Voice Conversion Algorithm

Akash I. Mecwan
Lecturer in E & C Department
Institute of Technology, Nirma University, Ahmedabad-382481,
Gujarat, India.
+91-02717-241911-15 (Ext:422)
akash.mecwan@nirmauni.ac.in

Vijay G. Savani
Lecturer in E & C Department
Institute of Technology, Nirma University, Ahmedabad-382481,
Gujarat, India.
+91-02717-241911-15 (Ext:422)
vijay.savani@nirmauni.ac.in

Shah Rajvi
Student, E &C
Institute of Technology, Nirma University, Ahmedabad-382481,
Gujarat, India.

Priya Vaya
Student, E &C
Institute of Technology, Nirma University
Ahmedabad-382481, Gujarat, India.

ABSTRACT

Recently, a lot of work has been done in the speech technology. The main concentration being on Text-to-speech and automatic speech recognition techniques, voice conversion is yet an undeveloped and naïve field in Speech Technology and a lot of contribution from speech researchers is expected in upcoming days. In this paper an approach for static voice conversion is discussed. Static speech parameters are the parameters over which speaker has least control such as vocal tract structure, natural pitch of speech etc. Here, two main parameters are considered Vocal Tract Structure and Pitch. For conversion process speech is resolved in two components, excitation component and filtered component using a Linear Predictive coding [LPC] based source-filter. The pitch contour is determined by an autocorrelation. The excitation component is generated using a set of signal generators generating the determined pitch and are driven by voicing detection. Filter coefficients are modified to approach target speaker coefficients for voiced segments and for unvoiced segments filter coefficients of source are used straightaway.

1. INTRODUCTION

Voice Conversion is a process of transforming the parameters of speech uttered by a source speaker such that a listener would perceive as if spoken by another target speaker. This has numerous applications such as in Karaoke applications, Text-to-speech synthesis, Dubbing Industry etc. The voice conversion process is carried out in three phases.

a. *Analysis*:

In this phase voice parameters corresponding to speaker identity are extracted from source as well as target speech. As stated vocal tract shaping function and pitch are the main parameters to be extracted in this phase.

b. *Mapping*:

In this phase the extracted parameters of source speaker are mapped such that they combine to produce target speech.

c. *Synthesis*:

The modified parameters are used to synthesize or reconstruct the new speech which shall sound like target speech. A source filter model is used to synthesize the speech in which the new vocal tract parameters are used to represent a filter, and using the pitch value a source/ excitation component is generated which is then passed through the vocal tract filter.

2. THEORY OF SPEECH PRODUCTION SYSTEM

The human speech production system is shown in "Fig. 1".

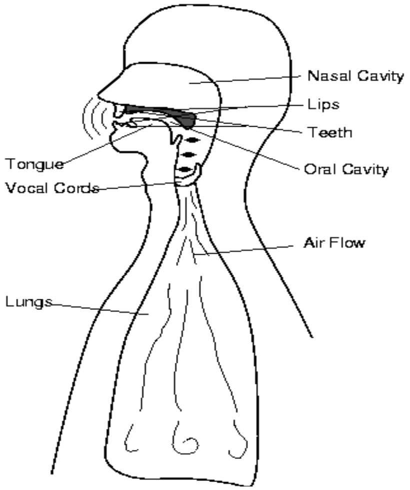

Fig.1. Human Speech Production System

Speech signals can be modeled as unstructured signals generated by a source and passed through interconnection of systems, which structures the signal to yield speech. The system can be modeled either as a linear or a nonlinear model. Though a linear model does not mimic the exact behavior, it is preferred as it provides a fair amount of accuracy with ease of implementation.

During human speech production, air flows from lungs, passes through vocal cords first. When the vocal cords are tensed, the

ICAC3'09, January 23–24, 2009, Mumbai, Maharastra, India.
Copyright 2008 ACM 978-1-60558-351-8.

airflow causes them to vibrate and hence output of this stage is a periodic signal, such signals are called voiced components of the speech. When the vocal cords are relaxed, air flows more freely through them resulting in turbulent flow of air and hence output airflow after this stage is very disordered. Such components are called unvoiced components. Amount of tension on vocal cords is driven by neural signal depending on what is to be spoken.

A common observation suggests that all the vowels are voiced; consonants whose production involves throat are also voiced. Rest of the consonants whose production is caused by oral [lips, tongue and mouth] and nasal cavities are unvoiced. So speech production model depends on whether a vowel is being spoken or a consonant. To be more precise whether the speech produced is voiced or unvoiced. So, periodic airflow coming out of the vocal cords can be described by a periodic pulse train with its period T, hence F0 [= 1/T] is the pitch of speech signal.

After passing through the vocal cords the thus produced airflow enters the mouth creating some amount of acoustic disturbances and exits through lips and some times through nasal cavity. The mouth, tongue, teeth, lips [oral cavity] and nasal cavity are named together as vocal tract. The cross-section of this vocal tract tube varies along its length because of varying positions of teeth, tongue and lips. These positions are determined by neural signals depending on which speech component is to be produced [2]. For example, the production of sound "ee" involves spreading the lips and bringing the teeth nearer. These variations result in a Linear Time Invariant System which has a frequency response as shown in "Fig. 2". The peaks on the frequency response curve are referred to as formants, so rather than naming it a filter, it is termed a shaping function which shapes the spectrum of airflow through vocal cords.

3. VOICE CONVERSION SYSTEM

In this paper source filter model based approach is discussed to carry out voice conversion. The source filter model is a model of speech where the spoken word is comprised of a source component originating from the vocal cords, which is then shaped by a filter imitating the effect of the vocal tract. "Fig 3" shows the source-filter model of speech. *x(t)* is the input to the filter and is called the excitation signal (output of vocal cords) since it excites the vocal tract. The vocal tract is a Linear Time-Invariant system with impulse response *h(t)*. This is sometimes called the shaping function of speech since it shapes the spectrum of excitation signal. The output of this shaping function is speech *y(t)*.

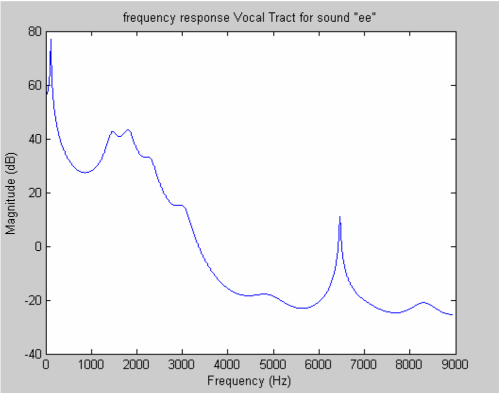

Fig 2. Frequency Response of Vocal Tract

As stated earlier the process of voice conversion is carried out in three phases. Analysis, Mapping and Synthesis.

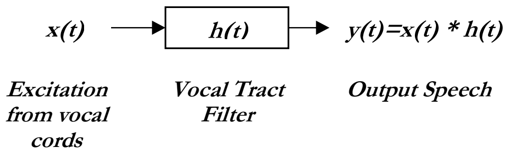

Fig. 3. Source Filter Model of Speech

a. Analysis Phase

In this phase voice parameters corresponding to speaker identity are extracted from source as well as target speech. For this a source-filter model based on LP Coefficients is used. In this model two parameters, excitation component and filter component (Vocal Tract model) are extracted from speech signal.

3.1. Vocal Tract Model

A generalized observation of human speech production states that a speech waveform is modeled as an output of a time varying all-pole filter driven by a source component (source can be a periodic signal, noise or mixture of this two). So, transfer function of vocal tract can be approximated as,

$$V(z) = \frac{Gz^{-p/2}}{1 - \sum_{j=1}^{p} a_j z^{-j}}$$

$$V(z) = \frac{Gz^{-p/2}}{A(z)}$$

$$V(z) \approx \frac{G}{A(z)} (delay - z^{-p/2} neglected)$$

The task of filter modeling is to approximate a_j coefficients such that the filter frequency response tracks the speech spectrum perfectly. Out of several available methods to obtain a_j coefficients such as LP Analysis, Cepstrum Analysis, Line Spectral Frequencies etc., LP Analysis is chosen to model filter coefficients [4]. The reason is this method is well documented in speech literature and computationally efficient then other methods. The drawback of this method is that it does not provide stability check.

Prior to analysis the speech signal is passed through a pre-emphasis filter in order to reduce the dynamic range of speech spectra. The pre-emphasized speech is then segmented into short-term analysis frames using a Hamming window. 30ms duration is chosen, to cover at least 2 pitch periods. Hamming window is used due to its tapered frequency response so that it reduces the effect of discontinuities at the beginning and end of each analysis frame. Hamming window is described by the following equation,

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N-1}\right)$$

$$0 \le n \le N-1$$

N is the window duration

Hamming window and its frequency response are shown in "Fig. 4". An important point to note is that the window affects the temporal gain characteristic of segment, and hence the next window is applied so that it has some amount of overlap with its previous window. The amount of overlap is kept 10ms here.

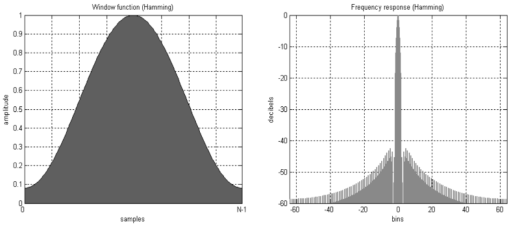

Fig 4.Hamming Window (left) & Frequency Response (Right)

Linear Predictive Analysis

Linear prediction is based on the fact that in slowly varying signals it is possible to predict the future sample based on the values of a few past samples [1], the number of samples used to predict the next value is called prediction order *p*. The Linear Prediction Equation is given by,

$$s(n) = Gu'(n) + \sum_{j=1}^{p} a_j s(n-j)$$

$$\therefore \hat{s}(n) = \sum_{j=1}^{p} a_j s(n-j)$$

$a_j s$ are called LP Coefficients and are obtained such that the Mean Square prediction error is minimum.

Hence Prediction Error e(n) is:

$$\therefore e(n) = s(n) - \hat{s}(n)$$

$$\therefore e(n) = s(n) - \sum_{j=1}^{p} a_j s(n-j)$$

Taking Z-Transform:

$$E(z) = S(z) - \sum_{j=1}^{p} a_j S(z) z^{-j}$$

$$\therefore E(z) = S(z).\left[1 - \sum a_j z^{-j}\right]$$

$$\therefore E(z) = S(z).A(z)$$

$$\therefore S(z) = \frac{E(z)}{A(z)} = Gu'(z).\frac{1}{A(z)}$$

1/A(z) represents vocal tract transfer function and Gu'(z) is excitation component. The coefficients a_j are obtained using either autocorrelation method or covariance method. These methods lead to a set of equations in terms of a_js known as Yale-Walker equations, which can be solved using Levinson-Durbin algorithm.

"Fig 5" Shows the result obtained by implementing the vocal tract filter as suggested by above method in MATLAB 7.0. Filter coefficients are obtained from a segment of speech by using a LP analysis of prediction order 20.

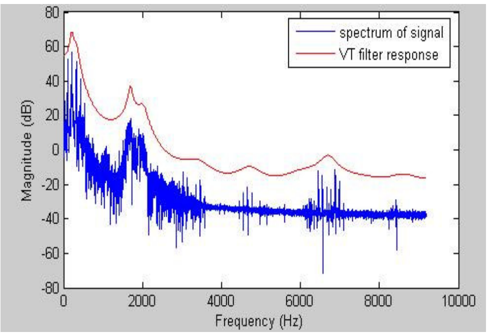

Fig 5. Response of filter characterized by LP coefficients

3.2. Pitch Period computation

Pitch detection has always been a complex issue in speech processing. With many pitch detection algorithms proposed in many years [6], a general observation states they are context specific algorithm and works well on specific content only. The basic time domain methods for pitch detection are zero-crossing rate, autocorrelation method and covariance method, but speech signal not being a *"pure periodic signal"*, these methods are largely affected by harmonics present in the signal and hence are not very robust. So, a mixed approach is used here for pitch computation. The method basically uses autocorrelation with voiced/unvoiced detection.

To detect the pitch, we take a window of the signal, with a length at least twice as long as the longest period that might occur. First we check if the window is voiced or not, this is done by comparing the mean energy level of the window with a threshold.

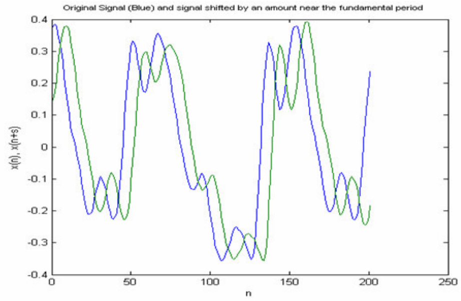

Fig.6. Representation of Autocorrelation at a particular shift

If the window is voiced then the method explained below is used. Human speech does not go below 50Hz in general. This corresponds to duration greater than 20ms. So window size of 30ms duration is considered. The next window has 10ms overlap with the previous window for the reason explained in LP analysis. Using this section of signal, we generate the autocorrelation function *r(s)* defined as the sum of the point wise absolute difference between the two signals over some interval. "Fig 6" shows how the signals begin to align with each other as the shift amount nears the fundamental period.

Intuitively, it should make sense that as the shift value *s* begins to reach the fundamental period of the signal T, the difference between the shifted signal and the original signal will begin to decrease. Indeed, this can be seen in the "Fig 7" in which the autocorrelation function rapidly approaches zero at the fundamental period.

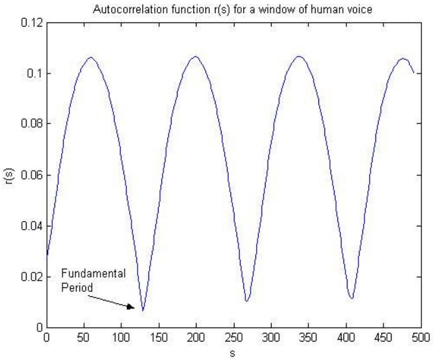

Fig.7. Autocorrelation function of a speech signal

As this involves difference at each sample, this becomes computationally very expensive, so in effort to improve the efficiency of this algorithm, an alternative called Fast Autocorrelation [3] is used. Here, the nature of the signal is exploited, specifically the fact that if the signal was generated using a high sampling rate and if the windows are narrow enough, it can be assumed that the pitch will not vary drastically from window to window. Thus, we can begin calculating the *r(s)* function using values of s that correspond to areas near the previous minimum. This means that, if the previous window had a fundamental period of 156 samples, calculate *r(s)* for s = 136. If the minimal s in this area cannot be found, we calculate further and further from the previous s until we find a minimum. Also, the first minimum (valued below the threshold) is always going to correspond to the fundamental frequency. Thus, we can calculate the difference equation *dr(s)/ds* can be calculated as *r(s)* is calculated and as the first minimum below threshold is encountered, next window is fetched.

b. Mapping:

After extracting the source and target parameters from source and target speech as explained in analysis phase, the extracted parameters of source speaker are mapped such that they combine to generate target speech. The key factor driving whole modification process is voiced/unvoiced decision for each window. After the pitch and filter coefficients are found, a codebook is made, containing five fields for each window,

1. Voiced/ Unvoiced Flag
2. Filter coefficients (target)
3. Mean value of gain (target)
4. Source Pitch
5. Target Pitch

c. Synthesis:

In this process, first excitation component is generated from extracted parameters. These parameters are target pitch, target filter coefficients and voicing detection [8]. Based on the voicing flag, the excitation signal is generated as a pulse train for voiced frames of speech. An impulse train generator is used to generate a pulse of unit amplitude at the beginning of each pitch period. For unvoiced excitation, a random noise generator produces a uniformly distributed random signal. The amplitude of the generated excitation signal is scaled by gain value and then passed through a filter characterized by LP coefficients of target speech. This process results in output speech windows, which are added with same amount of overlap used at the time of synthesis. The whole synthesis process is shown in the figure,

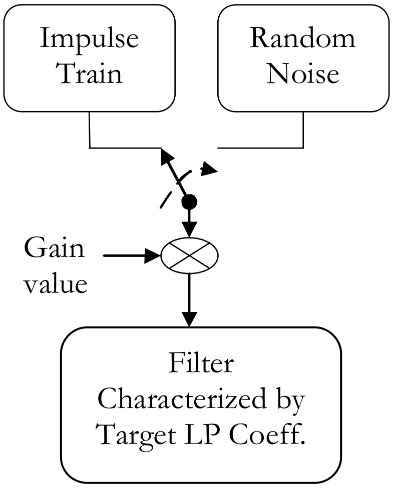

Fig.8. Synthesis Process controlled by voicing detection

4. CONCLUSION

A method developed here gives primitive insight into the field of voice conversion. The system discussed here processes on pre-time aligned speech samples. Also the current method can be used to modify the speech only when the target utterance is present. Numerous efforts and modifications can be implemented to make the present system more robust, efficient and generic. One of such modification is training. An ideal voice conversion system should include a training phase so that the system can be trained with target speech and can be used to convert any arbitrary speech uttered by source speaker. High quality transformations can be obtained with more complex and computationally expensive techniques. Also real-time voice conversion can be achieved with powerful Digital Signal Processors or similar Hardware. Voice conversion is yet an unexplored field in speech technology and expects a lot of contribution from speech researchers in future years.

6. REFERENCES

[1] Ben gold and Nelson Morgan, "Speech and audio signal processing". Wiley India, 2000.

[2] Don Johnson, "Modeling the Speech Signal", Unpublished.

[3] Gareth Middleton, "Pitch Detection Algorithms", Unpublished.

[4] Hui Ye and Steve Young, "Perceptually weighted linear transformation for voice morphing", Unpublished.

[5] L.R.Rabiner and R.W.Schafer, "Digital processing of speech signal" Pearson education, 1992.

[6] Lawrence R. Rabiner, Michael J. Cheng, Aaron E. Rosenberg And Carol A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", *IEEE Transactions On Acoustics, Speech, And Signal Processing*, vol.assp-24, no. 5, 1976.

[7] Levent Arslan , "Speaker Transformation Algorithm using Segmental Codebooks (STASC)", *Speech Communication*, vol.28, 1999.

[8] M.M.Hasan, A.M.Nasr and S.Sultana, "An approach to voice conversion using feature statistical mapping", *Applied Acoustics*, vol. 66, 2006.

[9] Oytun Türk and Levent M. Arslan, "Robust processing techniques for voice conversion", *Computer Speech and Language,* vol.20, 2006.

[10] Oytun Türk, "New Methods for voice conversion", Unpublished.