Verification of TD-PSOLA for Implementing Voice Modification

Vivek Vijay Nar, Alice N. Cheeran, Souvik Banerjee
(Electrical Department, VJTI, Mumbai, India)
(Electrical Department, VJTI, Mumbai, India)
(Electrical Department, VJTI, Mumbai, India)

ABSTRACT

Voice modification is conversion of a model input voice signal into desired output signal. This can be achieved by modifying basic parameters of voice viz. vocal tract, pitch and time scale. Four models which are used for this purpose are discussed here namely LPC model, H/S model, TD-PSOLA model and MBR-PSOLA model. In this paper, TD-PSOLA technique is studied and implemented. Results of this implementation are derived and reviewed. The voice modification system thus developed can contribute greatly to the Medical and Entertainment Industry where desire voice modification is required.

Keywords – Vocal tract, formants, pitch, voice.

I. INTRODUCTION

It is noteworthy fact that for people the voice is not only just a tool for communication, but also an identifying feature that allows expression of personality. Researchers observed changes in formant tracks and pitch contours, during clinical diagnosis of certain diseases. Diseases like Laryngeal cancer can have a drastic impact on the speech leading to disturbances of voice. Rehan A. Kazi concludes that the formant frequencies of laryngectomy patients are higher than the formant frequencies of normal subjects [1]. Gregory K. Sewall claims that patients with Parkinson's-related dysphonia usually have reduced pitch ranges and increased vocal tremor [2]. In such cases a need arises for voice modification. In voice modification a model input voice signal is modified by varying vocal tract, pitch and time scale. Men have a larger vocal tract, which essentially gives the resultant voice a lower-sounding timbre. Whereas Female have a smaller vocal tract, giving the resultant voice a higher-sounding timbre. Adult male voices are usually lower-pitched and adult female voices are higher-pitched. Several works have been done for achieving voice modification and many methods are available for the same. Some of these are The classical AutoRegressive (LPC) [3], The hybrid Harmonic/Stochastic (H/S) [4],[5], Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) [6] and Multi-Band Re-synthesis Pitch-Synchronous Overlap-Add (MBR-PSOLA) [7]. These methods can process stored data as well as data in real time. E.g. in karaoke, pitch shifting is used to tune the user's singing so that it reaches the original melody in real-time [8]. In music composition, composers process music using pitch shifting [9]. Also it is used in applications like public speech system, entertainment Industry where specific voice style, tone, expressions are required.

Voice modification process using TD-PSOLA is explained in subsequent sections. Section 2 describes speech signal fundamentals and importance parameters of voice. The algorithms used for modification are compared in section 3. Section 4 gives the details of TD-PSOLA method. Implementation of method and results are interpreted in section 5 and 6. Section 7 summarizes and throws light on future scope.

II. SPEECH SIGNAL FUNDAMENTALS

The vocal cords periodically vibrate to generate glottal flow. Human pronounces a vowel or a voiced consonant which is composed of glottal pulses. Pitch period is the period of a glottal pulse. Fundamental frequency is reciprocal of the pitch period. The vocal tract acts as a time-varying filter to the glottal flow. The characteristics of the vocal tract include the frequency response, which depends on the position of organs, such as the pharynx and tongue. The peak frequencies in the frequency response of the vocal tract are formants, also known as formant frequencies. A speech signal is a convolution of a time-varying stimulus (Glottal Flow) and a time-varying filter (Vocal Tract) as shown in Fig.1.

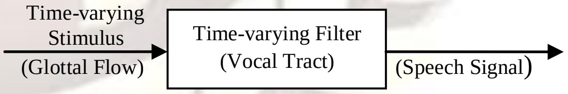

Fig.1. Modeling of speech signal

II. COMPARISON OF ALGORITHMS

The four models covered are:

1. The linear-prediction voice model is best classified as a parametric, spectral, source-filter model, in which the short-time spectrum is decomposed into a flat excitation spectrum multiplied by a smooth spectral envelope capturing primarily vocal formants [3].

2. The hybrid Harmonic/Stochastic (H/S) model proposed in [4] and [5]. In this speech signals expresses as the summation of slowly varying harmonic and stochastic components. This is based on Griffin's analysis algorithm, uses the OLA synthesis approach, the prosody matching and segment concatenation.

3. TD-PSOLA algorithms rely on a pitch-synchronous overlap-add approach for modifying the speech prosody and concatenating speech waveforms [6]. The modifications of the speech signal are performed in the time domain depending on the length of the window used in the synthesis process. The time domain approach provides very efficient solutions for the real time implementation of synthesis systems.

4. The Multi-Band Re-synthesis Pitch-Synchronous Overlap-Add (MBR-PSOLA) model is based on re-synthesis of the segments database of an original and efficient hybrid H/S [7]. It supports spectral interpolation between voiced parts of segments, with virtually no increase in complexity.

Result of Comparison

TD-PSOLA virtually exhibits no segments concatenation capabilities. LPC is slightly superior in this. LPC, hybrid H/S and MBR-PSOLA are superior to TD-PSOLA for automatic analysis procedures. An efficient segment database compression algorithm is ensured for LPC and hybrid H/S synthesizers. It is currently being developed for MBR-PSOLA. Fluidity is better of MBR-PSOLA and hybrid H/S due to superior concatenation capabilities. Prosody matching gives comparable results with all four models. TD-PSOLA is computationally simple compared to complexity of their respective synthesizers. It is much more intelligible than other synthesizers. It is much less sensitive to analysis V/UV errors than MBR-PSOLA. It is perceived as equally natural because it does not make use of any speech model. As seen from above, TD-PSOLA technique is better than other.

III. TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add)

This approach is based on the decomposition of the signal into overlapping frames synchronized with the pitch period. After modifications of the speech signal, consistency and accuracy of the pitch marks must be preserved [10]. For this pitch detection is performed to generate pitch marks through overlapping windowed speech.

Input signal $s[n]$ and $s_a[n]$ centered at t_a time is defined as:

$$s_a[n] = s[t_{a+n}]$$

Where, t_a is an analysis marks.

A short-time version of $s_a[n]$ by multiplying it by a window $w_a[n]$ is defined as $z_a[n]$:

$$z_a[n] = w_a[n] \times s_a[n]$$

The window length is two times of the local pitch period. To synthesize speech at different pitch periods, the Short Time signals (ST) are simply overlapped and added with desired spacing. The synthesized speech is defined as:

$$z[n] = \sum_{a=-\infty}^{\infty} z_n[n - t_a]$$

A good choice for the time marks t_a is to coincide with the instants of closing of the vocal folds which indicates the periodicity of speech. For unvoiced speech, these marks could be arbitrarily placed. This estimation from speech waveforms is a very difficult problem.

In speech analysis, a sequence of pitch-marks is provided after filtering the speech signal. Voiced/unvoiced decision is based on the zero-crossing and the short time energy for each segment between two consecutive pitch marks. A coefficient of voicement (v/uv) can be computed in order to quantize the periodicity of the signal [11]. To select pitch marks among local extreme of the speech signal, a set of mark candidates given with all negative and positive peaks. The OLA synthesis is based on the superposition-addition of elementary signals in the new positions. These positions are determined by the height and the length of the synthesis signal. To increase the pitch, the individual pitch-synchronous frames are extracted and given to Hanning window. Then output frame moved close together and added up, whereas output frame moved further apart to decrease the pitch. Increasing the pitch will result in a shorter signal, so to keep constant duration duplicate frames need to be added. A fast re-sampling method is used to shift the frame precisely, where it will appear in the new signal using the pitch mark and the synthesis mark of a given frame.

IV. Implementation

Following steps were implemented for achieving voice modification:

1. A wave file of mono sound quality, rate of 11025 with 8 bits per sample is given as input and amplitude vs. time graph was plotted.
2. Fast Fourier Transform of input file was taken and then magnitude (dB) vs. frequency graph of the same was plotted.
3. Input value of pitch scale, time scale and vocal tract were stored in a variable ‘a’, ‘b’, ‘c’ respectively which ranges from -1 to 1.
4. Using a polyphase filter implementation, resampling of the input at (P/Q) times of the original sampling rate is done. Resample applies an anti-aliasing (lowpass) FIR filter to input.
   Where,
   P = round (GValue*10); GValue = 2^c; Q=10.
5. Approximate pitch contour was calculated based on energy peaks for finding pitch marks and Path was found out using rridden function. By differentiating pitch

marks pitch period was found and first and last pitch marks were removed.

6. Analysis segment center was found. Then segment was stretched and new segment was added and amount of overlap between windows was defined.
7. Output signal so acquired, was plotted to get graphs mentioned in first two steps.
8. Finally both input and output files were played and result was concluded.

V. Results

In this study, a voice quality modification algorithm with TD-PSOLA modifier was implemented and tested. Normal female voice was taken as the reference and graphs plotted as shown in Fig.2 and Fig.3.

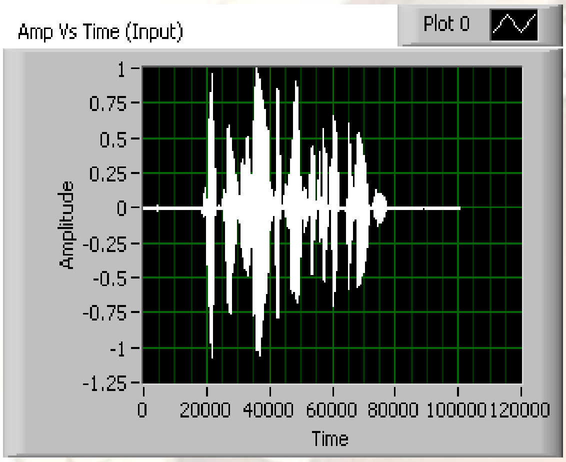

Fig. 2 Time domain graph of input

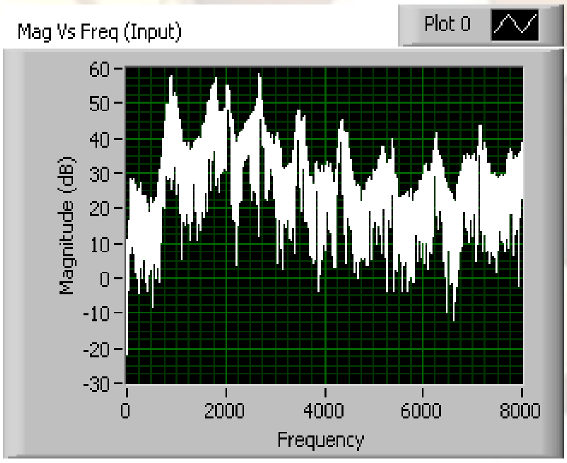

Fig. 3 Frequency domain graph of input

Above signal is then modified in different voices by varying parameters. After changing Vocal Tract parameter from 0 to +1 and from 0 to -1, the change in frequency domain graphs is shown in Fig. 4 and 5.

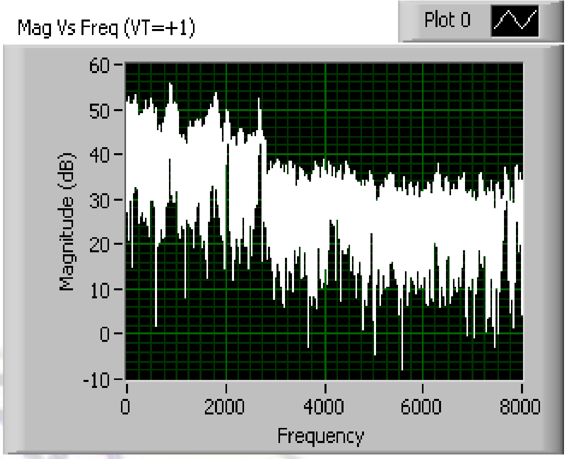

Fig. 4 FFT graph of VT= +1

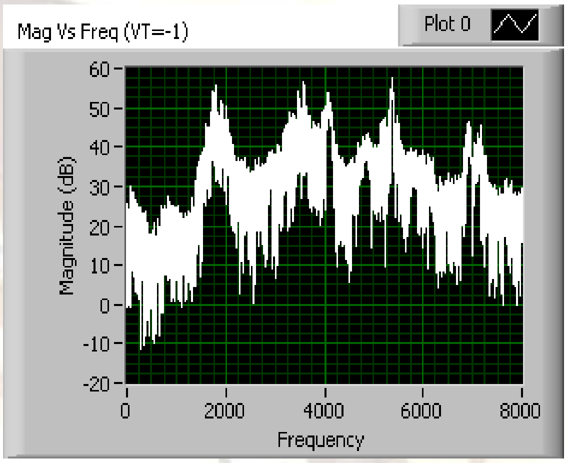

Fig. 5 FFT graph of VT= -1

After changing Time Scale parameter from 0 to +1, signal is expanded in time domain as shown in Fig. 6 and when change from 0 to -1 signal shrinks as shown in Fig. 7.

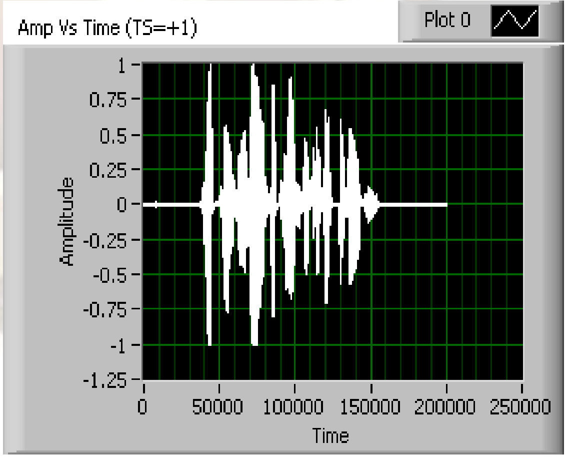

Fig. 6 FFT graph of TS= +1

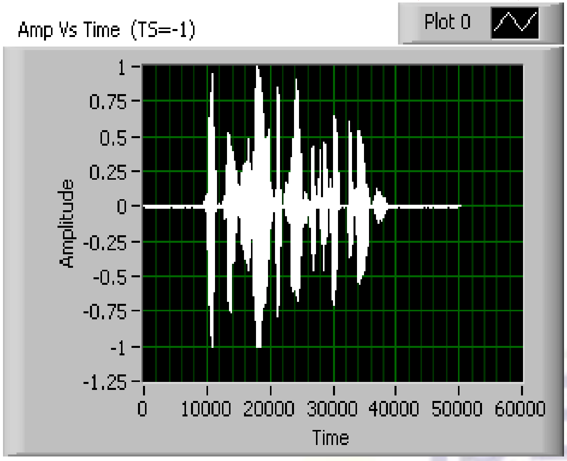

Fig. 7 FFT graph of TS= -1

After changing Pitch Scale parameter from 0 to +1 and from 0 to -1, the change in frequency domain graphs is shown in Fig. 8 and 9.

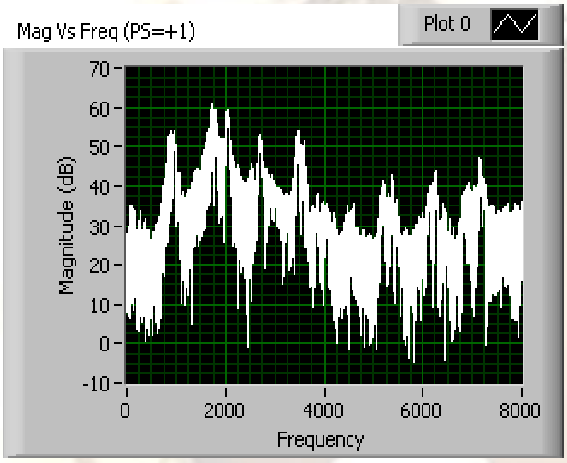

Fig. 8 FFT graph of PS= +1

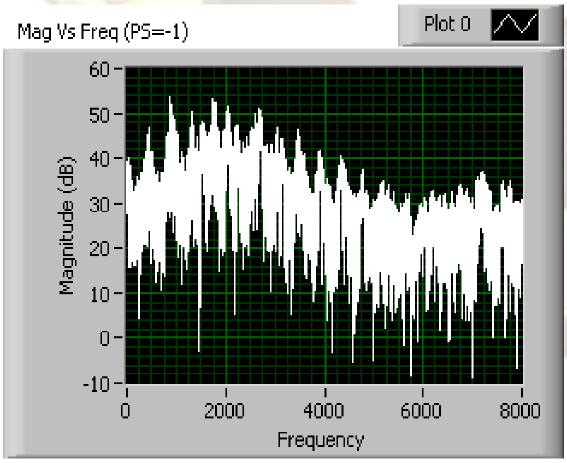

Fig. 9 FFT graph of PS= -1

VI. Conclusion and Future Scope

As seen from above results, since male have larger Vocal Tract, when VT parameter increased to +1, husky male voice is heard and decreasing it to -1 result into small child voice. Since female have higher pitch, when PS parameter increased to +1, nasal female voice is heard and decreasing it to -1 result into male voice. Same is summarized in Table below.

Table 1 Result of voice modification

Parameter	Parameter Values		
	-1	0	1
Vocal Tract	Small Child	Normal Female	Husky Male
Pitch Scale	Normal Male	Normal Female	Nasal Female
Time Scale	Fast	Normal Female	Slow

It is seen from above that vocal tract controls timbre of the voice signal whereas pitch varies tone of the same. Above results show that the algorithm can effectively modify input voice into the desired voice quality.

Results of the simulation verify that the quality of the synthesized signal by TD-PSOLA technique depends on the precision of the analysis marking and the synthesis marking. In future, Algorithm with better analysis and synthesis marking can be implemented for better voice quality and voice modification technique can be used for voice conversion applications.

References

[1] Kazi, Rehan A., Vyas M.N. Prasad, Jeeve Kanagalingam, Christopher M. Nutting, Peter Clarke, Peter Rhys-Evans, and Kevin J. Harrington, “Assessment of the Formant Frequencies in Normal and Laryngectomy Individuals Using Linear Predictive Coding”, *Journal of Voice 21, no. 6:661-668.*

[2] Sewall, Gregory K. MD; Jack Jiang, MD, PhD; and Charles N. Ford, MD, “Clinical Evaluation of Parkinson’s-Related Dysphonia”, *The Laryngoscope, 2006, 116:1740-1744.*

[3] J.D. MARKEL, A.H. GRAY Jr, ”Linear Prediction of Speech”, *Springer Verlag, New York, pp. 10-42, 1976.*

[4] D.W. GRIFFIN, J.S. LIM, “Multi-Band Excitation Vocoder”, *IEEE Trans. on ASSP, vol. ASSP-36, pp. 1223-1235, august 1988.*

[5] A. J. ABRANTES, J. S. MARQUES, I. M. TRANSCOSO, "Hybrid Sinusoidal Modeling of Speech without Voicing Decision", *EUROSPEECH 91, pp. 231-234.*

[6] E.MOULINES, F. CHARPENTIER, "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", *Speech Communication, Vol. 9, n°5-6. 1989.*

[7] T. DUTOIT, H. LEICH, "Improving the TD-PSOLA Text-To-Speech Synthesizer with a Specially Designed MBE Re-Synthesis of the Segments Database", *Proc.*

EUSIPCO 92, 25-28 august 92, Brussels, pp. 343-347.

[8] M. Ryynanen, T. Virtanen, J. Paulus, A. Klapuri, "Accompaniment Separation and Karaoke Application Based on Automatic Melody Transcription", *IEEE International Conference on Multimedia and Expo, 2008, pp. 1147-1420.*

[9] E. Gomez, G. Peterschmitt, X. Amatriain, P. Herrera, "Content-Based Melodic Transformations of Audio Material for a Music Processing Application", *Proc of 6th International Conference on Digital Audio Effects, London, UK, Sept 2003.*

[10] Abdelkader Chabchoub and Adnan Cherif, "Implementation of the Arabic Speech Synthesis with TD-PSOLA Modifier", *International Journal of Signal System Control and Engineering Application, 2010, Volume: 3, Issue: 4, pp. 77-80.*

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