Using Two- and Three-Dimensional Fleshpoint Measurements of Articulatory Kinematics in Concatenative Speech Synthesis

Hartmut R. Pfitzinger

Department of Phonetics and Speech Communication, University of Munich, Germany

hpt@phonetik.uni-muenchen.de

1. Abstract

The proposed contribution focuses on the variability of articulatory kinematics at concatenative speech synthesis segment boundaries. Especially, we answer two questions: i) how do the discontinuities introduced by imperfect concatenation synthesis look like in the articulatory domain, and ii) is it possible to estimate both the perceptual and the acoustic distance between two successive speech units from the articulatory trajectories, their velocities, and their accelerations?

In 2002, Sondhi [2] pointed out that "no speech synthesis system has been designed so far" that is based on the concatenation of articulatory units derived directly from the speech signal (see e.g. Yehia 1997 [4]). Other methods to estimate articulatory information (e.g. x-ray, MRI or EMMA) were only incidentally mentioned because: "Such methods are probably suitable more for studying speech production than for implementing a TTS system." (Sondhi 2002 [2], Sec. 4.1)

We agree only partially upon this claim: the pure sound quality of today's speech synthesis systems has reached a high-fidelity level and therefore would degrade considerably if the underlying speech database had been recorded while applying one of the above-mentioned articulatory measurement methods. In particular, when using EMMA a pellet glued near to the tongue tip causes some speakers to having a lisp, which is unacceptable in a today's speech synthesis system.

Nevertheless, articulatory measurements are suitable for speech synthesis research. The experiments which we conducted until now demonstrate that even the slightest acoustical or perceptual discontinuities at concatenation boundaries become obvious in the articulatory domain.

In the first experiment a phonetically rich read-speech database was used which also provides articulatory data recorded with an EMMA system (Carstens, AG-100 at 250 Hz sampling rate, see e.g. Hoole & Nguyen 1999 [1]). EMMA was applied because the current frame-rate of real-time MRI is too small to capture the kinematics of natural speech, and x-ray has adverse health effects, especially when recording a large single-speaker database. A standard concatenative speech synthesis method was used to systematically create new utterances with more or less obvious acoustical and perceptual discontinuities. Then, the related discontinuities in the articulatory data were analyzed. One of the basic results was that while both the velocity and acceleration profiles of the two concatenated segments matched, the horizontal and vertical displacements of the tongue and the jaw appeared to mismatch systematically. Finally, the acoustic distance measure of the synthesis system was tuned to minimize the articulatory displacement offsets which reduced the perceptual discontinuities evaluated by means of listening tests.

The second experiment was based on the same data as the first experiment. We replaced the original spectral distance measure by an articulatory-based distance measure. It included the positions, velocities, and accelerations of both the horizontal and vertical movements of the recorded articulators. This distance measure was suited to further reduce the perceptual discontinuities except in the frequent cases of velar activity and lip rounding because no pellets were attached to the velum or to the corners of the mouth. The latter clearly indicates that the experiments have to be repeated with a three-dimensional EMA system (Zierdt, Hoole & Tillmann 1999 [5]) which is also well-suited to measure the lateral tongue activity and additionally provides information on sensor orientation. These recordings as well as the analysis will be finished in October 2003, and the results will be presented at the Speech Production Seminar.

This work is highly interdisciplinary but it predominantly involves both speech production research as well as speech synthesis research. The proposed contribution presents a detailed analysis of the source of articulatory variability at segment boundaries and its influence on acoustical and perceptual discontinuity which was caused by imperfect concatenation.

Our approach is fundamentally new because in speech production research the articulatory discontinuities introduced by concatenative speech synthesis systems have not been investigated until now. This proposal shows that articulatory research is able to considerably contribute to speech synthesis research.

2. References

[1] P. Hoole and N. Nguyen, "Electromagnetic articulography," in *Coarticulation. Theory, Data and Techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge University Press, 1999, ch. 12, pp. 260–269.

[2] M. M. Sondhi, "Articulatory modeling: a possible role in concatenative text-to-speech synthesis," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, 2002.

[3] H. G. Tillmann and H. R. Pfitzinger, "Parametric High Definition (PHD) speech synthesis-by-analysis: The development of a fundamentally new system creating connected speech by modifying lexically-represented language units," in *Proc. of ICSLP 2000*, vol. 3, Beijing, 2000, pp. 295–297.

[4] H. C. Yehia, "A study on the speech acoustic-to-articulatory mapping using morphological constraints," Ph.D. dissertation, Graduate School of Engineering of Nagoya University, 1997.

[5] A. Zierdt, P. Hoole, and H. G. Tillmann, "Development of a system for three-dimensional fleshpoint measurement of speech movements," in *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 1, San Francisco, 1999, pp. 73–75.