COMPARISON OF THREE TECHNIQUES FOR VOICE TRANSFORMATION

Ingo R. Titze, Darrell Wong, Robert Lange, Brad Story

Abstract

This study investigates the importance of the physiological domain in voice transformation. Voice transformation is defined as the process of modifying the voice quality of sentence-level speech while maintaining the same phonetic content. Transformation occurs as a function of gender, age, emotional state, disordered state, or impersonation. The basic question is: relative to pure signal processing, can voices be transformed more effectively if biomechanical, acoustic, and anatomical scaling principles are applied? The work reported here is an extension of Childer's work (1989) into the physiologic domain.

Introduction

The task of voice transformation consists of three phases: 1) an analysis of the speech into a set of parameters that allow manipulation of voice quality, 2) a transformation of the parameters into a set that describe a different voice quality hopefully based on anatomical and physiological variations, and 3) a reconstruction of the speech utilizing the new parameters with the original articulatory (phonetic) content. The first phase of the process requires that the speech be separated into its articulatory component (that which determines what is being said) and its voicing component (that which determines how and by whom it is being said).

But this is not a simple source-filter separation. Much about voice quality is determined by lower vocal tract filtering (the pharynx, the epilarynx tube, the piriform sinuses, and the velopharyngeal port). Qualities such as twang, ring, sob (Colton and Estill, 1978) are based on more than vocal fold adjustments, but they do augment basic laryngeal changes defined by vocal registers and modes of vibration of vocal fold tissues.

Methods

The present study proposes 3 trial schemes for analyzing speech and compares them for the purposes of voice transformation (Figure 1). Method 1 uses a signal-domain representation of the speech only. Frame-by-frame linear predictive analysis (LPC) is performed on the microphone signal. The designation of articulatory states is relegated to the LPC coefficients. These coefficients are mapped to a pseudo-area function sequence. As for the voicing characteristic, an estimate of the glottal flow wave is derived from the LPC residual (Childers and Hu, 1994; Milenkovic, 1993). The voicing parameterization includes an identification of the voicing

![Simplified flow charts of three trial methods for voice transformation.](image-url)
(on/off) state, an estimation of pitch period markers from the glottal flow/residual (F0), and a measurement of energy in each voiced cycle (intensity of phonation). No other information about the natural voicing characteristic is retained. A synthetic flow wave is then constructed cycle-by-cycle, with the systematic addition or deletion of cycles performed as necessary for the modification of pitch. In the case of the voiceless portions of the utterance (or any interval which lacks sufficient periodicity in excitation), the full LPC residual is retained. The composite (voiced/voiceless) excitation signal is then introduced frame-by-frame into the time-varying LPC filter pole manipulation and pseudo-area manipulation in the lower vocal tract region is used to simulate qualities such as twang, sob, and ring. The new filter is determined by the inverse mapping of pseudo-area function to LPC-coefficient set. The output of the LPC-filtered synthetic excitation yields the transformed speech result.

As an example of Method 1, a transformation to the twang quality was performed. As shown in Figure 2, a target rule was first established that reduces the area function to 25% of its original (non-twang) value in the epilarynx region (dashed lines). But this area function is not known for Method 1. Instead, a pole rotation was performed as shown in the bottom half of Figure 3. The rules for this pole rotation were developed by solving the forward problem first (area function to pole mapping) with known areas. Note that the spectrum of a twanged /a/ vowel in the top half of Figure 3 shows a raising of F1, F2, and F3, but a lowering of F4. The moving together of F2 and F4 raises the high frequency energy around 3000 Hz, which helps create the perception of twang. In Figure 4, a spectrogram-like

![Figure 2. A scaling factor applied to the vocal tract area function in the epilaryngeal region (left side, dashed lines) to produce the twang quality.](image1)

![Figure 3. Method 1 transformation to the twang quality for a single vowel /a/, showing the original and transformed LPC spectrum (top) and the pole rotation (bottom) used to achieve this quality.](image2)

![Figure 4. Method 1 transformation to the twang quality for the entire sentence "The blue spot is on the key again". The lines shown are formant trajectories in time, with the solid lines being the original and the dotted-dashed lines being the transformed.](image3)
Figure 5. Interpolated area function sequence used to create the phrase: "the blue spot".

Figure 6. Speech waveforms for "the blue spot": a) simulation based on the area function sequence in Figure 5; b) natural recorded speech.

Formant tracking is shown for the full sentence "The blue spot is on the key again", with solid lines being the original and dot-dashed lines being the transformed twang speech.

The second trial method (middle of Figure 1) utilizes a mixture of signal processing and biomechanical simulation methods. A frame-by-frame LPC analysis is again used to map the pseudo-area functions. The parameterization of voicing in this method, however, is augmented by the use of the electroglottograph (EGG). The EGG signal augments the LPC residual signal as a reference for estimating voicing onset/offset and instantaneous pitch. For the specification of the unvoiced intervals, an estimate of the average flow through the glottis is derived in conjunction with the LPC residual signal. These voicing parameters, along with the pseudo-area functions, are then fed to a pressure-flow simulator with a dynamically varying vocal tract. This simulator consists of a 22-section wave-reflection cylindrical-tube model which incorporates yielding walls, frequency-dependent viscous losses, radiation from skin surfaces, and multiple branches in the vocal and nasal sections. The vocal tract is excited by the same periodic flow synthesizer as used in Method 1. It also incorporates an "in-tract" noise generator for the non-periodic excitations arising from consonantal constrictions. At this stage, no examples of this method are available.

The final method of voice transformation (bottom of Figure 1) uses measured vocal tract areas rather than pseudo-areas derived from LPC. The characteristics of voicing for the input speech are analyzed frame-by-frame in the same manner as Method 2, using a combination of the microphone signal and EGG. Three-dimensional MRI images of the vocal tract, measured specifically for the speaker, are used to construct a list of the target phonemes for the utterance. These 3-D images are converted to area functions and then time-aligned (manually) with the microphone/EGG signals obtained from the actual test sentence. (A third transducer, the electroarticulograph EAG, is also used to facilitate this manual alignment; Karlsson and Nord, 1970.) After time-alignment of the phoneme target area functions, a series of interpolations between targets is generated as a way of estimating the dynamic movement of articulators through the allophonic and coarticulated sounds. The net area function sequence consists of a regular sampling of vocal tract movement at a rate of 40 Hz or higher.

At this point, Method 3 is being used to recreate natural speech without transformation. When that has been satisfactorily accomplished, the transformation of voice qualities can proceed. The sequence of area functions shown in Figure 5, was used to simulate the phrase: "the blue spot". The x-axis represents the elapsed time from the beginning of the phrase, the y-axis is the distance from the lips, and the z-axis is the
cross-sectional area. Combining the area function sequence with an interpolation of voicing parameters (which drive a glottal flow pulse model) produced the simulated phrase shown in Figure 6a. For comparison, the natural (recorded) speech waveform for the same phrase is given in Figure 6b. The timing of the articulation appears to be reasonably well represented, based on the onsets and offsets of successive segments. However, the amplitude envelopes of each phrase segment are somewhat different for the two versions (natural versus simulated). This deviation from the natural speech could be partially remedied by manipulating the amplitude of the glottal flow pulse, but the fundamental frequency and harmonic content of the flow pulse as well as the formant structure of a given vowel also influence the output pressure. With the articulation in place, various voice characteristics can be applied to the flow pulse model to generate a voice transformation. For example, in the original recorded phrase, the fundamental frequency ranged from 100 Hz to 140 Hz. This range could now be lowered or raised by some percentage and injected into the same area function sequence as before (see Figure 5) effectively creating the same phrase but with a new voice.

Conclusions

The physiological domain transformation defined in Method 3 is presently in the initial stages of development and implementation. It provides a means for coordinating a combination of different vocal effects at various levels of production, and it appears to be a natural choice for simulating a variety of qualities exhibited in pathology and vocal performance. At present, the matching of MRI image targets to the acoustic signal sequence is performed manually. In the future, an (automated) neural network mapping from acoustic signal frames to vocal tract area functions will be used. A multilayer perceptron (Rahim, 1994) will be used to provide a mapping to continuous area function estimates for voiced intervals. For unvoiced sounds, a learning vector quantizer (Kangas, Torkkola and Kokkonen, 1992) can be used to classify consonant speech to normalized consonant-area estimates.

Acknowledgment

This work is supported by research grant number 5 R01 DC 02532-01 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

References


