Effective Implementation of Static Voice Alteration

Vivek Vijay Nar
Electrical Department,
VJTI, Mumbai, India

Alice N. Cheeran
Electrical Department,
VJTI, Mumbai, India

ABSTRACT
Voice alteration is the conversion of one speech signal into other, preserving the source content. For Voice alteration, two main parameters of speech must be considered viz. Static and Dynamic. In this paper, only static parameters are considered. The LP coefficients of source and target speech are extracted using LP analysis. The cross mapping of the extracted parameters is achieved by modifying source parameters in line with the target parameters using TD-PSOLA. Results illustrate that the TD-PSOLA method is reliable and efficient approach for voice alteration. The voice alteration system thus developed can contribute greatly to the Medical and Entertainment Industry where specific voice is essential.

General Terms
Speech Processing, Voice Alteration.

Keywords
Cross mapping, formants, pitch, static, voice.

1. INTRODUCTION
One of the unique functions of human being is ability to speak. Human speech production system is very important and complex process. Speech is an inherently human communication tool. Systems that concentrate on the intelligible content of speech have received widespread attention due to important applications in providing accessibility for disabled individuals [1]. Diseases like Laryngeal cancer, Parkinson's-related dysphonia can have a drastic impact on the speech in aspect of formant frequencies [2], pitch ranges and vocal tremor leading to disturbances of voice [3]. In such cases a need arises for voice alteration. So that such system can store parameters of particular patients and gives voices for them. It can also be used in many voice related applications such as public speech system, Voice restoration systems for people suffering from voice-impairing pathology. A survey of the current scenario in Speech Technology revealed that the main focus is on Text-to-speech and automatic speech recognition techniques. However, little attention has been provided to the field of voice conversion. This paper is an attempt to throw light on the methods for voice alteration. Voice alteration is the conversion of source speech signal to target speech without losing its information. This will result in a new audio signal of the same length with the information of the source signal. The major properties concerned for a speech signal are pitch and formant. These two reside in a convolved form in a speech signal. Hence some efficient methods for extraction of each of these are necessary. There are many feature extraction techniques such as LPC, MFCC, PCA [4]. Due to simplicity and efficiency of LPC approach [5], it is widely used for extraction of pitch and formant.

Voice alteration process is explained in subsequent sections. Section 2 describes human speech production system and important components of speech. The algorithms used for detection of formant and pitch are explain in section 3. Section 4 describes mapping of parameters required for voice alteration. Section 5 gives details of TD-PSOLA technique to modify excitation components. Implementation of method and results are interpreted in section 6 and 7 respectively. Section 8 summarizes and throws light on future scope.

2. HUMAN SPEECH PRODUCTION SYSTEM
Study of human speech production system is essential before starting with voice alteration process. This will also render the parameters which are responsible for voice distinction among all humans. The human speech production system begins with the lungs and end with mouth and nasal cavity. This complete process is controlled by neural signals from human brain [6].

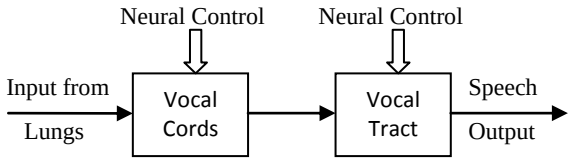

Figure 1: Block Diagram of the Speech Signal Production Process

Speech is comprised of two components viz. voiced component and unvoiced component. Voiced component is regularly spaced pulse-train; whereas unvoiced component is much noise-like. While modeling the excitation parameters of voiced components, pitch period is described by a periodic pulse train which is directly proportional to the resonance of vocal cords. The peak frequencies in the frequency response of the vocal tract are formants, also known as formant frequencies.

3. DETECTION OF FORMANTS AND PITCH
First stage of voice conversion requires extraction of formant and Pitch of source and target signal. The Linear Prediction Coding (LPC) method [7] applies an all-pole model to simulate the vocal tract. Figure 2 shows the flow chart of formant detection with the LPC method.

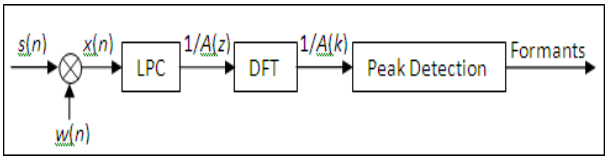

Figure 2: Formant Detection with the LPC Method [7]

In Figure 2, applying the window w(n) breaks the source signal s(n) into signal blocks x(n). Each signal block x(n) estimates the coefficients of an all-pole vocal tract model by using the LPC method. After calculating the discrete Fourier transform (DFT) on the coefficients A(z), the peak detection of 1/A(k) produces the formants.

Figure 2 shows the block diagram of pitch detection with the LPC method. This method [8][9] uses inverse filtering to separate the excitation signal from the vocal tract and uses the real cepstrum signal to detect the pitch.

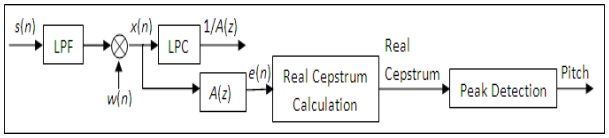

Figure 3: Pitch Detection with the LPC Method [8][9]

In Figure 3, the source signal s(n) first goes through a low pass filter (LPF), and then breaks into signal blocks x(n) by applying a window w(n). Each signal block x(n) estimates the coefficients of an all-pole vocal tract model by using the LPC method. These coefficients inversely filter x(n). The resulting residual signal e(n) passes through a system which calculates the real cepstrum. Finally, the peaks of the real cepstrum calculate the pitch.

4. MAPPING PARAMETERS FOR VOICE ALTERATION

After the LP analysis, LP Coefficients for both source and target have been extracted. These parameters are used to model a Vocal Tract Filter of Target Speech and an Inverse-Filter to extract excitation component of source speech. Now this excitation component is applied to the All Pole Vocal Tract Filter (Shaping Function), which 'Shapes' the spectrum of source excitation. Thus, by this filtering operation, spectral shaping of source signal is achieved. The modified spectrum has formants at the target formant frequencies [10]. Figure 4 shows the block diagram of same.

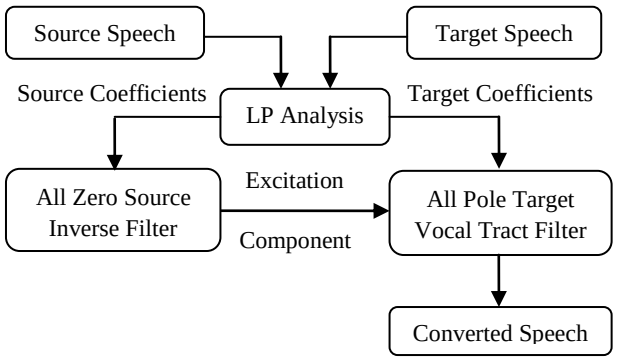

Figure 4: Block Diagram of Pitch and Vocal Tract Mapping Process

5. MODIFYING EXCITATION COMPONENT

TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) which modifies the source component so that the pitch of the signal approaches target pitch without any change in the time-scale [11]. The aim of this approach is to modify the source pitch to match the target pitch. Increment of the pitch value will lead to compression in timescale, whereas decrement of the pitch period will lead to expansion of time scale and the speech will no longer remain intelligible. The goal of pitch modification is to modify the pitch of a speech signal without losing its information. This will result in a new audio signal of the same length, which will have the information of the original signal at a desired target pitch. The classical Auto Regressive (LPC), the hybrid Harmonic/Stochastic and various SOLA methods are used conventionally. However PSOLA has proven to give the best results and comparison of the same is tabulated below considering different parameters [12]. PSOLA can be further classified as TD-PSOLA and FD-PSOLA. Time Domain Pitch Synchronous Overlap-Add is used because of lesser complexity involved.

Table 1: Comparison of LPC, hybrid H/S, TD-PSOLA [12]

	LPC	Hybrid H/S	TD-PSOLA
Analysis	automatic, easy	automatic, requires a careful design	semi-automatic (pitch marking)
Prosody matching	Good	Good	Good
Segments concatenation	Best	very good	poor
Modelization quality	poor	very good	perfect (no model)
Intelligibility	low	high	almost perfect
Fluidity	low	very high	fair
Naturalness	low	high	very high

6. IMPLEMENTATION

Following steps were implemented for achieving voice alteration:

1. A wave file of mono sound quality, rate of 11025 with 8 bits per sample was given as input and amplitude vs. time graph was plotted for the input speech signal.
2. Then a Fast Fourier Transform of input signal was taken and then magnitude (dB) vs. frequency graph of the same was plotted.
3. 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.
4. Framing is done using hamming window with fs = 16000.
5. The steps 1 to 5 are repeated for target speech signal.
6. Using source frame and order 18, excitation parameters were obtained.
7. Using target frame and order 18, filter coefficients were obtained.
8. Using excitation parameters and filter coefficient obtained from step 7 and 8 synthesis speech is generated.
9. Output signal so acquired, was plotted to get graphs mentioned in first two steps.
10. Finally both input and output files were played and result was concluded.

7. RESULTS

As discussed above, only static conversion is considered in this paper. An experiment was carried out with four different voices viz. husky male voice (M1), normal male voice (M2), normal female voice (F1) and nasal female voice (F2) in order to verify this technique. These voices were recorded in best possible noise-free environment. Keeping one of them as a source, same is converted into rest of the voices which are considered as target. This formed twelve different combinations. Time domain, frequency domain graphs of source, target and converted speech were plotted. One set of these plots is shown below.

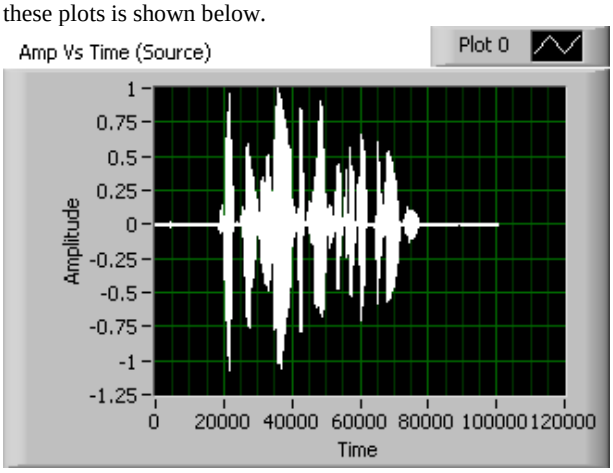

Figure 5: Time domain graph of source speech

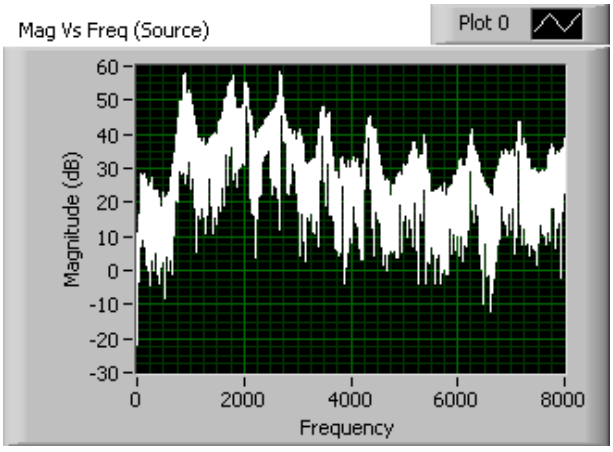

Figure 6: Frequency domain graph of source speech

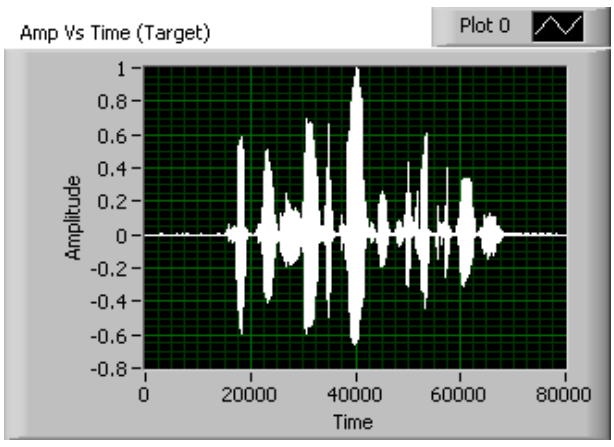

Figure 7: Time domain graph of target speech

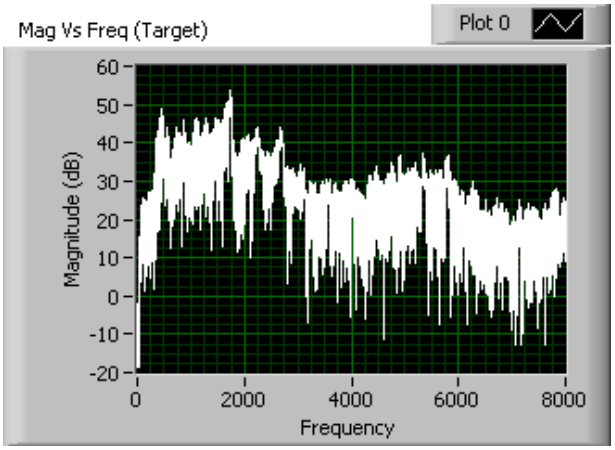

Figure 8: Frequency domain graph of target speech

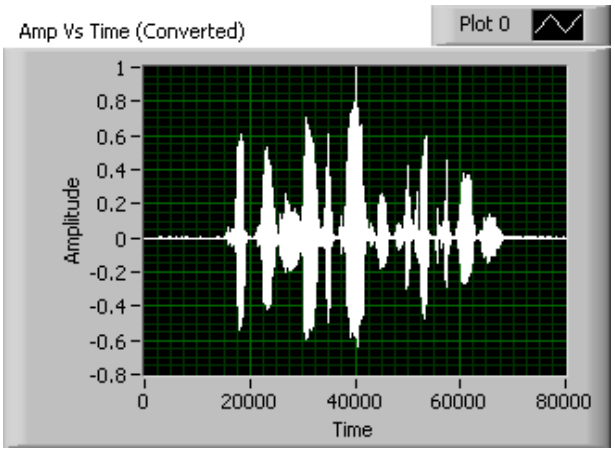

Figure 9: Time domain graph of converted speech

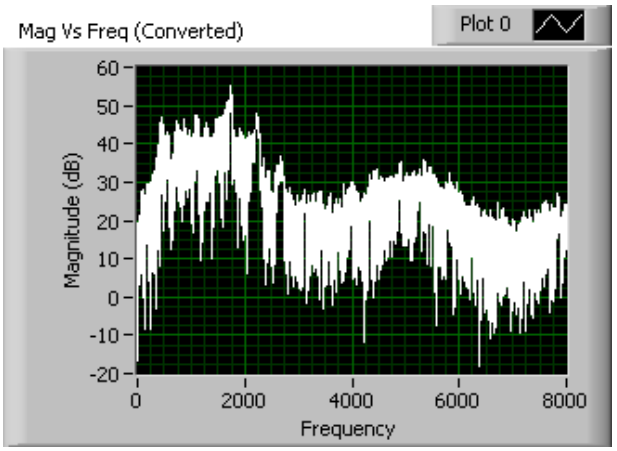

Figure 10: Frequency domain graph of converted speech

8. CONCLUSION AND FUTURE SCOPE

The output of this experiment was reviewed by ten people of different age groups, after listening to the output they rated the accuracy of conversion on the scale of 5. Average of these results was evaluated and figured out in following chart.

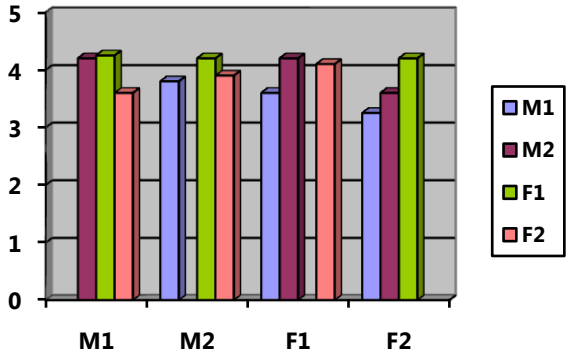

Figure 11: Result of voice alteration

As evident from the above results, overall accuracy of conversion is 3.9 out of 5. This satisfying result shows the TD-PSOLA method is a reliable and efficient approach for voice alteration.

It is also observed that the accuracy of voice alteration in husky male voice (M1) and nasal female voice (F2) is less as compared to normal male voice (M2) and female voice (F1). The reason is, being a static system, it does not account for the tone or other dynamic parameters of the voice. Same can be avoided using dynamic conversion system. It is further seen that, the accuracy of conversion diminishes due to the noise present in source voice while feeding as input as the same gets amplified during conversion. To achieve better quality of the conversion, noise should be removed. This can be removed at the input stage or can be filtered out in the system. So better voice quality can be acquired from such systems and accurate results can be achieved in many applications.

9. REFERENCES

[1] Anderson F. Machado and Marcelo Queiroz,“ voice conversion: a critical survey”. Pp 1-8, 2010.

[2] 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.

[3] 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.

[4] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar “A Review on Speech Recognition Technique”. International journal of computer application (0975-8887), volume10- No 3, November 2010.

[5] Oytun Turk, Levent M. Arslan, “Voice Conversion Methods for Vocal Tract and Pitch Contour Modification”. Pp 1-4.

[6] Markel, J. D. “Digital Inverse Filtering—A New Tool for Formant Trajectory Estimation”. IEEE Transactions on Audio and Electroacoustics 20, no. 2:129-137, 1972.

[7] Noll, A.M. “Cepstrum Pitch Determination”, Journal of the Acoustical Society of America 41 (February): 293-309, 1967.

[8] Markel, J.D. “The SIFT Algorithm for Fundamental Frequency Estimation”, IEEE Transactions on Audio and Electroacoustics 20 (December): 367-377, 1972.

[9] Ronald.W.Schafer, “A Survey of Digital speech processing Technique”. IEEE transactions on audio and electroacoustics volume AU-20, No 1 1997.

[10] Reinier W. L. Kortekaas and Armin Kohlrausch, “Perceptually weighted linear transformation for voice morphing”. J. acoust. Soc. Am. Volume 101, No 4 April 1997.

[11] 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.

[12] Thierry Dutoit, Henri Leich, “A Comparison of Four Candidate Algorithms in the context of High Quality Text-To-Speech Synthesis”,ICASSP’94, Adelaide, Australia.