedure in these regions. Only fully voiced parts will be marked and scaled.

The modified signal and the original one of Fig. 7 and 8 correspond to the vowel [a:] taken from the context /sa:/ extracted from the word /tusa: ɛi du/ at a duration of 10 ms. Figure 8, clearly shows that $F_1$ in the modified spectrum is increased by the same factor as $F_0$. The remainder spectral enveloped that include the other formants still unchanged in phase and amplitude. Simultaneous time scale and $F_0$-scale are allowed using both techniques and have the advantage to be based on a frame by frame processing from the fact that sound classification, duration modification, $F_0$ marking and scaling are performed in simultaneous fashion. This result offers the possibility of real time implementation of the proposed prosody modification techniques.

## CONCLUSION

In this study we have proposed two prosody modification techniques by combining SOLAFS algorithm with the multi level DWT using sentences in SA language sampled at an $F_S$ of 16 kHz.

The first technique consists of pitch synchronous processing of the third approximation level time segments used in SOLAFS algorithm. It aims at modifying the prosody of the input speech without affecting the spectral envelop.

The second one modifies duration and both local $F_0$ and $F_1$ of the instantaneous analysis time segments used in the SOLAFS algorithm at the third approximation level. It uses a re-sampling method of speech portions based on FFT interpolation principal. The technique explores the correlation between $F_0$ and $F_1$ of SA speeches in the frequency band from 0-1 kHz. For both techniques, $F_0$ shifting and marking are performed at the third approximation so that it provides independent control over the spectral envelope. Pitch marking is accomplished by comparative selection of peaks instants on multi level decomposition. Further, segments classification in the main regions (/suk:un/,

voiced, unvoiced and mixed) is also based on DWT processing.

The proposed techniques have the advantage to explore the appropriateness of multi level decomposition of speech signal, the reduced computational requirements, the suitability for real time implementation and the high quality time scale provided by the SOLAFS algorithm in order to manipulate both duration and $F_0$ for large scales. Further, the used of SOLAFS in duration modification aims at reduce buzziness that could appear in unvoiced sounds that uses the overlap and add principal.

Perceptual tests show that increasing $F_0$ and $F_1$ together by re-sampling the speech portions in the third approximation, improves significantly the naturalness of the $F_0$-scaled voices.

The use of high quality TSM algorithm in conjunction with multi level processing for $F_0$ shifting leads to high quality prosody modification and high control of speech bands. Future works may focus on the manipulation of formants structures at a multi level processing to increase the naturalness of higher pitch period modification factors.

## REFERENCES


Abou, H.L., 1994. Linguistic norms and dialectal variability: Spectral analysis of Arabic vocalic system, J. Applied Phonetic, 110: 1-15. http://runners.ritsumei.ac.jp/cgi-bin/swets/hold-query-e?mode=1&key=&idxno=08242043

Aissiou, M. and M. Guerti, 2009. Genetic supervised classification of standard arabic fricative sounds. Int. J. Speech Techno. 12:139-147. DOI: 10.1007/s10772-009-9061-5

Al-Ani and H. Salman, 1970. Arabic phonology an acoustical and phonological investigation. 1st Edn. Mouton and Co., Netherlands, pp: 583.

Al-Manie, M. A., M.I. Alkanhal and M.M. Al-Ghamdi, 2009. Automatic speech segmentation using the Arabic phonetic database. Proceedings of the 10th WSEAS International Conference on Automation and Information, Mar. 23-25, WSEAS Press, Czech Republic, pp: 76-79. http://portal.acm.org/citation.cfm?id=1562031

Alotaibi, Y.A. and A. Hussain, 2009. Formant based analysis of spoken Arabic vowels. Lecture Notes Comput. Sci., 5707: 162-169. DOI: 10.1007/978-3-642-04391-8_21

Belkaid, Y., 1984. Arabic vowels, modern literature, spectrographic analysis. Phonetic Works Strasbourg Inst., 16: 217-240. http://cat.inist.fr/?aModele=afficheN&cpsidt=12229518

Cheng, Y.M. and D.O. Shaughnessy, 1989. Automatic and reliable estimation of glottal closure instant and period. IEEE Trans. Acoust., Speech Signal Process., 37: 1805-1815. DOI: 10.1109/29.45529

Edgington, M. and A. Lowry, 1996. Residual-based speech modification algorithms for text-to-speech synthesis. 4th International Conference on Spoken Language, Oct. 3-6, IEEE Xplore Press, Philadelphia, pp: 1425-1428. DOI: 10.1109/ICSLP.1996.607882

Ykhlef, F., F.R. Ykhlef, M. Bensebti and M. Guerti, 2008. Pitch shifting of Arabic speech signal by source filter modeling for prosodic transformations. Int. J. Software Eng. Appli., 2: 2. http://www.sersc.org/journals/IJSEIA/vol2_no2_2008/6.pdf

Ghazeli, S., 1979. Status of vowels in Arabic: theoretical analysis, Arabic Stud., 3: 199-219. http://www.dur.ac.uk/daniel.newman/bib1.pdf

Hirahara, T., 1988. On the role of fundamental frequency in vowel perception. J. Acoust. Soc. Am., 84: 156-156. DOI: 10.1121/1.2025892

Hung, X. A. Acero and H.W. Hon, 2001. Spoken Language Processing, a Guide to Theory, Algorithm and System Development. 1st Edn., Prentice Hall Inc., USA., ISBN: 10: 0130226165, pp: 980.

Johnson A.J.I., 1996. Discrete wavelet transform techniques in speech processing. Proceeding of the IEEE TENCON Conference, Digital Signal Processing Applications, Nov. 26-29, IEEE Xplore Press, Australia, pp: 514-519. DOI: 10.1109/TENCON.1996.608394

Hejna, J.R., B.R. Musicus, Crowe and S. Rew, 1992. Method for time scale modification of signals. United States Patent, no. 5175769. http://www.freepatentsonline.com/5175769.html

Jun, Z., T. Wei, C. Yanpu and G. Yue, 2009. Parameters evaluation of SOLA algorithm for time scale modification. Int. J. Speech Technol., 10: 89-94. DOI: 10.1007/s10772-009-9019-7

Kain, A. and Y. Stylianou, 2000. Stochastic modeling of spectral adjustment for high quality pitches modification. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, June 5-9, IEEE Xplore Press, Istanbul, pp: 949-952. DOI: 10.1109/ICASSP.2000.859118

Kumar, A. and R. Jain, 2006. Speech pitch shifting using complex continuous wavelet transform. Proceeding of the IEEE Annual India Conference, Sept. 15-17, IEEE Xplore Press, India, pp: 1-4. DOI: 10.1109/INDCON.2006.302846

Laprie, Y. and V. Colotte, 1998. Automatic pitch marking for speech transformations via TD-PSOLA. Proceeding of the European Signal Processing Conference, Sept. 8-11, Typorama Patras Press, Rhodes, pp: 1133-1136. http://cat.inist.fr/?aModele=afficheN&cpsidt=1369614

Laroche, J., Y. Stylianou and E. Moulines, 1993. HNS: Speech modification based on a harmonic + noise model. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 27-30, IEEE Xplore Press, Minneapolis, MN., pp: 550-553. DOI: 10.1109/ICASSP.1993.319365

Lee, Y., K. Papineni, S. Roukos, O. Emam and H. Hassan, 2003. Language model based Arabic word segmentation. Proceeding of the 41st Annual meeting of The Association for Computational Linguistics, July 7-12, ACL Press, Japan, pp: 399-406. DOI: 10.3115/1075096.1075147

Mallat, S., 1989. A Theory for multiresolution signal decomposition: the wavelet representation. IEEE Patt. Anal. Mach. Int., 11: 674-693. DOI: 10.1109/34.192463

Mansour, M.A., 1998. Spectrographic analysis of Arabic vowels: A cross dialect study. J. King Saad Univ. Art, 10: 3-24. http://digital.library.ksu.edu.sa/paper1964.html

Moulines, E. and J. Laroche, 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Commun., 16: 175-205. DOI: 10.1016/0167-6393(94)00054-E

Moulines, E. and R.D. Francesco, 1989. Detection of the glottal closure by jumps in the statistical properties of the signal. Speech Commun. Neurospeech, 9: 401-418. DOI: 10.1016/0167-6393(90)90017-4

Muralishankar, R., A.G. Ramakrishnan and P. Prathibha, 2004. Modification of pitch using DCT in the source domain. Speech Commun., 42: 143-154. DOI: 10.1016/j.specom.2003.05.001

Newman, D. and J. Verhoeven, 2002. Frequency Analysis of Arabic Vowels in Connected Speech. Antwerp Papers in Linguist., 100: 77-86.

Peeters, G., 1998. Analysis and synthesis of musical sounds by the method PSOLA.

Popescu, B.P., 2001. Wavelets and their applications. Tech. Engender., TE5212: 1-25.

Prasad, K.P. and P. Satyanarayana, 1986. Fast interpolation algorithm using FFT. IEEE Elect. Lett., 22: 185-187. DOI: 10.1049/el: 19860129

Rao, K.S. and B. Yegnanarayana, 2006. Prosody modification using instants of significant excitation. IEEE Trans. Audio Speech Language Process, 14: 3. DOI: 10.1109/TSA.2005.858051

Roucos, S. and A. Wilgus, 1985. High quality time-scale modification of speech. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 1985, IEEE Xplore Press, Tampa, pp: 493-496.

Schnell, N., S. Lemouton, P. Manoury and X. Rodet, 2000. Synthesizing a choir in real time using pitch synchronous overlap add. Proceeding of the International Computer Music Conference, Sept. 200, Ircam Press, Berlin, pp: 102-108. http://articles.ircam.fr/index.php?Action=ShowArticle&IdArticle=159&ViewType=1

Strang, G. and T. Nguyen, 1996. Wavelets and Filter Banks. 2nd Edn., Wellesley-Cambridge Press, ISBN: 0961408871, pp: 520.

Syrdal, A.K. and S.A. Steele, 1985. Vowel F1 as a function of speaker fundamental frequency. J. Acoust. Soc. Am., 78: 56-56. DOI: 10.1121/1.2022883

Tanaka, K. and M. Abe, 1997. A new fundamental frequency modification algorithm with transformation of spectrum envelope according to $F_0$. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 21-24, IEEE Xplore Press, USA., pp: 951-954. DOI: 10.1109/ICASSP.1997.596095

Yegnanarayana, B., S. Rajendran, V.R. Ramachandran and A.S.M. Kumar, 1994. Significance of knowledge sources for TTS system for Indian languages. SADHANA, 19: 147-169. DOI: 10.1007/BF02760395