Voice Conversion Experiment and Conclusion*

Justin Chen

Abstract

The results of an experiment testing our voice conversion algorithm and possible ways to improve it.

1 Description of Experiment

To test our voice conversion algorithm, we administered a speaker identification test to twelve randomly selected people. Prior to the experiment, we recorded speech samples from four different speakers (two male and two female) and used our algorithm to convert between various combinations of their voices. For example, we took the sound of speaker #1 (the "source speaker") saying a certain phrase and converted it to the voice of speaker #2 (the "target speaker"). The participants listened to a series of these synthesized sounds, and we asked them to identify the speaker (the target) as well as the speaker's gender.

2 Results of Experiment

The target speaker was correctly identified **74%** of the time.

The target speaker's gender was correctly identified **93%** of the time.

*Version 1.4: Dec 21, 2004 5:13 pm -0600

†http://creativecommons.org/licenses/by/1.0

Graph of Gender-specific Conversion Accuracy

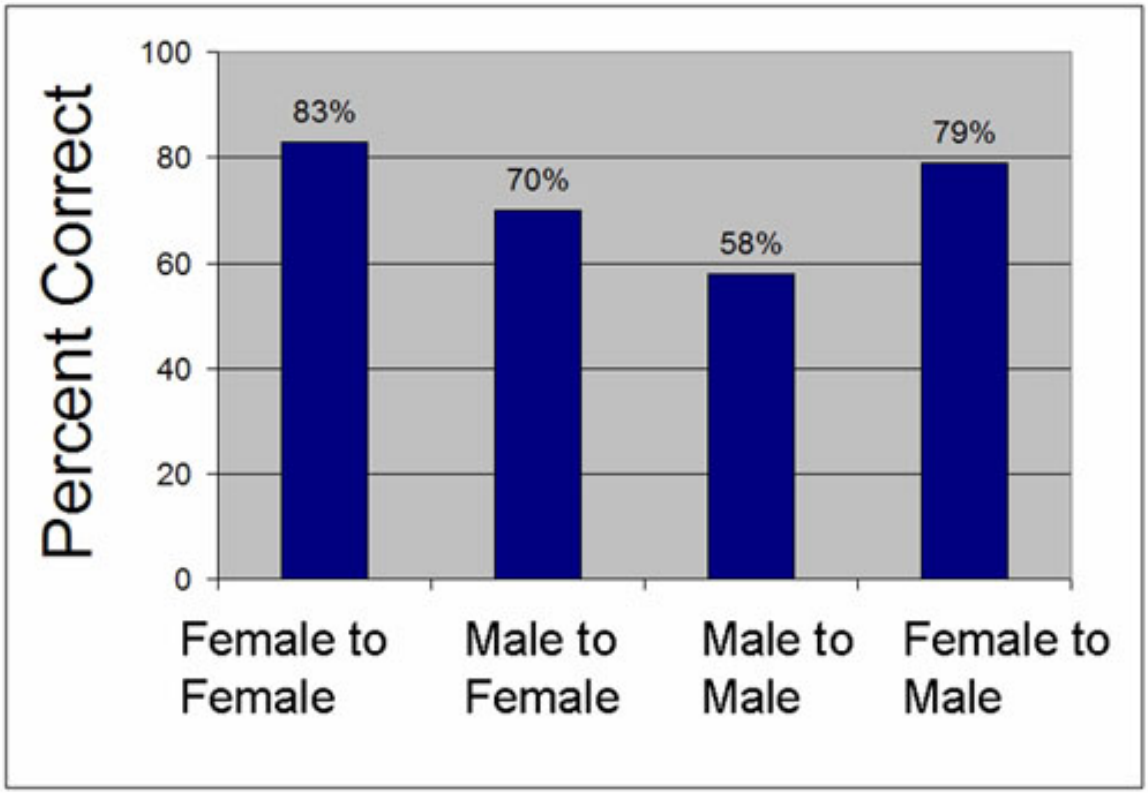

Figure 1: The first bar, "Female to Female," indicates a conversion from a female source speaker to a female target speaker was correctly identified 83% of the time.

3 Conclusions

Our voice conversion system was fairly effective at imitating a certain target speaker. From the "Gender-Specific Conversion Accuracy" graph, it can be implied that our system was better at converting female source speakers than male source speakers. One reason for this may be that the voices of the two male speakers used in the experiment had only a minor difference in pitch. The female speakers' voices, however, had a more noticeable difference.

4 Possible Improvements

At its current state, our system can only convert between two voices when it has samples of the speakers saying the same word or phrase. In order to make our system text-independent, we would need to implement **neural mapping**. This could be accomplished by using the **cepstrum** to identify certain characteristic sounds (such as vowel sounds) in the target speaker's speech sample and mapping their filters to the corresponding characteristic sounds in the source speaker's sample. In addition to adding text-independence to our system,

we could add a **band-pass filter** at the end of our system to help eradicate speech artifacts in our synthesized sounds. The filter would block out frequencies that are not in the range of human speech.