Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 31 (2015) 28–48

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Native and non-native class discrimination using speech rhythm- and auditory-based cues☆

S.-A. Selouani[a,*], Y. Alotaibi[b], W. Cichocki[c], S. Gharsellaoui[a,d], K. Kadi[e]

[a] *Université de Moncton, Shippagan Campus, New Brunswick E8S 1P6, Canada*
[b] *Department of Computer Engineering, King Saud University, Riyadh, Saudi Arabia*
[c] *Department of French, University of New Brunswick, Canada*
[d] *Department of Computer and Electrical Engineering, UQTR, Canada*
[e] *Faculty of Electronics and Computer Science, USTHB University, Algiers, Algeria*

Received 29 October 2013; received in revised form 3 October 2014; accepted 10 November 2014
Available online 22 November 2014

Abstract

In recent years, the use of rhythm-based features in speech processing systems has received growing interest. This approach uses a wide array of rhythm metrics that have been developed to capture speech timing differences between and within languages. However, the reliability of rhythm metrics is being increasingly called into question. In this paper, we propose two modifications to this approach. First, we describe a model that is based on auditory cues that simulate the external, middle and inner parts of the ear. We evaluate this model by performing experiments to discriminate between native and non-native Arabic speech. Data are from the West Point Arabic Speech Corpus; testing is done on standard classifiers based on Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) and a hybrid GMM/SVM. Results show that the auditory-based model consistently outperforms a traditional rhythm-metric approach that includes both duration- and intensity-based metrics. Second, we propose a framework that combines the rhythm metrics and the auditory-based cues in the context of a Logistic Regression (LR) method that can optimize feature combination. Further results show that the proposed LR-based method improves performance over the standard classifiers in the discrimination between the native and non-native Arabic speech.

Keywords: Speech rhythm; Auditory cues; Duration; Intensity; SVMs; GMMs; Hybrid GMM/SVM; Logistic regression

1. Introduction

Recent modeling in speech processing systems has used rhythm-based features to differentiate between speech varieties. One example is an automatic language identification study by Rouas et al. (2005), who modeled vocalic and consonantal durations with a Gaussian Mixture to distinguish among stress-timed, mora-timed and syllable-timed

☆ This paper has been recommended for acceptance by L. ten Bosch.

* Corresponding author. Tel.: +1 506 336 3401.
E-mail addresses: selouani@umcs.ca (S.-A. Selouani), yaalotaibi@ksu.edu.sa (Y. Alotaibi), cicho@unb.ca (W. Cichocki), Soumaya.Gharsellaoui@uqtr.ca (S. Gharsellaoui), kkadi@usthb.dz (K. Kadi).

http://dx.doi.org/10.1016/j.csl.2014.11.003
0885-2308/

language groups. The modeling was also able to identify the seven languages in the study with reasonable success. This rhythm-based approach uses a wide array of speech rhythm measures that have been developed to capture timing properties of different languages. Among the numerous other applications of these measures is the study of the rhythm component of second language phonology. White and Mattys (2007) found that certain metrics discriminate between native speakers of English and Spanish learners of English with more success than other metrics, specifically, %V and VarcoV vs. nPVI-V (these metrics are defined below).

The present study focuses on the classification of native (L1) and non-native (L2) Arabic speech. In our previous work on Arabic (Alotaibi and Selouani, 2009), we found that there are clear auditory differences between native and non-native speech; among the most noticeable features are speaking rate and the length of vowels and consonants. Besides this, it is worth noting that regional and foreign linguistic variation of Arabic speech rhythm have been the focus of only a small number of studies. Ghazali et al. (2002) documented rhythmic variation across geographic regions; Eastern Arabic dialects (Syria and Jordan) are distinct from Western dialects (Morocco and Algeria) with Tunisian and Egyptian dialects occupying an intermediate position. Meftah et al. (2013) found that even within the same dialect there can be rhythmic differences between male and female speakers. Improvements to the robustness of automatic speech recognition systems of Arabic require improved modeling of different accents, including those of L2 speakers.

In the next two subsections of this introduction, we will present the main concepts and definitions that we use in this study as well as a summary of our main contributions.

1.1. Measuring rhythm

In terms of rhythm, the traditional classification of languages is based on the principle of isochrony, attributed to Pike (1945) and Abercrombie (1967). Isochrony is defined as the property of speech to organize itself into portions of equal or equivalent durations. According to this perspective, languages such as English and Swedish are considered to be stress-timed, and their fundamental unit for equal-timed intervals is the foot. Syllable-timed languages, such as French and Italian, have the syllable as the fundamental isochronous unit. Based on this classification, Beckman (1992) investigated the basis of the rhythm class hypothesis; however, like other researchers, she found a lack of experimental support for isochrony in speech. Laver (1994) noted that, despite the popularity of the rhythm class hypothesis among linguists, researchers have not found support from duration measurements for isochronous timing on any absolute basis.

The weak empirical evidence for isochrony led Ramus et al. (1999) and Grabe and Low (2002) to propose an approach to the description of the rhythmic structure of languages that uses acoustic-phonetic measurements. They suggested a number of speech rhythm metrics, based on the durations of vocalic and intervocalic (consonantal) segments, that can capture the timing properties of different languages. Many studies have used these rhythm metrics to investigate cross-language differences and to classify languages into syllable-timed and stress-timed categories (Arvaniti, 2009; Wiget et al., 2010). Studies of the rhythm component in second language phonology have also made use of rhythm metrics (Gut, 2012; White and Mattys, 2007). Similarly, research on language dialects (Ferragne and Pellegrino, 2008; White et al., 2009) and on speech impairments (Liss et al., 2009) have taken this approach.

However, despite their popularity, researchers have noted a number of shortcomings of rhythm metrics. Arvaniti (2012) finds that rhythm metrics fail to capture consistent cross-linguistic differences because of large within-language variability that is due to effects such as elicitation methods, test materials and inter-speaker variation. Gut (2012) observes that most rhythm metrics do not yield similar results across different studies, even for the same language; this leads her to question the reliability of these measures.

Other considerations cast doubt on the validity of rhythm metrics. Some researchers suggest that the notion of rhythm is not limited to the duration of segments such as vowels and consonants (Dauer, 1983; Nolan and Asu, 2009). They argue that rhythm is located at a level above the segment; that is, rhythm is seen as the result of phonological, lexical and syntactic facts (Cummins and Port, 1998). Given these considerations, Bertinetto and Bertini (2010) conclude that the rhythm-metric model is not entirely predictive. Cumming (2008) suggests the inclusion of other prosodic features such as fundamental frequency in the definition of rhythm. Ferragne and Pellegrino (2008) and He (2012) study the inclusion of intensity. Thus, a more effective quantitative model of rhythm may lie in