Title:

Comparison of tongue contour extraction methods from ultrasound images for use in text-to-speech synthesis

Authors:

Tamás Gábor Csapó, csapot@tmit.bme.hu
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary

Steven M. Lulich, slulich@indiana.edu
Department of Speech and Hearing Sciences,
Indiana University, Bloomington, IN, USA

Abstract:

Text-to-speech (TTS) synthesis is a speech processing technique which can create human-like speech from any input text, typically with the help of a computer. TTS is used in screen reading applications for the vision impaired, car speech interfaces and automatic directory services. Although state-of-the-art TTS is highly intelligible, it is still far away from natural conversational speech and is often perceived to be robotic or buzzy. There is recent evidence that integrating information from the articulatory organs (e.g. tongue contour, lip motion, vocal tract shape) might improve the quality of text-to-speech synthesis, i.e. making it more human. For example, electromagnetic articulography has previously been used to integrate the movement of the speaking organs into a TTS system.

In this research, we use a 2D ultrasound for the tracking of the articulatory organs. Phonetic research has employed 2D ultrasound for a number of years, and there are therefore methods and algorithms which ease the recording and data processing of these data. The result of 2D ultrasound recordings is a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air.

In our experiments, we recorded parallel speech, video and 2D ultrasound signals with three speakers of American English and one speaker of Hungarian. The ultrasound transducer was held below the chin to record the movement of the tongue. In the recorded ultrasound images, the tongue contour was visible but the clearness of the data was found to be speaker dependent. We investigated manual tracings with seven tracers and calculated the differences in their measurements. We compared the manual tracing data with four automatic contour tracking methods which were able to extract the tongue contour from the 2D ultrasound images.

Future plans include the use of real-time 3D ultrasound for visualizing the whole tongue body, and we plan to integrate articulatory data based on ultrasound into American English and Hungarian TTS systems.