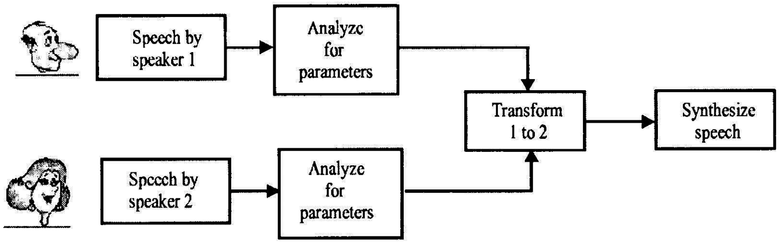
Voice Transformation

Mark Tse
Columbia University

EE6820 Speech and Audio Processing Project Report
Spring 2003



Abstract

Voice transformation is a technique that modifies a source speaker's speech so it's perceived as if a target speaker had spoken it. It falls into the general category of speech modification which is a subject of major interest today, with numerous applications including text-to-speech synthesis, preprocessing for speech recognition, voice editing, broadcasting, and entertainments, etc. Efficient speech synthesis and modification methods like Pitch-Synchronous-OverLap-Add (PSOLA) are widely used in many systems [1]. In recent years, other speech modification models such as sinusoidal model [2], and the harmonic plus noise model (HNS) [3] have also been presented. These advanced models, although computationally intensive, can produce very high quality transformed speech quality especially when used within the framework of concatenative speech synthesis.

This report presents the investigation and implementation of a low-order voice transformation system that is capable of transforming speech uttered by a male speaker to one that sounds as if it was uttered by a female speaker and vice versa. Like most of the other existing models, it is implemented within the popular LPC speech analysis/synthesis framework. It requires minimal training using only one pair of sentences from both the source and target speakers. Modification of other prosodic parameters such as duration and intensity, although important for some of the aforementioned applications, do not contribute significantly as far as the objective of this project is concerned, and therefore were not considered in this project due to time constraints.

1. Introduction

Voice transformation, in general, refers to the process of changing voice personality, i.e., speech uttered by a source speaker is modified to sound as if a target speaker had uttered it. Transformation is usually performed in two stages. In training stage, acoustic parameters of the speech signals uttered by both source and target speakers are computed and appropriate rules mapping the acoustic space of the source speaker onto that of the target speaker are obtained. In the transformation stage, the acoustic features of the source signal are transformed using the mapping rules such that the synthesized speech possesses the personalities and voice quality of the target speaker.

The particular voice transformation system presented in this report was developed with the specific aim of transforming only the voice characteristics of a speech utterance that are associated with gender identity. As such, throughout this report, ‘voice transformation’ refers to the narrower definition of gender voice transformation. And unless otherwise specified, ‘source’ and ‘target’ refer to speech from speakers of the opposite gender.

In this report, I will describe some of the techniques I have experimented with and present the results of informal listening tests. The report is organized as follows. In section 2, the source-filter model for speech synthesis is reviewed. In section 3, detail descriptions of the implementation approach and techniques are given. In section 4, results of informal listening tests are presented. Sections 5 and 6 summarize the work of this project and outline future work and investigations.

2. Source-filter Speech Production Model

The underlying model of speech production involving an excitation source and a vocal tract filter is implicit in many speech analysis methods. Physiologically speaking, voicing occurs in the larynx, where airflow from the lungs is pushed through the vocal cords and vocal tract and out from the lips and nose airways. For voiced sounds, the puffs of air produced by the opening and closing of the vocal folds generate a quasi-periodic excitation for the vocal tract. The fundamental frequency of the vocal fold vibration is known as F0 and the perceptual feature of speech corresponding to F0 is often called pitch. For unvoiced sounds, air flows through an open vocal cords and the air stream is forced through a narrow orifice in the vocal tract to produce a turbulent, noise-like excitation. Unvoiced speech sounds are usually characterized as aperiodic and noisy. Thus, from an engineering point of view, speech sounds are produced from a combination of this source of sound energy modulated by a time-varying acoustic filter determined by the shape and size of the vocal tract. This results in a shaped-spectrum with broadband energy peaks whose frequencies are known as formants. The spectrum of voiced sounds is primarily shaped by these resonant formant frequencies and has most of its power in the lower frequency bands, whereas the spectrum of unvoiced sounds is non-harmonic and usually has more energy in higher frequency bands. This model of speech production is known as the source-filter model and is shown in Figure 1. This source-filter concept leads directly to engineering methods to separate the source (the excitation signal) from the filter (the time-varying vocal tract transfer function) for independent manipulations. One of the procedures for implementing this separation is the Linear Predictive Coding (LPC) method [5].

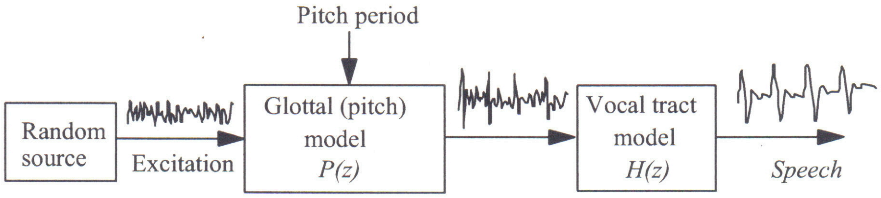


Figure 1. A source-filter model of speech production

3. Implementation Approach

The algorithms described here are based on the LPC analysis/synthesis framework. It should be noted that gender is conveyed in part by the vocal tract characteristics and in part by the pitch value of speech sounds. LPC analysis method allows parameterization of speech sounds by separating the source (excitation containing the pitch information) from the filter (vocal tract characteristics). The transformation procedure therefore involves mapping of both the pitch values and spectral parameters that characterize the vocal tract responses of the source and target speakers. The general steps that carry out the transformation are as follows. The same sentence uttered by both the source and target speakers are first broken into short speech frames. The frames are then time-aligned using dynamic time warp (DTW) technique that minimizes a distortion measure. Each pair of speech frames are analyzed and decomposed into the excitation (or residual) component and the filter component, which is described by a set of LPC filter coefficients. Linear regression least square estimation is then used to compute the transformation parameters for mapping LPC filter coefficients and pitch values. After these transformation parameters are obtained, a new 'test' sentence uttered by the source speaker is then input to the "trained" system for spectral envelope and pitch mapping on a frame-by-frame basis. The transformed speech is then synthesized using the LPC speech synthesis method.

As noted above, pitch periods associated with male speakers are generally quite different from those associated with female speakers, modifying the pitch contour of a speech signal alone can often result in some level of voice transformation. Indeed, increasing or decreasing the pitch periods of a speech signal had been shown to be capable of modifying the apparent gender of the speaker. On the other hand, it is also well known that the vocal tract transfer function (spectral characteristics) is the dominant factor associated with speaker individuality. In this project, I had experimented with voice transformation based on mapping LPC filters alone, pitch periods alone, and mapping of both LPC filters and pitch periods. The results are presented in section 4 below.

3.1 Spectral Transformation

During training phase of the voice transformation process, the time-aligned source and target speech frames are first decomposed into the filter and excitation components using linear prediction method (autoregression). The frame size here is chosen to be 128 speech samples, which corresponds to 16 ms of speech at a sampling rate of 8 KHz. This frame size should allow at least one pitch pulse but no more than a handful of pitch pulses to be processed per frame. A 12th order all-pole LPC filter is used in this system.

The filter transfer function for each 128-sample frame of speech is therefore characterized by a set of 12 LPC coefficients. Before training and mapping are performed, these coefficients are converted to Line Spectral Pairs (LSP) representation for its excellent interpolation properties. Linear regression using least square method is then employed to find the parameters that map the transfer function of the source signal to that of the target in the least square sense, i.e., it finds the values of *b0* and *b1* such that the square of estimation error is minimized. The error here is given by *E = Y – (X*b1 + b0),* where *X* and *Y* represent the source and target speech transfer functions respectively. The translated LSP coefficients are then converted back to the regular LPC coefficients for final synthesis. From experiments, I noticed that the mapped filters could occasionally become unstable and produce sporadic speech frames with much higher energy than the rest of the frames. This is despite using LSP representation during mapping. I was able to repair some of these rogue filters by using the Matlab 'polystab' function to reflect those filter polynomials with greater than unity magnitude back inside the unit circle. To further smooth the spectral transitions from frame to frame, I also used a median filter to interpolate the mapped filter transfer functions across speech frames.

3.2 Residual Modification

There are two steps to residual modification. The first is to classify the voice type of each speech frames as either voiced or unvoiced. Unvoiced frames are assumed to contain aperiodic noisy residual and will be modeled with white gaussian noise during synthesis. For voiced frames, the pitch periods of the excitation signal will be estimated using two different methods as described later in this section.

3.2.1.1 Voiced/Unvoiced Classification

A simple classification algorithm for voiced/unvoiced decision was given in [4] and is briefly described here. The energy of the prediction error (residual) and the first reflection coefficient are used to classify a speech frame as voiced or unvoiced. The first reflection coefficients is

$$r_1 = \frac{R_{ss}(1)}{R_{ss}(0)}$$

and

$$R_{ss}(0) = \frac{1}{h}\sum_{n=1}^{h} s(n)s(n),$$

$$R_{ss}(1) = \frac{1}{h}\sum_{n=1}^{h} s(n)s(n+1)$$

where h is the number of samples in the analysis frame and s(n) is the speech sample. The decision rules are as follows.

1. If the first reflection coefficient is greater than 0.2 and the residual energy is greater than a set threshold, then the current frame is classified as voiced.
2. If the first reflection coefficient is greater than 0.3 and the residual energy is greater than the set threshold used in rule 1 and the previous frame is also voiced, then the current frame is classified as voiced.

3. If the above conditions are not valid, then the current frame is classified as unvoiced.

The above algorithm generates a sequence of 1s and 0s. Patterns of 101 and 010 seldom occur in real speech and are corrected to strings of 111 and 000, respectively, to reduce the classification error rate.

3.2.1.2 Pitch Estimation

Because of the non-stationary nature of speech, irregularities in vocal cord vibration, interaction of the vocal tract and the glottal excitation, a perfect evaluation of the pitch periods is not always possible. However, many algorithms exist, some of them are performed in the frequency domain by measuring harmonic spacing, others are directly performed in the time domain. In this study, two different methods were experimented: autocorrelation and cepstral-deconvolution. The autocorrelation method of pitch detection is as implemented in the 'lpcBHenc' Matlab function and will not be described here. The cepstral-deconvolution method was described in [4] [5] and the steps are as outlined below:

1. Low-pass filter each frame of the prediction error (residual) waveform, *rsd(i)*. The filtered waveform is denoted as $rsd_{LP}(i)$.

2. Calculate the cepstrum-like sequence, $C_{rsd}(i)$.

   $$C_{rsd}(i) = IFFT(|\,FFT(rsd_{LP}(i))\,|) \qquad 1 \,\square\, i \,\square\, h\,,$$

   where *h* is the frame size, *FFT* is the fast Fourier transform and *IFFT* is the inverse operation.

3. Search for the index *m*, where $C_{rsd}(m)$ is the maximum amplitude in the subset $\{C_{rsd}(j) \mid 25 <= j <= h\}$.

4. Search for the index *k*, where $C_{rsd}(k)$ is the maximum amplitude in the subset $\{C_{rsd}(j) \mid 25 <= j = m\text{-}25\}$.

5. If $C_{rsd}(k) > 0.7\ C_{rsd}(m)$, *k* is the estimated pitch period, otherwise *m* is the estimated pitch period.

6. Low-pass filter (median filter) to smooth abrupt changes in pitch periods of successive frames.

The cepstral-deconvolution method seems to offer slightly better accuracy and was chosen for implementation in this project.

3.2.1.3 Pitch Mapping

Once the pitch periods for each frame of time-aligned source and target speech signals have been determined, linear regression estimation is again employed to obtain the pitch mapping parameters b0 and b1. From experiments, pitch mapping using this method produces only satisfactory results for male-to-female voice transformation. For female-to-male voice transformation, the average mapped pitch periods often remain low resulting in a transformed voice that still sounds like it was from a female speaker. A second mapping method that is based on simple scaling by the ratio of average source and target pitch values yields somewhat improved results but the transformed speech still exhibits the voice qualities of a female speaker. I think this is more of a result of less than optimal mapping of the transfer functions where the formant frequencies remain higher than they need to be. Shown in Figure 2 is an example of the transfer function frequency responses of a male (original) and female (transformed from original) speech frames using the spectral mapping method described in section 3.1. Shown in Figure 3 is an example of the transfer function frequency responses of a female (original) and male (transformed from original) speech frames. As can be seen in Figure 2, the formant frequencies of the transformed speech frame are higher than the original male speech frame as intended for a transformed female utterance. However, as seen in Figure 3, the formant frequencies of the transformed male speech frame are also higher than the original female speech frame. Not all of the female-to-male transformed speech frames exhibit this spectral mapping problem but it is fairly common. In addition, I also noticed that the first formant of the transformed (to male) speech is almost always higher than the original's first formant. Not knowing immediately how to correct this spectral mapping problem and due to time constraint, I opted to compensate by adjusting the pitch at the expense of synthesis voice quality and naturalness. I impose in my pitch-mapping algorithm that the average pitch periods of a female-to-male transformed excitation must be greater than 75 Hz. If this condition is not met, the pitch-scaling factor is readjusted so the above criterion is met. This crude method proves to be somewhat effective in that the transformed voice now possess a more hoarse quality consistent with that of a typical male speaker's voice. But the results are still not very satisfying. After giving this problem some more thoughts, I decided to modify the LPC mapping procedure for female-to-male speech transformation. I reduced the order of linear regression function to 0, i.e. b1 is now set to 1, and introduced additional bias to the b0 fitting parameters. The bias value was empirically determined to be –0.01. This fix seems to work very well, as the new formant frequencies of the female-to-male transformed speech signal are now consistently lower than the source speech signal.

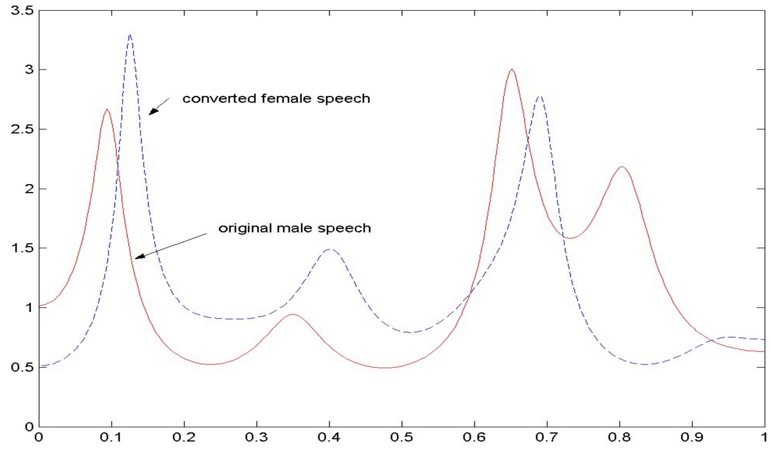


Figure 2. Transfer Function Frequency Responses of a Male and Corresponding Transformed Female Speech Frames

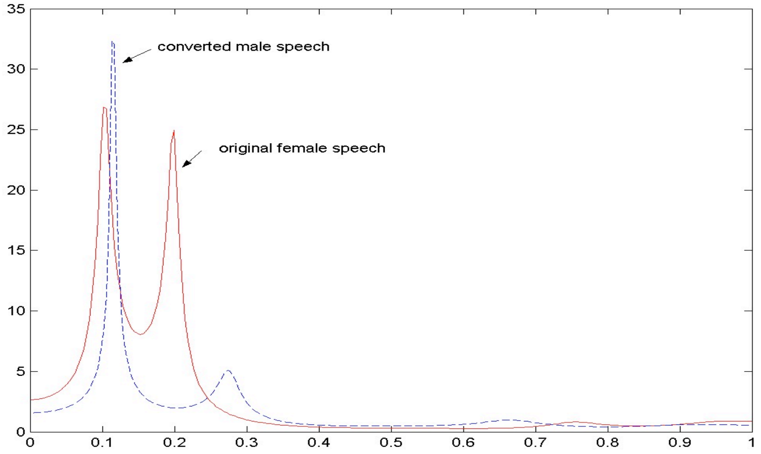


Figure 3. Transfer Function Frequency Responses of a Female and Corresponding Transformed Male Speech Frames

3.2.2 LP-PSOLA

In this project, I have also experimented with a second method of residual modification known as LP-PSOLA. In this technique, a time domain PSOLA [5] [6] [7] process is applied to the residual waveforms to modify the pitch-scale and time-scale of the residual signal. The modified residual waveform is then input to the vocal tract LPC filter to synthesize the new voice. Due to time constraints, I did not implement the prescribed PSOLA algorithm the way it was intended to be. The procedure I used to implement my particular "pseudo-PSOLA" algorithm is as follows.

1. The first and last frame of the speech signal is assumed unvoiced. Unvoiced frames will not be processed. White noise will be used for unvoiced frames during synthesis.

2. For each voiced frame, the instant of the main pitch pulse (with the largest amplitude) is determined.

3. A Hanning window of length that is twice the new pitch period for the current frame (computed from the pitch mapping stage) is centered around the pitch pulse located in step 2.

4. Segments of the windowed pitch waveform from step 3 are then repeated and overlap-added to produce the new modified residual waveform for the current frame. Care is taken to ensure pitch waveform continuity is maintained across successive frames.

Figures 4 and 5 illustrate actual examples of how the pitch of an excitation waveform was decreased and increased respectively using the implemented PSOLA algorithm.

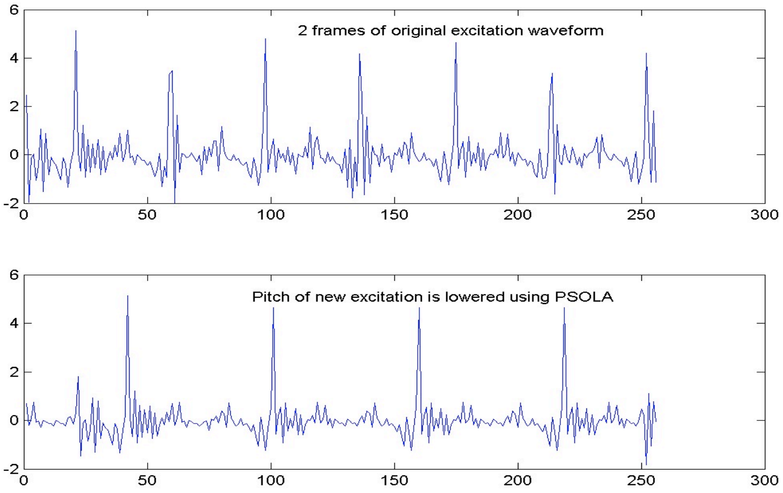


Figure 4. Pitch of an excitation signal is decreased using PSOLA

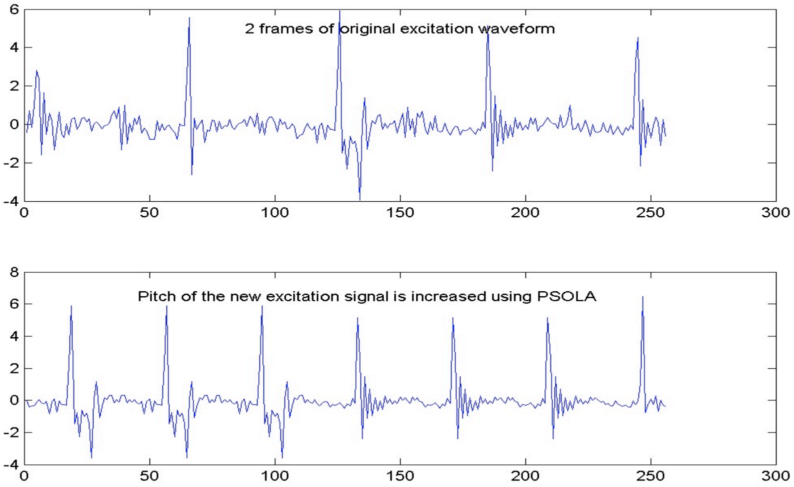


Figure5. Pitch of an excitation is increased using PSOLA

4. Experiment Results

The speech signals used for both training and testing are drawn from the TIMIT Speech Corpus made available by Columbia University for this speech-processing project. The TIMIT speech database consists of sentences uttered by 630 male and female speakers from many geographical regions of the United States. Of the many speech samples contained in this database, two of the sentences are identical and were spoken by all speakers. These two particular sentences were used for the training and testing phases of this project.

Informal listening tests were conducted to assess the effectiveness of the voice transformation algorithms. In this test, five pairs of speakers were randomly chosen from the TIMIT database. The two sentences “She had your dark suit in greasy wash water all year." and "Don't ask me to carry an oily rag like that." from each pair of speakers were selected for this test. In each case, the same sentence from both source and target speakers were used to train the system. After training, the second sentence (test sentence) from the source speaker is input to the system for conversion.

As mentioned previously, I experimented with voice transformation where only the vocal tract transfer functions were mapped. After listening to few transformed examples using this method, it was very apparent that the transformed voices remain very much like the original speakers' voices. I then decided not to pursue this further. I also experimented with mapping only pitch contours alone, this yielded mixed and very inconsistent results, I also decided to not to pursue this further due to lack of time. The PSOLA algorithm I implemented did not give robust and consistent results either. In isolated cases, the transformed voice did sound more natural. But in some other cases, I could hear both the original and the transformed voices at the same time. Admittedly, I did not spend a great deal of time implementing and debugging this algorithm. I believe that if I can spend more time to implement a true pitch-synchronous algorithm rather than the hybrid frame-based pitch-synchronous algorithm I came up with, and to work on improving the phase continuity of the overlap-add segments, this technique can yield promising results. For the above reasons, the informal listening tests were restricted to speech signals that were transformed through mapping of LPC filters and pitch values using the linear regression least square estimation method.

Five test subjects recruited from friends and family were asked to subjectively judge the gender of the speaker of the converted speech. In all cases, the subjects were not told the gender of the original speakers or allowed to listen to the original speech. For 4 out of 5 pairs of converted speech, all of the 5 test subjects judged the male-to-female converted speech as spoken by female speakers and the female-to-male converted speech as spoken by male speakers. For the remaining converted speech, the results were mixed with 3 out 5 subjects in one case and 4 out of 5 subjects in the other judged the converted speech to be spoken by speakers of the target gender. This gives an overall success rate of 94%. The test subjects were then asked to score the intelligibility of the converted sentences. The average score is 94.3% (with 100% being the best) indicating that no significant distortion was introduced during the transformation and synthesis process to degrade the intelligibility of the speech sounds. The test subjects were finally asked to compare the quality of the converted speech with the original. On a scale of 1 to 5 with 5 being the best, i.e. the quality of the original speech, the subjects scored an average of 2.39. When polled, the subjects complained about the occasional clicks and pops that are audible in the converted speech. They also cited the buzziness quality of the converted sounds. This is not all that surprising given the rather low-order voice conversion system that was implemented and the fact that LPC filter is an all-pole filter which doesn't model the zeros of the vocal tract response associated with nasal sounds. This also points to the need for more robust pitch estimation and spectral mapping methods. The results of the informal listening tests are summarized in Table 1.

Speakers		fajw0/mdb0	fedc0/mdmt0	ftmg0/mdac0	fpjf0/mdpk0	fdaw0/medr0
Gender	m to f	5/5	5/5	5/5	4/5	5/5
Perpception	f to m	5/5	3/5	5/5	5/5	5/5
Intelligi-	m to f	98	75	94	88	100
bility	f to m	98	92	100	100	98
Voice	m to f	3.1	2.2	2.1	2.1	2.6
Quality	f to m	3.0	2.7	2.3	3.1	2.8

Table 1. Informal Listening Test Results

5. Conclusion

In this project, I have investigated various voice transformation methods. I implemented a simple LPC-based gender voice transformation system that maps the residual and vocal tract transfer functions of a source speaker to those of a target speaker. Using only one pair of training sentences, the implemented system is capable of transforming a new sentence from the source speaker to give the perception that a speaker of the opposite gender had uttered it. Subjective listening tests indicate that the converted speech were highly intelligible but the synthesized speech quality were just below average due to the buzziness quality of the speech sounds and the occasional clicks, pops and squeals that were introduced in the transformation/synthesis process.

6. Further Investigations

The speech analysis and synthesis methods I implemented were based on fixed 128-saimple frames of speech data. I might be able to obtain better results if I do the analysis and synthesis pitch-synchronously. This should allow for the construction of a smoother pitch contour and better-fit LPC filters.

I would also like to spend more time on the PSOLA algorithm to improve the detection of pitch epochs and merging of the overlap-add segments to eliminate phase discontinuities. Another technique I would like to investigate is the modeling of excitation glottal pulses using polynomial model or LF model as suggested in [4], [8]. I expect to see improved naturalness of synthesized speech using this technique.

I would also like to investigate the viability of constructing a segment-based transformation system. Here, a speech recognition module such as HMMs would be used to segment training speech sequences into phoneme speech units. These speech units are then LPC analyzed and the corresponding LPC filter coefficients are stored to build a moderate-size inventory of filter samples. A new sentence for conversion will then be segmented into phonemes again and mapping of the corresponding LPC filters will then be based on best-matched spectral characteristics of the many samples stored in this inventory. In this scheme, linear regression mapping using Neural Nets would be more effective and meaningful and it should be interesting to see if this approach would produce better results than the scheme I adopted for this project. Of course, this technique would require a much larger training database and represents a more time-consuming undertaking.

References

[1] Moulines, E. and Laroche J., “Non-parametric techniques for pitch-scale and time-scale modification of speech”, Speech Comm. 16(1995) 175-205.
[2] George, E.B. and Smith, M.J.T., ‘Speech analysis/overlap-add sinusoidal model’, IEEE transaction on speech and audio proc. Vol. 5, No. 5, Sept (1997), 389-406.
[3] Laroche, J. Stylianou, Y. and Moulines, E. ‘HNM: A simple efficient harmonic+noise model for speech’ Proc. IEEE ICASSP-93, Minneapolis, Apr 1993.
[4] Childers,D.G., and Hu,T.H. (1994). Speech synthesis by glottal excited linear prediction. J. Acoust. Soc. Am.
[5] Gold, E., and Morgan, N., “Speech and audio signal processing”, John Wiley & Sons, Inc., 2000.
[6] Valbret,H., Moulines,E., and Taubach,J.P., “Voice transformation using PSOLA technique”, IEEE, 1992.
[7] Vergin,R., O’Shaughnessy,D., and Farhat,A., “Time domain technique for pitch modification and robust voice transformation” IEEE, 1997.
[8] Jiang,Y., and Murphy,P., “Voice source analysis for pitch-scale modification of speech signals”, University of Limerick, Limerick, Ireland.
[9] Mitra, S., “Digital Signal Processing, a Computer-based Approach”, McGraw-Hill, 2001.