Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 260–273

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Data driven articulatory synthesis with deep neural networks☆

Sandesh Aryal, Ricardo Gutierrez-Osuna*

Department of Computer Science and Engineering, Texas A&M University, United States

Received 2 July 2014; received in revised form 4 December 2014; accepted 24 February 2015
Available online 5 March 2015

Abstract

The conventional approach for data-driven articulatory synthesis consists of modeling the joint acoustic-articulatory distribution with a Gaussian mixture model (GMM), followed by a post-processing step that optimizes the resulting acoustic trajectories. This final step can significantly improve the accuracy of the GMM frame-by-frame mapping but is computationally intensive and requires that the entire utterance be synthesized beforehand, making it unsuited for real-time synthesis. To address this issue, we present a deep neural network (DNN) articulatory synthesizer that uses a tapped-delay input line, allowing the model to capture context information in the articulatory trajectory without the need for post-processing. We characterize the DNN as a function of the context size and number of hidden layers, and compare it against two GMM articulatory synthesizers, a baseline model that performs a simple frame-by-frame mapping, and a second model that also performs trajectory optimization. Our results show that a DNN with a 60-ms context window and two 512-neuron hidden layers can synthesize speech at four times the frame rate – comparable to frame-by-frame mappings, while improving the accuracy of trajectory optimization (a 9.8% reduction in Mel Cepstral distortion). Subjective evaluation through pairwise listening tests also shows a strong preference toward the DNN articulatory synthesizer when compared to GMM trajectory optimization.

Keywords: Articulatory synthesis; Electromagnetic articulography; Deep learning; Gaussian mixture models

1. Introduction

Certain speech modifications, such as changes in foreign/regional accents and articulatory styles are too difficult to perform in the acoustic domain, where voice quality information and speaking style interact in complex ways (Hermansky and Broad, 1989). By contrast, these two sources are readily decoupled in the articulatory domain through the position and dynamics of articulators (e.g., as measured via electromagnetic articulography). Following this intuition, in previous work (Felps et al., 2012; Aryal and Gutierrez-Osuna, 2014) we have shown how accent conversion can be performed by driving an articulatory synthesizer of a non-native speaker with articulatory gestures from a native speaker.

Our original articulatory-based method for accent conversion (Felps et al., 2012) was based on unit-selection synthesis. Given a corpus of articulatory-acoustic frames for a second-language (L2) speaker and a reference first-language

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

* Corresponding author. Tel.: +1 9798452942.
E-mail addresses: sandesh@cse.tamu.edu (S. Aryal), rgutier@cse.tamu.edu (R. Gutierrez-Osuna).

http://dx.doi.org/10.1016/j.csl.2015.02.003
0885-2308/

(L1) utterance, the approach consisted of finding a sequence of L2 diphones with similar articulatory configurations as those in the reference L1 utterance. Unfortunately, the approach provided only modest improvements in accent reduction due to the small size of the L2 acoustic-articulatory corpus and the fact that unit-selection cannot produce sounds that do not already exist in the L2 corpus. For these reasons, our recent work (Aryal and Gutierrez-Osuna, 2014) has focused on parametric statistical techniques proposed by Toda et al. (2004). The approach uses a Gaussian mixture to model the joint acoustic-articulatory distribution, followed by an optimization stage to find the maximum-likelihood acoustic trajectory for an articulatory sequence. This GMM framework is better suited for the limited size of articulatory-acoustic corpora and can interpolate sounds that do not exist in the L2 inventory. However, it is computationally intensive and impractical for real-time synthesis since the trajectory-optimization stage requires that the entire utterance be present at synthesis time.

To address the limitations of the above GMM framework for real-time articulatory synthesis, this paper explores the use of deep neural networks (DNN) to perform the articulatory-to-acoustic mapping. Following prior work on DNN for speech recognition, articulatory inversion and text-to-speech synthesis (Hinton et al., 2012; Uria et al., 2012; Zen et al., 2013), our approach uses a tapped-delay line to contextualize features by forming an input vector of short-term articulatory sequences from which to predict acoustic observations. Because temporal information is encoded in the tapped-delay line, the resulting acoustic sequence does not require the costly trajectory optimization of GMM articulatory synthesis. We compare the proposed DNN against two GMM implementations, a static GMM that performs a frame-by-frame mapping from articulators to acoustics, and a dynamic GMM that uses trajectory optimization. Our results show that the DNN provides a more accurate mapping, measured as Mel cepstral distortion, than either GMM implementation. A second comparison between the DNN and the GMM with different tapped-delay lengths shows that DNN accuracy increases monotonically for contexts of up to 60 ms, whereas GMM accuracy degrades drastically for contexts larger than 20 ms. Compared to trajectory optimization, which requires an average of 39 s of synthesis time per second of speech, the DNN requires only 267 ms per second of speech, making it suitable for real-time synthesis. A final subjective assessment through pairwise listening tests shows a strong preference (73%) toward the DNN synthesizer.

The remaining sections of this paper are organized as follows. Section 2 reviews related work in articulatory-to-acoustic mappings. Section 3 presents the proposed DNN architecture and the two GMM synthesizers used for comparison. Section 4 describes the experimental setup used for model evaluation, followed by experimental results in Section 5. The article concludes with a discussion of these results and directions for future work.

2. Related work

2.1. Articulatory-to-acoustic mappings

A significant amount of research has been performed toward understanding the forward mapping from articulators to acoustics and developing methods to build such mappings. These efforts can be grouped into two broad categories: physics-based models and data-driven models. Physics-based models approximate vocal tract geometry using a stack of cylindrical tubes with different cross section areas. Speech waveforms are then generated by solving the wave propagation equation in the approximated tube model (Mermelstein, 1973; Browman et al., 1984; Maeda, 1990; Birkholz et al., 2006). In contrast, data-driven approaches use machine learning techniques to build a forward mapping from simultaneous recordings of articulators and acoustics; it is this latter group of models that our review will focus on.

In one of the earliest studies, Kaburagi and Honda (1998) proposed a codebook algorithm for data-driven articulatory synthesis. Given a target articulatory frame, its acoustic observation was estimated by finding the closest articulatory frames in the corpus, and then computing a weighted average of their acoustic observations. Unfortunately, the nearest-neighbors search makes the synthesis process computationally expensive, and synthesis quality is limited by the relatively small size of articulatory corpora. Over the past decade, codebook techniques have been replaced with parametric models (Hiroya and Honda, 2004; Toda et al., 2004; Nakamura et al., 2006), which are better suited when the size of the corpus is limited. In a landmark study, Toda et al. (2004) used Gaussian mixture models (GMM) to learn the joint distribution of articulatory and acoustic parameters. Given a trained GMM and a vector of articulatory parameters, estimating the corresponding acoustic parameters in a frame-by-frame fashion involves a fixed number of operations independent of the database size.