Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 39 (2016) 108–128

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition☆

Abhay Prasad, Prasanta Kumar Ghosh *

Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India

Received 24 December 2014; received in revised form 26 March 2016; accepted 28 March 2016
Available online 6 April 2016

Abstract

We propose an information theoretic region selection algorithm from the real time magnetic resonance imaging (rtMRI) video frames for a broad phonetic class recognition task. Representations derived from these optimal regions are used as the articulatory features for recognition. A set of connected and arbitrary shaped regions are selected such that the articulatory features computed from such regions provide maximal information about the broad phonetic classes. We also propose a tree-structured greedy region splitting algorithm to further segment these regions so that articulatory features from these split regions enhance the information about the phonetic classes. We find that some of the proposed articulatory features correlate well with the articulatory gestures from the Articulatory Phonology theory of speech production. Broad phonetic class recognition experiment using four rtMRI subjects reveals that the recognition accuracy with optimal split regions is, on average, higher than that using only acoustic features. Combining acoustic and articulatory features further reduces the error-rate by ∼8.25% (relative).

Keywords: Mutual information; Phonetic recognition; Speech production; Region splitting

1. Introduction

The speech signal is encoded with both linguistic and para-linguistic information including speaker's characteristics, background noise condition, and emotional state of the speaker. Seeking representations from the speech signal has been one of the main challenges in speech research (Greenberg and Kingsbury, 1997; Paliwal, 1998). The nature of a representation could change depending on the type of task at hand. For example, a representation for speech recognition is expected to be invariant to recording environment, and speaker's characteristics (Greenberg and Kingsbury, 1997). On the other hand, for speaker recognition, a representation should mostly capture speaker specific information (Perez-Meana, 2007). Representations robust to noise and other channel distortions are also critical for reliable performance of a speech based system in different working conditions.

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

* Corresponding author. Tel.: +91 80 2293 2694; fax: +91 80 2360 0444.
E-mail address: prasantg@ee.iisc.ernet.in (P.K. Ghosh).

http://dx.doi.org/10.1016/j.csl.2016.03.003
0885-2308/

Speech is a time-varying signal with complex spectro-temporal characteristics. The temporal and spectral resolutions necessary to analyze different speech sounds often vary over a short period of time. Thus, most of the representations in the literature have been designed assuming quasi-stationarity of the speech signal within a short-time window. A number of representative features are available exploiting spectral-only, temporal-only, and joint spectro-temporal properties. Spectral features include Filter Bank Analysis (Potter et al., 1947; Nadeu et al., 2001), Mel Cepstrum Analysis (Davis and Mermelstein, 1980; Olli and Kari, 1998; Han et al., 2006), Linear Predictive Coding (LPC) (Itakura, 1975), Perceptually Based Linear Predictive Analysis (PLP) (Hermansky, 1990), and spectral subband centroids (Paliwal, 1998). Temporal features such as temporal envelope cues (Shannon et al., 1995), short-time energy and fundamental frequency are used as features for speech recognition. Benefits of joint spectro-temporal features for speech recognition have also been demonstrated (Kleinschmidt, 2003; Xavier et al., 2008).

In spite of several proposed speech signal-based features, obtaining a robust representation still remains a challenge. This has led researchers to look into obtaining features from other modalities that capture the speech production process directly. For example, Krishnamurthy and Childers (1986) have demonstrated the benefit of using representation from electroglottography (EGG) for classification of speech according to different voicing qualities. While EGG captures the glottal vibrations of the speaker, other modalities such as X-ray microbeam (Fujimura et al., 1973) and electromagnetic articulography (EMA) (Schönle et al., 1987) capture the movement of critical points on speech articulators in the vocal tract by placing sensors on the articulators. Representations derived from these sensor data in addition to the speech acoustic have been shown to be useful for speech recognition (Sun et al., 2000; Wrench and Richmond, 2000; Frankel and King, 2001; Ghosh and Narayanan, 2011; Markov et al., 2006). Unlike sensor tracking in EMA, ultrasound captures a complete spatial view of the tongue (Hueber et al., 2007b; Stone and Davis, 1995); representations derived from the ultrasound images have been shown to be effective for the continuous-speech recognition as well as speech synthesis (Hueber et al., 2007a; Denby and Stone, 2004). Features from electromyography (EMG) as well as electropalatography (EPG) have also been shown to be effective for speech recognition (Jou et al., 2006; Manabe et al., 2003; Jorgensen et al., 2003; Schultz and Wand, 2010; Soquet et al., 1999); these signals capture the muscle movements in the face while speaking. Representations from EMA and EPG data together have been shown to improve the recognition accuracy when used jointly with spectral features (Wrench and Richmond, 2000).

Unlike ultrasound, in real time magnetic resonance imaging (rtMRI) the complete upper airway of a subject is imaged including the subject's nose and upper palate in addition to the vocal tract region starting from the lips to the glottis (Narayanan et al., 2011). The rtMRI video frames also have regions that are outside the subject's face in the midsagittal plane. Thus, the rtMRI video directly captures time varying dynamics of the changes in the vocal tract shape. Since the vocal tract shapes cause the production of different sounds, a representation from such vocal tract images would be robust to noise present in the speech signal. Representation from the rtMRI images could also be complementary to the spectro-temporal features derived from the speech signal since the relation between the vocal tract dynamics and speech acoustics is highly non-linear (Deng, 2006).

1.1. Research question

rtMRI captures the air-tissue boundaries along the entire vocal tract region from the glottis to the lips. A commonly used protocol for recording rtMRI video captures MR images at a frame rate of 23.18 frames/s with a resolution of 68×68 pixels with simultaneous audio recorded at 20 kHz. Hence the rtMRI has $\sim$106,352 dimensional data captured per second, which is more than that captured in audio (sampling rate of 20 kHz). Thus, the entire image may not be an efficient representation of the corresponding sound due to its large dimensionality. The key information related to the sound may lie only in a few pixels or regions in the image. In this work, we address the problem of automatically obtaining optimal regions from the rtMRI images such that the articulatory features derived from these optimal regions best represent the respective broad phonetic classes. Features derived from the optimal regions in the rtMRI images could potentially remove the redundancy in representing various sounds.

1.2. Related works

Several approaches have been used to obtain optimal features from rtMRI. The most commonly used approach is to compute the mean pixel intensity in a defined region of interest (mostly, the region of the articulators) across all the recorded rtMRI frames; this captures the temporal change in pixel intensities of the defined regions. The defined