Speech Communication 89 (2017) 103–112

Contents lists available at ScienceDirect

Speech Communication

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

Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition

Vikramjit Mitra[a,*], Ganesh Sivaraman[b], Hosung Nam[c,d], Carol Espy-Wilson[b], Elliot Saltzman[c,e], Mark Tiede[c]

[a] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, United States
[b] Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, United States
[c] Haskins Laboratories, New Haven, CT, United States
[d] Department of English Language and Literature, Korea University, Seoul, South Korea
[e] Department of Physical Therapy and Athletic Training, Boston University, Boston, MA, United States

ARTICLE INFO

Article history:
Received 1 March 2016
Revised 31 January 2017
Accepted 29 March 2017
Available online 4 April 2017

Keywords:
Automatic speech recognition
Articulatory trajectories
Vocal tract variables
Hybrid convolutional neural networks
Time-frequency convolution
Convolutional neural networks

ABSTRACT

Studies have shown that articulatory information helps model speech variability and, consequently, improves speech recognition performance. But learning speaker-invariant articulatory models is challenging, as speaker-specific signatures in both the articulatory and acoustic space increase complexity of speech-to-articulatory mapping, which is already an ill-posed problem due to its inherent nonlinearity and non-unique nature. This work explores using deep neural networks (DNNs) and convolutional neural networks (CNNs) for mapping speech data into its corresponding articulatory space. Our speech-inversion results indicate that the CNN models perform better than their DNN counterparts. In addition, we use these inverse-models to generate articulatory information from speech for two separate speech recognition tasks: the WSJ1 and Aurora-4 continuous speech recognition tasks. This work proposes a hybrid convolutional neural network (HCNN), where two parallel layers are used to jointly model the acoustic and articulatory spaces, and the decisions from the parallel layers are fused at the output context-dependent (CD) state level. The acoustic model performs time-frequency convolution on filterbank-energy-level features, whereas the articulatory model performs time convolution on the articulatory features. The performance of the proposed architecture is compared to that of the CNN- and DNN-based systems using gammatone filterbank energies as acoustic features, and the results indicate that the HCNN-based model demonstrates lower word error rates compared to the CNN/DNN baseline systems.

1. Introduction

Spontaneous speech typically includes significant variability that is often difficult to model by automatic speech recognition (ASR) systems. Coarticulation and lenition are two sources of such variability (Daniloff and Hammarberg, 1973), and speech-articulation modeling can help to account for such variability (Stevens, 1960). Several studies in the literature (Kirchhoff, 1999; Frankel and King, 2001; Deng and Sun, 1994; Mitra et al., 2013a; Badino et al., 2016; and several others) have demonstrated that speech-production knowledge (in the form of speech articulatory representations) can improve ASR system performance by systematically accounting for variability such as coarticulation. A comprehensive exploration of speech-production features and their role in speech recognition performance is provided in King et al. (2007). Further studies (Richardson et al., 2003; Mitra et al., 2010a , 2011a) have demonstrated that articulatory representations provide some degree of noise robustness for ASR systems.

Deep learning techniques (Mohamed et al., 2012) involving neural networks with several hidden layers have become integral to current ASR systems. Deep learning has been used for feature representation (Hoshen et al., 2015), acoustic modeling (Seide et al., 2011), and language modeling (Arisoy et al., 2012). Given the versatility of the deep neural network (DNN) systems, it was observed in Mitra et al. (2014a) that speaker-normalization techniques such as vocal tract length normalization (VTLN) (Zhan and Waibel, 1997) are unnecessary for improving ASR accuracy, as the DNN architecture's rich, multiple projections through multiple hidden layers enable it to learn a speaker-invariant data representation. Convolutional neural networks (CNNs) (Sainath et al., 2013; Abdel-Hamid et al., 2012) motivated by the receptive field theory of the visual cortex are often found to outperform fully connected DNN architectures (Mitra et al., 2014a,b). CNNs are

* Corresponding author.
E-mail address: vikramjitmitra@gmail.com (V. Mitra).

http://dx.doi.org/10.1016/j.specom.2017.03.003
0167-6393/

known to be noise robust (Mitra et al., 2014a), especially in those cases where noise/distortion is localized in the spectrum. Speaker-normalization techniques are also found to have less impact on speech recognition accuracy for CNNs as compared to for DNNs (Mitra et al., 2014a). With CNNs, the localized convolution filters across frequency tend to normalize the spectral variations in speech arising from vocal tract length differences, enabling the CNNs to learn speaker-invariant data representations.

Studies have explored using DNNs (Mitra et al., 2010b, c; Uria et al., 2011; Canevari et al., 2013) for learning the nonlinear inverse transform of acoustic waveforms to articulatory trajectories (a.k.a. speech-inversion or acoustic-to-articulatory inversion of speech). Results have demonstrated that using articulatory representations in addition to acoustic features improves phone recognition (Badino et al., 2016; Canevari et al., 2013; Mitra et al., 2011a; Deng and Sun, 1994) and speech recognition performance (Mitra et al., 2011b , 2013a , 2014c). Learning acoustic-to-articulatory transforms is quite challenging, as such mapping is nonlinear and non-unique (Mitra, 2010; Richmond, 2001). Speaker variation adds complexity to the problem and makes speech-inversion even harder (Sivaraman et al., 2015; Ghosh and Narayanan, 2011). Ghosh and Narayanan (2011) demonstrated that speaker-independent speech-inversion systems can be trained and can offer similar performance to that of speaker-dependent speech-inversion systems.

This work has two principal aims:

(1) Explore deep learning approaches to learn "speaker-independent" speech-inversion mappings using synthetically generated parallel articulatory speech data;
(2) Create a suitable DNN architecture, where retrieved articulatory variables can be used for spontaneous speech recognition purposes.

Note that, this work to the best of our knowledge, for the first time explores the use of Vocal tract constriction variables (TVs) and filterbank features as input to DNN/CNN acoustic models. Hence we explore DNN/CNN configuration to best utilize the TV+filterbank features.

Specifically, we explore CNNs with the hope of achieving more robust speech-inversion performance compared to the traditional DNNs. We investigate using different parameters (such as neural network size and data contextualization (splicing) windows over the input acoustic-feature space) and show that impressive gains can be achieved through careful selection of parameters. We use the trained DNN/CNN models to predict the articulatory trajectories of the training and testing datasets of the *Wall Street Journal* (WSJ1) corpus, the noisy WSJ0 corpus Aurora-4, and the 265-h Switchboard-1 corpus, and we use the estimated trajectories to perform a continuous speech recognition task.

Further, we present a parallel CNN architecture, where time-frequency convolution is performed on traditional gammatone-filterbank-energy-based acoustic features, and time-convolution is performed on the articulatory trajectories. Each of the parallel convolution layers is followed by a fully connected DNN, whose outputs are combined at the context-dependent (CD) state level, producing senone posteriors. The proposed hybrid CNN (HCNN) architecture learns an acoustic space and an articulatory space, and uses both to predict the CD states.

The word recognition results from both the WSJ1 and Aurora-4 speech recognition tasks indicate that using articulatory information in addition to filterbank features provides sufficient complementary information to reduce the word error rates (WER) in clean, noisy, and channel-degraded conditions.

The novelties of this work are as follows:

Table 1
Constrictors and their vocal tract variables.

Constrictors	Vocal tract (VT) variables
Lip	Lip aperture (LA)
	Lip protrusion (LP)
Tongue tip	Tongue tip constriction degree (TTCD)
	Tongue tip constriction location (TTCL)
Tongue body	Tongue body constriction degree (TBCD)
	Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

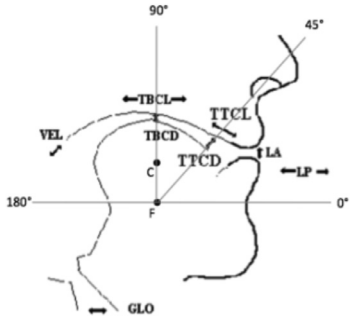

Fig. 1. Vocal tract variables at five distinct constriction organs.

(a) Use of large multi-speaker synthetic articulatory data to train robust speech-inversion models. Earlier work has used single-speaker model-based synthetic articulatory data (Mitra et al., 2014c). In this work, we simulate a diverse set of speakers, by varying vocal tract length, pitch, articulatory weights, etc.
(b) Investigate the effect of noise on speech-inversion performance, by comparing CNN and DNN models to analyze their robustness in noisy acoustic conditions.
(c) Explore joint-modeling of articulatory and acoustic spaces in speech recognition tasks. We show that the proposed HCNN better leverages articulatory information compared to simply combining articulatory and acoustic features and training one DNN or CNN acoustic model. In addition, we also explore fusing convolutional-layer feature maps generated from articulatory and acoustic features, and we found that such an approach yields promising results.

Overall, this study demonstrates that given articulatory trajectories and acoustic features are quite different in terms of the information they contain, it is useful to learn intermediate feature spaces from DNN hidden layers or CNN feature maps to fuse the information, rather than fusing the features on the input side and feeding the combined features to a single DNN or CNN acoustic model.

2. Vocal tract constriction variables

Articulatory Phonology (Browman and Goldstein, 1989, 1992) is a phonological theory that views speech as a constellation of articulatory gestures that can overlap in time. Gestures are defined as discrete action units whose activation results in constriction formation or release by five distinct constrictors (lips, tongue tip, tongue body, velum, and glottis) along the vocal tract. The kinematic state of each constrictor is defined by its corresponding constriction degree and location coordinates, which are called vocal tract constriction variables (the time-function output of these variables are typically identified as tract-variable trajectories, in short TVs). Please refer to Table 1 and Fig. 1 for more details regarding TVs. Table 2 presents the dynamic range and the measuring units for each TVs.