AUTOMATIC CLASSIFICATION OF BISYLLABIC STRESS PATTERNS FOR L2 LEARNING

Mostafa Ali Shahin, Beena Ahmed

Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar
mostafa.shahin@qatar.tamu.edu , beena.ahmed@qatar.tamu.edu

Introduction

The need of learning a second language is increasing as the world becomes more interconnected and millions of people are trying to acquire a new language. Most of the difficulties in correctly pronouncing words from a second language come actually from the pronunciation variations between the learner's mother tongue and the acquired language. As English is a stress-timed language, lexical stress plays an important role in the perception and processing of speech by native speakers. Learners whose first language is syllable-timed have difficulties in changing the stress level of the different syllables in the multi-syllabic words. In this paper we proposed a method to automatically classifying the stress patterns of each two consecutive syllables in the production of English multi-syllabic words into strong-weak (SW), weak-strong (WS), strong-strong (SS) and weak-weak (WW). This method can be used in a computer-assisted pronunciation training (CAPT) system to teach the user how to control stress level. The method is based on a deep neural network (DNN) classifier which is trained by a set of features derived from the duration, pitch and intensity of each of the two consecutive syllables along with a set of energies of different frequency bands.

Method

Our lexical stress classifier is applied on the speech signal along with the prompted word. Figure 1 shows a block diagram of the overall system. The speech signal is first force aligned with the predetermined phoneme sequence of the word to obtain the time boundaries of each phoneme. The alignment is performed using a Hidden Markov Model (HMM) Viterbi decoder along with a set of HMM acoustic models trained from the same corpus to reduce the error caused by inaccurate phone level segmentation. A set of features, listed in table 1, are then extracted from each syllable and the features of each pair of consecutive syllables combined by concatenating them into one wide feature vector used to train DNN classifier. The method was trained and tested with a speech corpus collected from ~500 children ranging from grad 0 to 10, each pronouncing 200 single words. The system achieved an overall accuracy of approximately 88%.

Feature	Description
f_1	Peak-to-peak amplitude over syllable nucleus
f_2	Mean energy over syllable nucleus
f_3	Maximum energy over syllable nucleus
f_4	Nucleus duration
f_5	Syllable duration
f_6	Maximum pitch over syllable nucleus
f_7	Mean pitch over syllable nucleus
f_8	27 Mel-scale energy bands over syllable nucleus

Table 1: *The extracted acoustic features*

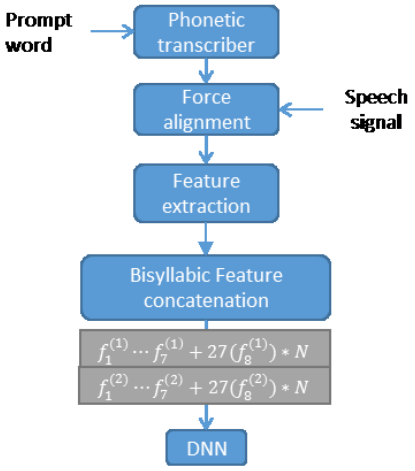

Figure 1: *System block diagram*