MARTTI VAINIO, TOOMAS ALTOSAAR,
MATTI KARJALAINEN, REIJO AULANKO (Helsinki)

MODELING FINNISH MICROPROSODY WITH NEURAL NETWORKS

In this study of Finnish microprosody, two prosodic parameters — pitch and loudness — were modeled with artificial neural networks. The networks are of the general feed forward type trained with backpropagation. For each phoneme, the network predicts a series of either pitch or loudness values on the basis of information of the phoneme's phonologically motivated features and its phonetic environment. The tests we have run so far indicate that the neural networks are highly successful and accurate in modeling the micro-level behavior of both pitch and loudness.

1. Introduction

Pitch-related microprosodic variation has been well attested for several languages including Finnish. For instance, the fundamental frequency difference between open and close vowels and the effect of immediate consonant context on the F0 of a vowel seem to be universal (Whalen, Levitt 1995; Aulanko 1985; Vilkman, Aaltonen, Raimo, Arajärvi, Oksanen 1989). Similar variation can be observed with regard to loudness. The most well known phenomenon is the difference between the inherent loudness levels of, e.g., open vs. close vowels and sonorant vs. obstruent consonants (Lehiste, Peterson 1959).

The microprosodic characteristics can be seen as the lowest level of a multi-layered prosodic system producing the final suprasegmental realization of speech. They are not generally seen as a part of the linguistic-prosodic pattern of the utterance, but rather to be segmentally conditioned. That is, they reflect the gestures necessary for producing the specific articulatory movements for various vowels and consonants.

In speech synthesis, microprosodic modeling has usually been fairly scarce — the developers have concentrated on more salient and urgent problems and the modeling has usually remained on a first approximation level. In general, speech synthesizers use some information about the intrinsic pitch, loudness and duration of speech sounds which are changed algorithmically according to certain rules that take the sounds' context into account. The microscopic changes within the time-varying parameters of the sounds have not been paid much attention to, although most synthesis systems do model the timing of F0 peaks and differences in F0 slopes and onsets after different consonants. It is probable that the inclusion of microprosodic variation would improve the naturalness and even the intelligibility of synthesized speech.

It can be argued that microprosodic variation is analogous to variation in other aspects of speech in that there are both phenomena that are extremely common