Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 173–195

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Integrating articulatory data in deep neural network-based acoustic modeling☆

Leonardo Badino*, Claudia Canevari, Luciano Fadiga, Giorgio Metta

Istituto Italiano di Tecnologia, via Morego 30, 16163 Genova, Italy

Received 19 June 2014; received in revised form 6 May 2015; accepted 26 May 2015
Available online 4 June 2015

Abstract

Hybrid deep neural network–hidden Markov model (DNN-HMM) systems have become the state-of-the-art in automatic speech recognition. In this paper we experiment with DNN-HMM phone recognition systems that use measured articulatory information. Deep neural networks are both used to compute phone posterior probabilities and to perform acoustic-to-articulatory mapping (AAM). The AAM processes we propose are based on deep representations of the acoustic and the articulatory domains. Such representations allow to: (i) create different pre-training configurations of the DNNs that perform AAM; (ii) perform AAM on a transformed (through DNN autoencoders) articulatory feature (AF) space that captures strong statistical dependencies between articulators. Traditionally, neural networks that approximate the AAM are used to generate AFs that are appended to the observation vector of the speech recognition system. Here we also study a novel approach (AAM-based pretraining) where a DNN performing the AAM is instead used to pretrain the DNN that computes the phone posteriors. Evaluations on both the MOCHA-TIMIT msak0 and the mngu0 datasets show that: (i) the recovered AFs reduce phone error rate (PER) in both clean and noisy speech conditions, with a maximum 10.1% relative phone error reduction in clean speech conditions obtained when autoencoder-transformed AFs are used; (ii) AAM-based pretraining could be a viable strategy to exploit the available small articulatory datasets to improve acoustic models trained on large acoustic-only datasets.

Keywords: DNN-HMM; Acoustic-to-articulatory mapping; Deep neural networks; Acoustic modeling; Electromagnetic articulography; Autoencoders

1. Introduction

The steady increase of training data and computational resources combined with the use of new machine learning strategies for acoustic modeling has been continuously improving ASR performance in the last few years. Deep neural networks (DNNs) (Hinton et al., 2006), either combined with HMMs or used in a recurrent architecture, are the best strategy for acoustic modeling (Mohamed et al., 2012; Dahl et al., 2012; Graves et al., 2013).

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

* Corresponding author. Tel.: +39 010 71781975.
E-mail address: leonardo.badino@iit.it (L. Badino).

http://dx.doi.org/10.1016/j.csl.2015.05.005
0885-2308/

However, despite the impressive results shown by DNN-based ASR, there are several real usage scenarios where ASR technology still needs large improvements. In general, ASR accuracy significantly decreases in mismatched training-testing conditions, as it has been shown for traditional Gaussian mixture model (GMM)-HMMs systems in, e.g., speaking style mismatched conditions (Yu et al., 1999), and for DNN-HMM systems in, e.g., environment and microphone mismatched conditions (Seltzer et al., 2013).

Other than simply increasing the number of training conditions we can explicitly address the speech modeling limitations responsible for the lack of generalization underlying the mismatched conditions problem. For example, context-dependent (CD)-DNN-HMMs, as well as GMM-HMMs, handle context effects (like, e.g., coarticulation effects) using hundreds/thousands of tied context dependent sub-phonetic states, i.e., senones (Dahl et al., 2012). The selection, either automatic or manual, of the number of senones (and, consequently, of learning parameters) may be affected by the number of conditions in the training dataset and, at the same time, by the invariance of the input feature set to those conditions (see, e.g., (Schaaf and Metze, 2010) where the portion of gender-dependent senones depends on the feature set used).

The senones themselves result from the need to reduce learning parameters and are created by exploiting some speech production knowledge in the form of speech production-based questions in the state clustering tree. However ASR may benefit from a more explicit use of speech production knowledge where speech production can be used as, e.g., additional observations appended to the vector of acoustic observations, or as hidden structure connecting the phonological level (i.e., the HMM hidden phonetic states) to the observed speech acoustics.

Such approaches are motivated by the fact that complex phenomena observed in speech, for which a simple purely acoustic description has still to be found, can be easily and compactly described in speech production-based representations (notably Browman and Goldstein, 1992; Jakobson et al., 1952; Chomsky and Halle, 1968). For example, in Articulatory Phonology (Browman and Goldstein, 1992) or in the distinctive features framework (Jakobson et al., 1952; Chomsky and Halle, 1968) coarticulation effects can be compactly modeled as temporal overlaps of few vocal tract gestures. The vocal tract gestures are regarded as invariant, i.e., context- and speaker-independent, production targets that contribute to the realization of a phonetic segment. Obviously the invariance of a vocal tract gesture partly depends on the degree of abstraction of the representation but speech production representations offer compact descriptions of complex phenomena and of phonetic targets that purely acoustic representations are not able to provide yet (see, e.g., Maddieson, 1997).

Additional motivations to the use of speech production in ASR come from theories of speech perception such as the well known Motor Theory of speech perception (Liberman et al., 1967; Galantucci et al., 2006) which assumes that the perception of speech is the perception of motor gestures and involves access to the motor system. Such claims are partly supported by neurophysiological studies that show the contribution of the activity of the motor cortex to speech perception (DAusilio et al., 2009; Bartoli et al., 2013).

In the last two decades several strategies have been proposed for an explicit use of speech production knowledge in ASR (see King et al., 2007, for an extensive review). Here we review studies where measured articulatory data are used for ASR. Such studies require simultaneous recordings of audio and articulatory data. Articulatory movements are recorded using techniques such as electro-magnetic articulography (EMA) (Wrench, 2000), X-rays (Westbury, 1994), ultrasounds (e.g., Grimaldi et al., 2008), and MRI (Narayanan et al., 2004).

The approaches that use measured articulatory data can be roughly grouped into two categories. In the first category (e.g., Stephenson et al., 2000; Markov et al., 2006; Mitra et al., 2012) articulatory information is represented as discrete latent variables which are observed during training but hidden during testing. The idea behind this approach is to explicitly and compactly model speech production processes that are among the main causes of acoustic variability (e.g., variability due to coarticulation effects). In the second category (e.g., Zlokarnik, 1995; Wrench and Richmond, 2000), which the present work belongs to, articulatory features (AFs) are recovered from speech acoustics and then appended to the vector of observed acoustic features. In this case the working hypothesis is that the recovered articulatory domain (combined with the acoustic domain) represents a transformation of the acoustic domain into a new speech-production constrained domain which is more invariant over different conditions and where phonetic-articulatory targets can be more easily discriminated.

We first review some of the studies belonging to the first category. In Stephenson et al. (2000) the articulatory information is represented by a single discrete articulatory variable within a dynamic Bayesian network (DBN). Its values are computed by clustering data points in a space defined by eight articulator sagittal positions (upper lip, lower lip, four tongue positions, lower front teeth, lower back teeth). The acoustic observation probability distribution is