Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 69 (2015) 31–45

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems

Luciana Ferrer *, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, Kristin Precoda

Speech Technology and Research Laboratory, SRI International, CA, USA

Received 15 May 2014; received in revised form 23 January 2015; accepted 5 February 2015
Available online 18 February 2015

Abstract

We present a system for detection of lexical stress in English words spoken by English learners. This system was designed to be part of the EduSpeak® computer-assisted language learning (CALL) software. The system uses both prosodic and spectral features to detect the level of stress (unstressed, primary or secondary) for each syllable in a word. Features are computed on the vowels and include normalized energy, pitch, spectral tilt, and duration measurements, as well as log-posterior probabilities obtained from the frame-level mel-frequency cepstral coefficients (MFCCs). Gaussian mixture models (GMMs) are used to represent the distribution of these features for each stress class. The system is trained on utterances by L1-English children and tested on English speech from L1-English children and L1-Japanese children with variable levels of English proficiency. Since it is trained on data from L1-English speakers, the system can be used on English utterances spoken by speakers of any L1 without retraining. Furthermore, automatically determined stress patterns are used as the intended target; therefore, hand-labeling of training data is not required. This allows us to use a large amount of data for training the system. Our algorithm results in an error rate of approximately 11% on English utterances from L1-English speakers and 20% on English utterances from L1-Japanese speakers. We show that all features, both spectral and prosodic, are necessary for achievement of optimal performance on the data from L1-English speakers; MFCC log-posterior probability features are the single best set of features, followed by duration, energy, pitch and finally, spectral tilt features. For English utterances from L1-Japanese speakers, energy, MFCC log-posterior probabilities and duration are the most important features.

Keywords: Computer-assisted language learning; Lexical stress detection; Mel frequency cepstral coefficients; Prosodic features; Gaussian mixture models

1. Introduction

Lexical stress is an important component of English pronunciation, as English makes a greater use of stress than many other languages. To understand spoken words, native[1] speakers of English rely not only on the pronunciation of sounds, but also on the stress patterns. Using the incorrect stress pattern can greatly reduce a speaker's intelligibility. This poses a big problem for English learners, especially for native speakers of languages that have more consistent lexical stress patterns or have different ways of incorporating timing and rhythm. This is especially

* Corresponding author at: Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

[1] A note on nomenclature: Throughout the paper we will use the word "native" to refer to the L1 of a speaker and, also, to whether the language being spoken is the speaker's L1. Hence, the phrase "native English speakers" refers to L1-English speakers, the phrase "native Japanese speakers" refers to L1-Japanese speakers, and the phrase "non-native English speakers" refers to speakers with L1 other than English. Furthermore, we will call "native data" any data where the language spoken is the same as the L1 of the speakers, and "non-native data" any data where the language spoken is not the same as the L1 of the speakers. When no language is specified, native and non-native refer to native English and non-native English (data or speakers), respectively.

http://dx.doi.org/10.1016/j.specom.2015.02.002
0167-6393/

true for native Japanese speakers learning English: in Japanese, the rhythm is more regular and syllables are more similar in prominence than in English. Computer-assisted language learning (CALL) software can then greatly benefit from the ability to provide feedback about stress pronunciation to the user.

A large variety of automatic systems that use different features and modeling techniques to classify stress have been proposed in the literature. Unfortunately, as we explain below, many of them are unsuitable for use in CALL systems because the assumptions they make do not apply to language learners. Many others were not tested on non-native speakers of the language for which the system was trained and, hence, their suitability for CALL systems is unknown.

Most proposed stress classification systems are based on prosodic features like pitch, energy and duration, which are normalized in different ways to make them independent of the speaker's baseline pitch, the channel volume, the speech rate and so on. Measurements are generally obtained only over the nucleus for each syllable. Examples of this kind of segmental features can be found in several papers (Tepperman and Narayanan, 2005; Chen and Wang, 2010; Deshmukh and Verma, 2009; Chen and Jang, 2012; Verma et al., 2006; Zhu et al., 2003). Spectral features, on the other hand, have been rarely used for stress detection. Li et al. (2007) and Lai et al. (2006) propose similar systems using mel-frequency cepstral coefficients (MFCCs) modeled by hidden Markov models (HMMs). Both papers address the problem of detecting English sentence-level stress rather than word-level stress and test only on data from native English speakers.

Modeling techniques for stress detection vary widely and include decision trees (Deshmukh and Verma, 2009), Gaussian mixture models (GMMs) (Tepperman and Narayanan, 2005; Chen and Jang, 2012), support vector machines (Deshmukh and Verma, 2009; Chen and Wang, 2010; Zhao et al., 2011), deep belief networks (Li et al., 2013), and HMMs (Lai et al., 2006; Li et al., 2007; Ananthakrishnan and Narayanan, 2005). In many cases, the task of stress detection is defined as the problem of locating the single primary stressed syllable in a word. Under this assumption, modeling techniques can make a single decision per word – rather than one decision per syllable – using features extracted from all syllables in the word (Chen and Wang, 2010; Chen and Jang, 2012) or obtain syllable-level scores and then choose the syllable with the largest score as the primary stress location (Tepperman and Narayanan, 2005; Zhao et al., 2011). Furthermore, some techniques require that words have correct phonetic pronunciation in order to make a stress level decision (Chen and Jang, 2012). Finally, the task of labeling each syllable in an utterance from a non-native English speaker as unstressed, primary stressed or secondary stressed is an extremely complex one. In our database, the observed disagreement for native Japanese children speaking English across three annotators is, on average, 21% (corresponding to an agreement of 79%). Given this difficulty, some researchers simplify the labeling task by asking annotators to assign "correct" versus "incorrect" labels to each word rather than actual stress pronounced on each syllable (Deshmukh and Verma, 2009; Verma et al., 2006) or by labeling only the location of the primary stress (Tepperman and Narayanan, 2005; Chen and Jang, 2012). Many of these modeling and labeling assumptions are inappropriate for language learners who will most likely mispronounce both phones and stress within a word and might pronounce more than one syllable with primary stress.

We describe a novel system for lexical stress feedback intended for use by native Japanese children learning English. We expect the learners to pronounce sounds poorly and to pronounce most syllables with more prominence than native English speakers would. In fact, according to our phonetician's annotations, in our Japanese children's database around one third of the incorrectly stressed words have primary stress in at least two syllables. Therefore, our system must allow more than one syllable with primary stress in a word. Furthermore, phonetic and stress pronunciations are tied together; pointing out a stress mistake might go a long way toward fixing the phonetic mistakes, and conversely. For this reason, we do not wish to assume correct phonetic pronunciation before giving feedback about the stress pronunciation.

The proposed system is designed to approximate the decisions a phonetician would make about the stress level pronounced for every syllable in a word. For the Japanese children data, the system is evaluated against decisions made by annotators. The goal is to approximate those decisions as well as possible. Hence, the most natural approach would be to train such a system using data from the same population of Japanese children speaking English. This way, the model would describe the stress level as pronounced by this population of speakers. Nevertheless, since the stress labeling task is costly and agreement is low, little amount of data is available with reliable labels for training the system. For this reason, we propose to use utterances from native English speakers to train our system. For this data, stress labels are obtained automatically, assuming that native English speakers pronounce stress in a predictable manner for selected words according to a dictionary. While this approach results in models that represent stress as pronounced by native English speakers, we show that it results in good performance on the Japanese children's data. Matched Japanese children's data can then be used to fine-tune the system through adaptation of the models.

The decisions made by the system are meant to be used as a tool within CALL software. The software could be designed to only correct the speaker when the stress mistake would result in intelligibility problems (for example, when the meaning of the word depends on the stress pattern). On the other hand, the software could aim at achieving native-like pronunciation, correcting the speakers every time they make a mistake, regardless of whether this would