Voice Conversion using Pitch Shifting Algorithm by Time Stretchingwith PSOLA and Re-Sampling

Article *in* Journal of Electrical Engineering · June 2011

DOI: 10.2478/v10187-010-0008-5

1 author:

Allam Mousa
An-Najah National University
42 PUBLICATIONS **335** CITATIONS

SEE PROFILE

Journal of ELECTRICAL ENGINEERING, VOL. 61, NO. 1, 2010, 57–61

VOICE CONVERSION USING PITCH SHIFTING ALGORITHM BY TIME STRETCHING WITH PSOLA AND RE–SAMPLING

Allam Mousa *

Voice changing has many applications in the industry and commercial filed. This paper emphasizes voice conversion using a pitch shifting method which depends on detecting the pitch of the signal (fundamental frequency) using Simplified Inverse Filter Tracking (SIFT) and changing it according to the target pitch period using time stretching with Pitch Synchronous Over Lap Add Algorithm (PSOLA), then resampling the signal in order to have the same play rate. The same study was performed to see the effect of voice conversion when some Arabic speech signal is considered. Treatment of certain Arabic voiced vowels and the conversion between male and female speech has shown some expansion or compression in the resulting speech. Comparison in terms of pitch shifting is presented here. Analysis was performed for a single frame and a full segmentation of speech.

K e y w o r d s: voice changing, PSOLA, pitch shifting, resampling, SIFT

1 INTRODUCTION

Speech is generated by pumping air from the lung through the vocal tract which consists of the throat, nose, mouth, palate, tongue, teeth and lips. Speech is usually characterized as voiced, unvoiced or transient forms [1].

Voiced speech is produced by an air flow of pulses caused by the vibration of the vocal cords. The resulting signal could be described as quasi-periodic waveform with high energy and high adjacent sample correlation. On the other hand, unvoiced speech, which is produced by turbulent air flow resulting from constrictions in the vocal tract, is characterized by a random aperiodic waveform with low energy and low correlation [1]. The main differences between voiced and unvoiced speech signals are illustrated in Fig. 1.

Voiced sounds as vowels have a periodic structure, *ie* , their signal form repeats itself after time, and this is called the pitch period T_P. Its reciprocal value $f_P = 1/T_P$ is called the pitch frequency [2].

There are a number of algorithms for pitch period estimation. The two broad categories of pitch-estimation algorithms are the time-domain and frequency-domain algorithms.

Time-domain algorithms attempt to determine the pitch directly from the speech waveform while frequency-domain algorithms use some forms of spectral analysis to determine the pitch period.

Pitch changes, pitch scaling, or pitch modification means transposing the pitch without changing the characteristics of the sound. In addition, it is defined as the process of changing the pitch without affecting the speech.

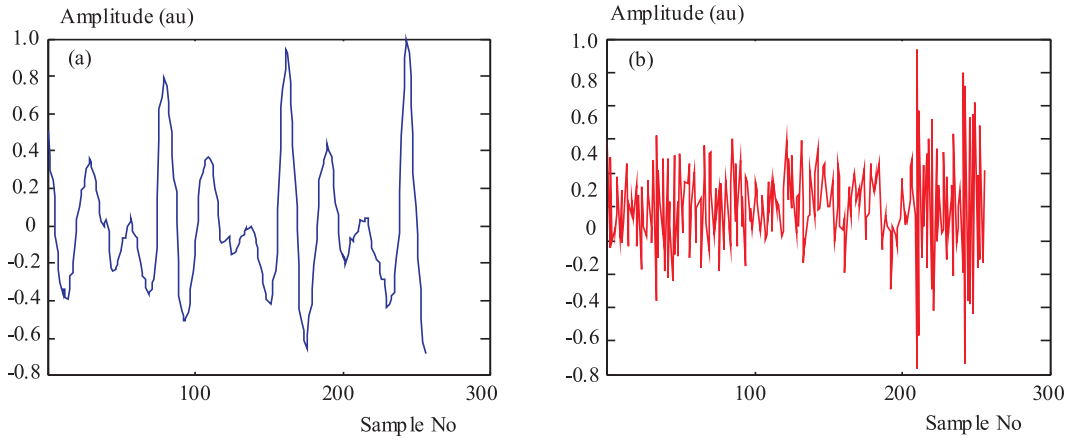

Fig. 1. Periodic nature versus aperiodic nature of: (a) – voiced speech, and (b) – unvoiced speech

* Electrical Engineering Department An Najah University, PO Box 7, Nablus, Palestinian Authority, allam@najah.edu

ISSN 1335-3632

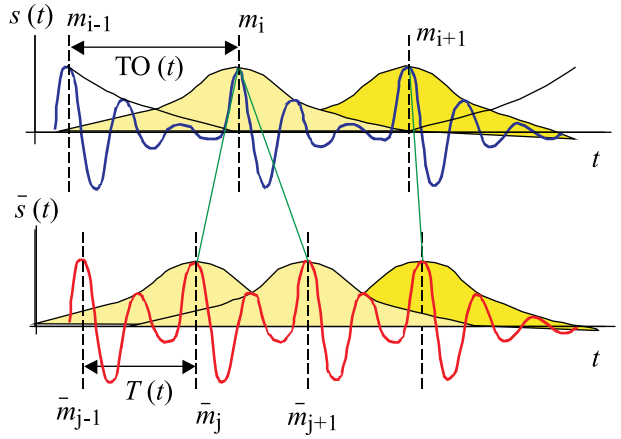

Fig. 2. Example of pitch-shifting and time stretching using PSOLA

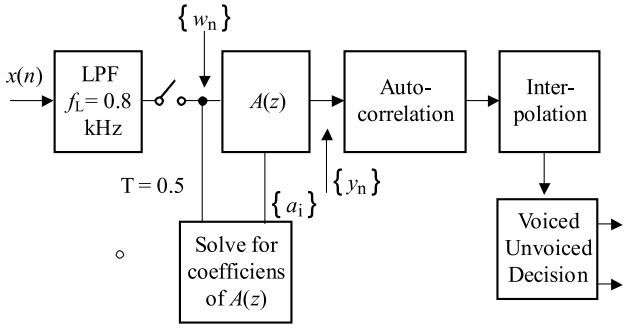

Fig. 3. Block diagram of a simplified SIFT block diagram

Pitch changing is an important algorithm which is used in voice conversion which has many applications like foreign language training and movie dubbing. It is also closely related to the process of converting a text into a spoken language. This has many applications especially in the field of assisting blind and deaf people, speaker verification and security.

2 VOICE CONVERSION USING PSOAL ALGORITHM

Speech morphing and voice conversion are obviously different: In the former case the source and target signals should be sufficiently similar to become reasonably aligned and interpolated for achieving new signals. In the latter case a source-target relationship is learned from a number of (not necessarily similar) utterances from two speakers [3]. This is then used to convert unseen signals of the source speaker towards the target speaker.

One of the methods in order to change the voice is to change the pitch of the voice; this is done by shifting the pitch of the voice using certain techniques like the Pitch Synchronous Over Lap-Add (PSOLA) algorithm.

2.1 Pitch shifting

The pitch period is responsible for making some sounds to be sharper than others. The number of vibrations produced during a given period determines the pitch period. This vibration rate of a sound is called its frequency, the higher the frequency the higher the pitch. The aim of pitch shifting algorithms is to create a change in pitch without creating a change in the replay rate. Pitch shifting can be done by performing a time stretch using PSOLA and resampling.

2.2 PSOLA

PSOLA is a method based on decomposition of a signal into a series of elementary waveforms in such a way that each waveform represents one of the successive pitch periods of the signal and the sum (overlap-add) of them reconstitutes the signal. PSOLA works directly on the signal waveform without any sort of model and therefore does not lose any detail of the signal [4].

There are several types of PSOLA such as Time Domain TD-PSOLA, Frequency Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA). TD-PSOLA is the most commonly used due to its computational efficiency but the others are more appropriate approaches for pitch-scale modifications because they provide independent control over the spectral envelope of the synthesis signal [5].

2.3 TD-PSOLA algorithm

The TD-PSOLA algorithm was proposed allowing pitch modification of a given speech signal without changing the time duration and visa versa [6]. The TD-PSOLA consists mainly of the following three steps:

1. The analysis step, where the original speech signal is first divided into separate but often overlapping short-term analysis signals (ST). Short term signals $x_m(n)$ are obtained from the digital speech waveform $x(n)$ by multiplying the signal by a sequence of the pitch-synchronous analysis window $hm(n)$ as in Eq. 1:

$$X_m(n) = h_m(t_m - n)x(n)\,, \tag{1}$$

   where m is an index for the short-time signal
2. The windows, which are usually Hanning type, are centered on the successive instants t_m, called pitch-marks. These marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate onthe unvoiced parts.
3. The modification step, where each frame is modified according to the target. The synthesis steps are performed such that these segments are recombined by means of overlap adding.

The main advantages of time-domain algorithms are:

- easy to be implemented,
- give good result when used on both speech signals given that a small pitch scale factor is used,

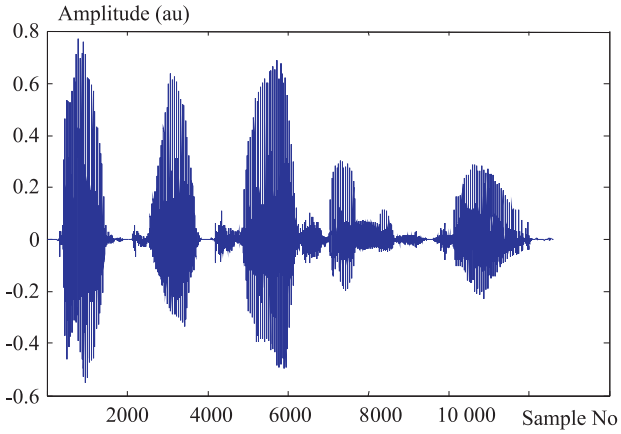

Fig. 4. Original speech signal waveform

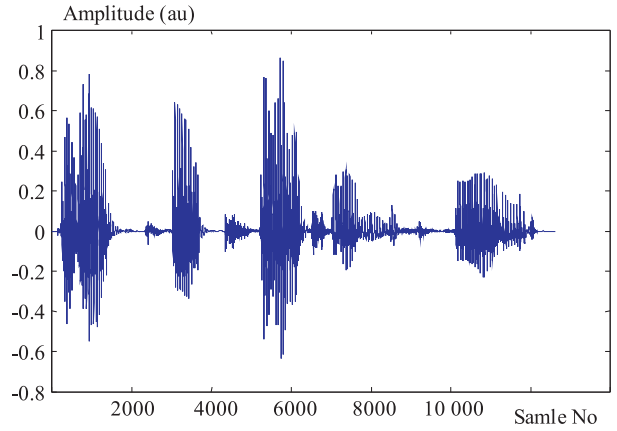

Fig. 5. Speech signal waveform after channging the pitch

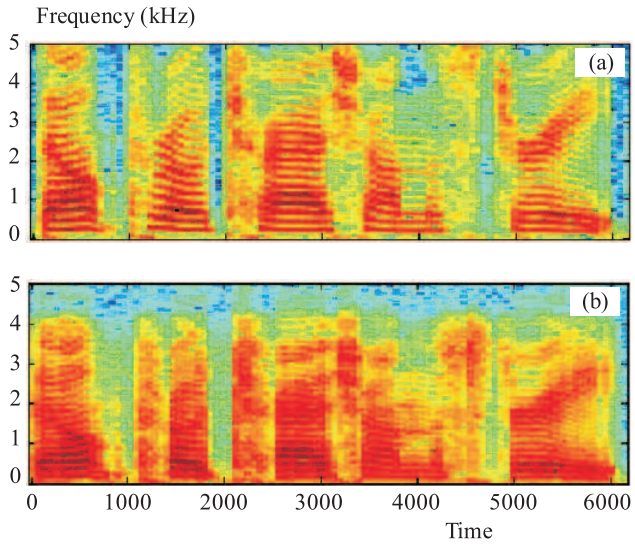

Fig. 6. Sectrogram of signal: (a) – before shifting the pitch, (b) – after the modification

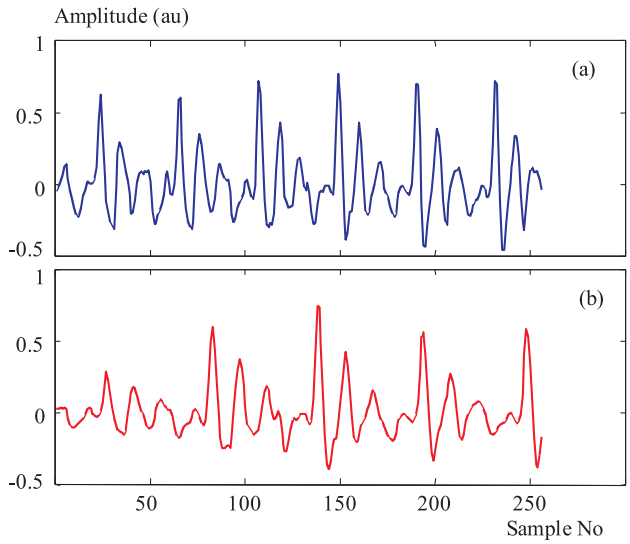

Fig. 7. 256 Samples of speech signal before (a), and after (b) changing the pitch

- pitch scaling in the time domain can be done by simply combining time-scaling and sample rate conversion.

2.4 Pitch Shifting by Time Stretching using PSOLA and Resampling

Pitch shifting by time stretching and resampling involves simply performing a time stretch as described earlier with the PSOLA and then resampling in order to return sound length to its original value. Expanding the sound by time stretch then resampling creates a higher pitch, while compressing and resampling creates a deeper pitch.

There are other methods of pitch shifting such as delay line modulation or by PSOLA and formant preservation [7]. A modification of the pitch of the signal from $T_0(t)$ to $T_1(t)$ is shown in Fig. 2.

2.5 Pitch Detection

The pitch period is usually defined as the time between the beginnings of two successive glottal excitations. Pitch determination is essential for many speech processing tasks and applications, this includes the classification of speech signal into voiced or unvoiced speech regions [8]. Pitch determination may not be a straightforward process due to the nature of the speech production mechanism. There are several types of pitch detection algorithms such as:

- Time-domain analysis like Autocorrelation method, AMDF (Average Magnitude Difference Function)
- Frequency-domain analysis like Cepstrum, Harmonic product spectrum, Chens heuristic method
- Others like Maximum likelihood, Simple inverse filter tracking (SIFT), Neural network approaches

More details and investigations about pitch determination are given in [8]. That work has considered several pitch determination algorithms like SIFT, Comb filter en-

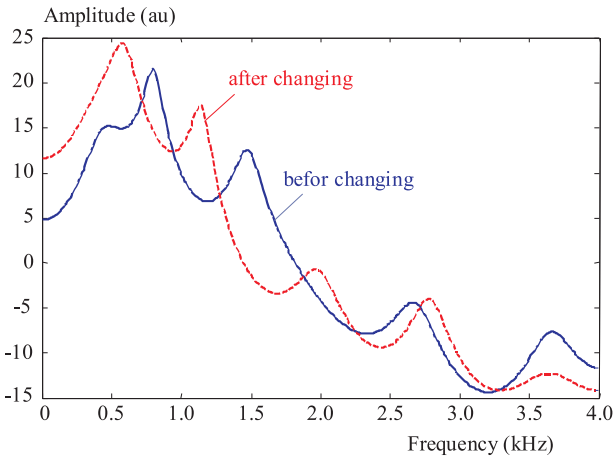

Fig. 8. Spectrum speech signal before (solid line), and after changing the pitch (dotted line)

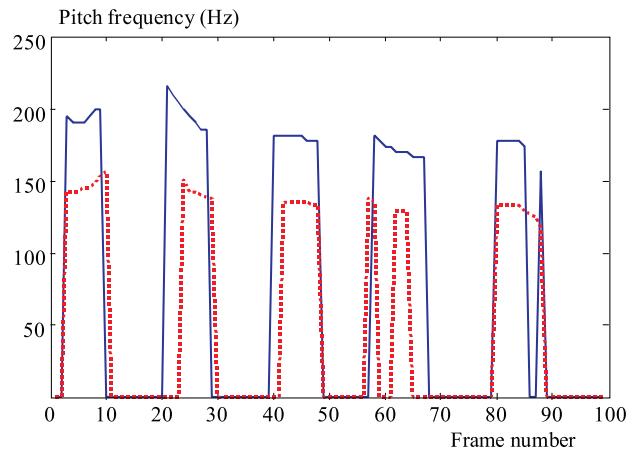

Fig. 9. Source pitch period per frame (solid line), and the new pitch period (dotted line)

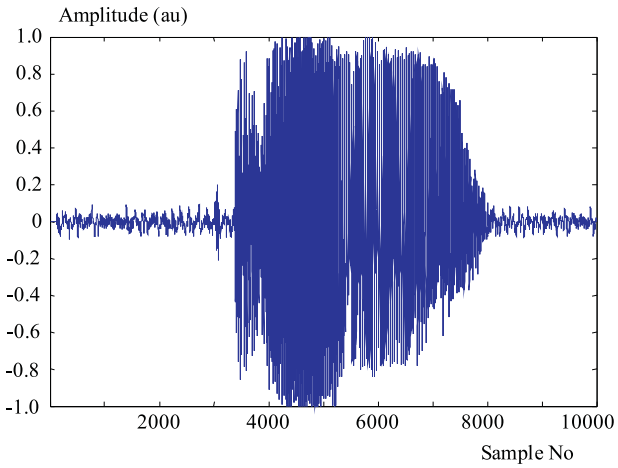

Fig. 10. Arabic female wave form (A'LI)

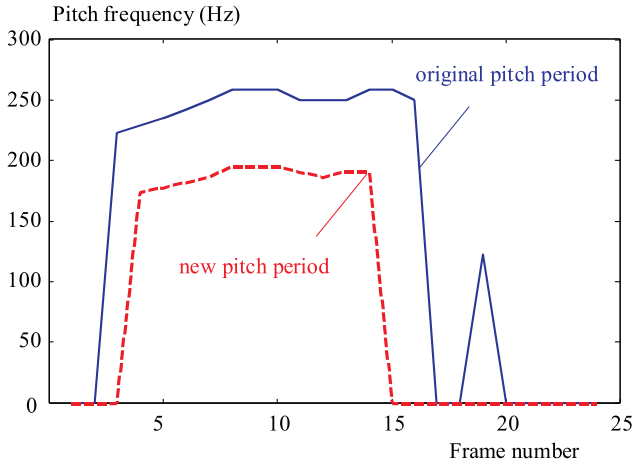

Fig. 11. Pitch in Hz for Arabic speech: original pitch (solid line), new pitch (dashed line)

ergy maximization, Spectrum decimation/accumulation, Optimal temporal similarity, Dyadic wavelet transform.

The algorithm which was used to detect the pitch of source and target signals is a simplified inverse filter tracking (SIFT) algorithm which was originally developed by Markel. The block diagram of the SIFT algorithm used by Markel [9] is shown in Fig. 3.

3 SIMULATION RESULTS:

Time stretching using the PSOLA algorithm and resampling in order to change the pitch were applied to a female speech signal as shown in Fig. 4. It is desired to decrease the pitch frequency of this signal, then the time stretching factor ? must be less than one, *ie*, the new pitch frequency/old pitch ratio is $\alpha = 133\,\text{Hz}/177\,\text{Hz} = 0.75$ and then resample the new signal in factor $= 1/\alpha = 1.33$.

The process is applied to the whole speech and the modified signal is shown in Fig. 5, where it is so clear that the modified signal has almost the same shape as the original one and also the same number of samples. This is mainly due to the resampling process which is applied in order to have the same play rate. A subjective test was performed on the new signal and it was clearly seen that this signal looks like a male speech signal saying the same original sentence. The spectrogram of the original and converted signals is shown in Fig. 6.

Looking closely at the signal requires focusing on a small data size as shown in Fig. 7 which shows the original speech of two frames illustrating clearly the original and modified pitch period of a voiced speech. The pitch period has been changed as desired.

The frequency domain may reflect these changes clearly. The spectra of these frames are shown in Fig. 8, where it is obvious that the formant frequency has been moved to the left due to pitch shifting. Actually the whole frequencies were shifted to the left. Some treatment, like formant correction, may be applied to improve the quality of the resultant speech.

The effect of changing the pitch period on the whole speech is shown in Fig. 9 for each individual frame of the signal. It is clear that the pitch period for voiced speech has been correctly transformed for a lower one as desired.

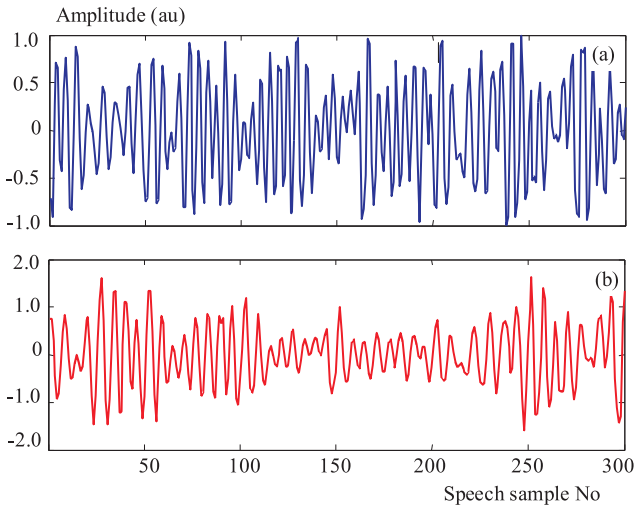

Fig. 12. Arabic speech signal after and before changing the pitch period in the time domain

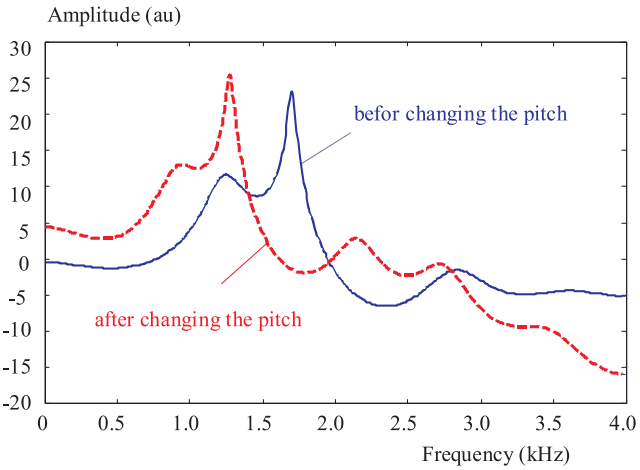

Fig. 13. Spectrum before (dashed line) and after (solid line) changing the pitch

Some of the Arabic speech vowels may have special characteristics such as the (A) and (kh). Hence, a special treatment is needed when voice conversion is applied. A typical Arabic speech (A'LI) is shown in Fig. 10 with emphasis on the (A') vowel for a female speaker.

The pitch shift applied to the whole speech is shown in Fig. 11 with success in achieving the shift as desired. A closer look at the data shows how much the pitch has been shifted and how the signal has been modified as shown in Fig. 12.

The spectrum of the new and old signals is illustrated in Fig. 13 which shows that the whole formant frequencies where shifted to the left and that the high frequency components were reduced in amplitude.

4 DISCUSSIONS AND CONCLUSION

Voice conversion depending on pitch shifting by time stretching using PSOLA and resampling was discussed. The pitch was detected in order to determine the pitch frequency of the source signal and the target, then the speech is converted from female speech to male speech or visa versa. The obtained results when converting the female speech to male one has shown that the pitch frequency will decrease and the signal must be expanded, while when converting the male signal to female one the pitch frequency will increase and so the signal must be compressed. This algorithm of voice conversion was applied to Arabic and English speech wave forms and the obtained results were similar.

References

[1] KONDOZ, A. M.: Digital Speech, Coding for Low Bit Rate Communication Systems, Wiley, 2004.

[2] JELINEK, M.—ADOUL, J.-P.: Frequency-Domain Spectral Envelope Estimation for Low Rate Coding of Speech, icassp, 1999.

[3] PFITZINGER, H. R.: Unsupervised Speech Morphing between Utterances of any Speakers, Proceedings of the 10th Australian International Conference on Speech Science & Technology, Sydney, December 8-10, 2004.

[4] SCHNELL, N. *et al*: Synthesizing a Choir in Real-Time using Pitch Synchronous Overlap Add (PSOLA), http//www.ircam.fr, dated 10/Sep/2006.

[5] JAU-HUNG CHEN—YUNG-AN KAO: Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method, Computational Linguistics and Chinese Language Processing **6** No. 2 (Feb 2001), 1–12.

[6] MARTIN, H.—GERNOT, K.: Poincare Pitch Marks, Speech Communication **48** (2006), 1650–1665.

[7] ARFIB, D.—VERFAILLE, V.: Driving Pitch-Shifting and Time-Scaling Algorithms with Adaptive and Gestural Techniques, Proc. of the 6th Int. Conference on Digital Audio Effects (dafx-03), London, UK, September 8-11, 2003.

[8] VEPREK, P.—SCORDILIS, M. S.: Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques, Speech Communication **37** (2002), 249–270.

[9] MARKEL, J. D.: The SIFT Algorithm for Fundamental Frequency Estimation, IEEE Transactions on Audio and Electroacoustics **20** No. 5 (Dec 1972), 367–377.

Received 17 July 2009

Allam Mousa is an Associate Professor in electrical engineering An Najah University. He received his BSc MSc and PhD in Electrical and Electronics Engineering from Eastern Mediterranean University in 1990, 1992 and 1996 respectively. From 1996 to 2000 he was with Al Quds University and served as the chairman of the Electronics Engineering Department then he joined An Najah University and was the chairman of the Electrical Engineering Department from 2004 to 2009. His current research interests are OFDM, Speech and Image Processing, Electromagnetic Radiation and Compatability. He teaches cources in Telecommunication Systems, Coding and Electrical Engineering. Dr. Mousa is also interested in higher education quality assurance and he is currently the director of Quality Assurance Unit. He is also a senior member of IEEE.