Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 165–172

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Speech Production in Speech Technologies: Introduction to the CSL Special Issue

Abstract

Aspects of speech production have provided inspiration for ideas in speech technologies throughout the history of speech processing research. This special issue was inspired by the 2013Workshop on Speech Production in Automatic Speech Recognition in Lyon, France, and this introduction provides an overview of the included papers in the context of the current research landscape.
© 2015 Published by Elsevier Ltd.

Keywords: Speech production; Speech recognition; Speech synthesis; Articulatory inversion; Articulatory features; Pronunciation modeling; Acoustic modeling; Acoustic features

1. Introduction

Speech production has inspired ideas in speech technologies throughout the last few decades of speech processing research (King et al., 2007; Ostendorf, 1999; Rose et al., 1996). For speech recognition, the motivations include potential robustness to noise (Eide, 2001; Kirchhoff et al., 2002; Mitra et al., 2011; Richardson et al., 2000) and pronunciation variation (Deng et al., 1997; Hasegawa-Johnson et al., 2005; Livescu, 2004; Mitra et al., 2011); multilinguality and language portability (Çetin et al., 2007; Lal and King, 2013; Siniscalchi et al., 2012; Stüker, 2003), and more generally better use of limited training data (King et al., 2007); better modeling of non-acoustic speech signals such as video (Cohen and Massaro, 1993; Hasegawa-Johnson et al., 2007; Saenko et al., 2009) and surface electromyography (Jorgensen and Dusan, 2010; Jou et al., 2006); and more effective processing of disordered speech (Rudzicz, 2011; Rudzicz, 2012). For speech synthesis, motivations include some of the same considerations, but also the potential for increased naturalness due to better modeling of coarticulation, as well as for increased flexibility in modifying speaker characteristics or affect (Black et al., 2012). Articulatory synthesis has also been used in analysis-by-synthesis approaches to recognition (Al Bawab et al., 2008; Blackburn and Young, 1995). Additional applications include speech technologies based on alternative input modalities, such as video, ultrasound, or neural measurements, which may be more directly connected to speech articulation (Bocquelet et al., 2014; Guenther and Brumberg, 2011; Hueber et al., 2010; Mesgarani et al., 2014). All of these applications have motivated work on certain sub-problems, including acoustic-to-articulatory inversion and articulatory feature classification, and extensive collection and analysis of articulatory data.

This special issue was inspired by the 2013 Workshop on Speech Production in Automatic Speech Recognition (SPASR), which was held as a satellite workshop of the Interspeech conference in Lyon, France. The goal of the workshop was to bring together researchers in this area, including both speech recognition and other technologies, to share ideas, results, and perspectives that can advance the field. The SPASR technical program included 5 invited

http://dx.doi.org/10.1016/j.csl.2015.11.002
0885-2308/© 2015 Published by Elsevier Ltd.

speakers, as well as spotlight talks and a poster session for the 11 contributed papers. The topics of the invited talks included new opportunities in the collection and use of diverse types of articulatory data, by Shri Narayanan; invariance in articulatory gestures, by Carol Espy-Wilson; models of dysarthric speech, by Frank Rudzicz; new articulatory data collections and acoustic-to-articulatory inversion, by Korin Richmond; and silent speech interfaces, by Bruce Denby.

The contributed papers and abstracts presented new work and reviews of work on related topics including computational models of infant language learning (Rasilo et al., 2013), human-machine comparisons in recognition errors (Juneja and Hasegawa-Johnson, 2013), estimation of articulatory parameters and primitives from acoustic and articulatory data (Kim et al., 2013; Li et al., 2013; Ramanarayanan et al., 2013; Schuppler et al., 2013), the use of articulatory classification/inversion in speech recognition (Canevari et al., 2013; Magimai-Doss and Rasipuram, 2013), silent speech interfaces (Freitas et al., 2013), the use of articulatory data for learning improved acoustic features (Andrew et al., 2013), and discriminative recognition with latent articulatory variables (Fosler-Lussier et al., 2013).

In this editorial introduction to the special issue, we will not attempt to review the entire history of speech production-based methods in speech technology. Earlier papers provide a thorough survey of the area (King et al., 2007; Rose et al., 1996). However, there have been certain recent developments that we wish to highlight, both within and outside the special issue.

The remaining sections provide an overview of the included papers in the context of the current research landscape. While some of the questions and approaches being addressed here are the same as those considered throughout the history of research in speech production, there have been advances in data collection and in machine learning techniques that have significantly advanced the state of the art.

2. Articulatory inversion-based models

Articulatory inversion has long formed a part of production-based approaches to automatic speech recognition (ASR). Since it was found that appending articulatory measurements to acoustic features can improve speech recognition (Wrench, 2000b; Zlokarnik, 1995), it seemed natural to attempt to predict articulation from the acoustics and append the predicted (inverted) articulatory measurements in the absence of measured ones. However, this approach is not straightforward since the articulatory inversion problem is challenging, and it has been quite difficult to obtain any improvements in ASR using this idea (Wrench, 2000b). One of the exciting developments of the last few years has been the marked improvement in performance of articulatory inversion models, based largely on the use of deep networks. As a result, a number of groups have now been able to significantly improve ASR performance using deep network-based articulatory inversion (Badino et al., 2016; Uria et al., 2012; Wang et al., 2015). In addition, recent work has given more attention to the use of articulatory inversion, or inversion-like models, across speakers (Arora and Livescu, 2013; Badino et al., 2016; Li et al., 2016; Wang et al., 2015).

In this special issue, Badino et al. present one such approach, where the output of a deep neural network (DNN) articulatory inversion model is appended to acoustic features in a hybrid hidden Markov model-deep neural network (HMM-DNN) recognizer, for improved phonetic recognition performance (Badino et al., 2016). More mixed results are obtained when using the inversion DNN to pretrain the recognizer's DNN, or when training and testing the inversion model on different speakers. An interesting result from this paper is that it is preferable not to predict articulatory measurements directly, but to first learn a transformation of the articulatory space and then predict articulatory representations in that space. Specifically, Badino et al. train an auto encoder to transform articulatory measurements into a lower dimensional space—intuitively encoding latent gestures that couple multiple measured articulatory dimensions—and use the resulting transformed articulatory vectors as the targets for articulatory inversion. Others have also made this observation, using canonical correlation analysis and its deep extension to simultaneously learn embeddings of the acoustics and articulation (Arora and Livescu, 2013; Wang et al., 2015).

Also in this special issue, Li et al. study the use of articulatory inversion for speaker verification (Li et al., 2016). The authors predict articulatory parameters—again in a derived space of articulatory parameters rather than the original measurement dimensions—using either a DNN or a method based on the generalized smoothness criterion (GSC) (Ghosh and Narayanan, 2010), and combine the predictions with standard acoustic features with a variety of feature and score fusion techniques.

Like Badino et al., Li et al. consider the use of different speakers for training the inversion models and for testing the verification systems. Overall, Li et al. find encouraging improvements in speaker verification using either GSC or

Download English Version:

https://daneshyari.com/en/article/558205

Download Persian Version:

https://daneshyari.com/article/558205

Daneshyari.com