Talk-Along Karaoke

Jason Andersen
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
janderse@andrew.cmu.edu

Anish Menon
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
anishm@andrew.cmu.edu

Takshak Desai
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
takshakd@ece.cmu.edu

Swetha Chigurupati
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
schiguru@andrew.cmu.edu

Abstract

Karaoke systems usually prompt users to "sing" out the lyrics of selected songs. The idea of a talk-along karaoke system is to have the user "talk" out the lyrics of a song and output a singing version of the recorded speech. This essentially means that the speech waveform must be modified to resemble the corresponding song waveform. Speech modification for such an application requires the pitch and duration to be modified in order to match the pitch and duration of the song. Several methods can achieve such modification. One simple algorithm to perform this is Pitch Synchronous Overlap Add (PSOLA). Modifying the pitch of the speech signal to approximate the note frequency in a song produces a satisfactory first-order representation of "singing speech". Further effects can be added by incorporating techniques like vibrato while modifying the pitch. This project opens up several interesting possibilities in speech transformation and speech-to-song conversion.

1 Introduction

Human speech is periodic in nature. Periodic oscillations during the process of speech generation give the speech a fundamental frequency which is perceived as the "pitch" of the speech signal. Pitch is not an objective physical property, but a subjective psychophysical attribute of sound. Pitch allows the construction of melodies. Pitches are compared as "higher" and "lower". Changing the pitch of a human's speech can make the speech sound higher/lower pitched and at the same time give it some sort of melody depending on how the pitch is varied.

The speech to song transformation process begins by capturing a speech segment from the user using a script program that was developed by the team, depicted by Figure 1. The script reads the user's input for the selection of the song and the script begins to display the lyrics of the song (Figure 1b).

(a) (b)

Figure 1: Karaoke program user interface provides (a) song selection with (b) lyrics displayed to the user

The lyrics are displayed to the user followed by a prompt for the user to begin speaking the song lyrics. The target songs are:

- "Doe Ray Me" – no artist, uses the Simpsons theme
- "Hey Jude" – The Beatles
- "Wouldn't It Be Nice" – The Beach Boys

The Sound eXchange (SoX) sound processing command line utility is used to record the user input sampled at 16 kHz in 16 bit PCM encoding on one channel. Figure 2 presents a sample speech waveform speaking "doe ray me fa so la tea doe."

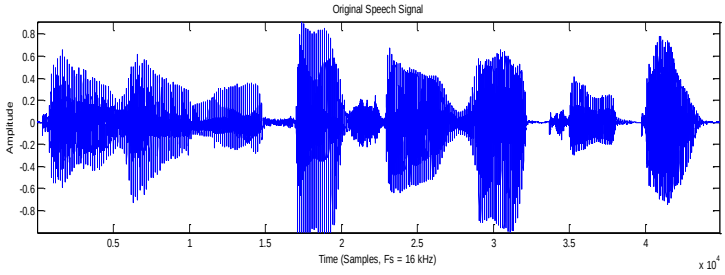

Figure 2: Speech waveform

1.1 Background

Speech contains voiced and unvoiced regions. Voiced regions are periodic with a frequency

$$F_0 = 1/\tau_0 \quad (1)$$

referred to as the fundamental frequency or pitch of the speech waveform. Unvoiced regions of speech are aperiodic and resemble a noisy source in the speech but are an artifact of human speech as depicted in Figure 3.

The periodic nature of voiced human speech and aperiodicity of unvoiced human speech can be observed in Figure 3 and Figure 4. The red lines denote periodic pitchmarks that depict times at which periodic peaks occur. As expected peaks in voiced speech coincide with the pitchmarks whereas for unvoiced speech the peaks do not match the pitchmarks.

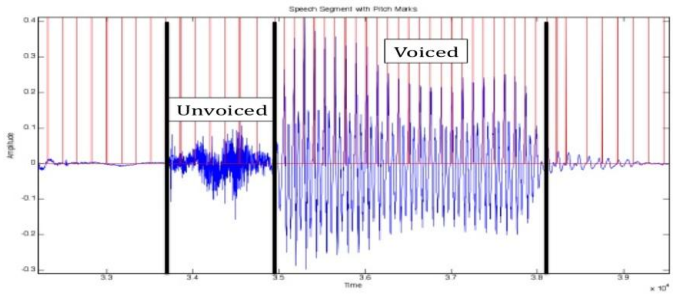

Figure 3: Unvoiced and Voiced Speech Segment

In order to transform the speech to song the pitch of the speech needs to be modified by adjusting the pitch and duration of the speech segments. The team has chosen the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) algorithm to perform the modification (TD-PSOLA and PSOLA are interchangeably throughout the paper and refer to the same algorithm). TD-PSOLA fails without proper pitch mark extraction. Pitch mark extraction extracts the fundamental frequency from the voiced regions of speech. The

Festival Speech Synthesis System, developed by University of Edinburgh and CMU, is used to extract pitch marks used by the karaoke system. Figure 4 shows a properly pitch marked segment of speech waveform that is used by the algorithm.

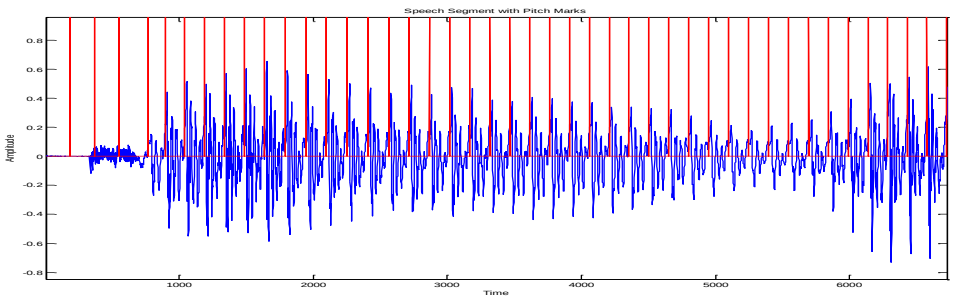

Figure 4: Pitch marked speech obtained from Festival

TD-PSOLA provides a method to adjust the pitch and duration; however, the speech then needs to be segmented so our system can appropriately map syllables of the lyrics to the target song. Syllables are composed of phonemes. Groups of phonemes are combined to form syllables and one or more syllables constitute a word. Figure 5 shows the phoneme labels provided by Festival. Additionally, Festival provides the system with the number of phonemes per syllable in the lyrics.

Usually in songs syllables get mapped to single notes. There are exceptions where more than one syllable gets mapped to one note or one syllable extends over two or more notes. However, for the selected songs, syllable-to-note mapping works. So, once the syllables and their constituent phonemes are identified, they can be mapped to separate notes. The pitch and durations of speech during these syllables are modified to mimic those of the notes being played in the music.

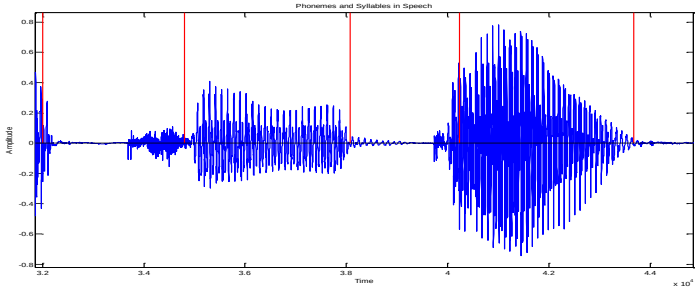

Figure 5: Phoneme labeled speech obtained from Festival Speech Synthesis software

The speech characteristics obtained above are all that is needed to transform the speech to song with one exception which is correctly mapping the source speech to the target song. The team has used *".mid"* formatted songs to extract the note number and note duration of the target song. Using MATLAB MIDI analysis scripts, developed by Ken Schutte, the note number and note durations were determined for the target songs.

The Talk-Along Karaoke system takes the following inputs: speech waveform in *".wav"* format, pitch marked speech, phoneme labels, phonemes per syllable, target note frequency and duration. The implementation of the system is detailed in the following sections.

2 Shell scripting

The team has used a bash shell script for our application to run fully automated with interaction from the user. There are two shell scripts 1) talkalong_windows.sh and 2) prompt_user_windows.sh.

2.1 Script talkalong_windows.sh

- Export the most frequently used paths
- Call the prompt_user_windows.sh to obtain the recorded song
- Copy the song to the data directory
- Build the data file needed by Festival to generate phoneme labels
- Generate the phoneme labels from Festival
- Generate the pitch marks from Festival
- Run the MATLAB file to perform PSOLA
- Mix the transformed speech with the target song (instrumental)
- Play the final song to the user

2.2 Script prompt_user_windows.sh

- Export the most frequently used paths
- Process the song selection from the user
- Generate the data file needed by Festival
- Generate phonemes per syllable from Festival using *text2syl* command
- Display the lyrics to the user
- Record the song using SoX
- Provide user with the option to rerecord if recording is not proper

3 Festival Speech Synthesis System

The Festival system includes 1) Festival Speech Synthesis, 2) Festvox, and 3) Speech Tools. The following are the commands that we used to generate the respective files using the festival software:

- *Festival –b PATH_TO_FESTVOX/build_1dom.scm '<build_prompts "etc/txt.done. data">'*
- *PATH_TO_LOCALDIR/bin/make_labs prompt-wav/song0001.wav*

These commands generate the user's phoneme labeled speech into the *song0001.lab* file.

- *PATH_TO_LOCALDIR/bin/make_pm_wave etc/txt.done.data*

This command generates the user's pitch marked speech into the *song0001.pm* file.

- *Echo "song lyrics" | ./ text2syl > filename.txt*

This command generates the number of phonemes per syllable from the song lyrics.

4 Pitch Synchronous Overlap and Add (PSOLA)

At the heart of the entire talk-along karaoke system is the algorithm that enables pitch modification i.e. PSOLA. It is the most widely used second generation signal-processing technique for pitch modification. PSOLA modifies the pitch and timing of speech but does so without performing any explicit source-filter separation. The main idea of the PSOLA techniques is to isolate individual pitch periods in the original speech, perform modification, and then re-synthesize to create the final waveform.

4.1 Time Domain PSOLA (TD-PSOLA)

TD-PSOLA is widely regarded as the most-popular PSOLA technique and the most-popular algorithm overall for pitch and timing adjustment. There is one analysis frame per pitch period. Thus it is necessary to identify epochs in the speech signal with high accuracy. The epoch position is the instant at which glottal closes at each period. The signal is separated

into frames using Hamming window, which is centered at the epoch and extends a little before and after the pitch period.

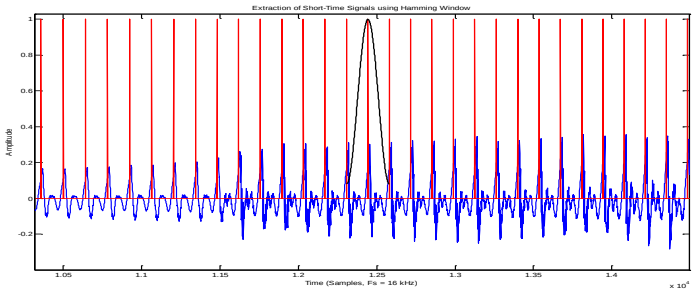

Figure 6: Hamming window centered on a pitch mark

These windowed frames can then be recombined by placing their centers back on the original epoch positions and adding the overlapping regions (hence the name, overlap and add). Time-scale modification is achieved by elimination or duplication of frames, as shown in figure below.

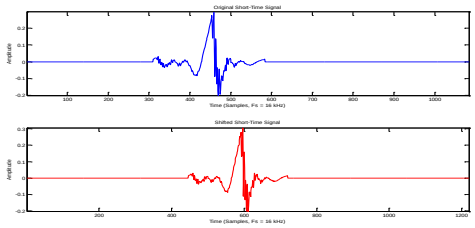

Figure 7: Windowed speech signals (window extends two pitch periods)

For a given set of frames, if we duplicate one of these frames and insert it back into the sequence and then perform overlap and add, then the desired effect of a longer stretch of natural speech is perceived. Listeners cannot detect that in the new signal two consecutive frames are identical, rather than slowly evolving, which is what we see in real speech. By eliminating a frame we can achieve the converse effect, and again listeners normally do not detect that a frame is missing. A rule of thumb is often quoted, namely that these processes can be used to lengthen or shorten a section of speech by a factor of about two without any or much noticeable degradation. The more modification one performs the more likely it is that the listener will notice.

Pitch-scale modification is performed by recombining the frames on epochs that are set at different distances apart from the analysis epochs. All other things being equal, if we take for example a section of speech with an average pitch of 100 Hz, the epochs will lie 10 ms apart. From these epochs we perform the analysis and separate the speech into the pitch-synchronous frames. We can now create a new set of epochs that are closer together, say 9 ms apart. If we now recombine the frames by overlap-add method, we find that we have created a signal that now has a pitch of 1.0/0.009 = 111.11 Hz.

Conversely, if we create a synthetic set of epochs that are further apart, and overlap and add the frames on those, we find that we generate non-synthetic waveform of lower pitch. This lowering process partly explains why we need frames that are twice the local pitch period; this is to ensure that, up to a factor of 0.5, when we move the frames apart we always have some speech to add at the frame edges.

4.2 Employing PSOLA in Talk-Along Karaoke

The basic idea of a talk-along karaoke system is to convert input speech to resemble a song. This essentially requires matching the pitch and duration of the speech waveform to those of the song waveform. This can be achieved by extending the PSOLA algorithm appropriately.

A recording of the target song with a single note is preferred as single notes provide better information for pitch and duration matching. For every note's duration, the target pitch frequency of the speech waveform should resemble the frequency of the note. This can be achieved by using PSOLA and placing the pitch-centered short-time signals at a frequency that is proportional to the note frequency. Overlap-add takes care of combining these short-time signals in a clean manner. The duration of the speech signal also needs to be simultaneously modified to match the duration of the song. As a first approximation, this can be done by mapping one syllable to one note. More complex mapping, say one syllable to several notes or several syllable to one note, can then be attempted and will vary with every target song.

Care needs to be taken when the pitch is being modified. Note frequencies are generally several orders higher than pitch frequencies of human speech. Using note frequencies directly can make the transformed speech sound instrument-like. The pitch frequency may be kept proportional to the note frequency but should be scaled down to match human speech. The fact that humans tend to sing in high-pitched voices can be used to advantage.

5 Future Work

The speech transformation can be improved by incorporating further pitch modification effects into the recorded speech and using Dynamic Time Warping for the matching process. Also the system can be improved to work with different styles of singers and songs.

5.1 Dynamic Time Warping

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This can be explained best with the help of the following example. Consider the DTW grid shown below.

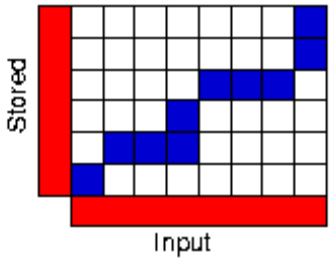

Figure 8: DTW Grid

The two sequences of observations can be arranged on the sides of a grid as shown in the above figure, with the unknown sequence on the bottom (six observations in the example) and the stored template up the left hand side (eight observations). Both sequences start on the bottom left of the grid. Inside each cell we can place a distance measure comparing the corresponding elements of the two sequences. The best match between these two sequences can be found by finding a path through the grid which minimizes the total distance between them. The path shown in blue in Figure 8 gives an example. Here, the first and second elements of each sequence match together while the third element of the input also matches best against the second element of the stored pattern. This corresponds to a section of the stored pattern being stretched in the input. Similarly, the fourth element of the input matches both the second and third elements of the stored sequence. Here a section of the stored sequence has been compressed in the input sequence. Once an overall best path has been found the total distance between the two sequences can be calculated for this stored template.

The procedure for computing this overall distance measure is to find all possible routes through the grid and for each one of these compute the overall distance which is the minimum of the sum

of the distances between individual elements on the path divided by the sum of the warping function. In this way, DTW can be used to create a mapping between the singer's voice and the user's voice which can then be used to modify the user's voice and make it sound like singing speech.

5.2 Digital Effects

The digital effects such as the Vibrato and Tremolo can be incorporated to give additional effects to the transformed speech.

Vibrato and Tremolo are separate musical effects. Vibrato is a musical effect obtained by periodic variation of a musical note. The two factors involved are amount of variation of pitch which is the extent of vibrato and the speed at which the pitch is varied which is the rate of vibrato. Vibrato adds warmth to the note. Tremolo is a musical effect obtained by the periodic variation in the volume i.e., the amplitude of a musical note.

5.3 Learning from Styles and Singers

Machine learning can be employed to learn from different styles of songs or different singers so as to more closely match the singing speech to the actual song. For example, Elvis' style can be learned from and then used to make a user sound more like Elvis. More can be learned from song styles and non-linear mapping of the speech waveform to the notes can be performed.

Acknowledgments
The Talk-Along Karaoke team would like to acknowledge Dr. Alan W. Black and Dr. Bhiksha Raj for their mentorship during the implementation of the system. The team utilized MIDI analysis scripts developed by Ken Schutte. The audio manipulation and recording was performed by the SoX command line utility provided for us by the SoX development community.

References
[1] Paul Taylor (2009) *Text-to-Speech Synthesis*. Cambridge University Press.
[2] Sami Lemmetty (1999) Review of Speech Synthesis Technology, Master's Thesis, Helsinki University of Technology.
[3] Werner Verhelst & Henk Brouckxon (2002) Voice Modification for Lip Synchronization, Voice Dubbing and Karaoke, *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium.
[4] Barbara Resch, Mattias Nilsson, Anders Ekman, & W. Bastiaan Kleijn (2007) Estimation of the Instantaneous Pitch of Speech, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3.
[5] Srikanth Mangayyagari & Ravi Sankar (2007) Pitch Conversion Based on Pitch Mark Mapping, *IEEE*, 1-4244-1029-0/07.