Articulatory features to address acoustic variability in Automatic Speech Recognition

Ganesh Sivaraman
Advisor: Prof. Carol Espy-Wilson

[1] University of Maryland College Park, MD
{ganesa90, espy}@umd.edu

Abstract

One of the main reasons why the performance of Automatic Speech Recognition (ASR) systems lag behind the human speech recognition performance is due to their lack of robustness against speech variability. The speech signal is extremely variable due to a large number of factors like speaker voice, speaking style, speaking rate, accents, casualness, emotions etc. The goals of this thesis are (i) to investigate the variability of speech from the perspective of speech production, (ii) put forth robust articulatory features to address this variability, and (iii) to incorporate the articulatory features in state-of-the-art ASR systems in the best way possible. The tasks towards achieving these goals include articulatory data collection, acoustic-to-articulatory speech inversion and the development of an articulatory feature based ASR system. This paper outlines the different objectives of my PhD thesis.

Index Terms: Acoustic to articulatory speech inversion, speech variability, Electromagnetic Articulometry, Articulatory phonology, Automatic Speech Recognition

1. Introduction

Current ASR systems perform poorly in conversational speech as due to the increase in acoustic variability as compared to clearly articulated speech. The reason for this is the limitation of the phone based acoustic modeling that is most commonly adopted in ASR systems. Although phonemes are distinctive units in the cognitive domain, their physical realizations are extremely variable due to coarticulation and lenition which are commonly observed in conversational speech. The traditional approaches deal with this issue by performing di-phone or tri-phone based acoustic modeling which often suffer from data scarcity (enough training data for each tri-phone) and are insufficient to model longer contextual dependencies. Analyzing speech in from the perspective of speech production provides a clear explanation for speech variability as compared to traditional phone based analysis. Articulatory phonology [1] analyzes speech as a constellation of coordinated articulatory gestures performed by the articulators in the vocal tract (lips, tongue tip, tongue body, jaw, glottis and velum). According to this theory, the acoustic variability can be explained by the temporal overlap of gestures and their reduction in space [2]. Coarticulation and lenition are due to the overlap of neighboring gestures. The persistence property of Articulatory gestures makes them a promising approach to better model the variability of speech. However, a major challenge to exploring this approach is to obtain a good groundtruth of these gestures for real speech utterances.

Previous approaches [3] have used the Task Dynamics and Application (TADA) model [4] which is a synthetic model of speech production to learn the mapping from speech acoustics to articulatory gestures using synthetic speech. However, the synthetic speech and the synthetic articulatory trajectories do not represent the amount of variability observed in real speech. This makes the TADA model based approach insufficient to analyze real speech variability. In this thesis I propose to explore methods of estimating articulatory gestures from real speech and articulatory data and developing a gesture based ASR system.

The first part of this thesis deals with the reliable speaker independent estimation of articulatory features from speech. Most previous studies [5][6] have focused on developing speech inversion systems for one or two speakers due to the vast acoustic and articulatory differences between speakers. The present study aims to develop a robust speech inversion system using data from multiple speakers. To overcome the speaker differences, a speaker normalization scheme has been developed to address the speaker variability in the acoustic and articulatory domains. More details of the speech inversion systems are given in section 2. The performance of the speech inversion system is analyzed on an articulatory dataset containing speech rate variations to see if the model is able to reliably predict the TVs in challenging coarticulatory scenario [7]. We propose to perform a deeper analysis of rate specific variations in the acoustic and articulatory domains and make our speech inversion system robust to speech rate differences.

The second part of the thesis deals with estimating discrete articulatory gestures from continuous time TV trajectories and the acoustic signal. This effort will enable us to annotate large speech datasets with articulatory gestures for ASR training. Section 3 details the proposed methods to estimate articulatory gestures from speech.

The third part of the thesis deals with the development of articulatory gesture based ASR systems. This thesis aims to explore different Deep Neural Network (DNN) based ASR architectures and develop the best method to combine the acoustic and articulatory features. The advantage postulated by articulatory features is their robustness against speech variability. In order to test this hypothesis, a cross-corpus speech recognition evaluation will be performed.

2. Acoustic-to-articulatory speech inversion

The speech inversion systems are trained using data obtained from the U. Wisconsin X-ray Microbeam (XRMB) dataset. The articulatory pellet trajectories from the XRMB dataset are

converted into vocal tract constriction variables (also known as Tract Variables (TVs)) [8]. Feedforward neural networks are trained to map acoustic features (contextualized MFCCs) to six TVs. The outputs were smoothed using a Kalman smoothing technique to obtain smooth TV estimates. Figure 1 shows the block diagram of our speech inversion system.

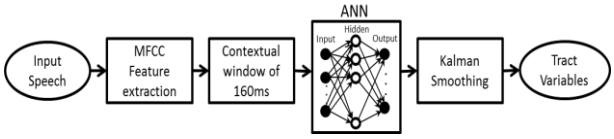

Figure 1: Block diagram of speech inversion system

The XRMB dataset consisting of 46 different speakers poses a great challenge to the robustness of the speech inversion system due to the vast interspeaker differences. The neural network based speech inversion system achieves an average correlation of 0.7 between the actual and estimated TVs. A Vocal Tract Length Normalization based speaker adaptation technique has been developed [9] which provides a 12% relative improvement in correlation over a speaker independent system.

3. Estimating articulatory gestures

In this thesis we will explore the estimation of articulatory gestures as defined by the TADA model. Although, the TADA model provides gestures, TVs and corresponding synthetic speech, there is no such definition for articulatory gestures in real articulatory data. This thesis proposes to explore methods for annotation of real speech with articulatory gestures.

The Mview [10] toolkit from Haskins laboratories defines gestures from the kinematics of articulatory trajectories. According to this software, gestural targets are achieved at minimal velocity of articulatory trajectories. The gestural onset and offset are defined based on velocity thresholds around the velocity minimum. In this thesis we borrow this definition of articulatory gestures and propose to develop an automatic gesture estimation system from real speech. The acoustic signal and the estimated articulatory trajectories from speech inversion systems will be used to infer the gestural patterns. Using this gesture estimation system, we plan to annotate a large speech recognition corpus with articulatory gestures.

4. Gesture based ASR architectures

The main reason for estimating articulatory gestures is to explore their robustness against speech variability. In spite of the challenges facing the state-of-the-art ASR systems, they have achieved very high performance in all domains of speech – from connected digits and read speech to the highly variable conversational speech. The state-of-the-art ASR systems rely on segmental units of speech known as tri-phones. Replacing the tri-phones with articulatory gestures will address the problem of speech variability in ASR in the following ways –

(a) Map the variabilities observed in the acoustic features to invariant articulatory gestures.

(b) Model coarticulation effects over a larger context as compared to triphones.

(c) Reduce the problem of training data insufficiency faced by tri-phones because the number of articulatory gestures is far less than the number of tri-phones.

However, going from the segmental tri-phone based models to the articulatory gesture based models, renders the variability in the form of gesture durations and inter-gestural overlaps. The ASR architectures need to be modified to suit the articulatory gestures taking into consideration their durational information. This thesis will explore various architectures for incorporating gestures in the DNN based ASR systems. In summary, Figure 2 shows the articulatory gesture based ASR architectures that work proposes to explore.

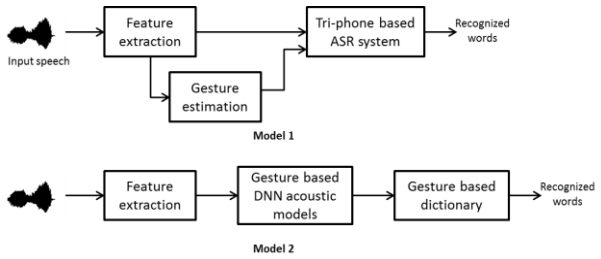

Figure 2: Gesture based ASR architectures proposed

5. Summary and Future directions

This thesis proposes to explore the challenging problem of articulatory gesture based ASR. Although articulatory features have been explored for speech recognition, articulatory gestures from real speech data have never been explored for ASR. The thesis also addresses the complex problem of speaker independent acoustic-to-articulatory speech inversion by combining articulatory data from several speakers. The vocal tract length normalization based speaker adaptation for speech inversion is also a novel approach towards improving the speaker independent speech inversion performance.

This thesis raises an important question as to whether speech can be analyzed as a series of overlapping articulatory gestures instead of segmental phoneme based approach. There has never been any effort to create a language dictionary based on articulatory gestures that are not derived directly from phoneme identities. We believe that the outcomes of this thesis can lead to approaches that can define a gesture based dictionary. In future, this gesture based approach to ASR can be explored to develop multilingual ASR systems.

6. Acknowledgements

This research was supported by NSF Grant # IIS-1162046.

7. References

[1] C. P. Browman and L. Goldstein, “Articulatory Phonology : An Overview *,” *Phonetica*, vol. 49, pp. 155–180, 1992.

[2] E. L. Saltzman and K. G. Munhall, “A Dynamical Approach to Gestural Patterning in Speech Production,” *Ecol. Psychol.*, vol. 1, no. 4, pp. 333–382, Dec. 1989.

[3] V. Mitra, “Articulatory Information For Robust Speech Recognition.” Ph.D. dissertation, University of Maryland, College Park, 2010.

[4] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, “TADA: An enhanced, portable Task Dynamics model in MATLAB,” *J. Acoust. Soc. Am.*, vol. 115, no. 5, p. 2430, May 2004.

[5] P. K. Ghosh and S. Narayanan, “A generalized smoothness criterion for acoustic-to-articulatory inversion.,” *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2162–72, Oct. 2010.

[6] B. Uria, S. Renals, and K. Richmond, “A Deep Neural Network for Acoustic-Articulatory Speech Inversion,”

2011.
[7] G. Sivaraman, V. Mitra, M. K. Tiede, E. Saltzman, L. Goldstein, and C. Y. Espy-Wilson, “Analysis of coarticulated speech using estimated articulatory trajectories,” in *INTERSPEECH*, 2015, pp. 369–373.
[8] H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, and L. Goldstein, “A procedure for estimating gestural scores from speech acoustics.,” *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3980–9, Dec. 2012.
[9] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, “Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion,” to appear in *Interspeech 2016*.
[10] M. Tiede, “MVIEW: software for visualization and analysis of concurrently recorded movement data.” Haskins Laboratories, New Haven, CT, 2005.