Voice Transformations: From Speech Synthesis to Mammalian Vocalizations[1]

Min Tang, Chao Wang, Stephanie Seneff

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA
{mtang, wangc, seneff}@sls.lcs.mit.edu

Abstract

This paper describes a phase vocoder based technique for voice transformation. This method provides a flexible way to manipulate various aspects of the input signal, e.g., fundamental frequency of voicing, duration, energy, and formant positions, without explicit F_0 extraction. The modifications to the signal can be specific to any feature dimensions, and can vary dynamically over time.

There are many potential applications for this technique. In concatenative speech synthesis, the method can be applied to transform the speech corpus to different voice characteristics, or to smooth any pitch or formant discontinuities between concatenation boundaries. The method can also be used as a tool for language learning. We can modify the prosody of the student's own speech to match that from a native speaker, and use the result as guidance for improvements. The technique can also be used to convert other biological signals, such as killer whale vocalizations, to a signal that is more appropriate for human auditory perception. Our initial experiments show encouraging results for all of these applications.

1. Introduction

Voice transformation, as defined in this paper, is the process of transforming one or more features of an input signal to new target values. By features we mean fundamental frequency of voicing, duration, energy, and formant positions. Aspects that are not targeted to change should be maintained during the transformation process. The reconstructed signal should also be of high quality, without artifacts due to the signal processing. This notion of voice transformation is related to but distinct from research on voice conversion, where a source speech waveform is modified so that the resulting signal sounds as if it were spoken by a target speaker [**?**]. A conversion of this type is often done by mapping between detailed features of the two speakers. Our research focuses instead on simultaneous manipulation along the dimensions listed above, in order to achieve interesting transformations of speech and other vocal signals. We will use conversion and transformation interchangeably in the rest of this paper.

In the past, research in speech synthesis has led to several voice transformation methods, which are mainly based on TD-PSOLA [**?**] or sinusoidal models [**?**]. The method we propose here is closely related to the above methods, and can perform general transformation tasks, while preserving excellent speech quality.

[1] This research was supported by a contract from BellSouth and by DARPA under contract N660001-99-1-8904, monitored through Naval Command, Control, and Ocean Surveillance Center.

Our system is based on a phase vocoder, where the signal is broken down into a set of narrow band filter channels, and each channel is characterized in terms of its magnitude and phase [**?**, **?**]. The magnitude spectrum is first flattened to *remove* the effects of the vocal tract resonances, and then a transformed version of the magnitude spectrum can be re-applied, to produce apparent formant shifts. The phase spectrum, once it is unwrapped, represents the dominant frequency component of each channel. It can be multiplied by a factor α to produce an apparent change in the fundamental frequency of voicing. Additional harmonics may need to be generated; these can be created by duplicating and shifting the original phase spectrum.

In this fashion, we can perform a variety of interesting transformations on the input signal, from simply speeding it up to generating an output signal that sounds like a completely different person. We believe that it will be useful in many possible application areas, including speech synthesis, language learning, and even in the analysis of animal vocalizations.

Voice transformation has the potential to solve several problems associated with concatenative speech synthesis, where synthetic speech is constructed by concatenating speech units selected from a large corpus. The unit selection criteria are typically complex, attempting to account for the detailed phonological and prosodic contexts of the synthesis target. Such a corpus-based method often succeeds in producing high quality speech, but there are several problems associated with it. First, the cost of collecting and transcribing a corpus is high, making it impractical to collect data from mutiple speakers. Second, there usually exist discontinuities in both formants and F_0 values at concatenation boundaries. Third, high level prosodic constraints are often difficult to capture, and an incorrect prosodic contour leads to the perception of a substantially inferior quality in the synthetic speech.

To deal with the problem of generating multiple voices, a voice transformation system can be used to generate speech that sounds distinctly different from the voice of the input corpus. The entire corpus can be preprocessed through a transformation phase that can change a female voice into a male-like or child-like voice, while preserving the temporal aspects. In this way, a corpus from a single speaker can thus be leveraged to yield an apparent multiple-speaker synthesis system.

The transformation system can also be utilized to alter the shape of the F_0 contour of a concatenated sequence, in order to gain explicit control over the desired intonation phrases. By applying voice transformation techniques to post-hoc edit concatenation units, we can relax the selection criteria and consequently reduce the size of the corpus. The constraints for unit selection can then focus on manner/place/voicing conditions, without addressing the issue of the prosodic contour.

Another area where we envision utility of the voice transformation system is as an aid to teaching pronunciation, partic-

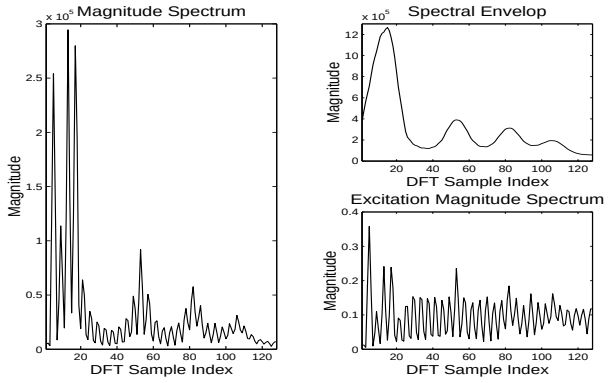

Figure 1: *Illustration of deconvolving the vocal-tract filter and excitation. The original spectrum is on the left and the spectral envelope and excitation spectrum are on the right.*

ularly for learners of a foreign language. The idea is to transform the student's own speech to repair inappropriately uttered prosodic aspects. This can be particularly effective for tone languages, where speakers who are unfamiliar with the concept of tone may not be able to perceive what is incorrect about their utterances. By transforming the student's speech to repair incorrectly uttered tones, the system allows the user to compare their own speech with the same speech tonally repaired. The student will hopefully be better able to distinguish which aspect is erroneous, because the speaker quality, temporal aspects, and phonetic content are held constant.

Finally, we have also used the voice transformation system to process other biological signals, in particular, killer whale vocalizations. These signals typically contain significant information at frequencies well beyond the human hearing range. The phase vocoder can be used to compress the frequency content while preserving the temporal aspects of the signal, bringing it into a range where human perception is acute.

The rest of the paper is organized as follows: in Section 2, we summarize briefly the methodology of our phase-vocoder based voice transformation system. In Section 3, implementation details are discussed. Finally, in Section 4, we present and discuss the experimental results.

2. Methodology

The source-filter model considers speech to be the convolution of an excitation source and a vocal-tract filter. In a phase vocoder, the input signal is first passed through a filter bank containing N contiguous band-pass filters, to obtain a magnitude spectrum and a phase spectrum. The spectral envelope of the magnitude spectrum, which characterizes the frequency response of the vocal-tract filter, can be obtained by low-pass filtering the magnitude spectrum. A time-domain deconvolution can be done by point-by-point dividing the magnitude spectrum by the spectral envelope samples. The result is the excitation magnitude spectrum. Figure **??** shows the effect of deconvolving the vocal-tract filter and the excitation magnitude spectrum.

The phase spectrum of each filter is unwrapped in time, in order to estimate the frequency of the energy contained in that filter. The first derivative of the unwrapped phases equals to the frequency f of the sinusoid component passing through the corresponding band-pass filter. Multiplying the first derivative by a factor α and then reconstructing the phase spectrum will

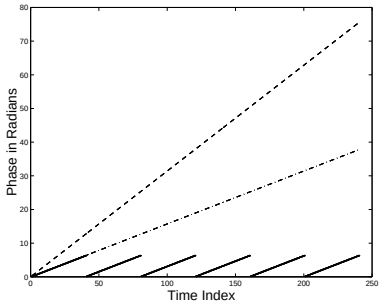

Figure 2: *Schematic of phase (in radians), unwrapped phase and doubled unwrapped phase, for a single channel [**?**].*

change the frequency of the particular sinusoid component to a frequency of αf. If the frequencies of all sinusoid components are changed by the same factor α, the effect is to change the pitch by a factor α. This process is illustrated in Figure **??**. Feature transformation can then be realized by modifying the spectral envelope and/or the phase spectrum. After modification, signals are reconstructed for each band-pass filter. Finally, the time-domain signal is reconstructed by combining the signals from each filter.

3. Implementation Issues

In this section, we discuss techniques for basic transformation tasks. A composite of these basic tasks can achieve more complex transformations, such as male/female conversion.

3.1. Iterative DFT

In our system, an N-point DFT of the input signal is computed at the original sampling rate, which is equivalent to passing the signal through a bank of $\frac{N}{2}$ contiguous band-pass filters. For efficiency, the system uses an iterative method, as described in [**?**], to compute the DFT, where each new output is obtained through incremental adjustments of the preceding output.

3.2. Deconvolving the DFT Spectrum

The $\frac{N}{2}$ DFT coefficients are then converted from rectangular to polar coordinates. The magnitude spectral envelope is estimated by low-pass filtering the magnitude spectrum. The excitation magnitude spectrum is obtained by dividing the magnitude spectrum by the smooth spectrum. The phase spectrum is unwrapped and first-differenced to obtain the instantaneous frequency.

The $\frac{N}{2}$-point excitation spectrum (excitation magnitude spectrum and phase spectrum) ranges from 0 to π. As we will soon see, sometimes we need to generate additional samples of the excitation spectrum beyond π. Generating these phantom excitation samples can be done by a cloning method described in [**?**].

3.3. Formant Modifications

The apparent positions of the formant frequencies can be altered by resampling the smoothed magnitude spectrum, and reconvolving it with the previously flattened excitation spectrum. For example, if we interpolate the spectral envelope by a factor of 1.2 and discard the extra points at the upper end, the formants will be moved up by roughly 20 percent. To move formants down, we need to decimate the spectral envelope. When reconstructing the speech signal, we discard some points at

the upper end of the excitation spectrum and phase spectrum. Correspondingly, the reconstructed speech will lose some high-frequency energy.

3.4. Pitch Modification

Pitch modification can be done by modifying the first derivatives of the unwrapped phases. To preserve the formant positions, we have to interpolate the spectral samples to compensate for the shifted phase spectrum. Thus, if we want to keep the spectral envelope intact when we change the pitch by a factor of α, the spectral envelope needs to be interpolated by a factor of $\frac{1}{\alpha}$ to hold the formants in position.

Another issue which needs to be considered is the change in the frequency range that will be caused by pitch modification. If the Nyquist frequency of the input speech is 4000 Hz (which means a sampling rate of 8000 Hz) and the pitch is changed by a factor of α, the frequency range after the pitch modification would be 4000α Hz. When $\alpha > 1$, the original 8000 Hz sampling rate is insufficient, and there will be frequency aliasing. A simple solution is to discard the signal above 4000 Hz. However, if $\alpha < 1$, the reconstructed signal will lose energy in the high frequencies. To make up for the loss, phantom excitation samples can be generated prior to the pitch modification.

3.5. Temporal Modification/Frequency Compression

Temporal characteristics of the input signal can also be manipulated. To slow down the speech, extra samples need to be generated. The magnitude spectrum of the generated samples can be obtained by interpolating the magnitude spectrum of the input signal in the time dimension. The α-scaled phase-derivative spectrum is interpolated similarly, and then the phase spectrum is restored using the interpolated phase derivatives. When the signal is reconstructed, there will be extra samples. If the signal is then played at the original sampling rate, the effect is that the speech is slowed down. Similarly, by decimating the magnitude and phase-derivative spectrum before reconstruction, we can reduce the number of samples, and consequently speed up the signal. By manipulating the sampling rate, we can convert temporal change to frequency change.

4. Experimental Results

4.1. Male/Female Conversion

To test the idea of voice conversion, we have conducted two experiments: (1) to convert a synthesis corpus of female recordings into a male-like voice, and (2) to convert a male Dectalk voice into a female voice. In the first experiment, all of the conversion can be performed in advance, creating an entire corpus with exactly the same temporal characteristics as the original one, but with a substantial change in the speaker characteristics.

The original synthesis corpus [**?**] was converted by lowering the fundamental frequency by 30% and the spectrum (formants) by 25%. Thus, the spectral envelope was interpolated by a factor of $\frac{1}{1.25*0.7}$. An additional thirteen phantom excitations were generated. The conversion preserves temporal characteristics, and hence we can reuse all the aligned transcription data from the original corpus. Figure **??** shows that, after conversion, the pitch and formant positions of the speech are similar to those of a corresponding natural male utterance.

Qualitative analysis reveals that male to female conversion achieves better quality than female to male conversion. We suspect this may be due to the extra complexity of generating the extra phantom excitation samples.

4.2. Language Learning

One of the difficulties in learning a foreign language is in mastering the prosodic aspects of the new language. For example, a native Chinese speaker generally has difficulty with stress and timing when speaking English. On the other hand, it is very hard for a native English speaker to speak Mandarin Chinese with correct tones, especially in complete sentences.

Voice transformation can potentially be used to enhance the experience of learning a foreign language. We can convert the prosody of a student's utterance to better match that expected from a native speaker, and use the modified utterance as a target for improvements. We believe that such a feedback mechanism would be more valuable to the student than an example spoken by a native speaker, because he or she can better perceive the differences and imitate the target speech, without the distractions from less relevant factors such as voice characteristics. The conversion can also be controlled to change only a certain aspect of the speech, to make the distinctions more apparent and easier for the student to follow.

We tested this idea by modeling the F_0 contours of Mandarin Chinese phrases spoken by a native English speaker. Mandarin Chinese is a tonal language, in which the syllable F_0 contour is essential to the meaning of sounds. There are four lexical tones in the language, each defined by a canonical F_0 pattern: high-level, high-rising, low-dipping, and high-falling. Although it is not too difficult for a non-native speaker to learn the tones in isolated syllables, the task becomes much more difficult for continuous speech due to tone coarticulation effects. The tone contours are perturbed in particular ways to form a smooth and coherent sentence F_0 contour. The process is very natural to a native speaker, but very hard for non-native speakers to learn, except by mimicking and practicing.

To predict a high-quality tone contour for continuous Mandarin speech, we obtained a complete profile of tones in all contexts from a spontaneous Mandarin speech corpus recorded from native Chinese speakers as examples [**?**]. The contextual effects are captured by means of context-dependent tone model parameters, which are discrete Legendre coefficients of syllable F_0 contours [**?**]. The conversion is carried out as follows. For each non-native Mandarin utterance, we first obtain its phonetic alignment and context-dependent tone labels. Then a basic target F_0 contour is constructed by piecing together the corresponding "standard" tone templates, with any remaining voiced gaps filled by linear interpolation. The new contour is smoothed and scaled to match the average F_0 of the speaker. A conversion ratio is thus obtained for each voiced frame of the speech signal, which is then used to modify the F_0 of the original waveform. Unvoiced frames are kept unchanged. Figure **??** illustrates this process with an example sentence. We tested this method on a few utterances and obtained good results. The tones in the reconstructed speech are significantly improved, judged by native Chinese speakers; while the other qualities of the waveforms are largely preserved.

4.3. Mammalian Vocalization

The voice transformation system can also be used to process non-speech signals. Killer whales produce a wide variety of distinct vocalizations[**?**], and a significant portion of the content is at frequencies that are well beyond the human hearing range. We have access to a large number of high-quality recordings obtained by Patrick Miller in Puget Sound using a beamforming

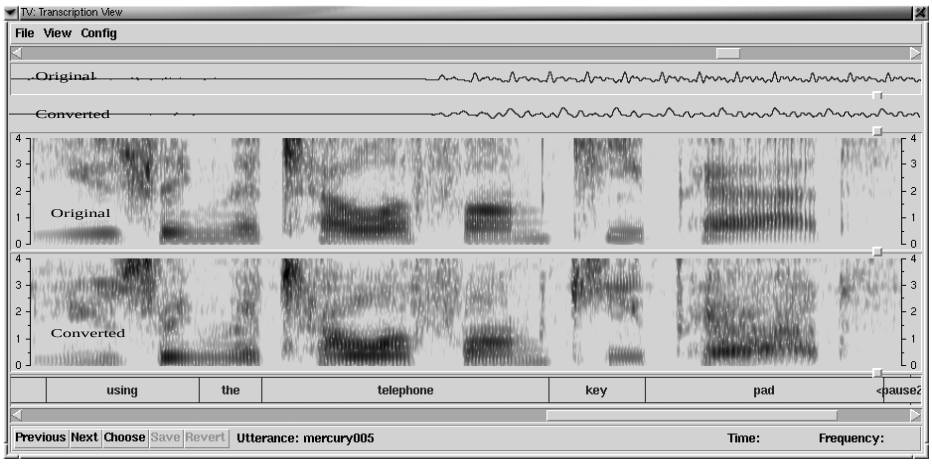

Figure 3: *Waveform and spectrogram of a female utterance before and after conversion. Notice that the formants in the converted spectrogram are shifted downward, and the higher frequency region is excited by phantom excitations.*

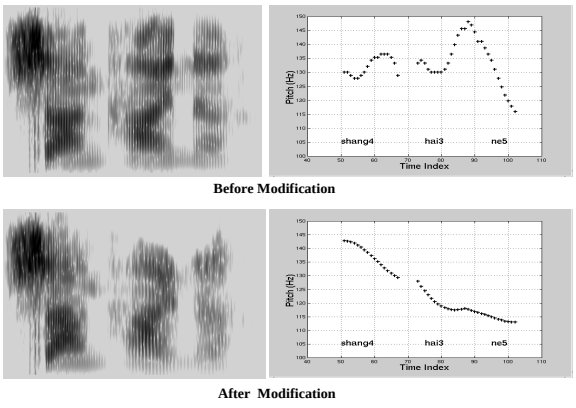

Figure 4: *Spectrogram and pitch contour (Hz) of the utterance "shang4 hai3 ne5?" (How about Shanghai?) before and after tone modification. The target pitch contour is generated using context-dependent tone model parameters. Some high frequency energy components are lost when the F_0 is lowered, which can be compensated by generating phantom excitations.*

array[**?**]. In a pilot study, we tried compressing frequencies by a factor of three, while preserving temporal aspects. The result is a signal that is much easier to study using standard tools for examining speech, and that also, perhaps more importantly, is more accessible to the human auditory system. A final advantage is that the storage requirements are reduced by a factor of three.

5. Summary

In this paper, we have proposed a voice transformation method based on the phase vocoder. This method demonstrates the ability to transform speech signals to various desired targets. The reconstructed speech has been found to be of high quality.

We have shown that a synthesis corpus can be transformed to generate speech of a different voice quality. In addition, we can adjust the intonation contour after concatenation to improve the prosodic quality. We have also demonstrated that the voice transformation system could help in language learning, particularly with regard to teaching non-native speakers of a tone language to utter the tones correctly. Finally, we have found the system to be useful for transforming killer whale vocalizations down into the human auditory frequency range.

6. Acknowledgements

Patrick Miller of the Woods Hole Oceanographic Institute provided us with example recordings from killer whales living in the Puget Sound area.

7. References

[1] Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Trans. Speech and Audio Proc., 6(2):131-142, 1998.

[2] E. Moulines, F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, V.9, pp. 453-467, 1990.

[3] M. P. Pollard, B. M. G. Cheeetham, C. C. Goodyear, M. D. Edgington, A. Lowry, "Enhanced shape-invariant pitch and time-scale modification for concatenative speech synthesis", ICCASP97, V.2, pp. 919-922, Munich, Germany, 1997.

[4] S. Seneff, "Speech Transformation System (Spectrum and/or Excitation) without Pitch Extraction", Technical Report 541, Lincoln Laboratory, MIT.

[5] J. L. Flanagan, R. M. Golden "Phase Vocoder", Bell System Tech. J., V.45 pp. 1493-1500, 1966.

[6] J. R. W. Yi, J. R. Glass, "Natural-sounding speech synthesis using variable-length units", ICSLP98, pp. 1167-1170, Sydney, Australia, 1998.

[7] C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, V. Zue, " Yinhe: a Mandarin Chinese Version of the Galaxy system", EUROSPEECH97, pp. 351-354, Rhodes, Greece, 1997.

[8] C. Wang, S. Seneff, "Improved tone recognition by normalizing for coarticulation and intonation effects", ICSLP00, V.2, pp. 83-86, Beijing, China, 2000.

[9] J. K. B. Ford, "Acoustic behavior of Resident Killer Whales (Orcinus orca) off Vancouver Island, British Columbia," Canadian Journal of Zoology 67, pp. 727-745, 1989.

[10] P. J. Miller, and P. L. Tyack, “A Small Towed Beamforming Array to Identify Vocalizing Resident Killer Whales (Orcinus orca) concurrent with focal behavioral observations,” Depp-Sea Research II, V. 45, 1998.