Generating Expression in Synthesized Speech

by

Janet E. Cahn

Abstract

This document is a revised version of my master's thesis, submitted in May, 1989 to the Media Arts and Sciences Section of the Department of Architecture, at the Massachusetts Institute of Technology. The revisions are as follows: grammatical and factual corrections, particularly in Chapter 2; revised table formats to better conform to IPA standards; the addition of a table in Appendix B; and the addition of Appendix C[1] containing pitch tracks of, energy tracks and spectrograms of synthesized speech. The grammatical and factual revisions were contributed by Susan E. Brennan of Hewlett–Packard Laboratories in Palo Alto in the summer of 1989, and by Professor Kenneth N. Stevens, of the Massachusetts Institute of Technology in the fall.

The document examines the proposal that affect can be reproduced in synthesized speech by imitating the effects of emotion in human speech. A program, the Affect Editor, was constructed to systematically vary the influence of the speech correlates of emotion and so direct the synthesis of the intended affect. The task raised and explored questions about: the appropriate representation of an emotion's effect on speech; the appropriate mapping from such a representation to synthesizer parameters; and the synthesizer features needed to generate convincing affect.

The true test of synthesized affect is perceptual. An experiment was performed to test whether the intended affect was reproduced by the Affect Editor. The results confirmed that the intended affect was recognized, and furthermore, bore out predictions about areas of confusion. This work supports the conclusion that affect can be generated and systematically controlled in synthesized speech. The limits of the Affect Editor indicate that more research is needed into the determination of useful taxonomies of emotion and of the speech correlates of emotion. Better synthesizers are also needed, to enable more precise testing and development, and eventually, real–time generation of affect in synthesized speech.

One audio cassette tape accompanies this document. *Side A* contains: (1) seventeen utterances from which figures in this document are generated and (2) the stimuli used in the experiment described in Chapter 6 (30 unique utterances). *Side B* contains: (1) sentences synthesized with varied affect specifications (2) seventeen utterance sequences — one per Affect Editor parameter — in which parameters values are cycled from low to high.

Thesis Supervisor: Chris Schmandt
Title: Principal Research Scientist

This work was supported by Nippon Telegraph and Telephone Company, as part of the Personal Computing and Telephony project, and by the U.S. Department of Air Force (Headquarters, Rome Air Development Center), as part of the Advanced Concurrent Interfaces project, contract number F30602-89-C-0022.

[1] At the time, the Media Lab's version of Latex could only handle one hundred and fifty pages; thus, the M.I.T. Library copy is missing Appendix C.