

Abstract

Voice transformation (also known as voice conversion or voice morphing) is a name given to techniques which take speech from one speaker as input and attempt to produce speech that sounds like it came from another speaker. One compelling argument for good voice transformation is that it reduces the difficulty in creating additional synthetic voices with new identities and styles once an existing voice has been created based on a full-sized corpus. There are further voice transformation applications for security, privacy, and assistive technologies.

Although current voice transformation techniques perform well in the sense that humans typically judge transformed speech to sound more like the target speaker than the source speaker, there is still room for improvement.

We investigate the use of articulatory position data to improve voice transformation. When a person speaks, motions of the articulators affect the shape of the vocal tract, which affects the produced sound. Recently, data that includes measurements of the positions of various articulators along with recordings of the produced speech has been made publicly available. This articulatory position data gives us new information about the production of speech and has already been used successfully to predict quantities such as Mel-frequency cepstral coefficients [Toda et al., 2004a]. Such data gives us a different source of information from typical features derived from speech signals and enables promising new approaches to voice transformation.

One of the current challenges of using articulatory position data is that it is difficult to collect, so little is available. In order for it to be useful for more than a few speakers, some strategy must be devised to estimate it for other speakers. We present a number of techniques to do this and demonstrate that they are plausible by showing that artificial estimates of articulatory positions can be used to improve phonetic feature predictions similar to actual articulatory positions. Then we proceed to the question of using articulatory position features for voice transformation. Modifying the voice transformation process and representation of the articulatory data enables us to show improvement according to an objective metric. Then we demonstrate that artificial articulatory position estimates can also be used to improve voice transformation for speakers for whom no articulatory position data has been collected, according to this same objective metric.

As we are attempting to improve voice transformation, we give further consideration to what this actually means. Although a number of objective and subjective tests have been used to judge voice transformation quality, the best way to evaluate it is still an open question. We present new subjective and objective measures for voice transformation and report the results and our observations.