VOICE MORPHING
K.Sivaraman,
Assistant Professor, Dept. of Computer Science
BIST, Bharath University, Chennai -600073

ABSTRACT

Voice morphing means the transition of one speech signal into another. Like image morphing, speech morphing aims to preserve the shared characteristics of the starting and final signals, while generating a smooth transition between them. Speech morphing is analogous to image morphing. In image morphing the in-between images all show one face smoothly changing its shape and texture until it turns into the target face. The major properties of concern as far as a speech signal is concerned are its pitch and envelope information. These two reside in a convolved form in a speech signal. Hence some efficient method for extracting each of these is necessary. We have adopted an uncomplicated approach namely cepstral analysis to do the same. Pitch and formant information in each signal is extracted using the cepstral approach.

INTRODUCTION

Voice morphing, which is also referred to as voice transformation and voice conversion, is a technique for modifying a source speaker’s speech to sound as if it was spoken by some designated target speaker. There are many applications of voice morphing including customizing voices for text to speech (TTS) systems, transforming voice-overs in adverts and films to sound like that of a well-known celebrity, and enhancing the speech of impaired speakers such as laryngectomees. To achieve high quality of voice conversion, include a spectral refinement approach to compensate the spectral distortion, a phase prediction method for natural phase coupling and an unvoiced sounds transformation scheme. Each of these techniques is assessed individually and the overall performance of the complete solution evaluated using listening tests. Overall it is found that the enhancements significantly improve.

Overall Framework

Transform-based voice morphing technology converts the speaker identity by modifying the parameters of an acoustic representation of the speech signal. It
normally includes two parts, the training procedure and the transformation procedure. The training procedure operates on examples of speech from the source and the target speakers. The input speech examples are first analyzed to extract the spectral parameters that represent the speaker identity. Usually these parameters encode the short-term acoustic features, such as the spectrum shape and the formant structure. After the feature extraction, a conversion function is trained to capture the relationship between the source parameters and the corresponding target parameters. In the transformation procedure, the new spectral parameters are obtained by applying the trained conversion functions to the source parameters. Finally, the morphed speech is synthesized from the converted parameters.

**Spectral Parameters**

As indicated above, the overall shape of the spectral envelope provides an effective representation of the vocal tract characteristics of the speaker and the formant structure of voiced sounds. Generally, there are several ways to estimate the spectral envelope, such as using linear predictive coding (LPC), cepstral coefficients, and line spectral frequencies (LSF). The main steps in estimating the LSF envelope for each speech frame are as follows.

1. Use the amplitudes of the harmonics determined by the pitch synchronous sinusoidal model to represent the magnitude spectrum. K is determined by the fundamental frequency, its value can typically range from 50 to 200.
2. Resample the magnitude spectrum nonuniformly according to the bark scale frequency warping using cubic spline interpolation.
3. Compute the LPC coefficients by applying the Levinson-Durbin algorithm to the autocorrelation sequence of the warped power spectrum.
4. Convert the LPC coefficients to LSF.
5. In order to maintain adequate encoding of the formant structure, LSF spectral vectors with an order of p=15 were used throughout our voice conversion experiments.

**REALTIME VOICE MORPHING**

In real time voice morphing what we want is to be able to morph, in real-time user singing a melody with the voice of another singer. It results in an “impersonating” system with which the user can morph
his/her voice attributes, such as pitch, timbre, vibrato and articulation, with the ones from a prerecorded target singer. The user is able to control the degree of morphing, thus being able to choose the level of “impersonation” that he/she wants to accomplish. In our particular implementation we are using as the target voice a recording of the complete song to be morphed.

System block diagram

The Voice Morphing System

Figure shows the general block diagram of the voice impersonator system. The underlying analysis/synthesis technique is SMS to which many changes have been done to better adapt it to the singing voice and to the real-time constrains of the application. Also a recognition and alignment module was added for synchronizing the user’s voice with the target voice before the morphing is done. Once a user frame is matched with a target frame, we morph them interpolating data from both frames and we synthesize the output sound. Only voiced phonemes are morphed and the user has control over which and by how much each parameter is interpolated. The frames belonging to unvoiced phonemes are left untouched thus always having the user’s consonants.

MORPHING

Depending on the phoneme the user is singing, a unit from the target is selected. Each frame from the user is morphed with a different frame from the target, advancing sequentially in time. Then the user has the choice to interpolate the different parameters extracted at the analysis stage, such as amplitude, thus always using the consonants from the user. This will give the user the feeling of being in control. In most cases the durations of the user and target phonemes to be morphed will be different. If a given user’s phoneme is shorter than the one from the target the system will simply skip the remaining part of the target phoneme and go directly to the articulation portion. In the case when the user sings a longer phoneme than the one present in the target data the system enters in the loop mode.
REFERENCES

1. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations Hui Ye and Steve Young
2. High quality Voice Morphing Hui Ye and Steve Young
3. www.wikipedia.org