A Prosodic Diphone Database for Korean Text-to-Speech Synthesis System

Kyuchul Yoon

The Ohio State University, Columbus OH 43220, USA,
kyoon@ling.osu.edu,
http://ling.osu.edu/∼kyoon

Abstract. This paper presents a prosodically conditioned diphone database to be used in a Korean text-to-speech (TTS) synthesis system. The diphones are prosodically conditioned in the sense that a single conventional diphone is stored as different versions taken directly from the different prosodic domains of the prosodically labeled, read sentences (following the K-ToBI prosodic labeling conventions [3]). Four levels of the Korean prosodic domains were observed in the diphone selection process, thereby selecting four different versions of each diphone. A 400-sentence subset of the Korean Newswire Text Corpora [5] were converted to its pronounced form as described in [8] and its read version was prosodically labeled. The greedy algorithm [7] identified 223 sentences containing 1,853 prosodic diphones (out of the 3,977 possible prosodic diphones) that can synthesize all four hundred utterances. Although our system cannot synthesize an unlimited number of sentences at this stage, the quality of the synthesized sentences strongly suggests that it is a viable option to use prosodically conditioned diphones in a text-to-speech synthesis system.

1 Introduction

Work on Korean shows that segmental properties are affected by the prosody of an utterance. In a study on the effect of prosodic domains on segmental properties of three Korean coronal stops /t, t^h, t*/, Cho and Keating [1] showed that initial consonants in higher prosodic domains are articulatorily stronger and longer than those in lower domains: the former has more linguopalatal contact and is longer in duration than the latter. Acoustic properties such as VOT, total voiceless interval, percent voicing during closure, and nasal energy minimum were also found to vary with prosodic position. Prosodic effects on segments have also been found in Korean fricatives. In her study on Korean coronal fricatives, Kim [4] observed prosodic effects on segmental properties. Linguopalatal contact was greater, acoustic duration was longer, centroid frequency was higher, and H1-H2 value for /s*/ was lower in higher domains than in lower domains. Yoon [9] looked at two Korean voiceless coronal fricatives and found that each fricative in different prosodic positions displayed characteristics that appear to signal its prosodic location by means of durational differences. Motivated by these findings,