

USC-SIPI REPORT #426

**Emotional speech production: From data to computational
models and applications**

by

Jangwon Kim

December 2015

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

EMOTIONAL SPEECH PRODUCTION: FROM DATA TO
COMPUTATIONAL MODELS AND APPLICATIONS

by

Jangwon Kim

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2015

Copyright 2015

Jangwon Kim

Dedication

Dedicated to my family and friends.

Acknowledgements

This dissertation would not have been possible without the help and support of my family, friends, and colleagues.

Foremost, I would like to thank my advisor Shrikanth Narayanan for his advice and support. For the past five years, Narayanan has been an exemplary mentor and inspirer. I would like to thank Sungbok Lee for many times of valuable discussion on my research and support. I also thank my other dissertation committee, Krishina Nayak and Louis Goldstein for their insightful comments and suggestions. I have spent great summer at the Qualcomm Inc. in 2014. Thank Erik Visser for hosting excellent and enriching summer internship in his Audio R&D team, with Juhan Nam and Laehoon Kim who offered me immense advice and support. My research interest in speech signal processing started from the internship under Jeung-Yoon Elizabeth Choi. I would like to thank her, too.

I am fortunate to have interacted with amazing colleagues in Signal Analysis and Interpretation Lab., and I would like to recognize their contribution to my dissertation and other research work. I have greatly enjoyed interactions with Panayiotis Georgiou, Naveen Kumar, Prasanta Kumar Ghosh, Asterios Toutios, Ming Li, Michael Proctor, Maarten Van Segbroeck, Adam Lammert, Kartik Audhkhasi, Nassos Katsamanis and Tanaya Guha. In addition, I would like to thank Donna

Erickson and Jeonhyung Kang who have been offering insights and helps for my research.

Finally, I thank all my family and friends for their support. I thank my wife Janet Hur who changed my life with inspiration, encouragement and love. I thank my parents Kangsoo Kim and Minhee Kim who have given me a lifetime of love and care. Thank you all my friends. Your emotional supports have salvaged my sanity on many occasions. Last but not least I want to thank everyone that I forgot to mention.

Contents

Dedication	ii
Acknowledgements	iii
List of Tables	viii
List of Figures	xi
Abstract	xv
1 Introduction	1
1.1 Overview and goals	1
1.2 Background	6
1.2.1 Direct articulatory measurements	6
1.2.2 Previous studies on emotional speech production	8
1.3 Dissertation outline	11
2 Data processing technologies	12
2.1 Automatic parameterization of real-time MRI data	12
2.1.1 Introduction	12
2.1.2 Methods	13
2.1.3 Construction of grid lines	15
2.1.4 Lips and Larynx detection	15
2.1.5 Airway-path detection	18
2.1.6 Airway-tissue boundary segmentation	19
2.1.7 Evaluation of estimated airway-tissue boundaries	20
2.1.8 Conclusion and future work	23
2.2 Co-registration of real-time MRI and EMA datasets	23
2.2.1 Introduction	23
2.2.2 Relation to prior work	26
2.2.3 Data	27
2.2.4 Spatial alignment	28
2.2.5 Temporal alignment	30

2.2.6	Results	33
2.2.7	Benefits of co-registered data	36
2.2.8	Discussion	39
2.2.9	Conclusions and future works	41
3	Vocal tract shaping of emotional speech	44
3.1	Introduction	44
3.2	The USC-EMO-MRI corpus	47
3.2.1	Speech stimuli	47
3.2.2	Data acquisition and processing	48
3.2.3	Evaluation of emotion quality	49
3.3	Methods	51
3.3.1	MR image parameterization	51
3.3.2	Principal feature analysis	53
3.3.3	Computing the vocal tract length	55
3.4	Results	56
3.4.1	Emotional variations of principal features	56
3.4.2	Emotional variations of the vocal tract length	60
3.5	Discussion	61
4	Articulatory variability, linguistic criticality, and emotion	64
4.1	Introduction	64
4.2	Data	67
4.3	Linguistic criticality of articulators	71
4.4	Landmarks-based analysis on syllable segments	74
4.4.1	Selection of syllables	75
4.4.2	Extraction of articulatory parameters	77
4.4.3	Statistical analysis of articulatory kinematics	78
4.4.4	Analysis at the landmark points	83
4.5	Articulatory analysis at phonetic targets	88
4.5.1	Experimental Setup	88
4.5.2	Inter-emotion variability	90
4.5.3	Within-emotion variability	96
4.6	Simulation experiment	101
4.6.1	Description of articulatory model	101
4.6.2	Synthesis of non-critical trajectories	103
4.6.3	Results	105
4.7	Discussion and Conclusions	109

5	Invariant properties and variation patterns in emotional speech production	113
5.1	Introduction	113
5.2	Iceberg metric	115
5.3	Methods	120
5.3.1	Data	120
5.3.2	Parameter extraction	122
5.4	Analysis on the invariant properties of the C/D model	122
5.4.1	Iceberg point	124
5.4.2	Shadow angle	125
5.5	Analysis of emotional variability in the C/D model	126
5.6	Discussion and future works	130
6	Rich inversion using co-registered multimodal speech production data	132
6.1	Introduction	132
6.2	Related works	134
6.3	Co-registration	135
6.4	MR image parameterization	138
6.5	Rich inversion model	139
6.6	Experimental setup	141
6.7	Results	142
6.8	Application to emotion classification	144
6.8.1	Experimental setup	145
6.8.2	Results	146
6.9	Conclusions and future works	147
	Reference List	149

List of Tables

1.1	Comparisons of articulatory data recorded by EMA and rtMRI and of speech audio	8
2.1	RMSE between the estimated and manually-labeled boundaries in pixel unit	22
3.1	List of sentence prompts for the USC-EMO-MRI corpus. ‘Index’ refers to the sentence index.	48
3.2	Summary of evaluation results of all evaluators. ‘Sentences’ indicates the sentence ID included. ‘Average’ and ‘STD’ denotes average and standard deviation of the matching ratio (%) between target emotion and the final emotion label for sentence-level utterances, respectively.	50
4.1	Confusion matrix between evaluated emotion determined by majority voting and target emotion of speakers. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. The numbers in bold are the greatest of evaluated emotion cell for each target emotion, each speaker and each intended style.	71
4.2	List of stop and fricative consonants in the EMA database and the flesh point sensors of critical articulators of them. Note that /s/ and /z/ have two critical articulators, because both tongue tip constriction and tongue dorsum wide opening gestures are critical for the production of the phones [Nam et al., 2004]. The list of vowels in the EMA database is [ɑ, æ, ə, ɔ, ʌ, aʊ, aɪ, ɛ, eɪ, ɪ, i, oʊ, ɔɪ, u]. The flesh-point sensors corresponding to critical or non-critical articulators of vowels are not specified here due to their less clarity than consonants.	73

4.3	Number of CVC syllable samples used for analysis. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. CAC1 indicates the critical articulator of the first consonant of its syllable, and CAC2 indicates the critical articulator of the second consonant of its syllable. TT is the tongue tip, L is the lips.	76
4.4	The results of one-tailed <i>t</i> -test with mean deviation measure on the hypothesis that non-critical articulator has greater range of articulatory target position for each phone within emotion than critical articulator. ‘x axis,’ ‘y axis’ indicates the results of test on mean deviation value of the horizontal articulatory position or the vertical articulatory position, respectively. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. Only consonants are included in this analysis. Numbers in bold are statistically significant ($p < 0.05$). <i>T</i> -statistic is out of parenthesis in each cell, and <i>p</i> -value is in parenthesis.	97
4.5	The number of utterances selected for simulation experiment.	103
4.6	The results of evaluation of the estimated articulatory trajectories. The mean of RMSE or correlation coefficient is shown without parenthesis. The standard derivation is shown in parenthesis.	104
5.1	Confusion between the target emotion and the final emotion label, i.e., the best (perceived) emotion. ‘Neu’ is neutrality, ‘Ang’ is anger, ‘Hap’ is happiness, ‘Sad’ is sadness.	121
5.2	Statistical test on linearity between excursion and speed of the critical articulator for consonant. ‘**’ denotes that p -value < 0.0000005 . ‘*’ denotes that p -value < 0.00005 . $N=25$	125
6.1	AWBDs of JAATA and baseline system on two sets of data: the entire utterance pairs (denoted by ‘All’) and a subset of utterance pairs (denoted by ‘Subset’). The unit of AWBD value is msec.	137
6.2	The Pearson’s correlation coefficient (%) of each articulatory parameters for the best DNN system before LPF (denoted by ‘None’) and after LPF (denoted by ‘LPF’). ‘Freq.’ denotes the cut-off frequency of LPF.	143
6.3	The (averaged) Pearson’s correlation coefficient (%) of each subset of articulatory parameters after smoothing. ‘EMA’ refers to parameters of anatomical point tracking; ‘DF’ refers to parameters of distance function. ‘LP’ refers to lip protrusion parameter. ‘LH’ refers to laryngeal height parameter. ‘VTL’ refers to the parameter of the vocal tract length. ‘SHP’ refers to the parameters of oropharyngeal airway shape.	143

6.4	The Pearson's correlation coefficient (%) of each articulatory parameters on JR's data.	146
6.5	Unweighted accuracy (%) of emotion classification. 'Baseline' is for the baseline openSMILE acoustic feature set; 'Arti' is for the predicted rich articulatory feature set; 'Fusion' is for the feature-level fusion of the two sets.	146

List of Figures

- 1.1 Example images of EMA and rtMRI data. Left plot shows the placement of EMA pellets in the mid-sagittal plane. Right plot shows an MR image of the upper airway in the mid-sagittal plane. 6
- 1.2 Example plots of the maximum tangential speed of critical articulators and the maximum f0. A circle indicates the Gaussian contour with 2-sigma standard deviation for each emotion (red-Anger, green-Happiness, black-Neutrality, blue-Sadness). Different emotions show distinctive variation patterns in the articulatory-f0 space 10
- 2.1 The MR image after each pre-processing step 13
- 2.2 The grid lines (cyan color) superimposed on an MR image. Four blue dots are the manually selected landmarks. The origins (green color) of the forward polar grid lines (19 ~ 61) and the reverse polar grid line (61 ~ 76) are determined based on the landmarks. 16
- 2.3 Estimated vocal tract parameters: (a) estimated locations of forward-most edge of the lips (yellow color) and top of the larynx (cyan color), (b) airway path (cyan color), (c) airway-tissue boundaries (red line for inner boundary, green line for outer boundary). 17
- 2.4 Distance function from the larynx to the lips for Figure 2.3 (c). Green line is the shortest distance from the estimated outer boundary point for each grid line to the closest point in the inner boundary to it. Blue line is the distance between inner and outer boundary points for each grid line. 19
- 2.5 Errorbar of the distance (in pixel unit) between manual airway-tissue boundary and estimated airway-tissue boundary. From left to right in each phone, each errorbar is for pharyngeal region (black color), velar and dorsal region (red color), palatal region (green color), and labial region (blue color). 21
- 2.6 (a) Top 3% highest variance pixels are highlighted (along with their bounding box), which includes articulatory movements in vocal tract region. (b) Spatial alignment result - dark blue line is the estimated palate trace on MRI image. 27

2.7	Four examples of optimum MRI regions whose mean pixel intensities show highest correlation with corresponding sensor trajectories. Automatically selected pixel region is marked by a blue square box on each MRI image. ‘x’ or ‘y’ after sensor name, i.e., LI, indicates the direction of sensor movement (in the x or y axis).	34
2.8	Alignment maps of 4 example sentences with acoustic only (MFCC) and acoustic-articulatory features (MFCC+Artic). Reference is for manually corrected phoneme boundary (baseline). (a) and (b) are when JAATA performs better than only MFCC, (c) is when benefits from JAATA is minimal, and (d) is when JAATA performs worse than only MFCC.	36
2.9	Clean speech waveform (top plot) for the word “harms” and corresponding time series of velic (the second plot), pharyngeal (the third plot) and labial (bottom plot) opening. The velic and pharyngeal opening parameters extracted from rtMRI are synchronized with the the labial opening parameter extracted from the EMA by JAATA.	37
2.10	Left: Six EMA sensors (circles) overlaid on MRI image with estimated vocal tract boundaries (outer and inner lines in the vocal tract) and grid lines after co-registration. Right: Constriction degrees of the tongue tip (top plot) and tongue dorsum (bottom plot) extracted from upsampled rtMRI data for the sentence “Publicity and notoriety go hand in hand.” The circle for each phone is placed on the trajectory of the critical articulator of the phone, indicating the frame index for the phone in the registered data.	38
3.1	Results of parameterization processes of a magnetic resonance image (speaker M1) as an example	52
3.2	Grid lines (principal features) overlaid on an MR image for each speaker. The indices of the principal features are noted next to the corresponding grid lines	54
3.3	Averaged time series of the first and the seventh principal features for each emotion. The averaged time series were temporally aligned. The utterances of sentence 6 “nine one five two six nine five one six two” are used.	56
3.4	Quantiles and quantile range of principal features for anger, happiness and sadness relative to neutrality in data of speakers M1 and M2	58
3.5	Boxplots of the vocal tract length of each emotion	60
4.1	Placement of EMA sensors in the mid-sagittal plane	67
4.2	The five landmark points on the trajectory of the lower lip in /p a p/	77

4.3	P-value of Kruskal-Wallis test on each articulatory parameters, such as horizontal and vertical positions, tangential speed and tangential acceleration at each landmark point. ‘POS x’ is the position in the x axis, ‘POS y’ is the position in the y axis, ‘LM’ is landmark, ‘SPD’ is tangential speed, ‘ACC’ is tangential acceleration.	79
4.4	Histograms of the vertical velocity of the tongue tip at releasing onset point (landmark 1) for each emotion in JN’s data.	81
4.5	Example plots (speaker JN) of sample distributions (represented by 2-sigma ellipses) of articulatory positions at different landmarks . . .	83
4.6	Averaged articulatory trajectories of each emotion for “nine tight night pipes” in the sentence 4 in JN’s data. The two plots from the top show the averaged trajectories of the tongue tip; the other two plots show those of the lower lip.	85
4.7	Correlation coefficients of two averaged trajectories of the tongue tip and the lower lip in the vertical direction	86
4.8	Box plots of the average of centroid distance among emotion cluster pairs. CA is critical articulator case, NCA is non-critical articulator case. C() denotes consonants. Vowels are analyzed separately from consonants, because of their different nature for determining critical or non-critical articulator in this study. The Value above each box plot is the mean of each case.	91
4.9	Relative mean (centroid) of the horizontal position of each emotion to the neutrality (which is aligned to 0 on the y axis) in SB’s lower lip data.	92
4.10	Relative mean (centroid) of the horizontal (top subplot) and vertical (bottom subplot) positions of each emotion cluster for each phone to the neutrality (which is aligned to 0 on the y axis) in JN’s tongue dorsum data.	94
4.11	The average of centroid distances of emotion cluster pairs for the horizontal (left plot) or vertical (right plot) tongue dorsum positions, contrasted based on critical constriction gestures, such palatal constriction and pharyngeal constriction of the tongue dorsum, for vowels.	95
4.12	Scatter plots of the mean deviation of articulatory positions of each emotion of SB. Divided by 2 gray solid lines, the left most block is critical articulator (noted as CA), middle block is consonant non-critical articulator (NCA), the right most block is vowel non-critical articulator.	99
4.13	Example plots of the true and estimated trajectories of the tongue tip, the tongue dorsum and the lower lip in the vertical direction. An utterance of neutral emotion in JN’s data is used.	106

4.14	Unweighted emotion classification accuracy (%) for true and estimated data.	107
4.15	t -statistic of the pair-sample t -test on two distributions, one of true data and the other of estimated data for non-critical articulators for each phone, each articulator, each emotion and each speaker. * indicates that p -value is less than 0.05 for the case.	108
5.1	Example of iceberg metric, mean and variance of slopes in each band, overlaid with vertical trajectories of CA, after normalized to the range of the trajectories of CA.	117
5.2	Syllable triangles constructed for a neutral utterance of “Pam said bat that fat cat at that mat.” The 1st panel is the speech waveform. The 2nd panel shows syllable triangles. In the other panels, the red dash-dot line denotes the iceberg time point for onset; the green dashed line denotes the iceberg time point for coda; the blue solid line denotes the syllable center point.	123
5.3	Example scatter plots for the excursion of the critical articulator (of consonant) and the articulatory speed at icebergs in CV/VC demisyllables. “ <i>Pam</i> ” is used in this plot.	124
5.4	Errorbar plot of the “shadow” angle for each emotion.	125
5.5	Syllable magnitude, as jaw excursion, for each mono-syllabic word in the utterance	127
5.6	Ratio of articulatory speed (at the iceberg point for CV/VC demisyllable) to the syllable magnitude for each demisyllable.	128
5.7	The top panel shows the time difference between the onset pulse point and the iceberg point.	129
6.1	AWBDs of the baseline system (DTW + MFCCs) and JAATA for each utterance	137
6.2	Rich inversion model training and testing processes	140
6.3	Averaged Pearson correlation coefficient (%) between true and predicted parameters of EMA, computed for different DNN structures and feature sets (only EMA parameters or entire rich articulatory parameters for output) for model training	142

Abstract

Speech is one of the most common and natural means of communication, conveying a variety of information, both linguistic and paralinguistic. The paralinguistic information is crucial in verbal communication, because rich meaning (e.g., nuance, tone) in spoken language, and the states (e.g., emotion, health, gender) and traits (e.g., personality) of the speaker are encoded and decoded in paralinguistic factors. These two aspects of information (linguistic and paralinguistic) are encoded into speech sound jointly and simultaneously by the actions of speech articulators. Hence, a better understanding of production aspects of speech can shed further light on the information encoding (and decoding) mechanism of verbal communication. This dissertation seeks a better understanding of the articulatory control strategy for the multi-layered information encoding process, as well as developing computational models for expressive speech production system. This describes my achievements in the research pathway on emotional speech production, including data collection, data processing, analysis, computational modeling and applications.

The first is development of algorithms and software for data processing: (i) robust parameterization of magnetic resonance images and (ii) co-registration of

real-time Magnetic Resonance Imaging (rtMRI) data and ElectroMagnetic Articulography (EMA) data. These algorithms allow automatic and robust extraction of articulatory information of interest from these speech production data.

The second is collection (and release) of the USC-EMO-MRI corpus: a novel multimodal database of emotional speech production, recorded using the rtMRI technology. This corpus is designed as a resource to study inter- and intra-speaker variability in both articulatory and acoustic signals of emotional speech.

The third is novel findings and insight on emotional speech production. The specific sub-topics are (i) the vocal tract shaping of emotional speech, (ii) articulatory variability of emotional speech, depending on the linguistic criticality of the articulator, and (iii) invariant properties and variation patterns in speech planning and execution components for emotional speech. Specifically, for (i) this dissertation investigates inter- and intra-speaker variability using the USC-EMO-MRI corpus. For (ii), this dissertation reports experimental results suggesting that the large variability of linguistically less critical articulators is an important source of emotional information, and its relationship with the controls of corresponding critical articulators. This also offers novel insight regarding the relationship, based on computational modeling and simulation experiments. For (iii), this offers novel findings on the invariant properties and variation patterns in the perspective of the Converter/Distributor model.

Finally, the fourth is development of a computational framework to predict rich articulatory information (anatomical point tracking, vocal tract shaping, morphology) from speech waveform. The articulatory information is extracted from production data recorded using multiple data acquisition modalities (rtMRI and EMA) after registering the data of different modalities. Deep learning model is used to learn the acoustic-to-articulatory (inverse) mapping. The benefit of using

rich articulatory parameters for inversion mapping and emotion classification application is discussed.

Chapter 1

Introduction

1.1 Overview and goals

Speech conveys a variety of different kinds of messages, including linguistic information and paralinguistic information. Linguistic information refers to discrete categorical information, e.g., lexical meanings of words, of language. Paralinguistic information refers to para-language or non-lexical elements, e.g., emotion, gender, age, personality and health condition of the speaker. Paralinguistic aspects are essential for natural communication and interaction. They are closely associated with complex, subtle and delicate meanings (e.g., nuance) in spoken language, thus the crux of *rich* information. Paralinguistic information processing is useful for many technical applications, including Human-Machine Interaction (HCI), healthcare, security and defense. This study particularly focuses on emotional aspects in speech signal. Studying speech emotion is important for not just developing advanced emotion processing technologies, e.g., emotion recognition and emotion synthesis, but also for providing quantitative ways of assessing human communicative behavior. Emotional information processing can impact a variety of applications including in commerce, education and learning, and healthcare.

Most of emotional speech studies have predominantly focused on speech acoustic properties, such as voice quality, prosody and speech spectrum [Schröder, 2001, Vroomen et al., 1993, Frick, 1985, Yildirim et al., 2004, Busso et al., 2009, Gobl and NiChasaide, 2003, Ververidis and Kotropoulos, 2006, Williams and Stevens,

1972, Lee and Narayanan, 2009]. Although such knowledge contributes to a better understanding of emotional speech acoustic variability and technical applications (e.g., emotion recognition and emotional speech synthesis), there remain open questions as to how these emotional variations are encoded into the speech acoustic signal. The elegant acoustic structure of speech is produced by dynamic shaping of the vocal tract in coordination with respiratory and laryngeal behavior. Complex, but choreographed actions of the speech articulators do not encode only the linguistically significant resonant structure and aerodynamic qualities for vowels and consonants, but they also do rich expressive quality jointly. However, little has been systematically and quantitatively established in regard of the production aspects of emotional speech.

A better understanding of how emotion affects articulatory behavior during speaking can be beneficial for both speech science and technologies. Speech variability has been one of the most prevalent and difficult problems to understand and deal with in speech signal processing [Benzeghiba et al., 2007, Mozziconacci, 1998, Perkell and Klatt, 2014]. In particular, emotional variations are closely related to both intra-speaker and inter-speaker variability; Emotion expression varies significantly across time and speakers. Studying production aspects is directly related to this problem in the sense that speech articulation is one of the major sources of generation and modulation of speech acoustics. Hence, a better understanding of the emotional influence on articulatory movements can shed further light on this variability problem. This can also inform various speech applications for recognizing and synthesizing more human-like expressive quality, as well as improving robustness of linguistic and other paralinguistic processing systems, e.g., automatic speech recognition, language identification, and speaker verification, against

emotional variations. Understanding the emotional states of speakers is also essential for choosing natural and proper ways to respond in HCI [Cowie et al., 2001, Picard and Picard, 1997]. Computational models and tools for speech emotion have broad impact on a wide range of applications, because emotion processing is employed in many domains including commerce (e.g., customer care quality control [Picard and Klein, 2002, Komiak and Benbasat, 2004]), child education [Finkelstein et al., 2009] and healthcare (e.g., characterizing atypical or distressed behavior in Autism [Hobson et al., 1988, Hobson, 1986], depression [Edwards et al., 2002] and post-traumatic stress disorders [Davidson and Irwin, 1999, Mazza et al., 2012]).

One of the goals of this dissertation is to discover the realistic rules of articulatory controls for the multi-layered (linguistic and emotional) message encoding in speech. Previous studies in emotional speech production [Erickson et al., 2004, 2006, Lee et al., 2005] reported surface-level articulatory characteristics depending on emotion. In this study, the kinematic aspects of articulators are further investigated systematically by exploring the relationship among emotional state of the speaker, and the roles of articulators, and the articulatory variability. This study also investigates emotional variations in the entire midsagittal plane of the vocal tract, and reports novel findings.

This dissertation also describes my contributions to the USC-EMO-MRI corpus and novel algorithms for speech production data processing. First, this dissertation describes the objectives, data acquisition configurations, and detailed contents of the novel multimodal dataset of emotional speech, which was recorded using real-time Magnetic Resonance Imaging (rtMRI) technology [Narayanan et al., 2004]. This dataset is released publicly and freely for assisting emotional speech production studies about various aspects, including inter- and intra-speaker variability of

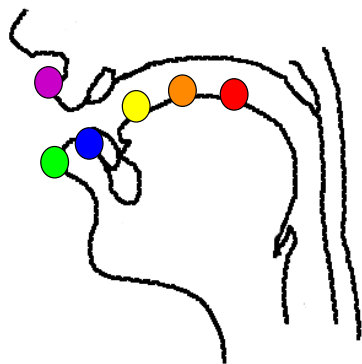
the vocal tract shaping for emotional speech, emotional effects on speech in the acoustic and articulatory domains, and their relationship. Second, this explains the robust algorithm that I developed for capturing linguistically important information from the rtMRI data automatically. A MATLAB implementation of this algorithm is released publicly and freely.

Another goal of this dissertation is to understand how emotional state of a speaker affects speech production process. Specifically, this dissertation is interested in emotional influence on the production components in both cognitive and surface-level stages. Under the framework of the Converter/Distributor (C/D) model, this explores invariant and variation aspects on both (pre-motor and motor) planning and execution stages [Perkell, 1999] in emotional speech. The pre-motor planning stage determines what (sequential lexical items of an utterance) and how (temporal organization of the items) to speak. The motor planning stage determines gestural organization of articulators. Finally, the execution stage converts the phonetic goals into articulatory muscle forces and eventually smooth articulatory motions under physiological constraints.

The C/D model proposed by Fujimura [2000] provides a theoretic framework for representing the comprehensive speech production processes, both temporal organization (planning) of an utterance and its realization (execution) to articulatory movements. This model contains speech variability factors in the individual stages and their relations. They are important components for the novel emotional speech production modeling at which this dissertation aims. Bonaventura [2003], Menezes [2003] have examined the validity of underlying assumptions of the C/D model and proposed algorithmic methods to infer temporal organization of an utterance from articulatory trajectories, only using neutral speech data. One contribution of this dissertation is to extend the C/D model to *emotional* speech

production model. In particular, this dissertation examines the validity of the underlying assumptions of the C/D model in emotional speech (and speech with contrastive emphasis), and discusses potential issues and methodology for computational implications for emotional speech generation using the C/D model.

The final goal of this dissertation is to develop a computational model for the inverse process of speech production, namely acoustic-to-articulatory inversion. In particular, this dissertation proposes the framework of estimating *rich* articulatory information using co-registered rtMRI and ElectroMagnetic Articulography (EMA) data. Despite recent progress in improving estimation accuracy [Liu et al., 2015, Najnin and Banerjee, 2015], there are only few studies [Aron et al., 2006] that succeeded improving the amount of articulatory information to be estimated. This may due to the limitation of currently available data acquisition modalities and the difficulty of simultaneous recording using multiple modalities. This dissertation proposes a methodology of using co-registered data for combining different types of articulatory information captured by multiple data acquisition modalities, in particular rtMRI data and EMA. The spatial and temporal alignment algorithm that I developed is used for generating the co-registered data. Then, deep neural network is adopted for learning the inverse mapping from clean speech audio (from EMA) to various kinds of articulatory parameters (from both EMA data and rtMRI data). This study examines the estimation performance of different kinds of articulatory information, and the benefit of offering rich information during model training in terms of prediction accuracy and for an application to emotion classification.



(a) EMA sensors attached on articulators (b) MR image of the upper airway

Figure 1.1: Example images of EMA and rtMRI data. Left plot shows the placement of EMA pellets in the mid-sagittal plane. Right plot shows an MR image of the upper airway in the mid-sagittal plane.

1.2 Background

1.2.1 Direct articulatory measurements

This section describes two data acquisition modalities for speech production research: One is EMA [Perkell et al., 1992] and the other is rtMRI [Narayanan et al., 2004]. This also discusses the difference, including advantages and disadvantages, of the individual modalities. The two modalities are chosen, because this dissertation uses speech production data recorded using them. For rtMRI, the data recording protocols and data specifications follows those of the USC-TIMIT corpus [Narayanan et al., 2014] and the USC-EMO-MRI corpus [Kim et al., 2014e], because the datasets were used for the studies in this dissertation. Similarly, the data information for EMA also based on the EMA data used in this dissertation.

EMA captures articulatory movements by tracking 3-dimensional (3D) coordinates of a handful of sensors attached on the surface of oral articulators. Speech waveform is often simultaneously recorded and later synchronized with the articulatory trajectories. Figure 1.1 (a) illustrates the placement of the six sensors in

the mid-sagittal plane for the USC-TIMIT corpus. Three sensors are placed on the tongue surface: the front-most sensor (yellow color in the figure) is placed about 0.5 - 1 cm behind the anatomical tongue tip for monitoring the movement of the tongue tip as well as minimizing its interference on the articulatory action; the rear-most sensor (red color) is attached as far back as possible for speakers (approximately 4 - 4.5 cm behind the tongue tip sensor) for capturing the movements of the tongue dorsum; and the third sensor (orange color) is positioned between the tongue tip and tongue dorsum sensors (typically at the center of the other tongue sensors). Sensors are also glued on the upper (violet color) and lower (green color) lips. Finally, a sensor is attached on the lower incisor for monitoring the movement of the jaw. 3D spatial coordinates of the sensors are recorded at a sampling rate of 100 Hz (using the NDI Wave Speech Research system). It is noted that the NDI Wave Speech Research system allows options of 100, 200, or 400 Hz and that the Carstens' AG500 EMA system allows 200 Hz.

The 3D coordinates of the six sensor position data are transformed for head movement correction and occlusal plane correction, based on reference sensor tracking. The orientation of the occlusal plane is measured using a half-rounded bite plane on which three reference sensors are attached. Typically, articulatory studies use the projections of the EMA sensors on the (horizontal) x-axis and the (vertical) y-axis as shown in Figure 1.1(a). Reference sensor trajectories and raw articulatory trajectories are smoothed by a 9th-order Butterworth filter, using 20 Hz cutoff frequency and 5 Hz cutoff frequency for articulatory trajectories and reference trajectories, respectively.

Recently, rtMRI technology has been employed for capturing the dynamics of the vocal tract shaping. Reconstructed Magnetic Resonance (MR) images offer the entire view of the upper airway, typically in the mid-sagittal plane. The rtMRI

Table 1.1: Comparisons of articulatory data recorded by EMA and rtMRI and of speech audio

	EMA	rtMRI
Frame rate (frame/sec.)	100	23.180
Monitored articulators	6 flesh points	The entire vocal tract in any plane
Data type	Motion capture	2D pixel image sequence
Audio quality	Clean speech	Noise-cancelled speech

data can offer the articulatory information in the vocal tract regions that are not easily monitored by using EMA. The frame rate of MR movies is 23.18 frames/sec. for the image resolution of 68×68 pixels (spatial resolution of 2.9 mm^2) in the USC-TIMIT corpus and the USC-EMO-MRI corpus. It is noted that the frame rate can be increased up to 162.23 frames/sec by finer sliding window reconstruction. It is also noted that sparse sampling and constrained reconstruction [Lingala et al., 2015] enables a frame of 83 frames/sec with spatial resolution of 2.5 mm^2 . Speech audio is recorded using a fiber-optic microphone at a sampling rate of 20 kHz simultaneously with MR image recording. Noise cancellation is performed as a post-processing procedure, using the normalized least mean square method based on the noise model [Bresch et al., 2006]. The audio and MR image sequences are synchronized.

Table 1.1 provides the specifications of EMA and rtMRI data in the USC-TIMIT corpus and the USC-EMO-MRI corpus.

1.2.2 Previous studies on emotional speech production

Human speech production system consists of two major components of vocal controls: voice source activity and articulatory movements. The prosodic and spectral aspects of speech sound are primarily governed by coordinated controls of voice source and articulatory modulations. Prosodic modulations have been considered

as the major component for emotion encoding in speech emotion research [Paeschke et al., 1999, Scherer, 2003]. Hence, previous studies [Mozziconacci and Hermes, Banse and Scherer, 1996] on emotional speech have been based on measurement and manipulation of prosodic parameters, such as duration, fundamental frequency (f_0) and intensity, which are obtained from the acoustic speech signal.

Compared to studies on acoustic properties of emotional speech, there have been only a few published studies on articulatory modulations in emotional speech. Erickson et al. [1998, 2004] reported that the positions of the tongue tip, the tongue dorsum, the jaw and the lips, can be characterized by emotion type of speakers. In addition, Erickson et al. [2006] reported the difference between acted and real sadness in terms of articulatory positioning. Lee et al. [2005] found that emotional speech articulation exhibits more peripheral or advanced tongue positions and that the movement range of the jaw is largest for anger. Both findings are in the line with the previous findings [Erickson et al., 2004, 1998]. Although these findings are valuable, they provide a limited view of emotional speech production, focused on the observation of a few anatomical points for vowels. Lee et al. [2005] reported better classification accuracy using articulatory parameters than using acoustic parameters for four acted emotions, such as neutrality, anger, sadness and happiness. Although it was obtained using a limited number of parameters and a relatively simple classifier (Fisher linear discriminant analysis), this result suggests that articulatory movements display significant emotion-dependent variations.

Later, Kim et al. [2012a] examined emotion-dependent information in both true articulatory trajectories and estimated trajectories (using inversion). A recently proposed algorithm based on the generalized smoothness criterion [Ghosh and Narayanan, 2010] was used for the inversion process. This study used EMA data

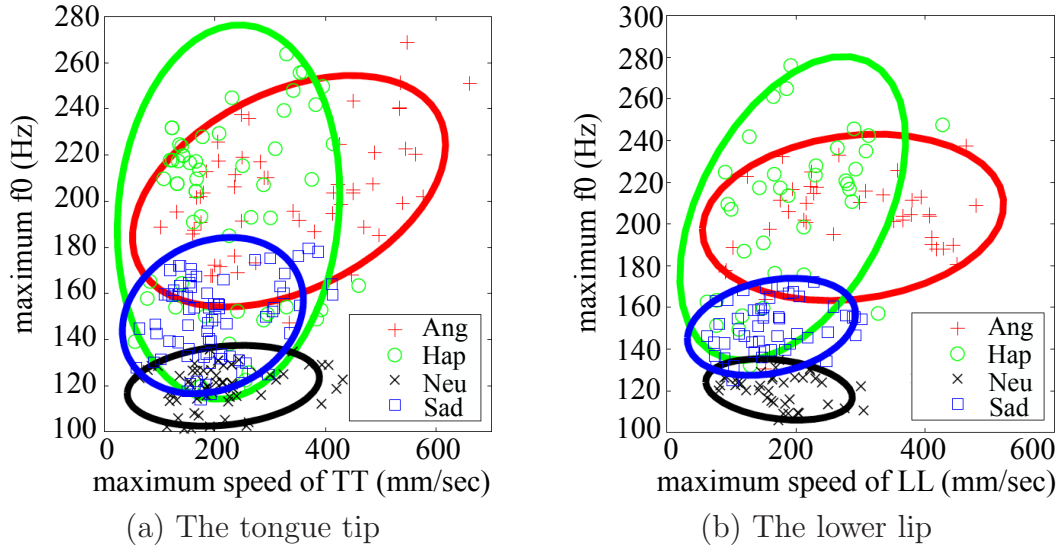


Figure 1.2: Example plots of the maximum tangential speed of critical articulators and the maximum f_0 . A circle indicates the Gaussian contour with 2-sigma standard deviation for each emotion (red-Anger, green-Happiness, black-Neutrality, blue-Sadness). Different emotions show distinctive variation patterns in the articulatory- f_0 space

for 5 elicited emotions, such as neutrality, hot anger, cold anger, happiness and sadness, spoken by 3 native speakers of American English. Experimental results suggest that the emotion-dependent information in the estimated trajectory, although smaller than that in the direct articulatory measurements, is found to be complementary to that in the prosodic features.

Kim et al. [2010] investigated the interplay between articulatory movements and f_0 patterns as a function of emotion. They found distinctive patterns in the prosodic-articulatory space, especially for happiness and anger. Figure 1.2 illustrates the distributions of 4 categorical emotions in the space of maximum f_0 and maximum articulatory speed in demisyllables. The examined speakers tended to emphasize articulatory speed modulations for angry speech, while emphasizing f_0 modulations for happy speech. Also, results indicate greater correlation between intensity statistics and articulatory speed statistics for high arousal emotions, such

as anger and happiness, than low arousal emotions, such as neutrality and sadness. These results suggest that the *joint* controls of prosody and articulation should be considered in the emotional speech production model.

1.3 Dissertation outline

The outline of this dissertation is as follows. This dissertation begins with my efforts on research creation for emotional speech production research. Chapter 2 presents my work on technical resource creation, such as automatic parameterization algorithm for rtMRI data, co-registration algorithm for EMA and rtMRI data, and their MATLAB implementations. Chapter 3 offers the details of the USC-EMO-MRI corpus. The objective, data acquisition protocols, data specifications, post-processing procedures and emotion quality evaluation are described. Initial analysis results on the vocal tract shaping depending on emotion are also provided. Chapter 4 discusses my scientific investigation on emotional speech production. In particular, this chapter presents empirical analysis and simulation results on the emotional variations in the articulatory variability, depending on the linguistic criticality of the articulator. Chapter 5 investigates emotional variations of speech production components in the planning and execution stages under the framework of the C/D model. Chapter 6 describes the inverse process of speech production, focusing on estimating rich articulatory information from speech acoustic signal. Finally, Chapter 7 concludes this dissertation with vision for future research.

Chapter 2

Data processing technologies

2.1 Automatic parameterization of real-time MRI data

2.1.1 Introduction

Real-time Magnetic Resonance Imaging (rtMRI) technology [Narayanan et al., 2004] is an important tool for studying human speech production. The vocal tract information that rtMRI offers encompasses the entire mid-sagittal view of the upper airway at a fast frame rate, which is spatially much richer than ElectroMagnetic Articulography (EMA). However, unlike EMA data, it is often required to extract interested vocal tract information, e.g., the vocal tract movements [Ramanarayanan et al., 2013] and the morphological structure of the vocal tract [Lammert et al., 2013], from the rtMRI data as a pre-processing for speech production research. Performing this parameterization automatically is essential for analyzing rtMRI data of speech production, that typically comprise hundred s or thousands of video frames; the complex structure of the vocal tract, non-uniform field sensitivity of the tissues in head and neck, grainy noise, magnetic resonance (MR) image artifact, and the rapidly varying irregular vocal tract shape, however, make this problem challenging. This study presents an algorithm for more robust segmentation of the MR images, which includes (1) retrospective pixel intensity correction of the MR images, (2) detection of the front-most edge of the lips and the top of

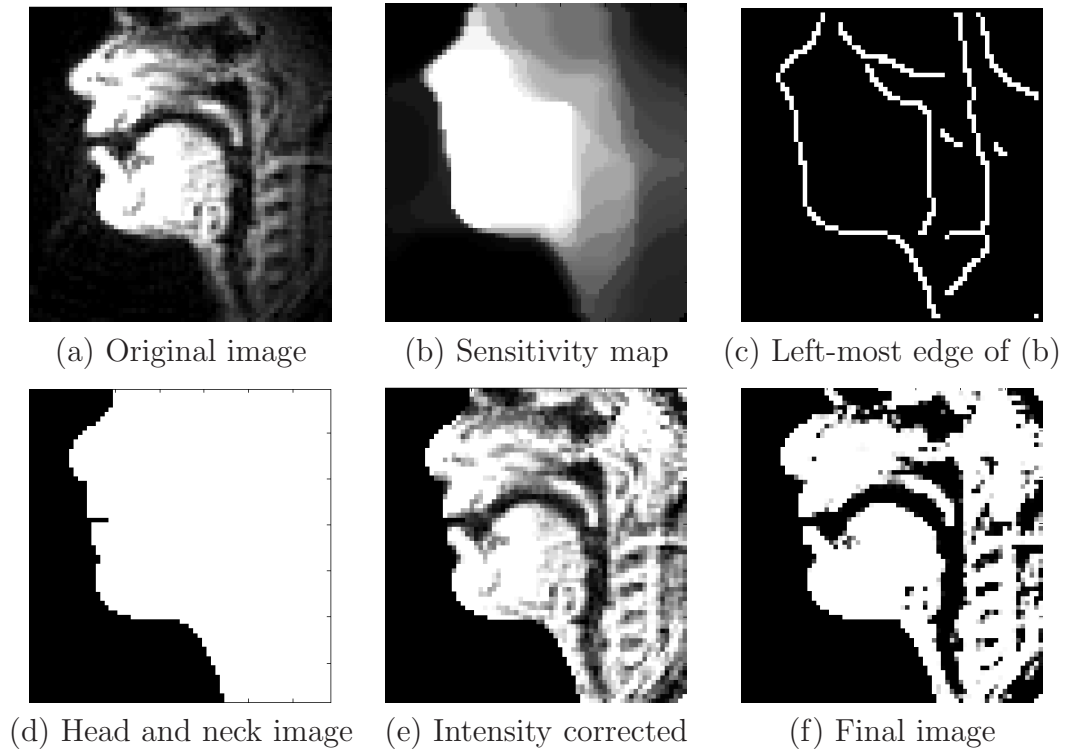


Figure 2.1: The MR image after each pre-processing step

the larynx, (3) segmentation of airway-tissue boundary in the vocal tract, and (4) measurement of the distance between the outer and inner boundaries. The current method improves the robustness of the airway-tissue boundary estimation over the previous method [Proctor et al., 2010] by using a combination of data-driven way of pre-processing of the MR images, robust airway path estimation, and model-based weighted linear curve fitting.

2.1.2 Methods

Pre-processing of MR images

Images of rtMRI data often suffer from grainy noise and non-uniform field sensitivity of the tissues, depending on recording configuration [Narayanan et al.,

2004]. Figure 2.1 (a) shows an example of the MR images in the USC-EMO-MRI corpus [Kim et al., 2014e], which was recorded at an image frame rate of 23.18 frames/sec and a spatial resolution of 68×68 pixels. The present algorithm uses a multi-resolution approach to minimizing the effects of the noise, artifacts, and non-uniform field sensitivity of the tissues. The details of the approach are as follows.

1. Create a field sensitivity map, denoted by S , of an original MR image using a morphological closing operation, followed by 2-dimensional median filtering. Figure 2.1 (b) shows the sensitivity map of the image in Figure 2.1 (a). The morphological closing operation selectively exclude low-intensity pixels (of grainy noise or artifacts in general) in the airway region when creating S .
2. Create the set of edge points, as in Figure 2.1 (c), of the sensitivity map using the Canny edge detector [Canny, 1986] implemented in MATLAB. Likewise, create the set of edge points of the original image. Let E_O and E_{SM} denote the sets of edge points of the sensitivity map and the original image, respectively.
3. Create the head and neck boundary line E_H by finding the left-most points of E_O and E_{SM} . Then, create a binary image, denoted by B , of the head-neck region by setting the pixel intensity to be 1 for pixels in the right side of E_H in each row and setting the pixel intensity to be 0 otherwise, as in Figure 2.1 (d).
4. Multiply the pixel intensity of the original image and the inverse of the pixel intensity of S for non-zero elements in B , while setting the non-tissue pixel intensity to be zero. Figure 2.1 (e) shows the result image, denoted by C .

5. Perform a sigmoid warping of the pixel intensity in C for suppressing grainy noise as well as highlighting tissue. Figure 2.1 (f) shows the final image.

2.1.3 Construction of grid lines

In order to detect the lips, the larynx, and the airway-tissue boundaries, the present algorithm constructs grid lines, adopting from the previous method [Proctor et al., 2010]. The previous method is motivated by the analysis of the upper airway image by Öhman [Öhman, 1967]. The grid construction method requires four manually chosen anatomical landmarks near the larynx, the highest point on the palate, the alveolar ridge, and the center of the lips, in one of the MR images. See [Proctor et al., 2010] for the details of grid construction that we follow. The differences from the previous method are (i) that a user chooses the distance between the center of adjacent grid lines in the present algorithm, not by the number of grid lines as in the previous method, and (ii) that the origin of the reverse polar grid lines is placed at the top of the image on the horizontal coordinate of the labial landmark point. This point offers more smooth transition from the forward polar grid lines to the reverse polar grid lines. Note that such method of the grid line construction assumes that the head and neck are aligned such that the subject faces the left side of the image and the neck is vertically straight.

2.1.4 Lips and Larynx detection

For each frame, the initial and the final grid lines correspond to the locations of the top of the larynx and the front-most edge of the lips, respectively. Since these articulatory positions vary slowly and smoothly over time, the present algorithm finds each of their optimal positions by constraining rapid change of the estimated locations of them.

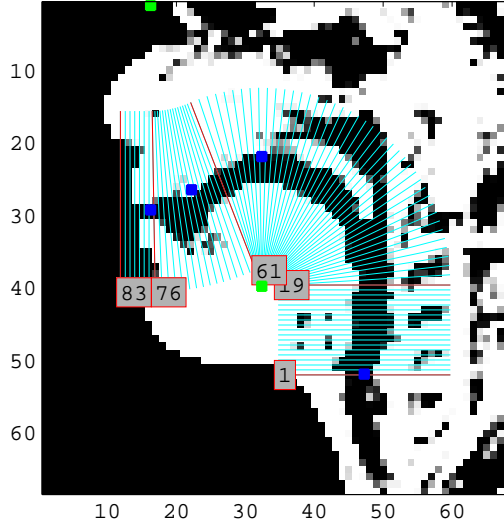


Figure 2.2: The grid lines (cyan color) superimposed on an MR image. Four blue dots are the manually selected landmarks. The origins (green color) of the forward polar grid lines (19 ~ 61) and the reverse polar grid line (61 ~ 76) are determined based on the landmarks.

Assume q_t is a state at instance t . N denotes the number of states. S_{q_i, q_j}^T denotes the transition score from q_i to q_j . $S_{q_i}^L$ is the likelihood score (of the observation) for q_i . P_i is the prior score of q_i . K is the number of instances. Q denotes a sequence of states q_1, q_2, \dots, q_K , one state for each instance. The objective score \mathcal{J} of Q is defined as follows:

$$\mathcal{J} = (P_1 S_{q_1}^L + w S_{q_2, q_1}^T) + \left(\sum_{u=2}^{K-1} S_{q_u}^L + w S_{q_{u+1}, q_u}^T \right) \quad (2.1)$$

where w is a weighting factor for S_{q_i, q_j}^T . The optimal sequence Q^* is obtained by finding Q associated with the minimum \mathcal{J} :

$$Q^* = \underset{[q_1, q_2, \dots, q_K]}{\operatorname{arg\,min}} \mathcal{J} \quad (2.2)$$

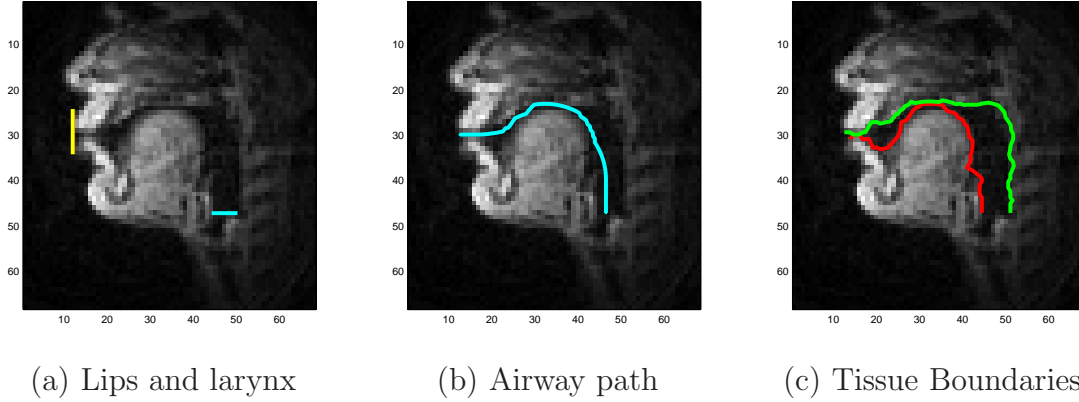


Figure 2.3: Estimated vocal tract parameters: (a) estimated locations of forward-most edge of the lips (yellow color) and top of the larynx (cyan color), (b) airway path (cyan color), (c) airway-tissue boundaries (red line for inner boundary, green line for outer boundary).

For detection problem of the edge of the lips, q_i corresponds to the i -th grid line, where $q_{N/2}$ is placed on the grid line of the labial landmark (the 77-th grid line in Figure 2.2). Also, $S_{x,y}^T$ is the Euclidean distance between the centers of grid lines x and y . S_x^L is the maximum pixel intensity of all pixels in the grid line x . Note that the length and width of the searching region for the lip detection are specified by users.

For the top of the larynx, q_i corresponds to the i -th grid line where $q_{N/2}$ is placed on the grid line of the larynx landmark (the first grid line in Fig. 2.2). $S_{x,y}^T$ is the same as defined for lips detection. Let D_x^L be the mean of the first-order derivatives of pixel intensities of x , computed along the grid lines. Then, $S_x^L = D_x^L \times W_x$, where W_x is an optional weighting term which gives more weight on higher grid line. W_x often helps for better estimation, especially when the low part of the larynx in MR images protrudes. This algorithm detects the point where the pixel intensity increases the most, searching from the top grid line.

The length and the width of searching regions (grid lines) for the lip detection and the larynx detection are specified by users. w is set to be 1 for these problems. One example of lips and larynx detection results is shown in Figure 2.3 (a).

2.1.5 Airway-path detection

The key idea behind improving airway-tissue boundary segmentation is to find an accurate and possibly approximate airway path in the upper airway first, from which the optimal airway-tissue boundaries can be determined easily and more robustly. The optimal airway paths passing through all grid lines in an MR image are determined by finding the paths of the minimum score, using the Viterbi algorithm. For this problem, each possible path in a grid line corresponds to a state, while each grid line corresponds to an instance. q_i corresponds to the i -th bin, where $q_{N/2}$ is located at the center of the grid line; $S_{x,y}^T$ is the Euclidean distance between bins x and y , where the bins are located in the adjacent grid lines, one bin for each grid line. S_x^L is the pixel intensity (observation) of the bin x , determined for each instance. Then, the optimal airway path is found by minimizing the score of possible bins as in the equation 2.2. The reason of using all bins, not only local minima as in the previous method [Proctor et al., 2010] is that all local minima are sometimes found outside the upper airway when some regions in the vocal tract is fully closed. The estimated airway path in our method can still stay within the region of interest during full contact in the upper airway, restricted by the transition costs between states.

Optionally, our algorithm performs a smoothing of the pixel intensity matrix (observations) using the mean of the 25% and 75% quantiles of the intensity values of neighboring pixels. We found that this smoothing is effective for reducing the estimation error caused by the low-intensity pixels outside the vocal tract walls,

because this smoothing tends to increase their intensity values. Also, the smoothing assists the airway path to stay inside the upper airway when a part of vocal tract is fully closed, by forcing the intensity of the present pixels of fully closed region to be low (because the past and future pixel intensities are low). Neighbors in the range of four instances, eight grids and four bins were used for the estimated airway path in Figure 2.3 (b). w in eqn. 2.1 was set to be 3.

2.1.6 Airway-tissue boundary segmentation

Two airway-tissue boundaries, i.e., the outer and inner boundaries of the vocal tract walls, are determined at the first bins whose pixel intensity is over a certain threshold in the outer direction and inner direction, respectively. The threshold was set to be 0.5, where the maximum pixel intensity of each MR image is 1.

The estimated airway-tissue boundary points are smoothed by the robust local regression using weighted linear least squares and a 1-st degree polynomial model, implemented in MATLAB, for each image frame. Figure 2.3 (c) illustrates the smoothed airway-tissue boundaries. Finally, a distance function for the airway-tissue boundaries is obtained by computing the Euclidean distance (in pixel unit)

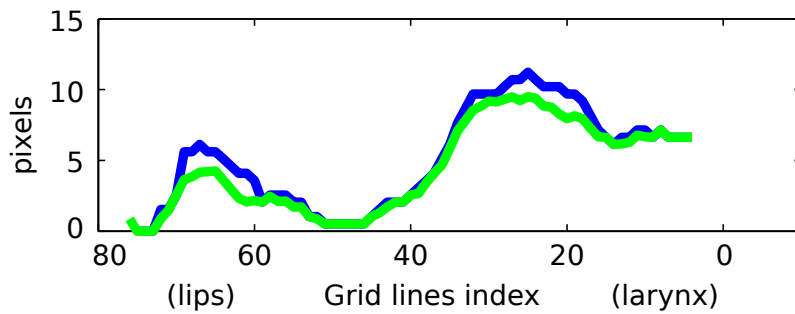


Figure 2.4: Distance function from the larynx to the lips for Figure 2.3 (c). Green line is the shortest distance from the estimated outer boundary point for each grid line to the closest point in the inner boundary to it. Blue line is the distance between inner and outer boundary points for each grid line.

between the outer and inner boundaries or between the outer boundary point and the closest inner boundary point regardless of their grid line. It was observed that the later (green line in Figure 2.4) is less erroneous, in particular near the lip region, than the former (blue line in Figure 2.4). The initial boundary point for computing the distance function is in the grid line of the estimated larynx. The final boundary point is in the grid line of the first local minimum distance from the final grid line. Figure 2.4 illustrates a distance function in the upper airway. The software package which contains MATLAB codes for the present algorithm and the subsets of data for demonstration is freely available at http://sail.usc.edu/old/software/rtmri_seg.

2.1.7 Evaluation of estimated airway-tissue boundaries

The estimated airway-tissue boundaries are evaluated against manually annotated tissue boundaries. For this purpose the annotators were instructed to sketch the inner and upper vocal tract walls using a continuous curve. For each of inner and outer boundaries, the Euclidean distance between each estimated boundary point and the closest point in the reference boundary for the estimated point is measured.

The statistics (mean and standard deviation) of the distance values are computed for each sub-region in the vocal tract and each phone. The sub-regions of the present algorithm are (1) grid lines 1 ~ 19 for pharyngeal region, (2) grid lines 20 ~ 52 for velar and dorsal constriction region, (3) grid lines 53 ~ 67 (alveolar ridge landmark) for the hard palate region, and (4) grid lines 68 ~ 77 for labial constriction region. The sub-regions of the previous algorithm are also determined in a similar way. The previous algorithm does not include the lip detection, thus large estimation error is observed in the grids after the lip landmark. For a fair comparison, the final grid line for analysis is fixed to the lip landmark point.

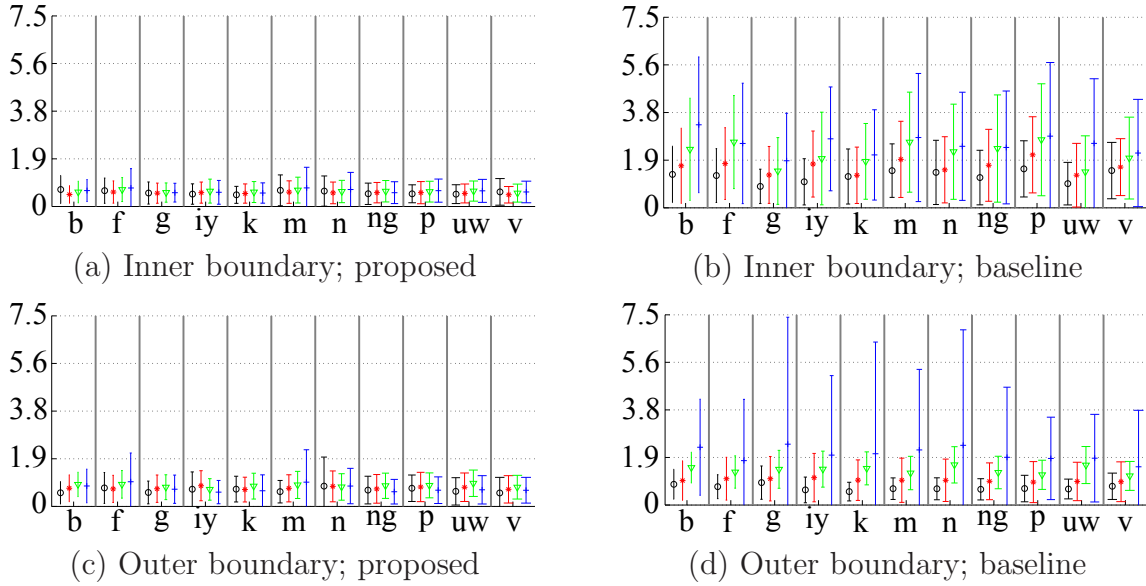


Figure 2.5: Errorbar of the distance (in pixel unit) between manual airway-tissue boundary and estimated airway-tissue boundary. From left to right in each phone, each errorbar is for pharyngeal region (black color), velar and dorsal region (red color), palatal region (green color), and labial region (blue color).

The palatal and dental corrections, and the mean pharyngeal wall were used as pre-processing for the baseline system [Proctor et al., 2010]. See [Proctor et al., 2010] for more details. For the present algorithm, the mean of the estimated boundary in the palatal region and the vertical position of the palate landmark is used in the final outer boundary. The reason for the palatal corrections in both algorithms is that the soft tissue in the hard palate region often shows significantly inner pixel intensity than other tissues, thus not sufficiently contrasted to the airway.

The list of phones used for evaluation is [b, f, g, iy, k, m, n, ng, p, uw, v]. Producing speech sound for these phones involves highly constricted or fully closed articulatory gestures, where the error of the estimated airway-tissue segmentation tends to be high. For each phone, 10 phone instances were randomly selected in a male subjects' data in the USC-EMO-MRI corpus. The acoustic phone boundary

Table 2.1: RMSE between the estimated and manually-labeled boundaries in pixel unit

Baseline		Proposed	
inner	outer	inner	outer
2.56	2.13	0.71	0.93

of each phone instance is obtained using an adaptive speech-text alignment tool, SailAlign [Katsamanis et al., 2011]. The image frames within the starting and final times with one marginal frame in each side were selected. In total, 492 image frames were used for evaluation.

Figure 2.5 shows the errorbar (as standard deviation) of the distance for each region and each phone for each estimated boundary. For inner boundary, the mean and standard deviation of the proposed algorithm is significantly smaller in all four regions than those of the baseline algorithm. Especially, the larger error in the front cavity (the palatal and labial regions) is significantly suppressed in the proposed algorithm. For the outer boundary, the baseline algorithm performs significantly better in the regions from the pharynx to the palate regions than for the inner boundary, presumably partially by the dental and palatal correction. However, the labial region still shows significantly large error. The amount of error in the labial region is significantly suppressed in the proposed algorithm. Table 2.1 shows the root-mean-squared-error (RMSE) for all estimated boundary points in each of the inner and outer boundaries. The proposed algorithm shows significantly inner RMSE for both inner and outer trajectories than the baseline algorithm. These results suggest that the proposed algorithm generates significantly more accurate airway-tissue boundaries than the baseline algorithm. In sum, the proposed algorithm generates more robust airway-tissue boundaries regardless of the phone and the vocal tract region than the baseline algorithm.

2.1.8 Conclusion and future work

The present algorithm estimates the airway-tissue boundaries from a robustly estimated airway path in each enhanced MR image. According to the quantitative evaluation on the estimated boundaries, the estimation error is significantly reduced by the present algorithm than the previous method [Proctor et al., 2010] in terms of RMSE (2.56 to 0.71 for the inner boundary; 2.13 to 0.93 for the outer boundary). A major advantage of the proposed method over the baseline is robustness across different regions in the vocal tract. The proposed algorithm also extracts the positions of the front-most edge of the lips and the top of the larynx automatically. This helps constrain the search space of the airway-tissue boundaries, resulting more robust boundary estimation. In addition, with the algorithm one can estimate the length of the vocal tract above the larynx.

Automatic head movement correction for each MR image is an on-going work that we would like to use for more robust and convenient tissue boundary estimation. In addition, this approach also calls for a pre-processing technique that is better suited to this imaging modality.

2.2 Co-registration of real-time MRI and EMA datasets

2.2.1 Introduction

Speech production research crucially relies on articulatory data acquired by various data acquisition modalities. Each modality has its advantage in terms of the nature of information it offers, while at the same time limited in important ways, notably in terms of the spatio-temporal details offered. Popular techniques

include ultrasound [Stone, 2005], X-ray microbeam [Fujimura et al., 1973], Electropalatography [Recasens, 1984], ElectroMagnetic Articulography (EMA) [Perkell et al., 1992] and recently introduced real-time Magnetic Resonance Imaging (rtMRI) [Narayanan et al., 2004]. For example, EMA offers motion capture of several flesh-point sensors in two (sagittal) or three dimensional (parasagittal) coordinates with high temporal resolution (100 samples/second in WAVE system), while real-time MRI (rtMRI) provides complete midsagittal (or along any arbitrary 2D scan plane) view of the vocal tract in relatively low temporal resolution (e.g., 68×68 pixel images at 23.180 samples/sec in the USC-TIMIT corpus [Narayanan et al., 2014]).

While it may be desirable to simultaneously acquire data with multiple modalities, it is not currently feasible due to technological limitations or incompatibility such as in the case of EMA and rtMRI. One possible way to obtain some of the combined benefits of EMA and rtMRI is by spatial and temporal alignment of datasets recorded with the same stimuli, by the same speaker, but at different times. However, differences in the dimensionality and quality of the measured articulatory and acoustic data across these two modalities make the alignment problem challenging. This study aims at obtaining the combined benefits of “multiple” data acquisition methods in modeling speech production dynamics by both spatial alignment and temporal alignment of these multimodal data. Specifically, it aims to obtain detailed vocal tract dynamics from MRI video aligned with EMA sensor trajectories. The alignment of multiple data will not only provide us finer and richer articulatory information, but also offer new opportunities for speech production research and modeling, i.e., temporal reconstruction (i.e., upsampling) of rtMRI based on EMA information, tongue reconstruction and complete tongue

movement representation from EMA pellets, palate reconstruction from EMA pellets, and their evaluations.

We use a corpus of TIMIT sentences collected from the same speakers, but at different times, with rtMRI and EMA as the basis for this study. The speech waveform and corresponding articulatory data (recorded simultaneously) within each dataset is provided as synchronized by the acquisition system itself (EMA by WAVE) or by an algorithm in the case of rtMRI [Bresch et al., 2006]. However, EMA TIMIT data and MRI TIMIT data need time warping alignment, because they were recorded separately. The temporal alignment of the two datasets is not straightforward due to several reasons. First, the nature of articulatory information of the two datasets is different: EMA is motion capture of flesh-point sensors and MRI is image stream. Second, rtMRI has grainy image noise and suffers from acoustic distortion in the speech audio signal. Lastly, the complex structure of articulators and their movements in rtMRI images make it hard to directly use spatio-temporal alignment techniques on the articulatory data.

In order to overcome the limitation of co-registering relying on any individual modality, such as using just acoustic feature based temporal alignment, we propose a novel temporal alignment using both acoustic and articulatory features, working with dynamic time warping (DTW) [Sakoe and Chiba, 1978]. The goal of this work is to examine how articulatory features can be used to improve temporal alignment. For instance, spatial alignment of articulatory data can be solved by transformation based on relatively stationary “reference” structures such as using palate tracking of both EMA TIMIT and MRI TIMIT. The automatic feature extraction technique in the novel temporal alignment formulation determines the set of pixels whose mean pixel intensity behaves similar to each EMA sensor trajectory. We demonstrate the performance of this alignment method on a subset

of the USC-TIMIT corpus elicited from a female speaker (denoted by F1 in the corpus) of American English.

This study is organized as following. Section 2.2.2 explains the relation of our new algorithm to prior work. Section 2.2.3 describes a multimodal speech production database, the USC EMA TIMIT and MRI TIMIT corpora, along with the details of post-processing them after acquisition. Section 2.2.4 describes our spatial alignment method and results. Next, Section 2.2.5 explains our temporal alignment method followed by its results in Section 2.2.6. Section 2.2.7 discusses the benefits of using co-registered data over data of individual modalities. Finally, discussions, conclusions and future works follow in Sections 2.2.8 and 2.2.9.

2.2.2 Relation to prior work

There have been spatio-temporal alignment studies in various domains including multimedia, medical imaging [Kovar and Gleicher, 2004, Ledesma-Carbayo et al., 2005, Kopp and Bergmann, 2007]. Although these methods have shown successful alignment results on their dataset of interest, they are not directly applicable to our multimodal data. This is mainly due to the different spatio-temporal nature of the multimodal data streams. Recently, canonical time warping (CTW) [Zhou and De la Torre Frade, 2009] was introduced for alignment task, which deals with different nature of data by alternating between the linear transformation of two original data spaces to a common latent space and temporal alignment. However, CTW based alignment is likely to fail when the two original feature streams have complex (nonlinear) relationships such as exhibited by the EMA sensor trajectories and MRI image streams. In fact we have found poor performance of CTW-based alignment on our corpus (see Section 2.2.8 for details).

Accurate information about the shape of the palate can be obtained by explicit measurements of the palate (i.e., taken from a dental cast), although in practice this can be labor intensive and uncomfortable for subjects. Previous work has tried to measure palate shape from flesh-point tracking data by asking subjects to sweep the tongue tip sensor across the palate, but this can be unreliable because subjects have trouble keeping the tongue tip sensor directly against the palate and precisely in the midsagittal plane [Westbury, 2005]. Palate shape can also be inferred from flesh-point tracking data, using all the sensor positions observed from an entire acquisition, for instance by taking the convex hull of those sensor positions [Tiede, 2010]. In the current study, palate shape is inferred from all tongue sensor positions in the data using a windowed technique which allows for more detail about palate shape to be preserved in the inference.

2.2.3 Data

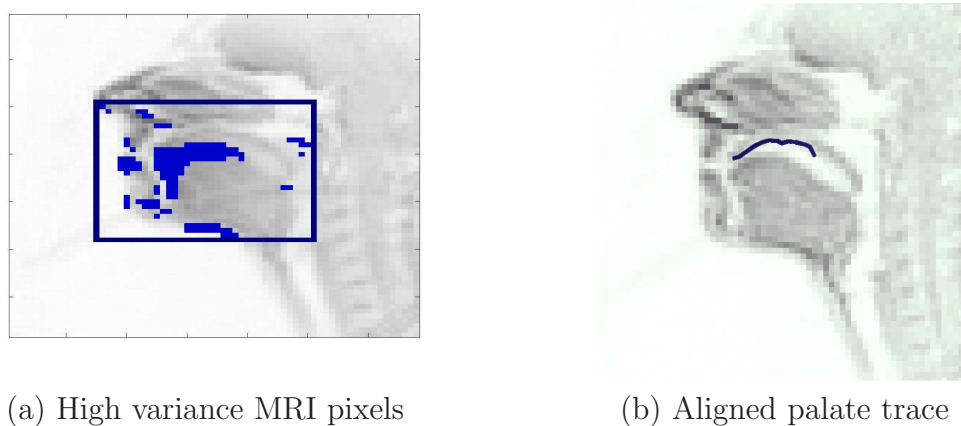


Figure 2.6: (a) Top 3% highest variance pixels are highlighted (along with their bounding box), which includes articulatory movements in vocal tract region. (b) Spatial alignment result - dark blue line is the estimated palate trace on MRI image.

We used the parallel EMA and rtMRI data of a female speaker (F1) in the USC-TIMIT corpus [Narayanan et al., 2014]. More details of the database, data collection and post-processing, including noise cancellation on speech audio, are offered in [Narayanan et al., 2014, Bresch et al., 2006], as well as in Section 1.2.1, thus omitted here.

Figure 2.6(a) shows a sample MRI video frame along with top 3% high variance pixels. With the same stimuli and subjects of MRI TIMIT we also collected, at a different time, flesh-point tracking EMA data using WAVE system (referred to as EMA TIMIT), which includes the trajectories of 6 flesh-point sensors on tongue tip (TT), tongue blade (TB), tongue dorsum (TD), upper lip (UL), lower lip (LL) and lower incisor (LI), at a sampling rate of 100 Hz and simultaneously recorded speech audio. Following the procedure outlined in [Kim et al., 2011a], we performed post-processing which includes smoothing and occlusal plane correction on EMA sensors. The x,y co-ordinate trajectories of six EMA sensors (i.e., 12 EMA trajectories) are used for our experiments. EMA TIMIT also contains palate tracking. In palate tracking, a subject scans the upper surface of the vocal tract from the alveolar ridge to the soft palate, using the TT sensor. This palate tracking along with MRI image is used for spatial alignment. For analyzing the performance of temporal alignment, we use identical set of 20 sentences (~ 40 sec) from the MRI TIMIT and EMA TIMIT such that they cover all phonemes.

2.2.4 Spatial alignment

The goal of spatial alignment is to align the reference midsagittal plane (i.e., x-y plane) in EMA recording with MRI scan plane such that EMA sensor coordinates on the midsagittal plane correspond to the respective points on the MRI image. The spatial alignment is achieved by estimating the transformation of EMA sensors

on the MRI image. We uses MRI image and palate tracking of EMA sensors for this task. The spatial alignment of articulatory sensors on MRI image can be done by applying the same transformation on the sensor coordinates. We estimate the palate contour from EMA palate tracking data as well as all tongue sensor data by choosing the highest vertical point in each adjacent bins ($L/20$ mm in length, no overlap), where L is the length of the palate tracking data, along x axis. To find a location for the palate trace in the MRI image plane, we firstly scaled down EMA sensors by 2.9 (Note that unit of EMA sensors is mm, and the pixel size of MRI image is 2.9 mm). Then, after manual initialization, we perform a grid search over a variety of translations, δ_x and δ_y (along x and y axis), from -5 to +5 pixels at increments of 0.5 and rotations θ from $-\pi/4$ to $\pi/4$ radians at increments of $\pi/32$ radians. The manual initialization is done at (horizontal pixel = 25th, vertical pixel = 23th, rotation = 0). The optimum translation and rotation is found to be ($\delta_x^* = 25.5$, $\delta_y^* = 24$, $\theta^* = -\pi/32$). δ_x^* , δ_y^* , and θ^* are found by maximizing the contrast across palate trace as follows:

$$\{\delta_x^*, \delta_y^*, \theta^*\} = \arg \max_{\delta_x, \delta_y, \theta} \sum_{\forall i, j \in \text{palate trace}} \frac{p_{i,j-1}}{p_{i,j+1}} \quad (2.3)$$

where $p_{i,j}$ is a pixel at (i, j) of standard deviation (SD) MRI matrix. The SD MRI matrix contains the standard deviations of MRI image pixels. In SD MRI matrix the palate is clearly visible as a region of high contrast just above the oral cavity and it also guards against the false palate problem unlike the raw MRI image matrix. Due to the unavailability of ground truth we visually examine the spatial alignment result. Figure 2.6(b) shows the optimum palate trace location of EMA on MRI image. Visually it appears that the transformation of EMA results in a good match between EMA palate trace and the palate visible in MRI image.

2.2.5 Temporal alignment

Below we describe our proposed automatic algorithm for temporal alignment of MRI and EMA recordings using both acoustic and articulatory features. We refer to this automatic algorithm as Joint Acoustic-Articulatory based Temporal Alignment (JAATA). A key feature of JAATA is that it computes EMA-like features from raw MRI video in order to achieve optimum alignment.

Objective function

Suppose we need to perform temporal alignment of MRI and EMA recording of F sentences. Suppose the f -th ($1 \leq f \leq F$) sentence has N_M and N_E frames in MRI and EMA recordings, respectively. Let $\mathbf{X}_{M,f} = [\mathbf{x}_{1,M} \ \cdots \ \mathbf{x}_{N_M,M}]$ denote the acoustic feature sequence matrix of MRI audio of the f -th sentence where $\mathbf{x}_{l,M}$ is the acoustic feature vector at the l -th frame. Similarly, let $\mathbf{X}_{E,f} = [\mathbf{x}_{1,E} \ \cdots \ \mathbf{x}_{N_E,E}]$ denote the acoustic feature sequence matrix of EMA audio. We vectorize MRI video in each frame, i.e., at l -th frame MRI video matrix $V_{l,M}$ (68×68) is converted to MRI video vector $\mathbf{y}_{l,M}$ ($68^2 \times 1$) such that $\mathbf{y}_{l,M}(68j + i) = V_{l,M}(i, j)$, $0 \leq i, j \leq 67$. Thus, for the f -th sentence, we obtain the MRI video sequence matrix $\mathbf{Y}_{M,f} = [\mathbf{y}_{1,M} \ \cdots \ \mathbf{y}_{N_M,M}]$. The 12 EMA sensor trajectory matrix is denoted by $\mathbf{Y}_{E,f} = [\mathbf{y}_{1,E} \ \cdots \ \mathbf{y}_{N_E,E}] = [\mathbf{z}_{E,f}^1 \ \cdots \ \mathbf{z}_{E,f}^{12}]^T$, where $\mathbf{y}_{l,E}$ (12×1) represents the 12 EMA sensor values at the l -th frame and $\mathbf{z}_{E,f}^q$ ($N_E \times 1$) is the trajectory of the q -th EMA sensor for f -th sentence. T is the matrix transpose operator. We obtain the best temporal alignment between

MRI and EMA recordings of all F sentences by minimizing the following objective function:

$$\begin{aligned}
& J(\lambda, \{\mathbf{W}_{M,f}, \mathbf{W}_{E,f}\}, \{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}) \\
&= \sum_{f=1}^F J_f(\lambda, \mathbf{W}_{M,f}, \mathbf{W}_{E,f}, \{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}) \\
&= \sum_{f=1}^F \left\{ \lambda \left(\left\| \mathbf{X}_{M,f} \mathbf{W}_{M,f} - \mathbf{X}_{E,f} \mathbf{W}_{E,f} \right\|_F^2 \right) \right. \\
&\quad \left. + (1 - \lambda) \left(\sum_{q=1}^{12} \left\| \frac{1}{A} \mathbf{s}_{q,M}^T \mathbf{Y}_{M,f} \mathbf{W}_{M,f} - (\mathbf{z}_{E,f}^q)^T \mathbf{W}_{E,f} \right\|^2 \right) \right\} \tag{2.4}
\end{aligned}$$

The objective function J is obtained by summing objective functions J_f corresponding to each sentence. J_f has two terms which are convexly combined using weight λ - the first term measures the Euclidean distance between acoustic features of MRI and EMA audio after alignment and the second term measures the same for articulatory features. $\|\mathbf{U}\|_F^2 = \text{Tr}(\mathbf{U}^T \mathbf{U})$ designates the Frobenious norm. $\mathbf{W}_{M,f}$, $\mathbf{W}_{E,f}$ encode the time alignment path for f -th sentence (see [Zhou and De la Torre Frade, 2009] for details).

$\mathbf{s}_{q,M}$ ($68^2 \times 1$) is a masking matrix, whose non-zero elements selects a submatrix (of size $K \times L, K, L \in \mathcal{Z}$) from the MRI image matrix. Thus, $\frac{1}{A} \mathbf{s}_{q,M}^T \mathbf{Y}_{M,f}$ is the articulatory trajectory derived from MRI video corresponding to q -th EMA trajectory. The number of pixels or the area of the submatrix is denoted by $A (= KL)$, which is user-specified before optimizing J . The elements of $\mathbf{s}_{q,M}$ can take value of 0 or 1. Thus, $\mathbf{s}_{q,M}^T \mathbf{1} = A$, where $\mathbf{1}$ is a column vector of all ‘1’s.

Optimization of the objective function

Minimization of J is a non-convex optimization problem with respect to the optimization variables $\mathbf{W}_{M,f}$, $\mathbf{W}_{E,f}$ (time alignment matrices), $\{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}$ and λ . Hence we use an iterative approach comprising two main steps - 1) Optimize $\mathbf{W}_{M,f}$, $\mathbf{W}_{E,f}$ using DTW given $\{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}$ and λ , 2) Given $\mathbf{W}_{M,f}$, $\mathbf{W}_{E,f}$ $\forall f$ and λ , optimize $\{\mathbf{s}_{q,M}\}$ sequentially $\forall q$ by searching over K, L such that $KL = A$. λ is optimized by performing a grid search. It is easy to show (from (2.4)) that in each of these steps J decreases monotonically. Thus the iterative process of optimization stops when the value of J reaches a local minima. The iterative process is initialized with the temporal alignment obtained by acoustic-only features using DTW and Euclidean distance between acoustic features as the distance measure.

Experimental setup

We use 13 dimensional Mel-Frequency Cepstrum Coefficient (MFCC) vector as the acoustic feature \mathbf{X}_M and \mathbf{X}_E for both MRI TIMIT and EMA TIMIT audio. MFCCs are computed at a frame rate of 100 Hz. Note that 12 EMA trajectories are also at a frame rate of 100 Hz. We applied smoothing on the EMA trajectories by butterworth filter with a cut-off frequency at 8 Hz. 8 Hz is chosen by the frequency analysis in a previous work in [Ghosh and Narayanan, 2010]. We have computed the derivative of EMA trajectories and denote them as \mathbf{Y}_E . Similar to the EMA trajectories, we also low-pass filtered MRI video pixel trajectories using a butterworth filter with a cut-off frequency at 8 Hz. Since MRI videos have a lower frame rate, we have upsampled the MRI video at a sampling rate of 100Hz such that both acoustic and articulatory data streams are at identical frame rate. This frame rate was chosen to match the frame resolution of the phone boundary,

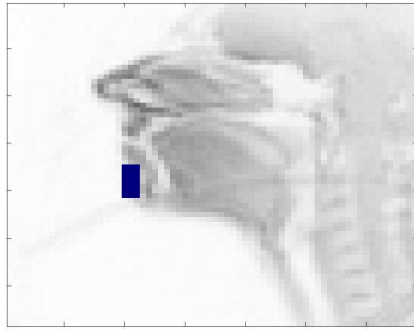
which is used for evaluation of temporal alignment. Derivatives of the upsampled MRI pixel trajectories are computed and used as \mathbf{Y}_M . We normalized both EMA and MRI articulatory feature trajectories between 0 and 1 for each sentence. We have found that derivative computation and normalization contribute to better temporal alignment performance.

As discussed in Section 5, for each EMA trajectory, the optimum rectangular region on the MRI image is estimated as a by-product of the temporal alignment formulation. Trajectory of the derivative of the mean pixel intensity of MRI in the optimized area is used for temporal alignment. To reduce the search space for finding the location of the optimum rectangular area, we restrict the search to a bounding box of the top 3% high variance pixels (see Figure 2.6(a)) which contains the surface movement of articulators. The λ values used for optimization are $\{(k - 1) \times 0.05, 1 \leq k \leq 20\}$.

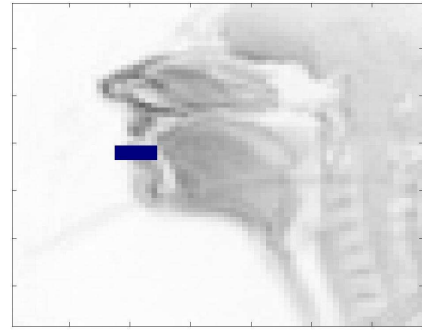
For evaluation of the temporal alignment, we have used an objective measure of how the phonetic boundaries of MRI audio correspond to those of the EMA audio when mapped using the optimized alignment path. We call this measure as Average Phonetic-boundary Distance (APD). Phonetic boundaries obtained from forced alignment [Katsamanis et al., 2011] are manually corrected to be used in this evaluation. APD is computed as the root mean square (RMS) value of the difference between the manually corrected phonetic boundaries and the estimated phonetic boundaries in EMA audio obtained by mapping phonetic boundaries of MRI audio using the temporal alignment.

2.2.6 Results

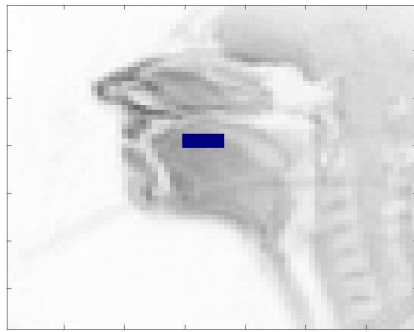
We experimented with different values of rectangular area A - 9, 12, 15, 18, 21, 24, 30, 32, 36. For all these different choices of A , the optimum value of λ turns out to



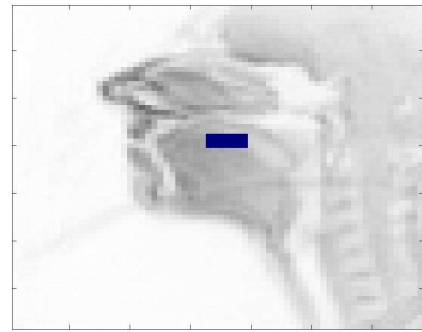
(a) for LIx ($\rho=0.68$)



(b) for LLy ($\rho=0.67$)



(c) for TTy ($\rho=0.65$)



(d) for TBy ($\rho=0.64$)

Figure 2.7: Four examples of optimum MRI regions whose mean pixel intensities show highest correlation with corresponding sensor trajectories. Automatically selected pixel region is marked by a blue square box on each MRI image. ‘x’ or ‘y’ after sensor name, i.e., LI, indicates the direction of sensor movement (in the x or y axis).

be 0.1. For different choices of A , APD averaged over all sentences reduces by ~ 6 msec when articulatory features are used in addition to MFCC by JAATA. The minimum APD, 44.198 msec occurs with $A=21$ compared to an APD of 50.101 msec using only MFCCs. To have deeper insights, we, therefore, investigate the quality of alignment for each sentence with $A=21$.

We firstly examine the optimum rectangular region on MRI image for each EMA trajectory. Figure 2.7 shows the estimated regions of MRI image with $A = 21$ for

four different EMA trajectories, namely Llx, LIy, TTy, TBy. From Figure 2.7 it is clear that the regions correspond to the respective articulators on the MRI image. The mean pixel intensity indicates the constriction degree in the region of selected pixels. Constriction degree measurement of a specific vocal tract region of rtMRI data has been used in earlier speech production studies i.e., [Hagedorn et al., 2011, Lammert et al., 2011]. However, finding the “best” region corresponding to each EMA trajectory by hand is not straightforward. Varying morphological structure of subjects sometimes makes it hard to decide the best region. Thus our proposed optimization for temporal alignment offers a solution in this regard. To examine how correlated the mean pixel trajectory is with the corresponding EMA trajectory, we also report correlation coefficient (ρ) between the two. ρ , when averaged over all articulators, is 0.59 with a SD of 0.10. ρ values for different articulators ranges from 0.36 (ULy) to 0.68 (Llx). ρ values suggest that, on an average, the features from the mean intensity over optimum MRI regions are linearly correlated to the respective EMA trajectories.

Figure 2.8 shows example alignment maps for four different sentences obtained using only MFCC and with both MFCC and articulatory features (‘MFCC+Artic’) using JAATA. As a reference alignment, we have also shown an alignment based on phonetic boundaries (‘Reference’). These four cases are chosen to illustrate the sentences where use of articulatory features led to better as well as worse alignment compared to only MFCC based alignment. For example, APD decreases by 134 msec for sentence 19 (Figure 2.8(b)) and by 34 msec for sentence 3 (Figure 2.8(b)) by using automatically extracted articulatory features in addition to MFCC. However for sentence 12, we observed that APD increases by 52 msec (Figure 2.8(d)).

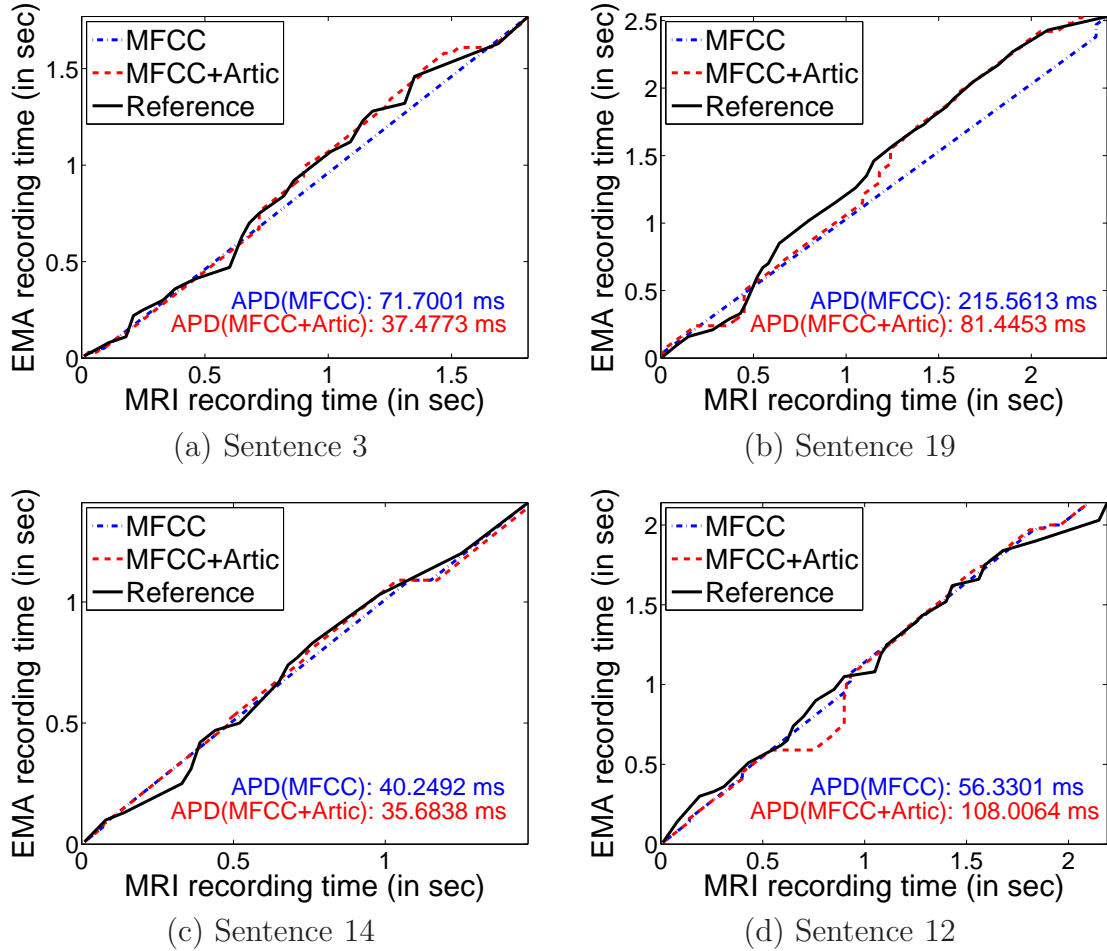


Figure 2.8: Alignment maps of 4 example sentences with acoustic only (MFCC) and acoustic-articulatory features (MFCC+Artic). Reference is for manually corrected phoneme boundary (baseline). (a) and (b) are when JAATA performs better than only MFCC, (c) is when benefits from JAATA is minimal, and (d) is when JAATA performs worse than only MFCC.

2.2.7 Benefits of co-registered data

The co-registered data can offer spatially or temporally richer articulatory information than either EMA or rtMRI data by themselves. This section illustrates some ways in which co-registered data can be used for taking advantage of both EMA and rtMRI data for speech production research.

Information from more speech articulators

Articulatory information that is not directly available from EMA sensors, e.g., constrictions in the velar and pharyngeal regions, can be measured from MR images in the co-registered dataset. An example of this can be seen in Fig. 2, which shows three articulatory time series extracted during articulation of the word “harms.” The velic and pharyngeal opening parameters were extracted from rtMRI data using Region-Of-Interest (ROI) analysis [Lammert et al., 2010]. Labial opening was extracted from EMA data as the Euclidean distance between the upper and lower lip sensors in the midsagittal plane. The action of the lips is accurately captured, and the closure of the lips during production of /m/ can be clearly seen. Moreover, labial closure is coordinated in time with the velic opening to produce the nasal sound, with both time series showing a similar time course. The

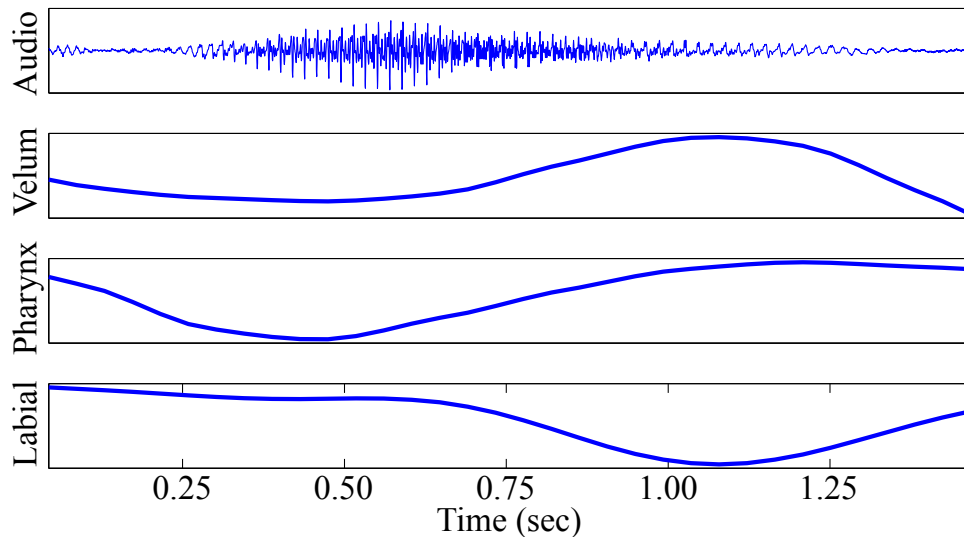


Figure 2.9: Clean speech waveform (top plot) for the word “harms” and corresponding time series of velic (the second plot), pharyngeal (the third plot) and labial (bottom plot) opening. The velic and pharyngeal opening parameters extracted from rtMRI are synchronized with the the labial opening parameter extracted from the EMA by JAATA.

pronounced pharyngeal constriction is also well captured, during the production of /ɑ/ and /ɪ/ and preceding the nasal.

Higher temporal information and tongue landmarks for rtMRI data

Spatio-temporal alignment of rtMRI and EMA can be used for articulatory landmark tracking in the MR images with improved temporal resolution as a result of co-registration. Anatomical landmarks are not always conspicuous in MR images (e.g., tongue tip) because certain speech articulators, particularly the tongue, change drastically in shape over time. These shape changes can obscure or make indistinguishable anatomical landmarks, and can present challenges for landmark tracking in rtMRI. The spatio-temporal alignment can provide information about which point in each MR image corresponds to each EMA sensor that was placed at an anatomical landmark in the vocal tract. In addition, the alignment map between rtMRI and EMA can assist in upsampling rtMRI data by utilizing the higher temporal resolution of EMA to interpolate between rtMRI

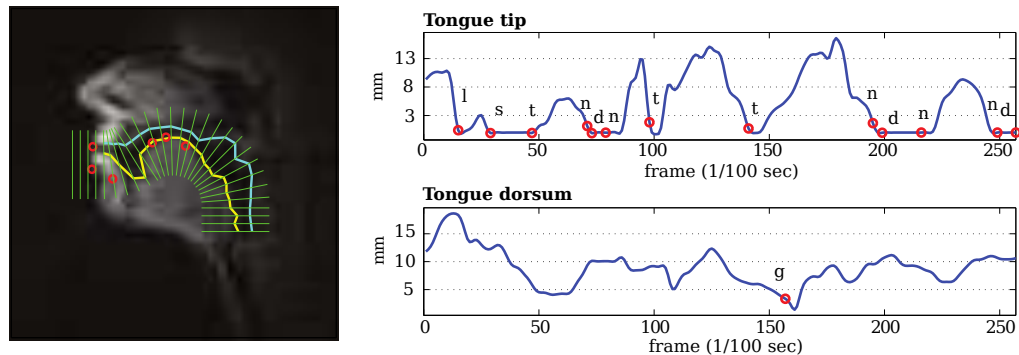


Figure 2.10: Left: Six EMA sensors (circles) overlaid on MRI image with estimated vocal tract boundaries (outer and inner lines in the vocal tract) and grid lines after co-registration. Right: Constriction degrees of the tongue tip (top plot) and tongue dorsum (bottom plot) extracted from upsampled rtMRI data for the sentence “Publicity and notoriety go hand in hand.” The circle for each phone is placed on the trajectory of the critical articulator of the phone, indicating the frame index for the phone in the registered data.

frames. The left-most plot in Fig. 3 shows an example plot of an MR image overlaid with EMA sensors (circles in the plot). Sample MRI videos with up-sampled rtMRI data, vocal tract tissue boundaries, overlaid EMA sensors before and after spatio-temporal alignment can be found at the ‘Demo video’ section in http://sail.usc.edu/old/software/Registration_EMA_rtMRI. The right-most plot in Fig. 3 illustrates the estimated constriction degrees of two landmark points (tongue tip and tongue dorsum) extracted from upsampled MR images for the sentence “Publicity and notoriety go hand in hand.” The mean of start and end times of each phone is indicated by a circle, where the phonetic boundaries were estimated by an adaptive speech-text alignment tool, SailAlign [Katsamanis et al., 2011], followed by manual correction. The constriction degrees from rtMRI were computed by measuring the Euclidean distance between upper (outer line in the vocal tract in the left-most plot in Fig. 3) and the lower (inner line) air-tissue boundary points on the closest vocal tract grid line to each EMA sensor of the tongue. Note that the tongue tip sensor is usually placed about 5 mm behind the anatomical tongue tip for minimizing its interference on its natural movement. For a rough comparison, the tongue tip constriction degree was measured on the next grid line anterior to the closest grid line to the superimposed tongue tip sensor position. Air-tissue boundaries were determined using a MATLAB-based software [Proctor et al., 2010] for analyzing rtMRI data. The right-most plot in Fig. 3 suggests that the estimated landmarks in the registered data capture the closure gestures of the tongue tip and the tongue dorsum well.

2.2.8 Discussion

This study includes two alignment tasks, spatial alignment and temporal alignment. The performance of our temporal alignment technique does not rely much

on spatial alignment. JAATA formulation does not use spatial alignment information directly. Even if we transform EMA sensor coordinates by spatial alignment before using them in JAATA, the temporal alignment performance may not change much. This is because the optimum spatial alignment parameter of rotation (θ^*) is small. However, the detailed information offered by spatial alignment, i.e., precise geometric relation between EMA sensor trajectories and the whole vocal tract in MRI could be beneficial for other speech production research problems.

Figure 2.8 shows that the temporal alignment of JAATA while promising, still has alignment error. Also, the temporal alignment of MRI and EMA recording using joint acoustic articulatory features improves APD for some sentences but decreases for others. This could be due to the temporal sparseness of articulatory information in rtMRI data. The frame resolution of rtMRI image is about 43 msec/frame, and the APD of temporal alignment using acoustic features is 50 msec. Therefore, the information gain for temporal alignment by incorporating articulatory features on top of acoustic feature might be limited. Error in manual phone boundary correction could be another possible reason for the limited performance of JAATA.

We have also investigated the benefit of using a subset of EMA sensors in temporal alignment using forward sensor selection approach. This was done by varying q (in eqn. 2.4) over a subset of sensor indices instead of all 12 EMA trajectories. The APD value was used to select the best EMA sensor trajectory in each iteration of forward selection approach. The lowest value of APD (44.106 msec) was achieved with $A=30$ and ULx, ULy, LLx, LIy, TDy trajectories. Thus, there was no significant benefit in APD by using forward sensor selection compared to using all sensor trajectories.

Finally, we tested the spatio-temporal alignment performance using CTW [Zhou and De la Torre Frade, 2009] on our corpus. Identical to JAATA evaluation, CTW performance is also measured by APD for each sentence. Articulatory features used in CTW are the direct 12 EMA sensor trajectories and the MRI image pixels (in the blue bounding box in Figure 1(a) for feature reduction without losing surface movements of articulators in the vocal tract). The mean (\pm SD) APD across all 20 sentences of CTW is 93.143 msec (\pm 56.026 msec), when CTW is initialized with uniform time warping [Fu et al., 2008] (the default initialization method of CTW). For fair comparison with JAATA, we also initialized CTW by DTW with MFCC. The mean APD of DTW with MFCC is 50.101 msec (\pm 40.659 msec). With MFCC based initialization, the APD of CTW with only articulatory data is 60.731 msec (\pm 39.427 msec). It indicates that CTW with articulatory data does not improve temporal alignment on top of MFCC based initialization. When both MFCC and articulatory data are used in CTW, the mean APD becomes 50.229 msec (\pm 40.617 msec). This result is worse than that of JAATA - 44.198 msec (\pm 19.949 msec) - which uses MFCC and automatically extracted articulatory features. This performance benefit suggests that the proposed JAATA formulation results in better temporal alignment performance. Additional benefit of JAATA is that it provides “interpretable” EMA-like articulatory features from MRI video.

2.2.9 Conclusions and future works

The goal of this study is to obtain spatial and temporal alignments of multi-modal speech production data, specifically MRI and EMA in order to gain the advantages of both types. For spatial alignment, we aligned the coordinates of EMA data to MRI images successfully by a grid search of estimated EMA palate tracking. For temporal alignment, we propose a novel algorithm, called JAATA,

which combines DTW-based temporal alignment with optimum articulatory feature extraction from MRI video. This technique also generates the best MRI image regions from which the EMA-like articulatory features are extracted for optimum alignment. We observed the benefits of using this technique experimentally using data from MRI and EMA articulatory corpora of English TIMIT sentences spoken by the same talker. Experiment on 20 sentences' data shows that JAATA reduces mean APD value from 50.101 msec (acoustic only alignment) to 44.198 msec, which is 12% improvement. Although results are reported on 20 sentences, the alignment algorithm developed in this work can be readily applied on all the sentences from MRI TIMIT and EMA TIMIT corpora.

Spatially and temporally aligned EMA and MRI data can assist speech production research by combining the advantages of both modalities. On top of the illustrated benefits of each articulatory measurement modality in Section 2.2.7, another possible advantageous combination would be to substitute the clean speech audio collected in conjunction with EMA data for the degraded rtMRI audio after temporal alignment. In addition, it may also be possible to reconstruct the tongue contour from EMA sensors, as Qin and Carreira-Perpinan [2010], by learning the statistical relationships between the EMA sensor positions and the midsagittal contours visible in rtMRI. The aligned data can also be used to extract articulatory features for subsequent modeling, including for automatic speech recognition [King et al., 2007] and speaker verification [Li et al., 2015] which use speech production knowledge.

The temporal alignment of EMA TIMIT and MRI TIMIT still has room for improvement. Although it was tested with more data in [Kim et al., 2014c], the robustness of JAATA against additional sources of intra- and inter-speaker variability needs to be examined. For example, the effects of the variation in speaking rate

and style (e.g., casual v.s. formal) need to be examined. Another future direction is to continue improving the proposed alignment techniques. For example, more flexible specifications (size, shape, numbers) of ROI selection might generate articulatory features leading to better alignment. Although the mean pixel intensity of some rectangular windows in rtMRI images behaves similarly to certain EMA sensors, pixel-wise tracking in rtMRI could be even more similar. Finally, our co-registration approach is potentially applicable for datasets collected by other modalities, e.g., ultrasound. Selecting a subset of EMA sensors (for alignment) depending on the corresponding available articulatory information in ultrasound data or proper feature engineering are needed so that the articulatory features from the two modalities behave similarly. These are part of our planned future work.

Chapter 3

Vocal tract shaping of emotional speech

3.1 Introduction

The emotional state of a speaker affects his/her articulatory and voice-source controls, by which emotional information is encoded in speech acoustic signal. Compared to the studies about acoustic variations in the prosodic, spectral and/or glottal feature spaces, there are much fewer studies regarding articulatory variations in emotional speech. One possible reason is that the direct measurements of articulatory movements often require expensive data acquisition devices and advanced signal processing technologies, which is not commonly available. Despite the difficulties, there have been efforts to understand articulatory variations and underlying control mechanisms for emotion expression. For example, Erickson et al. [2004, 2006], Lee et al. [2005] found postural variations of anatomical articulatory points, e.g., the tongue tip, the tongue dorsum, the jaw and the lips, using EMA data. Kim et al. [2011a, 2012a, 2015c] also found kinematic variations as well as postural variations of articulators in emotional speech, using EMA data. These preliminary studies suggest that emotional state of the speaker is reflected in the articulatory movements.

Recently, more comprehensive articulatory information than what EMA offers has been used to investigate articulatory variations during emotion expression. One

of such comprehensive articulatory data is upper airway image sequences recorded by real-time Magnetic Resonance Imaging (MRI) technology [Narayanan et al., 2004]. The real-time MRI is non-invasive articulatory data acquisition method that offers the complete view of dynamic vocal tract shaping in a plane, typically the mid-sagittal plane. Hence, Magnetic Resonance (MR) images are capable of providing spatially richer information of the vocal tract shaping than EMA. Using real-time MRI data of emotional speech, Lee et al. [2006] reported preliminary findings on the variations of the vocal tract shaping. Specifically, it was observed that vocal tract shape parameters, e.g., movement ranges in the pharyngeal region, and the vocal tract length, which are not measurable by using EMA, also vary depending on emotions. These findings were, however, obtained from limited amount of data from a single male speaker.

The present study investigates articulatory variability in real-time MRI data of ten speakers from the recently collected USC-EMO-MRI corpus. The USC-EMO-MRI corpus is a novel multimodal database of emotional speech, comprising Magnetic Resonance (MR) video data (sequences of upper airway images with synchronized speech audio after noise reduction) and perceptual evaluation results of speech emotion quality. This corpus is designed to serve as a resource in the context of diverse speech production studies, addressing, for example inter- and intra-speaker variability of the vocal tract shaping, resultant acoustic variations, and computational modeling of emotional speech production. Kim et al. [2014e] provided a brief summary of the corpus and reported preliminary analysis of articulatory variation across emotions, using a part of a single speaker’s data from the database. As an extended-version of the preliminary report, the goal of the present study is three-fold: (i) to provides the details of the USC-EMO-MRI corpus which are made publicly and freely available, (ii) to re-visit the preliminary findings of

articulatory variability in the literature, using data from multiple speakers, and (iii) to discover novel emotion-dependent variability in the vocal tract shaping.

The present study analyzes the vocal tract shape in terms of a distance function and the vocal tract length. In this study, the distance function refers to the collection of the Euclidean distances between inner and outer tissue-airway boundaries in the oropharyngeal vocal tract as a function of the distance from the lips. These vocal-tract-shape parameters will be automatically computed using a novel segmentation algorithm that we have developed [Kim et al., 2014b]. This algorithm performs tracking of the lips and the larynx, and detection of the tissue-airway boundaries in vocal-tract grid lines that are systematically spread over the oropharyngeal vocal-tract space [Öhman, 1967, Story, 2009, Proctor et al., 2010]. The details of this algorithm are provided in Section 3.3.1. The vocal tract length is computed based on the locations of the lips and the larynx, and the tissue-airway boundaries, which are discussed in Section 3.3.3.

The present study employs Principal Feature Analysis, or PFA [Lu et al., 2007] for a compact representation of the distance function. The distance function is generally redundant and highly correlated due to the physiological constraints of the vocal tract (e.g., smooth shape of the surface) and the coordinated controls of speech articulators. Hence, reduced set of parameters driven by decomposition techniques, e.g., Principal Component Analysis (PCA) and Fourier series, are often used for analysis and modeling of the vocal tract [Liljencrants, 1971, Harshman et al., 1977, Story et al., 1996, Story and Titze, 1998, Mokhtari et al., 2007, Cai et al., 2009]. These methods transform the initial parameters to a low-dimensional, compact parameter space in which the behavior of each component is often difficult to interpret. Also, the influence of the variation in a hidden parameter to the (interpretable) initial parameters is mostly complex, which makes it difficult to

analyze local variability of the vocal tract shaping. In contrast, PFA allows us to select most variable and least redundant locations in the vocal tract for each speaker, hence emotion-dependent variation captured in the compact feature set is easily interpretable in terms of its behaviors as a function of emotion. Analyzing most variable vocal-tract locations is important for the present study, because articulatory movement range varies depending on emotion [Lee et al., 2005, Kim et al., 2010, 2015c]. Using the compact set of distance values also allows efficient computation for temporal alignment which is needed for standardizing vocal tract shape across multiple utterances. The details of PFA are provided in Section 3.3.2.

This chapter is organized as follows: Section 3.2 describes the USC-EMO-MRI corpus. Section 3.3 describes the ways to compute vocal tract parameters from the real-time MRI data. Section 3.4 reports analysis results on the emotion-dependent variations of the vocal tract shaping in terms of the principal features and the vocal tract length. Section 3.5 provides discussion.

3.2 The USC-EMO-MRI corpus

The present study uses the USC-EMO-MRI corpus which comprises a real-time MRI data and corresponding speech audio from five female and five male speakers, and categorical emotion labels for each individual utterances. All speakers have had earlier professional acting experience and theatrical vocal training.

3.2.1 Speech stimuli

The speakers were recorded reading a short passage and a small set of sentences, shown in Table 3.1, while targeting to enact the three basic emotions of happiness, sadness and anger, or being in a neutral emotional state. The set of sentences was

Table 3.1: List of sentence prompts for the USC-EMO-MRI corpus. ‘Index’ refers to the sentence index.

Index	Prompt
1	John bought five black cats at the store.
2	The Leopard, skunk and peacock are wild animals.
3	Charlie, did you think to measure the tree?
4	The queen said the KNIGHT is a MONSTER.
5	Hickory dickory dock, the mouse ran up the clock. Hickory dickory dock.
6	9 1 5 (short pause) 2 6 9 (short pause) 5 1 6 2.
7	Ma Ma Ma (short pause) Ma Ma Ma (short pause) Ma Ma Ma Ma.

designed to investigate the effects of emotion expression on syntactic, prosodic, and rhythmic structure, including reiterant speech [Kelso et al., 1985].

In particular, speakers read the passage in normal speaking rate for all four emotions (including the neutral emotion) and additionally in fast rate only for the neutral emotion. They also read six sentences in seven repetitions for each of the four emotions, all in normal speaking rate. The order of presentation of the sentences was randomized in each repetition. For some subjects, a seventh “nonsense” sentence was added. That sentence was always presented right after the sixth sentence and speakers were asked to read it with the same intonation. When reading the fourth sentence, speakers were asked to emphasize the uppercased words in the stimuli, such as “KNIGHT” and “MONSTER.”

3.2.2 Data acquisition and processing

We used the MRI data acquisition and processing protocols of the USC-TIMIT database [Narayanan et al., 2014]. This section offers the summary of the acquisition and processing protocols. See [Narayanan et al., 2014] for technical details.

We collected upper airway MR images at the Los Angeles County Hospital using a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI). We used

a custom 4-channel receiver coil array, with two anterior coil elements and two coil elements posterior to the head and the neck. We recorded the MRI data of speakers while they lay supine in the scanner and read the stimuli.

The real-time MRI acquisition was performed using a spiral fast gradient echo sequence, where thirteen interleaved spirals form a single image. We used a sliding window technique [Narayanan et al., 2004] which allows view sharing, thus increases frame rate. The repetition time, or TR at data acquisition was 6.164 for each spiral, and the TR-increment for view sharing was 7 acquisitions [Narayanan et al., 2004, Bresch et al., 2008, Kim et al., 2011b]. Hence, MRI movies were generated with a frame rate of 23.18 frames/sec ($= 1 / (7 \times 6.164 \text{ msec})$). The field of view of imaging was 200×200 mm, and image resolution was 68×68 pixels (2.94×2.94 mm for each pixel). We used RTHawk (HeartVista, Inc., Los Altos, CA), which is a custom real-time imaging platform [Santos et al., 2004], for the scan plane localization of the mid-sagittal slice. We recorded speech audio at a sampling frequency of 20kHz, simultaneously with MR imaging, using a custom fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and a custom recording setup; the unblank TTL signal from the scanner, which is generated at the beginning of each MRI acquisition, triggers audio data recording.

During post-processing, we performed noise cancellation on the speech audio using a custom adaptive signal processing algorithm [Bresch et al., 2006], followed by the synchronization of the MRI video and speech audio.

3.2.3 Evaluation of emotion quality

Perceptual evaluation tests were performed to assess the emotional quality of the recorded data, i.e., how well the intended emotion (by speakers) is expressed in speech audio. At least ten evaluators tested each speaker’s data. Some of the

Table 3.2: Summary of evaluation results of all evaluators. ‘Sentences’ indicates the sentence ID included. ‘Average’ and ‘STD’ denotes average and standard deviation of the matching ratio (%) between target emotion and the final emotion label for sentence-level utterances, respectively.

	Subject ID (M: male, F: female)									
	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
# Evaluators	10	10	11	10	11	12	12	12	12	10
Sentences	1-6	1-7	1-7	1-7	1-7	1-6	1-6	1-6	1-7	1-7
Average	85.3	69.5	82.6	72.0	80.5	80.7	94.0	89.8	86.5	80.5
STD	9.9	11.8	8.5	11.1	10.0	11.1	5.4	8.4	8.2	11.8

evaluators were also actors or actress who have participated in the data collection. After listening to the recorded audio for each utterance, the evaluators were asked to: (i) choose the emotion that they perceive from the spoken sentence (neutral, anger, happiness, sadness, or other); (ii) judge their confidence in their choice; and (iii) judge the strength of the emotion expression. Confidence and strength were evaluated on a five-point Likert scale.

Table 3.2 presents the number of sentences recorded from each speaker of the database, the number of evaluators of each speaker’s data, and the average and standard deviation of the matching ratio between target emotion (intended emotion by the speaker) and perceived emotions (evaluated by listeners). The matching ratio counts the number of the utterances whose target emotion and the final emotion label match over the number of all utterances, where the final emotion label is determined by majority voting across all evaluators.

The USC-EMO-MRI corpus is publicly and freely available for research purposes at <http://sail.usc.edu/span/usc-emo-mri> with real-time MRI video (MR image sequences and speech audio) and emotion evaluation result for each utterance.

3.3 Methods

3.3.1 MR image parameterization

A tool for the automatic tissue segmentation in real-time MRI images has been developed for the purposes of this study. Kim et al. [2014b] presented an earlier version of the implemented method and its robustness on the segmentation task, compared to previous work [Proctor et al., 2010]. This method seeks pixel intensity thresholds distributed along tract-normal grid lines and defines airway contours constrained with respect to an estimated airway path from the glottis to the lips.

The method is initialized with a manually drawn reference line, roughly following the airway path, together with manually set points. The manually set points correspond to the midpoint of the lips, the highest point on the palatal surface, and the upper visible boundary of the arytenoid cartilage. The tool uses the arytenoid cartilage instead of the glottis, because the glottis is not well visible in the MR images. The reference line is smoothed using the discrete cosine transform technique. This initialization is done once; the upper boundary of the arytenoid cartilage in each frame is detected automatically, and so is the exit of the lips.

A set of equidistance grid lines is constructed automatically, perpendicular to the reference line. The first of these grid lines is located at the (automatically detected) upper boundary of the arytenoid cartilage and the last at the exit of the lips. A frame-specific airway path from the first to the last grid line is then determined, using dynamic programming algorithm, followed by spatio-temporal smoothing. Then, for each grid line two tissue-airway boundaries are determined as the first pixels whose intensity is above a certain threshold, along the grid line and in toward the two grid line edges starting from the intersection of the airway path and the grid line.

The method was applied on all non-silent regions of the previously described corpus. Silent regions were determined using the SailAlign [Katsamanis et al., 2011] which is a Hidden Markov Model based adaptive speech-to-text forced aligner. The method was initialized whenever significant differences in head posture or morphological structure were observed (e.g., different speakers and emotional states). The

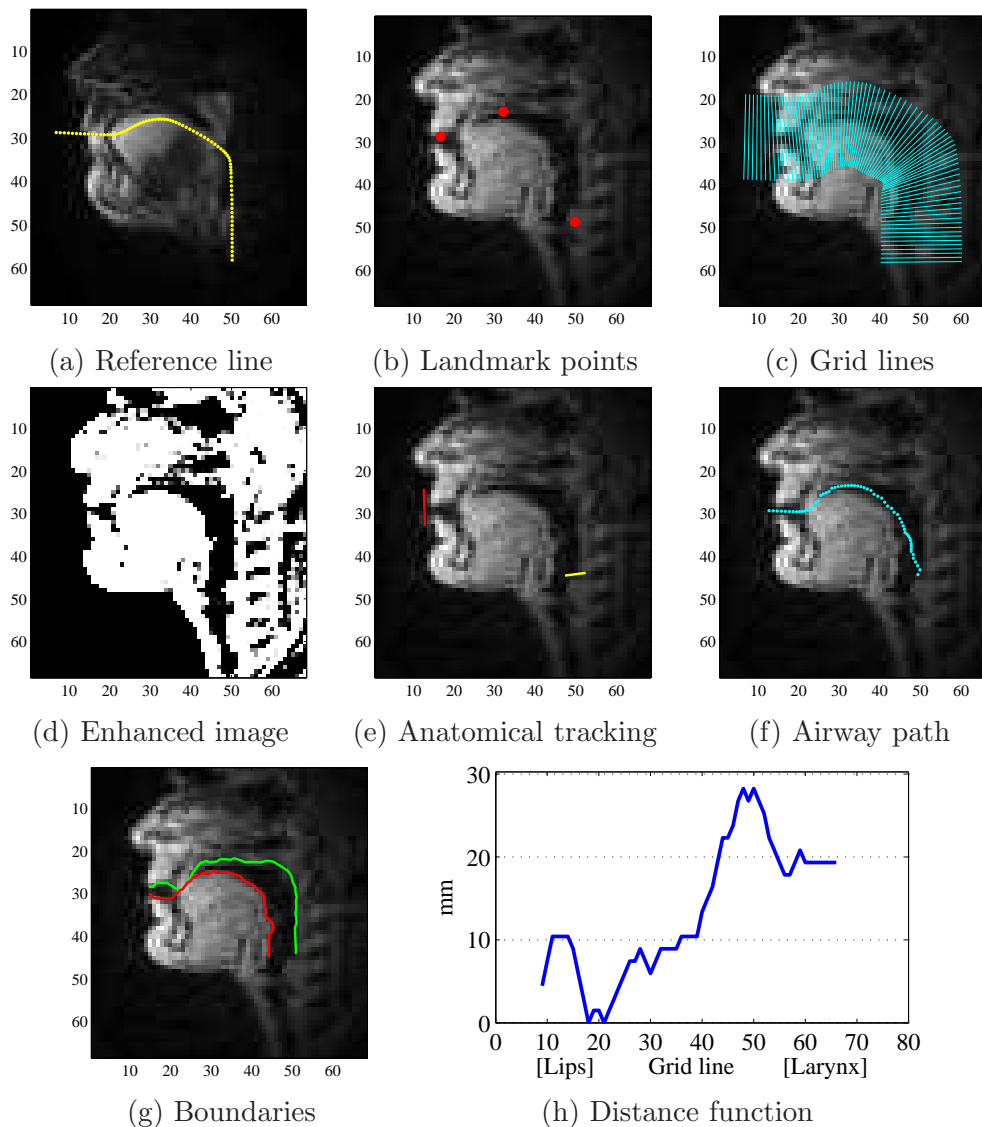


Figure 3.1: Results of parameterization processes of a magnetic resonance image (speaker M1) as an example

Euclidean distances between the two airway-tissue boundaries on each grid line were measured. Figure 3.1 illustrates the outputs of the procedures for MR image parameterization. The final vocal tract parameter computed from the distances (aka distance function), using PFA, is discussed in the following section.

3.3.2 Principal feature analysis

For a compact representation of the vocal tract shaping, we perform feature reduction using PFA [Lu et al., 2007]. This method selects a subset of the original features (i.e., distances between outer and inner boundaries on the grid lines), using the same feature reduction criteria as PCA. Sample points are maximally spread in the selected features, where the structure of the principal components is retained, thus preserving the variation of the original data. Hence, PFA offers a compact, effective and interpretable representation of the vocal tract shaping.

We computed the principal features of individual speaker’s data as follows: (1) Compute eigenvectors and eigenvalues from the covariance matrix of distance functions. (2) Choose the minimum number of dimension q for the subspace, where the cumulative sum of eigenvalues for the dimensions is greater than 90% of the total sum. (3) Perform k -means clustering on the row vectors of the subspace eigen matrix. In our case, k was 3 – 7 greater than q for retaining 90% of the total variation in the subspace. (4) Find the row vector corresponding to the mean of each cluster, where the index of the row vector becomes the index of the selected original feature. See [Lu et al., 2007] for full details of this algorithm. Figure 3.2 shows the subset (principal features) of the grid lines overlaid on an arbitrarily chosen midsagittal image of the corresponding speaker. There are more principal features in the front oral cavity (lips to the hard palate) than pharyngeal region, implying more complex movements of the tongue in the front oral cavity than the

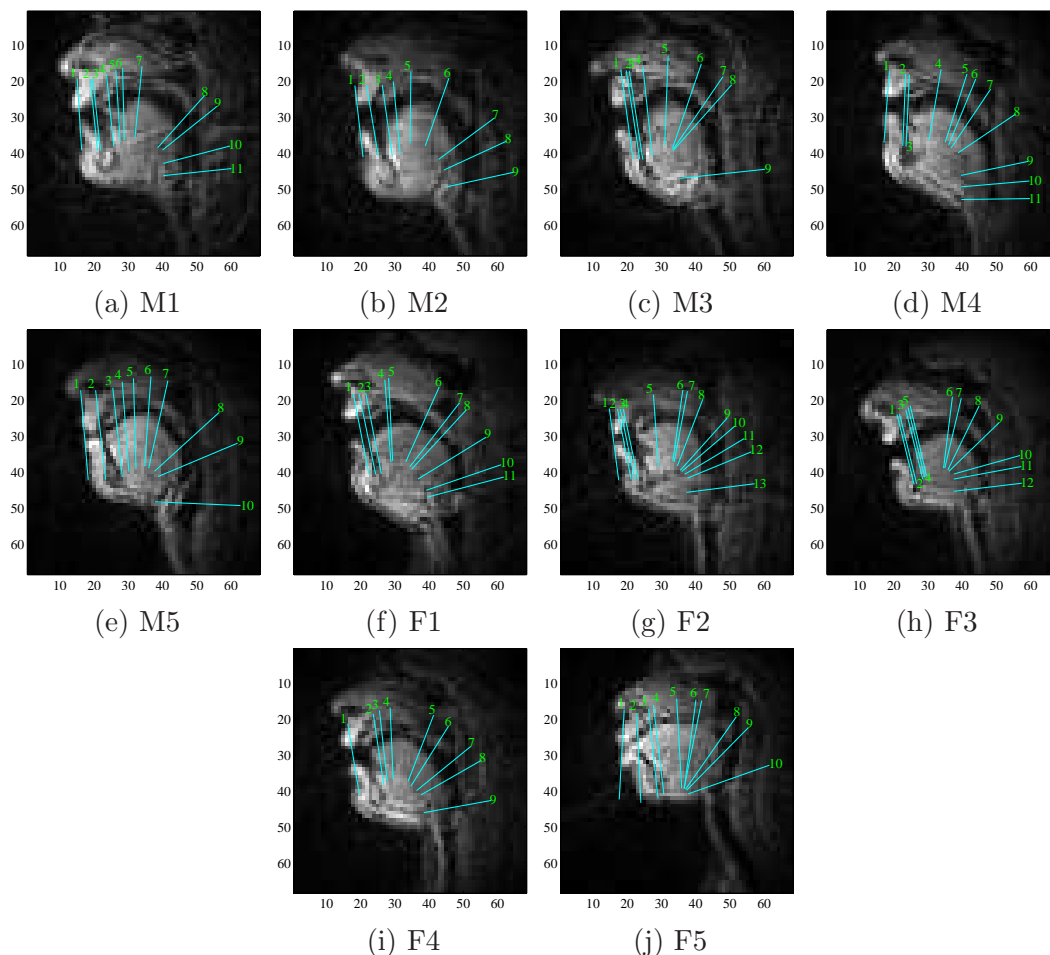


Figure 3.2: Grid lines (principal features) overlaid on an MR image for each speaker. The indices of the principal features are noted next to the corresponding grid lines

tongue in the pharyngeal region. Some principal features are sometimes clustered closely, especially near the alveolar ridge and the teeth. It is speculated that the delicate movements of the tongue tip causes neighboring grid lines in the region to be less correlated.

We performed within-speaker analyses, using data of the sixth utterance, “Nine one five, two six nine, five one six two (915 269 5162).” In order to compare time series of principal features for different emotions, we performed temporal alignment of each utterance to a reference utterance (arbitrarily selected utterance of neutral

emotion) using dynamic time warping algorithm [Sakoe and Chiba, 1978]. Finally, we computed averaged time series for each emotion, which was used to specify the analysis of emotion-dependent variation in vocal tract shaping along the midline of the vocal tract.

3.3.3 Computing the vocal tract length

The computation of the true vocal tract length requires the location of the lips (the initial point of the vocal tract), the location of the true vocal fold (the final point of the vocal tract), and the center line between outer and inner tissue-airway boundaries in the vocal tract. We selected the grid line of the smallest distance in the lip region for the initial point of the vocal tract. The true vocal fold is not captured in the MR images, but the arytenoid cartilage in the larynx is well observed. Hence, we selected the grid line of the top of the arytenoid cartilage for the final point of the vocal tract. We computed the geodesic distance (the sum of the Euclidean distance between the center points of adjacent grid lines) within outer and inner boundaries from the initial point to the final point, which is considered as an approximation of the vocal tract length in the context of this study. In order to minimize the variability by speaking rate, the time-series describing the dynamics of the vocal tract length were aligned using the alignment map created in Sec 3.3.2.

The MATLAB tool that computes distance function, principal features, and the vocal tract length from MR images is freely available at http://sail.usc.edu/old/software/rtmri_seg

3.4 Results

3.4.1 Emotional variations of principal features

This section discusses emotion-dependent variations of the vocal tract shaping. As an initial investigation, we compared the averaged time series of principal features for each of the four emotions. Figure 3.3 illustrates the averaged time series of two principal features as examples, using data of speaker M1.

Overall, the time series clearly show differences depending on emotions. For the first principal feature (located at the lips) in Figure 3.3 (a), the distances of anger and happiness are often greater than those of neutrality and sadness. It is

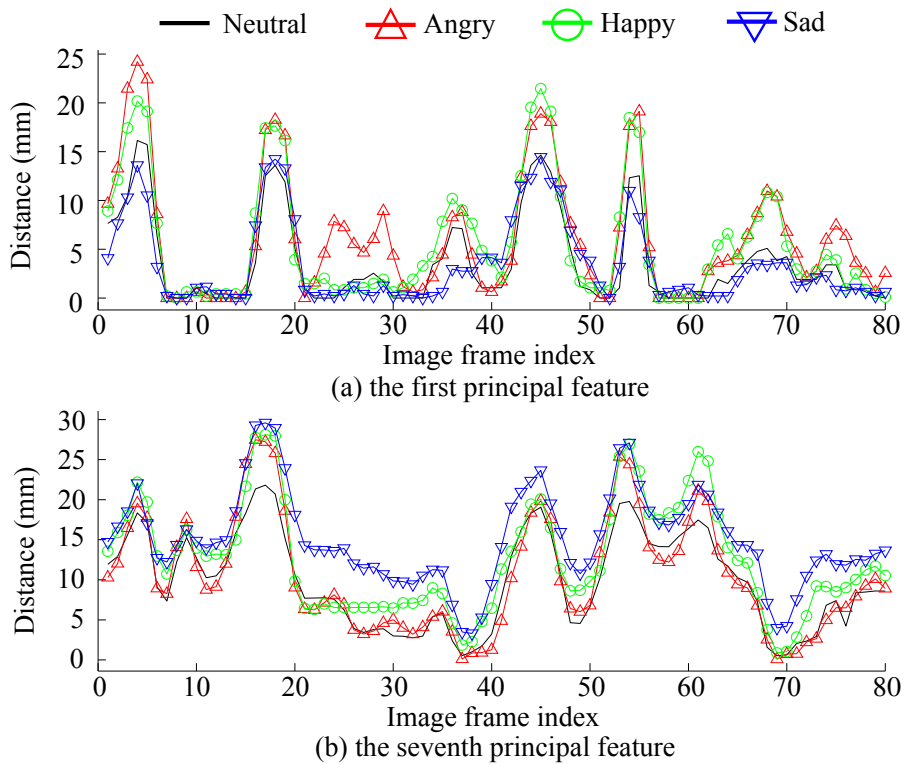


Figure 3.3: Averaged time series of the first and the seventh principal features for each emotion. The averaged time series were temporally aligned. The utterances of sentence 6 “nine one five two six nine five one six two” are used.

noted that in the USC-EMO-MRI corpus, anger and happiness are high arousal emotions, while sadness is low arousal emotion [Kim et al., 2014e]. Hence, this indicates that speaker M1 shows positive correlation between the degree of his lip opening and the arousal dimension of emotion. For the seventh principal feature (located on the hard palate) in Figure 3.3 (b), the distances for sadness is mostly greater than those for the other emotions, especially from the 20-th frame to the 35-th image frame. This region corresponds to the words “two six” that include high vowels /u/ and /i/. This suggests that the vertical constriction gesture of the tongue dorsum tends to be less strictly controlled for sadness, compared to the other emotions, and this pattern is more significant when the speaker utters the high vowels. Interestingly, happiness and anger show clearly different patterns in the range of the first principal feature for /u/, that is located near the image frames 23 and 75, while they show similar patterns in the seventh principal feature.

Motivated by the emotion-dependent patterns observed in the time series of principal features, we analyzed emotion-dependent variations relative to neutrality for all principal features for each speaker. First, we computed statistics ([0.1, 0.5, 0.9] quantiles, and 0.9 quantile – 0.1 quantile) of each principal feature for different emotions. [0.1, 0.5, 0.9] quantiles reflects constriction degree for high vowels and consonants, constriction degree for offset positioning and middle vowels, and constriction degree for low vowels, respectively. 0.9 quantile - 0.1 quantile reflects the movement range. Next, we computed relative statistics by subtracting the statistics of each emotion from that of neutrality. It is noted that we computed [0.1, 0.5, 0.9] quantiles instead of minimum, mean and maximum in order to minimize the effects of (possible) tissue-airway segmentation errors on this analysis.

Figures 3.4 illustrates emotion-distinctive behaviors captured in all principal features in terms of the four statistics. The plots in the left and the right columns

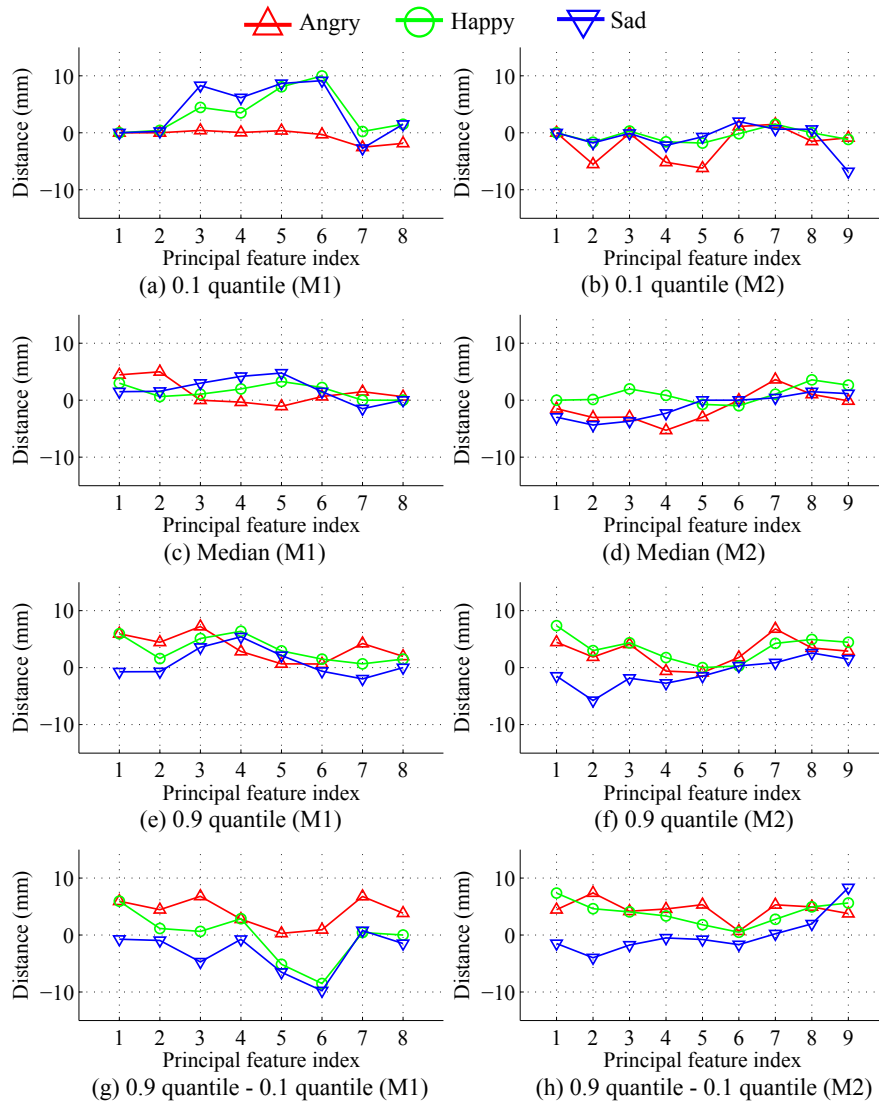


Figure 3.4: Quantiles and quantile range of principal features for anger, happiness and sadness relative to neutrality in data of speakers M1 and M2

in Figure 5 correspond to the results of speaker M1 and speaker M2, respectively. Although plots of the two speakers are provided in this paper, emotion-dependent patterns based on all ten speakers will be discussed.

Results show different control strategy of the vocal tract depending on emotions with respect to the neutrality. Here, we will discuss both speaker-dependent and

speaker-independent variation patterns in the vocal tract shaping. First, as shown in Figure 3.4 (a, b), most speakers tend to show similar or smaller constriction degree (tighter constriction) for anger than sadness, although such regions in the vocal tract vary depending on speakers. For example, such regions are observed in the front oral cavity (principal features 2, 4 and 5) for M1, while the tendency is significant in the pharyngeal region (principal features 5, 6 and 7) for F4. Compared to neutrality, the tighter constrictions for anger are significant for speakers M2, F4 and F5; looser constrictions for sadness are significant for speakers M1, M3, M5, F1, F2 and F3; both are significant for speaker M4. Second, as shown in Figure 3.4 (e, f, g, h), most speakers tend to show greater movement range and larger opening for high arousal emotions (anger and happiness) than neutrality in most of the vocal tract regions. In contrast, sadness tends to show smaller opening for the low vowels than the high arousal emotions in the front cavity more consistently than the other regions of the vocal tract. Finally, although offset vocal tract shaping shows emotion-distinctive patterns as shown in Figure 3.4 (c, d), speaker-independent patterns are not found.

3.4.2 Emotional variations of the vocal tract length

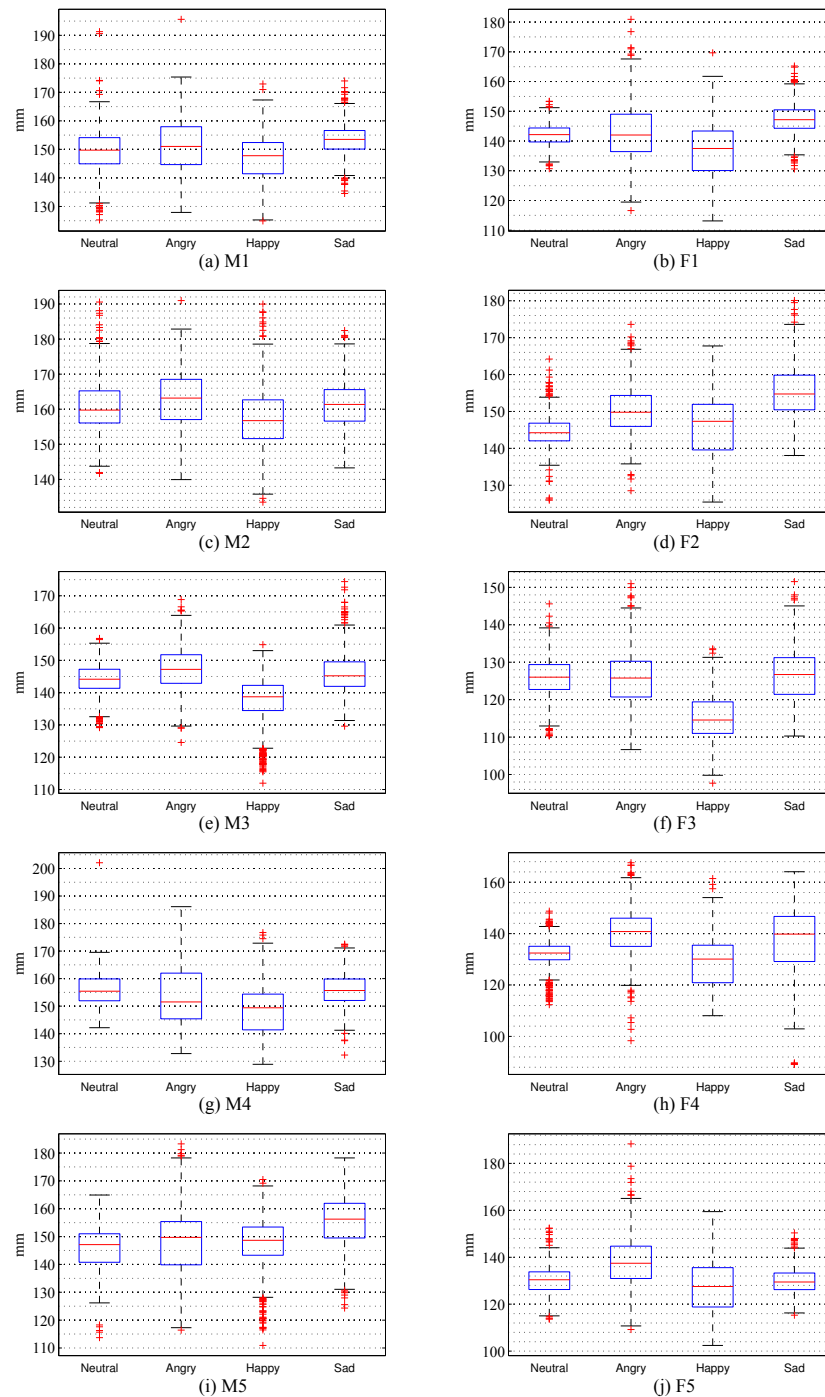


Figure 3.5: Boxplots of the vocal tract length of each emotion

Figure 3.5 shows boxplots of the vocal tract length for each emotion and each speaker. We observed that the vocal tract length also varies depending on emotions. Specifically, the vocal tract length tends to be shorter for happiness than anger or sadness across the ten speakers. We conducted one-tailed Welch’s t -test on the hypothesis that the mean of the vocal tract length for happy speech is shorter than the mean for angry/sad speech. Results indicates that on average, happy speech shows statistically significantly shorter vocal tract length than angry speech at $\alpha = 5 \times 10^{-6}$ level in general, except for M5 (t -statistic = 1.64, $p = 0.05$). Also, happy speech shows statistically significantly shorter vocal tract length than sad speech for all speakers’ data ($p < 0.005$).

3.5 Discussion

Based on new USC-EMO-MRI database and the MR image tracking software, the present study examines how the vocal tract shape changes depending on emotions, using automatically extracted vocal tract parameters which describe the distance between the inner and outer vocal-tract boundaries and (approximate) vocal tract length.

First, the present study provides supporting evidence for previous findings by Lee et al. [2006] for emotion-dependent variation patterns, using lexically richer data from the ten speakers in the USC-EMO-MRI corpus. Lee et al. [2006] found that anger shows wider opening in both oral cavity (the front side of the vocal tract) and pharyngeal region (the back side of the vocal tract) than neutrality. The result of the wider opening in the oral cavity for anger than neutrality was consistent with Lee et al. [2005], Kim et al. [2010, 2011a], where the movement range of a sensor attached on the tongue tip was analyzed. The present study found that

high arousal emotions (both happiness and anger) show greater movement range than neutrality, but the vocal tract region of significantly contrasted opening varies depending on speakers.

Also, the pattern that happy speech shows shorter vocal tract length than angry, neutral and sad speech is consistent with Lee et al. [2005]. Simulation experiments by Xu and Chuenwattanapranithi [2007] have suggested that the dynamic variations of the vocal tract length (and F0 jointly) are often perceived as expression of joy or anger, but depending on vowels. The present study provides evidence that relationship between the vocal tract length and emotion quality holds for continuous speech as well; this was consistent across speakers. In addition, lip spreading and larynx elevation are important factors contributing to the decreasing of the vocal tract length for happy speech. The relative contributions of these factors have been explored by Lasarcyk and Trouvain [2008], jointly with F0 raising, using isolated synthetic vowel sounds. That study found that lip spreading and laryngeal elevation often affect perceptual emotion quality in the dominance dimension. Similar simulation experiments in terms of perception of happiness would also be useful to understand the influence of the individual factors to emotion expression. We do not include simulation experiment in the present study due to the difficulty of robust spectral feature extraction from the speech audio in the USC-EMO-MRI corpus; although speech intelligibility is much enhanced by post-processing in Sec. 3.2.2, it still suffers from residual noise and/or post-processing artifact.

The present study also reports a novel finding that when low vowels are produced, sadness shows the smaller opening in the front oral cavity than anger and happiness. The articulatory characteristics for sadness have been studied by Erickson et al. [2004, 2006], where stronger constriction gesture for a high vowel /i/ was

observed in sad speech than neutral speech. However, such stronger lingual gestures for sad speech are not consistent for all speakers in the USC-EMO-MRI corpus. For example, Figure 5 (a) shows weaker constriction gestures in the oral cavity for sad speech than neutral speech. In fact, stronger lingual gestures for any particular emotion across all speakers were not observed. It could be speaker-specific, but more investigation is needed to understand further any such difference.

The vocal tract length is considered as an important morphological parameter that contains speaker-specific information. For example, Smith and Patterson [2005] found the interaction of vocal tract length is an important cue for differentiating speaker size, sex and age. The present study suggests that the vocal tract length is an important cue for differentiating the emotional state of the speaker. In fact, Kockmann et al. [2011] has reported that normalizing vocal tract length does not improve emotion recognition accuracy, although this normalization technique has been generally useful for reducing speaker variability [Lee and Rose, 1996]. However, in order to use (predicted) vocal tract length parameter as emotional cue, we need to understand how to decompose its variation into emotional and other speaker-dependent factors, e.g., age and gender. This is an open question that we will pursue in future.

Chapter 4

Articulatory variability, linguistic criticality, and emotion

4.1 Introduction

Previous studies have shown that the emotional states of speakers influence the acoustic and articulatory characteristics of their speech. While studies on the acoustic properties of voice quality and prosody of emotional speech abound in the literature as discussed in Section 1.2.2, there are considerably fewer studies about articulatory details of emotional speech, presumably due to difficulties in obtaining direct articulatory data. Although various data acquisition technologies have been used for the study of speech production, the data collection environment is not ideal for investigating natural emotion expression in speech. Nevertheless, it has been shown that distinctive emotional information is present in EMA data of *elicited* and *acted* emotional speech [Erickson et al., 2004, 2006, Lee et al., 2005]. In particular, it has been reported that the position and speed of articulators, particularly their properties at the syllable, word and utterance levels, are important emotional features [Lee et al., 2005, 2006, Kim et al., 2009, 2011a, 2012a]. Kim et al. [2010] reported empirical evidence for the interplay between prosodic characteristics and articulatory movements as a function of emotion.

Despite progress in understanding the articulatory aspects of emotional speech, there is still limited knowledge about the behavior of individual articulators for

achieving emotional goals in parallel with linguistic goals during speech production. The present study aims at investigating the relationship between the variability in the kinematic behavior of specific articulators for the achievement of specific linguistic and emotional goals during speech production. A better understanding of such details can shed further light on intra-speaker variability in speech production. Such knowledge can also benefit modeling and synthesis of emotional speech.

Linguistic criticality of articulators is an important factor to characterize articulatory variability during speech production. For achieving certain linguistic goals, (linguistically) critical articulators are more carefully controlled and display less variability, than non-critical articulators. For example, in producing /t/, it is essential that the tongue tip comes in instantaneous contact with the alveolar ridge, while the positions of the lips and the tongue dorsum may be more variable. In this case, the tongue tip is considered the critical articulator for /t/, while the lips and the tongue dorsum are considered non-critical articulators. The linguistic criticality of articulators can also be categorized based on the direction of movements, which depends on constriction locations. For example, horizontal constriction of the tongue body is critical for pharyngeal vowels, while vertical constriction is critical for palatal vowels.

The present study investigates the roles of non-critical articulators for emotional information encoding. Ananthakrishnan and Engwall [2008] reported that movement information from the critical articulators alone may be enough to almost fully encode the linguistic message. They showed that, from the viewpoint of linguistic encoding, the movements of non-critical articulators reflect temporal linguistic context, displaying interpolative motions of preceding and following critical points under vocal-tract physiological constraints. From the viewpoint of emotional encoding, we hypothesize that the motions of non-critical articulators may encode

emotion-distinctive information, in tandem with the aforementioned interpolative movements. In the present paper we test this hypothesis on the basis of the binary distinctions of articulator criticality in realizing speech gestures within the framework of articulatory phonology [Browman and Goldstein, 1989] and implemented in the task dynamics application model [Nam et al., 2004]. Further details of this setup is provided in Section 4.3.

We thus investigate the following three questions: (i) In what ways do the kinematics of critical and non-critical articulators vary as a function of emotions? (ii) Are the kinematics of non-critical articulators more emotion-distinctive compared to the kinematics of critical articulators? (iii) Is the emotion-dependent variability of non-critical articulators simply mechanical outcome of controls on critical articulators? In order to address the first and second questions we analyzed articulatory variability as function of emotion at the syllable and phone levels. At the syllable level, several static and dynamic articulatory parameters extracted from EMA data at manually labeled phone-target and transition points were analyzed using distribution plots. Phone-level analysis focused on understanding the variability of task-oriented articulatory trajectory formation in emotional speech. The emotional variation in articulatory position at phone targets was quantified by the average of centroid distance and mean dispersion, which represent the variability between emotions and within each emotion, respectively. Critical and non-critical cases are compared across phones using these two standardized measures. In order to address the third question, we compared the emotion-dependent postural variation of true and estimated articulatory trajectories, where the estimated trajectories were generated as a function of the movements of only critical articulators. High similarity of the emotion-dependent variation of the two trajectories (true

and estimated) implies considerable dependency of the emotional variation of non-critical articulators to controls of critical articulators.

This chapter is organized as follows: Section 4.2 describes methods for collecting, evaluating and processing articulatory data of emotional speech. Section 4.3 explains our binary categorization of articulators in terms of their linguistic task criticality. Sections 4.4 and 4.5 present our methods and results for the syllable level and phone level analyses, respectively. Section 4.6 presents our methods and results of the simulation experiment. Section 4.7 provides a discussion of the results.

4.2 Data

This study uses an articulatory database of emotional speech production collected at the University of Southern California. ElectroMagnetic Articulography (EMA)

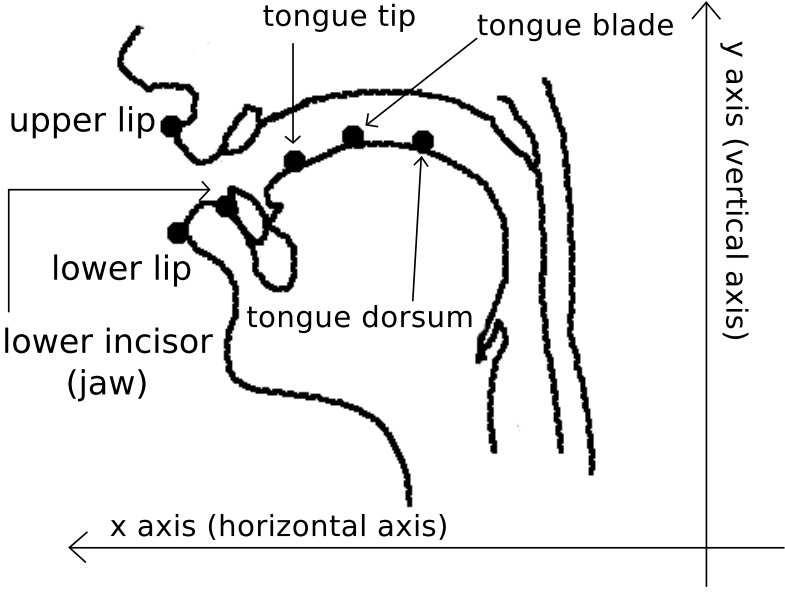


Figure 4.1: Placement of EMA sensors in the mid-sagittal plane

was used for the data collection as described in [Kim et al., 2011a]. The database includes speech waveforms, sampled at 16 kHz, and corresponding 3D coordinates of six sensors attached to oral articulators, sampled at 200 Hz.

Figure 4.1 shows the placement of the six sensors in the mid-sagittal plane. The six EMA sensors were placed on the tongue surface, lips and the lower incisor (jaw) as described in Section 1.2.1. Sensors were also glued on the upper and lower lips. Finally, a sensor was attached on the lower incisor for monitoring the movement of the jaw. The trajectories of the six sensors were recorded with a Carstens' AG500 EMA system. Three native speakers of American English, one male (referred to as SB) and two females (JR and JN) produced speech in five acted categorical emotions (neutrality, hot anger, cold anger, happiness, and sadness) and three speaking styles (normal, loud, fast). The speakers had previous training in theatre and acting. They were instructed to read four or five repetitions of seven sentences in every combination of the five emotions and the three speaking styles. The list of the sentences is:

- Say peep again? That's wonderful.
- It was nine one five two eight nine five seven six two.
- Say pop again? That's wonderful.
- I saw nine tight night pipes in the sky last night.
- Don't know how very joyful he was yesterday.
- Say poop again? That's wonderful.
- Native animals were often captured and taken to the zoo.

The order of the seven sentences was randomized at each repetition of data collection. There were 524 utterances for JR ($7 \text{ sentences} \times 5 \text{ emotions} \times 3 \text{ styles} \times 5$

repetitions - 1 erroneous recording which was discarded), 440 utterances for JN (7 sentences \times 5 emotions \times 3 styles \times 4 or 5 repetitions - 2 erroneous recordings) and 417 for SB (7 sentences \times 5 emotions \times 3 styles \times 4 repetitions - 3 erroneous recordings). Fast style utterances were excluded from analysis, because variation due to the intended speaking rate change was not within the scope of the present study. On the other hand, loud style utterances were included, because loudness of speech is an important factor of emotion expression and perception especially for distinguishing emotions in the arousal dimension [Kim et al., 2011a].

The 3D coordinates of the six sensor position data were corrected for head movement, and the orientation of articulatory trajectories was fixed to the occlusal plane. The orientation of the occlusal plane was measured using a half-rounded bite plane on which three sensors were attached. We use the projections of the EMA sensors on the (horizontal) x-axis and the (vertical) y-axis shown in Figure 4.1. Each raw articulatory trajectory (evolution of sensor position projected on the x- or y-axis) was smoothed by a 9th-order Butterworth filter with a 15 Hz cutoff frequency as in [Lee et al., 2005].

The emotion expressed in each audio utterance spoken by the three speakers of the database was evaluated by either four or five listeners, native speakers of American English and either undergraduate or graduate students at the University of Southern California (see http://sail.usc.edu/data/ema_eval_jr_short for the evaluation interface). For each speech audio, listeners were asked to choose (1) the best-representative emotion among six emotion categories (neutrality, hot anger, cold anger, happiness, sadness and other), (2) the degree of confidence in their evaluation and (3) the strength of emotion expression. Only speech audio was provided to the listeners in a randomized order, without showing any intended goal (loudness and emotion) of the speakers. They were asked to choose ‘other’

when none of the five given emotion categories was a good match to what they perceived. Confidence and strength were evaluated on a five-point Likert scale.

The most representative emotion for each utterance was determined by majority voting. If two emotions had the same evaluation scores, then the one with higher confidence score was chosen. The confidence of each evaluator was normalized by z-scoring across all utterances. Utterances that did not satisfy the majority voting criteria were discarded in order to maximize distinctive articulatory characteristics among the five categorical emotions for the present analysis. In the end, 312 utterances of JR's data, 281 utterances of JN's data, and 267 utterances of SB's data were used for analysis. It needs to be noted that the non-selected utterances are still important data for emotional speech research, since they may reflect ambiguous displays, and heterogeneity in judging emotions. For example, it would be important to understand what varies for emotional speech production and perception from one speaker to another because of individual differences in terms of gender, personality and prior experience.

Table 4.1 shows the confusion matrix among emotions perceived by the judges (i.e., the result of majority voting) and emotions intended by the speakers (target emotions). The matching ratios between intended and evaluated emotions are comparable to previous studies [Grimm et al., 2007, Shami and Verhelst, 2007]. In addition, we observed a significant degree of confusion between hot anger and cold anger across all speakers, indicating that the two emotions are quite similar in terms of perception and/or expression. Loudly spoken (intended) neutral speech was often perceived as cold or hot anger, especially for JN's data. Also, judges had a preference for hot anger to cold anger in loud style speech. These are in line with the observations of previous studies, e.g. [Kim et al., 2011a], which noted

Table 4.1: Confusion matrix between evaluated emotion determined by majority voting and target emotion of speakers. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. The numbers in bold are the greatest of evaluated emotion cell for each target emotion, each speaker and each intended style.

			Evaluated emotion														
			JN					JR					SB				
Target emotion	Style	Emo	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
		Target emotion	Normal	Neu	28	0	2	0	4	33	0	1	0	0	28	0	0
Han	0			13	12	4	0	1	10	20	0	0	4	0	12	1	1
Can	0			1	28	0	1	0	8	18	0	1	0	8	18	0	1
Hap	1			0	0	33	0	0	0	0	23	10	7	0	1	20	0
Sad	0			0	2	0	28	0	0	0	0	35	0	0	0	0	28
Loud	Neu		5	8	15	0	0	32	0	1	0	0	26	0	1	0	1
	Han		0	21	2	0	0	0	21	9	0	0	0	28	0	0	0
	Can		0	6	19	0	0	0	9	16	0	0	0	10	18	0	0
	Hap		0	3	0	18	0	0	2	1	26	1	0	2	0	24	0
	Sad		0	0	1	1	25	0	0	0	3	31	0	0	0	0	28

that loudness of speech is an important factor of emotion expression and perception, especially for distinguishing categorical emotions in the arousal dimension. Finally, loudly spoken (intended) happy speech was often perceived as hot anger, presumably due to the fact that they are close in the arousal dimension of emotion perception.

4.3 Linguistic criticality of articulators

In the present study, the linguistic criticality of speech articulators for the realization of each phone is determined on the basis of the framework of Articulatory Phonology [Browman and Goldstein, 1989] and its computational ancillary Task Dynamics Model [Saltzman and Kelso, 1987, Saltzman and Munhall, 1989]. Articulatory Phonology views the speech production process as composed of articulatory gestures. Specifically, the formation and release of constrictions in the vocal tract

is represented by gestures depending on linguistic context, hence lexical items are differentiated by different gestural composition. The gestures are defined in terms of specific tract variables (e.g., lip aperture, tongue tip constriction degree) in task dynamics, which specifies the sets of articulators contributing to each tract variable parameters. In our subsequent analysis, we consider as critical articulators for a given phone those that the task dynamics application model [Nam et al., 2004] regards as involved in the production of that phone; the rest are considered non-critical. Alternative ways of specifying the linguistic criticality of articulators have been previously proposed. Notably, Jackson and Singampalli [2008] proposed an empirical approach based on Kullback-Leibler divergence, based on the assumption that the variance of articulatory positions in the mid-sagittal plane is smaller for critical articulators than non-critical articulators. The validity of this assumption has been supported by several experimental results [Papcun et al., 1992, Frankel and King, 2001, Jackson and Singampalli, 2008]. This was done, however, only for the case of neutral speech; validity in para-linguistic quality, such as emotion, was never considered.

For consonants, we consider that the lower lip, tongue tip, and tongue dorsum sensors correspond to labial, apical, and dorsal articulators, respectively. Table 4.2 shows the list of consonants used for analysis and the sensors corresponding to their critical or non-critical articulators. Even though the motions of both upper and lower lips are considered important for the production of bilabial consonants, we consider here only lower lip sensor as critical for the sake of analytic simplicity, taking also into account that the motion of the upper lip is highly correlated with the motion of the lower lip during constriction and releasing gestures for the bilabial consonants.

Table 4.2: List of stop and fricative consonants in the EMA database and the flesh point sensors of critical articulators of them. Note that /s/ and /z/ have two critical articulators, because both tongue tip constriction and tongue dorsum wide opening gestures are critical for the production of the phones [Nam et al., 2004]. The list of vowels in the EMA database is [ɑ, æ, ə, ɔ, ʌ, aʊ, aɪ, ɛ, eɪ, ɪ, i, oʊ, ɔɪ, u]. The flesh-point sensors corresponding to critical or non-critical articulators of vowels are not specified here due to their less clarity than consonants.

phone	Critical articulator	Non-critical articulator
d	Tongue tip	Tongue dorsum, lower lip
ð	Tongue tip	Tongue dorsum, lower lip
f	Lower lip	Tongue tip, tongue dorsum
g	Tongue dorsum	Tongue tip, lower lip
k	Tongue dorsum	Tongue tip, lower lip
m	Lower lip	Tongue tip, tongue dorsum
n	Tongue tip	Tongue dorsum, lower lip
p	Lower lip	Tongue tip, tongue dorsum
s	Tongue tip, tongue dorsum	Lower lip
t	Tongue tip	Tongue dorsum, lower lip
v	Lower lip	Tongue tip, tongue dorsum
z	Tongue tip, tongue dorsum	lower lip

The situation is more complicated for vowels, because the critical gestures for vowel production do not rely on constriction in a single narrow region in the vocal tract, but on multiple regions or at least on a wider constriction, compared to consonants [Jackson and Singampalli, 2009, Recasens et al., 1997]. For example, it is not clear which tongue sensor is most representative, and how much, for the wide palatal region of most vowels. Also, it is not straightforward how to choose sensors for a pharyngeal constriction gesture critical for /æ/, /ɑ/ and /ɔ/. Although Jackson and Singampalli [2009] suggested a simple way (i.e., tongue-tip sensor for front, tongue-blade sensor for mid and tongue-dorsum sensor for back vowels) to achieve this, there are still other issues, such as the inter-subject variability in terms of vocal tract shape and articulatory controls or the inconsistency of attached sensor positions across speakers at data collection. For these reasons, we

analyze vowels separately from consonants and compare the relative articulatory variability within a relevant articulator for the production of the vowels. For example, the palatal constriction gesture is critical for /i/, hence we consider the vertical movement of the tongue dorsum to be more critical than the horizontal movement for the vowel. Because the pharyngeal constriction gesture is critical for /a/, we consider the horizontal movement of the tongue dorsum to be more critical than the vertical movement.

4.4 Landmarks-based analysis on syllable segments

In this section, we investigate how emotion affects articulatory kinematics during syllable production, conditioned on whether the articulators in question are deemed linguistically critical for the initial and final consonants in the CVC syllables considered. Syllable segments allow us to study emotion variation at phone-target points as well as consonant-vowel transitions for different levels of linguistic criticality of articulators. Our hypothesis is that emotion coloring is more prominently manifested in the kinematics of non-critical articulators than in the kinematics of critical articulators. Since inter-speaker differences in emotion expression are widespread but not fully understood, all experiments in this paper were done separately for each speaker of the database (within-speaker analysis). We begin with analyzing how emotion-related variability of articulatory movements in the syllable varies depending on the linguistic criticality of the articulators.

4.4.1 Selection of syllables

We chose CVC syllables with identical place of articulation for the first and second consonants for fair comparisons between constricting and releasing articulatory movements. More specifically, we used the syllables, /p i p/, /n a n/, /f a v/, /p a p/, /t a t/, /n a t/, /p a p/, /d o u n/ and /p u p/. Table 4.3 shows the number of CVC syllable segments used for analysis. The movements of critical articulators for the consonants in these syllables are captured by the tongue tip sensor or the lower lip sensor. Note that we consider the emotional quality of each monosyllabic word the same as that of the corresponding utterance produced by the actors.

Table 4.3: Number of CVC syllable samples used for analysis. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. CAC1 indicates the critical articulator of the first consonant of its syllable, and CAC2 indicates the critical articulator of the second consonant of its syllable. TT is the tongue tip, L is the lips.

			JN					JR					SB				
CVC	CAC1	CAC2	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
/p i p/	L	L	4	4	15	7	6	7	3	14	11	8	9	4	9	8	8
/n ai n/	TT	TT	20	33	29	19	22	32	35	15	20	34	29	26	19	18	24
/f ai v/	L	L	14	22	22	10	14	22	30	8	10	24	20	18	14	10	16
/p a p/	L	L	6	6	10	10	8	10	2	13	11	9	8	7	7	7	8
/t ai t/	TT	TT	6	11	7	9	8	10	5	7	10	10	9	8	5	8	8
/n ai t/	TT	TT	10	14	22	20	14	16	8	18	22	20	18	12	16	12	16
/p ai p/	L	L	6	11	7	9	8	10	5	7	10	10	9	8	5	8	8
/d ou n/	TT	TT	2	5	16	9	10	10	12	9	4	14	9	6	9	5	11
/p u p/	L	L	5	7	11	10	7	8	4	9	11	10	9	6	8	6	8

4.4.2 Extraction of articulatory parameters

We targeted five linguistically important landmark time points in each CVC: (1) onset of the release of the first consonant; (2) instant of maximum velocity during the release of the first consonant; (3) instant of maximum opening of the vowel; (4) instant of maximum velocity during the movement towards the second consonant closure; (5) onset of the second consonant closure. The landmark points were selected based on the vertical trajectory of the critical articulator (for onset/coda consonants). Figure 4.2 illustrates where the five landmark points are located on the lower lip trajectory on the vertical axis during a /p a p/ syllable segment. Landmarks 1 and 5 for each syllable were determined in the vertical movement range, algorithmically chosen to be at the position 3% lower from the highest position value in the CV regions and of the VC regions, respectively. Landmark 3 was determined at the maximal vertical displacement of the jaw from the occlusal plane in the vocalic region. Landmarks 2 and 4 were determined at the points of the maximal absolute first-order derivatives in the CV regions and in the VC regions, respectively. The landmarks serve as basis for the standardization of articulatory parameters of different speakers and emotions.

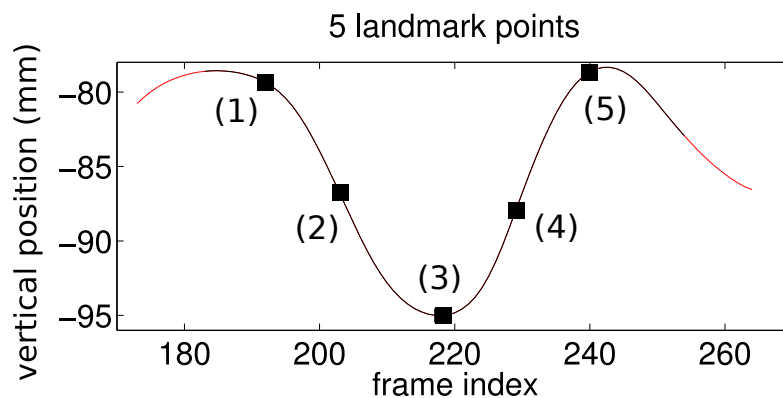


Figure 4.2: The five landmark points on the trajectory of the lower lip in /p a p/

At these five landmarks, articulatory kinematic parameters (position, speed and acceleration) of critical and non-critical articulators were measured. More specifically, positions in the horizontal and vertical directions, tangential speed and tangential acceleration were extracted from EMA sensor trajectories at each landmark. In order to minimize the effects of sensor tracking error, CVC syllables were discarded if any extracted position parameter was outside the $\pm 3\sigma$ range from the mean of the parameter, where σ and the mean were calculated over all data for each speaker. In total, 20 kinematic parameters (5 landmarks \times 4 measurements) were extracted for each articulator.

4.4.3 Statistical analysis of articulatory kinematics

In this section, we investigate what kinematic aspects of critical and non-critical articulators reveal significant emotional information and how they differ as function of criticality. A Kruskal-Wallis test [Kruskal and Wallis, 1952] was conducted on each articulatory kinematic parameter to reveal which parameters are emotionally significant. The null hypothesis of this test is that the samples (of each parameter) come from the same populations (of categorical emotions), while the alternative hypothesis is that the samples comes from populations, at least two of which differ with respect to location. The Kruskal-Wallis test is a nonparametric method that does not assume normal, or Gaussian, distribution of data points. We observed that sample distributions vary depending on articulatory parameters and that some features do not have normal distribution. For example, the distribution of tongue tip position in the x axis at landmark 1 (onset of release) was close to the normal distribution, while the distribution of tangential speed of the tongue tip at landmark 3 was considerably skewed to zero. Note that tangential speed at largest opening point is supposed to be small in general. The result of a Shapiro-Wilk

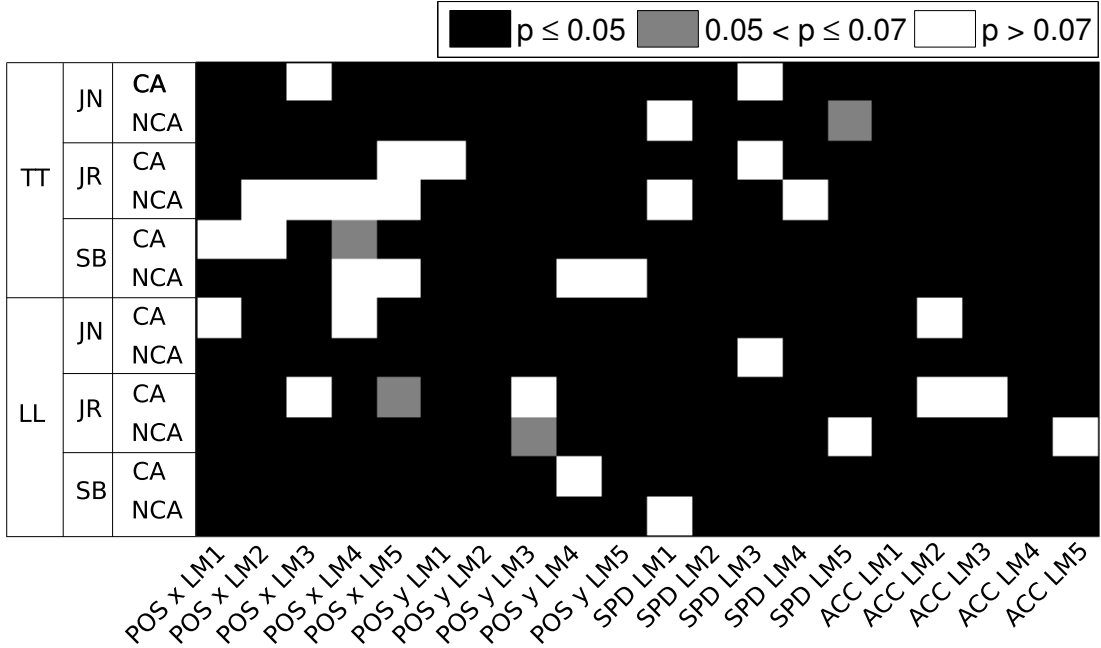


Figure 4.3: P-value of Kruskal-Wallis test on each articulatory parameters, such as horizontal and vertical positions, tangential speed and tangential acceleration at each landmark point. ‘POS x’ is the position in the x axis, ‘POS y’ is the position in the y axis, ‘LM’ is landmark, ‘SPD’ is tangential speed, ‘ACC’ is tangential acceleration.

test on the tangential speed of the tongue tip at landmark 3 of each emotion and each speaker also supported that the population of the feature was not normally distributed ($p < 0.05$).

Figure 4.3 shows the p-value of the Kruskal-Wallis test on each parameter for the five emotion classes. This figure shows, on top of large speaker dependence in emotional variation reflected in individual articulatory parameters, speaker-independent aspects of emotion-dependent articulatory variability in the given datasets of the three speakers. First, this figure shows that all tangential accelerations of the tongue tip are significantly different among five emotions for all speakers ($H(4) < 9.49, p < 0.05$). This result implies that emotion influences the variation of tangential acceleration of the tongue tip throughout the entire syllable,

regardless of the linguistic criticality of the tongue tip. Also, the vertical positionings of the tongue tip at landmarks 2 and 3 are statistically significant ($H(4) < 9.49$, $p < 0.05$), which indicates that the vertical positions of the tongue tip during releasing and at the largest opening are affected by emotion expression for all speakers. For lower lip parameters, horizontal position at the landmark 2, vertical position at landmarks 1, 2 and 5, tangential speed at landmarks 2 and 4, and acceleration at landmarks 1 and 4 are statistically significant for all speakers ($H(4) < 9.49$, $p < 0.05$). The difference of significant parameters of the tongue tip and the lower lip suggests that emotion-dependent variability appearing in articulatory movement margins (for the CVC syllables examined) is articulator-dependent.

Figure 4.3 also shows that some articulatory parameters of both tongue tip and lower lip are significantly different for the five emotions. Vertical position and tangential speed at the landmark 2 (maximum speed point during constriction release) and tangential acceleration at the landmark 4 (maximum speed point during constriction formation) are statistically significant parameters ($H(4) < 9.49$, $p < 0.05$) for all speakers, indicating that the movements of the two articulators during transition regions between two adjacent linguistic target positions are important sources of emotional information.

It is also observed in Figure 4.3 that some kinematic parameters of critical articulators are significant at landmarks 1 and 5, but not significant at landmark 3. For example, for speakers JN and JR, when the tongue tip is critical, tangential speed at the landmark 3 is not statistically significant ($H(4) > 9.49$, $p > 0.05$), while tangential speed at landmarks 1 and 5 is significant ($H(4) < 9.49$, $p < 0.05$). For a better understanding of this phenomenon, we examined the horizontal and vertical speed at the releasing onset and constriction/closure onset points separately. Figure 4.4 shows the histograms of vertical velocity for each emotion at

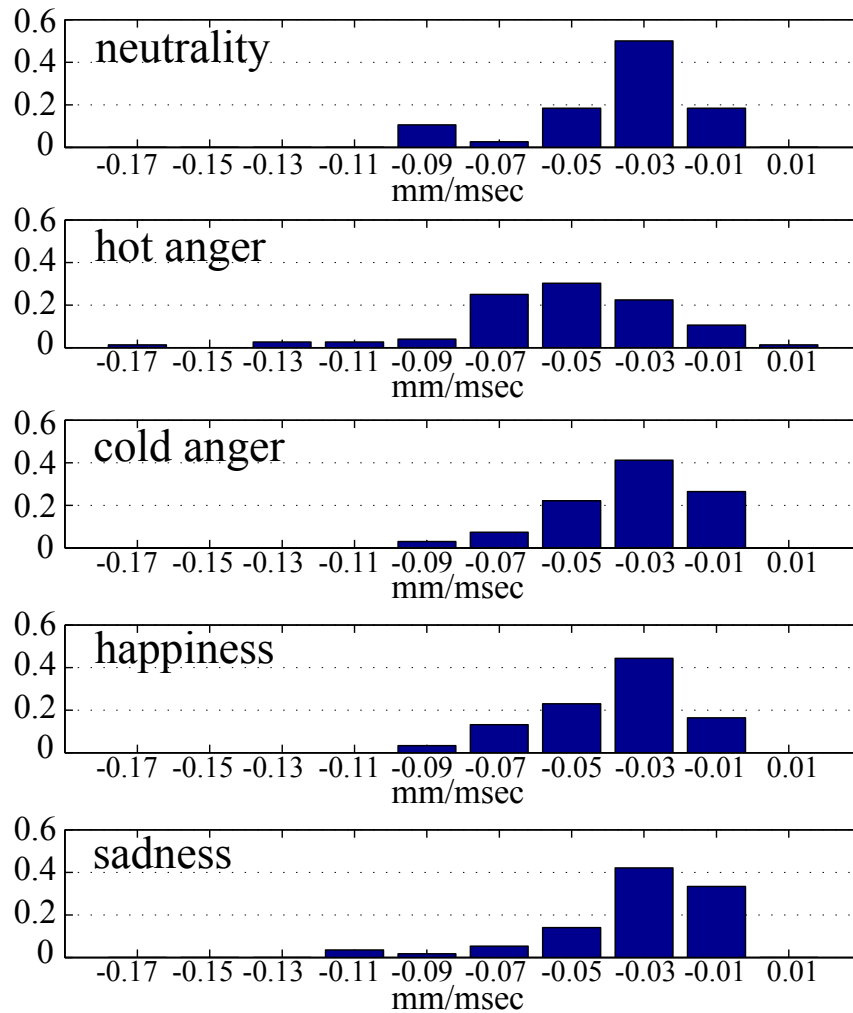


Figure 4.4: Histograms of the vertical velocity of the tongue tip at releasing onset point (landmark 1) for each emotion in JN’s data.

releasing onset (landmark 1) in JN’s data, as an example. We found that articulatory speed at onsets of release and closure were still significantly affected by emotion. More specifically, on average, higher horizontal and vertical speeds were detected at the two onsets in high arousal emotions, such as hot anger and happiness, than in the other emotions (e.g., sadness), while horizontal speed was lowest in sadness (low arousal emotion). In fact, tangential speed of critical articulators

at landmarks 2 and 4 is also statistically significant for both tongue tip and lower lip in all speakers data in Figure 4.3. These results indicate that, on average, initial and maximum articulatory speed during consonant-to-vowel transition, and maximum and final articulatory speed during vowel-to-consonant transition contain significant emotional information for critical articulators. These trends were not consistently detected in the non-critical articulators.

4.4.4 Analysis at the landmark points

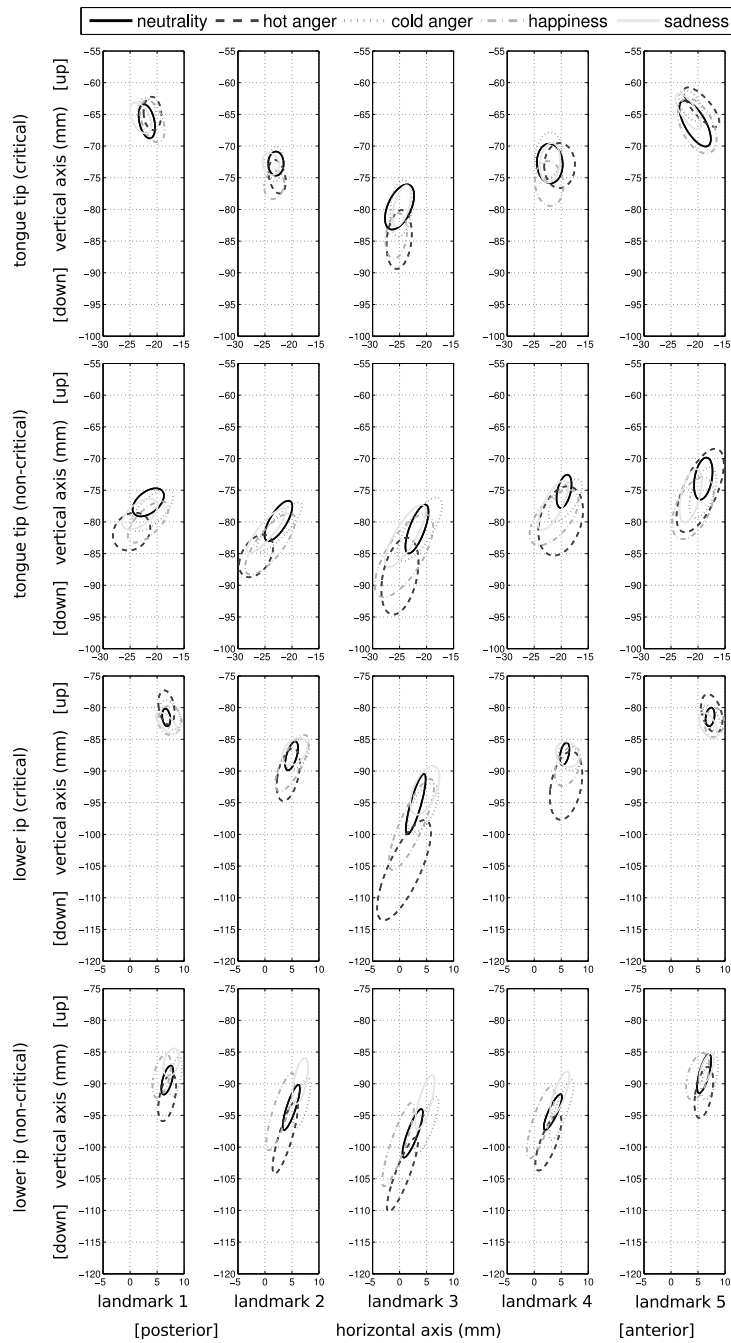


Figure 4.5: Example plots (speaker JN) of sample distributions (represented by 2-sigma ellipses) of articulatory positions at different landmarks

We also compared the distributions of the positions of critical and non-critical articulators at landmarks using two-sigma (two standard deviations) ellipses. Our goal with this analysis was to understand in what ways the movements of critical and non-critical articulators vary across emotions. Figure 4.5 illustrates emotion-dependent articulatory variability depending on the linguistic criticality. It is noted that the contrast of criticality of the tongue tip and the lower lip is greatest for consonants (landmarks 1 and 5), and the contrast decreases as being closer to the largest opening (corresponding to landmark 3) for vowels, since the criticality of articulator is categorized for consonants, not for vowels. Considering only the case of critical articulators (subfigures in the first and third rows of Figure 4.5), we observe that the divergence of non-neutral emotion ellipses from a neutrality ellipse was associated with the arousal dimension of emotions when the articulators were critical (for consonants), i.e. high arousal emotions showed greater divergence from neutrality than low arousal emotions. The dispersion of ellipse centers across emotions tended to be maintained throughout the CVC syllable regions for non-critical articulators. For example, when the lower lip is critical, the center of the hot anger ellipse is located higher than the center of the neutrality ellipse at landmark 1, then located lower at landmarks 2 and 3 (releasing and largest opening points, respectively), and finally located higher again at landmark 5 (closure formation). On the other hand, when the lower lip is non-critical, the relative locations of ellipse centers are consistent for all landmarks. This difference between critical and non-critical articulators is observed for all speakers.

Figure 4.6 illustrates emotional variations on articulatory trajectories of the tongue tip and the lower lip depending on their criticality. The articulatory segment is for ‘nine tight night pipes.’ In order to compare articulatory trajectories of

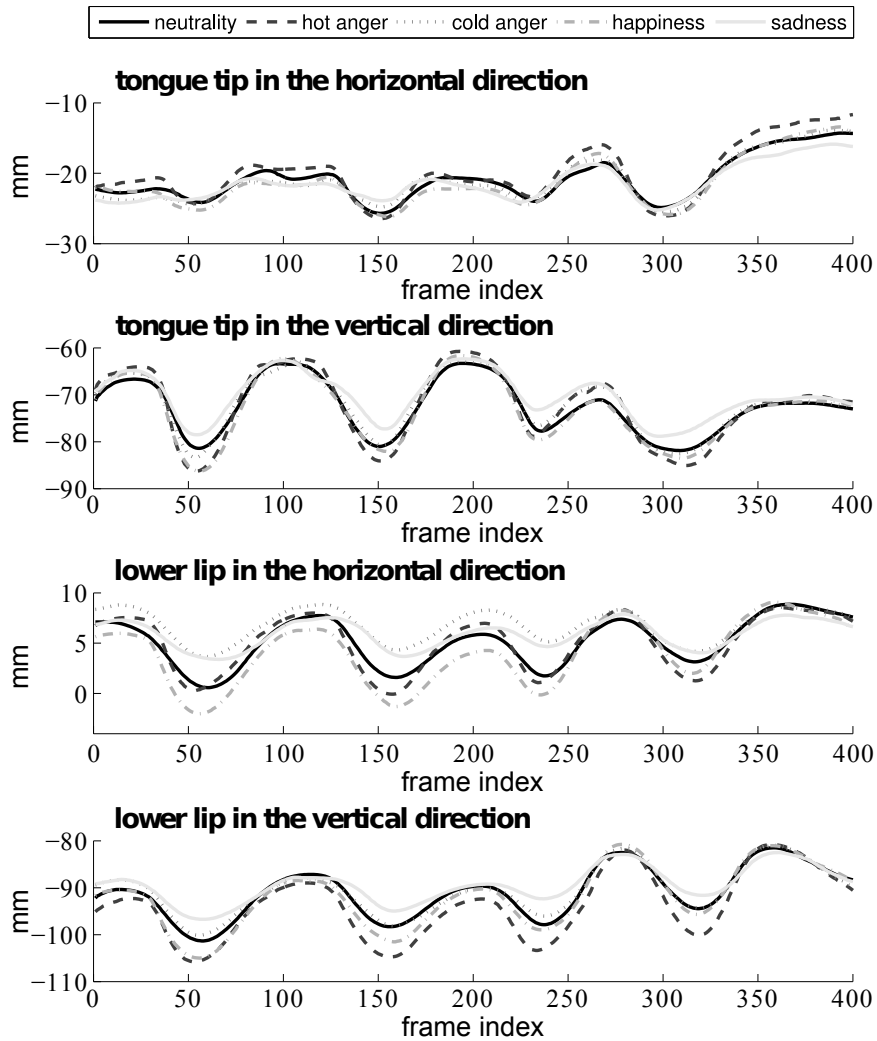


Figure 4.6: Averaged articulatory trajectories of each emotion for “nine tight night pipes” in the sentence 4 in JN’s data. The two plots from the top show the averaged trajectories of the tongue tip; the other two plots show those of the lower lip.

different emotions, articulatory trajectories of each instance were aligned to a reference trajectory which is arbitrarily chosen from neutral emotion data. Dynamic time warping [Sakoe and Chiba, 1978] with Euclidean distance between articulatory trajectories was used for the alignment. After alignment, average trajectories for each emotion were obtained by computing the frame-level mean of trajectories after a spline interpolation to 400 frames.

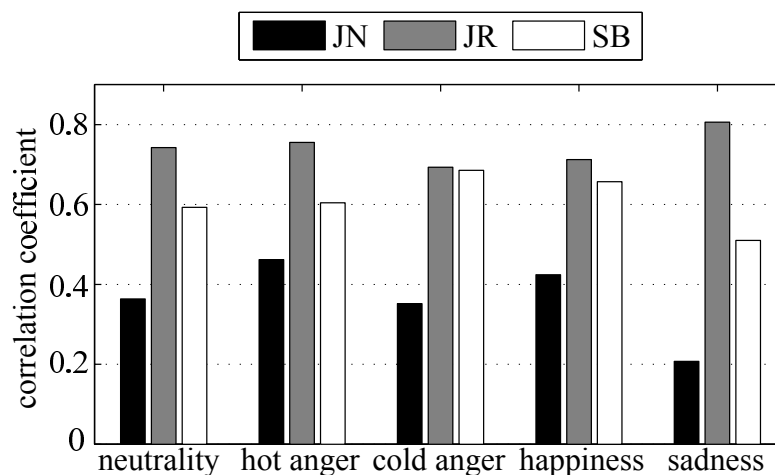


Figure 4.7: Correlation coefficients of two averaged trajectories of the tongue tip and the lower lip in the vertical direction

It was observed that when the articulator was critical for consonants, the articulations of hot anger and happiness often showed larger openings for vowels and exaggerated constrictions for consonants, while the articulations of sadness often showed smaller openings for vowels and smaller constrictions for consonants. When the articulator was non-critical, relative articulatory positioning among different emotions tended to be less sensitive to closure and opening gestures of the co-occurring critical articulators. This contrast of critical and non-critical cases was also observed in the data of all speakers, except the non-critical cases of JR’s data. However, the two-sigma ellipses plot (drawn with more data than the four words) of JR’s data supports that the relative articulatory positioning of different emotions for non-critical cases varies less in JR’s data. This may be due to the speaker-specific characteristic of JR, that is the significantly correlated vertical movements of the tongue tip and the lower lip as shown in Figure 4.7. This figure shows correlation coefficients of tongue tip and lower lip movements in the vertical direction,

which were computed from the average vertical trajectories of the two articulators for each emotion in “nine tight night pipes.” Figure 4.7 indicates that the vertical movements of the two articulators are most correlated for all emotions in JR’s data. Highly correlated movements of the two articulators of JR imply that the lower lip and the tongue tip movements are being coupled during the production of the monosyllabic words. This may suggest that the dependency between articulators is not a static parameter which is associated with only anatomical structure and coordination for linguistic encoding, but a dynamic parameter related to other factors, e.g., para-linguistic factors.

Figure 4.5 also shows the difference of articulatory variability during closure-to-releasing and approaching-to-closure motions. If the tongue tip is critical for consonants, the articulatory position shows greater variation at the landmark 1 than at the landmark 5 in terms of ellipse sizes and the dispersion of ellipse centers in the horizontal and vertical directions. In fact, such variation of articulatory position is greater at the landmark 2 than the landmark 4. This is counter-intuitive, because constriction formation of critical articulators for consonants is more actively and carefully controlled than releasing, hence approaching motions are likely to show less variability. The effect of co-articulation may be one possible reason for this phenomenon. The constriction gesture of the tongue tip for the second consonants in /n aɪ t/ and /n aɪ n/ may have become loosened by overlapping closure gestures of the lower lip for the following consonants, bilabial /p/ and labio-dental /f/, respectively.

In summary, we found that emotion affects the kinematics of both non-critical and critical articulators in the CVC syllable segments considered. The emotionally significant kinematic parameters vary depending on speakers and articulators. In addition, when articulators are critical for consonants in the CVC syllables, greater

openings for vowels and stronger constrictions for consonants for high arousal emotions are observed. When articulators are non-critical, relative articulatory positioning for different emotions tend to be less sensitive to closure and opening gestures of the co-occurring critical articulator. These results suggest that emotion affects the articulatory positioning for non-critical articulators and the movement range of critical articulators, controlled for achieving (short-term) linguistic goals.

4.5 Articulatory analysis at phonetic targets

Previous sections studied emotional variation in articulatory kinematics depending on the linguistic criticality of articulators with a limited dataset of CVC syllables. This section studies emotional variation of articulatory behaviors at phone target positions using the entire EMA database (excluding fast style speech as noted earlier).

4.5.1 Experimental Setup

The comparison of critical and non-critical articulators in terms of the articulatory variability of target phone position for different emotions requires determining when articulators reach the target position (steady-state point) for each phone. Manual selection of steady-state points in the whole EMA database is time-consuming. Acoustic boundaries determined by an automatic phonetic alignment do not inform articulatory target points directly. We determined the best steady-state point of each phone as follows. First, we determined phonetic boundaries using a hidden Markov model based automatic phonetic alignment toolkit, the Penn Phonetics Lab Forced Aligner [Yuan and Liberman, 2008], followed by manual correction of misalignment outputs. For each phone, we searched for the

best steady state point among multiple candidates. In the case of vowels, candidate points comprised the articulatory positions on the x and y axes of each articulator at three points: the middle frame, 20 msec before the middle frame, and 20 msec after the middle frame with respect to the manually corrected phone boundary. In the case of consonants, candidate point comprised the articulatory positions on the x and y axes of each articulator at the same three points, plus four additional points: (i) the highest point on the y axis in a large marginal region (20 msec before and after phone boundary); (ii) the highest point on the y axis in a small marginal region (10 msec before and after phone boundary); (iii) minimum tangential speed points in the large marginal region; (iv) minimum tangential point in the small marginal region. The reason for using the marginal regions instead of just phone boundaries is that acoustic phone boundaries do not always include the steady-state points of articulatory movements. Next, for each phone we calculated the mean of Euclidean distances from the median sample point to each articulatory position sample of each of the three frames for vowels or each of the seven frames for consonants. The frame of the least mean value was selected as the steady-state point for the phone.

Our foremost interest is the behavior at the phone level of (linguistically) critical and non-critical articulators in emotional speech production. However, from a statistical experimental design perspective the phone identity factor is nested in the criticality factor, which means that any analysis interested in the articulatory variability for different degrees of linguistic criticality requires a normalized representation of the variability across phones. To address this problem, we tried two different methods for parameterizing articulatory variability fairly applicable across phones, using (i) the centroid distances between emotion cluster pairs, and

(ii) the mean deviation within emotion for every phone. The details of the two methods are explained in the sections that follow.

4.5.2 Inter-emotion variability

In this section, we investigate how the degree of linguistic criticality of articulators is associated with inter-emotion variability. In particular, we examine which articulatory type (critical or non-critical) displays more inter-emotion variability in the articulatory phonetic target position. The inter-emotion variability in articulatory target positioning is quantified by the average of the centroid distances between emotion cluster pairs, where the centroid is the mean position of all samples of each emotion cluster. This parameter measures the averaged distance between the centers of different emotion clusters for each phone. Let x_i^k denote the arithmetic mean of all samples of the vertical or horizontal coordinate of a given EMA sensor’s position in emotion i and phone k . Suppose the number of emotion clusters is N . Then the average of the centroid distances between emotion cluster pairs of phone k , denoted by D_k , is calculated as

$$D_k = \frac{1}{C(N, 2)} \sum_{i,j} d(x_i^k, x_j^k), i \neq j \quad (4.1)$$

where $C(N, 2)$ is the number of 2-combinations from N elements; $d(a, b)$ is the Euclidean distance between a and b .

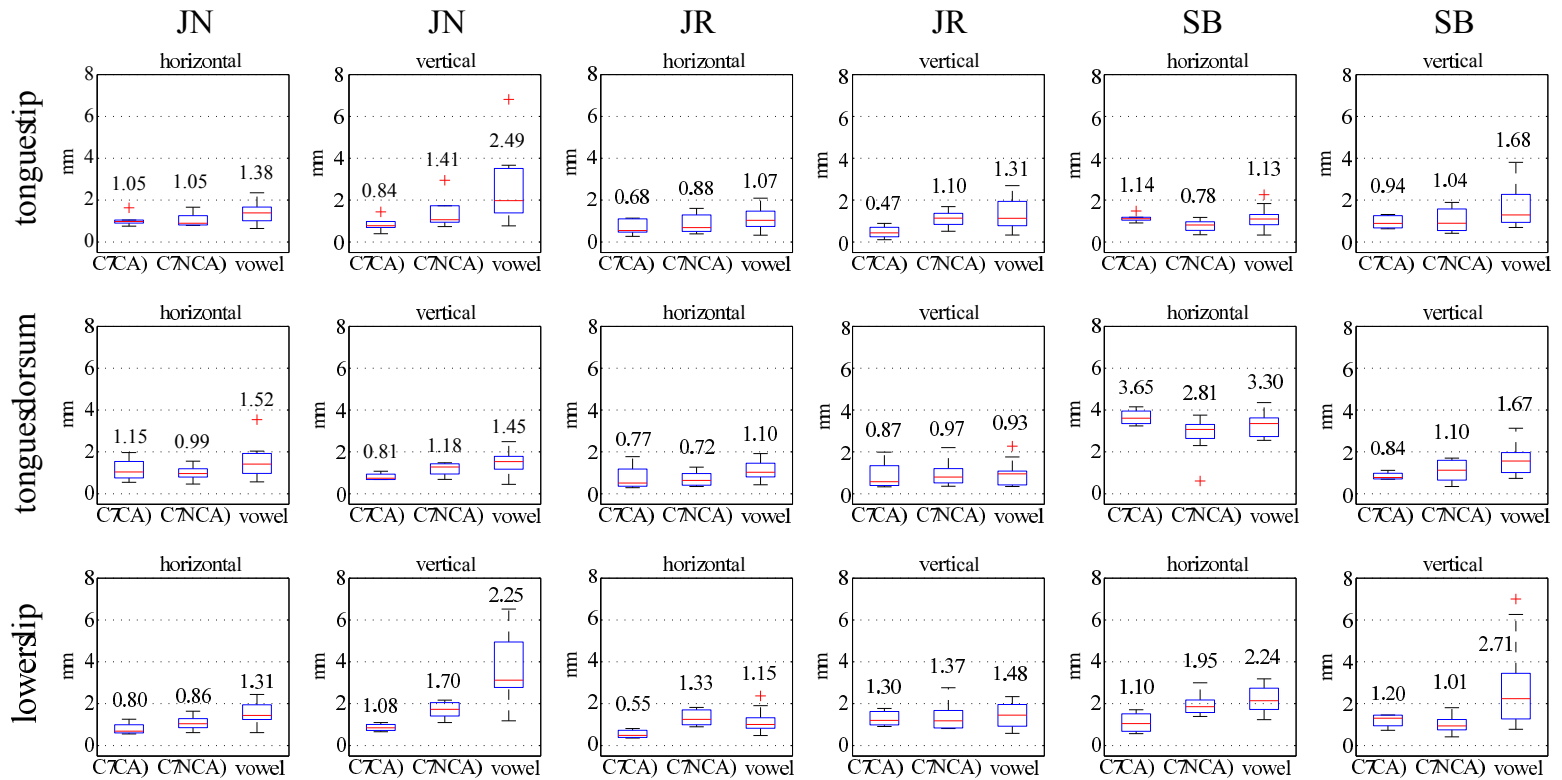


Figure 4.8: Box plots of the average of centroid distance among emotion cluster pairs. CA is critical articulator case, NCA is non-critical articulator case. C() denotes consonants. Vowels are analyzed separately from consonants, because of their different nature for determining critical or non-critical articulator in this study. The Value above each box plot is the mean of each case.

Figure 4.8 shows the box plots of D_k of critical and non-critical cases for each articulator. First, this figure shows that on average, the D_k of the horizontal lower lip positions at phone targets is greater for non-critical articulators than for critical articulators. We also conducted a one-tailed t -test with the hypothesis that the D_k of the horizontal lower lip positions at phone targets is greater for non-critical cases than for critical cases. Results indicated that the difference between critical and non-critical cases was significant for JR ($t=3.31$, $p=0.00$) and SB ($t=2.68$, $p=0.01$), but it was not significant for JN ($t=1.43$, $p=0.09$) at the 0.05 level. For the vertical position of the lower lip, the difference between critical and non-critical cases in terms of D_k was speaker-dependent: The average of D_k was greater for non-critical cases than for critical cases in the data of JN and JR, while it was the other way round in SB's data. These results suggest that on average, the mean positions of the lower lip for different emotions are more consistently dispersed for non-critical cases compared critical cases, only in the horizontal direction.

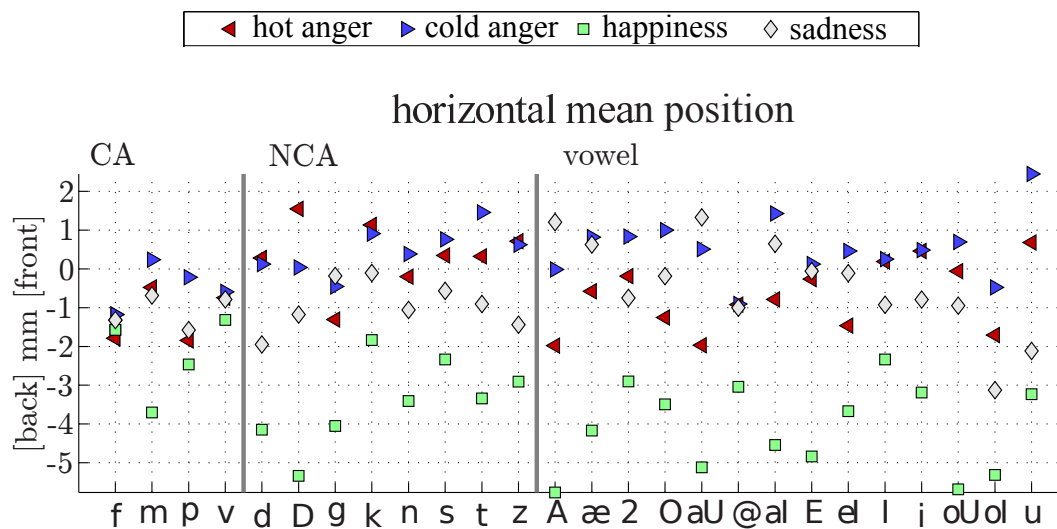


Figure 4.9: Relative mean (centroid) of the horizontal position of each emotion to the neutrality (which is aligned to 0 on the y axis) in SB's lower lip data.

We investigated the variation of the horizontal position of the lower lip more specifically for each emotion and each phone. Figure 4.9 shows the mean horizontal position of the lower lip of each emotion after aligning the mean of neutrality to be 0 for each phone in SB data, as an example. We found that the lower lip showed a posterior position for most phones when speakers expressed happy emotions than the other emotions. These trends were observed in the data of all speakers. The retraction of the lower lip for happiness was statistically significant in the result of one-tailed t -test ($t=7.38$, $p=0.00$). We also found that on average, the retraction of the lower lip for happiness occurred more significantly for non-critical cases than for critical-cases. One possible reason is that when speakers express happy emotion, their lip might have been stretched to the sides often, pulling the lower lip backward (smile-like gesture).

For each of the tongue sensors, the average D_k of the vertical position was larger for non-critical cases than for critical cases for all speakers. This indicates that on average, the vertical position of the tongue tip and the tongue dorsum at phone targets is more dispersed for non-critical cases than for critical cases. The difference between non-critical and critical cases was statistically significant at the 0.05 level for the tongue dorsum data of JN ($t=2.33$, $p=0.02$) and the tongue tip data of JR ($t=0.96$, $p=0.01$) by one-tailed t -test, while it was not significant for the other cases. On the other hand D_k of horizontal position was (even slightly) greater for critical cases than for non-critical cases for all speakers' data. The difference, however, was not significant for any of them at the 0.05 level of the one-tailed t -test. These results indicate that on average, the mean positions of the lower lip for different emotions are more dispersed for non-critical cases than critical cases, consistently only in the vertical direction.

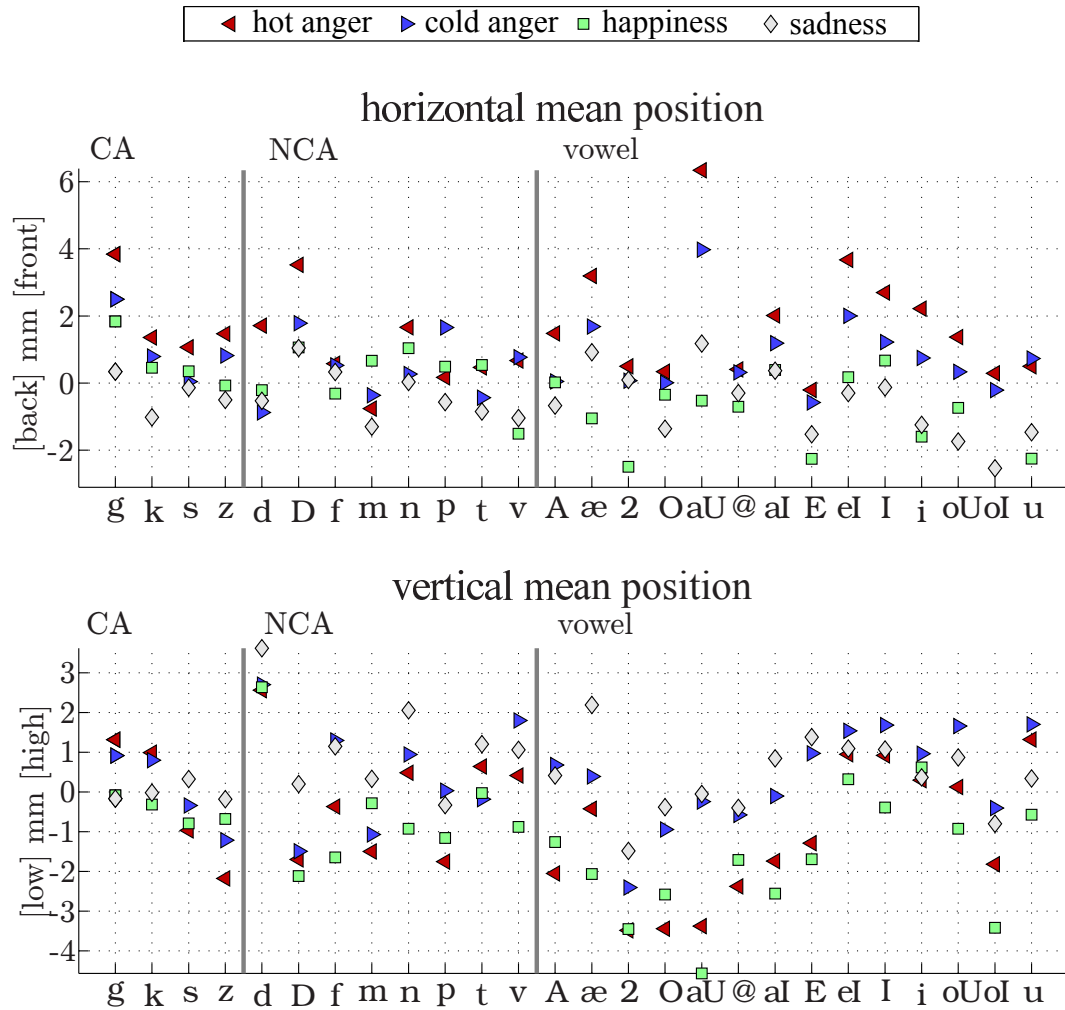


Figure 4.10: Relative mean (centroid) of the horizontal (top subplot) and vertical (bottom subplot) positions of each emotion cluster for each phone to the neutrality (which is aligned to 0 on the y axis) in JN's tongue dorsum data.

From our investigation on the tongue data for each emotion and each phone, it was observed that on average, the tongue dorsum as a critical articulator showed more upward constriction for velar stops, such as /g/ and /k/, when the emotional state of speakers were hot anger (and cold anger for JN and SB) as opposed to other emotions. More forward positioning of the tongue dorsum was also observed for hot anger, except for /k/ in SB's data. Figure 4.10 shows the horizontal and

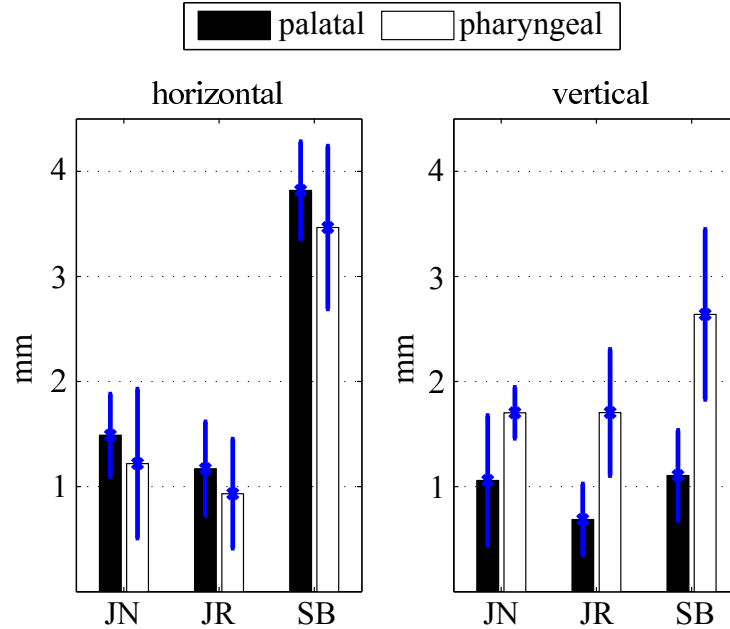


Figure 4.11: The average of centroid distances of emotion cluster pairs for the horizontal (left plot) or vertical (right plot) tongue dorsum positions, contrasted based on critical constriction gestures, such palatal constriction and pharyngeal constriction of the tongue dorsum, for vowels.

vertical positions of the tongue dorsum for each emotion and each phone in JN’s data, as an example. These results suggest that tongue dorsum closure gesture is stronger for hot anger than for the other emotions.

Finally, we also examined whether the tongue tip and the tongue dorsum showed different emotional variance depending on the linguistic criticality for vowels. According to the gestural description in the task dynamics application model [Nam et al., 2004], palatal constriction gesture is critical for /i/, /ɪ/ and /ε/, while pharyngeal constriction gesture is critical for /æ/, /ɑ/ and /ɔ/. It is reasonably assumed that the linguistic criticality of the vertical tongue dorsum position is higher for the palatal vowels than for the pharyngeal vowels. Also, it is assumed that the criticality of the horizontal tongue dorsum position is higher for the pharyngeal vowels than the palatal vowels. Figure 4.11 shows the average of

D_k of the tongue dorsum position in the horizontal and vertical directions for each of palatal and pharyngeal vowels. We found that on average, D_k of the tongue dorsum horizontal position was greater for the palatal vowels than for the pharyngeal vowels in all speakers' data, while D_k of the tongue dorsum vertical position was greater for the pharyngeal vowels than for the palatal vowels. This result implies that for vowels, the tongue dorsum position in less constrained direction displays more inter-emotion variation than in more constrained direction.

4.5.3 Within-emotion variability

In the previous section, we investigated inter-emotion variability of the mean position of articulators as a function of linguistic constraints. In this section, we investigate within-emotion variability in terms of the range of articulatory positions for critical and non-critical articulators. That is, we study in what way criticality associates with the range of articulatory positions for phone targets in each emotion. We also test our hypothesis that non-critical articulators have more emotional variation, particularly in terms of the within-emotion dispersion at articulatory target points, than critical articulators. We quantified the articulatory range variability of each emotion by the mean deviation of articulatory position samples in the horizontal or vertical direction.

Table 4.4: The results of one-tailed t -test with mean deviation measure on the hypothesis that non-critical articulator has greater range of articulatory target position for each phone within emotion than critical articulator. ‘x axis,’ ‘y axis’ indicates the results of test on mean deviation value of the horizontal articulatory position or the vertical articulatory position, respectively. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness and Sad is sadness. Only consonants are included in this analysis. Numbers in bold are statistically significant ($p < 0.05$). T -statistic is out of parenthesis in each cell, and p-value is in parenthesis.

Speaker	Emo	Tongue tip		Tongue dorsum		Lower lip	
		x axis	y axis	x axis	y axis	x axis	y axis
JN	Neu	0.46(0.33)	0.30(0.38)	0.42(0.34)	1.49(0.08)	1.32(0.11)	3.88(0.00)
	Han	1.13(0.14)	1.84(0.05)	0.35(0.37)	2.12(0.03)	3.08(0.01)	2.90(0.01)
	Can	-1.14(0.86)	1.66(0.06)	1.14(0.14)	2.35(0.02)	1.84(0.05)	4.85(0.00)
	Hap	-0.48(0.68)	0.62(0.27)	2.39(0.02)	1.45(0.09)	2.48(0.02)	2.68(0.01)
	Sad	-0.44(0.66)	0.47(0.32)	1.64(0.07)	2.49(0.02)	1.75(0.06)	2.89(0.01)
JR	Neu	1.34(0.11)	2.44(0.02)	2.41(0.02)	3.48(0.00)	0.52(0.31)	1.12(0.15)
	Han	1.87(0.05)	2.13(0.03)	2.11(0.03)	1.87(0.05)	2.13(0.03)	1.77(0.05)
	Can	1.78(0.05)	2.44(0.02)	2.78(0.01)	2.04(0.04)	2.24(0.03)	2.79(0.01)
	Hap	0.44(0.34)	1.38(0.10)	3.34(0.00)	2.31(0.02)	1.48(0.09)	-0.14(0.55)
	Sad	2.08(0.03)	2.55(0.02)	1.58(0.07)	2.65(0.01)	2.01(0.04)	1.42(0.09)
SB	Neu	-0.03(0.51)	1.86(0.05)	-1.29(0.89)	1.16(0.14)	2.61(0.01)	2.05(0.03)
	Han	-0.01(0.50)	1.92(0.04)	0.92(0.19)	1.71(0.06)	2.11(0.03)	2.72(0.01)
	Can	0.07(0.47)	1.85(0.05)	-0.03(0.51)	2.48(0.02)	2.31(0.02)	2.59(0.01)
	Hap	0.19(0.43)	1.82(0.05)	2.09(0.03)	1.44(0.09)	2.27(0.02)	1.57(0.07)
	Sad	0.03(0.49)	2.04(0.03)	0.73(0.24)	1.27(0.12)	1.99(0.04)	1.32(0.11)

First, we conducted one-tailed t -tests on the hypothesis that non-critical articulators had larger range of articulatory target position than critical articulators for each emotion. Table 4.4 shows the results. We found that on average, non-critical articulators show greater mean deviation than critical articulators in most cases (as indicated by positive t -statistic value in Table 4.4), which supports the hypothesis, overall. The statistical significance in Table 4.4 is largely speaker-dependent. For example, the difference between critical and non-critical in the vertical mean deviation of the tongue tip data of JN is not significant for any case, while the difference in the vertical mean deviation of the tongue tip data of JR is significant for four emotions (neutrality, hot anger, cold anger and sadness).

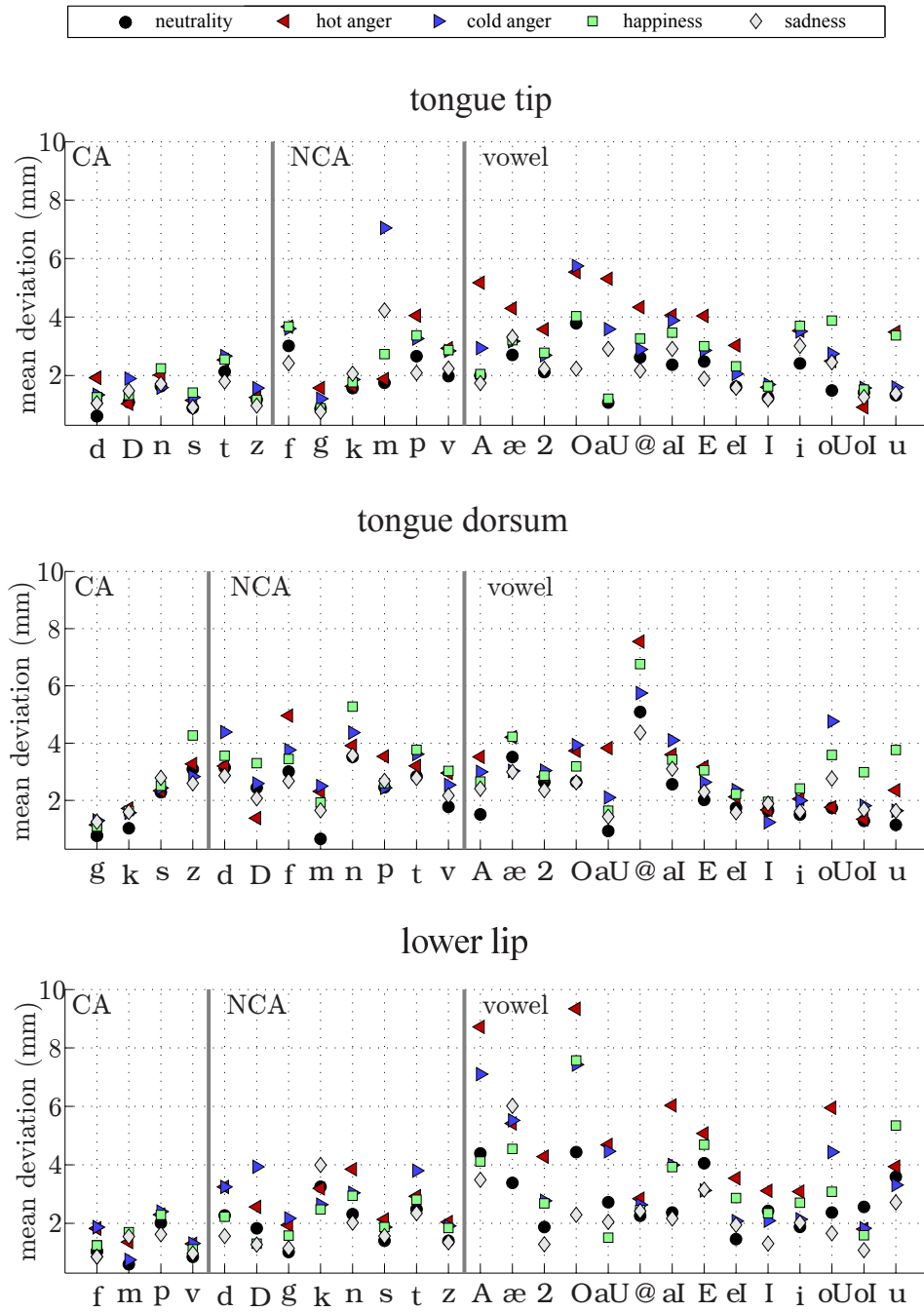


Figure 4.12: Scatter plots of the mean deviation of articulatory positions of each emotion of SB. Divided by 2 gray solid lines, the left most block is critical articulator (noted as CA), middle block is consonant non-critical articulator (NCA), the right most block is vowel non-critical articulator.

We also examined emotional variation of the positions of articulators depending on linguistic criticality more specifically for each phone. Figure 4.12 shows the scatter plot of the mean deviation of articulatory target positions in the vertical direction for each phone in SB's data as an example. We observed that overall, the mean deviations of the tongue tip, the tongue dorsum, and the lower lip at phone targets were larger when these articulators were non-critical than when they were critical. However, it is not always true for some cases. For vertical tongue tip position, the mean deviations of velar stops (/g/ and /k/) tended to be lower than those of the other non-critical cases (e.g., /f/, /m/, /p) and even similar to those of critical cases (e.g., /d/, /ð/, /n/). We note that two tongue sensors were placed about 1.5 ~ 2 cm closer to each other than the anatomical tongue tip and velar closure point on the tongue surface (tongue dorsum), which may have increased the dependency between the tongue tip and tongue dorsum motions. Hence, we speculate that tongue tip position was associated with tongue dorsum position more than the lower lip position, resulting in more limited tongue tip position variation for the velar stops than for the labial consonants. We also observe that, in the critical cases of the tongue tip, the maximal mean deviations in both horizontal and vertical directions for alveolar fricatives (/s/ and /z/) were always smaller than those for alveolar stops (/d/, /n/, /t/) in all speakers' data, presumably because alveolar fricatives require more careful maneuver than alveolar stops [Subtelný and Oya, 1972].

For the tongue dorsum, closure gesture for velar stops (/g/ and /k/) showed less mean deviation in the vertical direction compared to alveolar fricatives (/s/ and /z/) across all emotions. Although wide opening gesture of the tongue dorsum (as well as constriction gesture of the tongue tip) is essential for a production of

the alveolar fricatives [Nam et al., 2004], the result suggests that this wide opening gesture does not require a strict control of the constriction degree. Speaker-independent pattern contrasting the alveolar stops and the alveolar fricatives was not observed in the results of the tongue dorsum horizontal position.

4.6 Simulation experiment

In this section, we examine whether the large variability of non-critical articulators is the mechanical outcome of the controls of critical articulators. We first synthesize non-critical articulatory trajectories on the basis of the physiological constraints that govern the spatio-temporal relationships among all articulators and the time points when given articulators are linguistically critical. The emotional variations in the synthesized trajectories are compared to the emotional variation in the true data. If the two trajectories of non-critical articulators are similar in terms of emotional variation, it can be inferred that the emotion-dependent variability of non-critical articulators is a secondary effect of the control of critical articulators.

4.6.1 Description of articulatory model

This section describes an articulatory model that was used for the aforementioned simulation experiment. The details of this model is provided in [Kim et al., 2014d]. This model estimates trajectories of non-critical articulators based on only the following two factors: (i) the contextual constraints of the preceding or following time points when said articulators are critical, (ii) physiological constraints on said non-critical articulators from articulators that are critical at the time point in question. This estimation problem is formulated as the following: Let $f_i(t)$ denotes the position of i -th articulator at time t . $f_i(t_c)$ is the position of the i -articulator

at the nearest critical time point t_c from the current time t for the i -th articulator. Hence, this term represents the influence of the contextual constraints from the nearest critical point. $\hat{f}_i^p(t)$ represents the influence of the physiological constraints on the i -articulator. The estimated position of the i -th articulator, $\hat{f}_i(t)$ is modeled by convex combination of $f_i(t_c)$ and $\hat{f}_i^p(t)$, using a weighting function $K_i(t) \in [0, 1]$ as follows:

$$\hat{f}_i(t) = f_i(t_c)K_i(t) + \hat{f}_i^p(t)(1 - K_i(t)) \quad (4.2)$$

The weighting on the contextual factor should be negatively correlated to $|t - t_c|$, but the nature (linear or non-linear) of this function is unknown. Hence, $K_i(t)$ is modeled by the non-linear function as follows:

$$K_i(t) = \frac{1}{1 + \exp(-\eta(\lambda_i(t) - \xi))} \quad (4.3)$$

This sigmoid function can also be close to linear depending on the hyper-parameters η and ξ which are tuned on the development set. $\lambda_i(t) \in [0, 1]$ denotes a monotonically increasing function of $|t - t_c|$, thus $K_i(t)$ is monotonically decreasing.

$\hat{f}_i^p(t)$ is a function of the positions of *only* corresponding critical articulators at t as follows:

$$\hat{f}_i^p(t) = \sum_{\substack{l=1 \\ l \neq i}}^{N_C(t)} (\alpha_{i,l} f_l(t)) + \beta_i \quad (4.4)$$

where $N_C(t)$ is the number of the corresponding critical articulators at t ; $\alpha_{i,l}$ and β_i are the coefficients of the model. It is reasonably assumed that the effect of physiological constraints among articulators can be represented by an affine map. For example, the physiological influence from the position of the jaw to the position of the lower lip is computed by rotation, scaling and translation, those are affine

Table 4.5: The number of utterances selected for simulation experiment.

	Neutrality	Hot anger	Cold anger	Happiness	Sadness
JN	30	43	77	53	50
JR	62	44	55	48	67
SB	48	44	46	43	48

transformation. Note that the critical articulators' data used for representing $f_i^p(t)$ do not include the data of the i -th articulator itself.

Finally, the optimal $\hat{f}_i(t)$ is found by minimizing \mathcal{J} :

$$\mathcal{J} = \sum_{t=1}^M |f_i(t) - \hat{f}_i(t)|^2 \quad (4.5)$$

where M is the number of articulatory frames used for tuning the parameters of $K_i(t)$.

4.6.2 Synthesis of non-critical trajectories

In order to minimize the effect of erroneous articulatory data, we excluded utterances in which any of the articulatory data is out of empirically selected upper and lower boundaries. For each dimension of each speaker's data, the upper boundary is 0.95 quantile + 2 × standard deviation, while the lower boundary is 0.05 quantile - 2 × standard deviation. Then, each dimension of articulatory data of each speaker is scaled to the range of [0, 1] for fair evaluation. Table 4.5 shows the number of utterances selected for the simulation experiment.

The critical time point for each critical articulator for consonants is selected at the maximum constriction point of the articulator. We followed the distinction of critical and non-critical articulators in Table 4.2 for consonants. For vowels, the critical time point is decided based at the maximum opening point of the jaw and

Table 4.6: The results of evaluation of the estimated articulatory trajectories. The mean of RMSE or correlation coefficient is shown without parenthesis. The standard derivation is shown in parenthesis.

		Neutrality	Hot anger	Cold anger	Happiness	Sadness
JN	E_{RMSE}	0.079 (0.030)	0.074 (0.018)	0.074 (0.021)	0.072 (0.017)	0.071 (0.022)
	E_{CORR}	0.848 (0.116)	0.847 (0.096)	0.845 (0.122)	0.855 (0.102)	0.846 (0.130)
JR	E_{RMSE}	0.071 (0.023)	0.070 (0.022)	0.072 (0.024)	0.070 (0.020)	0.069 (0.023)
	E_{CORR}	0.869 (0.121)	0.864 (0.109)	0.856 (0.125)	0.865 (0.112)	0.862 (0.117)
SB	E_{RMSE}	0.079 (0.035)	0.071 (0.026)	0.077 (0.023)	0.071 (0.026)	0.075 (0.021)
	E_{CORR}	0.847 (0.135)	0.794 (0.141)	0.814 (0.116)	0.845 (0.119)	0.801 (0.162)

the tongue dorsum (in the vertical direction). The upper lip data is excluded in this experiment, because the upper lip is not anatomically constrained to any of the other articulators monitored in this dataset. Also, it was reported that the upper lip data did not improve the estimation performance [Kim et al., 2014d].

Our model is trained in leave-one-utterance-out setup for each emotion and each combination set of critical articulators, except the estimating articulator, because $\alpha_{i,l}$ and β_i of f_i^p in (4.4) depends on the combination. After the train and development sets are equally divided, $\hat{f}_i^p(t)$ is trained on the train set. The parameters of $K_i(t)$ are tuned on the development set. The performance of our final model $\hat{f}_i(t), \forall i$ is evaluated in terms of the mean of the root-mean-squared-error (RMSE), denoted by E_{RMSE} , and the mean of the correlation coefficient, denoted by E_{CORR} , between the true trajectory and the estimated trajectory of all utterances.

We first demonstrate that our model can estimate articulatory trajectories well for all emotions. Table 4.6 shows the evaluation results of the estimated articulatory trajectories in terms of RMSE and correlation coefficient. Our model shows satisfactory estimation performance (maximum $E_{RMSE} = 0.079$, minimum

$E_{CORR} = 0.794$) for data of all speakers and all emotions. This estimation performance is similar to one reported in our previous study [Kim et al., 2014d], which was performed with different sentences and emotions (neutrality, anger, happiness, sadness and fear). The result in Table 4.6 suggests that the trajectories of non-critical articulators is estimated reasonably well for the five target emotions in the dataset. It also suggests that the positions of non-critical articulators are considerably dependent on the positions of the corresponding critical articulators and the closest critical moment of the (non-critical) articulators. Figure 4.13 illustrates true and estimated vertical trajectories of the tongue tip, the tongue dorsum, and the lower lip for the sentence “I saw nine tight night pipes in the sky last night,” showing high similarity between the trajectories.

4.6.3 Results

In this section, we discuss the emotion-dependent variability of the non-critical articulators by comparing the true and simulated articulatory data. The emotion-dependent variability between the true articulatory data and the estimated data from the aforementioned model was compared by means of discriminant analysis and a statistical test. For the discriminant analysis experiment, emotion model was trained on the true data and tested in the estimated data. The classification accuracy on the estimated data was compared to the classification accuracy on the true data. Similar accuracy between the two results is an evidence for the high similarity between the true and estimation data in terms of emotion-dependent articulatory variability. For discriminant function, we used 2D normal density model, one mode for each emotion, in the Mahalanobis distance space.

The test statistic of the pair-sample t -test was used for the similarity metric of the two distributions: one for the true data and the other for the estimated data.

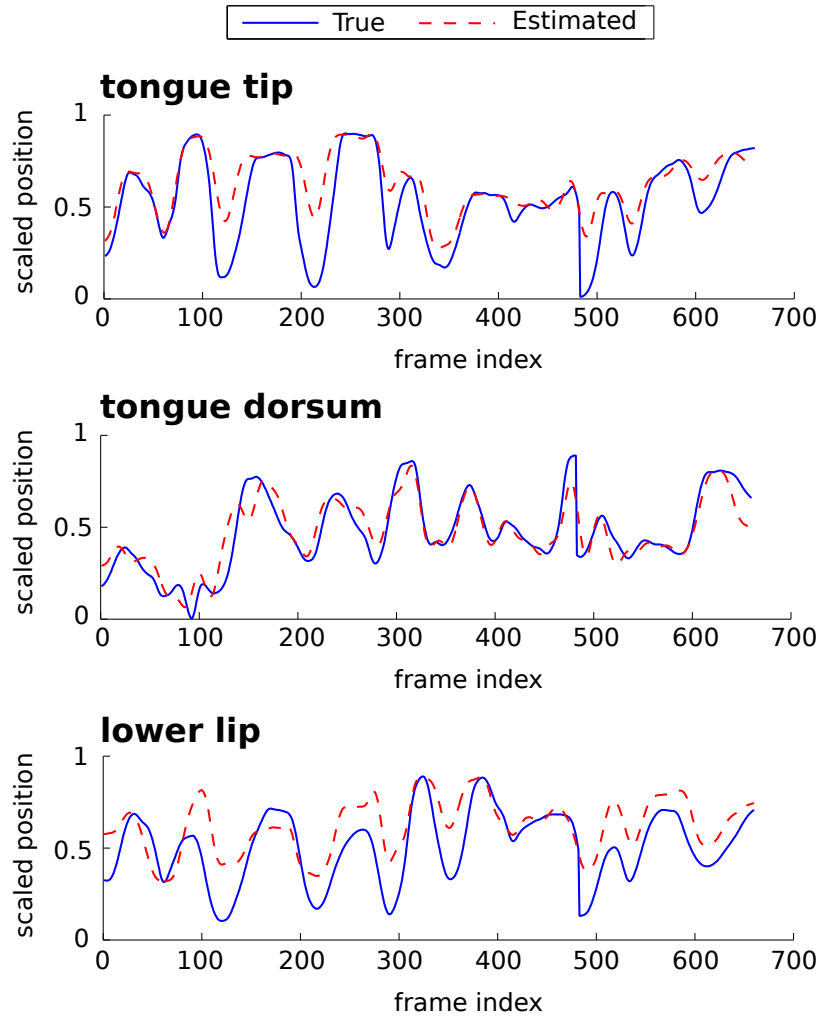


Figure 4.13: Example plots of the true and estimated trajectories of the tongue tip, the tongue dorsum and the lower lip in the vertical direction. An utterance of neutral emotion in JN’s data is used.

The analysis was performed for each phone, each emotion and each articulator. Each of true and estimated data was subtracted to the centroid of corresponding neutral data so that the distribution of each emotion represents the deviation from the neutral emotion.

Figure 4.14 shows the emotion classification results. In most cases, the classification accuracy of estimated data is similar to the classification accuracy of true

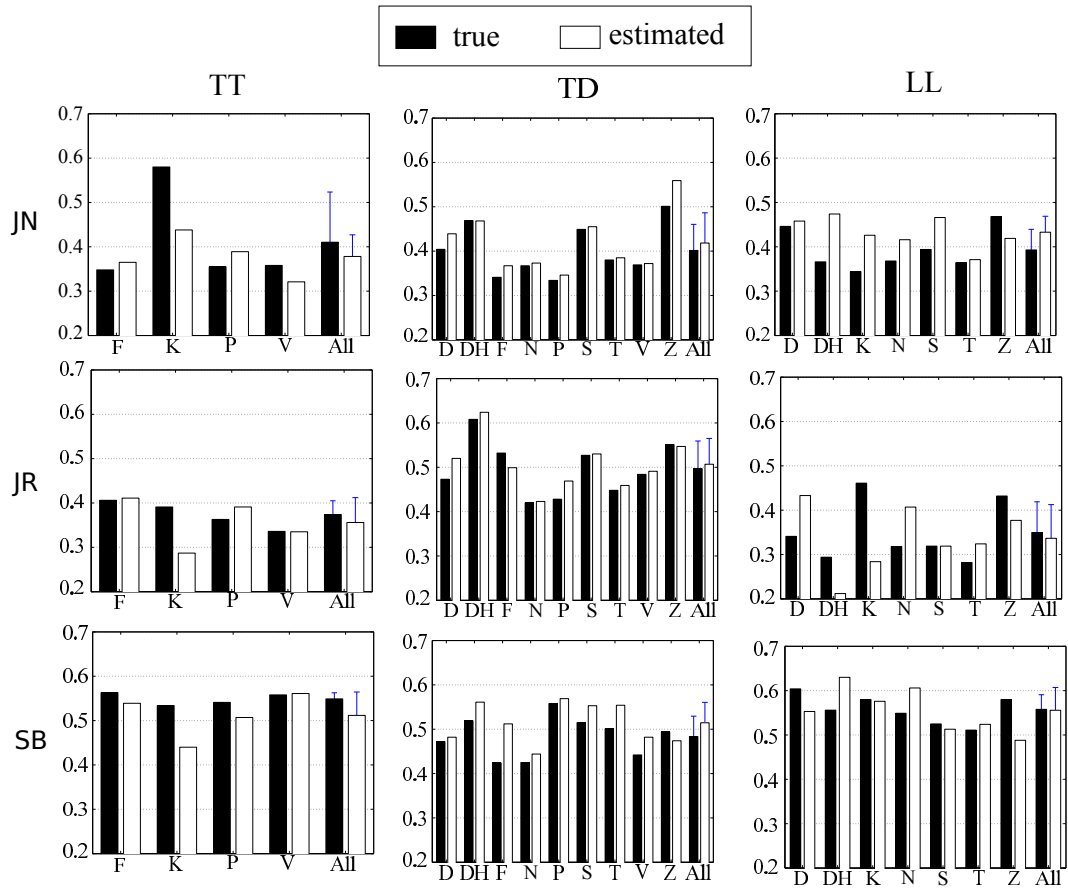


Figure 4.14: Unweighted emotion classification accuracy (%) for true and estimated data.

data. This indicates high similarity between true and estimated data in terms of the emotion-dependent variation of articulatory position distribution, and suggests that the large variability of non-critical articulators depending on emotion is significantly dependent on the controls of critical articulators.

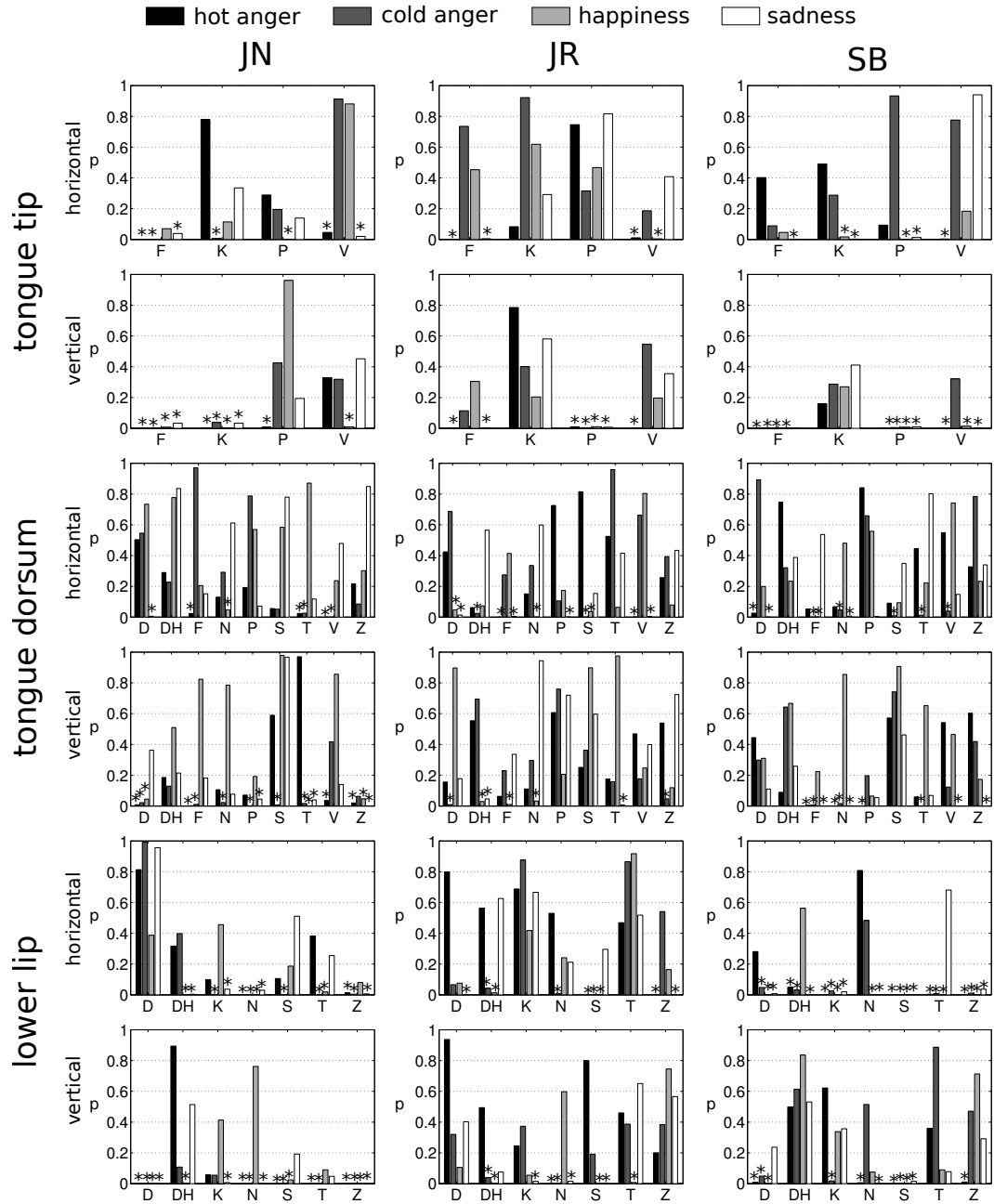


Figure 4.15: t -statistic of the pair-sample t -test on two distributions, one of true data and the other of estimated data for non-critical articulators for each phone, each articulator, each emotion and each speaker. * indicates that p -value is less than 0.05 for the case.

We also investigated the similarity between true and estimated data for *individual* emotions using the pair-sample *t*-test. The *p*-value less than 0.05 indicates that the means of the two distributions (true and estimated data) are statistically significantly different at the level of $\alpha = 0.05$, suggesting that overall, the emotional variations of the true and simulated data are significantly different in terms of their means. Although high *p*-value cannot be directly interpreted as a *statistical* evidence for the validity of null hypothesis, it can be used as a similarity metric of true and estimated data in terms of their mean. Figure 4.15 shows the *p*-value of pair-sample *t*-test for each speaker, each phone, each dimension and each articulator. Results of neutrality are omitted, because the means of their distributions are always 0. Note that the distribution of each emotion was normalized by subtracting to the centroid of neutrality. In many cases (Not marked by an asterisk in Figure 14) the *p*-value is greater than 0.05, so the null hypothesis that the mean of true and estimated data is significantly different cannot be rejected at the level of $\alpha = 0.05$ for these cases. In fact, the *p*-value is often considerably high. This is a supporting evidence for high similarity of the mean of true and estimated data in the cases. The result of high similarity suggests that the postural variation of non-critical articulators is often significantly dependent on the controls of critical articulators. The cases of high similarity are not consistent across speakers, suggesting large speaker variability on the dependency of non-critical articulators to critical articulators in emotional speech.

4.7 Discussion and Conclusions

This study provides evidence that the emotional variation pattern of articulatory positions during CVC syllables depends on the degree of linguistic criticality of

articulators for the first and the final consonants. When articulators are critical for the consonants in the CVC syllables, high arousal emotions show more peripheral articulatory movements with large movement range, especially in the vertical direction, while it was the other way around for low arousal emotions. The dispersion pattern of critical cases is in line with the experimental results of previous studies, i.e. large movement range and large opening for anger [Lee et al., 2005, 2006, 2008, Erickson et al., 2000]. Relative articulatory positioning of the five emotions for non-critical cases is not as sensitive to the manner of articulation for each phone as those for critical cases. One possible implication of these articulatory variations is the modulation of the vocal tract variables in the Task Dynamics as a result of emotion coloring. For example, when the tongue tip is critical for the initial or final consonants in a CVC syllable, tongue tip constriction degree in vowel regions can be higher (larger opening) for high arousal emotions than low arousal emotions. In summary, results suggest that the emotional variation pattern of articulators depends on the linguistic criticality of the articulators.

Considering speaker-independent behaviors observed in this study, our experimental results also support the hypothesis that the emotional variation of articulatory positioning for vowels is associated with linguistic criticality of the tongue body (the tongue dorsum sensor parameters were used) in terms of the average of distance between the mean positions of different emotions. Supporting evidence was found by comparison between palatal and pharyngeal vowels in their movement directions (critical and non-critical) in terms of the average of centroid distances between emotion cluster pairs in Section 4.5.2. Larger variance of non-critical articulatory trajectories when compared with critical articulatory trajectories were reported in literature [Papcun et al., 1992, Frankel and King, 2001], and this characteristic was employed for statistical identification of articulatory

roles [Jackson and Singampalli, 2009]. The experimental results in the present study provide additional information for vowels, i.e. the inter-emotion variance of articulatory positions is greater in the non-critical articulatory direction than in the critical articulatory direction. This suggests that the large variance of articulatory movements due to low linguistic criticality is an important factor of emotional modulation for vowels.

Previous studies have shown that tongue dorsum positions are dependent on emotion. For example, Erickson et al. [2000] reported that upward positioning and backward positioning of the tongue dorsum were observed for suspicion and admiration, respectively, in two vowels, /æ/ and /ʌ/ in an utterance “That’s wonderful.” The present study reports another speaker-independent characteristic in that the mean position of the tongue dorsum for velar stops is more forwarded and upward for hot anger than for other emotions. This tongue dorsum positioning for hot anger was consistently observed for all speakers only when the tongue dorsum were critical, implying that the exaggerated closure gesture of the tongue dorsum for velar stops is a characteristic of hot anger.

Non-critical articulators comprise dependent and redundant articulators according to the three-level categorization (critical, dependent, redundant) by Jackson and Singampalli [2009], Guenther [1995]. Dependent articulators refer to articulators whose movements are significantly dependent on the movements of critical articulators due to anatomical structure and/or coordinated articulatory controls for linguistic encoding. For example, the tongue blade is a dependent articulator of the tongue tip, and the jaw is a dependent articulator of the lower lip in general. Redundant articulators refer to the remaining articulators whose movements are little dependent on the critical movements. Although the present study considered only critical and redundant articulators, controls of dependent

articulators are also important to understand the detailed vocal tract shaping in emotional speech. Also, this study did not consider the jaw, although previous studies [Erickson et al., 2000, 2004, 2006] have shown that vertical jaw positioning is emotionally distinctive. In addition, jaw opening has been generally employed as a basic control of speech rhythm in literature [Nelson et al., 1984, Fujimura and Erickson, 2004], so a better understanding of jaw movement can be useful for a comprehensive model that incorporates articulatory and rhythmic aspects of expressive speech.

The results of our analyses still cast an open question: What are the acoustic and perceptual consequences of the emotional variations of critical and non-critical articulators, observed in the present study? Emotion perception tests with an articulatory synthesizer incorporating the controls of both critical and non-critical articulators will be useful for answering this question, although the articulatory synthesizer should be improved for minimizing potential loss of perceptual emotion quality first. In order to fully understand the variations of emotional speech production, it is important to know how the emotional variations of speech production components are related to each other, not only among articulators, but also with other emotionally crucial voice cues, i.e., prosody (pitch, energy and duration), intonation and voice quality. Articulatory synthesizers (e.g., [Rubin et al., 1996, Maeda, 1982, Toutios and Narayanan, 2013]) do not incorporate para-linguistic aspects of expressive speech yet. These remain topics for future research. Incorporating physiological constraint among non-critical articulators for the simulation experiment in Section 4.6 is also our future work.

Chapter 5

Invariant properties and variation patterns in emotional speech production

5.1 Introduction

The human speech signal is produced by the coordinated controls of vocal organs [Browman and Goldstein, 1992, Fowler and Saltzman, 1993] with large variability on their surface movements [Koenig et al., 2008, Jackson and Singampalli, 2009]. One of the main challenges in speech production modeling is, therefore, to represent articulatory behaviors in an effective, but simpler way. Despite the large variability of articulatory movements, previous studies have reported the presence of relatively invariant portions, called “iceberg” regions [Fujimura, 1986, Bonaventura, 2003], of the transient articulatory trajectories of demisyllables. More specifically, it has been observed that the speed of the (linguistically) critical articulator for producing the consonant in the demisyllable is relatively invariant at a certain excursion point regardless of prosodic change, e.g., different level of stress on the syllable, as long as the vowel of the demisyllable and para-linguistic factors, e.g., speaker-specific characteristics, gender and emotion,

are fixed. The iceberg region is roughly the fastest part of the critical articulatory trajectory in each demisyllable. This invariant characteristic at the iceberg is considered in the Converter/Distributor (C/D) model.

The C/D model is a comprehensive model of the speech production system, a part of which describes the (abstract) high-level temporal organization of speech, based on articulatory movements [Fujimura, 2000]. In this model, sequential syllable pulses represent the rhythmic pattern of consecutive syllables in the utterance. Syllable triangles are constructed based on the syllable pulses, where the height of the triangle reflects the syllable magnitude, i.e., syllable prominence, and the length of the base of the triangle reflects *abstract* syllable duration in the articulatory domain. Para-linguistic factors, e.g., speaker style, rate of speech and emotion, affect the variation of syllable pulse trains, resulting in the variation of the amplification and timing of the Impulse Response Function (IRF) for consonantal gestures [Fujimura, 1994b, 2002]. The IRFs are prototype time functions that represent inherent characteristics of elemental consonantal gestures, e.g., apical stop, labial fricative, velar stop. See [Bonaventura, 2003, Fujimura, 2002, Menezes, 2003] for the details of the entire C/D model that contains other components needed for generating articulatory signals from these variables.

The present study investigates the invariant properties and variation patterns of articulatory movements of emotional speech in the perspective of the C/D model. Such knowledge is valuable from both a theoretic standpoint (to shed further light on the articulatory control mechanism with emotion coloring) and application perspectives (such as in informing better articulatory modeling and (re-)synthesis with emotion). The invariant properties in the C/D model include (i) the strong linear relationship between the (vertical) excursion of critical articulators and the

articulatory speed at the iceberg and (ii) the linear relationship between the syllable duration and the syllable magnitude. The latter is based on the assumption that the acute angles of the two side lines of all syllable triangles, called “shadow” angles, are identical. This we refer to here as the consistency assumption.

The present study examines the variation of the timing and amplitude of the iceberg points found in the articulatory trajectories and the “shadow” angle. In the C/D model framework, emotion affects the parameters of IRFs, not the surface articulatory trajectories directly. However, the IRFs represent abstract articulatory gestural controls that are not directly observable due to the highly nonlinear nature between the IRFs and articulatory signals [Fujimura, 1994a], which makes the direct analysis on the IRFs harder. This study examines the variation of articulatory parameters that are influenced by the change of the IRFs as a function of emotion.

5.2 Iceberg metric

In most literature, the iceberg point is algorithmically determined at the minimum variance point of a number of trajectories of the same demisyllable. One approach is to find the point of the minimum root-mean-squared-error in the horizontal direction after optimal time shifting of the trajectories to the reference trajectory. Another approach is to choose the point of the minimum “iceberg metric” among multiple vertical movement bands of the critical articulator. The iceberg metric is proportional to the variance of articulatory speed and inversely proportional to the mean of articulatory speed in the band. Although these algorithmic approaches can find reliable iceberg points abiding in the invariability principle of the C/D model, these methods require a large number of trajectory samples to secure the

reliability. However, the number of trajectories of each demisyllable and each emotion is very limited in the present study.

This section offers discussion regarding the iceberg metric, presented in [Kim et al., 2015b]. Since the concept of the iceberg region is important in the C/D model, this section discusses the conventional way to determine the iceberg, following the guidelines in earlier work (e.g., [Fujimura, 1981, 1986, 1994b,a, 1996, 2002, Fujimura and Spencer, 1983, Menezes, 2003, Bonaventura and Fujimura, 2007]). According to the C/D model, a syllable consists of a nucleus (vowel) and onset and coda elements. For the sentence examined here, the crucial articulators for producing the onset and coda elements are the lower lip (for [p, m, b, f]), the tongue tip (for [s, d, t]) and the tongue dorsum (for /k/). If one overlays the vertical trajectories of the crucial articulator of all the instances for a given demisyllable containing a CV or VC pattern, for the same set of utterances for a given speaker, after shifting individual trajectories horizontally (in time) so that the sum of distance between individual trajectories and a reference trajectory in the horizontal dimension (time) is minimum, one can see the slopes in a certain region overlap very tightly. The longest trajectory is selected as the reference trajectory. This region is what is referred to as the iceberg. The iceberg metric is calculated for each band (1 mm interval) in the vertical range of motion of the crucial articulators, using the following formula [Menezes, 2003]:

$$Y(i) = \frac{\text{Var}\{X(i)\}}{\text{E}\{X(i)\}} \quad (5.1)$$

where i is the index of a small vertical region, $Y(i)$ is the iceberg metric at i -th band. $\text{Var}\{X(i)\}$ is the variance of vertical speed values of all data points in the i -th band, and $\text{E}\{X(i)\}$ is the mean of vertical speed values of all data points in

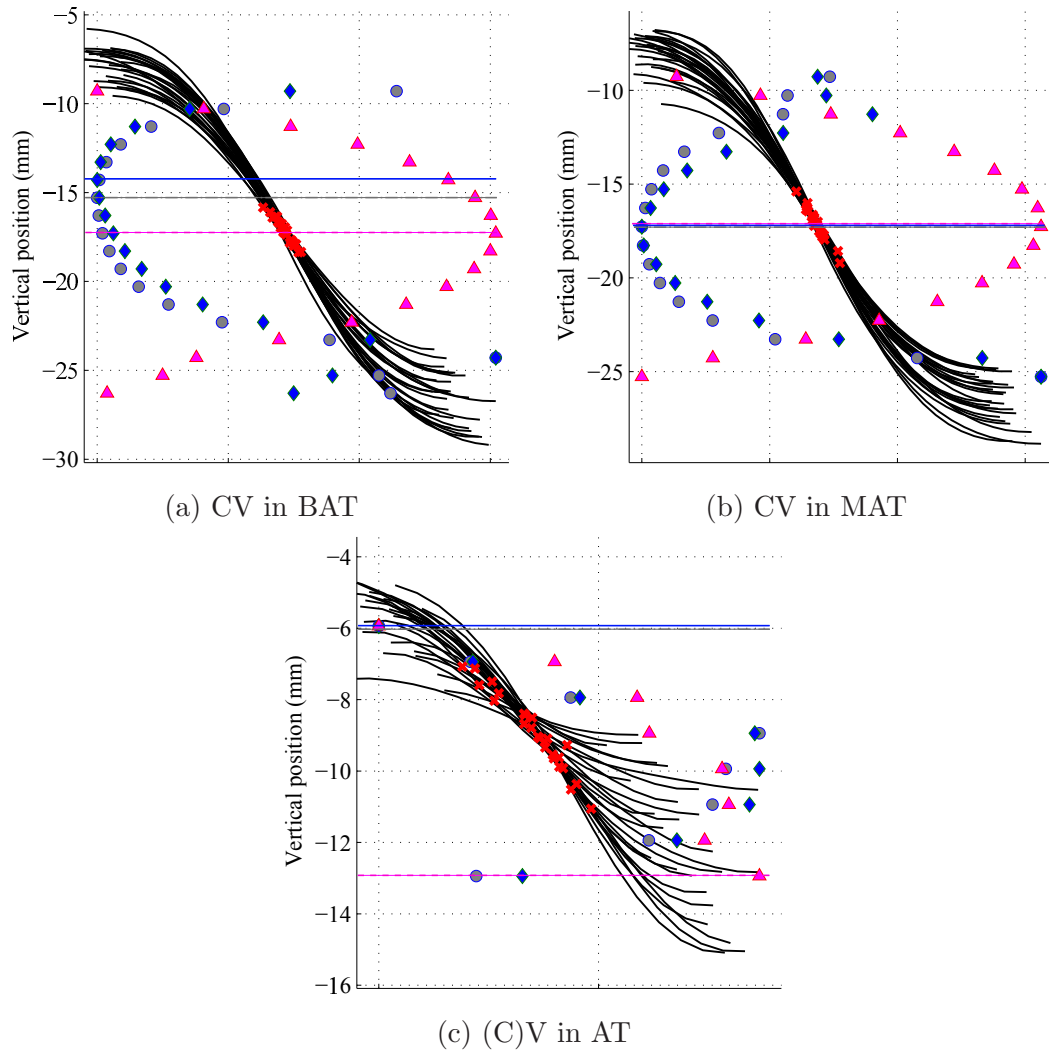


Figure 5.1: Example of iceberg metric, mean and variance of slopes in each band, overlaid with vertical trajectories of CA, after normalized to the range of the trajectories of CA.

the same band. The band of the smallest Y is selected as the iceberg region for the demissyllable.

According to Equation 5.1, the iceberg metric is a function of two factors: the variance of the slopes, as shown in the numerator, and the mean of the slopes, as shown in the denominator. The iceberg region is found where the variance term (in the numerator) is small and the speed term (in the denominator) is large.

Interestingly, as long as the demisyllable consists of a low vowel (/ae/ in our stimulus) and stop/fricatives ([p, m, b, f, s, d, t, k] in our stimulus) with complete closure or high constriction, the band of the minimum variances (in the numerator) and the band of the maximum mean speed (in the denominator) tend to be close, but do not always agree. This indicates that, depending on how the two factors are weighted, the algorithmically determined iceberg region can vary somewhat.

Figure 5.1 illustrates three different cases in terms of how well the bands match. Black curved solid lines are the horizontally aligned trajectories of the corresponding critical articulator: Trajectories for the initial demisyllable of BAT (Speaker A05) in (a), those for the initial demisyllable of MAT in (b), and those for the (co-articulated) initial demisyllable of (T)AT in (c). Pink triangles show the denominator values (mean of vertical speed) in (Equation 5.1), pink horizontal dashed line shows the center of the band for the maximum of the denominator values; blue diamonds show the numerator values (variance of vertical speed) in (Equation 5.1), blue horizontal solid line shows the center of the band for the minimum of the numerator values; gray circles show the iceberg metric of each band, gray horizontal dashed-dot line shows the center of the band for the minimum iceberg metric. Red ‘x’ dots show the maximum speed points of individual trajectories. Figure 5.1(a) is when the two optimal bands are not identical; the greatest mean speed (pink horizontal dashed line), the least speed variance (blue horizontal solid line) and the least iceberg metric (gray horizontal dashed-dot line) are located at different bands. Figure 5.1(b) illustrates the case when the three bands are identical, although the complete agreement among the three bands is not common in our data. We observed that the robustness of the iceberg metric depends on coda or onset for the demisyllable. Figure 5.1(c) shows results of the iceberg metric calculation for the first demisyllable of at, computed using the releasing movements

of TT for the preceding consonant /t/ in cat. For our stimulus, Pam said bat that fat cat at that mat, we have in the ensuing analysis assumed as a sort of temporary measure the final /t/ of cat to be the demisyllabic onset of at. For this iceberg region without a clear onset, the iceberg metric, and its numerator and denominator are very noisy, although the maximum speed point of individual trajectories are clustered in the center region. We also note that this is an important area of the model that needs to be addressed, especially for a language such as Japanese, which generally does not have syllable codas.

The present study chooses to use the maximum speed point of the crucial articulator for the onset or coda of each demisyllable. This method has been deployed by Erickson [2010], Kim et al. [2015a, 2014a], Erickson et al. [2015]. The justification of this is two-fold: (1) we found that the maximum vertical speed points of individual trajectories are very close to the point of the minimum iceberg metric, tightly clustered near the three bands, as shown in Figure 5.1(a,b), and (2) eventually, it is more desirable to be able to determine the iceberg points for each utterance independently, due to the difficulty and high cost of obtaining data with large repetitions. It is noted that using the maximum speed point of an individual trajectory can be noisier than using the point of the minimum iceberg metric from many repetitions, although our preliminary results in our (limited) data show that such noise is small. Remember that the center of the syllable is calculated as the midpoint between the iceberg regions, and based on these calculations, syllable triangles are then generated to derive not only syllable durations, but also phrasing/boundary patterns. Using this approach, we hope to examine the relationship between articulation of syllable duration, phrase boundaries, syllable magnitude, and the perception of them. The correspondence between articulatorily-generated phrase boundaries and perceived boundaries is discussed in [Erickson et al., 2015].

A discussion of articulatory syllable durations can be found in [Erickson et al., 2014], but are not discussed here.

5.3 Methods

5.3.1 Data

The ElectroMagnetic Articulography (EMA) dataset collected by the NDI WAVE system is used in this study. A sentence “*Pam said bat that fat cat at that mat*” was spoken by a female native speaker of American English. The stimulus was designed specifically for the study of the C/D model. For consonants, it contains only stops and fricatives in which the invariant properties of the C/D model have been shown in literature, e.g., in [Fujimura, 2000, 2002]. For vowels, it has only two vowels, eight /ae/ and one /eh/, so that the variation of the C/D model parameters due to vowels is minimized. The sentence was repeated five times for each of the five emotions, such as neutrality, anger, happiness, sadness and fear. The speaker was a professional actress who had theatrical vocal training. She was asked to start speaking after she had immersed herself in the target emotion.

A six Degree-Of-Freedom (DOF) sensor of the NDI WAVE system was used as the reference sensor, and six 5-DOF sensors were used for monitoring the movements of articulators, such as the tongue tip (TT), the tongue blade (TB), the tongue dorsum (TD), the upper lip (LL), the lower lip (UL) and the jaw. The 3-dimensional coordinates of the six 5-DOF sensors were recorded at a sampling rate of 100 Hz, and speech waveform was simultaneously recorded at a sampling rate of 22050 Hz. Occlusal plane correction was performed on the articulatory data of all utterances by using the recording of three 5-DOF sensors attached on the bite plate. After interpolating missing frames by the piecewise cubic Hermite

Table 5.1: Confusion between the target emotion and the final emotion label, i.e., the best (perceived) emotion. ‘Neu’ is neutrality, ‘Ang’ is anger, ‘Hap’ is happiness, ‘Sad’ is sadness.

		Final label					
		Neu	Ang	Hap	Sad	Fear	Other
Target	Neu	5	0	0	0	0	0
	Ang	0	5	0	0	0	0
	Hap	0	0	5	0	0	0
	Sad	0	0	0	5	0	0
	Fear	0	0	0	1	4	0
Total		5	5	5	6	4	0

interpolating polynomial, each sensor trajectory was smoothed with a 9th-order Butterworth low pass filter with a cutoff frequency of 20 Hz. Only tongue tip, tongue dorsum and lower lip sensors were selected as critical articulatory sensors for the sake of simplicity of analysis along with jaw contribution.

The best emotion of each utterance was judged by 11 native speakers of American English. After listening to each utterance, the evaluators were asked to choose (1) the best representative emotion among six categories, such as neutrality, anger, happiness, sadness and ‘other,’ where ‘other’ was for the case that none of the listed five emotions was the best, (2) confidence in their judgment, and (3) the strength of emotion expression. Confidence and strength were evaluated on a five-point Likert scale. The best emotion was determined by majority voting. If there were multiple emotions with the same evaluation score, the one of higher mean of confidence scores was chosen. Table 5.1 shows the confusion between the target emotion and the best (perceived) emotion used for analysis.

5.3.2 Parameter extraction

The critical articulators should be defined for computing C/D model parameters pertinent to this study. In this study, the critical articulator for each phone is determined based on the place of articulation, i.e., the tongue tip for coronals (/s/,/th/,/t/,/d/), the lower lip for labials (/p/,/m/,/b/,/f/), and the tongue dorsum for dorsals (/k/). Although there is no initial consonant for “AT,” the final consonant /t/ of the previous word “CAT” was used for extracting C/D model parameters, because “CAT” and “AT” were spoken continuously without pause.

The midpoint between the two iceberg points (for onset and coda) in each syllable is where the syllable pulse was placed. The excursion of the jaw at the midpoint was considered to be the height of the syllable pulse, which represents the syllable magnitude. The excursion of an articulator refers to the shortest distance between the occlusal plane and the position of the articulator [Bonaventura, 2003]. Then, the “shadow” angle of the triangle was calculated for each utterance in such a way that there is at least one pair of the close edges of adjacent triangles which meet with no overlap in between any adjacent triangles [Fujimura, 2000].

Figure 5.2 illustrates the iceberg time points, and syllable centers, syllable triangles in a neutral speech utterance. The time difference between the onset/coda pulse, i.e. syllable triangle edge, to the iceberg point of the demisyllable is referred to as τ (not shown in the figure, but discussed in Section 5.5).

5.4 Analysis on the invariant properties of the C/D model

This section discusses two invariant properties in the C/D model associated with emotional speech. One is the strong linear relationship between the excursion of

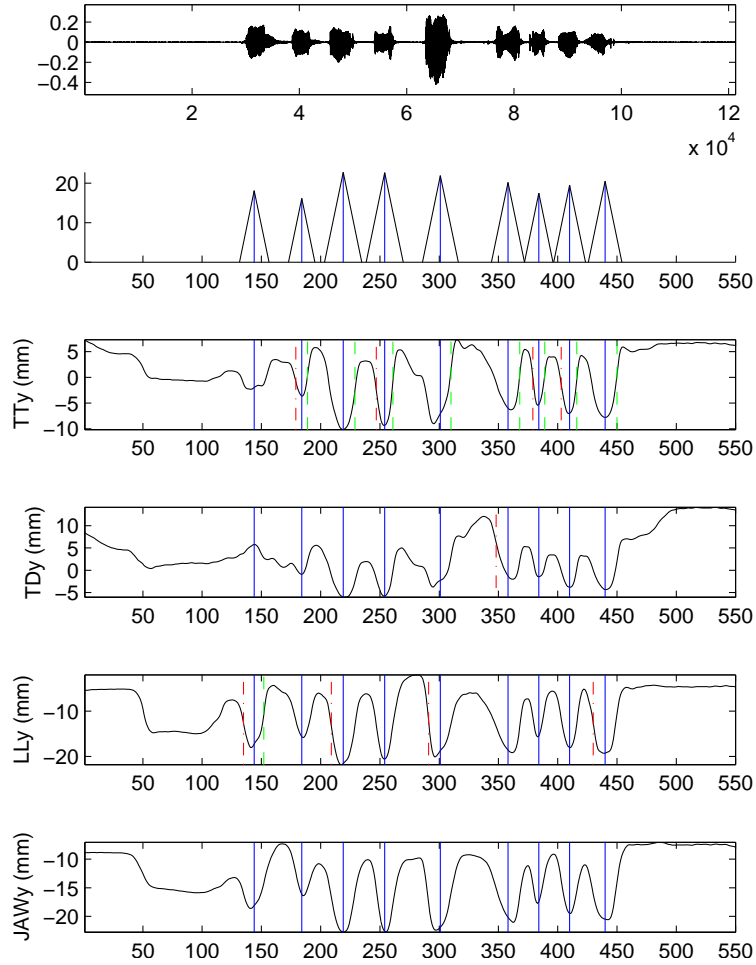


Figure 5.2: Syllable triangles constructed for a neutral utterance of “Pam said bat that fat cat at that mat.” The 1st panel is the speech waveform. The 2nd panel shows syllable triangles. In the other panels, the red dash-dot line denotes the iceberg time point for onset; the green dashed line denotes the iceberg time point for coda; the blue solid line denotes the syllable center point.

the critical articulator and the speed of the articulator at the iceberg, and the other is the consistency of the “shadow” angle values across utterances spoken with the same emotion. These invariant properties are examined across emotions and within emotion.

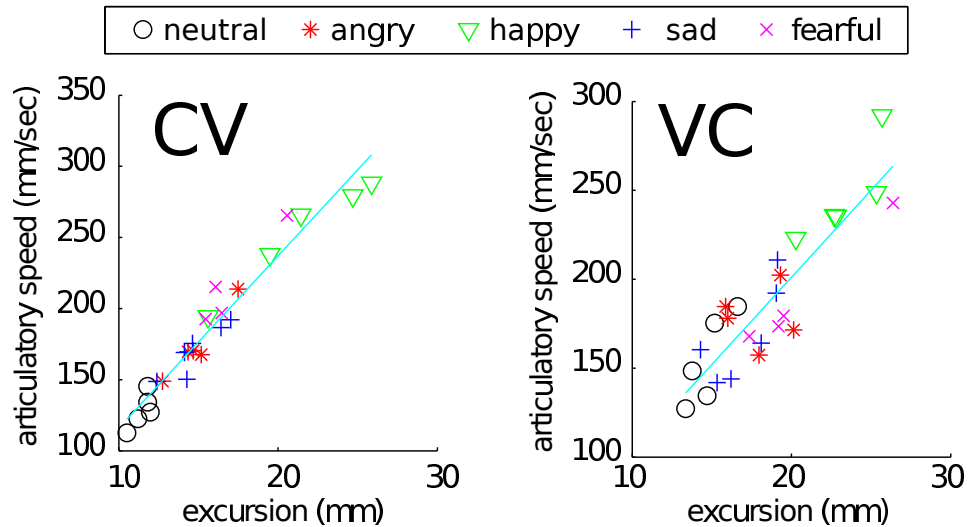


Figure 5.3: Example scatter plots for the excursion of the critical articulator (of consonant) and the articulatory speed at icebergs in CV/VC demisyllables. “*Pam*” is used in this plot.

5.4.1 Iceberg point

Visual inspection of the scatter plot (Figure 5.3) shows a strong linear relationship between the excursion of the critical articulator and the speed of the articulator at the iceberg point for CV and VC demisyllables, for all utterances regardless of the emotion condition. For demisyllables, happiness shows the greatest excursion and the highest articulatory speed, while neutrality shows the smallest excursion and the lowest articulatory speed. A linear regression analysis (Table 5.2) shows the F -statistic and p -value for all emotion conditions, including neutrality, for each CV/VC demisyllable. In Table 5.2, the p -value is significant at $\alpha = 0.00005$ level in all cases, indicating that a linear relationship between the two parameters is maintained across all emotion conditions, not just in neutral speech. This support of the C/D Model assumption of the linearity of articulatory speed and excursion is discussed further in Section 5.6.

Table 5.2: Statistical test on linearity between excursion and speed of the critical articulator for consonant. ‘**’ denotes that p -value < 0.0000005 . ‘*’ denotes that p -value < 0.00005 . $N=25$.

Syllable	CV				VC			
	β_1	β_2	F	p	β_1	β_2	F	p
PAM	12.1	-5.1	405	**	9.8	5.0	92	**
SAID	11.4	2.6	78	**	7.8	57.7	68	**
BAT	6.8	70.8	98	**	16.0	-83.2	92	**
THAT	15.0	-35.0	187	**	13.3	1.4	64	**
FAT	10.1	25.6	238	**	10.9	10.7	53	**
CAT	12.7	-52.6	48	**	12.2	12.8	124	**
(T) AT	11.9	23.3	139	**	10.6	43.8	26	*
THAT	14.9	-26.7	57	**	13.5	-1.3	88	**
MAT	11.1	-2.0	285	**	16.5	-66.2	55	**

5.4.2 Shadow angle

Next, we examine the shadow angle within each emotion condition in order to investigate the invariance of the shadow angle of the syllable triangle.

Figure 5.4 shows the errorbar plot of the shadow angle computed for each utterance. Note that the angle varies depending on the emotion: 36 degrees for

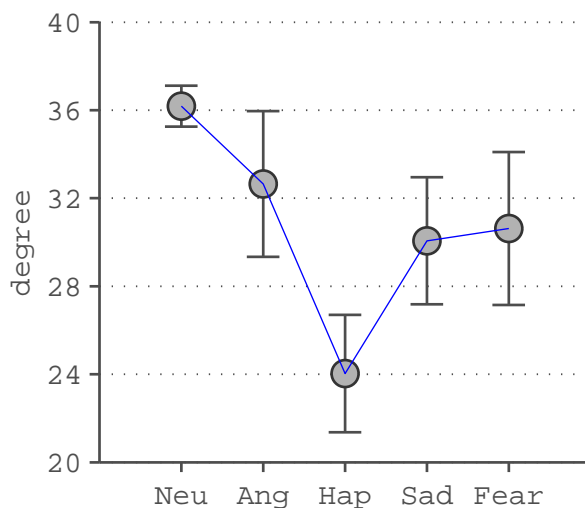


Figure 5.4: Errorbar plot of the “shadow” angle for each emotion.

neutrality, 32 for anger, 24 for happiness, 30 for sadness, and 30 for fear. The standard deviation of the angle is smallest for neutrality (0.93), and significantly greater for the other emotions: 3.31 for anger, 2.67 for happiness, 2.89 for sadness, and 3.48 for fear. This suggests that for a neutral speech condition, the shadow angle is fairly consistent, while it is relatively variable for emotional speech, within emotion as well as across emotions. This is an interesting finding and will be discussed in more detail in Section 5.6.

5.5 Analysis of emotional variability in the C/D model

In the C/D Model framework, emotional variation factors are a part of the utterance parameters, which cause variation of syllable magnitudes and IRF parameters (i.e., phase and magnitude of IRF peak). The variation of IRFs parameters, such as amplification (affected by the syllable magnitude) and timing (from the onset/coda excitation pulses), affect consonantal gestures. It follows from this that (i) the time-shifting of the IRFs influences the location of the maximum speed time points of the critical articulators and (ii) the amplification of the IRFs influences the speed (i.e., increases the speed) of the critical articulators at the iceberg point. Since the IRFs are hidden, the present paper analyzes the surface phenomenon directly. The goal is to understand the effects of emotion to the relative timing and speed of the iceberg points in syllables.

First, we investigate the effects of emotion on syllable magnitudes. Figure 5.5 shows the syllable magnitudes for each syllable, for each emotion condition. Overall, happiness shows the greatest syllable magnitude (jaw displacement), while anger shows the smallest. It would seem for happiness, the speaker uses greater

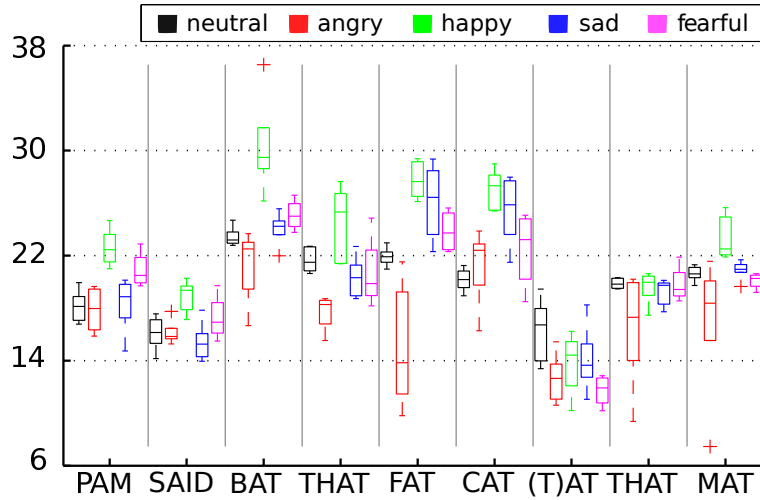


Figure 5.5: Syllable magnitude, as jaw excursion, for each mono-syllabic word in the utterance

jaw movement and for anger, this speaker speaks with a “clenched jaw,” a term often used in novels to describe expressions of cold anger. Note that although the syllable magnitude is smaller for anger compared to that of the other emotions, the speed and excursion of the critical articulators (as shown in Figure 5.3) is not significantly smaller. This finding hints that the emotional factor, e.g., the one resulting in the clenched jaw for anger, causes the variation of the relationship between the syllable magnitude and the amplitude of articulatory gesture (exhibited in the speed of the critical articulators at the iceberg point).

We further investigated the relationship between the speed of critical articulator and the syllable magnitude for different emotions. Figure 5.6 shows the ratio of the speed of the critical articulator (at CV/VC iceberg point) to syllable magnitude for each sample. This figure indicates that the ratio varies significantly depending on emotion. Note that the syllable magnitude is an indicator of syllable prominence in the articulatory domain. Also, note that the articulatory speed at the iceberg point is the maximum speed value of the critical articulator. Overall, the ratio for emotional speech (anger, happiness, sadness, fear) is greater than the ratio

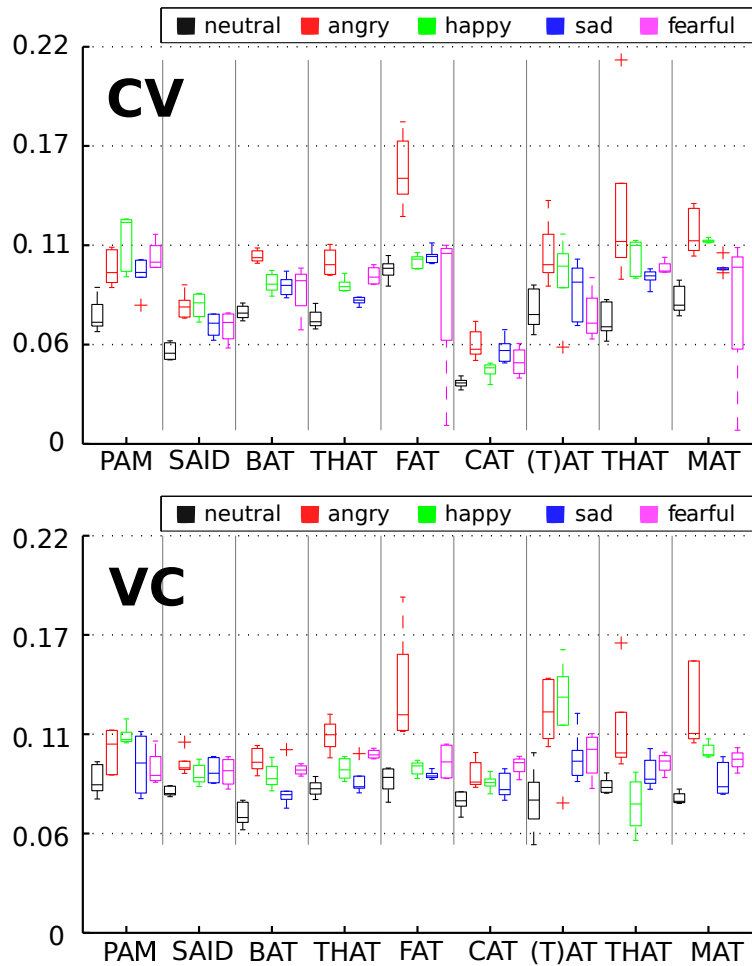


Figure 5.6: Ratio of articulatory speed (at the iceberg point for CV/VC demissyllable) to the syllable magnitude for each demissyllable.

for neutral speech, indicating that the ratio of the releasing speed of the critical articulator to the syllable magnitude is greater when the subject is emotional. This implies that the speaker tends to articulate with stronger consonantal gestures for critical articulators when the person is emotionally charged. This tendency is more consistent across CV demissyllables than VC demissyllables. In sum, results suggest that the maximum speed of critical articulators given syllable prominence varies depending on emotion.

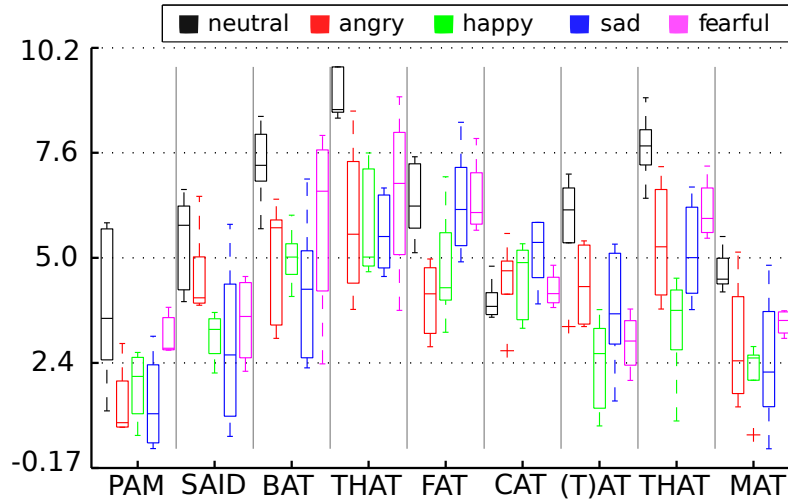


Figure 5.7: The top panel shows the time difference between the onset pulse point and the iceberg point.

Finally, we examined the time variation (τ) between the onset pulse, i.e., syllable triangle edge, to the iceberg point of CV demisyllable. Note that τ should be the same as the time difference between the coda and the iceberg point of VC demisyllable. This information is useful in the sense that it directly relates the abstract representation for temporal structure of an utterance to the surface phenomenon of articulatory movements. Figure 5.7 shows box plots of τ for each syllable. The mean of τ is greater for neutrality than for the other emotions in all cases, except ‘CAT.’ τ is a function of the shadow angle and the syllable magnitude: A larger shadow angle and greater syllable magnitude cause greater τ , which is in line with our previous observations in Figure 5.4 and Figure 5.5. For example, anger shows a smaller shadow angle and smaller syllable magnitude (due to idiosyncratic “clenched jaw” of the speaker) than neutrality, so τ of anger is also smaller than τ of neutrality.

5.6 Discussion and future works

In the analyses of this paper, we observed that emotion influences the shadow angle, syllable magnitude, the ratio of the maximum speed of the critical articulator in demisyllables to the syllable magnitude, and τ . One hypothetical reason for the variation of the shadow angle is that the assumption of the linear dependency between the syllable magnitude and the articulatory syllable duration is not valid in emotional speech. More specifically, jaw excursion may not be linearly dependent on the syllable duration in emotional speech, e.g., in clenched jaw for anger. This may point to the need of more comprehensive representation for the syllable magnitude in the C/D model framework for emotional speech. Another hypothetical reason is the conventionally applied assumption that the angles of the CV demisyllable and the VC demisyllable are identical is not valid in emotional speech. Bonaventura [2003] has raised the possibility of the two angles' asymmetry in the phrase-final elongation. In fact, the symmetry has been assumed for the simplicity of analysis, not for the theoretic or algorithmic necessity in the C/D model framework.

According to the C/D model, smaller shadow angle given the same syllable duration indicates greater syllable magnitude, thereby greater jaw excursion and faster maximum speed of critical articulator. This finding is in line with the faster articulatory movement and the greater movement range for happiness reported in previous studies [Kim et al., 2010, Lee et al., 2005, Kim et al., 2011a, Lee et al., 2006].

It should be noted that the variation of these parameters mentioned above is not independent from each other. The shadow angle is a function of the syllable magnitude and the time gap between the closest edges of adjacent syllable triangles. The maximum speed of the critical articulator is a function of the IRFs, which

is affected by the syllable magnitude. τ is also a function of the IRFs. Hence, a *joint* analysis and modeling for the variation of these parameters as a function of emotion, is important to represent emotional variability in the C/D model framework. An articulatory re-synthesis experiment with emotion transformation can be useful for evaluating the joint model. These constitute future work to be explored.

Chapter 6

Rich inversion using co-registered multimodal speech production data

6.1 Introduction

Recently, using articulatory information for recognition task has been drawing attention in the information processing for speech signal. Previous studies have shown that using speech articulatory signals can improve accuracy of various automatic recognition, e.g., linguistic and para-linguistic decoding tasks. For example, using the direct recording of articulatory signals on top of standard speech acoustic features, i.e., Mel-Frequency Cepstral Coefficients (MFCCs), improved phone recognition accuracy significantly [Zlokarnik, 1995]. Also, even predicted articulatory positions were useful to improve various recognition tasks, e.g., for phone [Ghosh and Narayanan, 2011], emotion [Kim et al., 2012b], speaker ID [Li et al., 2013a, 2015], palatal shape [Li et al., 2013b], and Parkinson’s condition [Hahm and Wang, 2015] of the speaker. Using articulatory information was also useful for the robustness against channel noise for speech recognition [Mitra et al., 2011].

One of the limitations for the inversion modeling lies on the difficulty of obtaining high quality of both acoustic and articulatory signals. Inversion modeling

requires multimodal data, both speech acoustic signal and articulatory signal. However, the state-of-the-art technologies for articulatory recording can offer only limited articulatory information. For example, ElectroMagnetic Articulography (EMA) offers the 3-dimensional (3D) coordinates of a handful of anatomical points monitored by pellets. Real-time Magnetic Resonance Imaging (rtMRI) [Narayanan et al., 2004] provides the complete view of a plane with fast frame rate, but simultaneous recording of speech audio involves scanning noise. Simultaneous recording using multiple data acquisition modalities is not often feasible due to technological limitations or compatibility issues such as in the case of EMA and rtMRI.

This study proposes a methodology of using co-registration technique for combining information of multiple modalities (e.g., EMA and rtMRI). The co-registered data will contain (i) clean speech audio (from EMA dataset), (ii) 3D tracking of a handful of anatomical landmarks on the vocal tract (from EMA) and (iii) the complete view of the upper airway (from rtMRI). We used the EMA and rtMRI datasets collected from the same speakers with the same stimuli, but in different time. The co-registered data from the two datasets was created by using recently proposed temporal alignment technique, namely Joint Acoustic and Articulatory Temporal Alignment (JAATA) [Kim et al., 2013]. For rtMRI data, linguistically important articulatory parameters will be extracted in batch using a robust vocal tract segmentation algorithm we developed. The advantage of learning the inversion mapping on the co-registered data is two folds: (i) The inversion model is capable of predicting various kinds of articulatory information, and (ii) the model is more useful for real applications, because clean speech audio (from EMA) can be directly usable as input signal (Speech audio in rtMRI dataset suffers from scanning noise or artifact from noise cancellation).

The present paper examines the capability of Deep Neural Network (DNN) regression model for the estimation of *rich* vocal tract information, i.e. vocal tract parameters obtained from co-registered data, from speech waveform. DNN has been shown satisfactory prediction accuracy for learning highly complicated relationship between parallel feature streams. The relationship between acoustic and rich vocal tract data is highly non-linear and complex, so DNN fits this problem well. We explore various structures of DNN for more accurate estimation of the vocal tract parameters. For an application, we perform preliminary test if the predicted rich articulatory information can boost emotion classification accuracy.

This study is organized as follows: Section 6.2 offers a summary of previous work on acoustic-to-articulatory inversion. Section 6.3 describes the co-registration method. Section 6.4 describes the articulatory parameter extraction method for the MR images Section 6.5 presents our rich inversion modeling scheme. Section 6.6 offers experimental setup. Section 6.7 discusses the prediction accuracy of the rich inversion model. Section 6.8 examines the benefits of using the predicted rich articulatory information on emotion classification. Section 6.9 concludes with the summary and future works.

6.2 Related works

Previous studies have evaluated various statistical models for acoustic-to-articulatory inversion. The inversion models in literature include a Gaussian mixture model [Toda et al., 2008], the Hidden Markov Model combined with maximum-likelihood parameter generation algorithm [Youssef et al., 2011], a generalized smoothness criterion [Ghosh and Narayanan, 2010], a DNN [Uria et al., 2011], a Recurrent Neural Network [Liu et al., 2015, Najnin and Banerjee, 2015],

a codebook [Ouni and Laprie, 2005], and a dynamic Kalman smoothing [Özbek et al., 2011]. These studies have performed inversion using EMA data, e.g., the MOCHA-TIMIT corpus [Wrench, 2000] and the MNGU0 corpus [Richmond, 2009]. However, EMA data contain spatially limited articulatory information (3-D coordinates of a handful of anatomical points), hence the articulatory information predicted in these studies is not rich.

The present paper focuses on another important problem for inversion, that is improving the amount of information to be predicted, aka. *rich* inversion problem. Training rich inversion model requires collecting rich articulatory information. The challenge for collecting rich articulatory information is technical limitations or incompatibility for using multiple modalities simultaneously. Although there was a success for ultrasound and EMA [Aron et al., 2006], and ultrasound and facial video [Hueber et al., 2012], it is still not feasible for many other modalities, e.g., EMA and rtMRI. The present paper consider an alternative approach for this case by collecting data using individual modalities, then registering the two collections afterwards. The registration algorithm we have developed will be discussed in the following section.

6.3 Co-registration

We used the EMA and rtMRI datasets of one female speakers (denoted as ‘F1’ in the corpus) in the USC-TIMIT corpus [Narayanan et al., 2014]. The EMA dataset contain the three-dimensional coordinates of six articulatory sensors and simultaneously recorded clean speech audio, while the rtMRI dataset contain the Magnetic Resonance (MR) images of the upper airway and simultaneously recorded and noise-cancelled speech audio. The frame rates of the EMA data and MR

images are 100 Hz and 23.180 Hz, respectively. See [Narayanan et al., 2014] for more details of the datasets.

Initially, identical utterance pairs from the two datasets are temporally aligned as follows: First, we created word sequences (transcription) for each sentence manually by listening to the speech audio. Second, we selected 411 utterance pairs whose word sequences are identical in EMA and rtMRI datasets. Third, we performed a temporal alignment for each pair of utterances using Joint Acoustic-Articulatory based Temporal Alignment (JAATA) method [Kim et al., 2013]. This algorithm utilizes both acoustic and articulatory information for aligning two feature streams from different modalities. Specifically, this iterates dynamic time warping and automatic feature extraction (from MRI data) in order to find an optimized alignment map. We evaluated the temporal alignment accuracy of JAATA based on the Averaged Word Boundary Distance (AWBD); AWBD is the root mean square value of the difference between true and estimated word boundaries in EMA audio. We considered word-final-time stamps that were automatically generated by forced alignment as the true word boundaries. For each of EMA and rtMRI audio, we used an adaptive forced aligner, SailAlign [Katsamanis et al., 2011]. The estimated word boundaries were obtained by mapping the word boundaries of MRI audio onto EMA data, using the alignment output of JAATA.

Finally, co-registered data is obtained by JAATA on a well-behaving utterance pairs. In order to obtain high quality of registered data, we performed temporal alignment twice: The first batch on the entire pairs, then the second batch on a subset after discarding poorly aligned pairs in the first batch. It is noted the automatic feature extraction in JAATA is dependent on the temporal alignment accuracy. Hence, discarding poorly aligned pairs can improve the alignment performance significantly. The utterance pairs were discarded for the second batch

Table 6.1: AWBDs of JAATA and baseline system on two sets of data: the entire utterance pairs (denoted by ‘All’) and a subset of utterance pairs (denoted by ‘Subset’). The unit of AWBD value is msec.

	All	Subset
Baseline	105.5	59.0
JAATA	100.4	48.6

if AWBDs of both JAATA and baseline system were equal to or greater than a threshold, that is $2 \times$ MR frame period ($2 / 23.180 = 86.28$ msec). Finally, 332 utterance pairs were selected for co-registered data.

The performance of JAATA was compared with the performance of a baseline system, that is dynamic time warping on the same acoustic features (13-dimensional MFCCs). Table 6.1 shows AWBDs for each batch and each alignment method. JAATA shows lower AWBD than the baseline on the entire utterance pairs. The improvement ratio by using JAATA than using the baseline system is even greater on the (cleaned) subset of utterance pairs: 17.6% lower AWBD for F1 and 31.6% lower AWBD for M1. The improvement ratio by using JAATA than using the baseline system is even greater on the (cleaned) subset of utterance pairs: 17.6% lower AWBD for F1. This improvement ratio and the boundary error value

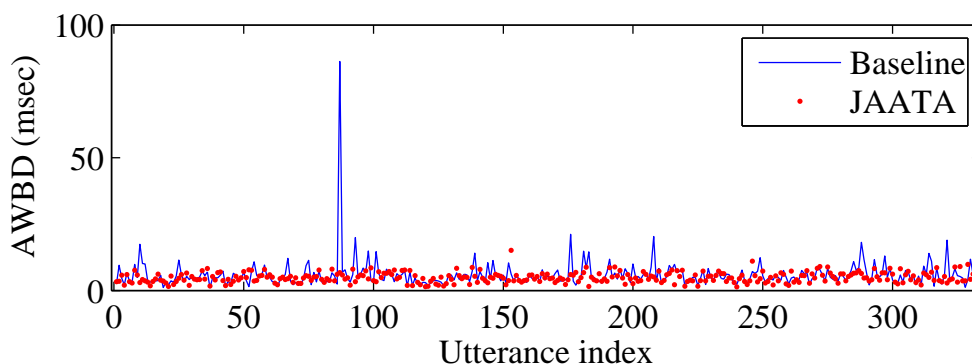


Figure 6.1: AWBDs of the baseline system (DTW + MFCCs) and JAATA for each utterance

are comparable to the cases when we used more strictly selected utterance pairs in previous studies [Kim et al., 2013, 2014c]. Figure 6.1 shows the AWBDs of the baseline system and JAATA for each utterance of the subset, indicating that in most cases, JAATA is capable of reducing significant temporal alignment error of the baseline system.

6.4 MR image parameterization

Rich vocal tract parameters were extracted automatically from the co-registered articulatory data. The rich information examined in this study comprises (i) 2D coordinates of six anatomical points (from EMA), (ii) distance function of the oropharyngeal airway (from rtMRI), (iii) lips and larynx positions (from rtMRI), (iv) vocal tract length and (v) oropharyngeal airway shape. A robust MR image segmentation algorithm [Kim et al., 2014b] was used in order to perform automatic tissue-airway segmentation in the oropharyngeal region. This algorithm initially estimates an oropharyngeal airway path between the outer and inner tissue-airway boundary robustly, using dynamic programming algorithm on smoothed the cost function. We computed the distance function by measuring the Euclidean distance between outer and inner tissue-airway boundaries. We considered the initial and final points in the oropharyngeal region as the lip and larynx positions, respectively. We measured the vocal tract length for each MR image by computing the sum of geodesic distance between adjacent center points of the upper airway. Finally, we computed the upper airway shape by measuring acute angles of three neighboring points on the center of the upper airway.

6.5 Rich inversion model

This study examines the capability of a DNN regressor for the rich inversion. We followed a conventional way of training the DNN regressor [Uria et al., 2011]. We initialized the weights of DNN by pre-training of a Deep Brief Network (DBN) model, followed by tuning the weights using stochastic gradient descent and the backpropagation algorithm. The DBN is a stacked Restricted Boltzmann Machines (RBMs), where a Gaussian-Bernoulli RBM and Bernoulli-Bernoulli RBMs are used for the bottom layer (visible units to hidden units) and higher layers (hidden units to hidden units) of the DBN, respectively. For efficient DBN training, we used Contrastive Divergence (CD) learning [Hinton, 2002] which typically reaches to a local minimum of the objective function (i.e., mean-squared-error form for regression) faster than the maximum likelihood learning. In order to minimize overfitting problem due to the limited amount of training data, we used the L2 regularization during DNN training. A linear regressor was used for the top layer of DNN.

Figure 6.2 shows the proposed training process for the inversion model.

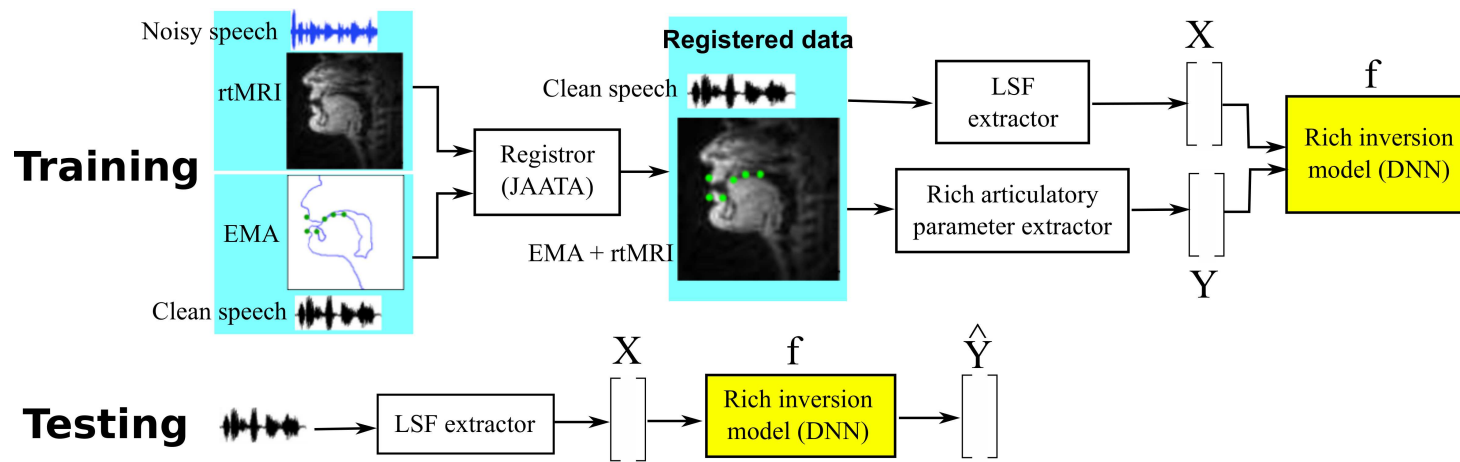


Figure 6.2: Rich inversion model training and testing processes

6.6 Experimental setup

The input (acoustic features) and output (articulatory features) features for the DNN regressor were created as follows. For acoustic features, we initially computed 41-dimensional acoustic features: 40 Linear Spectral Frequencies (LSFs) [Strube, 1980] and the gain of residual signal, which were extracted with a window size of 25 msec and a window shifting of 10 msec. In order to offer contextual information of acoustic feature stream, we created a context feature vector by concatenating 5 successive acoustic frames where the present frame corresponds to the third frame. Therefore, each context feature vector spans a period of 50 msec. We have compared 50-msec context and 100-msec context [Uria et al., 2011], but there was not significant difference in terms of prediction accuracy.

EMA data and the vocal tract parameters extracted using automatic process in Section 6.4 can be noisy. Hence, we discarded noisy feature frames empirically in order to minimize their effect as follows: We discarded frames if any articulatory feature values in the frames were outside $\pm 3 \times$ standard deviation from their means. The means were computed from the training set. Finally, the input (acoustic) and output (articulatory) features are normalized by z-scoring, where the mean and standard deviation of the training set were used for the normalization.

The co-registered data splitted into the training set (200 utterances, 60% of the total), the development set (66, 20%) and the test set (66, 20%). We have tested 6 different structures of DNNs: 2 or 3 layers of the same number of neurons, and 200 and 800 neurons for each layer. The optimal values of hyper-parameters, such as learning rate, momentum, the number of epochs, batch size, and L2 regularization parameter, for model training were empirically chosen based on the Pearson's correlation coefficient on the development set.

6.7 Results

Initially, we compared the prediction accuracy of two approaches: (i) training inversion model using only EMA parameters and (ii) training the model using rich articulatory parameters. The goal of this experiment is to check the benefits of offering richer articulatory information into the network in terms of prediction accuracy.

Figure 6.3 shows the prediction accuracy in terms of the average of Pearson correlation. Overall, training with rich articulatory features shows higher accuracy than doing with only EMA features. This suggests that the model is capable of using information from non-EMA features to predict EMA features better, indicating the benefit of using rich articulatory data for acoustic-to-articulatory inversion for higher prediction accuracy.

Next, the predicted articulatory parameters of the best DNN system (2 hidden layers with 800 neurons on each layer) were smoothed using a Butterworth low pass filter in order to improve the accuracy in terms of Pearson’s correlation. The trajectories of the articulatory parameters are smooth and slowly varying in

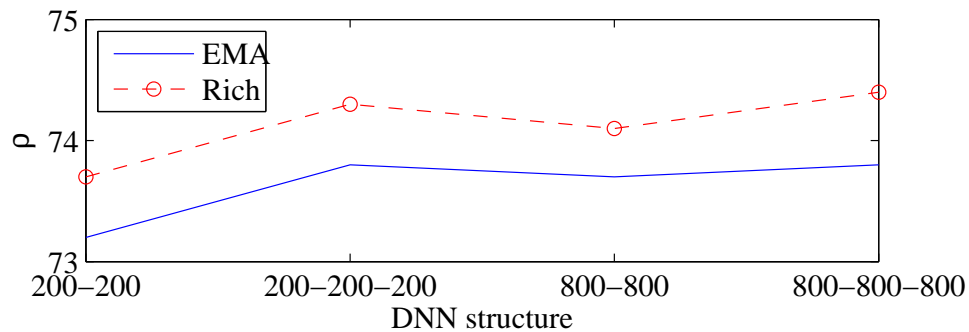


Figure 6.3: Averaged Pearson correlation coefficient (%) between true and predicted parameters of EMA, computed for different DNN structures and feature sets (only EMA parameters or entire rich articulatory parameters for output) for model training

Table 6.2: The Pearson’s correlation coefficient (%) of each articulatory parameters for the best DNN system before LPF (denoted by ‘None’) and after LPF (denoted by ‘LPF’). ‘Freq.’ denotes the cut-off frequency of LPF.

	TTx	TTy	TBx	TBy	TDx	TDy	ULx	ULy	LLx	LLy	Llx	Lly	Mean
None	76.8	84.7	77.8	85.9	79.1	76.8	56.3	53.8	72.0	75.1	69.5	83.3	74.3
LPF	80.2	87.3	81.2	88.6	82.2	80.5	61.4	57.3	75.8	77.4	77.4	85.6	77.5
Freq.	4.53	5.28	4.50	5.14	4.50	5.57	3.09	4.43	3.92	7.02	4.17	5.72	

nature due to physical constraints on the articulators. Also, temporal smoothing has been useful for achieving higher prediction accuracy in literature [Richmond, 2002, Toda, 2004, Uria et al., 2011]. The optimal cut-off frequency and the order of the filter was tuned on the development set. Finally, the Pearson’s correlation coefficient of the best system (DNN with 800-800 neurons) is improved to 77.5% by the smoothing. Table 6.2 shows the optimal cut-off frequency computed for each dimension of the EMA parameter set, and Pearson’s correlation before/after the smoothing.

Next, we examined the prediction accuracy for each subset of the other articulatory features after the smoothing. The articulatory parameters are categorized into (i) 2D coordinates of six anatomical points (from EMA), (ii) distance function of the oropharyngeal airway (from rtMRI), (iii) lip protrusion (from rtMRI), (iv) laryngeal height (from rtMRI), (v) vocal tract length and (vi) oropharyngeal airway

Table 6.3: The (averaged) Pearson’s correlation coefficient (%) of each subset of articulatory parameters after smoothing. ‘EMA’ refers to parameters of anatomical point tracking; ‘DF’ refers to parameters of distance function. ‘LP’ refers to lip protrusion parameter. ‘LH’ refers to laryngeal height parameter. ‘VTL’ refers to the parameter of the vocal tract length. ‘SHP’ refers to the parameters of oropharyngeal airway shape.

EMA	DF	LP	LH	VTL	SHP
77.5	68.8	37.2	10.4	55.6	14.0

shape. Table 6.3 shows the prediction accuracy in terms of the average of Pearson correlation of F1’s data as an example. First, prediction accuracy is highest on the EMA parameters. This is expected, because EMA parameters are free from any temporal alignment error and parameterization error. Overall, the Pearson’s correlation on the distance function is high (68.8%). In particular, the averaged correlation in the second-half of the distance function parameters is 66.5%. This is encouraging result, because the distance function offers the information of the vocal tract shaping for the locations that are hard to be monitored using EMA (e.g., back of the tongue). Vocal tract length parameter is also well predicted (55.6%), which is also information that the EMA data cannot offer.

However, the proposed system is not capable of well predicting the parameters of lip protrusion, laryngeal height and upper airway shaping. Although the lip protrusion parameter (‘LP’ in Table 6.3) was not well predicted (Pearson’s correlation is 37.2%), more accurate estimation of such information is still provided by the horizontal movements of the upper and lower lip, denoted by ULx and LLx in Table 6.2, respectively; The correlation values for ULx and LLx are 61.4% and 75.8%, respectively. The reasonably predicted upper airway shaping parameters (correlation is greater than 0.2) are limited to only 2-th to 7-th parameters, which are located from the lips to the alveolar ridge. This result indicates that the proposed model is not capable of predicting the oropharyngeal airway shape well. Other representation and prediction methods need to be explored in future study.

6.8 Application to emotion classification

This section explores an application of the rich inversion model for computational paralinguistics, in particular emotion classification. Specifically, we examine

whether the predicted rich articulatory parameters can boost the accuracy of emotion classification. This section also examines the prediction capability on the EMA data of a different speaker.

6.8.1 Experimental setup

The data of a female speaker JR, in Chapter 4, is selected, because the number of utterances is the greatest, compared to the other two speakers' data. This contains 470 speech utterances of five acted emotions: neutrality, hot anger, cold anger, happiness and sadness. See Section 4.2 in Chapter 4 for the detailed description of the data collection, post-processing, and emotion quality evaluation.

We will examine the benefit of using predicted articulatory information on top of acoustic features (directly computable from speech audio). The acoustic feature set was extracted using the openSMILE feature extractor [Eyben et al., 2010]. This comprises functionals of Low-Level Descriptor (LLD), i.e., prosodic, spectral, voice quality and voice source features. In total, 6374 features were initially extracted for each utterance.

The input acoustic features (LSFs) for DNN were computed using the same configuration in Section 6.6. Since speech region is of interest, we performed Voice Activity Detection (VAD) based on extracted f0 feature; The frames of non-zero f0 value with 50-msec margin were considered as speech region. The articulatory parameters were, then, estimated using the best DNN system (2 hidden layers with 800 neurons on each layer) in Section 6.7.

Some functionals are highly correlated, noisy and not-much useful for discriminating emotion. So, we performed feed-forward feature selection using 2-fold cross-validation on the training set. The Support Vector Machine (SVM) with a radial basis function kernel was used for the emotion classifier. Eighty percent of data

Table 6.4: The Pearson’s correlation coefficient (%) of each articulatory parameters on JR’s data.

TTx	TTy	TBx	TBy	TDx	TDy	ULx	ULy	LLx	LLy	Lix	Liy	Mean
36.4	35.3	27.2	65.6	38.3	62.3	33.0	9.6	42.3	54.0	33.1	58.0	41.3

was used for training the SVM classifier, and the remaining 20% was used as test set in a five-fold cross-validation setup.

6.8.2 Results

Table 6.4 shows the Pearson’s correlation between true EMA sensor parameters and predicted parameters using the best DNN, after smoothing. Overall, the prediction accuracy decreases significantly for the entire dimension, which indicates the needs for speaker normalization or speaker-independent prediction scheme. This will be discussed in Section 6.9.

Next, we examined the usefulness of the predicted rich articulatory information for improving the accuracy of emotion classification. Table 6.5 shows the unweighted classification accuracy on different set of features. The by-chance accuracy is 24.5%. the predicted articulatory features show lower classification accuracy on the test set than the baseline features. However, the best accuracy was achieved by the fusion system which improves the accuracy by 7.9% from the baseline features. McNemar’s test result indicates this improvement of performance from baseline system is statistically significant ($\chi^2 = 25.1$, $p < 5e-06$). This

Table 6.5: Unweighted accuracy (%) of emotion classification. ‘Baseline’ is for the baseline openSMILE acoustic feature set; ‘Arti’ is for the predicted rich articulatory feature set; ‘Fusion’ is for the feature-level fusion of the two sets.

Baseline	Arti	Fusion
70.6	69.8	78.5

result suggests that the predicted rich articulatory information contains complementary information than what acoustic features offer, resulting in boosting the emotion classification accuracy.

6.9 Conclusions and future works

This study explores a framework of acoustic-to-articulatory inversion where *rich* articulatory information from multiple modalities (EMA and rtMRI) is estimated. First, the optimal temporal alignment map between EMA and rtMRI data is obtained using JAATA, which generates more accurate alignment map than using DTW on MFCCs. Next, articulatory parameters from EMA and rtMRI, as well as acoustic features from clean speech audio of EMA are used for training the inverse mapping function. The proposed method of using DNN is capable of estimating articulatory parameters for EMA parameters, distance function for the oropharyngeal airway and the frame-level vocal tract length well, but not for the lip protrusion, laryngeal height and airway shaping parameters.

It is encouraging that the predicted articulatory parameters are capable of improving emotion classification accuracy even without speaker normalization. Speaker normalization [Afshan and Ghosh, 2015] and/or adaptation [Saon et al., 2013] may be able to improve the accuracy of articulatory prediction. Also, the state-of-the-art performance of inversion has been achieved by using recurrent neural network [Liu et al., 2015, Najnin and Banerjee, 2015], which has been tested on EMA data only. Since the recurrent neural network learns the dynamics of feature sequences, which is also important nature of the vocal tract shaping, it is worth to try on the rich articulatory parameters. Finally, a better representation of the articulatory parameter space can be explored. Embedding to a hidden space

using non-linear canonical correlation analysis [Andrew et al., 2013] or restricted Boltzmann machines [Ngiam et al., 2011] has shown its success on improving the inverse mapping. Exploring these methods for our rich inversion framework is a part of our on-going works.

Reference List

- Amber Afshan and Prasanta Kumar Ghosh. Improved subject-independent acoustic-to-articulatory inversion. *Speech Communication*, 66:1 – 16, 2015.
- G. Ananthakrishnan and O. Engwall. Important regions in the articulator trajectory. In *Proceedings of International Seminar on Speech Production*, pages 305 – 308, 2008.
- Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1247–1255, 2013.
- Michael Aron, Erwan Kerrien, Marie-Odile Berger, and Yves Laprie. Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition set up and preliminary results. In *Proceedings of International Seminar of Speech Production*, 2006.
- Rainer Banse and Klaus R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–636, March 1996.
- Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouviet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.
- Patrizia Bonaventura. *Invariant patterns in articulatory movements*. PhD thesis, Ohio State University, 2003.
- Patrizia Bonaventura and Osamu Fujimura. *Articulatory movements and phrase boundaries*, chapter 14, pages 209–227. Oxford linguistics. OUP Oxford, 2007. in *Experimental Approaches to Phonology*, edited by Sole, M.J. and Beddor, P.S. and Ohala, M.
- E. Bresch, Yoon-Chul Kim, K. Nayak, D. Byrd, and S. Narayanan. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *Signal Processing Magazine*, 25(3):123 – 132, 2008.

- Erik Bresch, Jon Nielsen, Krishna S. Nayak, and Shrikanth S. Narayanan. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, 120(4):1791–1794, Oct 2006.
- C. P. Browman and L. Goldstein. Articulatory gestures as phonological units. *Haskins Laboratories Status Report on Speech Research*, SR-99/100:69 – 101, 1989.
- Catherine P Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180, 1992.
- C. Busso, C. Lee, and S. S. Narayanan. An analysis of emotionally salient aspects of fundamental frequency for emotion detection. *Transactions on Audio, Speech, and Language Processing*, 7:4:582 – 596, 2009.
- Jun Cai, Yves Laprie, Julie Busset, and Fabrice Hirsch. Articulatory modeling based on semi-polar coordinates and guided PCA technique. In *Proceedings of Interspeech*, pages 56 – 59, Brighton, UK, 2009. ISCA.
- John Canny. A computational approach to edge detection. *Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.
- R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine*, 18(1):32 – 80, Jan 2001.
- Richard J Davidson and William Irwin. The functional neuroanatomy of emotion and affective style. *Trends in cognitive sciences*, 3(1):11–21, 1999.
- Jane Edwards, Henry J Jackson, and Philippa E Pattison. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review*, 22(6):789–832, 2002.
- D. Erickson, A. Abramson, K. Maekawa, and T. Kaburagi. Articulatory characteristics of emotional utterances in spoken english. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 365 – 368, 2000.
- D. Erickson, C. Menezes, and A. Fujino. Some articulatory measurements of real sadness. In *Proceedings of Interspeech*, pages 1825 – 1828, Korea, 2004. ISCA.
- D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya. Exploratory study of some acoustic and articulatory characteristics of sad speech. *Phonetica*, 63(1):1 – 25, 2006.

- D. Erickson, S. Kawahara, J. Moore, C. Menezes, A. Suemitsu, J. Kim, and Y. Shibuya. Calculating articulatory syllable duration and phrase boundaries. In *Proceedings of International Seminar on Speech Production*, pages 102–105, Cologne, Germany, May 2014.
- D. Erickson, J. Kim, S. Kawahara, I. Wilson, C. Menezes, A. Suemitsu, , and J. Moore. Bridging articulation and perception: The C/D model and contrastive emphasis. In *International Congress of Phonetic Sciences*, 2015.
- Donna Erickson. More about jaw, rhythm and metrical structure. In *Acoustical Society of Japan, fall meeting*, page 103, 2010.
- Donna Erickson, Osamu Fujimura, and Bryan Pardo. Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech*, 41(3-4):399–417, 1998.
- F. Eyben, M. Wöllmer, and B. Schuller. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, Firenze, Italy, 2010.
- Samantha L Finkelstein, Andrea Nickel, Lane Harrison, Evan Suma, Tiffany Barnes, et al. cmotion: A new game design to teach emotion recognition and programming logic to children using virtual humans. In *Virtual Reality Conference*, pages 249–250. IEEE, 2009.
- C. A. Fowler and E. Saltzman. *Language and Speech*, 36(2-3), 1993.
- J. Frankel and S. King. ASR - Articulatory speech recognition. In *Proceedings of Eurospeech*, volume 1, pages 599 – 602, 2001.
- Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412, 1985.
- Ada Wai-Chee Fu, Eamonn Keogh, Leo Yung Lau, Chotirat Ann Ratanamahatana, and Raymond Chi-Wing Wong. Scaling and time warping in time series querying. *The International Journal on Very Large Data Bases*, 17(4):899–921, Jul 2008.
- O. Fujimura and D. Erickson. The C/D model for prosodic representation of expressive speech in English. In *Proceedings of Acoustical Society of Japan. Fall meeting*, pages 271 – 272, Okinawa, September 2004.
- O Fujimura and WR Spencer. Effects of phrasing and word emphasis on transitional movements – location and stability of tongue blade iceberg patterns. *The Journal of the Acoustical Society of America*, 74(S1):S117–S117, 1983.

- O. Fujimura, S. Kiritani, and H. Ishida. Computer controlled radiography for observation of movements of articulatory and other human organs. *Computers in Biology and Medicine*, 3(4):371 – 384, 1973.
- Osamu Fujimura. Temporal organization of articulatory movements as a multidimensional phrasal structure. *Phonetica*, 38(1-3):66–83, 1981.
- Osamu Fujimura. *Relative invariance of articulatory movements: An iceberg model*, chapter 11, pages 226–242. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986. in *Invariance and Variability of Speech Processing*, edited by J. S. Perkell and D. Klatt.
- Osamu Fujimura. C/D model: A computational model of phonetic implementation. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 17: 1–20, 1994a.
- Osamu Fujimura. Syllable timing computation in the C/D model. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 519 – 522, Yokohama, Japan, September 1994b.
- Osamu Fujimura. Iceberg revisited. *The Journal of the Acoustical Society of America*, 99(4):2471–2500, 1996.
- Osamu Fujimura. The C/D model and prosodic control of articulatory behavior. *Phonetica*, 57(2-4):128–138, 2000.
- Osamu Fujimura. Temporal organization of speech utterance: A C/D model perspective. *Cadernos de Estudos Lingüísticos*, 43:9 – 35, 2002.
- Prasanta Kumar Ghosh and Shrikanth S. Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–2172, 2010.
- Prasanta Kumar Ghosh and Shrikanth S Narayanan. A subject-independent acoustic-to-articulatory inversion. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4624–4627. IEEE, 2011.
- Christer Gobl and Ailbhe NíChasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.
- Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10 - 11):787 – 800, 2007.

- Frank Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102:594 – 621, 1995.
- Christina Hagedorn, Michael I. Proctor, and Louis Goldstein. Automatic analysis of singleton and geminate consonant articulation using real-time magnetic resonance imaging. In *Proceedings of Interspeech*, pages 409–412. ISCA, 2011.
- Seongjun Hahm and Jun Wang. Parkinson’s condition estimation using speech acoustic and inversely mapped articulatory data. In *Proceedings of Interspeech*, pages 513–517. ISCA, 2015.
- Richard Harshman, Peter Ladefoged, and Louis Goldstein. Factor analysis of tongue shapes. *The Journal of Acoustical Society of America*, 62(3):693–707, 1977.
- Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.
- R Peter Hobson. The autistic child’s appraisal of expressions of emotion. *Journal of Child Psychology and Psychiatry*, 27(3):321–342, 1986.
- R Peter Hobson, J Ouston, and Antony Lee. Emotion recognition in autism: Coordinating faces and voices. *Psychological medicine*, 18(04):911–923, 1988.
- Thomas Hueber, Gérard Bailly, and Bruce Denby. Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. In *Proceedings of Interspeech*. ISCA, 2012.
- Philip J.B. Jackson and Veena D. Singampalli. Statistical identification of articulation constraints in the production of speech. *Speech Communication*, 51(8): 695 – 710, 2009.
- P.J. Jackson and V.D. Singampalli. Statistical identification of critical, dependent and redundant articulators. *The Journal of the Acoustical Society of America*, 123(5):3321 – 3321, 2008.
- Athanasios Katsamanis, Matthew Black, Panayiotis G. Georgiou, Louis Goldstein, and Shrikanth S. Narayanan. SailAlign: Robust long speech-text alignment. In *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, Jan 2011.
- J. A. S. Kelso, Eric Vatikiotis-Bateson, E., Elliot L. Saltzman, and Bruce Kay. A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *The Journal of the Acoustical Society of America*, 77(1):266–280, 1985. doi: <http://dx.doi.org/10.1121/1.392268>.

- J. Kim, S. Lee, D. Erickson, , and S. S. Narayanan. The C/D model and emotional speech: preliminary findings. In *Adventures in Speech Science Workshop*, page 43, Tokyo University, July 2015a.
- Jangwon Kim, Sungbok Lee, and Shrikanth S. Narayanan. A detailed study of word-position effects on emotion expression in speech. In *Proceedings of Interspeech*, pages 1987 – 1990, Brighton, United Kingdom, 2009.
- Jangwon Kim, Sungbok Lee, and Shrikanth S. Narayanan. A study of interplay between articulatory movement and prosodic characteristics in emotional speech production. In *Proceedings of Interspeech*, pages 1173 – 1176. ISCA, 2010.
- Jangwon Kim, Sungbok Lee, and Shrikanth S. Narayanan. An exploratory study of the relations between perceived emotion strength and articulatory kinematics. In *Proceedings of Interspeech*, pages 2961 – 2964, Florence, Italy, 2011a. ISCA.
- Jangwon Kim, Prasanta Ghosh, Sungbok Lee, and Shrikanth Narayanan. A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion. In *Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Los Angeles, USA, 2012a.
- Jangwon Kim, Prasanta Ghosh, Sungbok Lee, and Shrikanth S. Narayanan. A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion. In *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–4. IEEE, 2012b.
- Jangwon Kim, Adam C Lammert, Prasanta Kumar Ghosh, and Shrikanth S Narayanan. Spatial and temporal alignment of multimodal human speech production data: Real time imaging, flesh point tracking and audio. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3637–3641. IEEE, 2013.
- Jangwon Kim, Donna Erickson, Sungbok Lee, and Shrikanth Narayanan. A study of invariant properties and variation patterns in the Converter/Distributor model for emotional speech. In *Proceedings of Interspeech*, pages 413 – 417. ISCA, 2014a.
- Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth Narayanan. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *the 10th International Seminar on Speech Production (ISSP)*, pages 222 – 225, Cologne, Germany, 2014b.

- Jangwon Kim, Adam C. Lammert, Prasanta Kumar Ghosh, and Shrikanth S. Narayanan. Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 135(2):EL115–EL121, 2014c.
- Jangwon Kim, Sungbok Lee, and Shrikanth S. Narayanan. Estimation of the movement trajectories of non-crucial articulators based on the detection of crucial moments and physiological constraints. In *Proceedings of Interspeech*, pages 163 – 168, Singapore, 2014d. ISCA.
- Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan. USC-EMO-MRI corpus: An emotional speech production database recorded by realtime magnetic resonance imaging. In *Proceedings of the 10th International Seminar on Speech Production*, pages 226 – 229, 2014e.
- Jangwon Kim, Donna Erickson, and Sungbok Lee. More about contrastive emphasis and the C/D model. *Journal of the Phonetic Society of Japan*, 2015b. In Press.
- Jangwon Kim, Asterios Toutios, Sungbok Lee, and Shrikanth S. Narayanan. A kinematic study of critical and non-critical articulators in emotional speech production. *The Journal of the Acoustical Society of America*, 137(3):1411–1429, 2015c. doi: <http://dx.doi.org/10.1121/1.4908284>.
- Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak. Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order. *Magnetic Resonance in Medicine*, 65(5):1365–1371, 2011b.
- Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121(2):723–742, 2007.
- Marcel Kockmann, Lukáš Burget, et al. Application of speaker-and language identification state-of-the-art techniques for emotion recognition. *Speech Communication*, 53(9):1172–1185, 2011.
- Laura L. Koenig, Jorge C. Lucero, and Elizabeth Perlman. Speech production variability in fricatives of children and adults: Results of functional data analysis. *The Journal of the Acoustical Society of America*, 124(5):3158–3170, 2008.
- Sherrie Xiao Komiak and Izak Benbasat. Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. *Information Technology and Management*, 5(1-2):181–207, 2004.

- Stefan Kopp and Kirsten Bergmann. Towards an architecture for aligned speech and gesture production. In *Proceedings of the 7th international conference on Intelligent Virtual Agents, IVA*, pages 389–390, Paris, France, 2007. Springer-Verlag.
- Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. 23(3):559–568, 2004.
- William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- A. Lammert, M. Proctor, and S. Narayanan. Data-driven analysis of realtime vocal tract MRI using correlated image regions. In *Proceedings of Interspeech*, pages 1572 – 1575, Makuhari, Japan, 2010.
- Adam Lammert, Michael Proctor, and Shrikanth Narayanan. Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research*, 56(6):S1924–S1933, 2013.
- Adam C. Lammert, Michael I. Proctor, Athanasios Katsamanis, and Shrikanth S. Narayanan. Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact. In *Proceedings of Interspeech*, pages 2813–2816. ISCA, 2011.
- Eva Lasarcyk and Jürgen Trouvain. Spread lips + raised larynx + higher f0 = smiled speech? - An articulatory synthesis approach. In *International Symposium of Speech Production (ISSP)*, pages 43–48, 2008.
- M.J. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, M. Suhling, P. Hunziker, and M. Unser. Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. *Transactions on Medical Imaging*, 24(9):1113–1126, 2005.
- C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *Transactions on Speech and Audio Processing*, 13:2:293 – 303, 2009.
- Li Lee and Richard C Rose. Speaker normalization using efficient frequency warping procedures. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 353–356. IEEE, 1996.
- S. Lee, S. Yildirim, A. Kazemzadeh, and S. S. Narayanan. An articulatory study of emotional speech production. In *Proceedings of Interspeech*, pages 497 – 500. ISCA, 2005.
- Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, and Shrikanth S. Narayanan. A study of emotional speech articulation using a fast magnetic

- resonance imaging technique. In *Proceedings of Interspeech*, pages 2234 – 2237, Pittsburgh, PA, 2006. ISCA.
- Sungbok Lee, Tsuneo Kato, and Shrikanth S. Narayanan. Relation between geometry and kinematics of articulatory trajectory associated with emotional speech production. In *Proceedings of Interspeech*, pages 2290–2293, Brisbane, Australia, 2008. ISCA.
- Ming Li, Jangwon Kim, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth S. Narayanan. Speaker verification based on fusion of acoustic and articulatory information. In *Proceedings of Interspeech*, pages 1614–1618. ISCA, 2013a.
- Ming Li, Adam Lammert, Jangwon Kim, Prasanta Kumar Ghosh, and Shrikanth S. Narayanan. Automatic classification of palatal and pharyngeal wall shape categories from speech acoustics and inverted articulatory signals. In *Proceedings of ICASA Workshop on Speech Production in Automatic Speech Recognition*. ISCA, 2013b.
- Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth Narayanan. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. *Computer Speech & Language*, 2015. In Press.
- Johan Liljencrants. Fourier series description of the tongue profile. *Speech Transmission Laboratory-Quarterly Progress Status Reports*, 12(4):9–18, 1971.
- S.G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. S. Narayanan, and K. S. Nayak. High spatio-temporal resolution multi-slice real time MRI of speech using golden angle spiral imaging with constrained reconstruction, parallel imaging, and a novel upper airway coil. In *Proceedings of 23rd International Society of Magnetic Resonance in Medicine (ISMRM) Scientific Sessions*, page 689, 2015.
- Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, H. Meng, and Lianhong Cai. A deep recurrent approach for acoustic-to-articulatory inversion. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4450–4454. IEEE, April 2015.
- Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. Feature selection using principal feature analysis. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA, pages 301–304, New York, NY, USA, 2007. ACM.
- Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(34):199 – 229, 1982.

- Monica Mazza, Laura Giusti, Anna Albanese, Melania Mariano, Maria Chiara Pino, and Rita Roncone. Social cognition disorders in military police officers affected by posttraumatic stress disorder after the attack of An-Nasiriyah in Iraq 2006. *Psychiatry research*, 198(2):248–252, 2012.
- Caroline Menezes. *Rhythmic pattern Of American English: An articulatory and acoustic study*. PhD thesis, Ohio State University, 2003.
- Vikramjit Mitra, Hosung Nam, and Carol Y Espy-Wilson. Robust speech recognition using articulatory gestures in a dynamic Bayesian network framework. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 131–136. IEEE, 2011.
- Parham Mokhtari, Tatsuya Kitamura, Hironori Takemoto, and Kiyoshi Honda. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. *Journal of Phonetics*, 35(1):20 – 39, 2007. doi: <http://dx.doi.org/10.1016/j.jocn.2006.01.001>.
- Sylvie Jeannette Laure Mozziconacci. *Speech variability and emotion: Production and perception*. PhD thesis, Technische Universiteit Eindhoven, 1998.
- Sylvie JL Mozziconacci and Dik J Hermes. Expression of emotion and attitude through temporal speech variations. In *Proceedings of Interspeech*, pages 373 – 378. ISCA.
- Shamima Najnin and Bonny Banerjee. Improved speech inversion using general regression neural network. *The Journal of the Acoustical Society of America*, 138(3):EL229–EL235, 2015.
- H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. TADA: An enhanced, portable task dynamics model in Matlab. *The Journal of the Acoustical Society of America*, 115(5):2430 – 2430, 2004.
- Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 115(4):1771 – 1776, 2004.
- Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, and Michael Proctor. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America*, 136(3):1307 – 1311, 2014.

- W.L. Nelson, J.S. Perkell, and J.R. Westbury. Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations. *The Journal of the Acoustical Society of America*, 75(3):945–951, 1984.
- Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML)*, pages 689–696, 2011.
- Sven E. G. Öhman. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2):310–320, 1967. doi: <http://dx.doi.org/10.1121/1.1910340>.
- Slim Ouni and Yves Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 118(1):444–460, 2005.
- I. Yücel Özbek, Mark Hasegawa-Johnson, and Mübeccel Demirekler. Estimation of articulatory trajectories based on gaussian mixture model (GMM) with audio-visual information fusion and dynamic kalman smoothing. *Transactions on Audio, Speech, and Language Processing*, 19(5):1180–1195, 2011.
- A. Paeschke, M. Kienast, and W.F. Sendlmeier. F0-contours in emotional speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 929 – 932, San Francisco, U.S.A., 1999.
- G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92(2):688 – 700, 1992.
- J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M.T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078 – 3096, 1992.
- Joseph S Perkell. Articulatory processes. *The Handbook of Phonetic Sciences. Hardcastle, William J. and John Laver (eds).*, 1999.
- Joseph S Perkell and Dennis H Klatt. *Invariance and variability in speech processes*. Psychology Press, 2014.
- Rosalind W Picard and Jonathan Klein. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers*, 14(2):141–169, 2002.

- Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.
- Michael Proctor, Danny Bone, Nassos Katsamanis, and Shrikanth S Narayanan. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Proceedings in Interspeech*, pages 1576–1579. ISCA, 2010.
- Chao Qin and M.A. Carreira-Perpinan. Reconstructing the full tongue contour from EMA/X-ray microbeam. In *International Conference on Acoustics Speech and Signal Processing*, pages 4190–4193. IEEE, March 2010.
- Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth S. Narayanan. An investigation of articulatory setting using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 134(1): 510–519, 2013.
- Daniel Recasens. Vowel-to-vowel coarticulation in Catalan VCV sequences. *The Journal of the Acoustical Society of America*, 76(6):1624–1635, 1984.
- Daniel Recasens, Maria Dolors Pallars, and Jordi Fontdevila. A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, 102(1):544 – 561, 1997.
- Korin Richmond. *Estimating articulatory parameters from the acoustic speech signal*. PhD thesis, University of Edinburgh, 2002.
- Korin Richmond. Preliminary inversion mapping results with a new EMA corpus. In *Proceedings of Interspeech*, pages 2835–2838. ISCA, 2009.
- P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C Browman. CASY and extensions to the task-dynamic model. In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling - 4th Speech Prudction Seminar*, pages 125 – 128, 1996.
- Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43 – 49, 1978.
- E. Saltzman and J.A.S. Kelso. Skilled actions: A task dynamic approach. *Psychological Review*, 94:84 – 106, 1987.
- E. Saltzman and K.G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

- J.M. Santos, G.A. Wright, and J.M. Pauly. Flexible real-time magnetic resonance imaging framework. In *Proceedings of Engineering in Medicine and Biology Society*, volume 47, pages 1048–1051. IEEE, 2004.
- G. Saon, H. Soltan, D. Nahamoo, and M. Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 55–59. IEEE, 2013.
- Klaus R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227 – 256, 2003.
- Marc Schröder. Emotional speech synthesis: A review. In *Proceedings of Interspeech*, pages 561–564. ISCA, 2001.
- Mohammad Shami and Werner Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201 – 212, 2007.
- David R. R. Smith and Roy D. Patterson. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, 118(5):3177–3186, 2005. doi: <http://dx.doi.org/10.1121/1.2047107>.
- M. Stone. A guide to analyzing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 19(6-7):2128 – 2131, 2005.
- Brad H. Story. Vowel and consonant contributions to vocal tract shape. *The Journal of the Acoustical Society of America*, 126(2):825–836, 2009. doi: <http://dx.doi.org/10.1121/1.3158816>.
- Brad H. Story and Ingo R. Titze. Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics*, 26(3):223 – 260, 1998. doi: <http://dx.doi.org/10.1006/jpho.1998.0076>.
- Brad H. Story, Ingo R. Titze, and Eric A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–554, 1996.
- Hans Werner Strube. Linear prediction on a warped frequency scale. *The Journal of the Acoustical Society of America*, 68(4):1071–1076, 1980.
- J.D. Subtelny and N. Oya. Cineradiographic study of sibilants. *Folia phoniatrica*, 24(1):30–50, 1972.
- M. Tiede. Multi-channel visualization application for displaying dynamic sensor movements, 2010. In development.

- Black A. W. Tokuda K. Toda, T. Acoustic-to-articulatory inversion mapping with Gaussian mixture model. In *Proceedings of Interspeech*, pages 1129 – 1132. ISCA, 2004.
- Tomoki Toda, Alan W Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3):215–227, 2008.
- Asterios Toutios and Shrikanth Narayanan. Articulatory synthesis of French connected speech from EMA data. In *Proceedings of Interspeech*, pages 2738–2742. ISCA, 2013.
- Benigno Uria, Steve Renals, and Korin Richmond. A deep neural network for acoustic-articulatory speech inversion. In *Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain, December 2011.
- Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.
- Jean Vroomen, René Collier, and Sylvie JL Mozziconacci. Duration and intonation in emotional speech. In *Proceedings of Eurospeech*, 1993.
- J Westbury. *X-ray Microbeam Speech Production Database User’s Handbook*. 2005.
- Carl E. Williams and Kenneth N Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52:1238, 1972.
- Alan A. Wrench. A multichannel articulatory database and its application for automatic speech recognition. In *Proceedings of International Seminar of Speech Production*, pages 305–308, 2000.
- Yi Xu and Suthathip Chuenwattanapranithi. Perceiving anger and joy in speech through the size code. In *Proceedings of the International Conference on Phonetic Sciences*, pages 2105–2108, 2007.
- S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. S. Narayanan. An acoustic study of emotions expressed in speech. In *Proceedings of Interspeech*, pages 2193 – 2196. ISCA, 2004.
- Atef Ben Youssef, Thomas Hueber, Pierre Badin, and Gérard Bailly. Toward a multi-speaker visual articulatory feedback system. In *Proceedings of Interspeech*, pages 589–592. ISCA, 2011.

- J. Yuan and M. Liberman. Speaker identification on the scotus corpus. In *Proceedings of Acoustics*, pages 5687 – 5690, 2008.
- Feng Zhou and Fernando De la Torre Frade. Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems Conference (NIPS)*, December 2009.
- Igor Zlokarnik. Adding articulatory features to acoustic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, 97(5): 3246–3246, 1995.