MATTI KARJALAINEN, TOOMAS ALTOSAAR, MARTTI VAINIO (Espoo)

FINNISH SPEECH SYNTHESIS USING WARPED LINEAR PREDICTION AND NEURAL NETWORKS

A text-to-speech synthesis technique, based on warped linear prediction (WLP) and neural networks, is presented for high-quality individual sounding synthetic speech. Warped linear prediction is used as a speech production model with wide audio bandwidth yet with highly compressed control parameter data. An excitation codebook, inverse filtered from a target speaker's voice, is applied to obtain individual tone quality. A set of neural networks, specialised to yield synthesis control parameters from phonemic input in specific contexts, generate the detailed parametric controls of WLP. Neural nets are also used successfully to compute the prosodic parameters. We have applied this approach in prototyping improved text-to-speech synthesis for the Finnish language.

1. Introduction and motivation

After a long period of successful developments in text-to-speech (TTS) synthesis, voice quality still remains a challenge. No practical technique yields wide audio bandwidth, near human quality, and individual sounding speech.

Our effort in this study was to find a strategy to improve TTS synthesis for the Finnish language. Earlier achievements were first based on traditional formant synthesis with rule-based control, SYNTE 2 and 3 (Karjalainen, Laine, Toivonen 1980), and then concatenation synthesis called microphonemic synthesis (Lukaszewicz, Karjalainen 1987) similar to the PSOLA technique (Moulines, Carpenter 1990). Concatenative synthesis, based on samples from human speech, easily captures the features from individual speakers. In order to approach full naturalness, however, a huge inventory of samples in different contexts is needed. The algorithms to select concatenative units and to join them in synthesis tend to become complex.

Source-filter models for speech synthesis, such as those used in linear prediction, have more flexibility and allow for easy analysis of control data. The problem remains how to code the excitation (source) and the filter control parameters in a compact way and be able to recompute them from phonemic/phonetic information. Hand-tuned rules and tables, as used in early synthesis, cannot produce the highest quality of speech. Tables of parameter trajectories have similar problems as concatenative synthesis: the size of such inventories grows beyond practical limits when contextual details are included. Among the techniques that are used to