Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 294–306

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Principal differential analysis for detection of bilabial closure gestures from articulatory data☆

Farook Sattar[a], Frank Rudzicz[a,b,*]

[a] *Toronto Rehabilitation Institute, Toronto, Canada*

[b] *Department of Computer Science, University of Toronto, Toronto, Canada*

Received 12 June 2014; received in revised form 21 February 2015; accepted 6 July 2015

Available online 15 July 2015

Abstract

In this paper, a new statistical method for detecting bilabial closure gestures is proposed based on articulatory data. This can be surprisingly challenging, since mere proximity of the lips does not imply their involvement in a directed phonological goal. This segment-based bilabial closure detection scheme uses principal differential analysis (PDA) to extract articulatory gestures. The dynamic patterns of the *tract variables* (TVs) lip aperture, lip protrusion, and their derivatives, are captured with PDA and used to detect and quantify bilabial closure gestures. The proposed feature sets, which are optimized using sequential forward floating selection (SFFS), are combined and used in binary classification. Experimental results using the articulatory database MOCHA-TIMIT show the effectiveness of the proposed method demonstrating promising performance in terms of high classification accuracy (95%), sensitivity (95%), and specificity (95%).

Keywords: Bilabial closure gesture; Principal differential analysis (PDA); Articulatory data; Tract variables; Classification

1. Introduction

Bilabial closure often precedes sudden airflow out of the mouth, as in the plosive onset in *pie*, *buy*, or is concurrent with airflow out of the nose, as in the nasal onset in *my* (Lofqvist and Gracco, 2010). In either case, the resulting acoustics can be fairly similar to other phones in the respective classes, which can affect speech recognition (Rudzicz, 2011). However, articulatory features are robust across speakers, share commonalities (in many cases) across languages, and are generally insensitive to situational changes (Zhao et al., 2013). We propose a new scheme for detecting articulatory *gestures* based on principal differential analysis (PDA) and articulatory data. Here, gestures are directed reconfigurations of the vocal tract to achieve a discrete phonologically relevant goal, such as lowering the velum. For this purpose, we have focused on directed bilabial gestures in the MOCHA-TIMIT database (Wrench, 2000), which consists of electromagnetic articulography (EMA) data tracking the positions and velocities of point-sensors affixed to the articulators.

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author at: Toronto Rehabilitation Institute, Toronto, Canada. Tel.: +1 416 597 3422x7971.
E-mail address: frank@cs.toronto.edu (F. Rudzicz).

http://dx.doi.org/10.1016/j.csl.2015.07.002
0885-2308/

Task dynamics is a combined model of skilled articulator motion and abstract vocal tract configuration (Saltzman and Munhall, 1989) that provides a coherent and biologically plausible model of speech production with consequences for phonology (Browman and Goldstein, 1986), neurolinguistics, and the evolution of speech and language (Goldstein et al., 2006). In this theory, tract variables (TVs) generally refer to the locations and degrees of vocal tract constrictions, as functions of time. Each *gesture* is a directed motion to complete some phonologically or acoustically relevant task within one of the following TVs: lip aperture (LA), lip protrusion (LP), tongue tip constriction location (TTCL) and degree (TTCD), tongue dorsum constriction location (TDCL) and degree (TDCD), velum (VEL), glottis (GLO), and lower tooth height (LTH). For instance, a gesture to close the lips would occur within the LA variable and would set that variable to zero. The dynamic influence of each gesture in time on the relevant tract variable is modeled by a non-homogeneous second-order linear differential equation mimicking a highly coupled spring-mass system. Typically, the coefficients of these systems have been set empirically by experts (Saltzman and Munhall, 1989), but with several exceptions. For example, McGowan (1994) used a genetic algorithm to recover task dynamic parameters from acoustic speech signals, and Lammert et al. (2013) use both artificial neural networks and locally-weighted regression to estimate kinematic relationships of speech production in task dynamics. Similarly, Howard and Huckvale (2005) train an inverse mapping between an articulatory synthesizer's control parameters and their auditory consequences, in a manner similar to work done with the DIVA system (Guenther and Perkell, 2004), and Nam et al. (2012) use an iterative analysis-by-synthesis procedure using time-warping in task dynamics to learn relevant parameters. Although clear differences exist between these efforts, in general their aims were to realistically estimate and simulate gestural dynamics given data. By contrast, our approach of using principal differential analysis (which is unique among this work) is primarily a means towards an end, namely the classification of articulatory features.

2. Proposed method

The EMA data and their associated, time-aligned acoustics are segmented by phone annotations. These phone annotations allow us to simply extract phonological characteristics (Ali et al., 1999; Hosom, 2000; Ladefoged and Johnson, 2011), therefore all data are segmented by provided phone boundaries and labelled as either bilabial or non-bilabial. We note that, in this database, these phone boundaries have been determined using forced alignment on the acoustics, therefore some errors are possible. The raw EMA data are transformed to tract variables, as described in Section 3. Each tract variable is assumed to be governed by functional differential equations whose parameters must take into account the substantial variation inherrent in the articulators of speech (Rudzicz, 2012). Principal differential analysis optimizes the parameters of the following linear differential operator over a sample of functional data:

$$Ly(n) = w_0 y(n) + w_1 D^1 y(n) + \cdots + w_m D^m y(n), \tag{1}$$

where D^m denotes the mth derivative and $y(n)$ is an observed function. Specifically, m weighting functions w_j are optimized with point-wise minimization computed with standard least-squares estimation (Ramsay and Silverman, 2002; Ramsay, 1996). This model simplifies as $Ly_n = f_n$, $n = 1, \ldots, N$ where the forcing function f_n corresponds to the residual term in regression analysis and reflects variation that cannot be explained by the linear homogeneous differential equation $Ly_n = 0$. This approach has been successful in reducing bias in other types of quasi-stationary signals with additive Gaussian noise (Jin et al., 2013).

Relevant features are extracted and selected prior to classification, given the articulatory data segmented by phone and labelled with the class, as described in the following subsections.

2.1. Feature extraction

In this work, we use a second-order differential equation to capture the articulatory features in tract variables LA and LP and their derivatives, specifically:

$$LTV_n = w_0 TV_n + w_1 D^1 TV_n + w_2 D^2 TV_n, \quad n = 1, \ldots, N. \tag{2}$$

Given J observations where j, $1 \cdots J$, is the jth observation, we assume that each of the bilabial and non-bilabial groups can be described by uniquely parameterized differential equations of the form:

$$-D^2 TV^j(n) = w_0^j(n) TV^j(n) + w_1^j(n) D^1 TV^j(n) + r^j(n), \quad n = 1, \ldots, N; \quad j = 1, \ldots, J, \tag{3}$$