Speech Communication 92 (2017) 13–22

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Dialect classification using vowel acoustic parameters

Charalambos Themistocleous

Centre for Linguistic Theory and Studies in Probability, University of Gothenburg, Dicksonsgatan 4,412 56, Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 29 October 2016
Revised 11 May 2017
Accepted 22 May 2017
Available online 25 May 2017

MSC:
00-01
99-00

Keywords:
Dialect classification
Stress classification
Formant dynamics
C5.0
Vowels

ABSTRACT

This study provides a classification model of two Modern Greek dialects, namely Athenian Greek and Cypriot Greek, using information from formant dynamics of *F*1, *F*2, *F*3, *F*4 and vowel duration. To this purpose, a large corpus of vowels from 45 speakers of Athenian Greek and Cypriot Greek was collected. The first four formant frequencies were measured at multiple time points and modelled using second degree polynomials. The measurements were employed in classification experiments, using three classifiers: Linear Discriminant Analysis, Flexible Discriminant Analysis, and C5.0. The latter outperformed the other classification models, resulting in a higher classification accuracy of the dialect. C5.0 classification shows that duration and the zeroth coefficient of *F*2, *F*3 and *F*4 contribute more to the classification of the dialect than the other measurements; it also shows that formant dynamics are important for the classification of dialect.

1. Introduction

Spoken dialect identification refers to the process of determining the identity of a dialect based on acoustic evidence. There are two main theoretical and methodological approaches to spoken dialect identification. The first originates from studies on sociophonetics and aims to explain the variation that exists between dialects, sociolects, speech styles, and registers, and the causes that drive language variation and change (Foulkes and Docherty, 2006; Foulkes et al., 2010; Thomas, 2011; 2013). The second approach is automatic dialect classification, which aims to develop technologies for dialect identification in a wide range of speech processing applications, such as in speech-to-text systems, spoken document retrieval, spoken language translation, and in dialogue systems (see Li et al., 2013, for a review), and may result in high classification accuracy of dialects (see Glembek et al., 2009; Dehak et al., 2010; Behravan et al., 2015). Yet, to model speech variation, automatic dialect classification methods (e.g., Joint Factor Analysis and i-vector architectures), employ hyper-parameters that can be hard to interpret for the purposes of sociophonetic research (see Glembek et al., 2009; Dehak et al., 2010; Behravan et al., 2015).

The purpose of this study is to offer an account of dialect variation in terms of vowel formants and vowel dynamics, using machine learning methods often employed in automatic dialect classification. To this purpose, this study classifies two varieties, Athenian Greek (AG) and Cypriot Greek (CG), whose phonemic inventories contain the same vowels: /i e a o u/ (Kontosopoulos, 1968; Newton, 1972a; 1972b; Samaras, 1974; Botinis, 1981; Tseva, 1989; Jongman et al., 1989; Nicolaidis, 2003; Sfakianaki, 2002; Fourakis et al., 1999; Loukina, 2009; 2011; Themistocleous, 2017). Specifically, AG and CG vowels were produced in controlled settings and measured at multiple time points, which were then evaluated using a second degree (2nd) polynomial fit. This method provides a rich description of vowel formants, as it considers the fluctuation of a formant's frequency with respect to time and does not rely on a single measurement of formants at the middle of the vowel's duration (e.g. Cooper et al., 1952; Lisker, 1957; 1975; Stevens and Klatt, 1974; Rosner and Pickering, 1994; Themistocleous, 2017). The polynomial fit is appealing as each of the coefficients of the polynomial relates to characteristics of a formant contour, such as its position on the frequency axis (zeroth order coefficient) and the shape of the curve (see also Cohen, 1995). Earlier studies by McDougall (2005; 2006) and McDougall & Nolan (2007) using polynomial equations for regression showed that 2nd and 3rd degree polynomials perform better at 89–96% than raw data and static measurements of vowels (see also Van Der Harst et al., 2014). A key difference of this study with respect to most automatic language identification studies is that it employs a text-dependent approach, whereas most other studies on language identification employ a text-independent approach (see for a discussion of these approaches Atal, 1974; Doddington, 1985; Farrell et al., 1994; Furui, 1997; Gish and Schmidt, 1994; Reynolds and Rose, 1995; Larcher et al., 2014; Mporas et al., 2016).

E-mail addresses: charalambos.themistocleous@gu.se, themistocleous@gmail.com

http://dx.doi.org/10.1016/j.specom.2017.05.003

For the classification, we evaluated three different types of discriminative classifiers: Linear Discriminant Analysis (LDA), Flexible Discriminant Analysis (FDA), and C5.0. These classifiers as opposed to generative models, such as Naive Bayes and Hidden Markov Models (HMMs) do not rely on prior distributions and learned states (Zhang, 2014). The discriminative classifiers identify a class of a specific observation, e.g., the dialect by generalizing from previous measurements. Details for each classifier are provided below.

First, LDA is a classifier, which is very similar to multi-response regression. It permits the evaluation of a binary dependent variable using both continuous and categorical predictors (Harrington and Cassidy, 1999). Specifically, it attempts to find a linear structure in the predictors that can best separate two or more groups. LDA relies on the Bayesian probability, the maximum likelihood assumption, and requires that the data are normally distributed. A number of studies by Najim Dehak and colleagues showed that LDA can potentially provide good classification outcomes when employed in the reduction of i-vector dimensionalities of acoustic properties in state of the art i-vector architectures for speaker (Dehak et al., 2010; Sadjadi et al., 2016), accent (Behravan et al., 2015), and language classification (Dehak et al., 2011; Sizov et al., 2017).

FDA employs non-parametric techniques for the classification of categorical variables (Trevor et al., 1994). So unlike LDA, it does not require that the data are normally distributed. Because not all the predictors of this study are normally distributed, FDA is expected to offer a better classification accuracy than LDA.

C5.0 is a classification algorithm developed by Quinlan (1993). It assesses class factors, such as the dialect, based on a predefined set of predictors. C5.0 generates a decision tree and offers a ranking of features that can indicate the contribution of each acoustic feature in the classification. Specifically, it evaluates recursively the data and employs the predictors that can provide the best splitting of the data into more refined categories. The splitting criterion is the difference in information entropy (a.k.a., the normalized information gain). The predictor that provides the highest normalized information gain is the one selected for the decision (see also Woehrling et al., 2009, who provide a classification of regional French varieties, using a different decision tree method). Typically, each split is an interpretation of the variation or *impurity* in the data. The algorithm will stop when there are not enough data left to split. Finally, C5.0 provides both tree and rule models (for an application of C4.5, which is an earlier iteration of C5.0, on accent classification, see Vieru et al. (2011) and for the classification of stressed and unstressed fricatives using C5.0, see Themistocleous et al. (2016)).

To evaluate the effects of vowel acoustic properties on dialect classification, we also provide classification results for stress and vowel. A syllable in Modern Greek can be stressed or unstressed; the position of the stress in a Modern Greek word can change the meaning of the word, e.g., miˈlo ‘speak’ vs. ˈmilo ‘apple’. Stressed vowels are overall longer and more peripheral than the unstressed (e.g., Botinis, 1989; Arvaniti, 1991; Themistocleous, 2014; 2015). We also provide comparative classification models for vowels; yet, unlike previous studies that provide acoustic evidence mainly from AG vowels (Kontosopoulos, 1968; Samaras, 1974; Botinis, 1981; Tseva, 1989; Jongman et al., 1989; Nicolaidis, 2003; Sfakianaki, 2002; Fourakis et al., 1999; Loukina, 2009; 2011; Themistocleous, 2017), this study provides cross-dialectal evidence from AG and CG (see, however Themistocleous, 2017). Also, all previous studies on Modern Greek vowels rely on single acoustic measurements of formant frequencies at the middle of the vowel whereas this is the first study to analyze formant dynamics of Greek vowels.

Table 1
Speech material.

Vowel	stressed	unstressed	stressed	unstressed
/e/	ˈesa	eˈsa	ˈsesa	seˈsa
/i/	ˈisa	iˈsa	ˈsisa	siˈsa
/a/	ˈasa	aˈsa	ˈsasa	saˈsa
/o/	ˈosa	oˈsa	ˈsosa	soˈsa
/u/	ˈusa	uˈsa	ˈsusa	suˈsa

2. Methodology

This section presents the methods employed for the collection and analysis of the acoustic data. It also presents the selection criteria for the classification model reported in the paper.

2.1. Speakers

A large corpus of AG and CG vowels was recorded in Athens and Nicosia. These urban areas constitute the capital cities of Greece and Cyprus respectively. 45 female speakers between 19 and 29 years participated in the study: 25 CG speakers and 20 AG speakers. All speakers were born and raised in Nicosia and Athens respectively. Based on information from a demographic questionnaire, the participants from each dialect constituted sociolinguistically homogeneous groups: they originated from approximately the same socio-economic background and they were all university students, namely all CG speakers were students at the University of Cyprus and all AG speakers were students at the University of Athens. All participants knew English as a second language; four AG participants knew French as a third language. None reported a speech or hearing disorder.

2.2. Speech materials

The speech materials consisted of a set of nonsense words, each containing one of the five Greek vowels (/ e i a o u /) in both stressed and unstressed position, word initially and word medially. The nonsense words had the structure V̀*sa* (e.g., /ˈasa, ˈesa, ˈisa, etc./) or V*sà* (e.g., /aˈsa, eˈsa, iˈsa, etc./) sV̀sa (/ˈsasa, ˈsesa, ˈsisa, etc./) and sVsà (/saˈsa, seˈsa, siˈsa, etc./) and were embedded in the following carrier phrases (Table 1).

The AG carrier phrase was:

“ˈipes < *keyword* > ˈpali” (You told < *keyword* > again) and the CG carrier phrase was:

“/ˈipes < *keyword* > ˈpale/” (You told < *keyword* > again).

Each subject produced 80 utterances (i.e., 5 vowels × 2 stress conditions × 2 word placement conditions × 4 repetitions), resulting in a total of 3600 productions. To facilitate vowel segmentation and to control formant transitions at the beginning and the end of a vowel, the voiceless alveolar fricative [s] was selected as the immediate segmental environment–before and after–the designated vowel. Filler words were added in the carrier sentences to provide variation within the experimental material and to minimize speaker’s attention on the experimental words.

2.3. Procedures

The recordings were conducted in a recording studio in Athens, Greece and in a quiet room at the University of Cyprus in Nicosia, for the AG and CG speech material respectively. Two researchers, a female AG speaker and a male CG speaker (the author), provided standard instructions to the speakers before the recording, e.g., to speak at a normal pace, sit appropriately in front of the microphone, and keep a designated distance. The target words