SILENT SPEECH RECOGNITION FROM ARTICULATORY MOTION

By

Jun Wang

A DISSERTATION

Presented to the Faculty of

The Graduate School at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professors Ashok Samal and Jordan R. Green

Lincoln, Nebraska

November, 2011

SILENT SPEECH RECOGNITION FROM ARTICULATORY MOTION

Jun Wang, Ph.D.

University of Nebraska, 2011

Advisors: Ashok Samal and Jordan R. Green

Silent speech recognition is the process of converting motion data of articulators (e.g., tongue, lips, and jaw) into speech in the form of text. The primary objective of this dissertation was to develop new approaches for silent speech recognition from segmented and continuous input of tongue and lip movement data at three levels of speech units with increasing conceptual complexity - phonemes, words, and sentences. At each level, unique theoretical issues were addressed and plans for use in specific applications were described. This dissertation is motivated by the need for (1) speech movement-based treatment options for people with speech and voice impairments and (2) computational approaches for recognizing speech when acoustic data are not available or extremely noisy.

Machine learning and statistical shape analysis were used to classify and quantify the articulatory distinctiveness of phonemes, words, and sentences. The approach is unique in that it maps the motion data directly (instead of articulatory features) to speech units. Procrustes analysis, a statistical shape matching approach, provided an index of articulatory distinctiveness of vowels and consonants, which was used to derive quantitative articulatory vowel and consonant spaces. The derived vowel space resembles

long-standing descriptions of articulatory vowel space. The theoretical properties and practical applications in speech pathology (e.g., motor speech decline in amyotrophic lateral sclerosis) of these spaces were also discussed. In addition, support vector machine, Procrustes analysis, and Eigenspace approaches were used to classify a set of phonetically balanced words and functional sentences from articulatory motion. The direct mapping approaches resulted high classification accuracy levels, which were adequate for practical applications.

A near-time algorithm (Holistic Articulatory Recognition, HAR) to recognize the whole words and sentences from continuous (unsegmented) articulatory motion was proposed and evaluated. The accuracy and speed of HAR demonstrated its potential for practical applications. HAR is based on classification probabilities and hence any classifier that could estimate them can be incorporated seamlessly. HAR can serve as the recognition component of an articulation-based silent speech interface that may provide an alternate oral communication modality for persons with speech impairments.

ACKNOWLEDGEMENTS

It is very unlikely to accomplish a PhD dissertation without guidance and assistance from other people. During my PhD program, many people have helped and encouraged me. Frist and foremost, I am very grateful to my advisors, Drs. Ashok Samal and Jordan R. Green, for their guidance and research support throughout my dissertation study. Particularly, I thank Dr. Ashok Samal for guiding my research and teaching me data mining; I thank Dr. Jordan R. Green for guiding my research and training me interdisciplinary on speech science and communication disorders, as well as his support for my travels to top national and international conferences. I also thank all other members in my dissertation advisory committee. Specifically, I thank Dr. Tom D. Carrell for reading my dissertation, helping me in subject recruitment, and teaching me speech perception and acoustic analysis. I thank Dr. David B. Marx for reading my dissertation and teaching me statistics. I thank Dr. Sharad C. Seth for reading my dissertation. I thank Dr. Lisong Xu for reading my dissertation and teaching me automata theory.

I thank Dr. Ying Lu, Dr. Don Costello, Dr. Alvin Surkan, whom I worked with as a teaching assistant. I thank Dr. Jitender Deogun, Dr. Myra Cohen, Dr. Matthew Dwyer, Dr. Gregg Rothermel, Dr. Sebastian Elbaum, Dr. Steve Goddard, Dr. Hong Jiang, and Dr. Scott Stephen for teaching me other computer science courses.

I thank Dr. Tiffany Hogan, Dr. Yana Yunusova (University of Toronto, Canada), Dr. Frank Rudzicz (University of Toronto, Canada), and Dr. Tony Wilson (University of Nebraska Medical Center). I really enjoy the collaborations with them. I thank Dr. David

R. Beukelman, Dr. Charles E. Healey and other faculties in Barkley Center. I enjoy talking to them and thanks for their suggestions in my research. I also thank Dr. Michael H. Epstein for teaching me grant writing course.

I would also like to thank my colleagues Ms. Cara Ullman, Ms. Rebecca Hoesing, Ms. Kate Lippincott, Ms. Kayanne Hamling, Ms. Kelly Veys, and Ms. Rachel Egbert for their contribution to data collection, processing, and other work in my research. I would also like to thank all my beloved colleagues in the Speech Production Lab and the Barkley Center, Ms. Cynthia Didion, Ms. Kimber Green, Dr. Mili S. Kuruvilla, Dr. Lori Synhorst, as well as other undergraduate and graduate students. I really enjoy work with them. I also thank all my friends for their help in my life in Lincoln.

I am indebted to my family, my farther Huitu Wang, mother Guanping Bao, and sister Yan Wang, who is on the other side of the Pacific Ocean, for their encouragement and great support to my long-term education.

GRANT AND SUPPORT

I would like to acknowledge the support of Barkley Trust, Barkley Memorial Center, Department of Special Education and Communication Disorders, University of Nebraska-Lincoln, and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States, PI: Jordan R. Green). I thank the UNL Institutional Review Board for approving my data collection. I also thank the College of Education and Human Sciences and the Barkley Salary Savings, which supported my travels in the past years. I thank Dr. John Bernthal and Dr. Stephen Boney for approving my travel budgets.

I also thank the Department of Computer Science & Engineering, which supported the early years of my doctoral study through Dohrmann Fellowship and Graduate Teaching and Research Assistantships.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Silent speech recognition from articulatory motion (simply called silent speech recognition in the rest of this dissertation) is the process of converting non-audio motion data of articulators (e.g., tongue, lips, and jaw) into speech in the form of text. Silent speech recognition does not rely on information in the acoustic stream of speech (sound wave data), which is the major and significant difference between silent speech recognition and traditional speech recognition. This chapter introduces the concept of silent speech recognition, motivation, background knowledge, related work, data collection device, a formal problem definition, as well as a dissertation outline.

Speech recognition from acoustics (often referred as automatic speech recognition, ASR, or simply speech recognition) has been studied for several decades (Rabiner, 1989). Significant progress in ASR has led to many successful commercial applications in specific domains including voice dictation, voice dialing, and automatic banking service through phone call, although the general speech recognition still has room for improvement. This technology has largely focused on recovering speech from the acoustic signals. More recently, however, other aspects of speech (e.g., visual information and orofacial motion) have become increasingly interesting to speech recognition researchers because the visual or articulatory information can be used for

improving the robustness of acoustic speech recognition by providing an extra source of input, which is referred as to audio-visual speech recognition (Livescu et al., 2007; Potamianos, 2003).

Recognition of speech based on acoustic and visual information (i.e., facial and lip information and without tongue information) formed the dominant paradigm due to the logistic difficulty data collection from tongue. However, tongue is the major articulator, particularly for vowel sounds; therefore, it is unlikely to obtain high recognition accuracy without tongue information. A few recent studies have used both tongue and lip information for ASR, which is referred as articulatory speech recognition (King et al., 2007). Those researches have shown that articulatory information is helpful for improving the robustness of speech recognition particularly when speech has background noise or for unintelligible speech (Rudzicz, 2011).

Acoustic information is not available in some situations, for example, individuals cannot produce sounds after laryngectomy, a surgical removal of larynx due to the treatment of cancer. After laryngectomy, the persons cannot produce a voice, but they still retain full function over tongue, lip and jaw movements. In addition, millions of adults and children in the United States have other speech impairments in the United States (Connolly, 1990; Deller, Hsu, & Ferrier, 1988; Pausch & Williams, 1992) produce very unintelligent speech as judged by human listeners, e.g., cerebral palsy, a group of disorders that can involve brain and nervous system functions such as speech, body movements, etc. (U.S. National Library of Medicine, 2009). Silent speech recognition from articulatory movements is designed to help individuals with speech impairments to communicate orally by playing back prerecorded or synthesized sounds based on their

articulatory movements. However, silent speech recognition from articulatory movements is extremely challenging, because tongue function during speech is still poorly understood (Kent et al., 1996). In classic phonetics, sounds are descriptively distinguished by a set of categorical articulatory features (AFs), e.g., lip rounding, tongue tip position, manner of production, etc. Extant recognition approaches are often based on feature classification tables. These approaches, however, have not obtained a high recognition accuracy (i.e., greater than 90%) from articulatory information, due to co-articulation effects (Kent & Minifie, 1977) and because articulation can vary significantly within those categorical features (Uraga & Hain, 2006).

This goal of this dissertation research was to achieve a better understanding of speech production and then to develop novel techniques for accurately and efficiently recognizing speech in form of text from continuous tongue and lip movements, where articulatory movements were represented as time-series of 3D spatial coordinates. This research was unique from prior research because (1) it focused on the movement patterns of tongue and lips, rather than articulatory features, and (2) the recognition is word- and

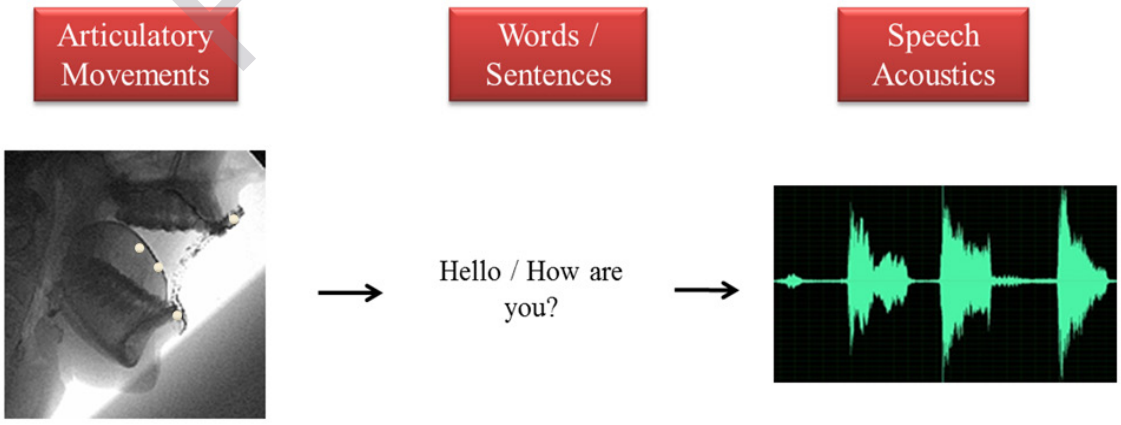

Figure 1.1 Schema of silent speech recognition from articulatory motion

sentence-level, rather than phoneme-level. In this dissertation, the production of speech was studied at multiple levels of speech units including phonemes, words, and sentences. Then we developed algorithms that not only achieve high recognition rates that are enough to be used in practical applications (i.e., greater than 90%), but also to perform this task efficiently for real-time applications. The long term goal is to employ the algorithms as the core of a recognition engine of an articulation-based silent speech interface (SSI) to enable individuals with laryngectomy to produced sounds using their tongue and lips (Fagan et al., 2010), which transforms articulatory movements to speech acoustics, as illustrated in Figure 1.1. The synthesis component can take many forms including a playback of recorded speech units or a text-to-speech (TTS) engine (Sproat, 1998) that outputs synthesized sounds. TTS is a well-researched topic in speech literature and is not a part of this research. The next subsection will give more details on silent speech interfaces.

1.1 Motivation

This research on silent speech recognition from articulatory motion is motivated by its significant scientific and clinic implications.

Scientific Knowledge: A better understanding of the mappings between articulatory movements and different units of speech including phonemes, words, and sentences will advance the understanding of articulatory basis of speech production. For example, the derived articulatory vowel space may be used together with the long established acoustic vowel space acoustic to study the relation between vowel articulation and acoustics The knowledge gained in this research will also facilitate efforts to

integrate articulatory models into speech synthesis (Shadle & Damper, 2001) and speech recognition (King et al., 2007; Rudzicz, 2011). As stated previously, articulatory movements have the potential to improve acoustic speech recognition when the quality of the input acoustics is low due to environmental noise or for unintelligent speech. The use of articulatory information to enhance acoustic speech recognition is becoming more prevalent and is referred to in the literature as articulatory speech recognition (Frankel & King, 2001; King et al., 2007) or audio-visual speech recognition (Livescu et al., 2007; Saenko, Darrell, & Glass, 2004).

Clinical Applications: The proposed recognition techniques also have the potential to provide a much needed objective and robust method for quantifying the degree of impairment in individuals with speech disorders. Currently, few objective methods exist to evaluate the degree of speech impairments. One testable hypothesis, for example, is that recognition accuracy from articulatory movement declines predictably with the degree of speech impairment. Thus, we can track the recognition accuracy of articulatory data collected a speaker with speech impairment at different times. The decline of the recognition accuracy may be highly correlated with the decline of speech motor control of the subject. In addition, quantitative articulatory vowel and consonant spaces we derived in this research (e.g., representing the articulatory distinctiveness between vowels and consonants) could be an indicator of the degree of speech impairment.

A long-term clinical application of this research was initially articulated by Pausch (Pausch & Williams, 1992), who discussed a research program centered on developing a real-time articulatory speech synthesizer to improve oral communication of

children with motor speech impairments. Specifically, an articulatory-to-acoustics synthesizer could be used to compensate for poor oral motor control in children with cerebral palsy (CP), or to enable laryngectomees to produce natural speech using their tongue and lips. In the United States, it is estimated 1.5-2.0 million children and adults have cerebral palsy. Approximately 10,000 infants and babies are diagnosed with cerebral palsy each year, and another 1200-1500 are diagnosed at preschool age (United Cerebral Palsy Research and Education Foundation, 2009). Each year, approximately 12,500 new cases of laryngeal cancer (Beenken et al., 2009) and an estimated 2,500 new cases of hyperpharyngeal cancer (Mendenhall, 2005) are diagnosed in the United States. After surgery, these individuals lose their ability to vocalize and, thus, communicate orally.

Currently, the commonly used alternatives for these patients are either Electrolarynx (a vibrating source held against the neck, which is the most common therapy currently (Bailey, 2006) or esophageal speech (burped speech). Both methods produce a very unnatural sounding voice. Moreover, when using the Electrolarynx, patients are required to hold the device against their neck. Esophageal speech (or Esophageal voice) is a speech production method that involves oscillation of the esophagus. This contrasts with traditional (healthy) laryngeal speech which involves oscillation of the vocal folds. Instead, air is injected into the upper esophagus and then released in a controlled manner to create sound used to produce speech. Persons with laryngectomy are usually trained to produce esophageal speech. However, esophageal speech is very hard to learn. Only 30% of the trained patients can produce perceptually acceptable sounds (Bailey, 2006).

Current challenges to developing a silent speech interface are both hardware and software related. Recent research on silent speech interfaces (Denby, Schultz, Honda, Hueber, Gilbert, & Brumberg, 2010) have shown the potential of developing portable, non-invasive, and affordable devices that can collect accurate data in real-time. Denby et al., (2010) provides a detailed comprehensive review of different silent speech approaches including Electromagnetic Articulograph systems that track the movement of sensor coils on articulators (Fagan et al., 2008; Perkell et al., 1992), ultrasound systems that track vocal tract shapes (Denby et al., 2010), non-audible murmur (NAM) microphone (Heracleous & Hagita, 2010; Nakajima, Kashioka, Shikano, & Campbell, 2003a, 2003b), surface electromyography (EMG) that record facial muscle activation patterns (Deng et al., 2009; Jorgensen et al., 2003; Jou et al., 2006; Schultz & Wand, 2010), electroencephalography (EEG) devices that track cortical activation patterns (Porbadnigk et al., 2007; Wolpaw & Birbaumer, 2002) and intracortical electrode approaches that record activity directly from cell in speech motor cortex (Brumberg et al., 2010). The interfaces between brain and computer interaction are also referred as brain-machine interface (BMI) or brain-computer interface (BCI) in literature (Lebedev & Nicolelis, 2006; Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007). Denby et al., (2010) also compared the strengths and weaknesses of different hardware technologies for practical real-time SSI applications. In this comparison, they evaluated each technology's potential for detecting speech in in silence, in noisy environment, for laryngectomy, and if they are non-invasive, ready for market, and in low cost. EMA-based data collection technology, which was used in this dissertation, is considered as one of the most promising technologies for practical SSI applications.

This dissertation targeted the development of algorithms (software) that accurately recognizing functional speech from articulatory motion for practical applications. Articulatory motion data were collected using an EMA (Carstens, Inc. Germany). Although EMA's (i.e., AG500) hardware is currently cumbersome for practical applications, it has the highest spatial resolution of all the electromagnetic tracking systems currently available. The spatial accuracy of data collected using the EMA is 0.5 mm (Yunusova, Green, & Mefferd, 2009). Thus, EMA AG500 was used for collecting data that was used to evaluate our algorithms. In the future, a more portable device can be used including that described by Fagan et al., 2008 and the Speech Wave System developed by NDI Inc. Canada (Berry, 2011). The methods and algorithms developed in this research will serve as the recognition component of a real-time articulation-based silent speech interface that could drive the playback of prerecorded or synthesized speech samples.

1.2 Speech Production Basics

Although speech is produced effortlessly by most talkers, the underlying coordination required to produce fluent speech is very complex involving dozens of muscles spanning the diaphragm to the lips. The speech production system includes the respiratory subsystem, the laryngeal subsystem, and the supraglottal subsystem (vocal tract). Exactly how speech is produced is still poorly understood (Kent et al., 1996).

The speech production literature contains diverse theoretical perspectives on the neurologic, physiologic, acoustic, perceptual, and environmental factors that govern speech. Examples are source-filter theory for speech acoustics (Fant, 1960; Stevens,