A Coarticulation Model for Articulatory Speech Synthesis

Anastasiia Tsukanova, University of Malta
Supervised by: *Yves Laprie*, Centre National de la Recherche Scientifique

MSc. Dissertation — 2016

Abstract

The state-of-the-art techniques for speech synthesis rely either on concatenation of acoustic units taken from a vast pre-recorded speech database noting the relevant linguistic information or on statistical generation of the necessary acoustic parameters and using a speech production model. These approaches yield synthesis of good quality, but are purely technical solutions which bring no or very little information about the acoustics of speech or about how the articulators (mandible, tongue, lips, velum...) are controlled.

In contrast, the articulatory approach generates the speech signal from the vocal tract shape and its modelled acoustic phenomena. The vocal tract deformation control comprises slow anticipation of the main constriction and fast and imperatively accurate aiming for consonants.

The system predicts the sequence of vocal tract consecutive configurations from a sequence of phonemes of the French language to be articulated and a model of the coarticulation effects in it. We use static magnetic resonance imaging (MRI) captures of the vocal tract shape when producing phonemes in various contexts, thus following an approach by Birkholz. The evaluation of the model is done both on the animated graphics representing the vocal tract shape evolution (how natural and efficient the movement is) and on the synthesised speech signals that are perceptively and—in terms of formants—qualitatively compared to identical utterances made by a human.

Our results show that there are a lot of effects in the dynamic process of speech that manage to be reproduced by manipulating solely static data. We discuss generation of pure vowels, vowel-to-vowel and vowel-consonant-vowel transitions, and articulators' behaviour in phrases, report which acoustic properties have been rendered correctly and what could be the reasons for the system to fail to produce the desired result in other cases, and ponder how to reduce the after-effects of target-oriented moves to obtain a more gesture-like motion.