Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 330–346

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories☆

Vikram Ramanarayanan*, Maarten Van Segbroeck, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, United States

Received 27 June 2014; received in revised form 2 March 2015; accepted 15 March 2015
Available online 21 March 2015

Abstract

How the speech production and perception systems evolved in humans still remains a mystery today. Previous research suggests that human auditory systems are able, and have possibly evolved, to preserve maximal information about the speaker's articulatory gestures. This paper attempts an initial step toward answering the *complementary* question of whether speakers' articulatory mechanisms have also evolved to produce sounds that can be optimally discriminated by the listener's auditory system. To this end we explicitly model, using computational methods, the extent to which derived representations of "primitive movements" of speech articulation can be used to discriminate between broad phone categories. We extract *interpretable* spatio-temporal primitive movements as recurring patterns in a data matrix of human speech articulation, i.e., representing the trajectories of vocal tract articulators over time. To this end, we propose a weakly-supervised learning method that attempts to find a part-based representation of the data in terms of recurring basis trajectory units (or primitives) and their corresponding activations over time. For each phone interval, we then derive a feature representation that captures the co-occurrences between the activations of the various bases over different time-lags. We show that this feature, derived entirely from activations of these primitive movements, is able to achieve a greater discrimination relative to using conventional features on an interval-based phone classification task. We discuss the implications of these findings in furthering our understanding of speech signal representations and the links between speech production and perception systems.

Keywords: Speech communication; Movement primitives; Phone classification; Motor theory; Information transfer

1. Introduction

Consider an information-based model of speech communication where the aim is to optimally and robustly convey a piece of information from speaker to listener. Scientists are still unclear about how the speech communication system has evolved in humans to achieve this task. One possibility is that the human auditory system has evolved to perceive speech produced by talkers, while another is that speakers' articulatory mechanisms have evolved to produce sounds that can be perceived by the listener's auditory system. A more likely possibility is that these systems have evolved

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author. Tel.: +1 2137403477.
E-mail address: vikram.ramanarayanan@gmail.com (V. Ramanarayanan).

http://dx.doi.org/10.1016/j.csl.2015.03.004
0885-2308/

together, the development of each bootstrapped by the other. This is because speech articulation is not the only action that can be produced by the human vocal organs, and likewise speech is not the only class of sounds that can be perceived by the auditory system. The human speech production system can perform actions other than those required for producing speech sounds (swallowing, chewing, etc.), while the auditory system can perceive natural sounds in the 20–20000 Hz range, including those that have distinct spectro-temporal characteristics not found in human speech. If we assume that the speech production and perception systems co-evolved to jointly optimize their (information encoding/decoding) performance with respect to each other, among other criteria, then this supposition would posit two broad predictions. First, the auditory system in listeners must process speech so as to preserve maximal information about the "intended" speech gestures of the speaker. Second, speakers must encode information – linguistic and/or paralinguistic – into speech gestures (and thereby speech) in such a manner that it can be robustly extracted by listeners.

With respect to the first prediction, researchers have presented evidence suggesting that the objects of speech perception are the intended gestures of the speaker, which could be represented, for instance, as invariant motor commands for linguistically significant movements (Liberman and Mattingly, 1985; Fowler and Galantucci, 2005). Though there is still debate among researchers regarding its validity, this theory, dubbed the Motor Theory of Speech Perception, is one popular theory that explicitly attempts to link speech production and perception. Smith and Lewicki (2006) found that the filterbank model of the cochlea has high coding efficiency for conveying maximal information to the brain for a wide range of natural sounds and, in particular, speech. It was in fact mathematically shown by Silva and Narayanan (2009) that a cochlear-like filterbank provides the Bayes optimal phonetic classification. Further, the research of Ghosh et al. (2011) and Bertrand et al. (2008) has shown that processing speech signals using an auditory cochlea-like filterbank preserves maximal mutual information between articulatory gestures and the processed speech signals. In other words, auditory filterbank-like transformations might improve speech perception/recognition performance because they maximize the articulatory information that speakers transmit. Hence there is some evidence in the literature in favor of the hypothesis that the human auditory system has evolved to maximally and robustly perceive information regarding the talker's speech gestures.

Now the second prediction posits that speech gestures must encapsulate information such that listeners can optimally perceive it. In particular, listeners must be able to derive categorical information regarding the underlying learned phonological structures (such as phonemes or syllables) of the language being spoken. This information must be discriminative such that these discrete constructs or categories can be teased apart from the continuous acoustic signal by listeners. Hitherto there has been little empirical evidence for such a claim for two reasons. For one, it was not until recently that major developments have been made in speech articulation measurement (see Ramanarayanan et al., 2012, for a review of recent developments in this field), which have allowed researchers to better explore hypotheses such as the two predictions mentioned above. Furthermore, speech gestures are theoretically defined in terms of abstract constriction-producing dynamical systems, and it is not clear how to extract these from speech articulation data in a principled manner. However, we recently showed qualitatively and quantitatively that one can robustly extract gesture-like movement primitives from speech articulation data using knowledge-informed machine learning techniques (Ramanarayanan et al., 2013). If these primitive representations are truly gesture-like and our hypothesis is true, they should contain discriminative information regarding underlying linguistic structure, such as phone categories. This leads us to the central question of this paper, that relates specifically on the second of the two predictions we presented earlier: *do directly data-derived "activation functions" of gesture-like movement primitives contain information to robustly discriminate between different phone categories?*

In addition to supporting scientific understanding, answering such a research question is important for speech technology applications such as automatic speech recognition; finding efficient representations is a key building block for such efforts (for a more detailed discussion, please see Ramanarayanan et al., 2012). Some reasons for this include: (i) improved noise robustness (Rose et al., 1996), (ii) better performance on spontaneous speech which exhibits a greater degree of coarticulation due to factored representations (Deng et al., 1997; Farnetani, 1997; Mcdermott and Nakamura, 2006), (iii) better modeling of different sources of variability, e.g., vocal tract morphology (Lammert et al., 2011), (iv) provision of a complementary view of the information captured by acoustic features alone (Arora and Livescu, 2013), and (v) the significantly lower-dimensional space of articulatory-based feature representations (Browman and Goldstein, 1995; King et al., 2007). To motivate the final argument in particular from a linguistics standpoint, Articulatory Phonology (Browman and Goldstein, 1995) theorizes that the act of speaking is decomposable into units of vocal tract action called "gestures" that are essentially low dimensional in nature, and suggests that lexical items are assembled from these dynamic primitive units, i.e., constriction actions of the vocal organs. Furthermore, Atal