Speech Segmentation in Synthesized Speech Morphing Using Pitch Shifting

Allam Mousa
Electrical Engineering Department, An Najah University, Palestine

Abstract: *This paper discusses the speech morphing process showing some limitations of using the directly obtained LPC and excitation parameters of speech. The algorithm here depends on changing the pitch of the source to match that of the target based on analyzing the speech signals to its basic components. Different experiments for changing the female to female, male to male, male to female and female to male speech were performed. Interesting results were obtained while dealing with children's speech. Difficulties of obtaining the pitch period were overcome but the obtained results have some diversity in the quality of performance even though the pitch has been changed correctly. The method for obtaining LPC and excitation used could be improved which could provide better results for this application.*

Keywords: *Speech morphing, linear prediction, pitch detection, and speech synthesis.*

Received August 3, 2009; accepted November 5, 2009

1. Introduction

Speech is generated by pumping air from the lung through the vocal tract which consists of throat, nose, mouth, palate, tongue, teeth and lips. Speech is usually characterized as voiced, unvoiced or transient forms [6].

It is important to have a good understanding of the speech production mechanism and so to have a good model for representing the speech signal. 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 and aperiodic waveform with low energy and low correlation [6]. Samples of voiced and unvoiced speech signals showing the main nature of the speech are depicted in Figure 1. Voiced sounds (like vowels) have a periodic structure, i.e., their signal form repeats itself after some time which is called pitch period (TP). Its reciprocal value fP=1/TP is called pitch frequency [1].

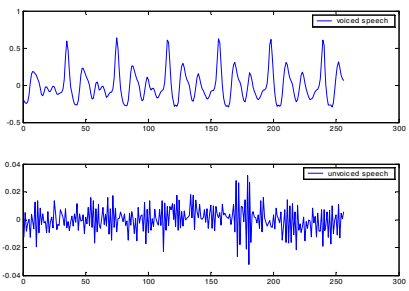

Figure 1. Periodic and aperiodic nature of voiced and unvoiced speech signals.

There exist a number of algorithms for pitch period estimation. Several pitch determination algorithms were discussed in [12]. The two broad categories of pitch-estimation algorithms are time-domain algorithms such as autocorrelation method and Linear Prediction Coding (LPC), and frequency-domain algorithms such as real cepstrum method. Time-domain algorithms attempt to determine pitch directly from the speech waveform. Frequency-domain algorithms use some form of spectral analysis to determine the pitch period. Basically, there are some advantages to the time domain algorithms and some other disadvantages to the Frequency domain algorithms. Changing, scaling or modifying the pitch means transposing the pitch without changing the characteristics of the sound. It can be seen as the process of changing the pitch without affecting the speed.

Speech morphing is one application of speech processing which is usually defined as the modification of speech signal of one speaker (source speaker) to sound as if it has been pronounced by a different speaker (target speaker). This converted speech can be recognized as the target speaker although the quality of the converted speech may be degraded. Some of the speech morphing techniques applications may be stated as [9].

- Text to speech customization systems where speech can be produced with a desired voice or email may be red out in the sender's voice.
- In replacing or enhancing the skills involved in producing sound tracks for animated characters, dubbing or voice impersonating which may be used in the entertainment industry.
- For voice disguising of a speaker especially in the

Internet chat rooms.

- Assisting hearing impaired persons where people with a hearing loss in the high frequency sounds may be able to benefit from the technique by changing appropriately the spectral envelope of the speech signal.
- Speakers of different languages may be able to communicate easier by certain systems that will first recognize the sentence uttered by each speaker and then translate and synthesize them in a different language using the original speaker sound.
- Concerning categorical perception, it would be interesting to use speech morphing between naturally spoken syllables, thus retaining all unknown cues and interpolating not only those cues known to be crucial [10].

The rest of the paper is organized such that; section 2 discusses the concept of speech processing with pre-emphasis and de-emphasis, it also discusses the LPC, pitch detection and pitch scaling. In Section 3 the speech morphing is discussed and Section 4 describes the simulation environment and discusses the results. Finally section 5 concludes this work.

2. Speech Processing

Speech processing is the study of the speech signals and hence the processing methods of these signals. The signals may be in analogue or digital format but usually it is processed in a digital representation. Speech processing can be divided into several categories like [6, 11].

- Speech recognition, dealing with analysis of the linguistic content of a speech signal.
- Speaker recognition, aiming to recognize the identity of the speaker.
- Enhancement of speech signals and noise reduction.
- Speech coding, which is mainly used in data compression and telecommunications.
- Voice analysis for medical purposes, such as analysis of vocal loading and dysfunction of the vocal cords.
- Speech synthesis to produce an artificial synthesis of speech like computer generated speech.

It is needed to analyze the speech signal into its components which are the excitation signal and the linear Prediction filter but before analysis the signal must be pre-emphasis using the pre-emphasis filter. Moreover, pitch determination usually has an important role in speech processing

2.1. Pre-Emphasis and de-Emphasis in Morphing

In speech processing, pre-emphasis should usually be applied to the input signal before the LPC analysis. During the reconstruction following the LPC synthesis, a de-emphasis process is applied to the signal to reverse the effect of pre-emphasis.

Pre- and de-emphasis are necessary because, in the spectrum of a human speech signal, the energy in the signal decreases as the frequency increases. Pre-emphasis increases the energy in parts of signal by an amount inversely proportional to its frequency. Thus, as frequency increases, pre-emphasis raise the energy of the speech signal by an increasing amount. This process therefore serves to flatten the signal so that the resulting spectrum consists of formants of similar heights. Moreover, digital speech waveforms have a high dynamic range and they suffer from the additive noise. Pre-emphasis is applied to reduce this range; this can be achieved by using a FIR filter of the form given by equation 1.

$$H(z) = 1 - az^{-1} \qquad 0.9 < a < 1.0 \qquad (1)$$

2.2. Linear Predictive Analysis

Linear Predictive Coding (LPC) is a method of predicting a sample of a speech signal based on several previous samples; the LPC coefficients may be used to separate a speech signal into two parts; the transfer function (which contains the vocal quality) and the excitation (which contains the pitch and the sound). The method of looking at speech as two parts is known as the source filter model of speech [7, 11]. The nth sample in a sequence of speech samples is represented by the weighted sum of the p previous samples as illustrated by equation 2.

$$S(n) = \sum_{k=1}^{p} a_k S(n-k) \qquad (2)$$

where p is the order of the prediction filter and a_k are the filter prediction coefficients (LPC coefficients) which are chosen in order to minimize the mean squared error between the real sample and its predicted value [7, 11].

Speech analysis into filter and excitation then speech synthesis process is shown in Figure 2 and explained in equations 3 and 4 respectively.

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (3)$$

$$E(z) = S(z)A(z) \qquad (4)$$

where A(z) represent the transfer function and E(Z) is the excitation function.

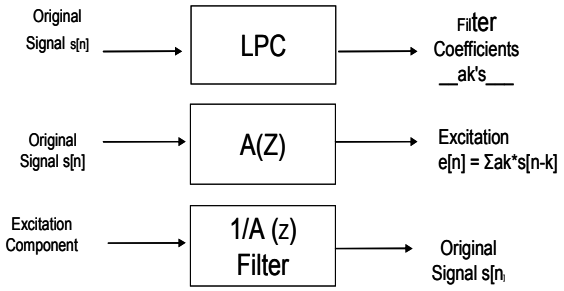

Figure 2. Speech analysis and synthesis process.

2.3. Order of the LPC Filter

In order to determine the optimum value of the number of predictor parameters, the minimum value of the prediction error, as a function of several values of the prediction order "p", has been studied [6]. The typical value for p was chosen to be 10-12 where the prediction error is minimized and almost not decreasing farther for values of p more than 12. These results suggest that p equal 12 is adequate for the voiced speech samples. Moreover, the SNR of the synthesized speech with respect to the original speech can be determined for different values of p. Again, the value of p equal 12 was adequate [6].

2.4 Pitch Detection

The pitch of the source and target signals must be detected in order to change the pitch of the source to match that of the target using pitch scaling algorithm. There are many possible methods for determining the pitch of a speech signal. A straight forward and simply implemented one depends on some sort of zero crossing algorithms [3]. However, this is complicated by the irregularity of the signal and so it becomes more appropriate to utilize a peak/valley detection algorithm and derive the pitch from there. Five basic pitch determination algorithms were discussed in [12]. These algorithms were.

- SIFT.
- Comb filter energy maximization.
- Spectrum decimation/accumulation.
- Optimal temporal similarity.
- Dyadic wavelet transform.

Some more details about pitch detection and high accuracy are illustrated in [2]. An algorithm which may be used to detect the pitch is Simplified Inverse Filter Tracking (SIFT) [8]. The block diagram of the SIFT algorithm is shown in Figure 3.

In this work, the pitch period of the speakers are defined and then inserted properly to fit such as to obtain the desired speaker sound.

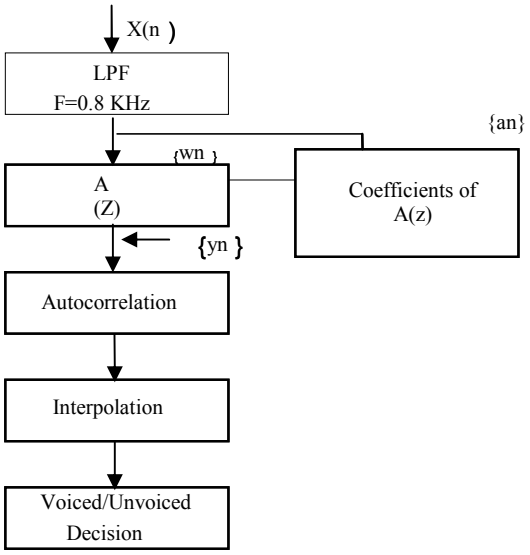

Figure 3. Simplified SIFT algorithm.

2.5. Pitch Scaling

Once the source and target pitch values are determined, the source pitch is scaled by the scaling factor α (α =pitch of the source/pitch of the target) in order for the source pitch value to match that of the target one. The original speech signal is first divided into separate, but often overlapping, short-term analysis signals (ST). Short term signals xm(n) are obtained from digital speech waveform x(n) by multiplying the signal by a sequence of pitch-synchronous analysis window hm(n) as shown in equation 5.

$$xm(n)=hm(tm-n)x(n) \quad (5)$$

where m is an index for the short-time signal the pitch modification was performed such that each frame is modified according to the target pitch by the scaling factor. Moreover, the synthesis procedure was such that these segments are recombined by means of overlap adding. This pitch scaling is performed per frame resulting in changing the shape of the speech signal. This change, for the whole speech and for each individual frame of the signal, is shown in Figure 4.

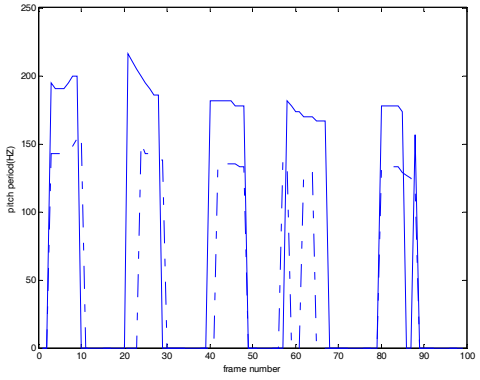

Figure 4. Pitch of speech signal, (-) old pitch of the signal, (-.-.) new pitch of the signal.

2.6. Speech Morphing

The concept of morphing relies heavily upon the fact that specific algorithms can synthesize the various characteristics of voice. If one had two speakers "A" and "B", and we wanted to take what "A" said, but make it come out in "B's".

In theory, one should take the excitation function of "A", map the pitch from the excitation function of "B" on to it using certain algorithms like the Dynamic Time Warping (DTW) algorithm, and then pass it through the filter created by the cavities and articulators of "B". This should synthesize "A" words using "B" pitch and formants [10].

The most important part of voice morphing is speech synthesis, since the quality of the synthesized speech is the ultimate aim of voice conversion. Speech signals will be synthesized by means of the same parametric representation that was used in the analysis. It can be synthesized from the linear predictive analysis parameters. A simplified flowchart showing the procedure used here to achieve speech morphing is illustrated in Figure 5.

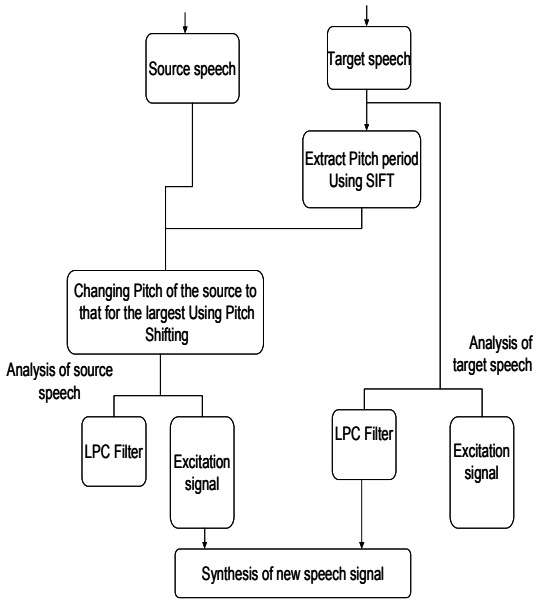

Figure 5. Flowchart of the speech morphing algorithm.

The speech signal is analyzed frame by frame and so is the synthesized speech, this is achieved by supplying the synthesizer with the control parameters, i.e. the excitation signal and LPC coefficients corresponding to each frame. The analysis was performed such as to get the LPC and excitation separately. More accurate results could be achieved for LPC coefficients, excitation and pitch values if more sophisticated techniques were performed like joint optimization discussed in [4, 5].

3. Simulations

Speech morphing described in this paper was fully implemented. The pitch periods of the speakers were determined and the speech signals were analyzed to obtain the LPC and excitation. Speech analysis was performed frame by frame and so was synthesis after modifying the pitch. The quality of speech morphing from certain sources to different targets were measured and the overall MOS was tabulated for different words, (like flower, apple, tree, come) as shown in Table 1. Speech ha been synthesized by exciting the LPC of the target signal by the excitation of the source one after changing its pitch period accordingly.

Table 1. Overall MOS for different words/speakers.

Test #	Female to Female	Female to Male	Male to Female	Male to Male
1	3.5	4	3.5	4
2	4	3.5	3	3.5
3	3.5	3.5	3	3.5
4	4	4	3	4
5	3.5	4	3.5	4
6	4	4	3	4
7	4	4	3	4

Although the overall performance may look satisfying, but it has been noticed that morphing for some speech cases was not that suitable. Children's speech morphing was lower in performance as shown in Table 2. This noticed degradation in the MOS could be mainly due to the excitation of an unstable filter as shown in Figure 6.

Table 2. MOS for morphing of children's speech.

Test #	Child1 to Child2	Child1 to Child3	Child2 to Child3	Child3 to Child2
1	0	0	2.5	2.5
2	0	0	3	0
3	2.5	2.5	0	4
4	2.5	2.5	0	4

Both female and male speech signals were processed and the speech morphing algorithm was applied on these two signals where the female is the source signal and the male is the target one. The time domain of one frame of the source, target and morphed signals is shown in Figure 7. It is noticed here that although the morphing signal has approximately the same period as that of the target one but still there are noticeable differences in shape, duration and energy distribution. The pitch period of the whole signal (source, target and morphed) was monitored as illustrated in Figure 8. Obviously, pitch of the morphed signal is similar to that of the target one.

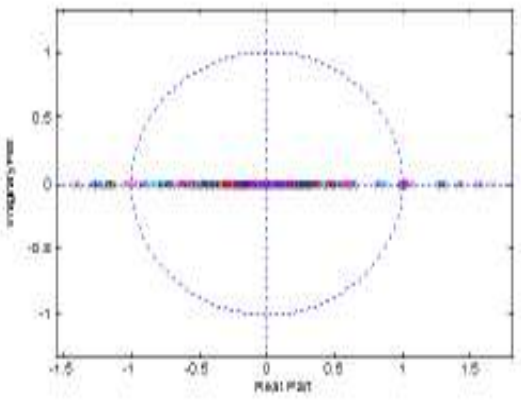

Figure 6. Filter coefficients for a child speech.

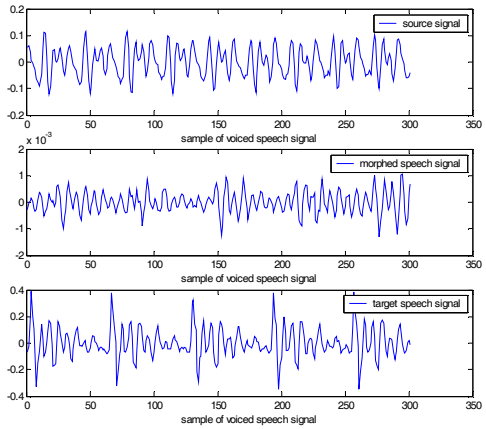

Figure 7. Speech signal of one frame for the source, the morphed and the target signals.

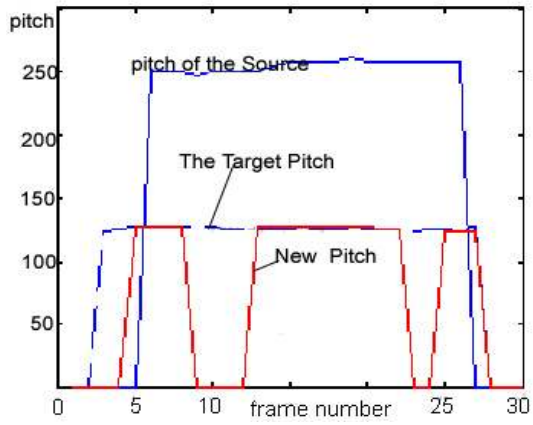

Figure 8. Pitches of source morphed and target speech signal.

The original complete source speech signal for the word "nine" pronounced by a female speaker, the target male speech and the synthesized male speech are illustrated in Figure 9. Obviously, speech morphing has changed the waveform but still can produce the desired target sound with some degradation in quality. The waveform differences could be due to the re-segmentation of speech or analysis and synthesis. The word "nine" has been processed several times and under several conditions and the results are shown in Table 3.

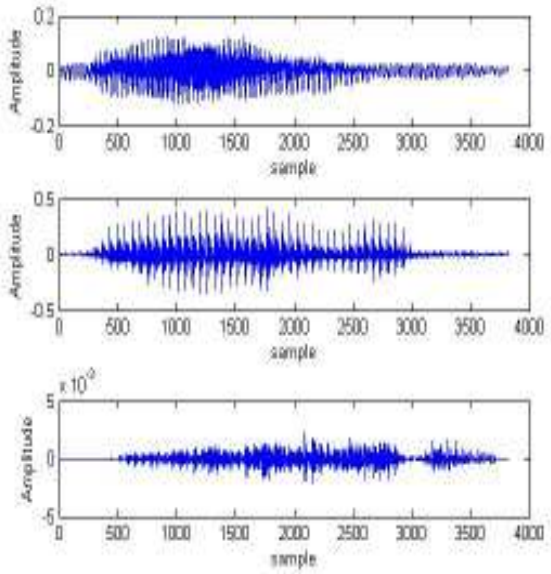

Figure 9. Source female, target male and synthesized speech signals respectively.

Table 3. MOS of morphing for the word "nine".

Test #	Female to Female	Female to Male	Male to Female	Male to Male
1	4	2.5	4	4
2	4	2.5	4	3.5
3	3.5	2.5	4	4
4	4	3	4	4
5	3.5	3	4	4
6	2.5	3	4	4

4. Discussion and Conclusions

Speech morphing was used to modify a source speaker sound to appear like another one's sound. Applying morphing to certain speech signals has achieved high MOS for many cases. However, for the male to female conversion case the performance was weaker than some other experiments, this could be since the pitch of the male has been changed to that of the female one while it is known that excitation function of male is stronger than the excitation of female. Children's speech was not that successful in some cases which could be due to using unmatched filter and excitation. The technique used here showed the ability to perform speech morphing and hence changing the speaker personality but did not preserve the naturalness. This could be due to the synthesis methodology used. Moreover, joint optimization techniques may be suggested here to determine the speech parameters in a better way.

References

[1] Ben G. and Nelson M., *Speech and Audio Signal Processing*, John Wiley and Sons, 2000.

[2] Dziubinski M. and Kostek B., "High Accuracy and Octave Error Immune Pitch Detection Algorithms," *Computer Journal of Archives of Acoustics*, vol. 29, no. 1, pp. 3-24, 2004.

[3] Ghulam M., Fukuda T., Horikawa J., and Nitta T., "A Noise-Robust Feature-Extraction Method Based on Pitch-Synchronous ZCPA for ASR," *in Proceedings of ICSLP04*, Korea, pp. 133-136, 2004.

[4] Hacioglu K. and Hasib A., "Pulse-By-Pulse Re-Optimization of the Synthesis Filter in Pulse Based Coders," *Computer Journal of IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 180-185, 1998.

[5] Hasib A. and Hacioglu K., "Source Combined Linear Predictive analysis in Pulse based Speech Coders," *Computer Journal of IEE Proceedings, Vision Image and signal Processing*, vol. 143, no. 3, pp. 143-148, 1996.

[6] Kondoz M., *Digital Speech, Coding for Low Bit Rate Communication Systems*, Press Wiley, 2004.

[7] Levinson S., *Mathematical Models for Speech Technology*, Press Wiley, 2005.

[8] Markel D., "The SIFT Algorithm for Fundamental Frequency Estimation," *Computer Journal of IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367-377, December 1972.

[9] Orphanidou C., "Voice Morphing," *MSc Thesis*, University of Oxford, 2001.

[10] Pfitzinger R., "Unsupervised Speech Morphing between Utterances of any Speakers," *in Proceedings of the 10th Australian International Conference on Speech Science and Technology*, Sydney, pp. 8-10, 2004.

[11] Rabiner R. and Juang B., *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[12] Veprek P. and Scordilis M., "Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques," *in Proceedings of Speech Communication*, pp. 249-270, 2002.

Allam Mousa is an associate professor in electrical engineering at An Najah National University and currently, the president assistant for P/D & quality. 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 became the chairman of the Electronics Engineering Department. In 2000, 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 processing, image and audio coding, and mobile communications. He has suppervised several MSc students in OFDM and telecommunication network planning. He teaches courses in telecommunication systems, digital communications, and DSP. He is also interested in higher education quality assurance and he is a senior member of IEEE.

.