International Journal of Signal System Control and Engineering Application 3 (4): 77-80, 2010
ISSN: 1997-5422

Implementation of the Arabic Speech Synthesis with TD-PSOLA Modifier

Abdelkader Chabchoub and Adnan Cherif
Laboratory of Signal Processing Science, Faculty of Tunis, 1060 Tunis, Tunisia

Abstract: This research describes techniques to improve the precision of prosodic modifications in the Arabic speech synthesis, using the TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) method. This approach is based on the decomposition of the signal into overlapping frames synchronized with the pitch period. The main objective is to preserve the consistency and accuracy of the pitch marks after prosodic modifications of the speech signal.

Key words: Speech processing, synthesis, pitch, TD-PSOLA, signal, system, power, frequency

INTRODUCTION

Several speech synthesis systems were developed such as vocoders and LPC synthesizers (Childers, 1995; Childers and Lee, 1991) but most of them did not reproduce high quality of synthetic speech when compared with that of PSOLA based systems (Acero, 1998) such as MBROLA synthesizers (Dutoit *et al*., 1996). Especially, TD-PSOLA method (Time Domain Pitch Synchronous Overlap-Add) is the most efficient method to produce criteria of satisfaction speech (Moulines and Charpentier, 1990) and is one of the most popular concatenation synthesis techniques now-a-days. LP-PSOLA (Linear Predictive PSOLA) and FD-PSOLA (Frequency Domain PSOLA), though able to produce equivalent result, require much more computati-onal power. The 1st step of the TD-PSOLA is to perform a pitch detection algorithm and to generate pitch marks through overlapping windowed speech. To synthesize speech, the Short Time signals (ST signals) are simply overlapped and added with desired spacing of the ST-signals.

TD-PSOLA PRINCIPLE

To describe the TD-PSOLA principle we would like 1st to define the input signal as x[n] and a local version of $x_a[n]$ centered at t_a time, t_a is an analysis marks:

$$x_a[n] = x[t_{a+n}]$$

We can then define $y_a[n]$ as a short-time version of $x_a[n]$ by multiplying it by a window $w_a[n]$ (Fig. 1):

$$y_a[n] = w_a[n] \times x_a[n] \quad (1)$$

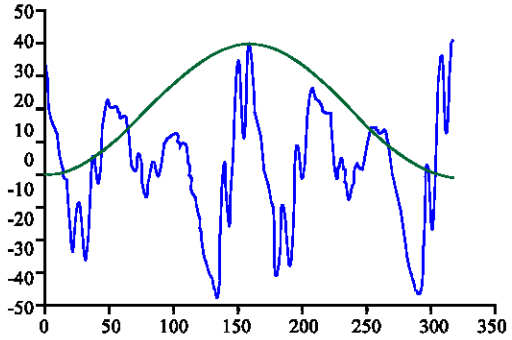

Fig. 1: A windowed speech signal using a hanning window $w_a[n]$

The window length is two times of the local pitch period (for that spectrum Si(n) approximates the spectral envelope x(n)). To synthesize speech at different pitch periods, the Short Time signals (ST) are simply overlapped and added with desired spacing. The synthesized speech is:

$$y[n] = \sum_{a=-\infty}^{\infty} y_a[n - t_a] \quad (2)$$

A good choice for the time marks (t_a) is to coincide with the instants of closing of the vocal folds which indicate the periodicity of speech.

For unvoiced speech, these marks could be arbitrarily placed. This estimation from speech waveforms is a very difficult problem but it can be done accurately by using EGG signals.

The use of a symmetric window makes perfect reconstruction impossible, unless time marks are equally spaced. In addition, truncation will occur if these time marks are spaced >N/2 apart (very long pitch periods). In

Corresponding Author: Abdelkader Chabchoub, Laboratory of Signal Processing Science, Faculty of Tunis, 1060 Tunis, Tunisia

synthesis, re-sampling is necessary at a time sequence t_s is a synthesis marks different from that of the analysis marks t_a.

SPEECH ANALYSIS AND SYNTHESIS

This study will describe the procedures of synchronous analysis and synthesis using TD-PSOLA modifier. Figure 2 shows the block diagram of these two stages.

Speech analysis: The 1st step in the speech analysis is to filter the speech signal by a RIF filter (pre-accentuation). The next step is to provide a sequence of pitch-marks and voiced/unvoiced classification for each segment between two consecutive pitch marks. This decision is based on the zero-crossing and the short time energy (Fig. 3a, b). A coefficient of voicement (v/uv) can be computed in order to quantize the periodicity of the signal (Cheveigne and Ahara, 1990).

Automatic segmentation: The segmentation of a speech signal is used in order to identify the voiced and un-voiced frames. This classification is based on the zero-crossing ratio and the energy value of each signal frame.

Speech marks: Different procedures of placed $t_a[i]$ are used according to the local features of components of the signal. A previous segmentation of the signal in identical feature zones permits to orient the marking toward the suitable method. Besides results of this segmentation will be necessary for the synthesis stage.

Reading marks: The idea of the algorithm is to select pitch marks among local extrema of the speech signal. Given a set of mark candidates which all are negative peaks or all positive peaks:

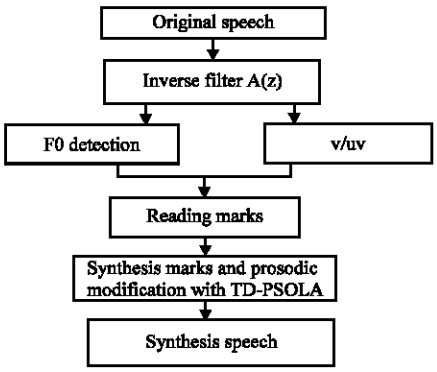

Fig. 2: Block diagram of speech analysis and synthesis

$$T_a = [t_a(i)] = t_a(1)...t_a(i).....t_a(N)$$

Where:
$t_a(i)$ = Sample of the ith peak
N = Number of peaks extracted

Laprie and Colotte (1998) explain how these candidates are found. Pitch marks are a subset of points out of T_a which are spaced by periods of pitch given by the pitch extraction algorithm. The selection can be represented by a sequence of indices:

$$J = [j(k) = j(1)..... j(k)..... j(K)$$

With K<N. J has to preserve the chronological order which requires the monotony of j: j(k)<j(k+1). The sequence of indices along with the corresponding peaks is defined to be the set of pitch marks:

$$T_a = [t_a(j(k))] = t_a(j(1))...t_a(j(k))...t_a(j(K))$$

The determination of j requires a criterion expressing the reliability of two consecutive pitch marks with respect to pitch values previously determined. The local criterion, we chose is:

$$d(c(1);c(i)) = |(c(i) - c(1)) - P_a(c(1))| \quad (3)$$

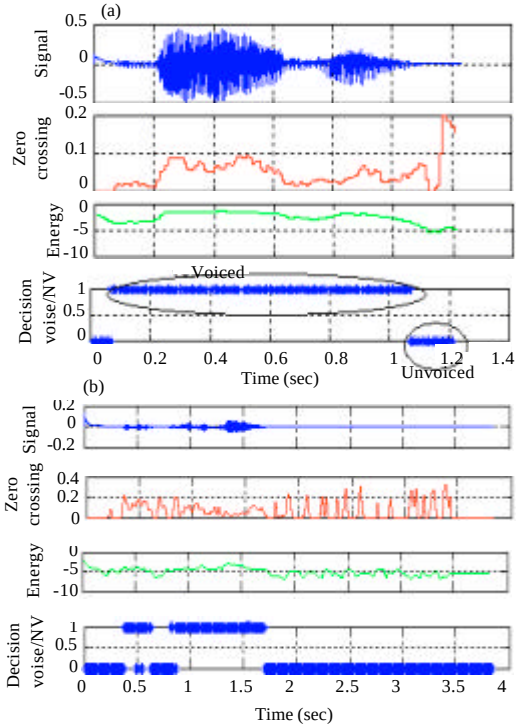

Fig. 3: Automatic segmentation of Arabic speech; a) babun; b) chamsun. This segmentation is used in order to identify the voiced and unvoiced frames

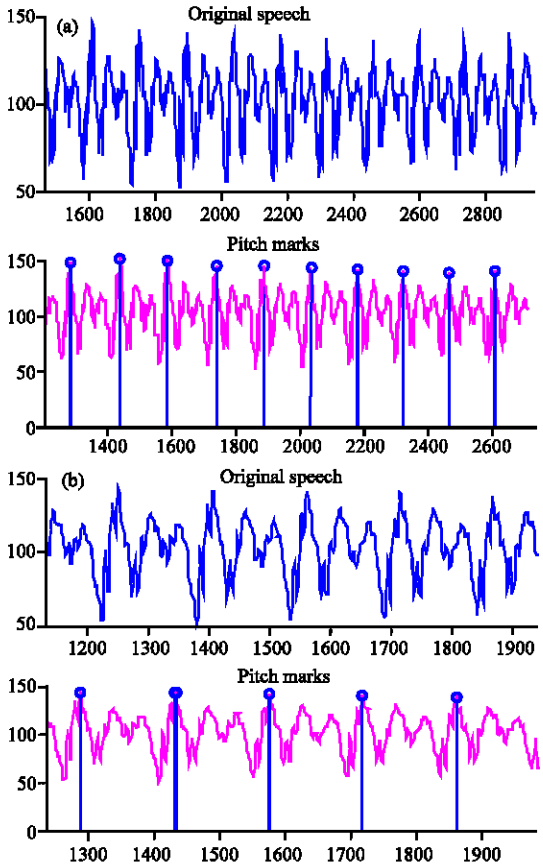

Fig. 4: Pitch marks of Arabic speech; a) babun; b) akala

We use the following algorithm for the marking: where, l<i. It takes into account the time interval between two marks compared to the pitch period P_a in samples. This criterion returns zero if the two peaks are exactly $P_a(c(l))$ samples away from one another and a positive value if the distance between these peaks is greater or less than the pitch period. The overall criterion is:

$$D = \sum_{K=1}^{K=l} d\left(t_a(j(k)), t_a(j(k+1))\right) - B\left(t_a(j(k+1))\right) \quad (4)$$

where, B is the bonus of selecting an extremum as a pitch mark. In a 1st time:

$$B(t_a(j(k)) = \delta\left|\text{amplitude}(t_a(j(k)))\right| \quad (5)$$

The coefficient δ expresses the compromise between closeness to pitch values and strength of pitch marks. Minimising D is achieved by using dynamic programming. The pitch marking results is shown in Fig. 4a, b.

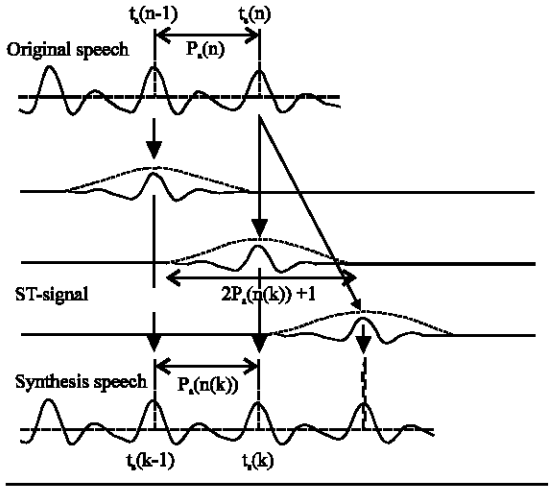

Fig. 5: TD-PSOLA for pitch (F0) modification

Synthesis marks: The OLA synthesis is based on the superposition-addition of elementary signals $Y_j(n)$, obtained from the $X_i(n)$ placed in the new positions $t_s[j]$. These positions are determined by the height and the length of the synthesis signal.

In such synthesis, one can modify the temporal scale by a coefficient t-scale .The positions $t_s(k-1)$ and the pitch period $P_a(k)$ are supposed to be known we can deduce $t_s(k)$ as (Mower *et al.*, 1991):

$$\begin{aligned} t_s(k) &= t_s(k-1) + t-\text{scale}\cdot P_a\left(n(k)\right) \\ n(k+1) &= n_s(k) + t-\text{scale} \end{aligned} \quad (6)$$

t-scale: Coefficient of length modification (Fig. 5). In order to increase the pitch, the individual pitch-synchronous frames are extracted, Hanning windowed, moved closer together and then added up. To decrease the pitch, we move the frames further apart. Increasing the pitch will result in a shorter signal so, we also need to duplicate frames if we want to change the pitch while holding the duration constant.

SYNTHESIS SPEECH

Therefore, given the pitch mark and the synthesis mark of a given frame, we use a fast re-sampling method described below to shift the frame precisely where, it will appear in the new signal. Let x[n] the original frame, the re-sampled signal is given by Oppenheim and Schafer (1975):

$$x(t) = \sum_{n=-\infty}^{\infty} x[n]\sin c\left(\frac{\pi(t-nTs)}{Ts}\right) \quad (7)$$

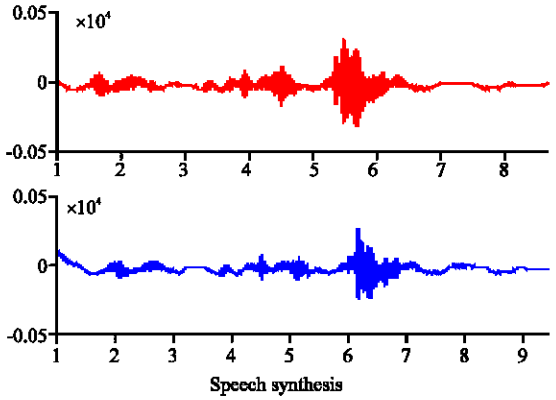

Fig. 6: Speech synthesis akala

where, Ts is the sampling period. Calculating the result frame y[m] corresponding to the frame x[n] shifted by a small delay δ amounts to evaluate x(mTs-δ). Therefore, y[m] = x(mTs-δ), i.e:

$$y[m] = \sum_{n=-\infty}^{\infty} x[n]\sin c\left(\pi fs\left[(mTs-\delta)-nTs\right]\right) = \sum_{n=-\infty}^{\infty} x[n]\sin c\left(\pi fs\left[(m-n)Ts-\delta\right]\right) \quad (8)$$

where, fs is the sampling frequency (1/Ts). Now by rewriting sin c as sin(x)/x and by using the following equation:

$$\sin(\pi fs[(m-n)Ts-\delta] = \cos(\pi fs\ \delta)\sin(\pi(m-n))$$

but $\cos\pi(m-n) = \pm 1$ and $\sin\pi(m-n) = 0$ we get (Fig. 6):

$$y[m] = \sum_{n=-\infty}^{\infty} x[n]\frac{(-1)^{(m-n+1)}\sin(\pi fs\delta)}{\pi fs\left[(m-n)Ts-\delta\right]} \quad (9)$$

As $0<\delta<Ts$ (resp.$-Ts<\delta<0$), we define $\delta = \alpha\ Ts$ where, $0<\alpha<1$ (resp.$-1<\alpha<0$). Then the synthesized speech is:

$$y[m] = \sum_{n=-\infty}^{\infty} (-1)^{(m-n+1)} x[n]\frac{\sin(\alpha\pi)}{\pi}\frac{1}{(m-n)-\alpha} \quad (10)$$

CONCLUSION

In this study, a voice quality conversion algorithm with TD-PSOLA modifier was implemented and tested under Matlab environment. The results of perceptual evaluation test indicate that the algorithm can effectively convert modal voice into the desired voice quality. Results of the simulation verify that the quality of the synthesized signal by TD-PSOLA with technique depends on the precision of the analysis marking as well as the synthesis marking which must be placed with precision to avoid errors in the phase. The higher precision algorithm for pitch marking during the synthesis stage increases the signal quality. This gain in accuracy, avoids the reduction of deference between original and synthetic signals.

REFERENCES

Acero, A., 1998. Source-filter models for time-scale pitch-scale modification of speech. Proceedings of the International Conference on Acoustics, Speech and Signal Processing. May 12-15, Seattle, USA, pp: 881-884.

Cheveigne, A. and H. Ahara, 1990. A comparative evaluation of Fo estimation algorithm. Proceedings of the Euro Speech Conference. (ESC'98), Norvege, pp: 453-467.

Childers, D.G. and C.K. Lee, 1991. Vocal quality factors: Analysis, synthesis and perception. J. Acoust. Soc. Am., 90: 2394-2410.

Childers, D.G., 1995. Glottal source modeling for voice conversion. Speech Commun., 16: 127-138.

Dutoit, T., V. Pagel, N. Pierret, F. Bataille and O. van der Vrecken, 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for noncommercial purposes. Proceedings of 4th International Conference on Spoken Language. Oct. 3-6, Philadelphia, PA , USA., pp: 1393-1396.

Laprie, Y. and V. Colotte, 1998. Automatic pitch marking for speech transformations via TD-PSOLA. Proceedings of the European Signal Processing Conference, Sept. 8-11, Typorama Patras Press, Rhodes, pp: 1133-1136.

Moulines, E. and F. Charpentier, 1990. Pitch-synchronous waveform processing techniques for TTS synthesis. Speech Commun., 9: 453-467.

Mower, L., O. Boeffard, B. Cherbonnel, White, 1991. An algorithm of speech synthesis high-quality. Proceeding of a Seminar SFA/GCP. (SSFA'91), SFA/GCP, pp 104-107.

Oppenheim, A.V. and R.W. Schafer, 1975. Digital Signal Processing. Prentice-Hall, Inc., New York.