

Student Research Symposium 2004

John Kominek

On the Road to High Quality Universal Speech Synthesis

Machine Translation has the Vaquois Triangle -- a famous high-level perspective that delineates the major approaches to MT, as well as their limitations. You can have either universality (through an Interlingua) or high quality (Direct translation), but not both. In between, trying to find a happy medium, reside Transfer techniques.

The field of Speech Synthesis also has such a triangle, with similarly frustrating trade-offs: either high quality or full flexibility, but not both. In this talk I begin by drawing the corresponding parallels, explaining where the three major approaches fit in, and their historical development. These three are unit-selection, spectrogram-based, and articulatory synthesis.

By directly employing segments of recorded speech, unit-selection synthesis can achieve excellent voice quality, but at the expense of flexibility. A universal synthesizer, ideally, can mimic any person in any language, in a full range of styles. Achieving this, though, demands precise modeling of the human vocal tract and articulators -- as yet an unsolved problem. In between, spectrogram-based synthesizers offer good controlability, but do not sound as natural as unit-selection techniques.

Two paths can thus be taken on the road to high quality universal synthesis. One can start with a flexible synthesizer and attempt to make it sound better. Or one can start with a good sounding synthesizer and try to make it more flexible. This talk will follow the second path.

To illustrate, we tackle the problem of "accent transformation" -- changing the accent of one person to sound more like that of another. This is made possible using CMU's recently created "Arctic Speech Databases," a parallel corpus of carefully spoken English sentences. Editions exist for American, Canadian, Scottish, Indian, and Japanese accented English. Grafting a new accent onto an existing voice is desirable for localizing a synthesizer to match that of a target region. Or, moving in the opposite direction, by making a native voice sound foreign, hence "exotic".