# Statistical identification of articulatory roles in speech production

Veena D Singampalli

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

January 2010

# Declaration

I declare that, apart from where properly indicated, the work presented in this thesis is entirely my own.

# Dedication

To my grandfather and my parents.

# Acknowledgements

I would like to thank my supervisor, Dr Philip Jackson, first of all, for giving me this opportunity to work with him. I would not have reached my goals with out his continuous guidance, support, encouragement, constructive criticism and feedback in every phase of this research. I have learnt a lot from him in every way and would like to express my sincere gratitude to him for supervising my research. My sincere thanks to my examiners, Dr Simon King and Prof Adrian Hilton, for their valuable comments, suggestions and feedback. Thanks to Prof Hilton for his support as a mentor during the course of my study.

Thanks to CSTR, University of Edinburgh, for providing the database for this work. Thanks to Dr Elliot Saltzman and Dr Philip Rubins for granting permission to reproduce some of their figures for this thesis. Thanks to Dr Jonathan Pincas, Dr Wenwu Wang and Ms Claire Turner for theoretical discussions and feedback. Thanks to colleagues and friends at CVSSP for all the help and support.

Thanks to the EPSRC and the CVSSP for the financial support which allowed me to undertake this research.

Thanks to all my family, especially my parents, Murali Mohan and Syamala Devi, my brother, Kalyan, and my husband, Saravan, for their unconditional love, support and encouragement.

# Abstract

The human speech apparatus is a rich source of information and offers many cues in the speech signal due to its biomechanical constraints and physiological interdependencies. Coarticulation, a direct consequence of these speech production factors, is one of the main problems affecting the performance of speech systems. Incorporation of production knowledge could potentially benefit speech recognisers and synthesisers. Hand coded rules and scores derived from the phonological knowledge used by production oriented models of speech are simple and incomplete representations of the complex speech production process. Statistical models built from measurements of speech articulation fail to identify the cause of constraints. There is a need for building explanatory yet descriptive models of articulation for understanding and modelling the effects of coarticulation.

This thesis aims at providing compact descriptive models of realistic speech articulation by identifying and capturing the essential characteristics of human articulators using measurements from electro-magnetic articulography. The constraints on articulators during speech production are identified in the form of critical, dependent and redundant roles using entirely statistical and data-driven methods. The critical role captures the maximally constrained target driven behaviour of an articulator. The dependent role models the partial constraints due to physiological interdependencies. The redundant role reflects the unconstrained behaviour of an articulator which is maximally prone to coarticulation. Statistical target models are also obtained as the by-product of the identified roles.

The algorithm for identification of articulatory roles (and estimation of respective model distributions) for each phone is presented and the results are critically evaluated. The identified data-driven constraints obtained are compared with the well known and commonly used constraints derived from the IPA (International Phonetic Alphabet). The identified critical roles were not only in agreement with the place and manner descriptions of each phone but also provided a phoneme to phone transformation by capturing language and speaker specific behaviour of articulators. The models trained from the identified constraints fitted better to the phone distributions (40% improvement) .

The evaluation of the proposed search procedure with respect to an exhaustive search for identification of roles demonstrated that the proposed approach performs equally well for much less computational load. Articulation models built in the planning stage using sparse yet efficient articulatory representations using standard trajectory generation techniques showed some potential in modelling articulatory behaviour. Plenty of scope exists for further developing models of articulation from the proposed framework.



**Key words:** Critical, dependent and redundant articulators, articulatory constraints, coarticulation, speech production, articulatory modelling

# Contents

# List of symbols

| | |
|---|---|
| $a$ | Number of articulatory coordinates |
| $k$ | Level at which algorithm operates |
| $m_I$ | Interpolated mean |
| $m^{\phi}, k$ | Model mean for phone $\phi$ at level $k$ |
| $\hat{m}^{\phi}$ | Final model mean for phone $\phi$ |
| $\bar{m}$ | Conditional mean (D-step) |
| $\dot{\mu}^{\phi}$ | 14D phone mean |
| $\mu^{\phi,k}$ | Phone specific mean for phone $\phi$ at level $k$ |
| $n^{\phi}$ | Number of model samples |
| $\nu^{\phi}$ | Number of phone samples |
| $\phi$ | Index for phone |
| $r$ | Correlation value |
| $\rho$ | Canonical correlation |
| $\sigma$ | Standard deviation |
| $\sigma_{xx}(i)$ or $\sigma_x^2(i)$ | variance associated with the $x$ movement of an articulatory coordinate $i$ |
| $\sigma_{yy}(i)$ or $\sigma_y^2(i)$ | variance associated with the $y$ movement of an articulatory coordinate $i$ |
| $\sigma_{xy}(i)$ | covariance between $x$ and $y$ movements of an articulatory coordinate $i$ |
| $\theta_C$ | Critical threshold |
| $\theta_D$ | Dependent threshold |
| $t$ | Time |
| $\varphi$ | Total number of phones in phone set $\Phi$ |
| $\omega$ | Critical articulator combination index (exhaustive search) |
| | |
| $C^{\phi,k}$ | Critical coordinate list for phone $\phi$ at level $k$ |
| $\hat{C}^{\phi}$ | Final list of critical coordinates for phone $\phi$ |
| $D^{\phi,k}$ | List of dependent coordinates for phone $\phi$ at level $k$ |
| $\hat{D}^{\phi}$ | Final list of dependent coordinates for phone $\phi$ |
| $\mathcal{D}_{KL}$ | Kullback information |
| $\Delta^{\phi}$ | Model statistics for phone $\phi$ |
| $\Gamma$ | Grand statistics |
| $I$ | Kullback information |
| $\mathcal{I}$ | Mutual information |
| $J$ | Symmetric Kullback Leibler divergence (identification divergence) |
| $K^{\phi}$ | Number of critical dimensions for phone $\phi$ |
| $\Lambda^{\phi}$ | Phone-specific statistics for phone $\phi$ |
| $M$ | Grand mean |

| | |
|---|---|
| $N$ | Number of grand samples |
| $P_k$ | Number of combinations (exhaustive search) at level $k$ |
| $\Phi$ | Set of phones |
| $R$ | Grand interarticulatory correlation matrix |
| $R^*$ | Grand interarticulatory correlation matrix with strong, significant correlations |
| $R^\phi$ | Phone specific interarticulatory correlations |
| $\hat{\mathcal{R}}^\phi$ | Final list of redundant coordinates for phone $\phi$ |
| $\bar{S}$ | Conditional variance (D-step) |
| $S^{\phi,k}$ | Model variance for phone $\phi$ at level $k$ |
| $\hat{S}^\phi$ | Final model variance for phone $\phi$ |
| $\Sigma$ | Grand variance |
| $\Sigma^\phi$ | Phone specifc variance for phone $\phi$ |
| $\dot{\Sigma}^\phi$ | 14D phone specific covariance matrix with 1D variances (2D covariances) |
| $\ddot{\Sigma}^\phi$ | 14D phone specific full covariance matrix |
| $\Theta$ | Set of critical and dependent thresholds |
| $\Upsilon_{conv}$ | Convergence scale |
| $\Upsilon_{eval}$ | Evaluation scale |
| $W$ | Histogram bin-width |

# List of abbreviations

| | |
|---|---|
| ACIDA | Articulatory constraint identification algorithm |
| ASR | Automatic speech recognition |
| BN | Bayesian networks |
| BY | Blackburn and Young's model |
| CDC | Critical-dependent-critical |
| CDDC | Critical-dependent-dependent-critical |
| CDRC | Critical-dependent-redundant-critical |
| CR | Coarticulation resistance |
| CRC | Critical-redundant-critical |
| CRDC | Critical-redundant-dependent-critical |
| CRRC | Critical-redundant-redundant-critical |
| CVC | Consonant-vowel-consonant |
| DAC | Degree of articulatory constraint |
| DBN | Dynamic Bayesian networks |
| DFS | Depth first search (ACIDA) |
| EMA | Electro-magnetic articulography |
| ES | Exhaustive search |
| HAMM | Hidden articulatory Markov models |
| HMM | Hidden Markov models |
| ICA | Independent components analysis |
| IPA | International Phonetic Association |
| IQR | Inter-quartile range |
| LDA | Linear discriminant analysis |
| LI | Lower incisor |
| LINT | Linear interpolation |
| LL | Lower lip |
| LTSHMM | Linear trajectory segmental hidden Markov model |
| MFCC | Mel-frequency cepstral coefficients |
| MOCHA | MultiCHannel Articulatory (database) |
| MRI | Magnetic resonance imaging |
| PCA | Principle components analysis |
| R | Redundant |
| RMSE | Root mean square error |
| SHMM | Segmental hidden Markov models |
| TADA | Task dynamic model of speech articulator coordination |
| TB | Tongue blade |

| | |
|---|---|
| TD | Tongue dorsum |
| TT | Tongue tip |
| TTS | Text to speech |
| UL | Upper lip |
| V | Velum |
| VCV | Vowel-consonant-vowel |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis focuses on identification of constraints on human speech articulators during the production of speech.

## 1.1 Context

Speech is an essential form of communication between humans. Initiation of speech takes place in the mind where lexico-grammatical structure of speech takes shape along with the motor plan for commands to be executed by the speech organs. A message conceptualised in the brain is encoded grammatically and phonologically which is then converted in to a phonetic plan for the articulators. At the production level, air expelled from the lungs passes through the larynx where phonation occurs. The articulation takes place in the mouth, where speech organs shape the vocal tract to produce different sets of sounds. The air is then expelled either through the nasal or the oral cavity. Different levels of processing for producting speech are explained in (Levelt, 1989). Figure 1.1 illustrates an outline of flow of information during speech production along with the vital speech organs. Some components of the speech production system that can move are called *active* articulators (Ladefoged, 2005), e.g., lips and tongue. Some articulators, for e.g., jaw, are heavy and resist being set in motion whereas other articulators such as tongue tip move rapidly from one target to the next. Rigid parts of vocal tract are called *passive* articulators (Ladefoged, 2005), e.g., alveolar ridge, palate. Speech articulators are constrained anatomically and have limited degrees of freedom. The movements of articulators are correlated due to presence of physiological links amongst them.

Speech sounds are distinguished based on the place and the manner of articulation (Ladefoged, 2005). The place of articulation refers to the regions in the vocal tract that are associated with an articulatory gesture, for example, a 'palatal' sound is made at the hard palate. The manner of articulation refers to the way in which the speech sound is produced, for e.g., for a 'nasal' , the air flow is only through the nose. Active articulators move towards target positions to shape the vocal tract for the production of a desired speech sound. For example, for producing a bilabial stop [p], the upper

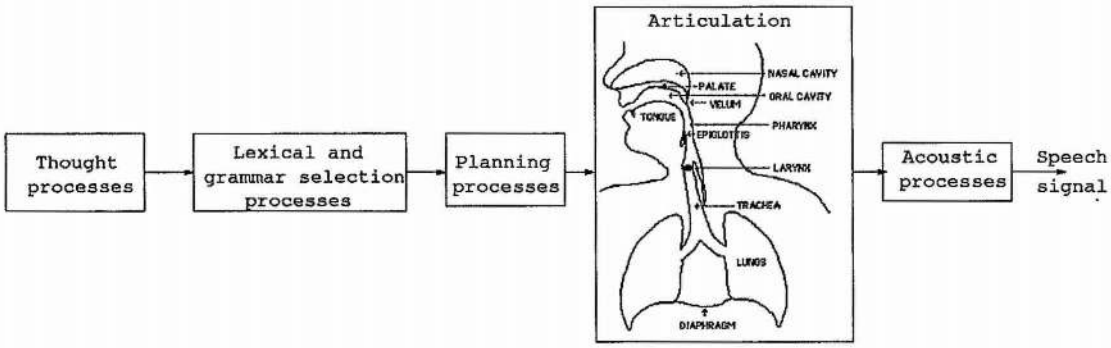



Figure 1.1: *Illustration of flow of information from initiation of thoughts to articulatory movements for producing speech. Vital speech organs which shape the vocal tract for production of various speech sounds are also illustrated. Illustration of human speech production system is taken from Rubin and Vatikiotis-Bateson (1998).*

and lower lips come together and the velum is raised to achieve a complete closure of the vocal tract required for generating the stop.

Realisation of the intended speech is affected by various factors at both higher and lower levels of processing (Levelt, 1989). Various social, pragmatic, syntactic, semantic and prosodic factors affect the generation of speech at the higher levels of speech production process. At lower levels, context sensitivity affects the plan for articulatory commands. Smooth and overlapping movements of articulators result in coarticulation whereas an inherent delay between the muscular command and the movements of articulators to achieve intended articulation leads to effects such as reduction. Though the effects occurring at the lower levels may only cause subtle changes to the perception, the acoustic realisation of the speech sounds may be affected to a greater extent. The performance of the speech recognition and synthesis systems is compromised due to the mismatch between the realised speech and the corresponding acoustic model. Both recognition and synthesis systems ignore the knowledge of speech production for modelling coarticulation. Context sensitive models such as triphone models are used for modelling coarticulation (Schwartz et al., 1985; Young et al., 2002). The parameters of available contexts are collected from the data and shared with other triphone models (parameter tying) to address the problem of data sparsity. Pseudoarticulatory information in the form of discrete or binary features (Chomsky and Halle, 1968) is used in parameter tying.

It is important to study and understand the process of speech production in relation to planning, articulation, physiological characteristics and context sensitivity for advances in speech technologies.

## 1.2 Motivation

Coarticulation is one of the main problems affecting the performance of speech recognition and synthesis systems. Due to coarticulation, the realisation of a speech sound

is not identical in all environments but varies depending on the context in which it occurs. For example, the position of lips for [s] will be rounded due to the influence of following rounded vowel [u] in [spun] and spread in [spin] due to the neighbouring sound [i]. During the production of a speech sound, some articulators are constrained to achieve target positions whereas the rest are unconstrained and are free to assume any uncontradicting positions. The unconstrained articulators are most susceptible to the coarticulation effects caused by neighbouring constrained articulators. In the above example, the lips are constrained to protrude and spread for vowels [u] and [i] respectively. The lips are unconstrained during the production of [s] and therefore assume rounded or spread position depending on the context. Some context sensitive effects can be planned in advance before the actual articulation takes place. In such cases, the motor commands for articulators are specified in such a way that the articulators can anticipate the positions required for the following sounds in the execution phase. On the other hand, competing demands are imposed on the articulators during the execution of motor commands in articulation stage in production of fluent speech. This causes variations in target positions due to the overlap of articulators which are characterised with varying degrees of inertia.

Though research and development of powerful training and decoding algorithms have increased the quality of state-of-the-art speech recognition systems, their performance is limited to certain tasks and environments. The performance of speech recognisers on spontaneous speech deteriorates due to coarticulation. The acoustic realisation of a speech sound varies in different contexts due to coarticulation and therefore it becomes difficult to match the acoustic observations with the appropriate speech sound. Context sensitive models such as triphone and quinphone models have been developed to model the coarticulatory effects (Schwartz et al., 1985; Young et al., 2002). Such models require collection of statistics and framework for sharing the parameters to avoid data sparsity problems. Modelling the span of coarticulatory effects is controlled by the length of the context sensitive models used.

The text to speech (TTS) systems commonly employ unit selection to synthesise speech from the text. Large amounts of databases of same speaker are essential to maintain good voice quality for speech synthesis. The search space for synthesis is restricted by the possible sounds and contexts present in the database. It is important to consider the effects of coarticulation to generate natural sounding speech. The discontinuity at the boundaries when different speech units are joined together makes synthesised speech sound unnatural. The spectral distance between features at either side of the join, known as the joint cost, is minimised when performing concatinative synthesis to reduce the discontinuity.

### 1.2.1 Problem statement

The human speech production is a rich source of information and offers many cues in the form of physiological constraints and biomechanical links. Acoustic modelling in the state-of-the-art speech recognition and synthesis systems uses no information of speech production process for modelling coarticulation. Models of coarticulation are built at the surface acoustic level while the underlying articulatory source is ignored. There are

potential benefits to be gained by incorporation of speech production knowledge in the architecture of recognition and synthesis systems. Each speech sound can be associated with a set of synchronised articulatory movements. The presence of constraints on some articulators during speech production reduces the variability in the articulatory space. The articulators have inertia and therefore their movements are smooth and slow. It is possible to generate smoother trajectories in the articulatory space which could potentially yield smoother spectral transitions in the acoustic space. Presence of physiological constraints and links makes it possible to generate compact and well defined models in the articulatory space.

Incorporation of speech knowledge into the structure of speech recognisers and synthesisers has attracted the interest of many researchers. Most of the existing production oriented models of speech recognition and synthesis rely on the phonological descriptions of the place and manner of each phone (for e.g. binary features by Chomsky and Halle (1968)) where measured articulatory data is unavailable. Coarticulation is modelled by specifying constraints on articulators using hand-coded rules and scores (for e.g. Richardson et al. (2000)). Such rules and codes are simple heuristic representations which do not take into account the variations due to language, speaker, style etc.. Availability of measured articulatory data in the form of X-ray recordings (Westbury et al., 1994) etc. made it possible to employ statistical modelling techniques in the articulatory domain to capture the knowledge of speech production. Though statistical models offer efficient algorithms for training and decoding, they lack explanatory power.

There is a need for building models of realistic speech articulation to identify and capture the essential characteristics of human articulators, such as target-driven behaviour, articulatory interdependencies and biomechanical constraints for efficiently modelling the effects of coarticulation. This thesis aims at providing a model of articulation that captures the essence of speech production by

- taking into account the biomechanical links between the articulators
- identification of the constraints on the articulators and thereby the invariance in the articulatory domain
- identification of partially controlled and totally redundant degrees of freedom of articulators which are prone to coarticulatory effects to a greater extent
- providing parsimonious representations by capturing the constraints on articulators
- providing a mapping from the phonological domain to the real-time phonetic domain by capturing the language, speaker, accent and style specific characteristics of each phone from the measured articulatory data
- employing entirely statistical and data-driven techniques.

The degrees of freedom of an articulator during production of each speech sound are identified using the knowledge of critical, dependent and redundant roles obtained from

the proposed approach. The data-driven constraints obtained from the proposed algorithm can be used to model the coarticulation caused by the critical articulators on the neighbouring dependent and redundant articulators. The proposed method provides parsimonious, statistical representations that can be trained from the data and can be incorporated into the architecture of speech synthesis and recognition systems.

## 1.3 Overview of the thesis

The remainder of this thesis is organised as follows

**Chapter 2** of this thesis presents the background information. The chapter is organised into two sections. The first part of the chapter presents the literature review covering previous work on different coarticulation modelling approaches in various speech recognition and synthesis systems. The second part of this chapter provides an introduction to the dataset used along with the preprocessing details. Assumptions underlying the model are presented and evaluated using statistical methods. Background information on the statistical measures used in the proposed approach is also presented.

**Chapter 3** presents the methodology of the proposed algorithm. Different stages in the algorithm are explained in detail along with the implementation details. Evaluation measures used for analysing the algorithm's performance are also introduced.

The results obtained from the proposed identification algorithm are presented in **Chapter 4**. Lists of critical, dependent and redundant roles obtained from 1D and 2D versions of the algorithm are presented for each speaker in the database. The obtained results are analysed and compared with the constraints from the phonological knowledge. The performance of the models estimated using the proposed algorithm is analysed using the evaluation measures introduced in Chapter 3.

**Chapter 5** presents evaluation of the role identification algorithm. The proposed method is evaluated against an exhaustive search based approach for the identification of roles. The performance of the algorithms are evaluated by comparing the results from the proposed algorithm against the exhaustive search results.

**Chapter 6** presents the models of articulation generated from the findings from the proposed role identification algorithm. The first half of this chapter focuses on derivation of different articulatory representations using orthogonal linear transforms, and evaluation of the usefulness of such representations in articulatory modelling. The second half of this chapter presents a statistical framework for modelling coarticulation in the planning stage of speech production using constraints from the proposed algorithm. Hypotheses for modelling different aspects of coarticulation are evaluated by trajectory synthesis. Synthetic trajectories are compared with the measured trajectories for evaluating the performance of models generated from different hypotheses.

**Chapter 7** of this thesis presents the summary, the conclusions, the contribution and the publications resulting from this work, and the future work.

# Chapter 2

# Background

This chapter provides relevant background information on different theories and models of speech articulation and introduces the proposed approach along with the preliminary evaluation of the data. This chapter is divided into two parts. The first half provides background literature on various approaches to modelling coarticulation in production oriented models of speech recognition and synthesis. Different aspects of coarticulation are introduced and existing theories for modelling coarticulation are explained. Various knowledge driven, data driven and hybrid approaches for incorporation of speech production information in speech recognition and synthesis systems are presented. Existing knowledge driven constraints on speech articulators are reviewed and the proposed statistical approach for identification of constraints from the measured articulatory data is explained. The second half introduces the dataset used for this study and various preprocessing stages. The assumptions made for the study also are introduced and evaluated using statistical methods.

## 2.1 Literature review

### 2.1.1 Introduction

The human speech production system is a valuable source of information for advances in speech science and technologies. Research has also shown links between speech perception and activation of relevant articulatory gestures (Fadiga et al., 2002; Wilson et al., 2004; Meister et al., 2007). One of the main problems faced by speech researchers in accurately modelling speech dynamics is coarticulation. The articulatory variability due to coarticulation makes it difficult to match the resulting acoustic realisations with the linguistic units for recognition. Coarticulation affects the naturalness of synthesised speech. It is widely accepted that production oriented representations of speech can potentially benefit the performance of speech synthesis and recognition systems (Rose et al., 1996; Deng et al., 1997; McDermott and Nakamura, 2006; King et al., 2007). However, in practice, finding a suitable representation for incorporation of production knowledge still remains an open problem. Speech production knowledge was incorporated into the structure of synthesisers and recognisers using many knowledge driven,

purely data driven and hybrid approaches. Many acoustic, articulatory and pseudo-articulatory approaches have been developed to model the effects of coarticulation. Movements of articulators during speech production have been recorded in many ways, e.g., using electro-magnetic articulography (EMA) (Wrench, 2001; Richmond, 2009), X-ray (Westbury et al., 1994; Soquet et al., 1999) and tagged MRI (Parthasarathy et al., 2007).

Before looking at these models in more detail, the two main aspects of coarticulation are introduced first.

### Aspects of coarticulation

There are two main aspects of coarticulation: (a) spatial and (b) temporal.

**Spatial coarticulation** : The degree to which the neighbouring context influences the target positions of speech articulators during the production of a phone is a measure of spatial coarticulatory effect. The target position of an articulator for a phone overshoots or undershoots due to coarticulation caused by the neighbours. Different factors such as articulatory inertia, competing demands on articulators, speaking rate and stress cause spatial coarticulatory effects. One of the most important theories explaining spatial coarticulatory effects is "the principle of economy of speech production and adaptive variability" proposed by Lindblom (1990). According to this theory, an articulator, when unconstrained to achieve a target position for a phone, tends to default to a low cost form of behaviour. Target undershoot occurs at shorter durations due to the lack of sufficient time to reach the ideal target position. The speaker can also adapt the behaviour at the rate of higher bio-mechanical cost. Locus equations were used to quantify the formant target undershoots of vowels in consonant-vowel-consonant (CVC) utterances as a function of duration by Lindblom (1963). His locus equation approach has been used to quantify the spatial coarticulation due to context by many since then. Apart from duration, the speaking style and stress also contribute to target undershoot (Moon and Lindblom, 1994). Öhman (1966, 1967) investigated the coarticulatory effects in vowel-consonant-vowel (VCV) utterances. In his view, the vocalic gesture involved in production of VCVs was assumed to be continuous. The consonantal gesture was superimposed on the continuous vocalic gesture. It was found that the articulators which were not actively involved in producing the consonantal gesture (i.e., unconstrained) were influenced most by the vocalic context. For example, for velars in VCV contexts, the tongue dorsum degree of constriction remained invariant whereas the place of constriction was modified due to the adjacent vocalic context. In his numerical model (Öhman, 1967), idealised consonant target shapes were modified using an estimate of coarticulation due to the vocalic context.

**Temporal coarticulation** : The extent in time to which the target position of an articulator required for a phone can influence the neighbouring unconstrained phonetic segments constitutes the temporal aspect of coarticulation. Depending on the direction of influence, the temporal coarticulation can be (i) anticipatory, where target

position required for a phone is anticipated by the neighbouring unconstrained phones, and (ii) carry forward, where the target position of a phone continues to affect the following unconstrained phones. The immediate neighbours are most susceptible to the anticipatory coarticulation effects. Moll and Daniloff (1971) found that the anticipatory coarticulation could also start a few segments before the influencing phone. The carry-forward coarticulation has mostly been attributed to the inertia associated with the articulators. Many theories (Henke, 1965; Moll and Daniloff, 1971; Saltzman and Munhall, 1989) were proposed to explain and model the anticipatory and the carry forward coarticulatory effects. For example, temporal coarticulatory effects on the adjacent segments have been quantified using electro-palatographic data and formant information by Recasens et al. (1997); Recasens and Pallarés (1999). Here, the anticipatory and carry forward effects caused by the consonant (C) on the neighbouring vowels (V) in VCV contexts were investigated. A rule based scale defined the degree to which an articulator resists coarticulation and was known as degree of articulatory constraint (DAC).

The following sections present various knowledge driven, data driven and hybrid approaches used for modelling different aspects of coarticulation and for incorporating the knowledge of speech production into recognition and synthesis systems. In the knowledge driven approach to modelling coarticulation, the constraints offered by human speech production system are derived from phonological knowledge. In the data driven approach, purely statistical models of coarticulation are built from acoustic and measured articulatory data. Phonological knowledge is combined with measured acoustic data in the hybrid approach for modelling coarticulation.

### 2.1.2 Knowledge driven models

In the knowledge driven models, the constraints of speech production are derived from phonological knowledge. Hand coded rules and scales are used to identify constrained and unconstrained components of speech production system for each phone. Different knowledge driven approaches can be classified into two categories: (i) feature based models, (ii) gesture based models.

#### Feature based models

In one of the theories of phonology, a distinctive set of binary features encode the place and manner of articulation for every phone (Chomsky and Halle, 1968; Fant, 1969). The presence of a feature is denoted by '+', and the absence of a feature is denoted by '−' as shown in Fig. 2.1. Any insignificant feature is left unspecified. In feature based approach to modelling coarticulation, discrete set of binary features were specified for each phone in the uttertance (Henke, 1965; Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973). Anticipatory coarticulation was modelled as the spread of features to the neighbouring unspecified phones in left to right direction at the phonological level (Henke, 1965; Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973). The feature spreading is blocked when the next specified feature is encountered. Henke (1965) implemented this feature spreading in the form of his look ahead model

| Features | before feature spread | | | after feature spread | | |
|---|---|---|---|---|---|---|
| | [o] | [p] | [m] | [o] | [p] | [m] |
| Nasal | | − | + | − | − | + |
| Voice | | − | + | − | − | + |
| Rounding | + | | | + | | |

Figure 2.1: *Binary features proposed by Chomsky and Halle (1968) to the left, and illustration of feature spread to the right.*

of coarticulation. The feature values were spread and then mapped to spatio temporal targets. The carry forward coarticulation was attributed to inertia of the articulators.

The extent to which an articulator's position is influenced by neighbouring phones was quantified using discrete feature values by Bladon and Al-Bamerni (1976). Coarticulation resistance (CR) was estimated from the formant space by measuring the consonant's resistance to the neighbouring vowels in VCV (vowel-consonant-vowel) contexts. The higher the CR value, the stronger the resistance to the coarticulation. Keating's window model (Keating, 1988) uses a range of values characterised by rectangular windows for each phone segment. Articulatory features in her model are still binary allowing for application of various phonological and phonetic fill in rules. Some speech recognition systems have been inspired by the feature based concept and used binary feature information for incorporating speech production knowledge (Kirchhoff, 1999; Metze and Waibel, 2002; Frankel et al., 2004; Eide, 2001; Koreman et al., 1998).

The feature based approach has many disadvantages. Feature based approach is poorly suited to represent the continuous and asynchronous articulatory movements. The features are non-overlapping, abstract and static representations and the feature boundaries are asynchronous. The feature specification is done at the phonological level and therefore the variations due to language, speaker and style are not considered. The extent to which the neighbouring segments are affected is unspecified. Discrete features such as coarticulation resistance are graded representations and are context specific. It is also difficult to interpret the binary features as instructions for the articulators (Dressler et al., 1992).

### Gesture based models

An alternative theory for modelling coarticulation is the gestural theory (Liberman, 1970; MacNeilage, 1970; Browman and Goldstein, 1986; Saltzman and Munhall, 1989). A phonetic gesture is defined as the movement of an articulator towards a phone specific goal. Articulatory gestures are associated with an intrinsic temporal structure that allows for the continuous and asynchronous movements. Each gesture is controlled by an activation wave which is associated with a gradual implementation phase followed by a relaxation phase. The overlap from coproduction of gestures results in coarticulation as shown in Figure 2.2. If an articulator is shared by two or more competing gestures, the resulting target position is subject to spatial coarticulation due to intergestural blending. The activation waves are shaped and prioritised using gesture scores obtained

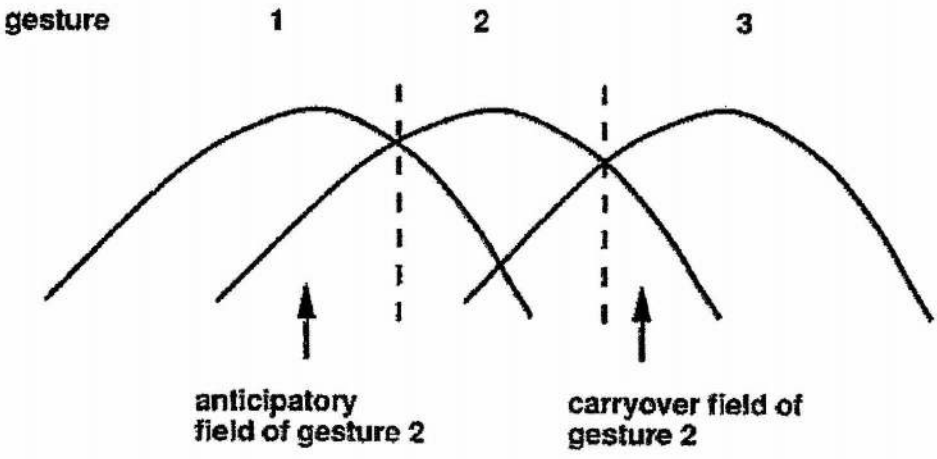



Figure 2.2: *Schematic representation of activation waves taken from Fowler and Saltzman (1993) for three overlapping phonetic gestures, with anticipatory and carry-forward coarticulatory fields indicated.*

from the phonetic knowledge. Dominance functions proposed by Löfqvist (1990) were one of the gesture based approaches for modelling coarticulation. The dominance of a segment over the vocal tract across time was defined using dominance functions. The dominance functions are analogous to the gesture activation waves depicted in Figure 2.2. The dominance functions of adjacent gestures overlap in time resulting in coarticulation. Cohen and Massaro (1993) developed exponential dominant functions for synthesising visual speech. Articulators are controlled by different exponential functions with variations in magnitude, time and offset for different phones.

Physiological models of speech production have been constructed by modelling muscle behaviour using differential equations from the classic mechanics (Coker, 1976; Ostry et al., 1996; Dang and Honda, 2004). Some speech recognition systems were inspired by the gesture models of coarticulation (Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000). Gestural patterns were estimated using dynamic programming and state model approaches from the speech synthesised using task dynamic model of speech articulator coordination (TADA) by Ghosh et al. (2009).

The gestures are more closely related to the speech production process than the distinctive binary features. However, in gesture based theory, the extent of anticipatory coarticulation is constant due to the fixed temporal structure of the gestures. Both feature and gesture based models rely on knowledge driven constraints for defining features and for prioritising gestures respectively. The following section presents purely data driven approaches to modelling coarticulation.

### 2.1.3 Data driven models

Statistical and data driven approaches have taken over the knowledge based approaches for modelling coarticulation in current ASR and synthesis systems. Context sensitive effects were modelled using acoustic data (Schwartz et al., 1985; Young et al., 2002) or measured articulatory data (Blackburn and Young, 2000) or a combination of both (Wrench, 2000, 2001). The context dependency due to coarticulation is modelled using context sensitive acoustic models such as triphones and quinphones (Schwartz et al., 1985; Young et al., 2002). An alternative solution was proposed by Sun (1997) where the context independent models were subject to linear interpolation techniques to derive smooth trajectories for ASR. Other techniques for generating smoother trajectories from the probabilistic descriptions of the acoustic data include dynamical models (Richards and Bridle, 1999) for ASR and trajectory HMMs (Tokuda et al., 2007) for synthesis.

Measured articulatory data has been used to model the source of coarticulation in various synthesis and recognition systems. Blackburn and Young (2000) used curvature and position estimates from the X-ray data to estimate context sensitive effects on the target positions. Triphone HMM models trained from EMA data were used to synthesise minimum-acceleration articulatory trajectories (Okadome and Honda, 2001). Parameters of dominance functions were estimated from the audio-visual data by Krňoul et al. (2006). The vocal tract spectrum was synthesised from the EMA data using target variance as a measure of susceptibility to coarticulation (Kaburagi and Honda, 2001; Kaburagi and Kim, 2007). Invariant articulatory features were determined statistically for each phone by performing eigenvalue decomposition on the ratio of within class variance to the total variance. The mode with smallest eigenvalue represented the constrained movements of articulatory coordinates for each phone.

Guenther (1995) proposed a neural network model, the parameters of which could be trained from measured speech data for modelling speech production. Speech was synthesised using articulatory HMMs trained on normalised palate positions by Hiroya and Mochida (2005). The EMA data was used for recognition using HMMs (Zlokarnik, 1993; Wrench, 2001, 2000; Uraga and Hain, 2006). Frankel used linear dynamical models to generate articulatory trajectories from measured articulatory data for recognition (Frankel, 2003; Frankel et al., 2000). Articulatory measurements from EMA were combined with acoustic data in an HMM/BN framework for recognition (Markov et al., 2006).

#### Multiple-level dynamical models

In the multiple-level architecture, articulatory or pseudo-articulatory models are defined in a hidden intermediate layer, the surface layer is acoustic. The mapping between the hidden and the surface layers could be forward or backward. Observations generated from the articulatory layer could be mapped to the surface acoustic layer or the articulation could be recovered from the surface acoustics (inversion mapping). In either case, the difference is fed back to train the models. Bakis (1991) used an HMM based approach with a hidden abstract articulatory layer generated from a lookup table. A

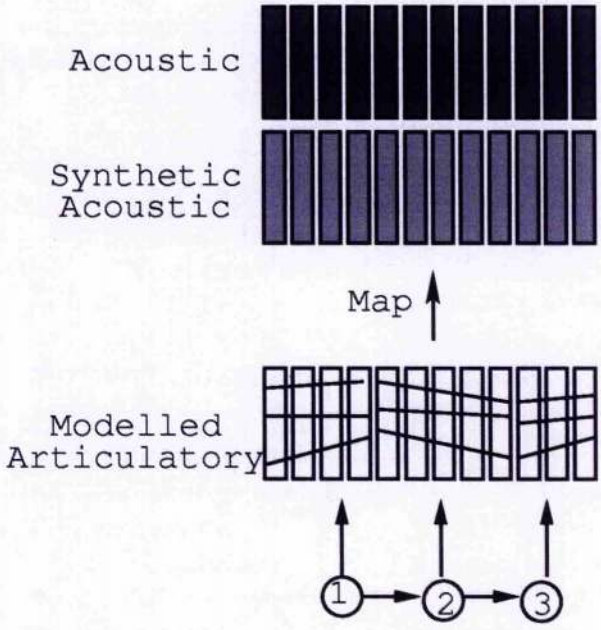



Figure 2.3: *Multi-level linear segmental HMM with a hidden articulatory layer and a visible acoustic layer with linear mapping between them (Russell and Jackson, 2005).*

hidden dynamical model was proposed by Richards and Bridle (1999) where the sequence of targets generated in the hidden layer was mapped into the acoustic space non-linearly using a multilayer perceptron. It was observed that the hidden target space closely resembled the formant representation of the utterance. This concept was developed by others. Deng and Ma (2000) used a first order linear state space equation to model the formant trajectories in the hidden layer. The phones were divided into different categories and a multilayer perceptron was used for each category to map the trajectories to the surface acoustic space. This approach was further extended by using a mixture dynamical model (Ma and Deng, 2004) and a linear mapping for each phone to make the models adaptable to different speakers and speaking rates.

Segmental trajectory HMMs (Holmes and Russell, 1995, 1996, 1997; Russell and Holmes, 1997) were developed and embedded into a multi-layer system to overcome the beads-on-string limitation (Ostendorf, 1999) of conventional HMMs. Each state of a segmental HMM generates a sequence of observations instead of single observation. Figure 2.3 shows multilevel linear SHMMs where the states in the articulatory layer generate linear trajectories. Jackson et al. (2002) used a nonlinear mapping to map formants into acoustic domain. Fixed linear trajectory SHMMs with measured articulatory based (Russell and Jackson, 2002) and formant based (Russell and Jackson, 2005) intermediate layers were proposed. The trajectories associated with the hidden states of the model are linear and are mapped into the surface acoustic layer using a set of linear mappings. Different trajectory shapes (Singampalli, 2006) and pruning methods to improve computational efficiency for recognition (Shiga and Jackson, 2008) were developed.

An alternative approach to mapping between the acoustic and articulatory spaces is the inversion mapping, where the articulatory information is retrieved from the acoustic data. Inversion mapping poses a one to many problem since a single acoustic effect can be generated from multiple articulatory configurations. The problems in acoustic to articulatory inversion are discussed in (Atal et al., 1978; Bailly et al., 1992). A comprehensive review of several inversion mapping techniques is provided in (Schroeter and Sondhi, 1994). More recent models use machine learning methods on measured articulatory data, such as codebooks (Hogden et al., 1996; Okadome et al., 2000), self organising HMMs (Roweis, 1999), mixture density networks (Richmond, 2006, 2007a). An attempt to determine non-uniqueness of acoustic to articulatory inversion from peaks of conditional distribution of the articulatory space was proposed by Ananthakrishnan et al. (2009).

Statistical models are powerful and use available data to generate probabilistic representations of the phones, but fail to identify the constraints offered by human speech production system. State-of-the-art synthesis and recognition systems use context sensitive models to capture the coarticulatory effects. Such models require sufficient training data and clustering techniques for sharing parameters. Any knowledge of articulatory constraints have to be input explicitly in the form of phonological knowledge. The following section presents all such models, called the hybrid models, which are partly knowledge based and partly data driven.

### 2.1.4 Hybrid models

The knowledge based and the data driven approaches have been combined together to generate hybrid recognition and synthesis systems. Such hybrid systems aim at improving the performance of recognisers and synthesisers by making the best use of both approaches.

#### Feature based models

Distinctive binary features were used in ASR systems for improving performance and robustness to noise. Several

Eide (2001) used "feature-present" and "feature-absent" information to discriminatively score the acoustic input for HMM based recognition. A reduction in the word error rate (34%) was reported by Eide (2001) under noisy conditions (car engine noise). Metze and Waibel (2002) also used a HMM based recogniser for detecting articulatory features on noisy and spontaneous speech and reported a small improvement (2%) over the baseline system. Koreman et al. (1998) reported that acoustic-phonetic representations obtained by mapping acoustic information onto binary features using neural networks gave improvements (39%) in consonant recognition when used with a HMM recogniser over the baseline (no phonetic feature information is used).

Kirchhoff (1999) classified acoustic data into feature based groups using independent classifiers as shown in Figure 2.4 and tested the recognition potential of combining feature information with acoustic features in noisy and clean conditions. When a hybrid

HMM/ANN model was used for small vocabulary recognition using articulatory feature information under noisy conditions (10db SNR), the improvement obtained over was 8%. However, when tested on a large vocabulary *conversational* speech, the baseline models were slightly better (1.4%). Dependencies were introduced between otherwise independent features for feature recognition using DBNs (Chang et al., 2001; Frankel et al., 2004). Here, the features were derived from the phone labels. The DBN approach for articulatory feature recognition was extended by Frankel et al. (2007) by eliminating the dependency on phone derived labels by using an embedded training scheme to learn asynchronous feature changes from the data.

The performance of models trained on acoustic and measured articulatory data in predicting phonetic features was analysed by Toth and Black (2005). It was found that using articulatory positional data improves the prediction of phonetic features over using acoustic features. Articulatory features were also used to derive factored state representations of phone models (Livescu et al., 2003).

The speech recognition systems which use articulatory features were found to be robust to noise and showed potential on small vocabulary/digit recognition tasks. The feature based recognition systems could perform well in noisy environments since there are small number of features to detect than the phone classes. Phonemes are defined using a compact feature values which encode place and manner of articulation (Chomsky and Halle, 1968) and different phonemes could have same features in common. The training data could be efficiently used in the feature based approaches. However, when conversational speech or spontaneous speech is considered (for e.g., Kirchhoff (1999)), the feature based representations did not give improvements over acoustic only models. The feature based representations are coarse and there is a need for a finer grain representation of articulatory information for modelling coarticulation effects. The following section presents speech production models inspired by gestures, which are more closely related to the articulatory domain than the features.

### Gesture based models

Gestural information was also combined with the statistical modelling techniques for synthesis and recognition.
Gestural scores were derived for critical articulators involved in production of each phone from articulatory data using temporal decomposition techniques (Jung et al., 1996; Collins et al., 1999) for synthesis; the specification of critical articulators for phones was from phonological knowledge. Similarly, articulatory priorities for phones were derived from phonological knowledge in Coker's model of speech synthesis (Coker, 1976). Quantised articulatory configurations representing the shape of vocal tract for each phone were used for speech synthesis (Larar et al., 1988).

Dang and others (Dang et al., 2004, 2005) proposed a statistical model for modelling coarticulation in VCV utterances using measured articulatory data. Dang's model was based on Öhman's model of VCV coarticulation, where the vocalic gesture is treated as continuous on which a consonantal gesture is superimposed. The temporal effects of coarticulation due to the position of a crucial articulatory point were quantified using standard deviation. Scales such as coarticulation resistance (Bladon and Al-Bamerni,

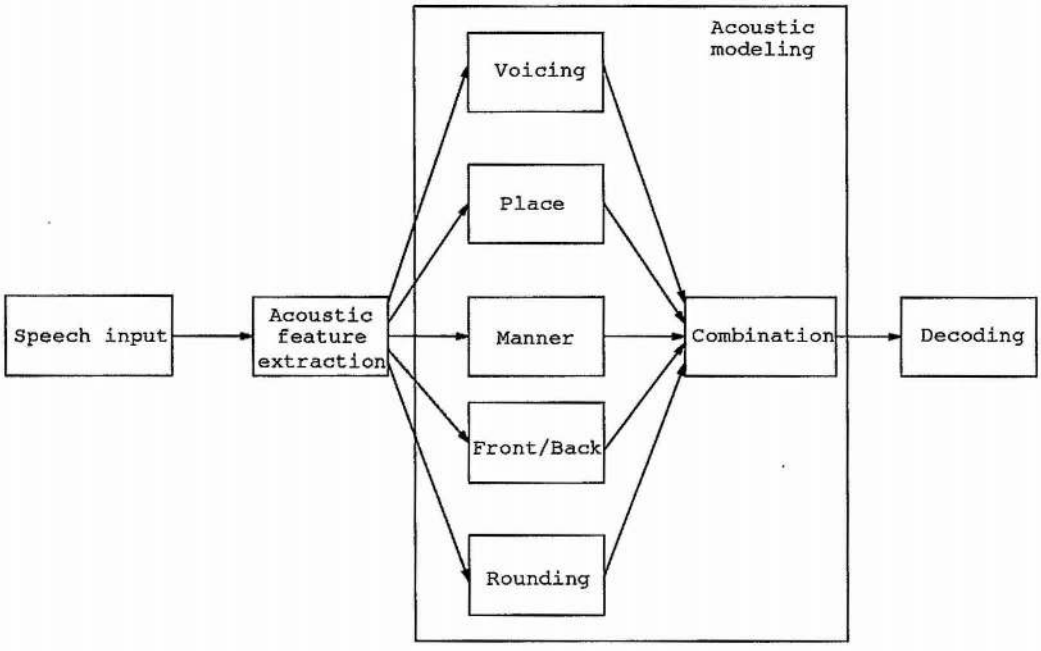



Figure 2.4: *Acoustic modelling using phonological binary feature information (Kirchhoff, 1999). An intermediate articulatory feature based representation is derived from acoustic features for improving performance of speech recognisers.*

1976) and degree of articulatory constraint (Recasens et al., 1997) were used in his model and have been estimated from the data. However, crucial articulatory points in his models were defined from phonological knowledge.

For speech recognition tasks, quantised articulatory configurations were represented by the internal states of HMMs by Erler and Freeman (1996). Transition from one state to another represents the articulation of speech. A set of static constraints allow only phonetically relevant configurations and dynamic constraints ensure that the movements are physically plausible. Such models were extended to include diphone models and were called hidden-articulator Markov models (HAMMs) by Richardson et al. (2000). Model by Deng and Sun (1994) allowed for rule based overlap of quantised gestures for generation of state transition graphs. Figure 2.5 shows the state transition graph representing quantised articulatory configurations for [t] in 'ten'. Anticipatory coarticulation caused by the right context [eh] on the tongue body is shown as R(9) where R indicates right context and 9 indicates the quantised location. The gesture based models showed some potential in isolated word speech recognition tasks.

The representations in the form of quantised gestures are heuristic in nature and need rule based constraints to model speech articulation. They donot consider articulatory interdependencies due to biomechanical links. There is a need for more accurate, explicit models of speech articulation, than symbolic knowledge driven representations for modelling coarticulation.

Different aspects of coarticulation and different knowledge and data driven, and hy-

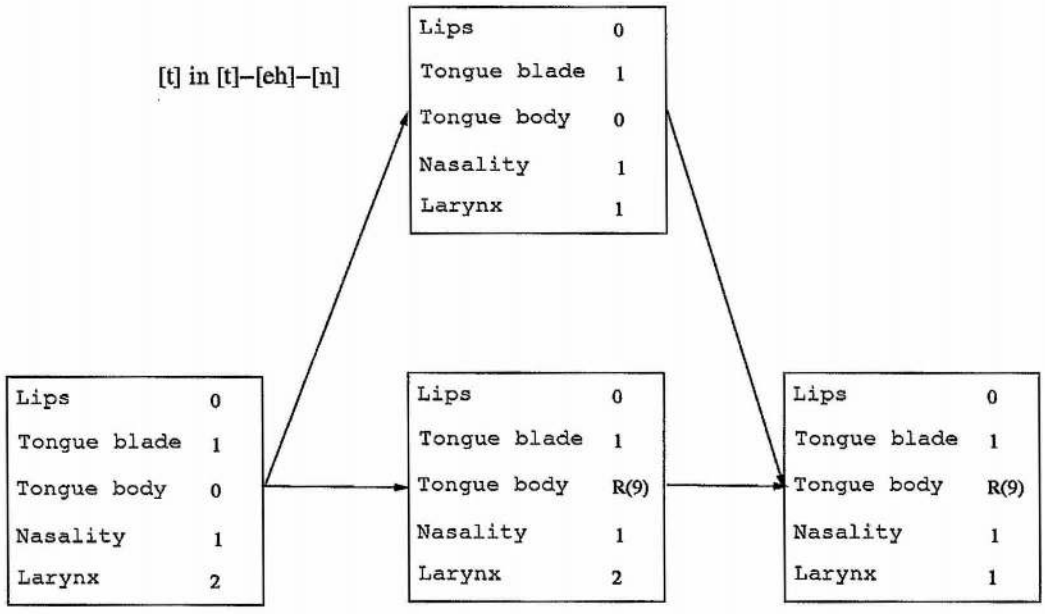



Figure 2.5: *HMM state transitions representing various quantised articulatory configurations of lips, tongue blade, tongue body, velum and larynx as in Deng and Sun (1994). Different state transitions are shown for [t] in [tehn] (ten), R indicates that the specified articulatory configuration is the resultant of the anticipatory effects caused by neighbouring [eh].*

brid models used for modelling coarticulation and for incorporating speech production knowledge have been presented so far. The following section looks at the relationship between the articulatory constraints and coarticulation and different ways in which the articulatory priorities have been defined.

### 2.1.5 Articulatory constraints vs coarticulation

In this section, the relationship between different aspects of coarticulation and articulatory constraints are explained. Different ways in which the articulatory priorities have been set in the literature are explained. The proposed approach of modelling constraints by articulatory roles using statistical and data driven ways is introduced.

#### How do constraints explain different aspects of coarticulation?

The articulators move continuously from one configuration to the next during production of phones. For each phone, some articulators are constrained to achieve target positions while some others are free to assume any uncontradicting position. The constrained articulators cause temporal coarticulatory effects on the neighbouring unconstrained segments. The position of the constrained articulator is subject to spatial

coarticulation due to the neighbouring phones, leading to target undershoots and overshoots. When modelling coarticulation in the speech planning stage, it is important to establish the degrees of freedom of articulators for each phone. In the articulation stage, the smoothness constraints play an important role in generating smooth and continuous articulatory trajectories. The different ways in which the articulatory constraints are derived from the knowledge of place and manner of articulation are explained below.

### Knowledge driven constraints

The binary features used in feature based theory are derived from the place and manner information. Unimportant features for a phone are left unspecified. When feature spread is implemented, the spreading feature is propagated to neighbouring underspecified features (Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973; Keating, 1988). In gesture based theory, gesture scores are used to prioritise articulatory gestures (Browman and Goldstein, 1986; Cohen and Massaro, 1993). Scales such as degree of articulatory constraint (DAC) (Recasens et al., 1997; Recasens and Pallarés, 1999) are used to describe the extent to which the tongue dorsum position is constrained for phones; the higher the position, the larger the resistance to coarticulation. Mermelstein (1973) used a graded approach to rank how critical an articulatory gesture was to a given phone. The concept of crucial points were introduced by Dang and his colleagues (Dang et al., 2004, 2005) from phonetic knowledge. The crucial points were defined to be resistant to coarticulatory effects and cause a maximum coarticulatory influence on its neighbours. The critical articulators were found to have smaller variance when compared with non-critical articulators (Papcun et al., 1992; Frankel and King, 2001). Active, passive and neutral attractors were used to shape the articulator movements when it is fully constrained, partially constrained and unconstrained respectively (Saltzman and Munhall, 1989).

Knowledge driven features rely on the phonological information and are heuristic in nature. Though binary features provide a compact and complete description of place and manner information of phones, they are difficult to convert to commands for articulators. Quantised articulatory configurations and rules are crude representations of the speech production process.

Other than knowledge driven approaches, crucial points were also identified in a data driven way by Ananthakrishnan and Engwall (2008). Critical articulators that already reached target position were identified as the locations in the articulatory trajectory that are associated with minimum change in velocity or maximum change in the angle. This approach considers only the mean positions of the articulators.

### Proposed approach: statistical identification of data driven constraints

A statistical approach for identification of articulatory constraints is proposed in this thesis. The constraints are established in the form of articulatory roles. The biomechanical relationships between the articulators are incorporated in the form of correlations. The correlated movements of articulator in space are also considered. During

speech production, an articulator can play (i) a critical role or (ii) a dependent role or (iii)a redundant role.

**Critical:** If an articulatory gesture or movement plays an important role in the production of a phone, it is considered to be critical for that phone. A phone can have more than one critical articulator, for e.g. upper and lower lips for [b]. Previously, the critical articulators were only associated with smaller variances (Papcun et al., 1992; Frankel and King, 2001). In the proposed approach, the critical articulators are also characterised by a shifted mean position when compared with their grand positions. To incorporate the information of direction of the movement of articulator from its grand position, the changes in the covariance are also considered along with the variance.

**Dependent:** A dependent articulator shares a bio-mechanical correlation with a critical articulator. The position of a dependent articulator is partially controlled by the critical articulator(s) due to the presence of correlation(s) between them, while the remaining degrees of freedom are prone to coarticulatory effects. This is closely related to the passive gesture proposed by Saltzman and Munhall (1989).

**Redundant:** A redundant articulator is free to move and its position does not affect the phone's production in a critical way. A redundant articulator is maximally prone to coarticulation due to the neighbouring critical articulators in time.

The proposed algorithm makes use of the bio-mechanical correlations and spatial correlations to identify the critical, dependent and redundant roles for each phone using statistical approaches. The model is entirely data-driven and generates parsimonious representations of the target configurations of articulators for every phone. The proposed approach provides scope for building and improving data-driven coarticulation models and has the potential to improve the performance of speech synthesis and recognition systems. The following section presents the introduction to the data, approaches and preliminary analysis.

## 2.2 Preliminaries

### 2.2.1 Data

The Electro-Magnetic Articulograph (EMA) data from MOCHA-TIMIT database (Wrench, 2001) was used for this work. The data has 14 channels representing the horizontal (x) and vertical (y) movements of 7 articulatory points with the upper incisor and bridge of the nose as reference points. The articulatory points were located on upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD and velum V as shown in Fig. 2.6. The EMA data from one male (msak) and one female (fsew) speaker were used. The 7 points chosen on the articulators were found to represent the articulatory configurations with reasonable accuracy (Badin and Serrurier, 2006; Qin et al., 2008). The recordings were made when the users uttered 460 TIMIT

sentences in English. The acoustic data was also recorded simultaneously. Recordings were also obtained from a Laryngograph. The EMA data was sampled at a rate of 500Hz while the acoustic and the laryngograph data were sampled at a rate of 16kHz.

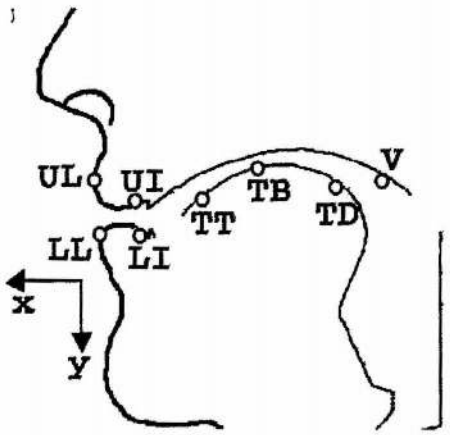



Figure 2.6: *Midsaggital section of the human speech production system highlighting the articulatory coordinates used for the EMA recordings: upper lip* UL, *lower lip* LL, *lower incisor* LI, *tongue tip* TT, *tongue blade* TB, *tongue dorsum* TD *and velum* V. *The outline was taken from (Saltzman and Munhall, 1989).*

The phone set comprised of a total of 24 consonants, 13 vowels and 7 diphthongs. The phone labels for silence, [sil] and breath, [breath], were included at the beginning and the end of each sentence. The IPA notation is used for all speech sounds. The consonants group consisted of bilabial stops [p], [b] and [m], alveolar stops [t], [d] and [n], velar stops [k], [g] and [ŋ], labio dentals [f] and [v], inter-dental fricatives [θ] and [ð], sibilants [s] and [z], post-alveolar sibilants [ʃ] and [ʒ], affricates [ʧ] and [ʤ],lateral [l], approximant [ɹ], palatal sound [j], labio-velar sound [w] and glottal sound [h].

The vowel group consisted of front vowels [æ], [ɛ], [ɪ], [iː], [i], mid vowels [ə], [ɚ], [ʌ] and back vowels [ɑ], [ɒ], [ɔ], [ʊ], [u]. The diphthongs present in the data were [aɪ], [eɪ], [ɛə], [ɪə], [ɔɪ], [oʊ] and [aʊ].

### 2.2.2 Preprocessing

A few annotation errors were detected in the existing transcripts. The phonetic transcriptions of 8 sentences (numbered 173, 317, 332, 340, 352, 354, 357 and 369) for each speaker were found to be erratic. New labels were generated for each of the above sentences based on the phonetic transcriptions of the words and aligned to acoustic input manually. The EMA recordings were found to be corrupt for one sentence (268) and therefore was excluded from the experiments. There were some cases of failed forced alignment and mismatches between the dictionary transcripts and utterances. It was found that the alveolar stop consonants /t/, /d/ and /n/, suffered high levels of elision and deletion. The label file entries for these phones were corrected manually. Full details of the changes can be found online (Jackson et al., 2004).

| Consonants | Male | Female | Vowels | Male | Female |
|---|---|---|---|---|---|
| [p] | 379 | 379 | [æ] | 229 | 229 |
| [b] | 312 | 312 | [ɛ] | 301 | 301 |
| [m] | 434 | 446 | [ɪ] | 957 | 960 |
| [t] | 852 | 861 | [iː] | 309 | 308 |
| [d] | 522 | 516 | [i] | 167 | 167 |
| [n] | 835 | 816 | [ə] | 1391 | 1397 |
| [k] | 552 | 549 | [ɚ] | 91 | 90 |
| [g] | 197 | 197 | [ʌ] | 190 | 190 |
| [ŋ] | 151 | 155 | [ɑ] | 108 | 107 |
| [f] | 267 | 267 | [ɒ] | 233 | 233 |
| [v] | 228 | 228 | [ɔ] | 205 | 205 |
| [θ] | 75 | 75 | [ʊ] | 57 | 57 |
| [ð] | 327 | 330 | [u] | 263 | 263 |
| [s] | 708 | 712 | | | |
| [z] | 497 | 496 | Diphthongs | | |
| [ʃ] | 149 | 149 | | | |
| [ʒ] | 17 | 17 | [aɪ] | 255 | 256 |
| [ʧ] | 99 | 99 | [eɪ] | 254 | 253 |
| [ʤ] | 144 | 144 | [ɛə] | 33 | 33 |
| [l] | 674 | 672 | [ɪə] | 28 | 28 |
| [ɹ] | 629 | 630 | [ɔɪ] | 46 | 46 |
| [w] | 249 | 248 | [oʊ] | 201 | 201 |
| [j] | 196 | 197 | [aʊ] | 87 | 86 |
| [h] | 154 | 154 | | | |

Table 2.1: *Sample size for consonants and vowels obtained from midpoint locations, and for diphthongs from samples taken at 1/3rd and 2/3rd (not shown) points of the duration for male and female speakers.*

Inconsistency in the EMA recordings of female speaker (`fsew0`) were reported by Richmond (Richmond, 2001, 2009). Several factors such as reattaching coils, movement of head within the helmet during the recording session were found to cause shift in the mean velum position across sentences. The female speaker data was z-score normalised using the underlying mean pattern to minimise such effects by Richmond (2001). In the present study, though similar effects were observed in the female speaker data, no such normalisation was used for either speakers.

The EMA data was smoothed and converted to time frames of 10ms duration each. Mel-Frequency Cepstral Coefficient (MFCC) features were extracted from the acoustic data. The dimensionality of the MFCC features was 14 which included the zeroth coefficient. The log-energy of the Laryngograph data was appended to the EMA features for recognition experiments.

For every consonant and vowel, it was assumed that the articulators reach their target positions at the midpoint of its duration. Therefore, a sample at the midpoint of the phone duration was selected for each consonant and vowel. For each diphthong, two

samples were taken, at 1/3rd and 2/3rd locations respectively. It was assumed that the target for the first vowel of a diphthong occurs at 1/3rd of the total duration and for the second vowel occurs at 2/3rds of the duration. Table 2.1 shows the number of samples per each phone in the data for male and female speakers. Of all phones, the phone [ʒ] had the least number of samples for both male (17) and female (17) speakers. The neutral vowel [ə] had the highest number of samples for both male (1391) and female (1397) speakers. The positions of articulators during the pause before and after each sentence could include phonetically irrelevant configurations. Therefore, the [sil] and [breath] frames occurring at the beginning and the end of each sentence were excluded from the analysis. The set of consonants, monophthongs and initial and final vowels of diphthongs is denoted by $\Phi$. The total number of phones in $\Phi$ is $\varphi = 51$.

The proposed critical articulator algorithm was implemented for two cases: (i) In the 1D case, the x and y movements of every articulatory coordinate were treated independently. Therefore, the number of articulatory coordinates $a = 14$ (ii) In the 2D case, where the correlation between x and y movements was considered. Therefore, $a = 7$. Grand mean $\mathbf{M}_i$ and variance $\mathbf{\Sigma}_i$ were computed from the data sampled from all phones for each articulatory coordinate $i \in \{1..a\}$. In the 1D case, the grand distribution for each articulatory coordinate $i$ was assumed to be univariate Gaussian in nature, i.e., $\mathcal{N}(M_i, \Sigma_i)$. The grand distributions in the 2D case were considered to be bivariate Gaussian, $\mathcal{N}(\boldsymbol{M}_i, \mathbf{\Sigma}_i)$.

Phone-specific means and variances were estimated from the data for each phone $\phi \in \Phi$. In the 1D case, phone specific distribution for each articulator $i$ was univariate Gaussian, $\mathcal{N}(\mu_i^\phi, \Sigma_i^\phi)$. The bivariate distribution in the 2D case is denoted by $\mathcal{N}(\boldsymbol{\mu}_i^\phi, \mathbf{\Sigma}_i^\phi)$.

### 2.2.3 Evaluation of Gaussian assumption

Graphical and statistical methods were used to check for validity of Gaussian assumption of grand and phone-specific distributions. The fit of the distributions to the Gaussian assumption was graphically checked by plotting histograms along with the Gaussian curves. The histograms provide information about the mean of the data, spread of the data, skewness and kurtosis along with the modes in the data. Single sample Kolmogorov-Smirnov (KS) test with Lilliefor's correction (Massey, 1951) was used for checking the goodness of fit of the Gaussian distributions to the data. The histogram plots were also used for checking the validity of the Gaussian assumption.

#### Kolmogorov-Smirnov goodness-of-fit test

The one-sample KS test is a statistical significance test used to find out how well a theoretical distribution describes a given set of data (Massey, 1951). The algorithm for the test is given in Appendix B.3.2. This test uses the maximum absolute difference between the actual cumulative distribution $F(x)$ of the data and the hypothesised distribution $F_0(x)$ as a measure of goodness-of-fit.

A benchmark statement about the population represented by a given set of the data against which the outcomes of the test are measured is known as the null hypothesis.

The null hypothesis of this test assumes that the actual distribution of the data matches the hypothesised distribution, for e.g., the grand data matches grand univariate Gaussian distribution. The research hypothesis is the statement about what can be inferred from the data when the null hypothesis does not hold true. In this case, the research hypothesis assumes that the actual and hypothesised distributions are different.

The level of significance is the amount of risk associated with rejecting the null hypothesis due to error when it actually holds true for the given set of data. For example, if the level of significance is chosen to be 5%, it is assumed that if the test is performed 20 times, the null hypothesis is rejected once on average due to error when it actually holds true.

The evaluation measure computed from the samples used to test the null hypothesis is known as the test statistic. In this test, the difference between the actual and hypothesised distributions is used as the test statistic. If the obtained difference exceeds a threshold, it is concluded that the fit of the hypothesised distribution to the data is poor. In other words, the null hypothesis is rejected. This threshold value is also know as the critical value and it varies with the degrees of freedom and the level of significance chosen for the test. The critical values of the K-S test are given in Table A21 in (Sheskin, 2000).

The grand distributions of all articulators failed the KS test. The sample size of the grand data was very large (13609 samples for the male and 15800 samples for the female speaker). So the histograms were plotted for verification of Gaussian assumption. Figure 2.7 shows the histograms for grand and phone specific distributions of $\text{UL}_y$ for [b] and [g] with their normal probability curves overlaid on the top. The width of histogram bins was calculated according to (Izenman, 1991) as

$$W = 2(IQR)N^{1/3} \tag{2.1}$$

where IQR is the InterQuartile Range and $N$ is the number of the samples in the data. The histogram plots showed that the grand distributions of all tongue and velum coordinates were in agreement with the Gaussian assumption. The grand distributions of $\text{UL}_y$ and $\text{LL}_y$ appeared to be bimodal whereas the grand distributions of $\text{LL}_x$ and $\text{LI}_x$ were skewed to the right. The grand distribution of $\text{LI}_y$ was skewed to the left.

The phone specific distributions were also subject to the KS significance test. The distribution of $\text{UL}_x$ was found to be Gaussian for most phones for both speakers (78% for male and 67% for female). The KS test results showed that relatively a few phone distributions were Gaussian for $\text{TT}_x$ for male speaker (37% of phones) and $\text{V}_y$ for female speaker (18% of phones). When the data from consonants was analysed, the distribution of more than 70% articulatory coordinates were Gaussian for post-alveolar sibilants and affricates. The distributions of all articulatory coordinates failed the test for [n] for male speaker. For vowels, the distributions of all articulatory coordinates failed the test for [ə] for female speaker, whereas only the distribution of $\text{TB}_y$ was Gaussian for male speaker. More than 70% of articulatory coordinates were Gaussian for 4 vowels for male speaker (2 for female). The distributions of more than 70% of articulatory coordinates were Gaussian for some diphthongs (7 for male and 5 for female).

Though some deviations from Gaussian assumption were found, the articulatory data in this thesis was modelled using univariate and bivariate Gaussian representations.

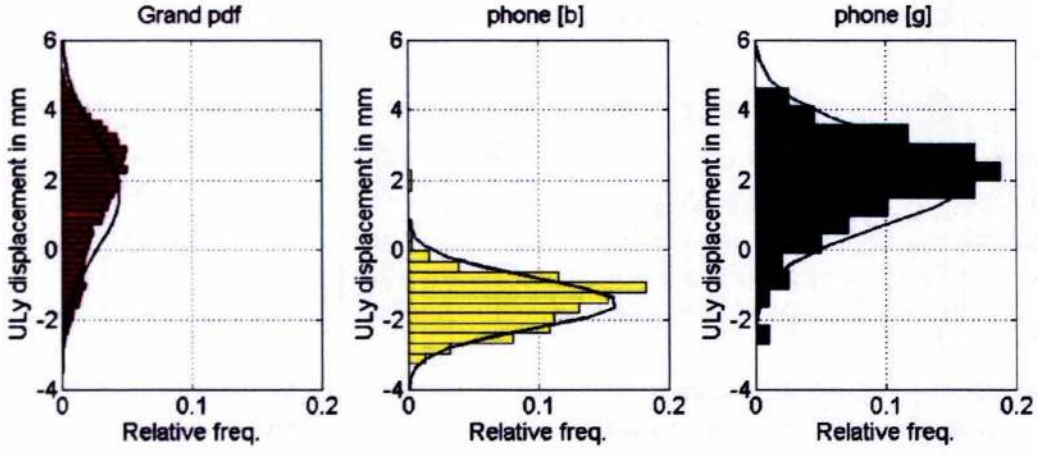



Figure 2.7: *Histograms and bell curves illustrating the grand (red, left), and phone specific data distributions of* $UL_y$ *for phones [b] (yellow, centre) and [g] (blue, right).*

Log transformations or Box-Cox transformations (Box and Cox, 1964) could be used to make the data normal. Alternatively, more complete models based on multimodal Gaussian distributions could be used to generate better fitting representations to the data and the proposed algorithm could be extended accordingly. However, such work is beyond the scope of this thesis.

### 2.2.4 Illustration of grand and phone distributions

Figure 2.8 shows the midsaggital display for phones [g] and [s] generated from the EMA data of male speaker. The covariance ellipses in red are centred at the grand means of articulators and depict their grand covariances in x and y. Similarly, the phone specific distributions shown by the green covariance ellipses were centred at the phone means and depict the phone covariances. The palate was formed as an outline of the maximum displacement of the tongue in the vertical direction. The different points that make up the lip were positioned manually and the point of contact of the lips was roughly approximated using the readings of upper and lower lips of bilabial phones [b, m, p]. The outline of the upper and lower incisors were manually generated and the position of the pellets on them were roughly approximated.

It can be seen for the phone [g] that the tongue dorsum distribution is characterised by a shifted mean position and a different covariance direction when compared with its grand configuration and can be termed as critical for that sound. The tongue blade and tip distributions are affected by the tongue dorsum position due to the correlations between them and hence can be termed as dependent articulators for that phone. There is no significant difference in the distributions of the UL and LI and thus can be treated as redundant for phone [g]. Similarly for phone [s], the distribution of TT differs from its grand state configuration. The affect of the position of TT can be seen on the distributions of TB and TD more so in the y direction. The distribution of jaw LI and velum V are also different from their grand configurations.

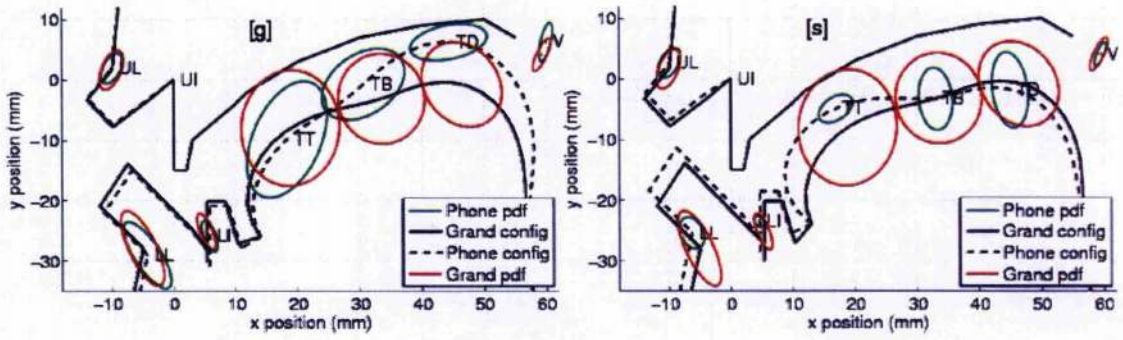



Figure 2.8: *Midsaggital illustration of articulators for phones [g] and [s] depicting grand (red) and phone-specific (green) covariance ellipses (±2 standard deviations) centred at the respective means.*

The distribution of the articulatory coordinate critical in the production of a phone differs significantly from its grand distribution. This difference between the distributions could be due to the difference in either in their means or covariances or both. The articulatory movement that is not critical in production of speech sound is prone to the context based variations and hence its covariance would be larger than or equal to that of its grand covariance. The critical articulators affect the positions of articulators with which there are

### 2.2.5 Statistical distance measures

The distance between the distribution of the critical articulator for a phone and its grand distribution can be quantified using several distance measures for 1D and 2D cases. For the univariate case, two independent samples student's t-test (Sheskin, 2000) can be used to check for the difference between the means of two univariate Gaussian distributions. It is assumed that the groups are independent of each other and the samples in each group can be described using a univariate Gaussian distribution with sample mean and sample variance as the parameters. The test also assumes that the amount of variability in each of the two groups is equal, however unequal consequences can be compensated. The test statistic used for this test is known as the t-value. The t-value is the difference between the observed difference and the hypothesised difference between the means of the two groups with respect to the standard error. Levene's F statistic (Field, 2005) can also be used to test for the equality of variances.

Hotelling's $T^2$ test (Anderson, 1984; Field, 2005; Morrison, 1990), an extension of the simple univariate t-test, can be used for finding the difference between the bivariate means of two groups. It is assumed that the two groups are independent of each other and the samples in each group follow a bivariate Gaussian distribution. It is also assumed that the covariances of the two groups are equal and this assumption is verified as a part of the implementation of the test. Box's M test for covariance (Anderson, 1984; Morrison, 1990) can be used to check for the validity of this assumption.

The statistical significance tests mentioned above the statistical significance of the differences between either the means or the variances (assuming that the other is equal

)of any two Gaussian distributions. There is a need for a distance metric which measures the difference between the grand and the phone specific distributions rather than just means or just (co-)variances. Kullback-Leibler (KL) divergence (Kullback, 1968) is one of the most commonly used statistical distance measure and measures the distance between the distributions using their log-likelihood ratio. We use the Kullback Leibler divergence as the statistical distance measure.

### Kullback Leibler Divergence

Given any two distributions $F_1(x)$ and $F_2(x)$, the Kullback information between their densities $f_1(x)$ and $f_2(x)$ is given as

$$\mathcal{D}_{KL}(1||2) = I(1:2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \tag{2.2}$$

The Kullback information, is a special case of $f$ divergence (Liese and Vajda, 2006), and is non-negative and asymmetric, i.e., $I(1:2) \neq I(2:1)$ if $f_1(x) \neq f_2(x)$. It is also related to the mutual information and other statistical distance measures such as Mahalanobis distance. The mutual information between two random variables x and y can be expressed using Kullback information as,

$$\mathcal{I}(X;Y) = \mathcal{D}_{KL}(f(x,y)||f(x),f(y)) \tag{2.3}$$

where $f(x)$ and $f(y)$ denote the mariginal distributions and $f(x,y)$ denotes the joint distribution.

We need a symmetrical distance measure for determining the distance between the grand distributions and the phone specific distributions. The Kullback Leibler divergence or J divergence is a symmetric version of the Kullback information and is given as

$$\begin{aligned} J(1,2) &= I(1:2) + I(2:1) && (2.4)\\ &= \int f_1(x) \log \frac{f_1(x)}{f_2(x)} + f_1(x) \log \frac{f_1(x)}{f_2(x)} && (2.5) \end{aligned}$$

The KL divergence between two multivariate normal distributions, $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, is given as

$$J(1,2) = \frac{1}{2} tr(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}) + \frac{1}{2} tr(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \tag{2.6}$$

The divergence equals zero if the distributions match exactly. The larger the divergence, the greater the difference between the two distributions. The KL divergence is equal to the Mahalanobis distance when $\Sigma_1 = \Sigma_2 = \Sigma$.

The KL divergence was used as the distance measure throughout this study to compute the difference between multivariate unimodal Gaussian distributions. Numerical methods such as Monte Carlo simulations can be used to extend the KL divergence to compute the difference between multi-modal Gaussian distributions (Hershey and Olsen, 2007; Chen et al., 2008).

## 2.3 Conclusion

The literature review section could be summarised as follows:
Coarticulation is one of the main problems in speech research. Coarticulation occurs naturally in fluent speech due to efficient planning and execution of articulatory movements. It is commonly agreed that incorporation of production knowledge could benefit coarticulation modelling in speech recognition and synthesis systems. Two important theories of modelling coarticulation are feature based and gesture based. The feature based theories use abstract phonological representations such as discrete binary features and coarticulation is modelled as a spread of such features. The gesture based theories model coarticulation as coproduction of gestures. Both features and gesture scores have to be derived heuristically from the phonological knowledge and they are incomplete and coarse representations of the speech production process.

Articulatory knowledge in the form of pseudoarticulatory representations (such as features and gestures) when incorporated in the structure of speech recognisers showed some potential in small vocabulary recognition tasks in noisy conditions. However, more fine grained representations are needed to capture the essence of speech production system and thereby, to model coarticulation in spontaneous and conversational speech. Purely statistical models built from the measured articulatory data fail to explain the constraints and characteristics of speech production system.

There is a need for building models of realistic speech articulation to identify and capture the essential characteristics of human articulators, such as target-driven behaviour, articulatory interdependencies and biomechanical constraints for efficiently modelling the effects of coarticulation. We propose a statistical yet explanatory approach for extracting and modelling the essence of speech articulation. Articulatory roles are identified as critical (i.e., target-driven and constrained), dependent (partially constrained due to the interarticulatory dependencies and partially prone to coarticulation) and redundant (completely unconstrained and hence maximally prone to coarticulation) using the EMA data from the MOCHA-TIMIT database (Wrench, 2001).

In the second half of this chapter, measured articulatory data from the MOCHA-TIMIT database was introduced along with the preprocessing information. Derivation of statistics from the data was explained along with the evaluation of Gaussian assumption. The distance measure used for identification of articulatory roles, Kullback Leibler (KL) divergence, was introduced.

The following section presents the proposed algorithm for identification of articulatory roles.

# Chapter 3

# Articulatory constraint identification algorithm

## 3.1 Overview

The algorithm for the identification of data-driven constraints from the EMA data is presented in this section. The EMA fleshpoint coordinates are used as a low dimensional representation of the articulators. Although they are continuously deformable, a few well-selected points can faithfully represent the full shape of the articulators with reasonable accuracy (Badin and Serrurier, 2006; Qin et al., 2008). The articulatory constraint identification algorithm (ACIDA) identifies critical, dependent and redundant roles played by articulatory coordinates for each phone. For the 1D case, the EMA data is treated as 14 separate 'articulators'; for the 2D case, x and y coordinates are combined to 7 'articulators'.

The KL divergence is used to quantify the distance between the grand and the phone specific distributions of each articulatory coordinate. The articulatory coordinate that is associated with the maximum divergence, if greater than a threshold value, is identified as critical. The algorithm uses the knowledge of correlations amongst the articulatory coordinates to identify the articulatory coordinates dependent on the critical coordinate(s). It is assumed that the grand distributions and correlations estimated from a phonetically balanced database reflect the gross biomechanical properties of human speech production. The articulators that are not correlated with the critical articulators are identified as redundant. The model distributions of articulatory coordinates were also estimated for critical, dependent and redundant coordinates.

The grand univariate and bivariate correlations amongst the articulatory coordinates were computed for 1D and 2D cases. The correlations are analysed in Section 3.2. The methodology of the algorithm is presented in Section 3.3. The working of the algorithm is presented in Section 3.3.3. Implementation issues and evaluation measures are discussed in Section 3.4.

| | ULx | ULy | LLx | LLy | LIx | LIy | TTx | TBx | TDx | TTy | TBy | TDy | Vx | Vy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ULx | 1.00 | 0.51 | 0.35 | -0.18 | 0.00 | -0.20 | -0.24 | -0.22 | -0.25 | 0.00 | 0.00 | 0.00 | 0.00 | -0.17 |
| ULy | 0.51 | 1.00 | 0.24 | -0.28 | 0.00 | 0.00 | -0.17 | -0.13 | -0.15 | 0.30 | 0.24 | 0.11 | 0.00 | -0.12 |
| LLx | 0.35 | 0.24 | 1.00 | 0.73 | 0.60 | 0.56 | -0.12 | -0.25 | -0.18 | -0.36 | 0.00 | 0.16 | -0.14 | -0.16 |
| LLy | -0.18 | -0.28 | 0.73 | 1.00 | 0.48 | 0.66 | 0.00 | 0.13 | 0.00 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 |
| LIx | 0.00 | 0.00 | 0.60 | 0.48 | 1.00 | 0.79 | 0.00 | 0.00 | 0.00 | 0.46 | -0.37 | 0.00 | 0.00 | 0.00 |
| LIy | -0.20 | 0.00 | 0.56 | 0.66 | 0.79 | 1.00 | 0.00 | 0.00 | 0.00 | 0.63 | 0.42 | 0.00 | 0.00 | 0.00 |
| TTx | -0.24 | -0.17 | -0.12 | 0.00 | 0.00 | 0.00 | 1.00 | 0.90 | 0.85 | 0.00 | 0.00 | 0.00 | 0.12 | 0.10 |
| TBx | -0.22 | -0.13 | -0.25 | 0.13 | 0.00 | 0.00 | 0.90 | 1.00 | 0.93 | 0.16 | 0.00 | -0.24 | 0.00 | 0.00 |
| TDx | -0.25 | -0.15 | -0.18 | 0.00 | 0.00 | 0.00 | 0.85 | 0.93 | 1.00 | 0.11 | 0.00 | -0.20 | 0.00 | 0.00 |
| TTy | 0.00 | 0.30 | -0.36 | 0.34 | 0.46 | 0.63 | 0.00 | 0.16 | 0.11 | 1.00 | 0.56 | 0.11 | 0.00 | 0.00 |
| TBy | 0.00 | 0.24 | 0.00 | 0.00 | -0.37 | 0.42 | 0.00 | 0.00 | 0.00 | 0.56 | 1.00 | 0.75 | 0.00 | 0.14 |
| TDy | 0.00 | 0.11 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | -0.24 | -0.20 | 0.11 | 0.75 | 1.00 | 0.14 | 0.29 |
| Vx | 0.00 | 0.00 | -0.14 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.79 |
| Vy | -0.17 | -0.12 | -0.16 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 | 0.14 | 0.29 | 0.79 | 1.00 |

Table 3.1: *The 1D grand correlation matrix $R^*$ generated from the male speaker data depicting statistically significant correlations ($\alpha = 0.05$), also $|r_{ij}| > 0.1$, for $i, j \in 1..a$.*

## 3.2 Inter-articulatory correlations

### Univariate correlations

Grand univariate correlations were computed from 1D articulatory data, $R = \{r_{ij}\} \forall i, j \in \{1..a\}$ where $a$ is the number of articulatory coordinates (14 in the 1D case and 7 in the 2D case). The statistical significance of the univariate correlations was tested using Pearson's test at level of significance, $\alpha = 0.05$. The procedure for Pearson's test is given in B.3.3. Statistically insignificant and very weak ($|r_{ij}| < 0.1$) correlations were set to zero. Table 3.1 depicts the grand 1D correlation matrix $R^*$ with the remaining correlations for the male speaker. Recall from Chapter 2 that $M_i$ and $\Sigma_{ii}$ denote the grand mean and variance of the 1D positional distribution of any articulator $i$, $i \in \{1..a\}$. The covariance between any two articulators $i$ and $j$ was estimated from $R^*$ as $\Sigma_{ij} = \Sigma_{ii}^{1/2} r_{ij} \Sigma_{jj}^{1/2}$.

Table 3.1 shows statistically significant and strong correlations between tongue tip, blade and dorsum in the x direction for the male speaker. There was little correlation between the x and y movements of TT, TB and TD which indicates the independent movements of the tongue in horizontal and vertical directions. The correlation between $\text{TT}_y$ and $\text{TD}_y$ was small (0.11), which shows that the vertical movement of the tongue dorsum has a very small effect on the movement of the tongue tip. The magnitude of the correlations between $\text{TT}_y$ and $\text{TB}_y$, $\text{TB}_y$ and $\text{TD}_y$ were less than those in the x direction. The lower lip and incisor movements are strongly correlated in the x and y directions. The correlations between the upper lip and the lower lip were stronger than those between the upper lip and the jaw (LI). Some correlations existed between the jaw and the tongue tip in the y direction. Strong correlation existed between velum x and y movements. The velum almost had no correlation with other articulators. The articulatory system behaved like three largely independent components: the lip and the

| | ULx | ULy | LLx | LLy | LIx | LIy | TTx | TBx | TDx | TTy | TBy | TDy | Vx | Vy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ULx | 1.00 | 0.11 | 0.55 | -0.38 | 0.00 | -0.15 | -0.24 | -0.28 | -0.21 | -0.12 | 0.00 | 0.00 | 0.00 | 0.00 |
| ULy | 0.11 | 1.00 | -0.21 | -0.27 | 0.00 | 0.18 | 0.00 | 0.00 | -0.15 | 0.32 | 0.22 | 0.00 | 0.00 | 0.00 |
| LLx | 0.55 | -0.21 | 1.00 | -0.36 | 0.53 | 0.46 | 0.00 | 0.00 | 0.11 | -0.32 | 0.00 | 0.12 | 0.00 | 0.00 |
| LLy | -0.38 | -0.27 | -0.36 | 1.00 | -0.27 | 0.54 | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 | -0.13 | 0.14 | 0.00 |
| LIx | 0.00 | 0.00 | 0.53 | -0.27 | 1.00 | 0.48 | 0.19 | 0.33 | 0.37 | -0.33 | 0.00 | 0.13 | -0.30 | 0.00 |
| LIy | -0.15 | 0.18 | 0.46 | 0.54 | 0.48 | 1.00 | -0.17 | -0.18 | -0.23 | 0.59 | 0.33 | 0.00 | 0.15 | 0.00 |
| TTx | -0.24 | 0.00 | 0.00 | 0.00 | 0.19 | -0.17 | 1.00 | 0.86 | 0.79 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 |
| TBx | -0.28 | 0.00 | 0.00 | 0.00 | 0.33 | -0.18 | 0.86 | 1.00 | 0.91 | 0.00 | -0.12 | 0.00 | -0.14 | 0.00 |
| TDx | -0.21 | -0.15 | 0.11 | 0.00 | 0.37 | -0.23 | 0.79 | 0.91 | 1.00 | -0.11 | 0.00 | 0.00 | 0.00 | 0.00 |
| TTy | -0.12 | 0.32 | -0.32 | 0.30 | -0.33 | 0.59 | 0.14 | 0.00 | -0.11 | 1.00 | 0.42 | 0.00 | 0.23 | 0.16 |
| TBy | 0.00 | 0.22 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | -0.12 | 0.00 | 0.42 | 1.00 | 0.74 | 0.20 | 0.00 |
| TDy | 0.00 | 0.00 | 0.12 | -0.13 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.74 | 1.00 | 0.22 | 0.17 |
| Vx | 0.00 | 0.00 | 0.00 | 0.14 | -0.30 | 0.15 | 0.00 | -0.14 | 0.00 | 0.23 | 0.20 | 0.22 | 1.00 | 0.58 |
| Vy | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.17 | 0.58 | 1.00 |

Table 3.2: *The 1D grand correlation matrix $R^*$ generated from the female speaker data depicting statistically significant correlations ($\alpha = 0.05$), also $|r_{ij}| > 0.1$, for $i, j \in 1..a$.*

jaw group, tongue and velum.

Statistically significant correlations generated from the female speaker data in a similar way are shown in Table 3.2. The correlation patterns in female speaker data were found to be slightly different from those found in male data. The correlations between x and y movements of UL, LL and LI were relatively weaker and $\text{UL}_y$ and $\text{LL}_x$ were negatively correlated. The x movement of the jaw was correlated with the x movements of the TT, TB and TD which were absent in the male speaker data. These slight variations in the correlations could be due to the difference in the style of speaking between the speakers. Different points on the tongue were strongly correlated with each other in the x direction in a similar fashion to that of the male speaker, also velum had very small or no correlations with the rest of the articulators. No correlation was present between $\text{TT}_y$ and $\text{TD}_y$ which indicates the independent movements of tongue tip. For both speakers, about 37% of the total correlations were found to be statistically insignificant or very weak.

### Bivariate correlations

The bivariate correlations between articulators were computed using canonical correlation analysis (Johnson and Wichern, 1998) from the 2D articulatory data for male and female speakers. This analysis employs singular value decomposition to find the direction in which every pair of articulators are maximally correlated. The eigenvectors indicate the directions of the correlations and the strength of correlations is given by eigenvalues. The maximum number of canonical correlations that can be computed between a pair of articulators is equal to the dimensionality of the data, here 2. The statistical significance of canonical correlations was tested at level of significance $\alpha = 0.05$. Statistically insignificant and very weak canonical correlation values ($|\rho| < 0.15$) were set to zero. Let $\rho_{ij} = \text{diag}(\rho^1_{ij}, \rho^2_{ij})$ be the pair of statistically significant canonical

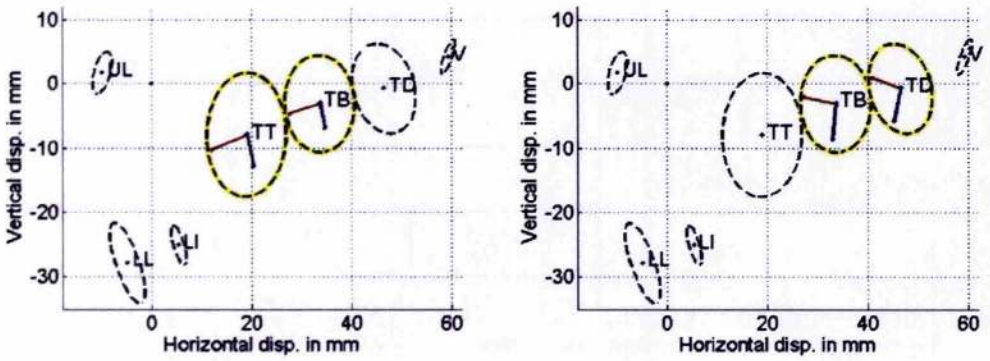



Figure 3.1: *Directions of first (red) and second (blue) canonical correlations between* TT-TB *(left) and* TB-TD *(right) for male speaker data generated from corresponding eigenvectors and correlation values. The articulatory pairs are highlighted in yellow.*

correlations between any two articulators $i$ and $j$. The direction in which articulatory coordinates $i$ and $j$ are correlated is denoted by the eigenvectors $\mathbf{U}_i^1$ and $\mathbf{V}_j^1$ for canonical correlation $\rho_{ij}^1$; eigenvectors, $\mathbf{U}_i^2$ and $\mathbf{V}_j^2$, denote the direction of canonical correlation $\rho_{ij}^2$.

Figure 3.1 shows the canonical correlations between two different pairs of articulators for male speaker data: TT-TB and TB-TD. The covariance ellipse of each articulatory coordinate indicates its grand covariance. The canonical correlation value is a fraction of the grand covariance when a pair of articulatory coordinates are maximally correlated in the directions shown by the eigenvectors. The eigenvector directions for TT-TB and TB-TD indicate the strong forward/backward movements tangential to the tongue surface and the raising and lowering of the tongue. The canonical correlations for the forward/backward movements of the tongue ($\rho_{\text{TT,TB}}^1 = 0.93, \rho_{\text{TB,TD}}^1 = 0.93$) were stronger than the correlations for raising and lowering of the tongue ($\rho_{\text{TT,TB}}^2 = 0.53, \rho_{\text{TB,TD}}^2 = 0.75$). Some articulatory pairs (33%) had two significant and strong canonical correlations, no canonical correlation was found for 14% of the total articulatory pairs and the other pairs (52%) had one significant correlation value each. Canonical correlation analysis was also done on the female speaker data. No canonical correlations were found between two articulatory pairs: V-UL and TB-LL. Approximately 38% of all articulatory pairs had two significant canonical correlations, other 52% of the articulatory pairs had one significant canonical correlation value.

Bivariate correlations were estimated from statistically significant canonical correlations and corresponding eigenvectors for both male and female speakers. Tables 3.3 and 3.4 depict the bivariate correlations for the male and the female speakers respectively. For both speakers, the bivariate correlations were similar to the univariate correlations in absolute value.

| | ULx | ULy | LLx | LLy | LIx | LIy | TTx | TBx | TDx | TTy | TBy | TDy | Vx | Vy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ULx | 1.00 | 0.51 | 0.35 | -0.18 | 0.08 | -0.22 | -0.24 | -0.22 | -0.25 | 0.00 | -0.02 | -0.07 | -0.10 | -0.17 |
| ULy | 0.51 | 1.00 | 0.24 | -0.28 | -0.02 | 0.05 | -0.17 | -0.13 | -0.15 | 0.30 | 0.24 | 0.11 | -0.07 | -0.12 |
| LLx | 0.35 | 0.24 | 1.00 | 0.73 | 0.60 | 0.58 | -0.09 | -0.25 | -0.18 | -0.36 | -0.04 | 0.16 | -0.15 | -0.15 |
| LLy | -0.18 | -0.28 | 0.73 | 1.00 | 0.48 | 0.66 | 0.08 | 0.14 | 0.09 | 0.32 | 0.02 | -0.08 | 0.06 | 0.06 |
| LIx | 0.08 | -0.02 | 0.60 | 0.48 | 1.00 | 0.70 | 0.03 | -0.08 | 0.00 | 0.46 | -0.37 | 0.00 | 0.00 | 0.00 |
| LIy | -0.22 | 0.05 | 0.58 | 0.66 | 0.70 | 1.00 | -0.04 | 0.09 | 0.00 | 0.63 | 0.42 | 0.00 | 0.00 | 0.00 |
| TTx | -0.24 | -0.17 | -0.09 | 0.08 | 0.03 | -0.04 | 1.00 | 0.90 | 0.85 | 0.03 | 0.07 | 0.01 | 0.00 | 0.00 |
| TBx | -0.22 | -0.13 | -0.25 | 0.14 | -0.08 | 0.09 | 0.90 | 1.00 | 0.93 | 0.16 | -0.02 | -0.24 | -0.01 | -0.06 |
| TDx | -0.25 | -0.15 | -0.18 | 0.09 | 0.00 | 0.00 | 0.85 | 0.93 | 1.00 | 0.13 | -0.02 | -0.20 | -0.02 | -0.04 |
| TTy | 0.00 | 0.30 | -0.36 | 0.32 | 0.46 | 0.63 | 0.03 | 0.16 | 0.13 | 1.00 | 0.56 | 0.00 | 0.00 | 0.00 |
| TBy | -0.02 | 0.24 | -0.04 | 0.02 | -0.37 | 0.42 | 0.07 | -0.02 | -0.02 | 0.56 | 1.00 | 0.75 | 0.02 | 0.12 |
| TDy | -0.07 | 0.11 | 0.16 | -0.08 | 0.00 | 0.00 | 0.01 | -0.24 | -0.20 | 0.00 | 0.75 | 1.00 | 0.15 | 0.29 |
| Vx | -0.10 | -0.07 | -0.15 | 0.06 | 0.00 | 0.00 | 0.00 | -0.01 | -0.02 | 0.00 | 0.02 | 0.15 | 1.00 | 0.79 |
| Vy | -0.17 | -0.12 | -0.15 | 0.06 | 0.00 | 0.00 | 0.00 | -0.06 | -0.04 | 0.00 | 0.12 | 0.29 | 0.79 | 1.00 |

Table 3.3: *The 2D grand correlation matrix $R^*$ generated from the male speaker data depicting statistically significant correlations ($\alpha = 0.05$), also $|r_{ij}| > 0.1$, for $i, j \in 1..a$.*

| | ULx | ULy | LLx | LLy | LIx | LIy | TTx | TBx | TDx | TTy | TBy | TDy | Vx | Vy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ULx | 1.00 | 0.11 | 0.55 | -0.38 | 0.02 | -0.16 | -0.24 | -0.28 | -0.21 | -0.12 | 0.02 | 0.03 | 0.00 | 0.00 |
| ULy | 0.11 | 1.00 | -0.21 | -0.27 | -0.03 | 0.18 | 0.05 | -0.08 | -0.15 | 0.32 | 0.22 | 0.02 | 0.00 | 0.00 |
| LLx | 0.55 | -0.21 | 1.00 | -0.36 | 0.53 | 0.48 | -0.05 | 0.00 | 0.11 | -0.32 | 0.00 | 0.12 | 0.02 | 0.01 |
| LLy | -0.38 | -0.27 | -0.36 | 1.00 | -0.27 | 0.54 | 0.05 | 0.00 | 0.06 | 0.30 | 0.00 | -0.13 | 0.14 | 0.06 |
| LIx | 0.02 | -0.03 | 0.53 | -0.27 | 1.00 | 0.40 | 0.12 | 0.33 | 0.37 | -0.36 | -0.06 | 0.10 | -0.30 | -0.07 |
| LIy | -0.16 | 0.18 | 0.48 | 0.54 | 0.40 | 1.00 | -0.19 | -0.18 | -0.21 | 0.58 | 0.33 | -0.06 | 0.14 | 0.03 |
| TTx | -0.24 | 0.05 | -0.05 | 0.05 | 0.12 | -0.19 | 1.00 | 0.86 | 0.76 | 0.14 | 0.01 | -0.03 | -0.04 | -0.02 |
| TBx | -0.28 | -0.08 | 0.00 | 0.00 | 0.33 | -0.18 | 0.86 | 1.00 | 0.91 | -0.02 | -0.12 | -0.10 | -0.13 | -0.05 |
| TDx | -0.21 | -0.15 | 0.11 | 0.06 | 0.37 | -0.21 | 0.76 | 0.91 | 1.00 | -0.11 | -0.06 | 0.10 | -0.04 | -0.03 |
| TTy | -0.12 | 0.32 | -0.32 | 0.30 | -0.36 | 0.58 | 0.14 | -0.02 | -0.11 | 1.00 | 0.42 | 0.00 | 0.23 | 0.13 |
| TBy | 0.02 | 0.22 | 0.00 | 0.00 | -0.06 | 0.33 | 0.01 | -0.12 | -0.06 | 0.42 | 1.00 | 0.74 | 0.20 | 0.07 |
| TDy | 0.03 | 0.02 | 0.12 | -0.13 | 0.10 | -0.06 | -0.03 | -0.10 | 0.10 | 0.00 | 0.74 | 1.00 | 0.23 | 0.16 |
| Vx | 0.00 | 0.00 | 0.02 | 0.14 | -0.30 | 0.14 | -0.04 | -0.13 | -0.04 | 0.23 | 0.20 | 0.23 | 1.00 | 0.58 |
| Vy | 0.00 | 0.00 | 0.01 | 0.06 | -0.07 | 0.03 | -0.02 | -0.05 | -0.03 | 0.13 | 0.07 | 0.16 | 0.58 | 1.00 |

Table 3.4: *The 2D grand correlation matrix $R^*$ generated from the female speaker data depicting statistically significant correlations ($\alpha = 0.05$), also $|r_{ij}| > 0.1$, for $i, j \in 1..a$.*

## 3.3 Articulatory constraint identification algorithm

### 3.3.1 Outline

The algorithm for identifying the critical articulators for each phone using 2D data ($a = 7$) is presented in this section. This algorithm was also implemented for the 1D case by assuming independence between x and y movements of articulatory coordinates and treating each articulatory coordinate as a separate channel ($a = 14$). As a precursor to running the algorithm, the grand and the phone statistics are gathered along with the correlations amongst the articulatory coordinates. The statistics of the grand distribution, $\mathcal{N}(M_i, \Sigma_i)$, are estimated for each articulator $i$ from samples of all phones. The statistics of phone specific distribution, $\mathcal{N}(\mu_i^{\phi}, \Sigma_i^{\phi})$, are estimated from the samples specific to each phone $\phi$. Figure 3.2 depicts the 2D grand and phone specific distributions of TT for phones [s] and [b] from male speaker data.

The algorithm identifies a list of critical articulators for each phone iteratively until a stopping criteria is met. The iterative nature of this algorithm helps identify all the critical dimensions essential for producing each phone in the phone set. The KL divergence computation is one of the crucial stages in the process of identification of articulatory roles. Recall from Chapter 2 that KL divergence is a measure of distance between two distributions. It can be seen from Fig.3.2(a) that the phone [s] distribution for TT has a smaller variance when compared with the grand distribution. There is also a difference between the grand and phone specific means of TT for [s]. Hence, the covariance ellipse of TT is tightly constrained in both x and y directions. Here, the KL divergence between the grand and phone specific distributions for TT (with magnitude of 22) was found to be the maximum of all divergences computed from other articulatory coordinates. For [s] , the proposed algorithm identified TT as the first critical coordinate.

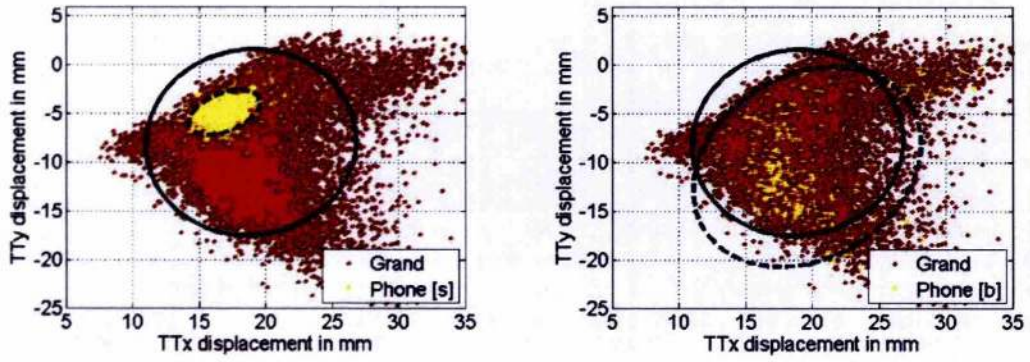



Figure 3.2: *Covariance ellipses for phones [s] (left) and [b] (right) depicting the grand (solid) and phone (dashed) distributions of* TT *in x and y directions. The samples of grand distribution are plotted in red and of phone distribution are plotted in yellow.*

Figure 3.2(b) shows that no such differences between the means and variances of the grand and the phone-specific distributions of TT exist for phone [b]. This is further

backed up by the fact that the KL divergence between the grand and the phone [b] distributions of TT was found to be very small (0.4). The large covariance ellipse of phone [b] distribution indicates that the tongue tip position is not constrained to be in any particular location and is free to move. Instead, the algorithm identified the upper lip coordinate which had the highest divergence as critical for [b].

The algorithm not only identifies the roles played by the articulators but also simultaneously estimates their target distributions, also known as *model distributions*. For each phone, $\phi$, the model distribution is defined for every articulator $i$ as $\mathcal{N}(m_i^\phi, S_i^\phi)$. The model distribution is initialised to the grand distribution before running the algorithm. As the algorithm runs, the model distribution is updated for each articulator $i$ depending on the its role in the production of phone $\phi$.

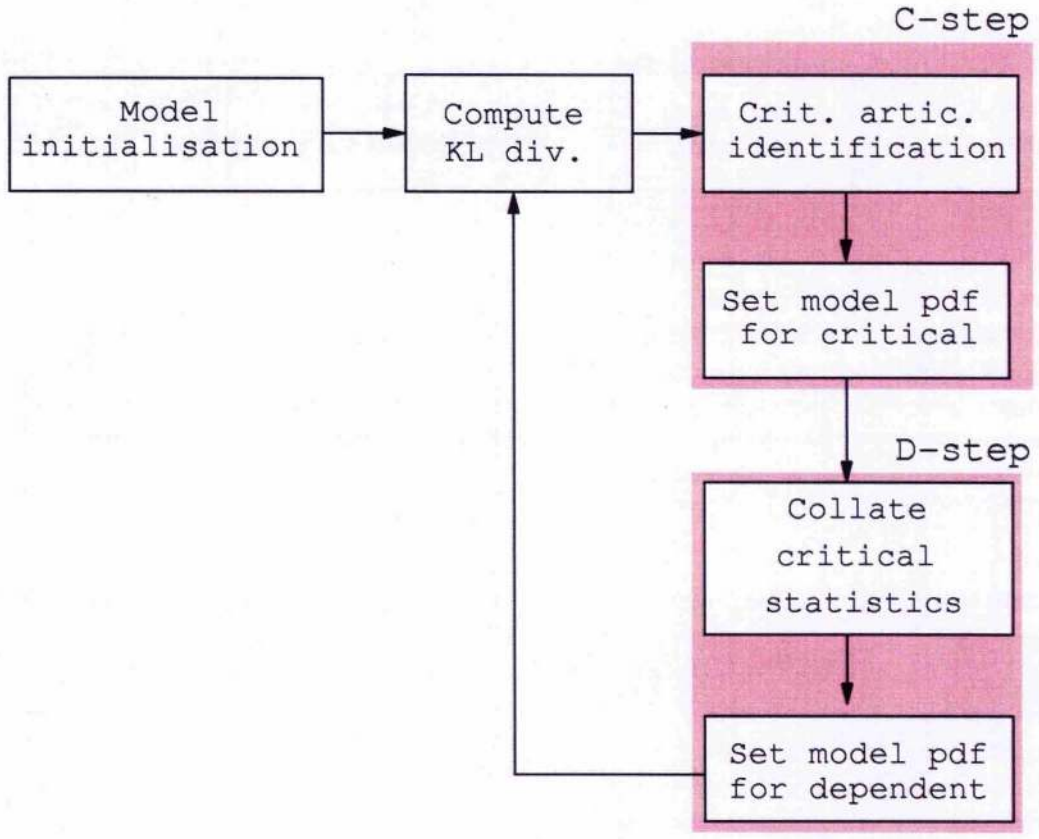



Figure 3.3: *Flow chart depicting different stages in the algorithm for identifying critical articulatory coordinates and updating model distributions for a phone $\phi$.*

Figure 3.3 shows the stages and flow of the algorithm for identification of articulatory roles. The algorithm operates in the following four stages

1. Model initialisation: In this stage, the model means and variances of all articulatory coordinates for each phone are set to the grand means and variances.

2. Divergence calculation: The symmetrical KL divergence between the model distributions and the phone-specific distributions of all articulatory coordinates for each phone, known as *identification divergence*, is calculated.

3. Critical identification step (C-step): Here, the articulatory dimension associated with the maximum divergence is identified as critical. The model distribution of the critical coordinate is updated by setting it to the phone specific distribution.

4. Dependent update step (D-step): The dependent coordinates are identified using grand articulatory correlations and their distributions are updated conditioned on the critical dimensions. This reflects the statistical properties of muscle and tissue linkages in speech production.

As shown in Fig.3.3, steps 2 to 4 are repeated until the identification divergence is greater than a threshold value known as the **critical** threshold. For each phone, the algorithm identifies a set of critical articulatory coordinates and estimates the model distributions of articulators according to the role played by them for that phone. The algorithm is explained in detail in the following section.

### 3.3.2 Algorithm

The 2D version of the algorithm is presented in this section which can be simplified for the 1D implementation by treating x and y movements of the articulatory coordinates as separate and independent channels. Figure 3.4 shows the pseudocode for the 2D implementation of the critical articulator identification algorithm.

#### Model definition

Let the phone set be denoted by $\Phi$, where $\Phi = \{[\text{ə}],[\text{b}],...[\text{ʒ}]\}$. Let the number of phones in $\Phi$ be $\varphi$. Recall from the previous section that the number of articulatory coordinates for the 2D case be denoted by $a$ ($a$=7 for the 2D and 14 for the 1D cases). The grand distribution for any articulatory coordinate $i \in \{1..a\}$ is denoted by $\mathbf{\Gamma}$ and defined using the 2D mean $\boldsymbol{M}_i$ and covariance $\mathbf{\Sigma}_i$ where

$$\boldsymbol{M}_i = \begin{bmatrix} M_x(i) \\ M_y(i) \end{bmatrix} \tag{3.1}$$

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma^2_{xx}(i) & \sigma_{xy}(i) \\ \sigma_{yx}(i) & \sigma^2_{yy}(i) \end{bmatrix} \tag{3.2}$$

Here the grand mean of the articulator $i$ in the x direction is denoted by $M_x(i)$ and in the y direction by $M_y(i)$. The grand variances in the x and y directions are given by $\sigma_{xx}(i)$ and $\sigma_{yy}(i)$ respectively. Note that $\sigma_{xx}(i)$ and $\sigma_{yy}(i)$ denote the variances but not standard deviations and $\sigma_{xx}(i) = \sigma^2_x(i)$ and $\sigma_{yy}(i) = \sigma^2_y(i)$. The grand covariance between x and y movements of $i$ is denoted by $\sigma_{xy}(i)$ (or $\sigma_{yx}(i)$). Number of grand samples is denoted by $N$. The set of global statistics along with the sample size is denoted by $\mathbf{\Gamma} = \{\boldsymbol{M}, \mathbf{\Sigma}, N\}$. The dimensionality of grand mean is $\boldsymbol{M}$ is $a \times K$ where K =

1 for 1D and 2 for 2D case. The dimensionality of grand variance is $\boldsymbol{\Sigma}$ is $a \times K \times K$. The grand correlation matrix with statistically significant and strong correlations derived from the canonical correlation analysis (section 3.2) is denoted by $\boldsymbol{R}^*$.

Similarly, the phone specific distribution, denoted by $\boldsymbol{\Lambda}^\phi$, estimated from the data of every phone $\phi \in \{\Phi\}$ for an articulator $i$ is defined using the 2D phone mean $\boldsymbol{\mu}_i^\phi$ and covariance matrix $\boldsymbol{\Sigma}_i^\phi$. The sample size of each phone $\phi$ is denoted by $\nu^\phi$. The set of phone specific statistics is denoted by $\boldsymbol{\Lambda}^\phi = \{\boldsymbol{\mu}^\phi, \boldsymbol{\Sigma}^\phi, \nu^\phi\}$. The phone-specific correlation matrix representing the correlations amongst the articulatory coordinates estimated from the data from each phone is denoted by $\boldsymbol{R}^\phi$.

The algorithm iterates through $k$ levels where $k$ denotes the length of critical articulator list, $0 \leq k \leq a$. At level $k$, the 2D model distribution for phone $\phi \in \{\Phi\}$ for an articulator $i$ is denoted by $\boldsymbol{\Delta}_i^{\phi,k}$. The model mean and the covariance are denoted by $\boldsymbol{m}^{\phi,k}$ and $\boldsymbol{S}^{\phi,k}$ respectively. The following sections explain the stages of the algorithm shown in Fig. 3.3 in more detail.

### Model initialisation

In the model initialisation stage shown in Fig. 3.4, the counter for levels, $k$, was initialised to zero. The model mean $\boldsymbol{m}_i^{\phi,k}$ was set equal to the grand mean $\boldsymbol{M}_i$ and the model covariance $\boldsymbol{S}_i^{\phi,k}$ to the grand covariance $\boldsymbol{\Sigma}_i$ for each articulator $i \in \{1..a\}$. The critical articulator list $\boldsymbol{C}^{\phi,k}$ and the dependent articulator list $\boldsymbol{D}^{\phi,k}$ were also initialised to null. The number of samples used to estimate the model distribution of each articulator $i$ at this stage $n_i^{\phi,k}$ was also set to the grand sample size $N$.

### Divergence calculation

The distance between model and phone pdfs of every articulator $i$ at any level $k$ was measured using KL divergence, called **identification divergence**, $J_i^{\phi,k}$. The function *computeIdiv* in Figure 3.4 computes the KL divergence between the model and phone-specific distributions. The effect of the sample sizes of grand and phone distributions on the estimation of the phone and grand means was incorporated before computation of the identification divergence. The standard error of sample mean provides an estimate of error in the estimating the population mean from the samples. The standard error of mean is given as

$$SE = \frac{\sigma}{\sqrt{n}} \tag{3.3}$$

where $\sigma$ is the standard deviation of a distribution with $n$ samples. Larger sample sizes provide a more accurate estimate of the mean and result in a smaller standard error value. The effect of different sample sizes on the estimation of grand and phone means was incorporated by adding the square of the standard error, known as variance of mean, to the respective covariances as shown in *computeIdiv* in Fig.3.4,

$$\boldsymbol{S}_1 = \boldsymbol{S}_i^{\phi,k} + (\boldsymbol{S}_i^{\phi,k}/n_i^{\phi,k}) \quad \boldsymbol{S}_2 = \boldsymbol{\Sigma}_i^\phi + (\boldsymbol{\Sigma}_i^\phi/\nu^\phi) \tag{3.4}$$

**Derive statistics**
Global statistics $\mathbf{\Gamma} = \{M, \mathbf{\Sigma}, N\}$, means $(a \times K)$, variances $(a \times K \times K)$ and sample size $(a \times 1)$
Grand correlation $\mathbf{R}^*$
Phone statistics $\mathbf{\Lambda}^\phi = \{\mu^\phi, \mathbf{\Sigma}^\phi, \nu^\phi\}$, means $(a \times K)$, variances $(a \times K \times K)$ and sample size $(a \times 1)$
Phone correlation $R^\phi$
Model statistics $\mathbf{\Delta}^{\phi,k} = \{m^{\phi,k}, S^{\phi,k}, n^{\phi,k}\}$, means $(a \times K)$, variances $(a \times K \times K)$ and sample size $(a \times 1)$
Threshold $\Theta = \{\theta_C, \theta_D\}$
**Model initialisation**
level $k = 0$
$m_i^{\phi,k} = M_i$, $S_i^{\phi,k} = \mathbf{\Sigma}_i$, $n_i^{\phi,k} = N$, for all articulators $i \in \{1..a\}$
Critical articulator list: $C^{\phi,k} = \{\}$
Dependent articulator list: $D^{\phi,k} = \{\}$
Model convergence: $isConverged$ =FALSE
WHILE($(k \leq a)$ AND ($!isConverged$))
  **Compute identification divergence**
  $J_i^{\phi,k}$ = computeIdiv($\mathbf{\Delta}_i^{\phi,k}, \mathbf{\Lambda}_i^\phi$), for all articulators $i \in \{1..a\}$
  Find articulator with maximum divergence: $j = \text{argmax}\{J_1^{\phi,k}.., J_n^{\phi,k}\}$
  **C-step**
  IF($J_j^{\phi,k} > \theta_C$)
    Increment level: $k \leftarrow\!\!\!| \; k+1$
    Replicate model: $\mathbf{\Delta}^{\phi,k} = \mathbf{\Delta}^{\phi,k-1}$
    Add critical articulator: $C^{\phi,k} \leftarrow\!\!\!| \; \{C^{\phi,k-1}\} \cup \{j\}$
    Set distribution: $m_j^{\phi,k} \leftarrow\!\!\!| \; \mu_j^\phi$, $S_j^{\phi,k} \leftarrow\!\!\!| \; \mathbf{\Sigma}_j^\phi$
    $n_j^{\phi,k} \leftarrow\!\!\!| \; \nu^\phi$
    **D-step**
    [ $\mathbf{\Delta}^{\phi,k}, D^{\phi,k}$] =updateDep($\mathbf{\Gamma}, \mathbf{R}^*, \mathbf{\Lambda}^\phi, R^\phi, \Theta, J^{\phi,k-1}, D^{\phi,k-1}, C^{\phi,k}$)
  ELSE
    $isConverged$ =TRUE
    Store final critical articulator list: $\hat{C}^\phi = C^{\phi,k}$
    Store final dependent articulator list: $\hat{D}^\phi = D^{\phi,k}$
    Obtain final redundant articulator list by elimination: $\hat{\mathcal{R}}^\phi = \{1..a\} - \{\hat{C}^\phi\} - \{\hat{D}^\phi\}$
    Store model statistics: $\hat{m}^\phi = m^{\phi,k}$, $\hat{S}^\phi = S^{\phi,k}$
    Store no: of critical articulators: $K^\phi = k$.
  END IF
END WHILE

**function computeIdiv($\mathbf{\Delta}_i^{\phi,k}, \mathbf{\Lambda}_i^\phi$)**
Incorporate standard error: $S_1 = S_i^{\phi,k} + (S_i^{\phi,k}/n_i^{\phi,k})$, $S_2 = \mathbf{\Sigma}_i^\phi + (\mathbf{\Sigma}_i^\phi/\nu^\phi)$
$J = \frac{1}{2}\left(tr(S_1 - S_2)(S_2^{-1} - S_1^{-1}) + tr(S_1^{-1} + S_2^{-1})(m_i^{\phi,k} - \mu_i^\phi)(m_i^{\phi,k} - \mu_i^\phi)'\right)$
RETURN $J$

**function updateDep($\mathbf{\Gamma}, \mathbf{R}^*, \mathbf{\Lambda}^\phi, R^\phi, \Theta, J^{\phi,k-1}, D^{\phi,k-1}, C^{\phi,k}$)**
Initialise dependent list at current level by eliminating any critical articulators if present:
$D^{\phi,k} = \{D^{\phi,k-1}\} - \{C^{\phi,k}\}$
Get critical grand statistics from $\mathbf{\Gamma}$ and $\mathbf{R}^*$: $M_{\{C\}} = \{M_i\}_{i \in C^{\phi,k}}$; $\mathbf{\Sigma}_{\{C\}\{C\}} = \{\mathbf{\Sigma}_{ij}\}_{i,j \in C^{\phi,k}}$
Get critical phone statistics from $\mathbf{\Lambda}^\phi$ and $R^\phi$: $\mu_{\{C\}}^\phi = \{\mu_i^\phi\}_{i \in C^{\phi,k}}$; $\mathbf{\Sigma}_{\{C\}\{C\}}^\phi = \{\mathbf{\Sigma}_{ij}^\phi\}_{i,j \in C^{\phi,k}}$
FOR $i \in \{1..a\} - \{C^{\phi,k}\}$
  IF($J_i^{\phi,k-1} > \theta_D$)
    Update dependent list: $D^{\phi,k} \leftarrow\!\!\!| \; \{D^{\phi,k}\} \cup \{i\}$
    Get dependent covariance : $\mathbf{\Sigma}_{i\{C\}} = \{\mathbf{\Sigma}_{i,j}\}_{j \in C^{\phi,k}}$
    Update mean: $m_i^{\phi,k} \leftarrow\!\!\!| \; M_i + \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}_{\{C\}\{C\}}^{-1}(\mu_{\{C\}}^\phi - M_{\{C\}})$
    Update variance: $S_i^{\phi,k} \leftarrow\!\!\!| \; \mathbf{\Sigma}_i + \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}_{\{C\}\{C\}}^{-1}(\mathbf{\Sigma}_{\{C\}\{C\}}^\phi - \mathbf{\Sigma}_{\{C\}\{C\}})\mathbf{\Sigma}_{\{C\}\{C\}}^{-1}\mathbf{\Sigma}_{i\{C\}}'$
    Update sample size: $n_i^{\phi,k} \leftarrow\!\!\!| \; \nu^\phi$
  END IF
END FOR
RETURN $\mathbf{\Delta}^{\phi,k}$ and $D^{\phi,k}$

Figure 3.4: *Algorithm for articulatory constraint identification for phone $\phi$, including functions for computing KL divergence and updating model distributions using critical articulator information and inter-articulatory correlations. For 1D (K=1) or 2D (K=2) versions, use scalar or vector means, $M$, $\mu^\phi$ and $m^{\phi,k}$ and scalar or matrix (co-) variances $\mathbf{\Sigma}$, $\mathbf{\Sigma}^\phi$ and $S^{\phi,k}$.*

### C-step

In the **critical identification step** or the **C-step** of the algorithm, the articulator $j$ with maximum identification divergence $J^{\phi}_{k,j}$ is identified and the corresponding model distribution is updated. If the identification divergence is greater than the critical threshold value $\theta_C$, the algorithm progresses to the next level, otherwise terminates. The model information from the previous level $\mathbf{\Delta}^{\phi,k-1}$ and the critical articulator list $\boldsymbol{C}^{\phi,k-1}$ are propagated to the current level $k$. The articulator $j$ is identified as critical and added to the list of critical articulators identified up to the level $k$.

$$\boldsymbol{C}^{\phi,k} \leftarrow\!\shortmid \{\boldsymbol{C}^{\phi,k-1}\} \cup \{j\} \tag{3.5}$$

The model distribution for the critical articulatory coordinate $j$ is updated by setting it to the phone specific distribution. The sample size of the model distribution for $j$ is also updated to the phone sample size $\nu^{\phi}$.

$$\boldsymbol{m}_j^{\phi,k} \leftarrow\!\shortmid\ \boldsymbol{\mu}_j^{\phi} \tag{3.6}$$

$$\boldsymbol{S}_j^{\phi,k} \leftarrow\!\shortmid\ \boldsymbol{\Sigma}_j^{\phi} \tag{3.7}$$

$$n_j^{\phi,k} \leftarrow\!\shortmid\ \nu^{\phi} \tag{3.8}$$

$$\tag{3.9}$$

### D-step

In the **dependent update step** or the **D-step**, the distributions of the articulatory coordinates other than the critical articulatory coordinates are updated using the function *updateDep* shown in Figure 3.4. The distribution of the dependent articulator is updated conditioned on the critical articulator information and the inter-articulatory correlations. A **dependent threshold** value, $\theta_D$, is introduced in the D-step to prevent the models of dependent articulators from getting over-updated when critical threshold, $\theta_C$ on identification divergence is set to a very small value (less than 0.1). Only the articulators with divergence greater than $\theta_D$ are updated. The distribution of the redundant articulator remains unchanged, i.e., equivalent to the grand distribution.

Let the list of critical coordinates upto and including level $k = 2$ be $\{\boldsymbol{C}^{\phi,k}\} = \{j_1, j_2\}$. If $j_1$ and $j_2$ are correlated, then there is a good chance that $j_2$ will be in the dependent list at level $k-1$, i.e., $\boldsymbol{D}^{\phi,k-1}$. The dependent coordinate list at level $k$ is initialised by propagating all previous dependent list after excluding any new critical dimensions if present, to the current level, i.e.,

$$\boldsymbol{D}^{\phi,k} = \{\boldsymbol{D}^{\phi,k-1}\} - \{\boldsymbol{C}^{\phi,k}\} \tag{3.10}$$

The D-step is carried out in two stages by

- collating critical statistics
- updating model distribution for dependent articulatory coordinates

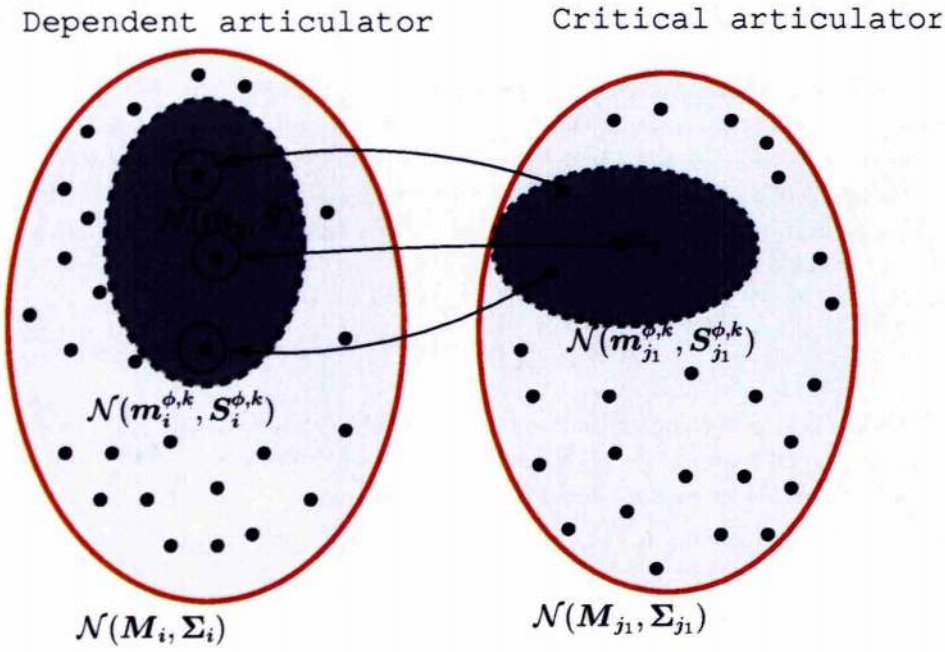



Figure 3.5: *Illustration of the estimation of the statistics of dependent articulatory coordinate $i$ from the knowledge of critical articulatory coordinate $j_1$. Grand covariance (solid red) and model covariance (dashed blue, grey background) ellipses representing grand distributions $\mathcal{N}(\boldsymbol{M}, \boldsymbol{\Sigma})$ and model distributions $\mathcal{N}(\boldsymbol{m}^{\phi,k}, \boldsymbol{S}^{\phi,k})$ are shown along with the samples $x$ (black dots).*

The statistics of critical articulators identified up to and including level $k$ are collated to estimate their combined effect on the positions of the rest of the articulatory dimensions. If $j_1$ and $j_2$ were identified as critical dimensions up to and including level $k = 2$, the grand statistics for 2D case are collated in function *updateDep* as

$$\boldsymbol{M}_{\{\boldsymbol{C}\}} = \begin{bmatrix} M_x(j_1) \\ M_y(j_1) \\ M_x(j_2) \\ M_y(j_2) \end{bmatrix} \tag{3.11}$$

$$\boldsymbol{\Sigma}_{\{\boldsymbol{C}\}\{\boldsymbol{C}\}} = \begin{bmatrix} \sigma_{xx}(j_1) & \sigma_{xy}(j_1) & \sigma_{xx}(j_1, j_2) & \sigma_{xy}(j_1, j_2) \\ \sigma_{yx}(j_1) & \sigma_{yy}(j_1) & \sigma_{yx}(j_1, j_2) & \sigma_{yy}(j_1, j_2) \\ \sigma_{xx}(j_2, j_1) & \sigma_{xy}(j_2, j_1) & \sigma_{xx}(j_2) & \sigma_{xy}(j_2) \\ \sigma_{yx}(j_2, j_1) & \sigma_{yy}(j_2, j_1) & \sigma_{yx}(j_2) & \sigma_{yy}(j_2) \end{bmatrix} \tag{3.12}$$

The grand covariance matrices $\boldsymbol{\Sigma}_{j_1}$ and $\boldsymbol{\Sigma}_{j_2}$ form the principal diagonal elements of the matrix $\boldsymbol{\Sigma}_{\{\boldsymbol{C}\}\{\boldsymbol{C}\}}$ and the rest of the covariances are estimated from the grand correlation matrix $\boldsymbol{R}^*$. Note that $\sigma_{xx}(j_1)$ denotes the variance but not the standard deviation and $\sigma_{xx}(j_1) = \sigma_x^2(j_1)$. The term $\sigma_{xy}(j_1, j_2)$ in matrix $\boldsymbol{\Sigma}_{\{\boldsymbol{C}\}\{\boldsymbol{C}\}}$ represents the covariance between x dimension of $j_1$ and y dimension of $j_2$. The phone statistics of critical articulators are also collated in a similar way to form phone mean matrix $\boldsymbol{\mu}_{\{\boldsymbol{C}\}}^{\phi}$

and covariance matrix $\mathbf{\Sigma}^{\phi}_{\{C\}\{C\}}$.

If $i$ represents a non-critical coordinate and the identification divergence $J_i^{\phi,k} > \theta_D$, then the covariance between $i$ and $\{C\}$, known as the dependent covariance $\mathbf{\Sigma}_{i\{C\}}$, is estimated from grand correlation matrix $\boldsymbol{R}^*$ as

$$\mathbf{\Sigma}_{i\{C\}} = \begin{bmatrix} \sigma_{xx}(i,j_1) & \sigma_{xy}(i,j_1) & \sigma_{xx}(i,j_2) & \sigma_{xy}(i,j_2) \\ \sigma_{yx}(i,j_1) & \sigma_{yy}(i,j_1) & \sigma_{yx}(i,j_2) & \sigma_{yy}(i,j_2) \end{bmatrix} \tag{3.13}$$

Figure 3.5 illustrates a simple case of updating the model distribution of a dependent articulator $i$ based on the distribution of just one critical articulator $j_1$. Each sample $x_\eta$, $\eta \in \{1..n^{\phi,k}\}$ in the model distribution of critical dimension $j_1$, $\mathcal{N}(\boldsymbol{m}_{j_1}^{\phi,k}, \boldsymbol{S}_{j-1}^{\phi,k})$ is mapped to the dependent space using grand correlations. The resultant position in the dependent space after mapping $\mathcal{N}(\bar{\boldsymbol{m}}, \bar{\boldsymbol{S}})$ is given as

$$\bar{\boldsymbol{m}}_\eta = \boldsymbol{M}_i + \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}^{-1}_{\{C\}\{C\}}(\boldsymbol{x}_\eta - \boldsymbol{M}_{\{C\}}) \tag{3.14}$$

$$\bar{\boldsymbol{S}} = \mathbf{\Sigma}_i - \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}^{-1}_{\{C\}\{C\}}\mathbf{\Sigma}_{\{C\}i} \tag{3.15}$$

The above estimate is obtained by applying the theory of the multivariate conditional distributions presented in (Anderson, 1984). The complete derivation of Eq. 3.14 and Eq. 3.15 is given in appendix B.2 (Eq. B.20 and Eq. B.21 resp.).

The mean of the dependent coordinate distribution is then estimated as the mathematical expectation of the resultant sample distributions. The variance is estimated by averaging the squared distance of its possible values from the mean and adding the sample variance $\bar{\boldsymbol{S}}$,

$$\boldsymbol{m}^{\phi}_{k,i} = \mathscr{E}(\bar{\boldsymbol{m}}_\eta) \tag{3.16}$$

$$\boldsymbol{S}^{\phi}_{k,i} = \mathscr{E}(\bar{\boldsymbol{m}}_\eta - \boldsymbol{m}^{\phi}_{k,i})(\bar{\boldsymbol{m}}_\eta - \boldsymbol{m}^{\phi}_{k,i})' + \bar{\boldsymbol{S}} \tag{3.17}$$

Solving Eq. 3.16 and Eq. 3.17 using Eq. 3.14 and Eq. 3.15 results in the expressions presented in the *updateDep* function in the Fig. 3.4,

$$\boldsymbol{m}^{\phi}_{k,i} \leftarrow \boldsymbol{M}_i + \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}^{-1}_{\{C\}\{C\}}(\boldsymbol{\mu}^{\phi}_{\{C\}} - \boldsymbol{M}_{\{C\}}) \tag{3.18}$$

$$\boldsymbol{S}^{\phi}_{k,i} \leftarrow \mathbf{\Sigma}_i + \mathbf{\Sigma}_{i\{C\}}\mathbf{\Sigma}^{-1}_{\{C\}\{C\}}(\mathbf{\Sigma}^{\phi}_{\{C\}\{C\}} - \mathbf{\Sigma}_{\{C\}\{C\}})\mathbf{\Sigma}^{-1}_{\{C\}\{C\}}\mathbf{\Sigma}'_{i\{C\}} \tag{3.19}$$

The dependent list is updated by adding $i$ to the existing list

$$\boldsymbol{D}^{\phi,k} \leftarrow \{\boldsymbol{D}^{\phi,k}\} \cup \{i\} \tag{3.20}$$

**Critical, dependent and redundant lists**

The algorithm iterates through computation of KL divergence step, C-step and D-step, as long as maximum identification divergence at any level $k$, $max\{\boldsymbol{J}^{\phi,k}\} > \theta_C$. The execution stops at level $k$ for phone $\phi$ when $max\{\boldsymbol{J}^{\phi,k}\} < \theta_C$. The final list of critical articulatory coordinates for every phone $\phi$ is given by $\hat{\boldsymbol{C}}^{\phi}$, the number of critical coordinates identified is given by $\hat{k}^{\phi}$. The final list of dependent articulatory coordinates is given by $\hat{\boldsymbol{D}}^{\phi}$. The list of redundant articulators is estimated by eliminating the critical and dependent coordinates from the set of all articulatory coordinates

$$\hat{\boldsymbol{\mathcal{R}}}^{\phi} = \{1..a\} - \{\hat{\boldsymbol{C}}^{\phi}\} - \{\hat{\boldsymbol{D}}^{\phi}\} \tag{3.21}$$

The model distribution of each phone $\phi \in \{\Phi\}$ after identification of critical, dependent and redundant lists is given as $\mathcal{N}(\hat{\boldsymbol{m}}^{\phi}, \hat{\boldsymbol{S}}^{\phi})$.

The following section provides a graphical illustration of the working of the algorithm.

### 3.3.3 Working of the algorithm

Figure 3.6 illustrates operation of 1D and 2D versions of the algorithm of male speaker data for [g]. Grand, phone and model distributions are represented by dotted red, dashed green and solid blue covariance ellipses respectively. The major and minor axes of each ellipse depict $\pm 2\sigma$ about the mean in x and y directions respectively. The axes are aligned in the direction of eigenvectors of their covariances.

The model distributions after initialisation, C-step and D-step are illustrated in Figure 3.6. In the model initialisation stage, the model distributions are set to the grand distributions for both 1D and 2D cases. Figure 3.6(b) represents the distributions after the C-step. In the 1D case, $\text{TD}_y$ was identified as the first critical coordinate in the C-step. The identification divergence of $\text{TD}_y$ ($J = 15$) was greater than that given by other articulatory coordinates and hence was identified as critical. The model distribution of $\text{TD}_y$ was set to the phone distribution in the C-step. In the 2D case, TD was identified as critical ($J = 18$). The model distributions of both x and y coordinates of TD were set to phone-specific distributions in the C-step.

Figure 3.6(c) shows the distributions after the D-step. The model distributions of the rest of the articulatory coordinates were updated conditioned on the distribution of $\text{TD}_y$ in the 1D case. The distributions of $\text{UL}_y$, $\text{LL}_x$, $\text{TT}_y$, $\text{TB}_x$, $\text{TB}_y$, $\text{TD}_x$, $\text{V}_x$ and $\text{V}_y$ were updated in D-step. The critical articulator $\text{TD}_y$ influenced $\text{TB}_y$ to a greater extent when compared with other coordinates, the correlation between $\text{TD}_y$ and $\text{TB}_y$ was 0.75. The position of $\text{TT}_y$ was not affected due to the absence of correlation between $\text{TT}_y$ and $\text{TD}_y$. The distributions of other articulators were not significantly affected since the correlations with $\text{TD}_y$ were weak. In the 2D case, only LI was identified as redundant, The distributions of the remaining articulators were updated in the D-step. Similar to the findings in the 1D case, the distribution of TB was most effected by the critical articulator TD.

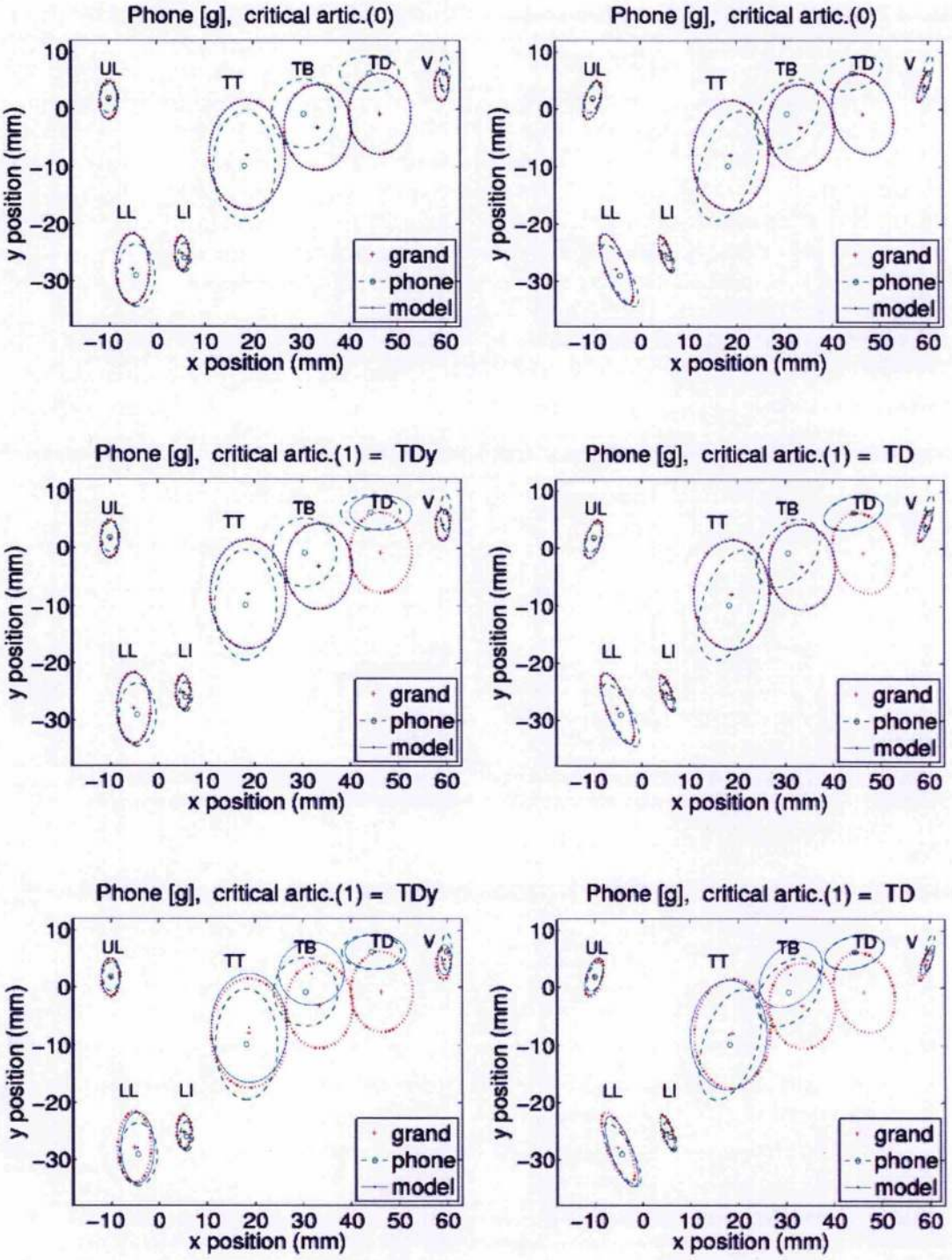



Figure 3.6: *Mid-saggital display of convergence of 1D (left) and 2D (right) phone models of [g] using grand (dotted red), phone (dashed green) and model (solid blue) distributions as critical articulator was identified up to and including level $k = 1$. From top to bottom, the figure illustrates the distributions after model initialisation, C-step and D-step respectively.*

## 3.4 Effect of critical threshold

The critical threshold $\theta_C$ used in the C-step of the algorithm is related to the number of critical dimensions identified by the algorithm for each phone . Decreasing the value of the critical threshold would result in an increase in the number of critical articulatory coordinates identified by the algorithm. The effect of varying critical threshold values on the performance of the algorithm was evaluated using two KL divergence based metrics known as the **convergence scale**, $\Upsilon_{conv}$ and **evaluation scale**, $\Upsilon_{eval}$. The convergence scale quantifies the goodness of fit of 1D and 2D model distributions to the respective 1D and 2D phone distributions and hence measures the model convergence. The evaluation scale is an indication of how well the 1D and 2D models fit the actual phone specific distributions. The following section defines the convergence and evaluation scales.

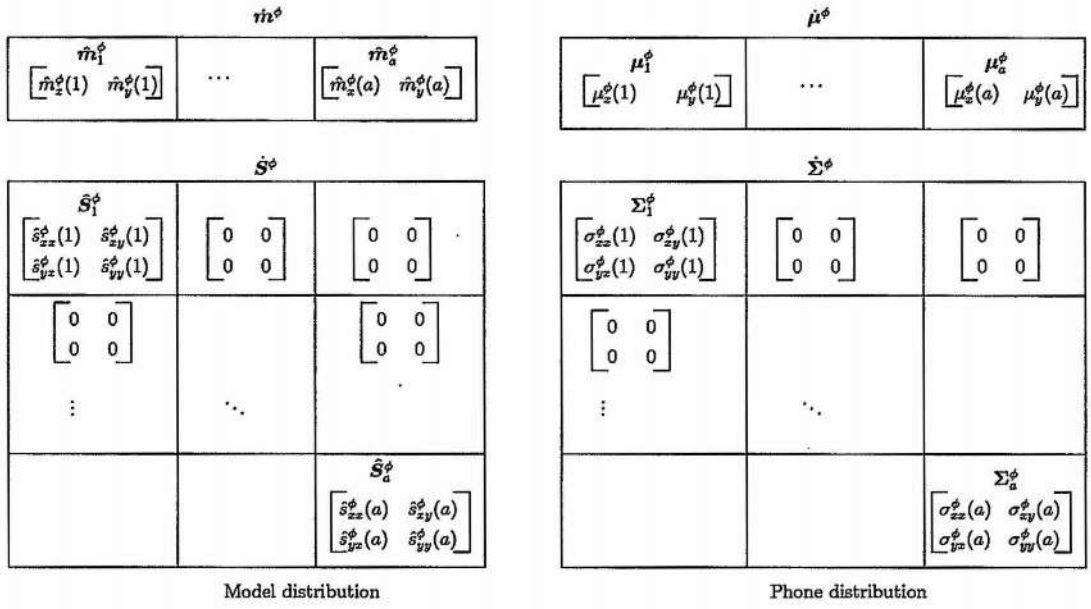



Figure 3.7: *Two dimensional model distributions (left) arranged to form a 14D mean vector $\dot{m}^{\phi}$ and 14D covariance matrix $\dot{S}^{\phi}$ and 2D phone distributions (right) with 14D mean vector $\dot{\mu}^{\phi}$ and 14D covariance matrix $\dot{\Sigma}^{\phi}$ used for computation of the convergence scale $\Upsilon_{conv}$.*

### 3.4.1 Convergence and evaluation scales

Convergence of 1D and 2D model distributions to the respective phone distributions was calculated using a KL divergence based metric known as **convergence scale**, $\Upsilon_{conv}$. Though convergence scale was estimated for both 1D and 2D cases, the method for estimation of only the 2D convergence scale is presented here. The 2D model means and covariances of all $a$ articulators were collated to form a 14D mean vector and 14D covariance matrix respectively as shown in Fig. 3.7. For the 2D case, model

covariances of each phone $\phi$ for all $a$ articulators, $\hat{\boldsymbol{S}}_i^{\phi}$, $\forall i \in \{1..a\}$, form the principal diagonal elements of the 14D model covariance matrix.

For the 1D case, the x and y movements of each articulator are treated independently and therefore $\hat{s}_{xy}^{\phi}(i)$ and $\hat{s}_{yx}^{\phi}(i)$ were set to zero. The 1D and 2D phone means and covariances were also arranged in a similar fashion as shown in Fig. 3.7 to form 14D mean vector and covariance matrix respectively. If $\mathcal{N}(\dot{\boldsymbol{m}}^{\phi}, \dot{\boldsymbol{S}}^{\phi})$ represents the 14D model distribution and $\mathcal{N}(\dot{\boldsymbol{\mu}}^{\phi}, \ddot{\boldsymbol{\Sigma}}^{\phi})$ represents the phone distribution shown in Fig. 3.7, the convergence scale was calculated as the 14D KL divergence according to the following equation

$$\Upsilon_{conv}^{\phi} = \frac{1}{2} tr(\dot{\boldsymbol{S}}^{\phi} - \ddot{\boldsymbol{\Sigma}}^{\phi})(\ddot{\boldsymbol{\Sigma}}^{\phi^{-1}} - \dot{\boldsymbol{S}}^{\phi^{-1}}) + \frac{1}{2} tr(\dot{\boldsymbol{S}}^{\phi^{-1}} + \ddot{\boldsymbol{\Sigma}}^{\phi^{-1}})(\dot{\boldsymbol{m}}^{\phi} - \dot{\boldsymbol{\mu}}^{\phi})(\dot{\boldsymbol{m}}^{\phi} - \dot{\boldsymbol{\mu}}^{\phi})' \quad (3.22)$$

The goodness of fit of the model distributions under 1D and 2D assumptions to the actual phone distribution with full 14D covariance matrix was measured using the **evaluation scale**, $\Upsilon_{eval}$. The arrangement of model and phone statistics for the computation of the **evaluation scale** for the 2D case is shown in Figure 3.8. Here the phone covariance matrix is 14D and full with the inclusion of off-diagonal covariances between articulators.

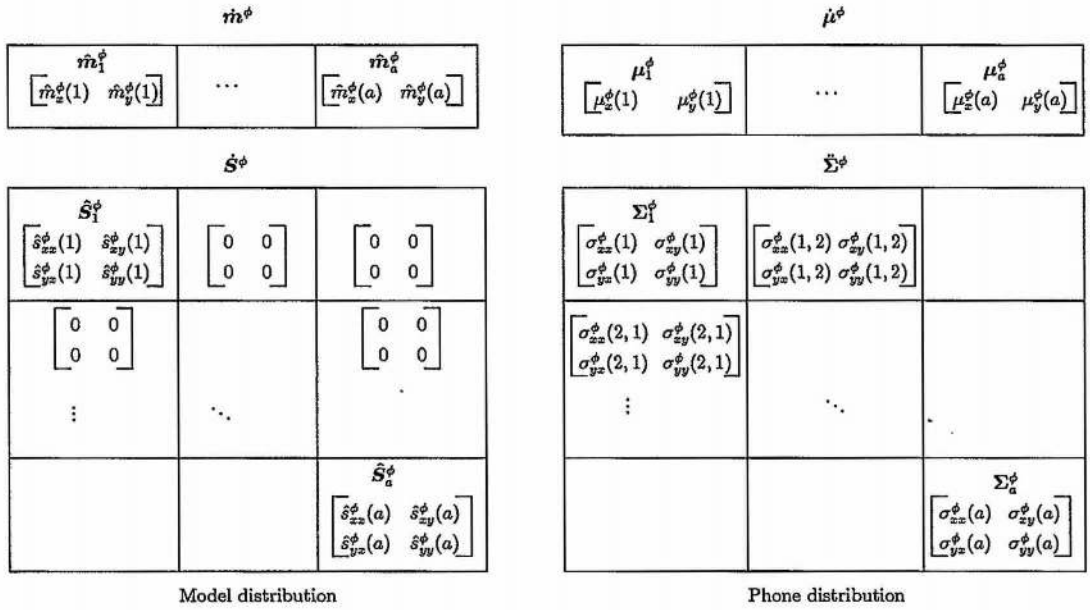



Figure 3.8: *Two dimensional model distributions (left) arranged to form a 14D mean vector $\dot{\boldsymbol{m}}^{\phi}$ and 14D covariance matrix $\dot{\boldsymbol{S}}^{\phi}$ and 2D phone distributions (right) with 14D mean vector $\dot{\boldsymbol{\mu}}^{\phi}$ and 14D covariance matrix $\ddot{\boldsymbol{\Sigma}}^{\phi}$ used for computation of the evaluation scale $\Upsilon_{eval}$.*

The evaluation divergence between the model distribution $\mathcal{N}(\dot{\boldsymbol{m}}^{\phi}, \dot{\boldsymbol{S}}^{\phi})$ and the phone distribution with full covariance matrix $\mathcal{N}(\dot{\boldsymbol{\mu}}^{\phi}, \ddot{\boldsymbol{\Sigma}}^{\phi})$ is calculated as the 14D KL diver-

gence averaged across all phones according to the equation

$$\Upsilon^{\phi}_{eval} = \frac{1}{2}tr(\dot{S}^{\phi} - \ddot{\Sigma}^{\phi})(\ddot{\Sigma}^{\phi^{-1}} - \dot{S}^{\phi^{-1}}) + \frac{1}{2}tr(\dot{S}^{\phi^{-1}} + \ddot{\Sigma}^{\phi^{-1}})(\dot{m}^{\phi} - \dot{\mu}^{\phi})(\dot{m}^{\phi} - \dot{\mu}^{\phi})' \quad (3.23)$$

The following sections describe the effect of varying critical thresholds on the model convergence and evaluation scales.

### 3.4.2 Trade off between model convergence and $\theta_C$

The convergence scale measures how well the 1D and 2D model distributions of each phone match respective 1D and 2D phone distributions at different levels of critical threshold values. The convergence scale was averaged across all phones for both 1D and 2D cases. The initial convergence scale was computed at level $k = 0$ between model and phone distributions before application of the algorithm. The initial convergence scale averaged across all phones for 1D case was found to be 20 for the male speaker and 21 for the female speaker. Figure 3.9 shows the plots depicting the average convergence scale and the average number of critical articulatory coordinates at various threshold values for 1D and 2D models for both male and female speakers. The x axis shows the average number of critical dimensions per phone and the y axis shows the convergence scale averaged across all phones for different values of critical thresholds, $0.1 \leq \theta_C \leq 5$. For 1D case, after running the algorithm at $\theta_C = 5$, the model convergence improved by 50% for both speakers when compared with the initial convergence. The average number of critical dimensions per phone at this level was 0.5 for both speakers. Incorporating information of critical dimensions and updating the model distributions of dependent articulators resulted in the improvement of the model convergence. Decreasing the critical threshold to 1 increased the average number of critical dimensions per phone to 2.5 for both speakers and the improvement obtained over the convergence at $\theta_C = 5$ was 72%. Decreasing the threshold beyond this point ($\theta_C = 1$) significantly increased the number of critical dimensions but only small improvement in convergence was obtained. At $\theta_C = 0.1$, a further 25% of improvement in model convergence was obtained but the number of critical dimensions per phone rose to 7.

Similar observations were made for the 2D case for both speakers. As the threshold was lowered from 5 to 1, the model convergence increased by 78% and the number of critical dimensions per phone doubled in number. Decreasing the critical threshold on identification divergence beyond 1 resulted in significant increase in the average number of critical dimensions per phone but only small improvements in convergence were obtained. This analysis showed that a minimum of 2 critical dimensions per phone are required to estimate model distributions with reasonable accuracy for both 1D and 2D cases. Increasing the dimensionality of critical articulator space beyond a certain point, i.e., at $\theta_C = 1$ for both 1D and 2D cases, does not improve the convergence of the models but increases the parameter space of the models.

### 3.4.3 Evaluation scale vs $\theta_C$

The goodness of fit of the 1D and the 2D model assumptions to the actual phone distributions measured using **evaluation scale**, $\Upsilon^{\phi}_{eval}$. The evaluation scale was computed

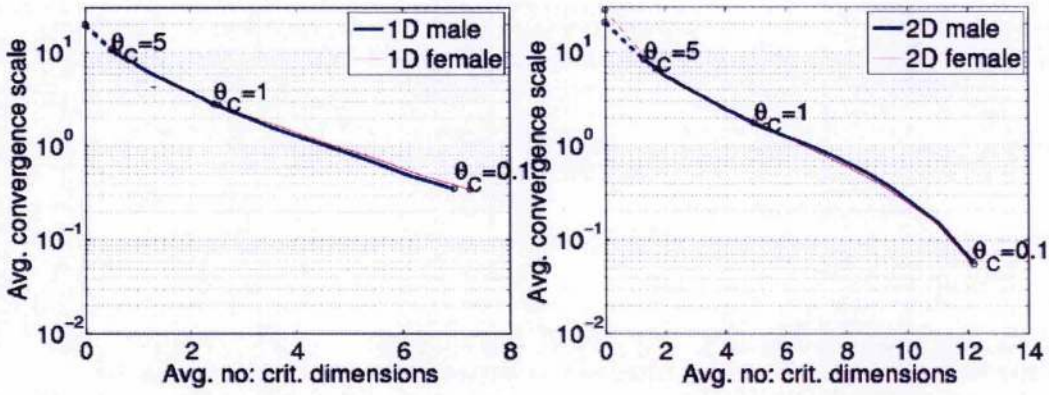



Figure 3.9: *Convergence of 1D (left) and 2D (right) models: the convergence scale* $\Upsilon^{\phi}_{conv}$ *between 14D model and phone pdfs averaged across all phones for* $\theta_C = \{0.1, 0.2, ..., 5\}$ *for both male (blue) and female (red) speakers.*

at a range of thresholds, $0.1 \leq \theta_C \leq 5$ for every phone $\phi \in \{\Phi\}$. Figure 3.10 shows the plots depicting the average evaluation scale values and the average number of critical articulatory coordinates at various threshold values for 1D and 2D models for male and female speakers. The divergence lies between infinity and zero (for perfectly matching distributions). The goodness of fit of the models was evaluated as a function of critical threshold $\theta_C$, which is applied to the 1D and 2D identification divergence.

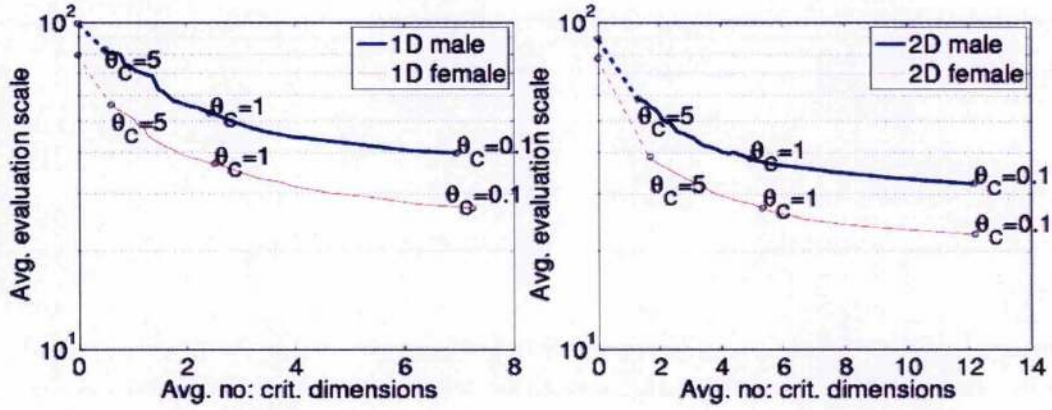



Figure 3.10: *Evaluation of 1D (left) and 2D (right) models: the evaluation scale* $\Upsilon^{\phi}_{eval}$ *between 14D model and phone pdfs averaged across all phones for* $\theta_C = \{0.1, 0.2, ..., 5\}$ *for male (blue) and female (red) speakers.*

The x-axis of the graph represents the average number of critical dimensions per phone obtained by averaging the number of dimensions across all the phones. Each point on the plot corresponds to a particular critical threshold value within the range $0.1 \leq \theta_C \leq 5$. Considering the 1D models first, the initial value of evaluation scale between the 1D models with diagonal covariances and the phone models with full phone covariances before applying the algorithm (at level $k = 0$) was found to be 99 for the male and 79

for the female speaker. The initial evaluation scale of the female speaker was smaller than that of the male speaker and the difference in the values between the speakers was persistent at all values of critical thresholds. It was found that as the threshold was lowered, the goodness of fit improved at the expense of increased critical dimensions. For example an improvement of 36% was achieved when the threshold was lowered from 5 to 1 for the male speaker data whereas the average critical dimensions per phone increased from 0.5 to 2.5. This trend was also observed in the female speaker data. Further lowering the threshold to 0.1 improved the goodness of fit by 17% on average between male and female speaker data but the number of critical dimensions per phone on average rose to 7 (half of the total available).

The fit of the 2D models to the phone models with full phone covariances is also shown in Fig. 3.10. With the inclusion of the correlations between the x and y dimensions, the initial divergence between the 2D models with 2D diagonal covariances and the phone models with full phone covariances was improved by 10% for the male speaker and 2% for the female speaker data when compared with the 1D initial evaluation scale values. In the 2D case, the fit of the models improved significantly as the average number of critical dimensions per phone increased from 1 to 5 for the male speaker and from 2 to 5 for the female speaker when $\theta_C$ was lowered from 5 to 1. Within this range, the improvement in the fit achieved was 34% for the male and 44% for the female speaker data. The fit improved by a mere 4% when the critical threshold was lowered from 1 to 0.1 but the average number critical dimensions per phone increased to 12 (x and y coordinates of 6 articulators) for both male and female speakers. As expected with the additional flexibility to describe correlations between x and y movements of each articulator, the fit of the 2D models to the phone pdfs with full covariances was better than the fit of the 1D models at all levels of threshold.

## 3.5 Conclusion

The methodology for articulatory constraint identification algorithm (ACIDA) was presented. The proposed algorithm identifies critical, dependent and redundant roles played by articulators for each phone. The algorithm also updates the model distributions of all articulators from the knowledge of identified constraints. Grand inter-articulatory correlations between the articulators used for identifying dependent articulatory roles were computed and the correlation patterns were identified. The articulatory coordinates could be separated into the lip and jaw group, the tongue group and the velum based on the strength of the correlations. The four stages in the algorithm, model initialisation, computation of identification divergence, C-step and D-step were explained. The working of the algorithm was illustrated using the grand, phone-specific and model distributions of a phone ([g]). The effect of critical threshold on the performance of the algorithm was also evaluated using convergence and evaluation scales. It was found that the convergence and the fit of the models improved as the critical threshold was lowered due to increased model complexity. After a certain threshold ($\theta_C = 1$) no improvement in the goodness of fit and convergence of the models to the phone distributions was achieved though the model complexity increased. In the following chapter (Chapter 4), lists of critical coordinates identified for each phone are

presented and analysed using phonological information.

# Chapter 4

# Identified critical coordinates and their phonetic analysis

## 4.1 Overview

Speech articulators are constrained to achieve target positions during the production of speech sounds. Identification of degrees of freedom of articulators during speech production plays an important role in modelling coarticulation effects. Unconstrained articulators are most susceptible to the coarticulation caused by the neighbouring constrained articulators (Mermelstein, 1973; Recasens and Pallarés, 1999). Speech articulators are correlated with one another due to the presence of physiological connections between them. Constrained articulators also influence the positions of other unconstrained articulators partially due to the presence of such inter-articulatory correlations. Different approaches were used for specification of constraints on articulators for each phone in the form of binary features (Henke, 1965; Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973), gesture scores (Browman and Goldstein, 1986; Saltzman and Munhall, 1989) and quantised configurations (Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000). Constraints in the form of binary features (Chomsky and Halle, 1968) are phonological and static in nature and are difficult to convert to commands for articulators. Hand coded gesture scores and quantised articulatory configurations are heuristic and incomplete descriptions. The proposed articulatory constraint algorithm (ACIDA) identifies constraints on the articulators in the form of critical, dependent and redundant roles for each phone from the EMA data. The movements of critical articulators towards targets are critical for production of speech sounds. The dependent articulators are partially constrained due to their relationship with the critical articulators and the remaining degrees of freedom are prone to context sensitive effects. The redundant articulators are unconstrained and are free to assume any uncontradicting position.

In this chapter, critical coordinates identified using the proposed ACIDA approach are presented. It is assumed in throughout this study that the available articulatory coordinates are treated as low dimensional representations of articulators. It has been shown that a few well-selected points can faithfully represent the full shape of the articula-

tors with reasonable accuracy (Badin and Serrurier, 2006; Qin et al., 2008). Expected critical coordinates are derived from the IPA chart. The identified critical coordinates are compared with the expected critical coordinates to analyse the performance of the models. Differences between results from the proposed approach and the knowledge driven approach are analysed. Model distributions estimated from the expected critical coordinates are also compared with the model distributions from the proposed approach for analysing the fit of the models to the actual phone distributions.

The rest of this chapter is organised as follows: Section 4.2 presents the motivation behind choosing the IPA chart for this analysis. Section 4.3 presents the derivation of expected critical coordinates for each phone in the database. Identified critical coordinates are presented in Section 4.4. Phonetic analysis of results is presented in Section 4.5. Conclusions from this analysis are presented in Section 4.6.

## 4.2 Why IPA?

The articulatory features such as discrete binary features (Chomsky and Halle, 1968) derived from phonological knowledge represent the place and manner of articulation of speech sounds. The IPA chart (International Phonetic Association, 2003) can be viewed as a short-cut representation depicting the intersection of different binary features. The IPA is a widely accepted representation and takes language specific phonetic variations into consideration. The IPA provides a good basis for comparison because it is an internationally agreed summary of the knowledge built up over many generations. Its main purpose involves the transcription of human speech by phoneticians, and is therefore tailored (i) to encapsulate meaningful distinctions in the context of language, (ii) for utterances produced by humans and (iii) observed by phoneticians. Increasingly, there is a need for phonetic descriptions for use in speech technologies that need (i) to model the characteristics of typical phones within a language (ii) to include the implicit effects found in human phoneme-to-phone realisation, such as coarticulation and (iii) to incorporate knowledge from other types of observations, such as X-ray and articulography data. The proposed ACIDA algorithm identifies the constraints on articulators from EMA data using statistical techniques. The identified critical coordinates are compared with the expected critical coordinates to analyse the performance of the model.

The following section (4.3) presents the list of expected critical dimensions derived from the IPA chart for the analysis of results.

## 4.3 Derivation of expected critical coordinates

A list of expected critical articulatory coordinates for all speech sounds was generated from the IPA chart for both 1D and 2D representations. The expected critical coordinates for consonants were obtained from the knowledge of the active articulator involved in the production of the consonant sounds. For vowels and diphthongs, the IPA vowel chart was used to estimate the critical coordinates. The expected critical coordinates are derived using the available articulatory coordinates. Recall from Chapter 2 that

the EMA data used in this work comprised measurements from x and y movements of upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD and velum V.

### 4.3.1 Consonants

The expected critical coordinate list derived from IPA for the 1D case, where the x and y dimensions of each articulator were treated as separate and independent coordinates, is shown in Table 4.1. For all bilabial sounds, [p], [b] and [m], $\text{UL}_y$ and $\text{LL}_y$ coordinates essential for lip closure were made critical. For labio-dental sounds, [f] and [v], $\text{LL}_y$ and $\text{LL}_x$ were chosen as critical. For inter-dental fricatives, [θ] and [ð], sibilants, [s], [z], lateral [l] and approximant [ɹ], the active articulator is the tongue tip. Hence, tongue tip x and y dimensions were marked as critical. For alveolar stops, [t], [d], [n], the $y$ movement of TT was chosen as critical. Tongue tip $x$ and $y$ dimensions were also chosen as critical for post-alveolar sibilants [ʃ], [ʒ], and affricates [ʧ], [ʤ]. For palatal sound [j], the tongue blade x and y dimensions were made critical. For velar stops, [k], [g] and [ŋ], the y movement of TD was made critical. For labio-velar sound, [w], $\text{UL}_x$, $\text{LL}_x$ and $\text{TD}_y$ were chosen as critical. No critical articulators were derived for glottal sound [h] from the available articulatory dimensions. For all nasal sounds, [m], [n] and [ŋ], in addition to the aforementioned critical coordinates, $\text{V}_x$ was also marked as critical.

The critical coordinates for the 2D case are shown in Table 4.1. In the 2D case, the x and y movements of each articulator are considered together for incorporating the spatial correlations. Therefore, for [p], [b] and [m], the UL and LL were made critical. For [f] and [v], the LL was made critical. For [t], [d], [n], [θ], [ð], [s], [z], [ʃ], [ʒ], [ʧ], [ʤ], [l] and [ɹ], the expected critical coordinate was TT. For palatal sound [j], tongue blade was critical. For [w], the UL,LL and TD were chosen as critical. For [k], [g] and [ŋ], the expected critical coordinate was [TD]. Velum was also chosen as critical for [m], [n] and [ŋ].

### 4.3.2 Vowels

Unlike consonants, it is difficult to describe vowels in terms of articulatory positions and vocal tract configurations. The vocal tract is unconstricted during the production of vowels making it difficult to describe the positions of articulators. Moreover, the vowels tend to overlap and merge into each other. Also, the effects of pronunciation, language and accent are greater on the realisation of vowels (Rosner and Pickering, 1994; Dirven and Verspoor, 2004).

The vowels present in the database are highlighted in the IPA vowel chart shown in Figure 4.1. The configurations of vowels shown on the vowel chart are part articulatory and part acoustic (as well as auditory) (Rosner and Pickering, 1994; Ladefoged, 2005), therefore it is difficult to describe articulatory positions for the vowels. In the articulatory domain, the tongue height is expected to play a key role in producing open, mid or close vowels. The extent of backness of the tongue is important for producing front, central or back vowels. The horizontal movement of lips is crucial for lip rounding. The

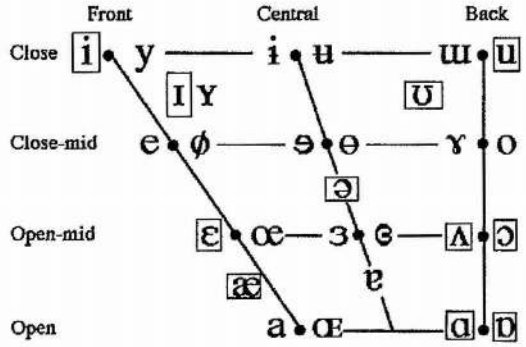



Figure 4.1: *Illustration of IPA vowel chart (International Phonetic Association, 2003), the vowels in the database are highlighted. Where ever vowels appear in pairs, the vowels to the left are unrounded and vowels to the right are rounded.*

articulatory flesh point marked as critical was assumed to reflect the backness of the tongue. Therefore, for all front vowels [æ, ɛ, ɪ, iː, i], TT was marked as critical. For mid vowels [ə, ɚ, ʌ], tongue blade was marked as critical and for back vowels [ɑ, ɒ, ɔ, ʊ, u], tongue dorsum was made critical. In the 1D case, tongue height was represented using the $y$ coordinate of the flesh points on the tongue. The $x$ movements of lower and upper lips were chosen as critical for rounded vowels. Table 4.2 shows the list of 1D and 2D expected critical articulators for vowels.

### 4.3.3 Diphthongs

Each diphthong was treated as a combination of two vowels and critical coordinates were specified for the initial and final vowels of each diphthong. Tables C.7 and C.10 show the list of critical coordinates derived for diphthongs. The first entry for each diphthong specifies the crucial coordinate for the initial vowel and the second entry for the final vowel. Similar to the approach taken for specifying critical coordinates for vowels, for all front vowels, TT was made critical, for all mid vowels TB was made critical and for back vowels, TD was made critical. In the 1D case, the $y$ movement was made critical since tongue height is discriminatory between open and close vowels. For all rounded vowels, the $x$ movements of upper and lower lips were made critical.

## 4.4 Identified critical coordinates

In this section, identified critical articulatory coordinates $\hat{C}^{\phi}$ for every phone $\phi \in \{\Phi\}$ are presented for the 1D and the 2D versions of the algorithm for both male and female speakers. Identification of critical articulatory coordinates is determined by the critical threshold value, $\theta_C$. Recall from Chapter 3 that an articulator is classified as critical only if the KL divergence between its grand and the model distributions (identification divergence) exceeds the critical threshold value $\theta_C$. It was found from the convergence

and evaluation scale analyses that lowering the critical threshold value increases the number of critical dimensions per phone. It is important to determine the threshold value at which the algorithm operates for a fair comparison of the identified and the expected results.

### 4.4.1 Selection of critical threshold $\theta_C$

The critical threshold for obtaining the results for this analysis was determined using the information of expected critical dimensions for vowels and consonants. The value of $\theta_C$ was adjusted such that the average number of identified critical dimensions were equal to the average number expected critical dimensions. Only the expected critical coordinates from vowels and consonants were considered for determining the threshold. For the 1D case, the critical threshold was set to 1.7 for both speakers. For the 2D case, the critical threshold was set to 2.3 for the male speaker and 2.0 for the female speaker. At this threshold level, the 1D version of the algorithm identified critical dimensions for 86% of phones for the male speaker and 88% for the female speaker. In the 2D case, the algorithm identified critical coordinates for 86% of phones for the male speaker and 88% for the female speaker.

Both 1D and 2D results for the male speaker for consonants are presented in Table 4.1, for vowels in Table 4.2 and for diphthongs in Table 4.3. The female speaker results are in Tables C.5, C.6 and C.7 for the 1D case and Tables C.8, C.9 and C.10 for the 2D case. Tables C.1 and C.2 show the dependent list $\hat{\boldsymbol{D}}^{\phi}$ and redundant list $\hat{\boldsymbol{\mathcal{R}}}^{\phi}$ for male and female speakers for the 1D case. The 2D case dependent and redundant lists are shown in Tables C.3 and C.4 for the male and the female speaker respectively.

## 4.5 Phonetic analysis of results

In this section, phonetic analysis of identified critical coordinates is presented. The analysis was done (i) using evaluation scale $\Upsilon_{eval}$, and (i) by comparison of expected critical coordinates and identified critical coordinates. Recall from Chapter 3 that evaluation scale is a measure of the goodness of fit of the model distributions with 1D variances or 2D covariance matrices to the respective actual phone specific distributions with full phone covariance (Figure 3.8). For analysis using evaluation scale, the 1D and 2D model distributions were updated in three ways, (a) using the expected critical coordinate information from IPA, (b)by applying D-step on models updated using the expected critical coordinate information and (c) using the identified critical coordinates. In the IPA based representations, only critical coordinates are derived for phones from the IPA chart. So the rest of the articulators other than critical were classified as non-critical articulators. So the dependent update step (D-step) was used to introduce the dependencies amongst the articulators using grand correlations and the performance of the resultant models was evaluated. Evaluation scale was averaged across all phones in all cases. Expected and identified critical coordinates were then compared to find if the identified critical coordinates are in agreement with the place and manner information of phones and to analyse any differences between identified and expected critical coordinates.

| Phones | 1D results | | 2D results | |
|---|---|---|---|---|
| | Expected | Identified | Expected | Identified |
| [p] | $UL_y$ $LL_y$ | $UL_y$ $LL_y$ | UL LL | UL LL |
| [b] | $UL_y$ $LL_y$ | $UL_y$ $LL_y$ | UL LL | UL LL |
| [m] | $UL_y$ $LL_y$ $V_x$ | $UL_y$ $LL_y$ $V_x$ | UL LL V | UL LL V |
| [t] | $TT_y$ | $TT_y$ | TT | TT |
| [d] | $TT_y$ | $TT_y$ | TT | TT |
| [n] | $TT_y$ $V_x$ | $TT_y$ $V_x$ | TT V | TT V |
| [k] | $TD_y$ | $TD_y$ | TD | TD |
| [g] | $TD_y$ | $TD_y$ | TD | TD |
| [ŋ] | $TD_y$ $V_x$ | $TD_y$ $V_x$ | TD V | TD V |
| [f] | $LL_y$ $LL_x$ | $LL_y$ $UL_y$ | LL | LL |
| [v] | $LL_y$ $LL_x$ | $LL_y$ $UL_y$ | LL | LL |
| [θ] | $TT_y$ $TT_x$ | $TT_x$ $TT_y$ $LL_y$ | TT | TT LL |
| [ð] | $TT_y$ $TT_x$ | $TT_x$ $TT_y$ | TT | TT |
| [s] | $TT_y$ $TT_x$ | $LI_y$ $TT_x$ $TT_y$ | TT | TT LI |
| [z] | $TT_y$ $TT_x$ | $LI_y$ $TT_x$ $TT_y$ | TT | TT LI |
| [ʃ] | $TT_y$ $TT_x$ | $TT_y$ $TB_x$ $LI_y$ $TD_y$ | TT | TT LI TD |
| [ʒ] | $TT_y$ $TT_x$ | $LI_y$ $TT_y$ $TD_y$ $TT_x$ $LL_y$ | TT | LI TT TD LL |
| [ʧ] | $TT_y$ $TT_x$ | $LI_y$ $TT_y$ $TB_x$ $TB_y$ | TT | TT LI |
| [ʤ] | $TT_y$ $TT_x$ | $TT_y$ $TB_y$ $TT_x$ $LI_y$ | TT | TT TB LI |
| [l] | $TT_y$ $TT_x$ | - | TT | - |
| [ɹ] | $TT_y$ $TT_x$ | $TB_x$ | TT | TT |
| [w] | $UL_x$ $LL_x$ $TD_y$ | $UL_y$ | UL LL TD | UL TT |
| [j] | $TB_y$ $TB_x$ | $TB_y$ | TB | TB |
| [h] | - | - | - | TT |

Table 4.1: *Expected and identified 1D and 2D critical coordinates for consonants for the male speaker.*

### 4.5.1 Comparison with IPA using evaluation scale

Model distributions updated using expected critical coordinate information from the IPA were used for computation of the evaluation scale. For each phone, the model distributions of all articulatory coordinates were initialised to the grand distributions. Then, only the distributions of critical coordinates derived from IPA were updated by setting their model distribution to the respective phone specific distributions (as in C-step). The evaluation scale was computed between the model and yje phone pdfs (with full phone covariance) and averaged across all phones as illustrated in section 3.4.1.

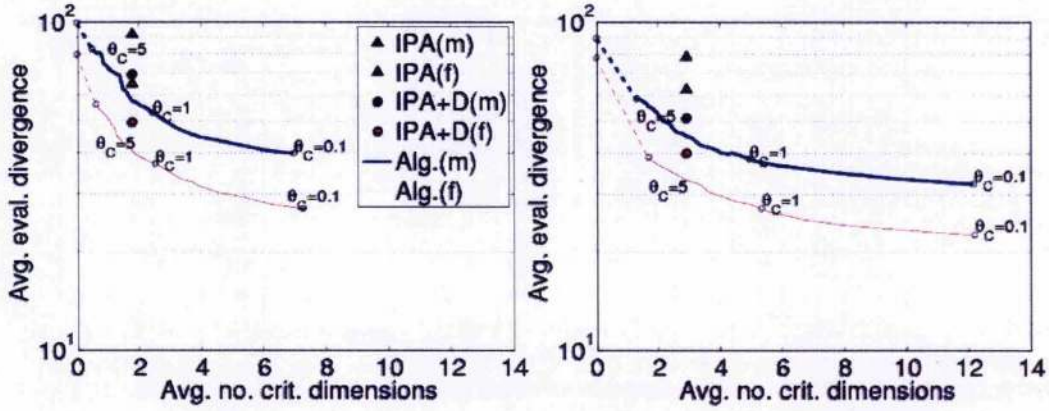



Figure 4.2: *Average evaluation scale computed using 14D phone pdfs and model pdfs for 1D (left) and 2D (right) cases for the male (m) and the female (f) speaker. Models are trained (a) using the expected critical coordinate information only (IPA), (b) by combining expected knowledge with the D-step (IPA+D) and (c) using the proposed algorithm (Alg.) at various values of critical threshold* $\theta_C = \{0.1, 0.2, ..., 5\}$.

Figure 4.2 shows the average evaluation scale computed using IPA-based distributions for male and female speakers. The x axis for both 1D and 2D versions represents the average number of critical dimensions per phone. In the 2D version, $x$ and $y$ dimensions of each flesh point were considered together, for example, TT in the 2D case represents both $\text{TT}_x$ and $\text{TT}_y$. Therefore, to achieve same scale on x axis of both plots in Figure 4.2, the average number of 2D critical dimensions was multiplied by 2.

From Figure 4.2, it can be seen that setting the model distributions of critical articulatory coordinates specified by IPA gave little benefit. The goodness of fit of the model distributions to the full phone distributions improved by including dependencies between the articulatory coordinates. For the 1D case, an improvement of 24% was obtained for both male and female speakers. For the 2D case, the improvement achieved was 35% for both speakers. Evaluation scale computed from the model distributions obtained from the 1D and 2D versions of the algorithm averaged across all phones is also shown in Figure 4.2. At the same number of average critical dimensions per phone, 1D algorithm improved the goodness of fit of the models by 37% over the distributions updated using IPA information alone. For the 2D case, an improvement of 44% was obtained. At the same level of complexity, the models estimated from the algorithm were found to be a better fit to the actual phone distributions. The following

| Phones | 1D results | | 2D results | |
|---|---|---|---|---|
| | Expected | Identified | Expected | Identified |
| [æ] (near open) | $TT_y$ | $LL_y$ | TT | LL |
| [ɛ] (open mid) | $TT_y$ | $LL_y$ | TT | LL |
| [ɪ] (close) | $TT_y$ | - | TT | - |
| [iː] (close) | $TT_y$ | $TD_y$ $LL_y$ $TT_x$ | TT | TB TT |
| [i] (close) | $TT_y$ | $TD_y$ | TT | TB |
| [ə] (mid) | $TB_y$ | - | TB | - |
| [ɚ] (rhotacized) | $TB_y$ | $LL_y$ | TB | - |
| [ʌ] (mid) | $TB_y$ | - | TB | - |
| [ɑ] (open) | $TD_y$ | $LL_y$ $TB_y$ $TB_x$ | TD | LL TB |
| [ɒ] (open rounded) | $TD_y$ $UL_x$ $LL_x$ | $TB_y$ | TD UL LL | TB |
| [ɔ] (mid rounded) | $TD_y$ $UL_x$ $LL_x$ | $LL_y$ $TB_x$ $TT_y$ | TD UL LL | TB LL |
| [ʊ] (near close rounded) | $TD_y$ $UL_x$ $LL_x$ | - | TD UL LL | - |
| [u] (close rounded) | $TD_y$ $UL_x$ $LL_x$ | $TD_y$ | TD UL LL | TD |

Table 4.2: *Expected and identified 1D and 2D critical coordinates for front, mid and back vowels for the male speaker.*

section presents an analysis of findings of the critical articulator identification of the algorithm. The expected and identified critical articulatory coordinates are compared for identification of the differences that resulted in the improvement in goodness of fit.

### 4.5.2 Comparison of identified and expected critical coordinates

#### Consonants

This comparison aims to find how similar the identified critical coordinates are to the expected critical coordinates for consonants. Since consonants have well defined active articulators and places of articulation, we expect to find the identified results to be in agreement with the expected results for most cases. In case of any differences, the information provided by the algorithm is expected to supplement the IPA results. The data-driven nature of the algorithm could also help in identification of speaker specific patterns.

Table 4.1 shows the expected and identified critical coordinates for the male speaker for both 1D and 2D cases (the female speaker results are in C.5 and C.8). In the 1D case, the algorithm identified critical coordinates for 92% of consonants for the male speaker and 96% of consonants for the female speaker. In the 2D case, the algorithm identified critical coordinates for 96% of consonants for both speakers. Thus, the proposed algorithm was able to capture the constraints on articulators during the production of most of the consonants. The identified critical coordinates were in general agreement with the expected critical coordinates though there were some notable differences.

| Phones | | 1D results | | 2D results | |
|---|---|---|---|---|---|
| | | Expected | Identified | Expected | Identified |
| [aɪ] | [a] (front close) | $TT_y$ | $LL_y$ $TD_y$ | TT | LL TD |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ | TT | LL |
| [eɪ] | [e] (front close) | $TT_y$ | $LL_y$ | TT | LL |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ $TT_x$ $TD_y$ | TT | LL TT |
| [ɛə] | [ɛ] (front mid) | $TT_y$ | $LL_y$ | TT | LL |
| | [ə] (centre mid) | $TB_y$ | $LL_y$ | TB | LI |
| [ɪə] | [ɪ] (front close) | $TT_y$ | $LL_y$ $TT_x$ | TT | LL TT |
| | [ə] (centre mid) | $TB_y$ | $LL_y$ | TB | LL |
| [ɔɪ] | [ɔ] (back mid rounded) | $TD_y$ $UL_x$ $LL_x$ | $TT_x$ $LL_y$ $TT_y$ | TD UL LL | LL TT |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ $TD_y$ $UL_x$ | TT | LL |
| [oʊ] | [o] (back mid rounded) | $TD_y$ $UL_x$ $LL_x$ | - | TD UL LL | - |
| | [ʊ] (back close rounded) | $TD_y$ $UL_x$ $LL_x$ | $UL_y$ $LL_y$ | TD UL LL | - |
| [aʊ] | [a] (front close) | $TT_y$ | $LL_y$ $TD_y$ $TT_y$ | TT | LL TB |
| | [ʊ] (back close rounded) | $TD_y$ $UL_x$ $LL_x$ | $TB_y$ $TB_x$ | TD UL LL | TB |

Table 4.3: *Expected and identified 1D and 2D critical coordinates for diphthongs for the male speaker.*

When the expected and the identified critical coordinates were compared, identical lists were found for 46% (1D) and 58% (2D) of consonants for the male speaker. For the female speaker, similar lists (i.e., same coordinates but in different order) were found for 25% (1D) of consonants of which 21% were identical. In the 2D case for the female speaker 50% of the consonants had similar lists of which 42% were identical. No critical coordinates were identified for [l] for both male and female speakers in both 1D and 2D cases. For glottal sound [h], no expected critical coordinates were specified from the available articulatory dimensions. The tongue tip was identified as critical for [h] for both speakers (except for the male speaker in the 1D case). Here the tongue tip lowered to allow the air flow for generating [h]. No critical coordinates were identified for [h] for the male speaker in the 1D case.

The velum x movement was identified as critical for all nasals [m, n, ŋ] for the male speaker in both 1D and 2D cases. For the female speaker, the identification divergence $J^{\phi}$ (computed between grand and model distributions) for V ($V_x$ for 1D), was higher than values given by other articulatory coordinates but fell below the chosen threshold level. Hence, the velum was not chosen as critical for any nasals for the female speaker in both 1D and 2D cases. Also, some inconsistencies in the velum measurements of the female speaker data (Richmond, 2001, 2009) could possibly affect the distributions of the velum coordinates.

For all sibilants [s, z, ʃ, ʒ] and affricates [ʧ,ʤ], the tongue tip TT and the jaw LI were identified as critical. For the 1D case, only the y coordinates of TT and LI were identified. The movement of TT is considered to be the primary articulation for these consonants. Since strong correlations exists between $TT_y$ and $LI_y$, identifying one of the coordinates as critical would update the distribution of the other in the D-step of the algorithm. Yet, in spite of the presence of this interdependency, the movement of LI was identified as the second most important critical articulation for these sounds. Here, the algorithm identified the articulatory coordinates responsible for the two important mechanisms required for a sibilant (Shadle, 1985), (i) forming a narrow channel which creates a fast moving jet of air, and (ii) locating the obstacle in the way of air flow created by positioning of the jaw.

More 1D critical coordinates (twice the average) were identified for phone [ʒ] for both male and female speakers. Phone [ʒ] had the least number of samples of all phones. The affect of small sample sizes on the estimation of the distribution statistics was compensated in the algorithm by adding the variance of mean to the distribution variance in the function *computeIdiv* (Figure 3.4). Yet, phone [ʒ] had more 1D and 2D critical dimensions for both speakers. The next highest number of critical dimensions were identified for post-alveolar sounds [ʃ, ʧ,ʤ] . Other points on the tongue were also identified other than the expected tongue tip. This shows that the shape of the tongue plays an important role in production of these sounds.

A few insertions and substitutions were made of correlated articulators. Some of the expected critical coordinates had strong correlations with identified critical coordinates. For example, for [ɹ], $TB_x$ was identified as critical for the male speaker which has strong correlations with the expected $TT_x$. Similarly for [w], the y movements of UL and LL were identified which are highly correlated with the expected x coordinates. Some differences between the expected and identified critical dimensions were inconsistent across the speakers. For [θ], the tongue blade also lowered to achieve the expected tip position and hence was chosen as critical for the female speaker. There was no significant change in the position of TB for the male speaker which was not identified as critical for [θ].

To summarize the findings, the identified critical coordinates were in general agreement with the active articulators of consonants. The 2D results gave a more clearer picture than the 1D results because of the inclusion of correlations between x and y movements of articulators. The lower incisor was identified as secondary articulation for sibilants. Few substitutions were made of correlated articulators, for e.g., $TT_x$ by $TB_x$ for [ɹ]. Some insertions were made by the proposed algorithm which supplemented the results. Some speaker specific differences were found which could be due to the speaking style variations. The algorithm also identified the details which are not explicit form the place of articulation descriptions from the IPA chart.

### Vowels

It is difficult to determine the shape of the oral cavity for vowel sounds. There are no clear boundaries between the vowels and lack of constriction makes it difficult to estimate which part of the tongue plays a critical role in the production of vowels (Rosner

and Pickering, 1994; Dirven and Verspoor, 2004). Targets for vowels are part articulatory and part acoustic (as well as auditory) (Rosner and Pickering, 1994; Ladefoged, 2005). In the articulatory space, the vowels are characterised using tongue height (from open to close), degree of backness (front to back) and lip rounding (rounded/unrounded) (Ladefoged, 1975). We expect the y dimension of the tongue to be identified as critical for high to low vowels, the flesh point coordinate on tongue would reflect the backness of the tongue, (TT for front, TB for mid and TD for back vowels respectively) and the lips to be identified as critical for rounded vowels.

Table 4.2 shows the expected and the identified 1D and 2D critical coordinates respectively for the male speaker (refer to Tables C.6 and C.9 for the female speaker results). At the same level of critical threshold, the algorithm identified critical coordinates for 77% of vowels for the female speaker in both 1D and 2D cases. For the male speaker, 69% of vowels in the 1D case and 62% in the 2D representations had critical coordinates.

For both speakers, no critical coordinates were identified for close-mid vowels which included the neutral vowel [ə], front vowel [ɪ] and back vowel [ʊ]. At the chosen critical threshold, none of the central vowels had 2D critical coordinates for the male speaker.

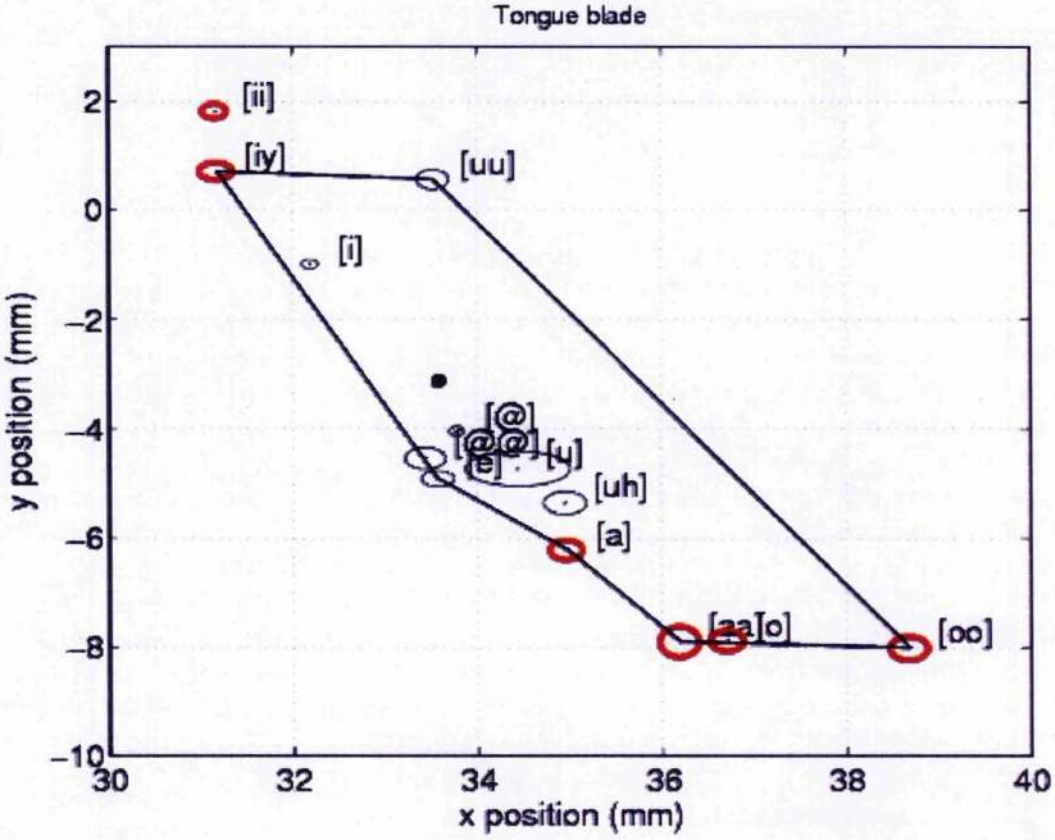



Figure 4.3: *Articulatory vowel quadrilateral for* TB *obtained by joining the model means of vowels closest to the primary cardinals in the database for the male speaker. The standard errors of model means are depicted using covariance ellipses for vowels for which* TB *is critical (thick red) and not critical (thin black).*

For open vowels, [æ, ɛ, ɑ], the jaw opening gesture was defined as critical using $\text{LL}_y$ (1D) and LL (2D) for the male speaker. The lower lip was identified as critical only for front vowels of all open vowels for the female speaker. Of all rounded vowels, the lower lip $\text{LL}_y$ (1D) and LL (2D) was identified as critical only for [ɔ]. The movements of upper lip and lower lip in the x direction were small and the rounding was perhaps not evident from the midsaggital data.

Amongst the three flesh points on the tongue, mostly TB followed by TD was chosen as the critical coordinate involved in shaping the vocal tract for production of vowel sounds. For both speakers, tongue tip was not identified as critical for any vowels in the 2D case. In the 1D case, $\text{TT}_x$ (which is highly correlated with $\text{TB}_x$ and $\text{TD}_x$) was identified as critical for front close vowels, [i] for the female and [iː] for the male speaker. The $y$ coordinate of TT was identified as critical only for [ɔ] for the male speaker and [ɚ] for the female speaker. The position of tongue (TT, TB and TD) for these two sounds was found to be lower than the neutral configuration. In the 1D case, the covariance between the x and y movements is ignored. The 1D identification divergence amongst the three points was found to be higher for TT in y direction when compared with TB and TD. In the 2D case, the TB was identified as critical.

For close front and back vowels, mostly $\text{TB}_y$ and $\text{TD}_y$ (1D) and TB (2D) were identified as critical. For open back vowels, $\text{TB}_x$ and $\text{TB}_y$ (1D) and TB (2D) were chosen as critical. The back vowel [ʊ] had no critical coordinates for both speakers in 1D and 2D cases.

The vowels were also analysed using the IPA vowel chart (Figure 4.1). The vowel chart is partly acoustic and partly articulatory in nature. There is a relationship between certain articulatory characteristics of the vowels and their acoustic features, the vowel height is inversely proportional to the frequency of the first formant (F1) and the backness is proportional to the difference between first two formants (F2-F1) (Ladefoged, 1975). The IPA vowel classification posits cardinal vowels that are based on two 'primary' positions (the theoretical extreme positions that can be achieved by the articulatory apparatus). Cardinal vowel (1), [i], is produced when the tongue is as high and forward as possible and the lips are spread. Cardinal vowel (5), [ɒ], is produced with tongue as low and back as possible. The rest of the cardinal vowels, [e, ɛ, a, u, o, ɔ], are placed with equal acoustic distance between the primary cardinals. The secondary cardinal vowels have opposite amount of lip rounding to that of primary cardinals. Other vowels can be specified relative to these cardinals and the IPA representation provides a number of such vowels which can be used for more accurate representations. The analysis of vowel quadrilaterals in the articulatory space aims at finding the relationship between the pseudo-articulatory vowel characteristics depicted in the vowel chart and the measurements from the flesh point data. Dang et al. (2009) also used articulatory data (X-ray) for analysing the structure of vowels.

Vowel quadrilaterals similar to the IPA vowel chart shown in Figure 4.1 were generated using the grand and phone means for front, mid and back vowels for both speakers. Vowel quadrilateral derived from measured articulatory data of TB (the most commonly identified critical coordinate) for the male speaker is shown in Figure 4.3 (vowel quadrilaterals of all points on the tongue for both speakers are shown in Figure C.1). Vowels are represented in MOCHA-TIMIT notation and corresponding IPA symbols can be found in Table A.2. The standard error of the mean was also plotted. The closest vow-

els to the primary cardinals (highlighted in Figure 4.1) were joined to obtain the vowel quadrilaterals. The shape of the vowel quadrilateral changed for each tongue coordinate and was different from the IPA vowel quadrilateral depicted in Figure 4.1. The vowels shown in Figure 4.3 are represented in purely articulatory domain and correspond to the real articulation data.

The shapes of the quadrilaterals for the male speaker were a closer resemblance to the IPA vowel quadrilateral than those from the female speaker. The position of the y dimension of tongue was found to be discriminatory between the vowels and was in agreement with the open/mid/close descriptions of the vowels. The x dimension looks less discriminatory since only the tongue blade coordinate (the most commonly identified critical coordinate for the vowels) is depicted for all front, mid and back vowels. The position of the tongue for front close vowels, [ɪ, iː, i], was high and forward as indicated in the vowel chart. The position of the tongue for mid vowels was central in terms of height and backness. The x position of the tongue for back vowels [u] ([uu] in Mocha-Timit) and [ɐ] ([o] in Mocha-Timit) was similar whereas the y position indicated the close and open nature of the vowels respectively for the male speaker. The magnitude of backness was greatest for vowel [ɔ] when compared with all other vowels for both speakers.

To summarise, the analysis of identified critical coordinates for the vowels showed that the tongue blade and dorsum play an important role in shaping the tongue for generation of vowels. The algorithm identified no critical coordinates for the neutral vowel [ə]. The lower lip was identified as critical for open vowels and one rounded vowel. The lip rounding was not clearly evident from the measurements of x movements of lips. The analysis of articulatory vowel quadrilaterals showed that the y dimension of the tongue measurement is in agreement with the tongue height feature of the vowels.

### Diphthongs

Diphthongs are the sounds that have changing vowel quality during the course of the syllable (Ladefoged, 2005). Hence each diphthong was treated as a sequence of initial and final vowels. It is assumed that the target is reached at the midpoint location of initial and final vowels. Therefore, the data sampled at one third and two third positions of the total diphthong duration was used for estimating distributions of initial and final vowels respectively. The initial vowel the diphthong is thought of as the most prominent than the final vowel Ladefoged (1975). The higher the prominence, the greater its influence on the quality of the realised diphthong. From this analysis, we expect to find the affect of the prominent part on the less prominent part of a diphthong, the relationship between the prominence and the identified critical coordinates and the differences between the identified critical coordinates of a vowel in its diphthong and its pure realisations.

The number of types of diphthongs present in the data set was 7. The initial and final vowels were both front vowels for [aɪ] and [eɪ]. The initial vowel was a front vowel and final vowel was a mid vowel for [ɛə] and [ɪə]. For the other three diphthongs [ɔɪ], [oʊ] and [aʊ], the transitions were from back rounded to front, back rounded to back rounded and front to back rounded respectively. Expected and identified critical coordinates for

diphthongs are shown in Table 4.3 for the male speaker. For the female speaker, the 1D and 2D results are presented in Tables C.7 and C.10 respectively. The analysis of identified critical dimensions for diphthongs was done by comparison of the transition from initial to final vowels of a diphthong and with their pure vowel (monophthong) realisations.

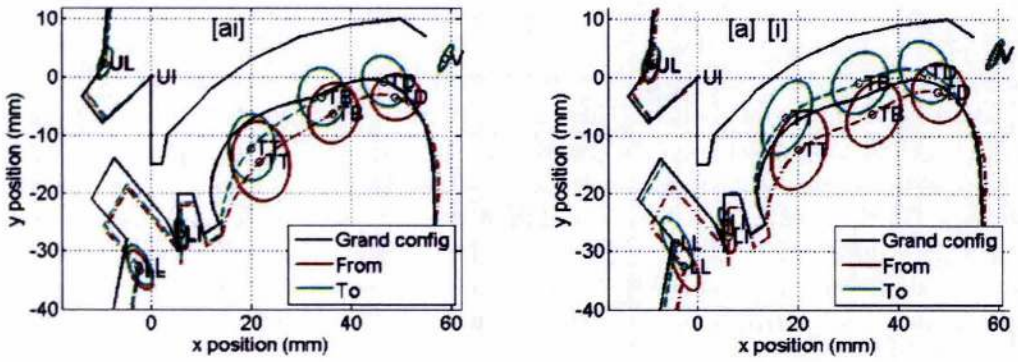



Figure 4.4: *Outlines of distant neighbour diphthong (left) and monophthong (pure vowel) (right) realisations for [aɪ] (MOCHA symbol is [ai]) obtained from the male speaker data. In the distant neighbour group, the initial and final vowels are characterised by larger contrast in the vowel quality. The grand (solid black), initial vowel (dashdot, red) and final vowel (dashed green) configurations are plotted.*

To analyse the nature of diphthongs, comparisons of articulatory configurations of initial and final vowels occurring as pure vowels and as diphthongs were made using midsaggital representations and covariance ellipses. Wherever an initial or a final vowel of a diphthong could not be found in the existing pure vowel set, it was compared with its closest neighbour in the vowel space, for example, [a] in [aɪ] with with [æ]. Figure 4.4 shows the diphthong and monophthong realisations for [aɪ] for the male speaker. It can be seen from Figure 4.4 that the magnitude of the transition from initial to final vowels was smaller than that when both monophthongs occur in a sequence (certainly some of the vowel reduction can be attributed to effort minimisation). The distributions of both initial and final vowels were tightly constrained and were different from their monphthong realisations. Also the articulatory configurations of initial and final vowels greatly influenced each other. As a result, the critical dimensions identified for initial and final vowels of a diphthong were different from the critical coordinates obtained from their monophthong realisations. These findings were also true for other diphthongs in the dataset.

Based on the contrast between the expected articulatory positions of initial and final vowels, the diphthongs were grouped in to two categories: (a) distant neighbour group consisting of [aɪ], [ɔɪ], [aʊ], [ɪə] and (b) close neighbour group consisting of [eɪ], [oʊ], [ɛə]. The diphthongs in the close neighbour group have a smaller contrast in vowel quality than the sounds in distant neighbour group (Ladefoged, 2005). On similar lines, contrast between the articulatory configurations of initial and final vowels was found to be very small for [ɛə] and [oʊ].

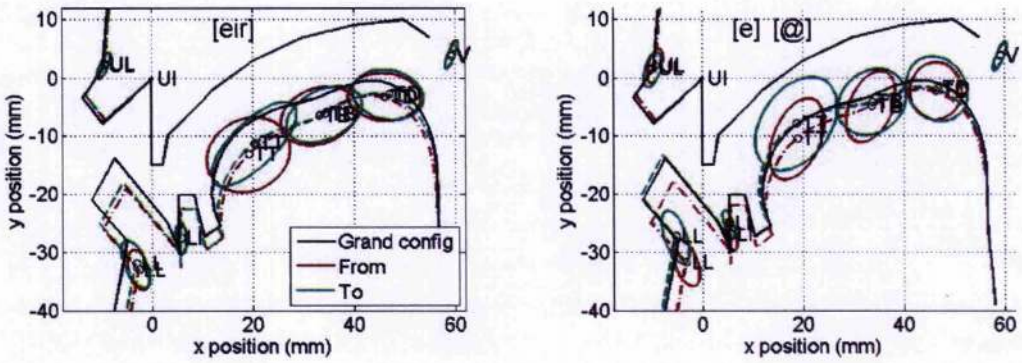



Figure 4.5: *Outlines of close neighbour diphthong (left) and corresponding monophthong (right) realisations for [ɛə] (MOCHA symbol is [eir]) obtained from the male speaker data. The grand (solid black), initial vowel (dashdot, red) and final vowel (dashed, green) configurations are plotted. The contrast in the vowel quality is small for initial and final vowels in the close neighbour group.*

For diphthongs in the distant neighbour group, crucial articulatory positions of the most prominent component influenced the configuration of the less prominent component. The analysis of the identified critical coordinates for the initial and final vowels of diphthongs showed that the initial vowel is most prominent part of a diphthong. Carry forward effect was caused by the prominent initial vowel on the final vowel for other diphthongs ([aɪ], [ɔɪ], [ɪə]). For a very few diphthongs, the final vowel was prominent. For [aʊ], the prominent component was the final vowel and it caused anticipatory effect on the configuration of the initial vowel. In all cases, the critical coordinates of the strongest vowel component of the diphthong were either similar to or strongly correlated with the critical coordinates identified from its pure vowel realisation. The critical coordinates of the weakest vowel component of a diphthong were different from those obtained in its pure realisation due to influence of the strongest component. For example, for [ɔɪ], identified critical coordinates for the prominent initial vowel of the diphthong were almost similar to the critical dimensions of monophthong [ɔ]. The tongue dorsum was identified as critical for the less prominent final vowel [ɪ] (which is a front vowel in its pure realisation) due to the affect of the initial vowel. Similarly, for [ɪə], the identified critical coordinates for [ɪ] were similar to its closest monophthong realisation [iː]. But for the final neutral vowel [ə], the position of the jaw continued to be critical due to the carry forward effect caused by the initial vowel. Similarly, the jaw was also identified as critical for the final vowel [ɪ] in [aɪ]. For [aʊ], the articulatory configuration of the initial vowel was affected by the final vowel due to the anticipatory effect. Therefore, the tongue blade and dorsum were identified as critical for the initial vowel [a].

Though the contrast between initial and final articulatory configurations was smaller for the diphthongs in the close neighbour group, the findings were similar to those in the distant neighbour group. The neutral vowel [ə] in [ɛə] had critical coordinates due to the carry forward effect induced by the initial vowel [ɛ]. Figure 4.5 shows the midsaggital plot for diphthong and corresponding monophthongs for [ɛə]. The identified

critical dimensions for the final vowel [ɪ] in [eɪ] were similar to the critical dimensions of monophthong [iː] and resulted in the identification of similar critical dimensions for its initial vowel as well. The phone [oʊ] had the smallest contrast of all and no critical dimensions were identified for the final vowels in most cases. Only the lips were identified as critical for the final vowel [ʊ] for the male speaker. The articulatory configurations of the initial and the final vowels for this diphthong did not match the configuration of any pure vowel.

To conclude, the articulatory configurations of the pure vowels and corresponding components of diphthongs were compared. The available diphthongs were categorized into close and distant neighbour groups based on the strength of vowel-quality contrast. It was found that the prominent vowel component of a diphthong shared similar critical coordinates with its corresponding pure vowel. The critical coordinates of the weaker components were influenced by the prominent components and hence differed from those obtained from similar pure vowels. The first vowel of the diphthong was found to be the stronger component for most diphthongs (60%).

## 4.6 Conclusion

In this chapter, identified critical coordinates for consonants, vowels and diphthongs were presented and analysed. The IPA standard was used to derive expected critical coordinates for 1D and 2D versions. Identified critical coordinates were analysed using the evaluation scale measure and by comparing them with the expected critical coordinates for both speakers. Evaluation scale analysis showed that, at the same level of complexity, the models updated using the proposed algorithm outperformed the IPA-based models. Introducing the dependent update step of the algorithm improved the performance of the IPA-based models. Comparison of expected and identified critical coordinates was done for consonants, vowels and diphthongs. The identified critical coordinates compared well with the expected critical coordinates for consonants. Some speaker specific patterns were also identified. The analysis of consonants was most straight-forward when compared with vowels and diphthongs. The identified critical coordinates were mostly in agreement with the active articulators involved in shaping the vocal tract for production of consonants. The tongue blade and the dorsum were identified as the articulatory coordinates playing critical role in shaping the tongue for production of vowels. No critical coordinates were identified for centralised vowels. The initial and final vowels of each diphthong were compared with their monophthong realisations. The analysis showed that the critical coordinates of the prominent vowel component of a diphthong influence the articulatory configuration of the less prominent component. The critical coordinates of the prominent component of a diphthong were similar to those identified from its pure vowel realisation.

The following chapter (Chapter 5) presents the evaluation of the search procedure employed by the proposed critical articulator identification algorithm.

# Chapter 5

# Analysis of the algorithm

The algorithm for identification of articulatory roles was evaluated using an exhaustive search procedure. Critical coordinates identified using the proposed algorithm were compared with the findings of the exhaustive search to determine (i) if the identified constraints vary based on search procedure, (ii) if the order of critical articulators makes any significant difference to the model convergence, and (iii) the best fitting models. The exhaustive search procedure was implemented for both 1D and 2D versions. The fit of the models estimated using the proposed algorithm and the exhaustive search to the actual phone distributions was also analysed.

The rest of this chapter is organised as follows: The procedure for identification of critical articulators using exhaustive search is presented in Section 5.1. Comparison of results from both approaches is presented in Section 5.2. The findings of the analysis are summarised in Section 5.3.

## 5.1 Evaluation by exhaustive search

The proposed articulatory constraint identification algorithm (ACIDA) follows a depth first search (DFS) approach for identification of critical articulatory coordinates. At any level $k$, the identification procedure is conditioned on the critical articulators identified upto and including level $k-1$. Therefore, the proposed algorithm can also be referred as the DFS procedure. On the contrary, the exhaustive search (ES) procedure searches all possible paths at each level to identify the best contender. The search space at the current level is not constrained by what has been identified as critical upto and including previous levels.

Figure 5.1 illustrates the difference between the functioning of the DFS and the ES methods. Only three nodes denoted by $\{a_1, a_2, a_3\}$ are considered in this particular example. The number of nodes (i.e., articulatory coordinates) considered for the implementation of the DFS (proposed algorithm) and the ES procedures on EMA data could be less than or equal to the dimensionality of the coordinate space, $a$ (14 for 1D and 7 for 2D). Both DFS and ES algorithms could progress upto level $k$ where $k \leq a$. Considering Figure 5.1, the number of valid paths for initial transition from

$k = 0$ to $k = 1$ in both DFS and ES cases is 3, nodes $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$. Assume that both search procedures identify node $\{\mathbf{a}_2\}$ as critical at level $k = 1$. In the DFS case, the identification of next critical coordinate at level $k = 2$ is conditioned on the best node at $k = 1$, i.e., the possible combinations upto and including level $k = 2$ could be $\{(\mathbf{a}_2, \mathbf{a}_1), (\mathbf{a}_2, \mathbf{a}_3)\}$. Here, the combination $(\mathbf{a}_2, \mathbf{a}_1)$ is chosen as critical.

In the ES case, all possible paths from $k = 1$ to $k = 2$, i.e., $\{(\mathbf{a}_1, \mathbf{a}_2), (\mathbf{a}_1, \mathbf{a}_3), (\mathbf{a}_2, \mathbf{a}_1), (\mathbf{a}_2, \mathbf{a}_3), (\mathbf{a}_3, \mathbf{a}_1), (\mathbf{a}_3, \mathbf{a}_2)\}$, are searched to determine the best combination $(\mathbf{a}_1, \mathbf{a}_3)$. Similarly, all possible combinations of nodes are searched to determine the best critical combination at level $k = 3$.

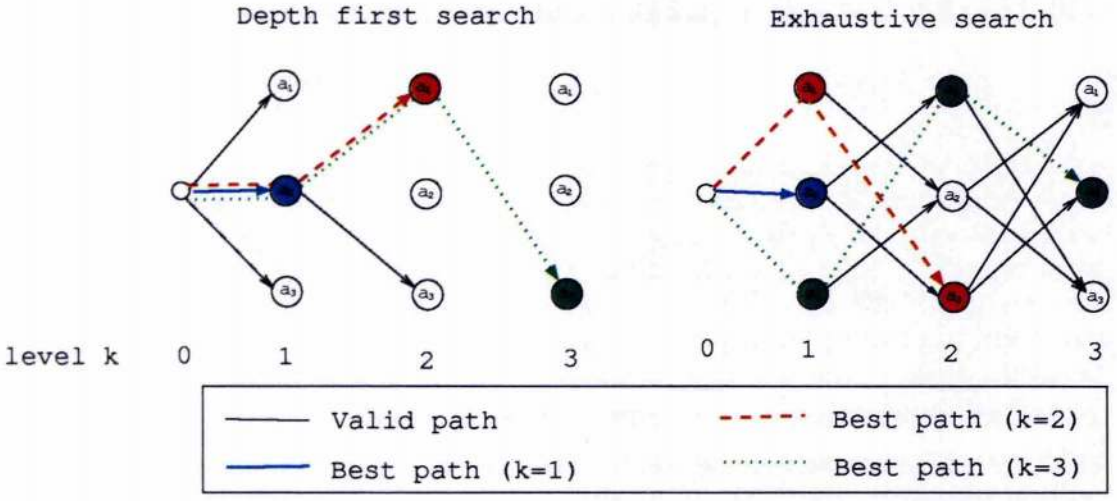



Figure 5.1: *Search trees for the proposed and the ES procedures depicting the path followed by both procedures in identifying articulatory roles for level* $k = 0, 1, 2, 3$.

Recall from Chapter 3 that a coordinate $j$ is selected as critical if its identification divergence $J_j$ is the maximum divergence at the level and is greater than the critical threshold value $\theta_C$. In the ES search, in addition to the critical threshold constraint, the best combination is identified using the minimax criterion (Coppin, 2004). According to minimax, the combination of articulatory coordinates which minimises the maximum identification divergence is chosen as critical. Figure 5.2 shows various stages in the ES method for identification of the critical articulatory coordinates,

- **Model initialisation**: This stage is similar to the model initialisation stage in the DFS method. The model means and variances of all articulatory coordinates are set to the respective grand means and variances.
- **C step**: In this stage, every combination of articulatory coordinates at each level is made critical. The model distributions are updated by setting them to the phone-specific distributions.
- **D-step**: The grand and the phone specific statistics of the rest of the articulators are collated. The dependent articulators are identified using inter-articulatory correlations and their distributions are updated.

- **Minimax**: The minimum of maximum identification divergence from each articulatory combination is identified and passed onto the next stage where it is compared with the critical threshold value.

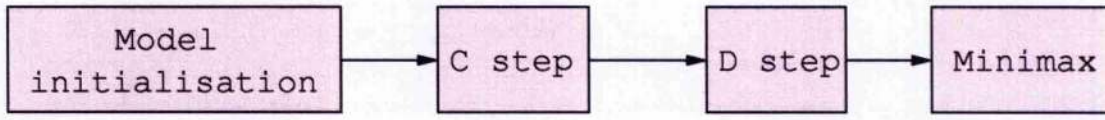



Figure 5.2: *Illustration of the flow of data for the ES approach for identification of critical articulatory roles.*

This process is repeated at each level until all possible combinations are searched exhaustively to find the best set of critical articulatory coordinates.

### 5.1.1 Algorithm

The pseudocode of the ES procedure for identifying critical articulators is presented in Fig. 5.3. The notation used for the ES algorithm is similar to that presented in Figure 3.4. The grand information is denoted by $\Gamma = \{\boldsymbol{M}, \boldsymbol{\Sigma}, N\}$, phone-specific statistics for each phone $\phi$ are denoted by $\Lambda^{\phi} = \{\boldsymbol{\mu}^{\phi}, \boldsymbol{\Sigma}^{\phi}, \nu^{\phi}\}$, the model statistics at each level $k$ are denoted by $\Delta^{\phi,k} = \{\boldsymbol{m}^{\phi,k}, \boldsymbol{S}^{\phi,k}, n^{\phi,k}\}$.

The number of combinations at each level $k$ is denoted by $P_k$. At level $k = 0$, the model distributions are initialised to the grand statistics and the accumulators for storing critical and dependent articulators are initialised. Initial identification divergence value $J_i^{\phi,0}$ is calculated for each articulator $i \in \{1..a\}$. Identification divergence is the 1D or 2D KL divergence between the model and the phone specific distributions. All paths leading to and including level $k-1$ are evaluated to identify the best set of critical coordinates at level $k$. Adding articulator $j \in \{1..a\}$ existing critical articulators from previous level $\boldsymbol{C}_p^{\phi,k-1}$ results in new critical articulator combination $\boldsymbol{C}_{\omega}^{\phi,k}$ where $\omega$ indexes that particular combination. The model information from each combination $p \in \{1..P_{k-1}\}$ at level $k-1$ is propagated to the level $k$

$$\boldsymbol{\Delta}_{\omega}^{\phi,k} = \boldsymbol{\Delta}_{p}^{\phi,k-1} \tag{5.1}$$

In the C-step, the model distributions represented by $\omega$ are updated as

$$\boldsymbol{m}_{\omega,j}^{\phi,k} \leftarrow \boldsymbol{\mu}_j^{\phi} \tag{5.2}$$

$$\boldsymbol{S}_{\omega,j}^{\phi,k} \leftarrow \boldsymbol{\Sigma}_j^{\phi} \tag{5.3}$$

The dependent articulators, $\boldsymbol{D}_{\omega}^{\phi,k}$, are identified and their model distributions are updated using grand and phone-specific distributions and correlations in function *upDateDep*. The identification divergence, $J_{\omega,i}^{\phi,k}$ at level $k$ for each articulator $i \in \{1..a\}$ for combination $\omega$ is calculated using function *computeIdiv*.

Of all combinations, $\omega \in \{1..P_k\}$, the critical articulator set, $\boldsymbol{C}_{\omega_b}^{\phi,k}$, yielding minimum of maximum identification divergence is selected. The minimax identification divergence,

$J^{\phi}_{max}$, is compared with the critical threshold value $\theta_C$. The best critical coordinates are stored and the algorithm progresses to the next level only if the divergence is greater than $\theta_C$.

The number of articulatory combinations and permutations searched by the ES algorithm increases at a factorial rate as the algorithm progresses from one level to the next. For example at level $k = 6$, the number of articulatory combinations considered for the search were over 2 million. The following section presents a comparison of the proposed and the ES procedures using evaluation scale and identified critical coordinates.

## 5.2 Comparison of the proposed and the ES methods

The performance of the DFS and the ES procedures for identifying critical articulators was analysed by comparison of

- evaluation scales
- identified critical articulators
- computational effort

Evaluation scale, $\Upsilon_{eval}$, is the 14D KL divergence between the 1D or 2D model distributions with scalar or matrix (co-) variances and the actual phone distributions with full 14D covariances (refer to Figure 3.8). The evaluation scale averaged across all phones was used to analyse the goodness of fit of the models from the DFS and the ES methods to the actual phone distributions. The two search procedures were also compared on the basis of economy of computational effort.

The proposed algorithm identifies critical coordinates conditioned on the information from the previous level, whereas the ES algorithm identifies articulatory roles independently of any previous critical articulator information.The list of critical articulatory coordinates identified using the DFS and the ES procedures were compared for consonants, vowels and diphthongs identify the effect of the search procedure on the identification of articulatory roles. The search procedures were compared at two values of critical threshold: (a) **IPA level of complexity**, where the average number of critical dimensions per phone equals to that of average number of expected dimensions per phone derived from IPA chart (b) **2×IPA level of complexity**, where the complexity of the models is twice that at IPA level. The ES search was implemented upto and including level $k = 6$ in both 1D and 2D cases.

### 5.2.1 Evaluation divergence

To calculate the evaluation scale, the model and phone-specific statistics were collated to form 14D mean and covariance vectors as explained in Chapter 3 (see Figure 3.8). The model distributions updated using the ES procedure and the proposed DFS procedure were computed at a range of critical threshold values ($0.1 \leq \theta_C \leq 5$) up to and including level $k = 6$.

**Derive statistics**
Global statistics $\mathbf{\Gamma} = \{\boldsymbol{M}, \boldsymbol{\Sigma}, N\}$, means $(a \times K)$, variances $(a \times K \times K)$ and sample size $(a \times 1)$
Grand correlation $\mathbf{R}^*$
Phone statistics $\mathbf{\Lambda}^\phi = \{\boldsymbol{\mu}^\phi, \boldsymbol{\Sigma}^\phi, \nu^\phi\}$, means $(a \times K)$, variances $(a \times K \times K)$ and sample size $(a \times 1)$
Phone correlation $\boldsymbol{R}^\phi$
Model information at level $k$: $\boldsymbol{\Delta}_p^{\phi,k}$ $\forall p \in \{1..P_k\}$
**Model initialisation**
level $k = 0$
No: combinations at level $k$, $P_k = 1$
FOR $p \in \{1..P_k\}$
    $\boldsymbol{m}_{p,i}^{\phi,k} = \boldsymbol{M}_i$, $\boldsymbol{S}_{p,i}^{\phi,k} = \boldsymbol{\Sigma}_i$, $n_{p,i}^{\phi,k} = N$, $\forall i \in \{1..a\}$
    Critical articulator list: $\boldsymbol{C}_p^{\phi,k} = \{\}$
    Dependent articulator list: $\boldsymbol{D}_p^{\phi,k} = \{\}$
    Compute identification divergence: $J_{p,i}^{\phi,k} = \text{computeIdiv}(\boldsymbol{\Delta}_{p,i}^{\phi,k}, \boldsymbol{\Lambda}_i^\phi)$, $\forall i \in \{1..a\}$
END FOR
Model convergence: *isConverged*=FALSE
**Exhaustive search**
WHILE $(k \leq a)$ AND (!*isConverged*)
    Go to next level: $k = k + 1$
    Initialise counter: $\omega = 0$
    Initialise maximum divergence value to a constant: $J_{\max}^\phi = MAX$
    FOR $p = \{1..P_{k-1}\}$
        **C-step**
        FOR $j = \{1..a\} - \{\boldsymbol{C}_p^{\phi,k-1}\}$
            Increment counter $\omega = \omega + 1$
            Replicate model $\boldsymbol{\Delta}_\omega^{\phi,k} = \boldsymbol{\Delta}_p^{\phi,k-1}$
            Add $j$ to critical articulator list: $\boldsymbol{C}_\omega^{\phi,k} \leftarrow\!\shortmid \{\boldsymbol{C}_p^{\phi,k-1}\} \cup \{j\}$
            Update model: $\boldsymbol{m}_{\omega,j}^{\phi,k} \leftarrow\!\shortmid \boldsymbol{\mu}_j^\phi$, $\boldsymbol{S}_{\omega,j}^{\phi,k} \leftarrow\!\shortmid \boldsymbol{\Sigma}_j^\phi$
            $n_{\omega,j}^{\phi,k} \leftarrow\!\shortmid \nu^\phi$
            **D-step**
            $\boldsymbol{\Delta}_\omega^{\phi,k}$ =updateDep$(\mathbf{\Gamma}, \boldsymbol{R}^*, \mathbf{\Lambda}^\phi, \boldsymbol{R}^\phi, \Theta, \boldsymbol{J}_p^{\phi,k-1}, \boldsymbol{D}_p^{\phi,k-1}, \boldsymbol{C}_\omega^{\phi,k})$
            $J_{\omega,i}^{\phi,k} = \text{computeIdiv}(\boldsymbol{\Delta}_{\omega,i}^{\phi,k}, \boldsymbol{\Lambda}_i^\phi)$, $\forall i \in \{1..a\}$
            **Minimax criterion**
            IF$(\max_{i \in \{1..a\}}(J_{\omega,i}^{\phi,k}) < J_{\max}^\phi)$
                Update maximum divergence: $J_{\max}^\phi = \max_{i \in \{1..a\}}(J_{\omega,i}^{\phi,k})$
                Store the index of the best combination: $\omega_b = \omega$
            END IF
        END FOR
    END FOR
    **Critical threshold check**
    IF $(J_{\max}^\phi > \theta_C)$
        Store the best combination at level $k$: $\hat{\boldsymbol{C}}^{\phi,k} \leftarrow\!\shortmid \boldsymbol{C}_{\omega_b}^{\phi,k}$
        Store the best model statistics at level $k$: $\hat{\boldsymbol{m}}^{\phi,k} \leftarrow\!\shortmid \boldsymbol{m}_{\omega_b}^{\phi,k}$, $\hat{\boldsymbol{S}}^{\phi,k} \leftarrow\!\shortmid \boldsymbol{S}_{\omega_b}^{\phi,k}$
        *isConverged* =TRUE
    END IF
    Store number of combinations at the current level: $P_k = \omega$
END WHILE

Figure 5.3: *Pseudocode for exhaustive search procedure for identifying critical articulators. For 1D or 2D versions, use scalar or vector means,* $\boldsymbol{M}$, $\boldsymbol{\mu}^\phi$ *and* $\boldsymbol{m}^{\phi,k}$ *and scalar or matrix (co-) variances* $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^\phi$ *and* $\boldsymbol{S}^{\phi,k}$.

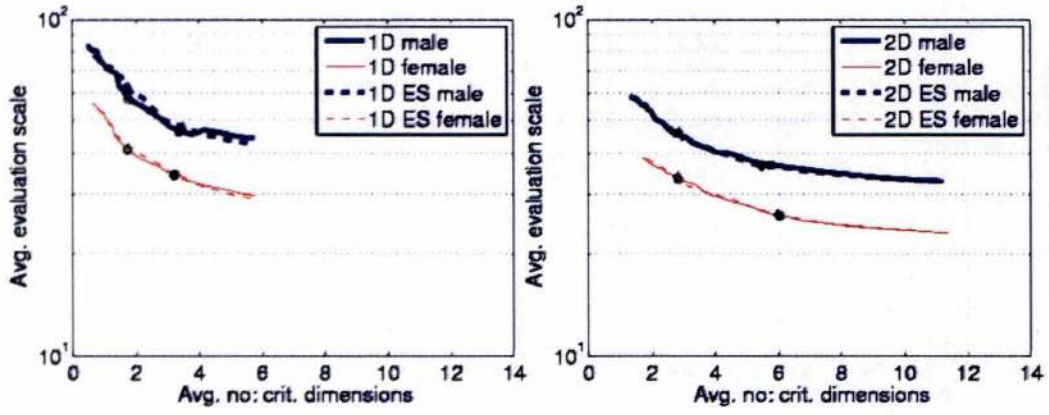



Figure 5.4: *Average evaluation scale computed from the DFS (solid) and the ES (dashed) methods for the male (blue, thick) and the female (red, thin) speakers. The IPA level of complexity and the 2×IPA level of complexity (filled) are indicated.*

Figure 5.4 shows the evaluation scale computed between model and phone pdfs as $0.1 \leq \theta_C \leq 5$, averaged across all phones for the DFS and the ES procedures for identifying critical articulators for 1D and 2D versions respectively. At all levels of critical threshold, the DFS and the ES procedures gave similar results. At the IPA level of complexity, the critical threshold was found to be 1.5 for both speakers in the 1D case for the ES procedure whereas setting $\theta_C$ to 1.7 yielded the IPA level of complexity for the DFS procedure. Here, the minimax identification divergence given by the ES procedure averaged across all phones was 8% smaller than that given by the proposed method when averaged across both speakers. However, this reduction in the divergence due to the application of minimax criterion did not improve the goodness of fit of the ES models over the DFS models. For the male speaker, the fit of the models updated using the DFS method was 9% better than that using the ES procedure. Similarly for the female speaker, the evaluation scale given by the DFS method was 2% better than that given by the ES method (though 8% reduction of maximum identification divergence was given by the ES method). In the 2D case to achieve the IPA level of complexity , the critical threshold, $\theta_C$ was set to 2.2 for the male and 1.9 for the female speaker for the ES procedure whereas for the DFS procedure $\theta_C$ was 2.3 for the male and 2.0 for the female speaker. For the male speaker, the average evaluation scales for both search procedures were similar where as for the female speaker, the DFS method performed better than the ES method by 2%.

For the ES method, setting $\theta_C$ to 0.5 for the male speaker and 0.6 for the female speaker yielded twice the number of average critical dimensions per phone (2×IPA level of complexity) in the 1D case. Here the percentage reduction in the maximum identification divergence obtained using the ES method over the DFS was 21% in the 1D case when averaged across both speakers. For the male speaker, the DFS method gave 2% improvement over the ES method whereas for the female speaker, the ES method was better than the DFS by 3%. Critical threshold was set to 0.7 for both speakers in the 2D case for the ES method. Similar values of evaluation scale were given by both 2D DFS and ES methods for both speakers at the 2×IPA level of complexity. Reducing

the threshold beyond the 2×IPA level of complexity increased the average number of critical dimensions per phone with a relatively small change in the evaluation scale (10% to 12% reduction).

### 5.2.2 Identified articulatory roles

The identified articulatory coordinates were compared at IPA and 2×IPA levels of complexity for both male and female speakers for both 1D and 2D cases. This analysis is presented for consonants, vowels and diphthongs. The male speaker results for both 1D and 2D cases at 2×IPA level of complexity are shown in Tables 5.1 and 5.2. Results for all speakers for 1D and 2D cases in more detail (i.e., with maximum identification divergence and evaluation scale values) are presented in Tables from C.15 to C.22.

The following terminology is used for the rest of this section: when the ES method identifies at least one additional critical dimension than the DFS method for a phone, the additional dimension is 'inserted' by the ES method. On the contrary, when the DFS method identifies at least one additional critical dimension than the ES method, the ES method 'deleted' the corresponding dimension. If the ES method identifies a critical coordinate correlated with that identified by the DFS method, it is referred as 'substitution'.

#### Consonants

At the IPA level of complexity, for the male speaker, similar sets of critical coordinates were obtained using both DFS and ES methods for 62.5% (1D) and 83% (2D) of consonants. Of consonants with similar critical coordinates, identical results were obtained for 80% (1D) and 65% (2D) of phones. For the female speaker at IPA threshold, similar sets of critical coordinates were identified by both search procedures for 79% (1D) and 92% (2D) of consonants, of which 58% (1D) and 64% (2D) were identical. The results from the 2D case at the IPA level of complexity were easier to interpret than those from the 1D case. There were relatively few differences between the results of ES and DFS methods for both speakers. When similar critical coordinates were identified by both search procedures, the order of the critical dimensions made *negligible* difference to the evaluation scale value.

In the 1D case, some speaker specific differences were found between the results of the DFS and the ES methods. However, a majority of cases included substitutions by strongly correlated critical dimensions. Such substitutions were found for 33% (1D) and 4% (2D) of consonants for the male speaker and 17% (1D) and 8% (2D) for the female speaker. For example, for [v], in the 1D case for the male speaker, the proposed algorithm identified $\text{UL}_y$ as critical where as the ES identified $\text{UL}_x$ dimension as critical. Substitutions by strongly correlated articulatory coordinates made negligible difference to the evaluation scale and thereby to the goodness of fit of the models. For phone [ʒ], which has the smallest sample size of all phones, only one critical coordinate was identified in common by both DFS and ES methods for the male speaker. Here the total number of critical coordinates for the DFS method were 5 and for the ES method

| Phone | Identified critical coordinates (1D) DFS | ES |
|---|---|---|
| [p] | ULy LLy Vx | ULy Vx LLy |
| [b] | ULy LLy | ULy Vx LLy |
| [m] | ULy LLy Vx | LLy ULy Vx |
| [t] | TTy TTx | TTy TBx |
| [d] | TTy TTx | TTy TBx |
| [n] | TTy Vx TBx | TTy TBx Vx |
| [k] | TDy TBy | TTy TDy |
| [g] | TDy TBy Vy | TTy TDy Vy |
| [ŋ] | TDy Vx LIx | TBy TDy Vx |
| [f] | LLy ULy LIy | LIx TBy LLy ULy |
| [v] | LLy ULy LIy LLx Vx | ULx ULy LLy LIy |
| [θ] | TTx TTy LLy LIy | TBx TBy LLy TTy |
| [ð] | TTx TTy TBy | LLy TBy TTy TBx |
| [s] | LIy TTx TTy LLy ULy | ULy LLy LIy TTx TTy |
| [z] | LIy TTx TTy LLy | LLy LIy TTy TTx |
| [ʃ] | TTy TBx LIy TDy LLy | ULy LIy TTy TDx TBy |
| [ʒ] | LIy TTy TDy TTx LLy Vy | TDx TDy TTy LLy LIy Vy |
| [ʧ] | LIy TTy TBx TBy Vx LLy | LLy TTy TBx Vx LIy TDy |
| [ʤ] | TTy TBy TTx LIy Vx TDy | TTy TBx Vx TBy LLy LIy |
| [l] | TBy TTy LIx TBx | LIx TTx TDy TTy |
| [ɹ] | TBx TTy | TTy TBx |
| [w] | ULy TTy LLy TTx TDy | ULy LLy TTx TTy TDy |
| [j] | TBy TTy | TTy TBy TDx |
| [h] | LLy TTx | LIy TBx |
| | | |
| [æ] | LLy TBy TBx | LLy TDy |
| [ɛ] | LLy | LLy |
| [ɪ] | TBy TTx | TBx TBy |
| [i:] | TDy LLy TTx TTy | TTx TTy LLy TDy |
| [i] | TDy TTx LLy TTy | LLx TTx TTy TDy |
| [ə] | | |
| [ɚ] | LLy LIy TTy TBx TDy | LIy TDy TTy TBx LLy |
| [ʌ] | LLy | LLy |
| [ɑ] | LLy TBy TBx | ULy LLy TBy TTx |
| [ɒ] | TBy TDx LLy | LLy TBy TDx |
| [ɔ] | LLy TBx TTy ULy ULx | TBx ULy LLy ULx TTy |
| [ʊ] | LLy ULx | ULx LLy |
| [u] | TDy ULy LLy | ULy LLy TDy |
| | | |
| [aɪ] | LLy TDy TBx TTy | LLy TBx TBy |
| [aɪ] | LLy TTx TTy Vy | TDx TTy LLy Vx |
| [eɪ] | LLy TDy TTx | LLy TBy TBx |
| [eɪ] | LLy TTx TDy | LLy TTx TDy |
| [ɛə] | LLy TBy | LLy TBy |
| [ɛə] | LLy TBy | ULx LLy TBy |
| [ɪə] | LLy TTx TBy TTy ULx | LLy TBy TTx TTy ULx |
| [ɪə] | LLy ULx TDy | ULx LLy TBy |
| [ɔɪ] | TTx LLy TTy ULx TDy ULy | LLy TDy TTy ULx TBx ULy |
| [ɔɪ] | LLy TDy ULx TTx TTy | LLy ULx Vy TTy TDy TDx |
| [oʊ] | LLy ULy | ULx LLy TBy |
| [oʊ] | ULy LLy TTx | ULy LLy TTx |
| [aʊ] | LLy TDy TTy TBx Vy | TDx TTy Vy TDy LLy |
| [aʊ] | TBy TBx TTy Vx | LIy TTx TBy Vx |

Table 5.1: *1D critical modes identified using the proposed depth-first search algorithm (DFS) and exhaustive search (ES) at 2×IPA critical threshold for the male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (bold, magenta), deleted (bold, blue) and inserted (bold, green) critical coordinates.*

| Phone | Identified critical coordinates (2D) DFS | ES |
|---|---|---|
| [p] | UL LL | LL UL |
| [b] | UL LL | UL LL |
| [m] | UL LL V TT | V UL TT LL |
| [t] | TT | TT |
| [d] | TT | TT |
| [n] | TT V | TT V |
| [k] | TD TB | TB TD |
| [g] | TD V TB | TD LI V |
| [ŋ] | TD V LI UL | V LL TD |
| [f] | LL UL TT | UL TT LL |
| [v] | LL UL LI | LL LI UL |
| [θ] | TT LL LI UL | TB LL UL TT |
| [ð] | TT TD LL | TT TB |
| [s] | TT LI LL | LI LL TT |
| [z] | TT LI LL | LI TT LL |
| [ʃ] | TT LI TD LL | UL TT LI TB |
| [ʒ] | LI TT TD LL V | TT LL V LI TD |
| [ʧ] | TT LI TB V LL | LI V TT LL TD |
| [ʤ] | TT TB LI V LL | TD LI TT V LL |
| [l] | TB TT LI | LI TB TT |
| [ɹ] | TT | TT |
| [w] | UL TT LL | LL TT UL |
| [j] | TB TT UL | TT TD |
| [h] | TT LL | TT LL |
| | | |
| [æ] | LL TB | LL TB |
| [ɛ] | LL TB | LL TB |
| [ɪ] | TT | TT |
| [i:] | TB TT LL | LL TT TD |
| [i] | TB TT | TB TT |
| [ə] | | |
| [ɚ] | LL TB LI TT | LL TB LI TT |
| [ʌ] | LL TB | TD LL |
| [ɑ] | LL TB | LL TB |
| [ɒ] | TB LL | LL TB |
| [ɔ] | TB LL UL TT LI | LI LL TB TD |
| [ʊ] | LL UL | UL LL |
| [u] | TD UL | UL TD |
| | | |
| [aɪ] | LL TD TT | TB LL |
| [aɪ] | LL TT | LL TT |
| [eɪ] | LL TT TD | TT LL TD |
| [eɪ] | LL TT TD | TB LL TT |
| [ɛə] | LL TD UL LI | LL UL LI TB |
| [ɛə] | LI UL TT TD | LI UL TD TT |
| [ɪə] | LL TT TB UL | TT LL TD |
| [ɪə] | LL TB UL V | V TB UL LL |
| [ɔɪ] | LL TT UL TD | TT TD LL UL |
| [ɔɪ] | LL TD UL TT | LL UL TT TD |
| [oʊ] | TT LL | LL TT |
| [oʊ] | UL LL TT | UL LL TT |
| [aʊ] | LL TB TD UL | TD LL TT |
| [aʊ] | TB TT V | V TD TT |

Table 5.2: *2D critical modes identified using the proposed depth-first search algorithm (DFS) and exhaustive search (ES) at 2×IPA critical threshold for the male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (bold, magenta), deleted (bold, blue) and inserted (bold, green) critical coordinates.*

were 3. The fit of the DFS models was greater than that of the ES models. However, for the female speaker, similar sets of critical coordinates were identified for [ʒ]. The ES method identified one critical dimension for [l] where as none were identified for the DFS method for the male speaker. For the female speaker, both search procedures identified no critical dimensions for [l]. Velum was identified as critical for nasal sound [n] for the female speaker when the ES method was used, but was not identified as critical for any nasal consonant when the DFS method was used. One critical dimension each was inserted by the ES method for [l] for the male speaker, and [n] and [ʧ] for the female speaker.

At 2×IPA threshold level, the number of substitutions were much higher, 21 critical dimensions from 15 consonants (male) and 12 critical dimensions from 8 consonants (female) were substituted by correlated dimensions in the 1D case. Substitutions by correlated critical dimensions made negligible difference to the evaluation scales and hence to the goodness of fit of ES and DFS models. One extra critical dimension was identified for 4 consonants for the male speaker (3 consonants for the female) by the ES method when compared with the DFS method and here the evaluation scale computed from the ES models was smaller than that from the DFS models.

The pattern of results at the 2×IPA level of complexity for the 2D case were similar to those at the IPA level of complexity. Similar sets of critical coordinates were identified for most of the consonants (67% for the male and 87.5% for the female). Substitutions by correlated articulations were made by the ES method for 8 consonants for the male and 2 consonants for the female speaker. In all these cases, there were negligible differences between the evaluation scales generated by ES and DFS methods.

### Vowels

For vowels at IPA threshold, ES and DFS methods identified similar critical dimensions for 53% (1D) and 77% (2D) for the male speaker. Substitutions by correlated critical dimensions were made by the ES method for 46% (1D) and 15% (2D) of vowels. Both ES and DFS methods gave similar evaluation scale values when similar and correlated sets of critical coordinates were identified. For [ɑ] in the 1D case, the ES method identified $\text{LI}_y$ and $\text{TD}_y$ as critical which are correlated with $\text{LL}_y$ and $\text{TB}_y$ respectively. Here, the DFS models were a better fit (33% improvement) to the phone models than the ES models. In the 2D case, the DFS method identified no critical dimensions for [ɚ] at IPA threshold, whereas TT was chosen as critical by the ES method. Here, the improvement achieved by the ES models over the DFS models was 36.5%.

For the female speaker at IPA threshold, similar sets of critical coordinates were identified for both speakers for 61.5% (1D) and 69% (2D) of vowels. Substitutions by correlated critical dimensions were made by the ES method for a few phones (31% for 1D, 15% for 2D). The difference between evaluation scales generated by the ES and the DFS methods were negligible when similar and correlated critical coordinates were identified by both search procedures. The ES method identified one extra critical dimension when compared with the DFS method for [ʊ] in the 1D case. In the 2D case, one extra dimension was identified by the DFS method over the ES method for [iː, i].

At the 2×IPA level of complexity for the male speaker, similar sets of critical coordinates were identified for most of the vowels (69% for 1D and 77% for 2D). For the female speaker, similar sets of critical coordinates were found for 38% (1D) and 61.5% (2D) of phones. In the 1D case for the female speaker, the ES method identified more critical coordinates than the DFS method for 5 vowels [æ, ɪ, i, ɒ, ʊ]. For all these vowels, the ES method gave a small improvement (8% on average) over the DFS method.

### Diphthongs

Each diphthong was treated as a combination of initial and final vowels; critical coordinates were derived independently for initial and final vowels of a diphthong. At IPA threshold, for the male speaker, both ES and DFS methods identified similar sets of critical coordinates for initial and final vowels of [aɪ] and [eɪ] in the 1D case. Similar critical coordinates were identified for initial vowel of [ɪə] and final vowels of [ɔɪ] and [oʊ]. Here both ES and DFS methods generated identical evaluation scale values. The ES method substituted $\text{LI}_y$ for $\text{LL}_y$ for both initial and final vowels of [ɛə], $\text{LL}_x$ for $\text{LL}_y$ for final vowel of [ɪə] and initial vowel of [ɔɪ]. The ES method also substituted tongue blade for dorsum and tip for the initial and final vowels of [aʊ]. For all these phones, the DFS models performed better than the ES models. In the 2D case, for the male speaker, ES and DFS methods gave identical results for initial vowels of [eɪ, ɛə, ɪə, aʊ, oʊ], final vowel of [aɪand both initial and final vowels of [ɔɪ]. For the remaining phones, the ES method made substitutions by correlated dimensions for [aɪ, eɪ, aʊ], insertions for [ɛə, ɪə] and deletions for [oʊ]. Differences in evaluation scales were negligible when substitutions were made. The ES models performed well when insertions were made and worse in case of deletions.

At the IPA level of complexity, for the female speaker, no critical dimensions were identified for both initial and final vowels of [oʊ] by the ES and the DFS methods. Similar critical coordinates were identified for final vowels of [aɪ, ɪə, ɪɔ, aʊ] and initial vowel of [eɪ]. Substitutions by correlated critical dimensions were made by the ES method for the rest of the phones in this set. The DFS method identified one additional critical dimensions for [aɪ, ɔɪ] (initial vowels) and [eɪ] (final vowel). In the 2D case, for the female speaker, both ES and DFS methods identified no critical dimensions for [oʊ]. Similar sets of critical coordinates were identified for final vowels of [aɪ, ɛə, ɪə, aʊ] and initial vowel of [ɔɪ]. Substitutions by correlated critical dimensions were made by the ES method for the remaining phones. The ES method inserted one additional critical dimension LL for final vowel of [ɔɪ] and deleted one critical dimension each from initial vowels of [aɪ, ɪə].

At the 2×IPA level, in the 1D case, similar sets of critical coordinates were identified for fewer phones (4 for the male and 5 for the female speaker) than at the IPA level of complexity. For the remaining phones, substitutions by correlated critical dimensions were made by the ES method when compared with the DFS critical coordinate lists. The ES method identified more critical dimensions than DFS method for 3 phones for the male speaker and 5 phones for the female speaker respectively. The goodness of fit of the ES models was better than the DFS models for the above phones (10% improvement for the female speaker and 5% for the male speaker). For the final vowel

in [ɔɪ], the DFS method gave best performance (12%) with 5 critical dimensions over the ES method with 6 critical dimensions.

The pattern of 2D results at the 2×IPA level of complexity was similar to that at 1D case. Both search procedures identified similar sets of critical coordinates for most of the phones (8 for the male and 10 for the female speaker). The ES search method identified critical coordinates that were correlated with the DFS critical coordinates for 2 phones for the male speaker and 5 phones for the female speakers. The number of critical coordinates identified by the DFS method were more than that identified using the ES method for 3 phones for the male speaker and 4 phones for the female speaker. For these phones, the fit of the DFS models was better than that of the ES models.

### 5.2.3 Computational load

The ES procedure was found to be computationally very expensive when compared with the proposed algorithm. It took 277.8 hours (1D) and 1.6 hours (2D) to run exhaustive search for one speaker upto level 6 when implemented in Matlab v7.5.0 on a machine with a 3.3GHz processor with 32GB RAM. The DFS search took less than 1s in both cases. Hence, any slight benefit from the ES procedure was out weighed by the computational load. Thus the DFS is an efficient and effective algorithm.

### 5.2.4 Summary

When DFS and ES results were compared at the IPA level of complexity, similar lists of critical coordinates were identified for the majority of consonants, vowels and diphthongs for both speakers. The analysis showed that both search procedures identified similar critical dimensions for more consonants, which have well defined places of articulation, than vowels and diphthongs. A few differences between the critical coordinates of DFS and ES methods were found due to substitutions by correlated critical dimensions, insertions and deletions. Substitution by correlated critical dimensions made negligible difference to the evaluation scale. The pattern of results at the 2×IPA level of complexity for 2D case was similar to that at the IPA level. In the 1D case, at 2×IPA, the number of substitutions, insertions and deletions were greater than the similarities, all the more for consonants and diphthongs. The number of similarities between ES and DFS results for the female speaker were slightly higher than those for the male speaker. Also more similar critical coordinates were identified in the 2D case when compared with the 1D case at both levels of critical threshold. The order of the critical dimensions made no difference to the evaluation scale and therefore to the goodness of the fit of the models to the full phone distributions. Insertions by the ES method reduced the evaluation scale and deletions increased the evaluation scale. A few cases of insertions and deletions were found mostly when complexity was increased at the 2×IPA level of complexity than the IPA level.

## 5.3 Conclusion

In this chapter, the proposed algorithm for identification of critical dimensions was evaluated using an exhaustive search procedure. The minimax criterion was used to select the best set of critical articulators in the ES method at each level. The algorithm for the ES method was presented. The performance of the proposed and the ES procedures was evaluated using evaluation scale measure and by comparison of identified critical coordinates at two levels of thresholds. The performances of the proposed and the ES methods were very similar with small differences in performance (under 5% on average) when compared using evaluation scale. Comparison of critical dimensions identified by the search procedures showed that similar results were generated by both search procedures for a significant number of the phones, 75% (IPA) and 70% (2×IPA) in the 2D case, 61% (IPA) and 42% (2×IPA) in the 1D case when averaged across consonants, vowels and diphthongs. The order of the critical dimensions made no difference to the fit of the models. The ES procedure was found to be computationally very expensive. The proposed algorithm performed as well as the ES method and has the added advantage of faster execution times.

The following chapter presents the application of the articulatory role information and model distributions obtained from the proposed ACIDA algorithm in articulatory modelling.

# Chapter 6

# Analysis of articulatory representations

Finding suitable *feature* (not discrete features described in Chapter 2, but feature vectors derived by transforming the measured articulatory data) representations of the data is an important task in most speech related applications. Most of the approaches aim at providing compact representations of the available data while maximising the information content. In this study, linear orthogonal transforms are used for generating various articulatory feature representations. Orthogonal transforms are desirable since they tend to extract underlying non-overlapping components in the data. In this chapter, different feature spaces are derived from the measured articulatory data and analysed in terms of

- optimisation criterion
- power of interpretation
- informational efficiency
- compactness
- recognition performance

The optimisation criterion plays an important role in choosing suitable transforms for generating articulatory feature spaces, for e.g., transformation such as principle components analysis (PCA) rotates the data in the direction of maximum variability. The power of interpretation determines how interpretable are the feature representations generated from the measured EMA data, for e.g., could the modes resulting from the transform be related to the underlying independent components or the muscle groups controlling the speech articulators? The proposed ACIDA algorithm is used on different feature spaces to determine the critical modes for each phone and their corresponding movements are determined using the mode shapes. The model distributions for each phone are trained simultaneously using the algorithm in each feature space. The information contained by the modes in each articulatory feature space is then analysed using corresponding model distributions to determine the most efficient representation.

Most linear orthogonal transforms concentrate most of the information in the first few modes. The last few modes represent a small fraction of information and are noisy. Compact representations are derived by discarding the noisy modes which contain little information. The proposed algorithm identifies the constraints in the data and models the articulatory distributions in a compact way. The compactness of each feature representation when used with the proposed algorithm is also analysed along with the information contained in each mode. The performance of the proposed articulatory feature representations on a simple speech recognition task is also analysed.

Organisation of this chapter is as follows: Section 6.1 introduces the linear orthogonal transforms used in this study and derivation of various articulatory feature representations using the transforms. The interpretation power of various feature representations used in this study is analysed in Section 6.2. Identification of critical gestures from different feature spaces is presented in Section 6.3. The informational efficiency of each feature representation is analysed in Section 6.4. The compactness of each feature representation is analysed in Section 6.5 and recognition performance in Section 6.6.

## 6.1 Articulatory feature representations

Linear orthogonal transformations have been successfully used to derive compact yet informationally rich representations of the data. The work presented in this chapter uses two such representations for generating articulatory features namely, (i) Principal Components Analysis (PCA) and (ii) Linear Discriminant Analysis (LDA).

PCA (Jackson, 1991) is one of the commonly used transformations for feature reduction. PCA removes correlations between variables in the data and transforms it in the direction of maximum variability. The first few modes of PCA account for most of the variance found in the data. PCA is also used as a compression technique since compact representation could be achieved by eliminating the last few noisy modes. PCA has also been used on measured articulatory data for feature extraction (Wrench, 2000, 2001; Uraga and Hain, 2006). LDA is one of the most commonly used techniques in pattern recognition (McLachlan, 2004). LDA optimises separability between the classes in the data and is used for feature extraction and dimensionality reduction. Articulatory feature vectors were derived from measured EMA data using linear discriminant analysis by Wrench (2001).

Various factor analysis and PCA based approaches were also used for extracting underlying independent components of the articulatory system by Maeda (1990) and Badin et al. (2002). In the present approach, the knowledge of independence between the components is derived from grand inter-articulatory correlations. Independent components analysis (ICA) (Hyvärinen and Oja, 2000.), which is an extension of PCA, uses higher order statistics to find the underlying statistically independent sources from the data. However, in the present work, only the absence of correlations is used to establish independence between the articulatory coordinates, and linear orthogonal transforms are employed to obtain uncorrelated and informationally efficient representations. The proposed work could be extended in the future using techniques such as linear components analysis (Kirirani et al., 1977; Maeda, 1990; Badin et al., 2002) and independent components analysis (Hyvärinen and Oja, 2000.).

In the present study, the articulatory coordinates were divided into independent groups using the knowledge of correlations such that the intra-group correlations are maximised and the inter-group correlations are reduced to zero. Different kinds of PCA and LDA based feature spaces were derived from different combinations of articulatory groups.The following section, Section 6.1.1 presents the approach for generating different feature representations from raw articulatory data.

### 6.1.1 Generation of articulatory feature sets

Samples taken from midpoint positions for vowels and consonants, from third and two-third locations for diphthongs were centred and used for estimating the transformations. The articulatory feature vector resulting from the application of a single PCA transformation on all articulatory coordinates is denoted by PC1. This is the most commonly used method for deriving PCA based feature vectors (Wrench, 2000).

Articulatory coordinates were pooled into independent groups based on the strength of grand correlations between them as shown in Table 6.1. Along with PC1, four other types of PCA based feature sets denoted by PC3, PC4, PC5 and PC7 were derived from the articulatory groups. For PC3, the articulatory coordinates were pooled into three independent groups, the lip and jaw group UL+LL+LI, the tongue group TT+TB+TD and the velum V. Each of these three groups were transformed individually using PCA and the resulting modes were combined and sequentially indexed. The PC3 representation has 14 modes of which modes 1 to 6 are estimated from UL+LL+LI group, 7 to 12 are from TT+TB+TD group and 13 to 14 are from V group. For PC4, the upper lip was separated from the lower lip and jaw group, while the tongue and velum groups remained similar to those in PC3. The x and y coordinates of tongue coordinates were separated into x only and y only groups in PC5. The PC7 feature space was obtained by considering correlations between x and y coordinates of each flesh point while the rest of the correlations were ignored. All PCA based feature spaces were 14 dimensional. PCA based transformation matrices were obtained by performing eigenvalue decomposition on the correlation matrix of each independent articulatory group.

Linear discriminant analysis based feature sets were derived in a similar way from the raw articulatory coordinate space. The number of phone classes considered for performing LDA was 51. The data sampled at midpoints of 24 consonants and 13 vowels, at 1/3rd location of 7 diphthongs and 2/3rd location of 7 diphthongs was centred and whitened. Articulatory coordinates were pooled in to 5 groups as shown in Table 6.1 and LDA based feature vectors denoted by LD1, LD3, LD4, LD5 and LD7 are extracted from the measured articulatory data. LDA based transformation matrices were obtained by performing eigenvalue decomposition on the ratio of within class variance to the between class variance of each articulatory group.

| Feature sets | Articulatory groups (mode indices) | | | | | | |
|---|---|---|---|---|---|---|---|
| PC1/LD1 | All: UL+LL+LI+TT+TB+TD+V (1 to 14) | | | | | | |
| PC3/LD3 | Lip&Jaw: UL+LL+LI (1 to 6) | | | Tongue: TT+TB+TD (7 to 12) | | | V (13 to 14) |
| PC4/LD4 | UL (1 to 2) | LL+LI (3 to 6) | | Tongue: TT+TB+TD (7 to 12) | | | V (13 to 14) |
| PC5/LD5 | UL (1 to 2) | LL+LI (3 to 6) | | $T_x$: $TT_x$+$TB_x$+$TD_x$ (7 to 9) | $T_y$: $TT_y$+$TB_y$+$TD_y$ (10 to 12) | | V (13 to 14) |
| PC7/LD7 | UL (1 to 2) | LL (3 to 4) | LI (5 to 6) | TT (7 to 8) | TB (9 to 10) | TD (11 to 12) | V (13 to 14) |

Table 6.1: *Articulatory groups and corresponding PCA and LDA based feature sets. Each group was transformed using a separate PCA (LDA) and the resulting modes were indexed sequentially. The number of modes in every feature set is 14.*

## 6.2 Interpretation power of different PCA and LDA based transformations

The mode shapes of the transformations indicate the directions in which the data is rotated according to the optimisation criterion. The directions of maximum variability are indicated by the mode shapes of PCA based transformations and the directions of separability between phone classes are indicated by the mode shapes of LDA based transformations. To analyse the power of interpretation of PCA and LDA based transformations, mode shapes were estimated by projecting the respective eigenvectors onto the raw articulatory coordinate space. Sections 6.2.1 and 6.2.2 present the analysis of PCA and LDA mode shapes respectively.

### 6.2.1 PCA based transformations

The eigenvectors from each representation were projected on to the raw articulatory coordinate space to determine the shapes of different PCA modes. The proportion of total variance represented by each mode was also analysed simultaneously. The shapes of all PCA based feature modes for both speakers are shown in Figures C.27 to C.36.

Figure 6.1 shows the shapes of the first three modes of PC1 for the male speaker. The first PC1 mode showed the tongue moving forward/backward tangentially to its surface, and the opening/closing movement of the lips for both speakers. The second PC1 mode depicted the upward/downward movement of the tongue along with the jaw. The third mode of PC1 showed tongue tip moving independently of the blade and the dorsum, while the fourth mode showed some kind of tongue bunching. The fifth mode for the male and the sixth mode for the female speaker depicted the raising/lowering movement of the velum. The mode shapes beyond this appeared to be noisy and were difficult to interpret for both speakers. The proportion of grand variance (in %) represented

by each PC1 mode was calculated for both male and female speaker and is shown in Fig. 6.3. It was found that the first five modes explain 70% of the total grand variance (averaged across both speakers) whereas the first 11 modes of PC1 account for 95% of the total variance. Modes 8 to 14 accounted for less than 5% of total variance each.

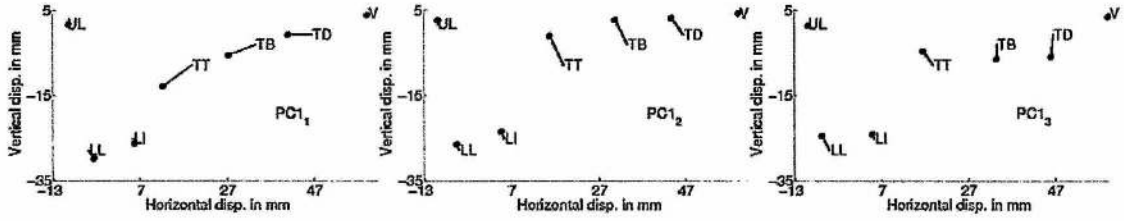



Figure 6.1: *Shapes of the first three modes of PC1, i.e., $PC1_1$ (left), $PC1_2$ (middle) and $PC1_3$ (right) for the male speaker.*

The PC3 transformation was derived from the three main independent groups the articulatory coordinates fell into depending on the pattern of the correlation amongst them. The articulatory coordinates were pooled into UL+LL+LI, TT+TB+TD and V groups. Mode shapes were calculated for all the three articulatory groups of PC3. First six modes were extracted from the lip and the jaw group. The first mode in this group depicted the open/close movement of lips and jaw and accounted for 13% of total variance when averaged across both speakers as shown in Fig. 6.3. The second and third modes accounted for 6% and 4% of total variance respectively when averaged across both speakers. The second mode depicted correlated upward/downward movements, and the third mode depicted the forward/backward movements of the lips and the jaw. The other three modes were noisy and represented a small fraction of the total variance (2% for the male and 3% for the female). Six modes numbered 7 to 12 respectively were extracted from the tongue group. Modes 7, 8 and 9 showed in Figure 6.2 represented 20%, 18% and 13% of total variance respectively when averaged across both speakers. The forward/backward movement of the tongue tip, blade and dorsum was represented by mode 7. Mode 8 showed the upward/downward movement of the tongue, and mode-9 showed the independent movement of TT with respect to TB and TD. The other 3 modes (10, 11, 12) accounted for less than 5% of total variance each and their shapes were noisy. The last two modes, numbered mode-13 and mode-14 were extracted from PCA on x and y movements of the velum. The two modes depicted opening/closing movement of velum (5% of grand variance) and forward/backward movement of velum (2% of grand variance).

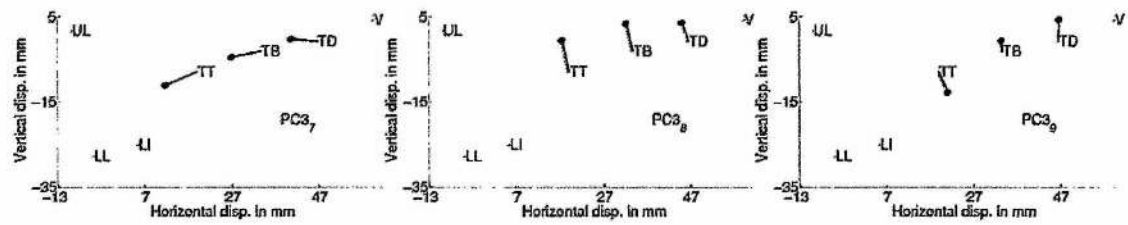



Figure 6.2: *Shapes of the three mode from tongue group,* TT+TB+TD *of PC3, i.e., $PC3_7$ (left), $PC3_8$ (middle) and $PC3_9$ (right) for the male speaker.*

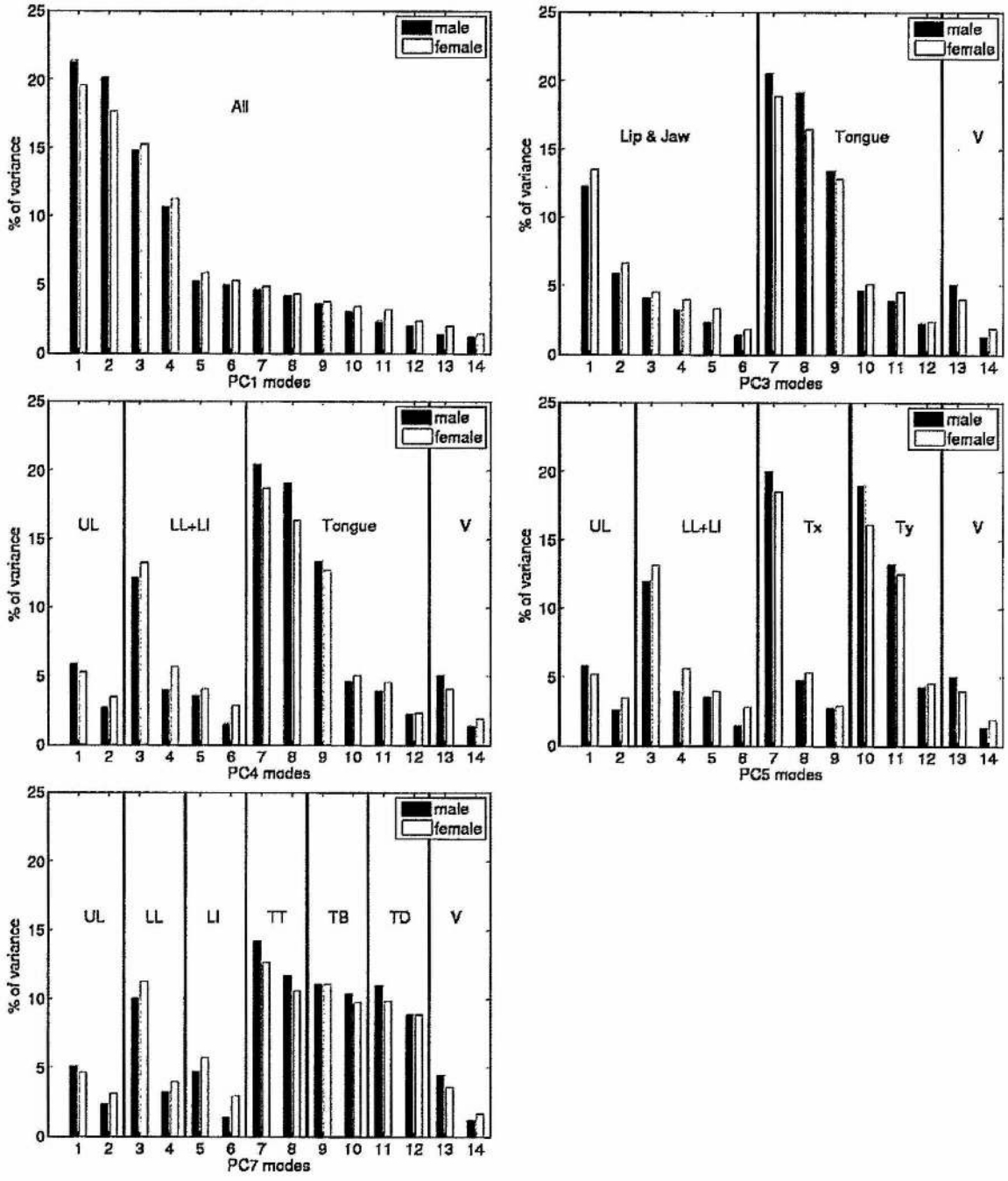



Figure 6.3: *Proportion of grand variance in percentage represented by each mode of PC1, PC3, PC4, PC5 and PC7 for male (shaded) and female (unshaded) speakers.*

Lack of strong correlations between the upper lip and the jaw for the female speaker was taken into consideration while grouping articulators for PC4 transformation. The upper lip was separated from lower lip and the jaw group to form two of the four articulatory groups for PC4 as shown in Table 6.1. The first articulatory group consisted of x and y movements of UL, whereas the lower lip and jaw were grouped together to form the second articulatory group. The tongue and the velum groups from PC3 were retained for PC4. The proportion of grand variances represented by PC4 modes are shown in Fig. 6.3. The first two mode shapes of PC4 extracted from UL showed the upward/downward (5% of total variance) and forward/backward movement (3% of total variance) of upper lip for both speakers. The next four modes, numbered 3 to 6, were extracted from LL+LI group. Mode 3 represented 13% of total variance and showed correlated upward/downward movement of LL and LI. The other modes in this group (4, 5 and 6) each accounted for less than 5% of total variance and their shapes were difficult to interpret. The mode shapes and variances of the tongue group and the velum group were similar to those in PC3. The shape of mode 8 depicted the upward/downward movement of tongue.

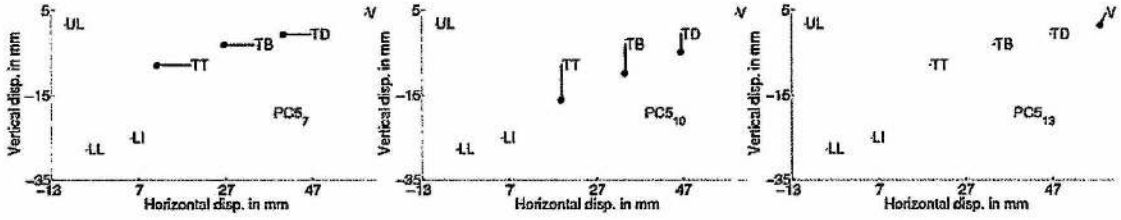



Figure 6.4: *Shapes of the PC5 modes: $PC5_7$ from $TT_x$+$TB_x$+$TD_x$ (left) group, $PC5_{10}$ from $TT_y$+$TB_y$+$TD_y$ group (middle) and $PC5_{13}$ from the* V *group (right) for the male speaker.*

The correlations between the $x$ and $y$ movements of different flesh points on the tongue were absent or rather weak for both male and female speakers. Therefore, the x and y movements of TT, TB and TD were treated independently in PC5. The tongue group (in PC3 and PC4) was disintegrated into x-only group consisting of $TT_x$+$TB_x$ +$TD_x$ and y-only group made up of $TT_y$+$TB_y$ +$TD_y$. The other three groups, UL, LL+LI and V were retained from the PC4 transform. The mode shapes and the proportional variance representations of the retained modes was identical to those in PC4. Three modes, numbered mode 7, 8 and 9 were extracted from the x-only group of tongue. The mode-7 shown in Fig. 6.4 represented the forward/backward movement of the tongue and accounted for 20% of the total variance (Fig. 6.3). The other two modes (8 and 9) in the x-only group were noisy. Three modes (10, 11 and 12) were derived from the tongue y-only group. The modes 10 and 11 represented 16% and 13% of total variance when averaged across both speakers and represented upward/downward movement of the tongue and independent movement of TT respectively. Mode $PC5_{10}$ is shown in Figure 6.4.

For PC7, each articulatory flesh point coordinate was assumed to be independent of the other coordinate and therefore, 7 articulatory groups were derived. Two modes were extracted from each articulatory group leading to a total of 14 modes. Two mode shapes were found in common across all articulatory groups, one indicating the for-

ward/backward movement and the other depicting the upward/downward movement. The proportion on grand variance represented by each PC7 mode is shown in Fig. 6.3. Of all groups, maximum variance was represented by the tongue coordinates for both speakers.

### 6.2.2 LDA based transformations

The mode shapes of all LDA based features were also generated using eigenvectors from the respective transformations. For LD1, all articulatory coordinates were pooled together to calculate a 14 dimensional feature space. It was difficult to identify distinctive gestural patterns for the modes of LD1 unlike PC1. The first three modes of LD1 are shown in Figure 6.5. For both speakers, the mode 1 direction identified the raising/lowering of tongue dorsum with velum moving forward/backward as the most discriminatory gesture between the phone classes. For the male speaker, the direction of the mode 2 depicted the forward/backward movement of tongue which was also one of the commonly identified gestures in various PCAs. For the female speaker, the mode 2 depicted the upward/downward movement of upper and lower lips and the jaw. The third mode showed the independent movement of tongue tip with respect to blade and dorsum for the male speaker. For the female speaker, the upward/downward movement of TD while the TT moved forward/backward was identified as the shape of the third mode. The movement of lower lip was opposite to that of the jaw for the first two modes for the male speaker. The fourth mode of LD1 depicted the upward/downward movement of upper lip and tongue tip for both speakers. Raising of tongue blade gesture was identified as the shape of mode 5 for both male and female speakers. The up/down movement of the TT was identified as the shape of the sixth mode for both speakers. Beyond this level, the mode shapes were difficult to interpret.

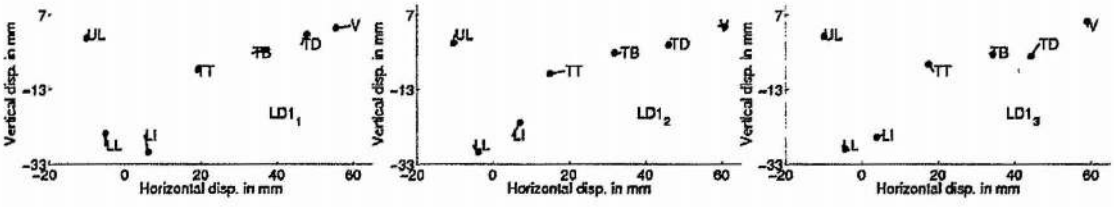



Figure 6.5: *Shapes of the first three modes of LD1, i.e., $LD1_1$ (left), $LD1_2$ (middle) and $LD1_3$ (right) for the male speaker.*

For LD3, the first 6 modes were extracted from the lips and the jaw group. For the male speaker, the first four mode shapes depicted open/close movement of mouth, lips moving up/down together, jaw and lower lip moving up/down and jaw and upper lip moving up/down respectively. For the female speaker, the first two modes depicted lips and jaw moving forward and backward in opposite directions and open/close gesture of mouth respectively. The shapes of rest of the modes from the UL+LL+LI group were difficult to interpret. Modes 7 to 12 were extracted from the TT+TB+TD group. For both speakers, each mode represented a unique tongue shape. For the male speaker, modes 7 to 10 depicted TT moving up/down when TD moves forward/backward, TT

moving forward/backward irrespective of blade and dorsum, TT+TB+TD moving forward/backward, TB up/down movement respectively. For the female speaker, modes 7 and 8 depicted TT moving up/down with TD (and TB for mode-7) moving in opposite directions. Shapes of other modes in this group were difficult to interpret. The last two modes (13 and 14) were extracted from the velum group which depicted velum moving up/down and forward/backward respectively.

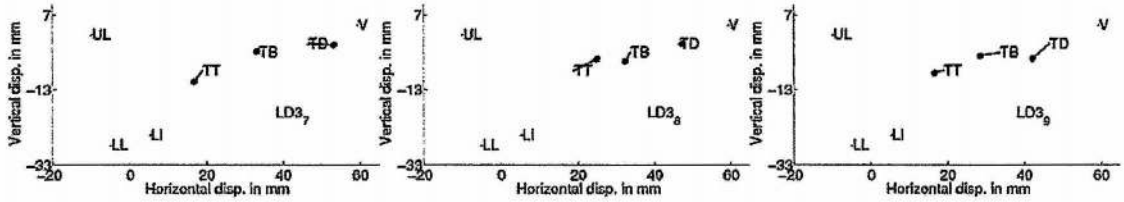



Figure 6.6: *Shapes of the first three modes of LD3, i.e., $LD3_7$ (left), $LD3_8$ (middle) and $LD3_9$ (right) for the male speaker.*

The upper lip was separated from lower lip and incisor to form two groups UL and LL+LI for LD4 like in PC4, the tongue and velum groups were retained from LD3. The first two modes were extracted from UL group which depicted the upper lip moving forward/backward and up/down respectively for both speakers. Modes 3 to 6 were extracted from LL+LI group. For the male speaker, only the shapes of modes 3 and 6 could be interpreted. The mode-3 showed lower lip moving up/down while the jaw moved forward/backward and the mode-6 showed both lip and jaw moving up/down together. For the female speaker, modes 3 and 4 depicted the jaw moving up/down in a direction opposite to that of lower lip and lower lip moving up/down while the jaw moved forward/backward respectively. The shapes of modes from TT+TB+TD and V groups were identical to those from LD3 transformation.

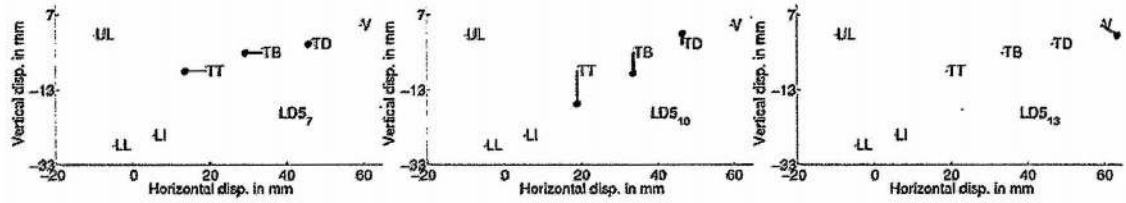



Figure 6.7: *Shapes of the LD5 modes: $LD5_7$ from $TT_x$+$TB_x$+$TD_x$ (left) group, $LD5_{10}$ from $TT_y$+$TB_y$+$TD_y$ group (middle) and $LD5_{13}$ from the* V *group (right) for the male speaker.*

For LD5, the upper lip, the lower lip and the jaw, and the velum groups were retained from LD4, where as TT+TB+TD group was spilt into the x-only ($TT_x$+$TB_x$+$TD_x$) and the y-only ($TT_y$+$TB_y$+$TD_y$) groups. Two modes from UL and four modes from LL+LI groups formed the first six modes of LD5. The shapes of these modes were identical to those in LD4. Three modes each were extracted from tongue x-only (7 to 9) and y-only (10 to 12) groups. For the male speaker, mode-7 represented tongue tip and blade moving forward/backward together, whereas mode-8 represented tongue tip and dorsum moving forward/backward together while blade moved in opposite direction. For the female speaker, mode-7 represented tongue blade and dorsum moving

forward/backward together whereas the shape of mode-8 was similar to that for the male speaker. The modes 10 to 12 from y-only group for the male speaker represented correlated movement of tongue tip and blade opposite to that of tongue dorsum, tongue blade and dorsum moving up/down in opposite directions and in similar directions respectively. For the female speaker, modes 10, 11 and 12 represented tongue blade and dorsum moving up/down in a direction opposite to that of tongue tip, tongue dorsum moving up/down independently of tongue tip and blade and correlated up/down movements of tongue tip and dorsum respectively. Modes 13 and 14 extracted from velum group were identical to those from LD3 and LD4 transforms.

Two modes were extracted from each of the 7 groups for LD7. For both speakers, the first mode indicated forward/backward movement and the second mode indicated the upward/downward movement of the articulator in each group. Before the comparison of feature sets, analysis of different critical modes identified using the proposed algorithm in PCA and LDA feature spaces was carried out. The following section, Section 6.3 presents the evaluation of critical modes in different PCA and LDA feature spaces.

## 6.3 Applying the ACIDA

Critical modes were identified using the proposed algorithm for different PCs and LDs and the corresponding model distributions were estimated for each phone using the algorithm. The critical threshold for all PCs and LDs was set to the IPA level of complexity, where the average number of critical modes per phone were equivalent to that derived from the IPA chart for articulatory coordinates. The critical threshold at the IPA level of complexity was equal to 1.5 for all other PCA features except for PC1 for the male and PC4 for the female ($\theta_C = 1.4$), PC1 for the female ($\theta_C = 1.3$). The critical threshold was set to 1.6 for LD1, LD5 and LD7 for the male speaker to obtain same number of critical dimensions at the IPA level of complexity for the male speaker. The critical threshold was set to 1.8 for LD3 and 1.7 for LD4. For the female speaker, the critical threshold at the IPA level of complexity was set to 1.5 for LD1 and LD3, and 1.4 for LD4 and LD5.

Table 6.2 shows the top three most commonly identified modes various PCA and LDA based feature spaces. For most of the PCA features, the shapes of the top 3 modes for both speakers were similar. The mode shapes of PC7 were easy to interpret of all PCA based transformations. The top three critical modes for PC7 indicated the upward/downward movement of lower lip, tongue tip and tongue dorsum for the male speaker. In the raw articulatory space, $\text{LL}_y$, $\text{TT}_y$ and $\text{TD}_y$ were identified as critical for most of the phones. The results were also similar for the male speaker, except that the tongue blade was identified as critical instead of the tongue dorsum. For PC5 and PC4, all 3 modes were from the tongue group for the female speaker whereas one lip&jaw mode made it to the top three list for the male speaker. For PC3, two of the top three modes indicated the open/close movement of the mouth and tongue tip moving independently of tongue blade and dorsum in the y direction for both speakers. The other mode represented the tongue x movements for the female and tongue y movements for the male speaker. The findings for PC3, PC4, PC5 and PC7 were in

agreement with the expected articulatory coordinates from the raw articulatory space. The mode shapes of PC1 were difficult to interpret for both speakers. Modes 2 and 3 which depict different y movements of tongue were the most commonly identified critical modes for both speakers. For the male speaker, mode 4 which represents the open/close movement of the lips was identified as the 3rd most commonly identified critical mode where as mode 1 which represents tongue x movement was identified for the female speaker.

| Features | Modes | | Features | Modes | |
|---|---|---|---|---|---|
| | (m) | (f) | | (m) | (f) |
| PC1 | 2, 3, 4 | 3, 2, 1 | LD1 | 1, 2, 3 | 1, 2, 3 |
| PC3 | 1, 8, 9 | 9, 1 , 7 | LD3 | 1, 7, 8 | 7, 8 , 1 |
| PC4 | 3, 9, 8 | 9, 7, 8 | LD4 | 3, 7, 8 | 7, 8, 3 |
| PC5 | 3, 11, 7 | 10, 11, 7 | LD5 | 3, 11, 10 | 11, 12, 3 |
| PC7 | 3, 7, 9 | 9, 3, 11 | LD7 | 4, 8, 10 | 4, 8, 10 |

Table 6.2: *The top three most frequently identified modes for different PCA and LDA based representations for male (m) and female (f) speakers at the IPA level of complexity. Refer to Table 6.1 for articulatory groups and corresponding mode indices for each representation.*

For LD7, the most commonly identified critical modes were similar for both speakers. From LD5 onwards, though the top 3 modes came from similar articulatory groups, the mode shapes for the male speaker were different from those of the female speaker. For example, the mode 7 of LD4 for the male speaker depicted the tongue tip and tongue dorsum moving horizontally in opposite directions, whereas the same mode for the female speaker indicated the tongue tip moving independently of blade and dorsum in y direction. However, the tongue shape of mode 10 of LD5 for the male speaker was similar to that of mode 11 of LD5 for the female speaker. It was difficult to correlate the mode shapes of the male speaker with those of the female speaker for LD1.

### 6.3.1 Critical modes

A brief summary of the critical mode shapes identified for consonants, vowels and diphthongs is presented in this section. Lists of critical modes are given in Tables C.23 to C.25.

The modes from the velum group were identified as critical for nasalised stop consonants. For all bilabial stops [p, b, m], critical modes were identified from the articulatory groups formed from upper lip, lower lip and jaw. For alveolar stops [t, d, n], critical modes were identified from the TT+TB+TD group for PC3, PC4 and LD3, LD4 features. For PC5 and LD5, the identified critical modes were from tongue y-only group. For PC7 (male only) and LD7, the identified critical mode shapes indicated up/down movement of tongue tip. For the female speaker, the mode indicating up/down movement of the jaw which is highly correlated with that of the tongue tip was identified as critical. For velar sounds [k, g, ŋ], critical modes were also identified from the tongue group.

The identified critical modes indicated the movement of tongue dorsum (i) along with tongue tip and blade, (ii) along with the tongue blade, but independent of the dorsum. For interdental sounds [f, v], the identified critical modes belonged to the lips and jaw groups and the mode shapes indicated the forward/backward movement of the lips. For interdentals, [θ, ð], the critical mode shapes indicated the forward/backward movement of tongue tip , along with the raising of the jaw. For alveolar sibilants [s, z], post-alveolar sibilants [ʃ, ʒ] and alveolo-palatal sibilants [ʧ, ʤ], the mode shapes of identified critical modes were in agreement with the two mechanisms required for uttering sibilants, (i) forming a narrow channel using tongue and (ii) creation of obstacle using jaw position. Identified critical modes were from tongue group for [ɹ] (forward/backward movement of tongue), [j] and [w] (upward/downward movement of tongue). For [w], critical modes were also identified from lip and jaw groups for rounding gesture.

For all vowels, the critical coordinates came from the lip and jaw groups and the tongue group. No critical modes were identified for neutral vowel [ə] for all PCs and LDs. Relatively few critical modes were identified for other near-central vowels [ɪ, ʌ, ʊ]. The modes from the lip and jaw groups were mostly identified as critical for front open vowels [æ, ɛ], from tongue group for front close vowels [iː, i], close back vowel [u] and from both lip and tongue groups for the rest of the vowels. None of the critical coordinates were identified from the velum group for any vowel sounds.

For diphthongs, the majority of the critical coordinates were from the tongue group, followed by the jaw and lip groups. Critical modes were also identified from the V group for [ɔɪ] for both initial and final vowels. When compared with other diphthongs, relatively few critical coordinates (rather none for some PCs) were identified for [oʊ]. Critical coordinates were also identified for centralised vowels unlike in monophthongs due to the influence of initial vowel and final vowels on each other.

| Phones | Raw | PC7 | PC4 | PC1 | LD1 |
|---|---|---|---|---|---|
| b | $UL_y$, $LL_y$ | 1, 3 | 1,3 | 4, 3 | 2 |
| t | $TT_y$ | 7 | - | 2 | 1 |
| k | $TD_y$ | 11 | 9 | 3, 5 | 4, 3 |
| θ | $TT_x$, $TT_y$, $LL_y$ | 8, 3, 7 | 7, 3 | 1, 3, 2 | 1, 6, 3 |
| v | $LL_y$, $UL_y$ | 3, 4, 1 | 3, 4, 1 | 7, 5 | 7, 2 |

Table 6.3: *Critical coordinates identified using ACIDA algorithm in raw, PC7, PC4, PC1, and LD1 feature spaces at the IPA level of complexity for the male speaker. Refer to Table 6.1 for articulatory groups and corresponding mode indices for each representation.*

Table 6.3 depicts critical modes identified for PC7, PC4, PC1 and LD1 in comparison with the raw articulatory space for the male speaker. The critical modes from the PC7 and LD7 groups were easy to interpret and correlate to the critical articulatory coordinates in the raw articulatory space. For example, for [k], the critical mode $PC7_{11}$ was derived from the TD group and depicted the y movement of tongue dorsum. Similarly modes $PC4_1$ and $PC4_3$ which depict the open/close movements of upper lip, lower lip and jaw were identified as critical which are in agreement of the critical coordinates from the raw articulatory space.

The models lost power of interpretation as more strong inter-articulatory correlations were taken into account for generating transformations using PCA and LDA. For PC1 and LD1, the dependent update step had no effect on the identification of critical modes due to the absence of correlations amongst the modes. Modes 2, 3 and 4 for PC1 were identified as critical for most phones for the male speaker. For the female speaker, modes 3, 2, and 1 from PC1 representation were identified as critical for 27%, 23% and 22% of total phones respectively. Phones with different expected coordinates shared some common critical modes. For example, for a bilabial stop [b] for PC1, mode 3 which depicts open/close movements of lips was identified as one of the critical modes. This mode was also identified as critical for a velar stop [k], since the mode also depicts the correlated up/down movement of the tongue dorsum and blade independent of the tongue tip. In cases where more than one critical mode was present for a phone, the critical modes either shared a common gesture or depicted two essential but different gestures. For example, for phone [θ], PC1 modes 1, 2 and 3 were identified as critical. Here modes 1 and 3 commonly share tongue tip moving up/down gesture and where as mode 2 depicts the tongue tip moving forward/backward which is also essential for producing this interdental phone. For LD1, the first three modes were identified as critical for most number of phones for both male and female speakers. The LD1 critical modes painted a similar picture to that of PC1 critical modes. For both PC1 and LD1, it was difficult to relate the shapes of critical modes of some phones to their respective expected critical gestures. For example, it was difficult to interpret $LD1_7$ (Table 6.3) which was identified as the first critical mode for [v].

Model distributions, $\mathcal{N}(\hat{m}^{\phi}, \hat{S}^{\phi})$ were obtained for each phone $\phi$ in all PCA and LDA based feature representations. The informational efficiency of each feature representation was analysed using the respective model distributions in the following section, Section 6.4.

## 6.4 Informational efficiency

In this section, different articulatory representations were compared to identify the best feature set which yields the most informationally rich model distributions. To analyse the information content of the model distributions in each feature space, the evaluation scale measure (Section 3.4.1) was used. Evaluation scale value, $\Upsilon_{eval}$, was computed as the 14D KL divergence between the model distributions and the phone-specific distributions for each representation. Smaller values of evaluation scale indicate that the model distributions (computed from the knowledge of articulatory roles and correlations) closely match the actual phone specific distributions. From this analysis, we intend to find

- whether transforming the data using PCA and LDA gives any advantage over the raw articulatory representation,
- what kind of optimisation criterion, the direction of maximum variability or the direction of maximum separability, gives informationally efficient models, and

- how do the models obtained from the articulatory groups (independent components) fare with the traditional LD1 and PC1 representations.

### 6.4.1 Evaluation scale

The identification divergence was subject to a range of critical thresholds, $0.1 \leq \theta_C \leq 5$, for each feature representation and the evaluation scale values were estimated from the resulting model distributions. The evaluation scale value at each threshold was averaged across all phones.

Figure 6.8 shows the average evaluation scale plotted across average number of critical modes per phone for a range of critical thresholds for the male speaker. The pattern of results for the female speaker was also similar to that of the male speaker (refer to Figure C.2). The average evaluation scale values were plotted for all PCA based (solid) and LDA based (dashed) representations along with the raw EMA representation (solid black). Recall from Chapter 3 that evaluation scale is a measure of the distance between the model and actual phone distributions and therefore smaller values are desirable.

The evaluation scale values at different critical threshold values show the pattern of the fit of the models. The performance of all LDA and PCA based models was better than the models from the raw articulatory features at all values of critical threshold. Also LDA based models performed better than PCA based models when features from similar articulatory groups were compared. For example, the performance of LD1 was better than PC1 and so on. Therefore, transforming the raw articulatory data in the direction which ensures maximum separation between groups i.e., LDA, gives the best performance over the transformation in the direction of maximum variance, i.e., PCA. The results also showed that representations generated by ignoring weak correlations, i.e., LD3 and PC3, give slightly worse performance over LD1 and PC1 models respectively. The goodness of fit of the models deteriorates as more correlations are ignored, i.e., PC3>PC4>PC5>PC7 (similarly LD3>LD4>LD5>LD7). The PC7 and LD7 models gave the worst performance of all PCA and LDA representations which was close to the performance of models from the raw articulatory space.

The evaluation scale values at the IPA level of complexity are shown in Figure 6.9 for both speakers. Here, the fit of the models to the full phone distributions was improved by 73% (78%/68%-m/f) using LD1 over raw articulatory coordinate space. The next best performance was given by PC1 where improvement obtained over raw articulatory space was 61% when averaged across both speakers. The next best performance was given by the LD3 models (followed by the PC3 models) which are estimated from three articulatory groups UL+LL+LI, TT+TB+TD and V formed by ignoring weak correlations amongst articulators. The fit of the models worsened as more correlations between articulators were ignored in forming articulatory groups. The worst performance of all PCA and LDA based groups was given by PC7 (followed by LD7) at the IPA level of complexity. Only 7-8% improvement on average was given by PC7 and LD7 transforms over the raw feature space.

It can be concluded from this analysis that transforming the data in the direction of better separability between phone classes generates more informationally efficient

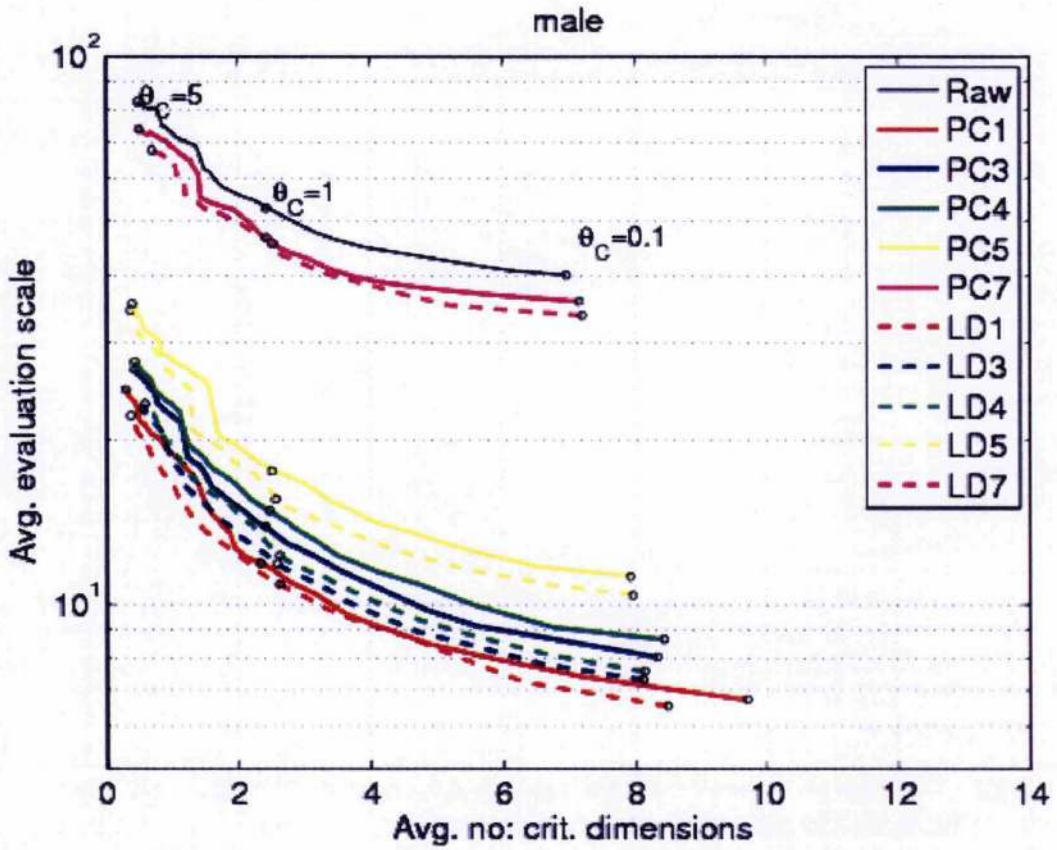



Figure 6.8: *Evaluation scale $\Upsilon_{eval}$ averaged across all phones on the y axis and the average number of critical dimensions per phone on the x axis. Evaluation scale was computed from PC1, PC3, PC4, PC5, PC7 and LD1, LD3, LD4, LD5, LD7 features at various critical thresholds, $0.1 \leq \theta_C \leq 5$ for the male speaker.*

models than in the direction of maximum variability. Models from LDA and PCA based features are more efficient than those from the raw articulatory space. However, ignoring the correlations amongst the articulatory coordinates and performing PCA or LDA on the resulting independent articulatory groups gives only a small improvement (7-8%) over the raw articulatory models.

## 6.5 Compactness of the models

The PC1 and LD1 representations had highest power of model compactness due to the absence of correlations between the modes and also lowest transparency of interpretation. The interpretation power was higher for representations generated from independent articulatory groups (by eliminating correlations amongst articulators). For all such representations, the intra-group correlations amongst the modes were absent but the intergroup correlations amongst the modes were present. For example, no correlations were present amongst modes 1 to 6 from the UL+LL+LI group of PC3, but modes

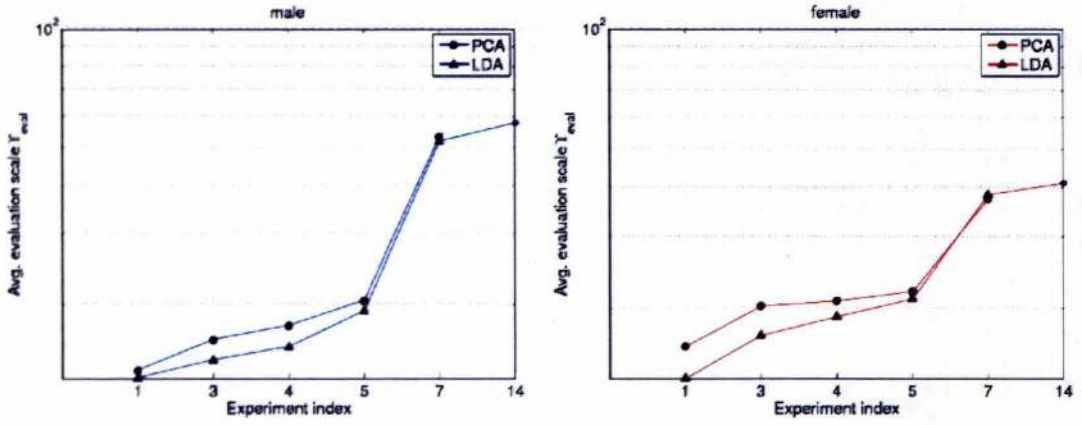



Figure 6.9: *Average evaluation scale computed from PCA and LDA based features in comparison with the raw articulatory space (experiment index 14) at the IPA level of complexity for the male (left) and the female (right) speakers.*

1 to 6 were correlated with modes 7 to 12 from the TT+TB+TD group of PC3. Of all PCA and LDA based representations, PC7 and LD7 features were easy to interpret and were least compact. The compactness of $PC1 > PC3 > PC4 > PC5 > PC7$, similarly, $LD1 > LD3 > LD4 > LD5 > LD7$.

Assume that in the *conventional* approach to estimating model distributions, neither the knowledge of articulatory roles nor the interdependencies are considered, and model distributions are simply set equal to the phone specific distributions. If $a$ (14) represents the number of modes, the number of parameters (means and variances) needed for modelling $\varphi$ phones in the *conventional* approach is $2a\varphi$.

When the proposed algorithm (ACIDA) is used for estimating the model distributions in each feature space, if $\hat{k}^{\phi}$ represents the number of critical dimensions per each phone $\phi$ , then number of parameters required are

$$2a + \frac{a(a-1)}{2} + \varphi\left(2\hat{k}^{\phi} + \frac{\hat{k}^{\phi}(\hat{k}^{\phi}-1)}{2}\right) \tag{6.1}$$

which includes grand means and variances ($2a$) along with grand inter-articulatory correlations ($a(a-1)/2$), phone means and correlations ($2\hat{k}^{\phi}$) and phone correlations ($\hat{k}^{\phi}(\hat{k}^{\phi}-1)/2$).

At the IPA level of complexity, for PC1 and LD1, the reduction in number of parameters required for estimating model distributions was 82% over conventional approach due to the absence of correlations between the modes. For PC1 and LD1, Eq. 6.1 reduces to

$$2a + \varphi\left(2\hat{k}^{\phi} + \frac{\hat{k}^{\phi}(\hat{k}^{\phi}-1)}{2}\right) \tag{6.2}$$

For other PCA and LDA based representations, though correlations between modes within each articulatory group were absent, the inter-articulatory group correlations

were present. The reductions in number of parameters over conventional models were: 78% for PC3/LD3 and PC4/LD4, 77% for PC5/LD5 and 76% for PC7/LD7 and raw articulatory features.

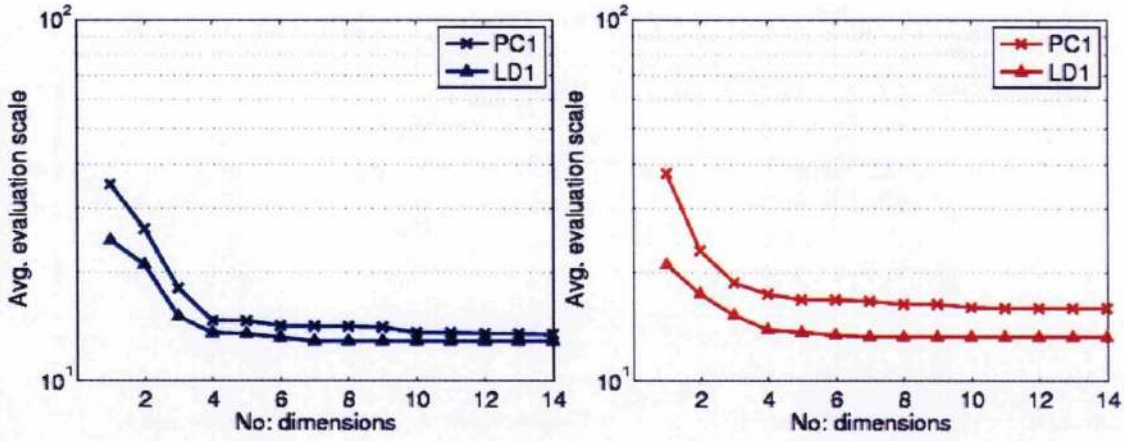



Figure 6.10: *Evaluation scale computed from model distributions obtained from ACIDA at the IPA level of complexity averaged across all phones for male (blue) and female (red) speakers plotted for PC1 and LD1 as the feature dimensionality is varied from 1 to 14.*

### 6.5.1 Dimensionality reduction

The information contained in the modes of the most compact representations of all, PC1 and LD1, are analysed. This findings of this analysis could help in making the models more compact by identifying the last few noisy modes for dimensionality reduction. Figure 6.10 shows the information conveyed by the PC1 and LD1 modes as the dimensionality of representation is varied from 1 to 14. Evaluation scale computed from models trained using ACIDA algorithm at the IPA level of complexity was averaged across all phones. The dimensions were increased from 1 to 14. At dimensionality level 14, the information of all modes was available whereas at level 1, only the information from the 1st mode, i.e., $PC1_1$ (or $LD1_1$) was made available. The critical modes were selected by ACIDA algorithm from the information available at each dimensionality level. The knowledge of critical modes was used to train the model distributions of each phone. The evaluation scale plot shows that the first 4 modes of PC1 and LD1 contain most of the information. Figure 6.3 shows that the first 4 modes of PC1 account for almost 65% of total variance for both male and female speakers. The information contained by the models improved by 56% for PC1 (45 % for LDA) as the dimensionality was increased from 1 to 4 for the male speaker. The improvement for the female speaker was 54% for PC1 and 34% for LD1. The improvement obtained when the dimensionality was increased from 4 to 10 was very small, 8% for both speakers for PC1 and 5% for the male (4% for the female) for LD1. No improvement was achieved when dimensions were increased from 10 to 14 for both speakers in PC1 and LD1 representations.

This analysis showed that the first four modes of PC1 which represent a significant portion of total variability had most information. Similarly, the first four modes of LD1 convey most of the information which maximises the separability between phone classes.

## 6.6 Recognition performance

Preliminary phone classification (the phone boundaries are known) experiments were performed on acoustic and articulatory data using the SEGVit speech recogniser (Singampalli, 2006). Linear trajectory segmental HMMs (LTSHMMs) were used for each speaker along with a bigram language model to test the phone classification performance of male and female speakers. The best classification accuracy reported on the acoustic data (MFCCs+$\delta + \delta\delta$) was 72% (averaged across male and female speakers) given by LTSHMMs. The classification accuracy of monophone HMM models on MFCC data with first and second ordered features was 65%. The classification performance of articulatory models trained on PC1 feature set with first and second order features was comparable to acoustic models for LTSHMMs and slightly worse for HMMs (63%).

| | Our experiments (Recognition accuracy %) | | Wrench (2001) (Recognition accuracy %) | |
|---|---|---|---|---|
| Experiments | Male | Female | Male | Female |
| LD1 | **62** | **62** | | **63** |
| PC1 | 60 | **62** | | 55 |
| LD3 | **62** | 61 | | |
| PC3 | 60 | 61 | | |
| LD4 | **62** | 60 | | |
| PC4 | 60 | 60 | | |
| LD5 | 61 | 60 | | |
| PC5 | 60 | 60 | | |
| LD7 | 59 | 60 | | |
| PC7 | 59 | 60 | | |
| EMA | 58 | 60 | | |
| MFCC | 52 | 47 | **63** | **65** |
| $MFCC_{PC}$ | 54 | 47 | | |
| $MFCC_{LD}$ | 56 | 50 | | |

Table 6.4: *Recognition accuracy values averaged across all 5 jack-knife results for different LDA, PCA, articulatory and acoustic features using monophone models for both speakers in comparison with values reported by Wrench (2001) using triphone Gaussian mixture models.*

Phone recognition experiments (unknown phone boundaries) were performed on various PCA and LDA feature sets and raw articulatory coordinate features using HTK v3.4.1 (Young et al., 2002). The recognition performance of articulatory representations was also compared with that obtained from the acoustic data. The first 14 MFCC coefficients were extracted from the acoustic data and were transformed using PCA and LDA

methods. Phone recognition results were obtained from MFCC, MFCC transformed using PCA ($MFCC_{PC}$) and LDA ($MFCC_{LD}$). All articulatory and acoustic feature spaces were augmented with delta and acceleration features and monophone models were used for all acoustic and articulatory features. All available data was divided in to 5 training and test groups and 5 sets of recognition results were obtained in each feature space. A bigram language model was estimated from each training set and was used in recognition experiments. For all articulatory spaces, the log energy of the laryngograph was also included in the feature set, whereas for acoustic data, the zeroth MFCC coefficient was included. The dimensionality of all feature sets including energy coefficients along with first and second order time derivatives was equal to 45.

Table 6.4 shows the recognition accuracy averaged across all jackknife results from each feature set for both male and female speakers. The best recognition accuracy (62%) was given by LD1, LD3 and LD4 features for the male speaker, and LD1, PC1 features for the female speaker. The recognition performance of monophone models in PC1 feature space (62%) was better than the results obtained from triphone models (55%) reported by Wrench and others (Wrench, 2000, 2001) on female speaker data. The performance of monophone LD1 models (62%) was also comparable to the best performance reported by Wrench (Wrench, 2001) (63%) using triphone models on the female speaker data. The improvements in performance could be due to better transcriptions generated by correcting some of the labelling errors (Chapter 2).

The statistical significance of difference between the recognition accuracy values was tested using student's t-test and f-test at level of significance $\alpha = 0.05$. The results from male and female speakers from similar feature sets were combined for testing statistical significance. The recognition performance of EMA features (58%/59% -m/f) was only comparable to the performance of PC7 and LD7 models and no significant differences was found between raw (EMA) , PC7 and LD7 results as expected. Statistically significant difference was found between the recognition accuracy of LD1 when compared with all other PCA representations except PC1. No significant difference was found between the performance of PC1 and other features except PC7 and LD7. The overall performance of articulatory feature sets was better than the acoustic features.

The poorest recognition performance of all feature sets was given by MFCC features (51%/47% - m/f). This was worse than the recognition accuracy using triphone Gaussian mixture models reported by Wrench (2000) (57% - f) and Wrench (2001) (63.5%/ 65% -m/f). Performing PCA on MFCC data improved the recognition performance of acoustic features only for the male speaker, whereas LDA improved performance for both speakers. The recognition accuracy from $MFCC_{PC}$ features (54%/47% - m/f) was slightly worse than that from $MFCC_{LD}$ features (57%/50% - m/f). However, the improvements achieved using LDA and PCA on MFCCs were not statistically significant. The differences between all acoustic and articulatory results were statistically significant.

## 6.7 Conclusions

Articulatory feature representations were derived by applying two linear orthogonal transforms, PCA and LDA, on the measured articulatory data. Apart from the con-

ventional PCA and LDA feature representations, other PCA and LDA based feature vectors were derived from independent articulatory groups. The independence between the groups was established by ignoring correlations and each group was transformed independently using PCA or LDA. Different feature representations were compared based on power of interpretation, informational efficiency, compactness and recognition performance. More conventional representations such as PC1 and LD1 were found to be most compact but difficult to interpret. The power of interpretation increased as a greater number of articulatory coordinates were pooled into independent groups (i.e., from PC3 to PC7 and LD3 to LD7). The analysis demonstrated that the LDA and PCA based representations outperformed the raw articulatory representation and LDA gave slight improvement over PCA in all aspects. However, PCA and LDA gave little benefit on articulatory groups obtained by discarding all correlations other than those between x and y movements of each articulator (i.e., PC7 and LD7). The phone recognition performance of articulatory feature vectors was better than the acoustic feature vectors when a simple monophone model frame work was used along with a bigram language mode.

The following chapter, Chapter 7, analyses different approaches to modelling coarticulation using articulatory roles and model distributions from various feature representations discussed here, in a trajectory synthesis framework.

# Chapter 7

# Modelling coarticulation and trajectory generation

## 7.1 Introduction

Speech production is a complex process where the information is processed at different mutually interacting levels, starting from the speaker's intention to communicate to fluent articulation for generation of speech (Levelt, 1989; Garrett, 1980). At the higher levels, the speech production is influenced by various social, pragmatic, semantic, syntactic, prosodic and phonetic factors. The model of speech articulation presented here is concerned only with the lower levels of speech production where the intended utterance in the form of a sequence of phones is converted to articulation, the affects of the higher processing units on the speech are not considered. A simple production oriented model comprising of planning and execution stages shown in Figure 7.1 is considered for modelling speech articulation.

In the model planning stage, articulatory targets are specified for each phone in the utterance. Trajectories representing smooth, blended and continuous movements are generated from the articulatory targets in the execution stage. In the hierarchical model of speech production, the representation between the planning and execution phases where targets are specified belongs to the empirical level. Synthesis of speech (physical level) is beyond the scope of this work. Each level in the speech production process is governed by different set of constraints (Butterworth, 1980). In the planning stage, it is important to know the constraints on articulators as well as their degrees of freedom. Typically, the model planning stage is conditioned on the static constraints derived from phonological knowledge.

Target specification plays an important role in modelling coarticulation in the execution stage. The target representations resulting from the planning stage have been modelled as spatial targets (Henke, 1965; MacNeilage, 1970), tract variables (Saltzman and Munhall, 1989), vocal tract shape and area functions (Lindblom, 1990), muscle length targets (Cohen et al., 1988), windows (Keating, 1988) and convex regions (Guenther, 1994). The window based target representations allow a range of articulatory move-

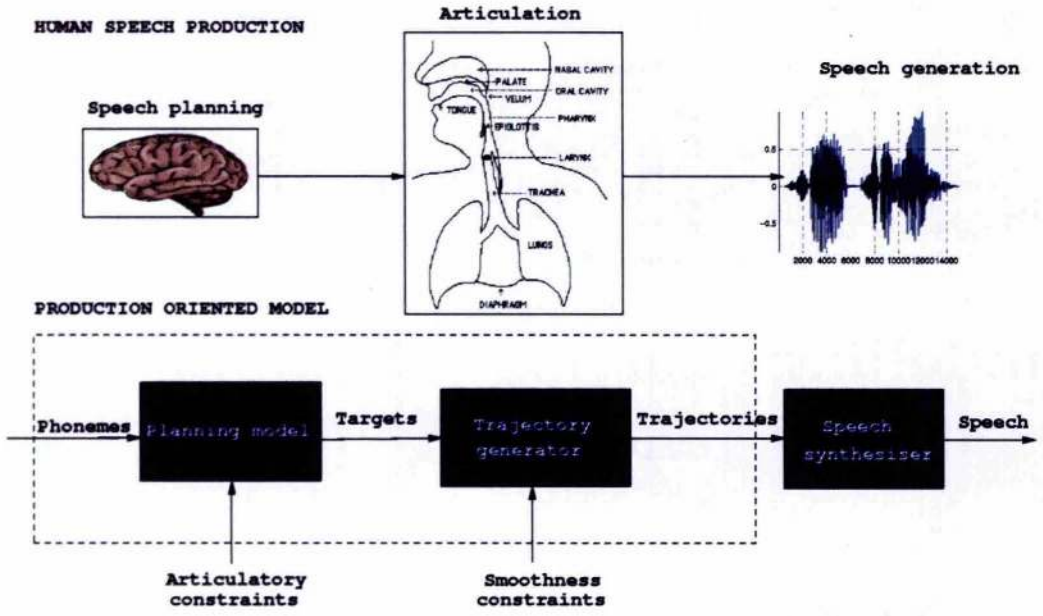



Figure 7.1: *Illustration of overview of human speech production process (top) and production-oriented speech synthesis (bottom). The midsaggital outline is taken from Rubin and Vatikiotis-Bateson (1998).*

ments for each phone to facilitate the variation in the articulatory positions due to coarticulation (Perkell, 1980).

In the model execution phase, the articulatory movements are synthesised from the target representations derived from the planning stage. Smoothness constraints are essential to generate slow and blended movements of articulators. Commonly used techniques to generate trajectories are linear interpolation (Blackburn and Young, 2000), neural networks (Richmond, 2006), MLPG algorithm (Tokuda et al., 2000), trajectory HMMs (Tokuda et al., 2007) etc..

Coarticulation occurs naturally in fluent speech and is one of the main sources of variability is speech. Though coarticulation affects both planning and execution stages of speech production, coarticulation theories have focused largely on either planning or the execution stages. Temporal coarticulation spanning longer durations, for e.g., anticipatory velum lowering in nasal contexts, has been better explained by models built in the planning stage. Coarticulation due to overlap of smooth and continuous articulatory movements, variation due to compensatory articulation and minimisation of effort have been attributed to the execution stage. In the execution stage, economy of effort refers to the minimised expenditure of physical and bio-mechanical energy. As explained by Lindblom (1990), articulators naturally tend to minimise effort when speaking. There is a trade off between the amount of articulation (ranging from hypo to hyper) and the need to communicate robustly. Adjustments to minimise the biomechanical effort lead to spatial coarticulation effects such as target overshoot/undershoot. Transitions from one target to another are more rapid for some articulators than others due to varying degrees of inertia. Heavy articulators such as the jaw have greater inertia and

resist being set in motion, whereas articulators such as the tongue tip can move rapidly. Minimisation of bio-mechanical effort due to inertia, stiffness and damping characteristics of an articulator leads to gesture reduction and assimilation of some phones in fluent speech. In the proposed approach, the concept of economy of effort is applied in the planning stage. Sparse and efficient representations are derived for modelling the redundant degrees of freedom from the knowledge of constraints.

There is scope for improving the performance of articulatory synthesis when coarticulation is modelled in both planning and execution stages of speech production. In the planning stage, better representations of the constraints could potentially lead to better models of coarticulation. Most of the existing representations of constraints categorize the degrees of freedom in to constrained/unconstrained categories and do not consider the partial constraints on articulators due to biomechanical links. Articulatory constraints in the form of critical, dependent and redundant articulatory roles derived statistically from the measured articulatory data using articulatory constraint identification algorithm (ACIDA) capture the essence of speech production mechanism. Recall from Chapter 2 that a **critical** articulator is constrained to achieve a target position to produce a speech sound and causes coarticulatory effects on the neighbouring articulators. Note that the target position of a critical articulator could be subject to undershoot/overshoot due to inertia and damping characteristics of articulators or due to lack of sufficient time to reach the idealised target position. A **dependent** articulator is partially controlled by critical articulator(s) due to the presence of bio-mechanical correlation(s) between them. The remaining degrees of freedom of a dependent articulator are free to move. A **redundant** articulator is completely unconstrained and is prone to coarticulatory effects to a higher degree. Also the evaluation scale analysis (Figure 4.2) in Chapter 4 demonstrated that the models trained from critical, dependent and redundant knowledge (ACIDA) are a better fit to actual phone distributions than those from constrained/unconstrained (IPA) representations. Also inclusion of interarticulatory correlations in the form of D-step improved the performance of IPA models. This section is aimed at analysing the potential of identified roles in modelling coarticulation in the planning stage using a statistical framework.

On the other hand, variation in the execution phase can be better modelled using statistical targets when compared with spatial targets and rectangular windows (Blackburn and Young, 2000). Powerful statistical models such as trajectory HMMs, neural networks etc. which make efficient use of existing data can be used to generate trajectories from such target distributions. In the proposed approach, the target representations derived from the planning stage are modelled as point targets (means) and could be extended to statistical window based representations by incorporating the target variance along with the mean.

The main aim of this work is to provide a statistical framework for modelling coarticulation using role information in the planning stage of speech production and to evaluate its potential in trajectory synthesis. In this chapter, two aspects of coarticulation are modelled using role information in the model planning stage. The approach is introduced as follows:

- Temporal coarticulation: Recall from Chapter 2 that the temporal coarticulation refers to the extent in time to which the target position of an articulator required

for a phone can influence the neighbouring unconstrained phonetic segments. Temporal coarticulation effects could be anticipated or carried forward in time. In the proposed approach, the critical articulator is considered to be the source of temporal coarticulation on the neighbouring non-critical, i.e., dependent and redundant articulators. The extent to which the neighbouring articulator gets affected is modelled differently for dependent and redundant articulators. The position of a dependent articulator is partially determined by its correlation with the critical articulator(s). The remaining degrees of freedom of a dependent articulator are prone to anticipatory and carry forward coarticulatory effects. A redundant articulator is completely unconstrained and can be completely subject to the temporal coarticulatory effects due to the neighbouring critical articulators. As a result, the redundant articulator can completely anticipate or carry forward a target position of the neighbouring critical articulator.

- Modelling redundant degrees of freedom using minimum effort principle: During speech articulation, the articulators assume a low cost behaviour to economise the bio-mechanical effort. Consider an example where an articulator becomes briefly unconstrained before reaching the next target from the previous target position. During the unconstrained phase, the articulator can either relax completely or assume a position which satisfies requirements of the target sequence but using limited resources; the former refers to relaxation and the later constitutes the economy of effort behaviour of the articulator. In either case, the position of the articulator during the unconstrained phase does not contradict the production of the speech sound. The concepts of relaxation and economy of effort were applied to model the redundant degrees of freedom in the planning stage of speech. The relaxation principle was implemented by imposing no constraints (from neighbouring phones) on redundant degrees of freedom of dependent and redundant articulators. A dependent articulator's position is only controlled by the biomechanical constraints (i.e., its dependency on current critical articulator) and a redundant articulator is modelled as grand distribution. The economy of effort principle was implemented by deriving the current target position in the planning stage by predicting smooth transitions from the previous target to the next. For a dependent articulator, the economical position during the transition would be intermediate between the fully constrained (effect of neighbouring phone + dependency on current critical articulator) and the partially constrained position (dependency only). For a redundant articulator, the economical position would be intermediate between completely relaxed (grand) and completely constrained (due to neighbouring phone) positions.

This work is not aimed at providing a complete explanation for each aspect of coarticulation but to identify the single strongest theory of the existing explanations. Therefore, the theories were tested individually in the proposed framework whereas in reality different aspects of coarticulation occur together.

The theories of speech articulation presented above were implemented in the planning stage using articulatory role information and were evaluated by generating synthetic trajectories in the execution phase. The span of coarticulation was limited to immediate

neighbours. Triphone contexts where the articulator plays a critical role for at least one of the phones in the sequence were considered. Other than triphones, redundancy modelling constraints were also applied to quadphones only when the articulator is critical at the either ends of the quadphone sequence. Target distributions with mean and variance as parameters were generated for each articulator.

Two simple trajectory generation methods were used to evaluate the theories implemented in the planning stage. Trajectories were generated using linear interpolation and Blackburn and Young's model of coarticulation. Linear interpolation is a simple and one of the most commonly used methods for generating smooth articulatory movements and uses only the target mean positions for synthesis. In the Blackburn and Young's model (Blackburn and Young, 2000), the target undershoots and overshoots are modelled using articulatory curvature information. In their model, new position distributions are generated using both target mean and variance. However, for synthesis, successive position means are linearly interpolated and the position variance is ignored. Though both techniques used for synthesis have the drawback of treating the targets as points rather than distributions, they serve as the means of testing the coarticulation and redundancy modelling theories in the planning stage of articulation. The framework generated by modelling the articulatory behaviour in the planning stage provides a foundation for building more complete models such as trajectory HMMs which use both target mean and variance for synthesis.

Methodology for building coarticulation models and trajectory generation is presented in Section 7.2. Implementation details are presented in section 7.4. Results obtained after trajectory synthesis in various feature spaces are presented and discussed in Section 7.5.

## 7.2 Method

The notation and terminology used in this section is identical to that used for describing the proposed algorithm in Chapter 3. Recall that the grand distribution for an articulator $i$, $\mathcal{N}(M_i, \Sigma_i)$, was estimated using data from all phones and represents that the articulator is unconstrained and hence associated with a larger covariance matrix. The phone-specific distribution of $i$ for a phone $\phi$, $\mathcal{N}(\mu_i^\phi, \Sigma_i^\phi)$ was estimated from the phone specific data. The model distribution, $\mathcal{N}(\hat{m}_i^\phi, \hat{S}_i^\phi)$ was estimated using articulatory constraint identification algorithm.

### 7.2.1 Target specification

Targets distributions were specified for every phone based on 5 hypotheses

1. *Conventional*: In the conventional approach, no explicit constraint information is used in modelling articulation. It is assumed that the statistics of each articulator capture the target information and variations due to coarticulation implicitly. Therefore, the phone-specific distributions were used for specification of articulatory targets for each phone. No role information was used here to derive any models of coarticulation.

2. *Relaxation/Baseline*: The model distributions derived from the articulatory role identification algorithm reflect the critical (fully constrained), dependent (correlations incorporated through D-step) and redundant (grand configuration) roles. Here, the knowledge of articulatory constraints are implicitly incorporated but no explicit models of coarticulation are derived. Therefore, the redundant degrees of freedom of non-critical articulators are not subject to context sensitive effects. This hypothesis is also referred as *baseline* since it serves as a means for validating the model distributions derived from the algorithm in comparison with other theories.

The following three hypotheses for modelling coarticulation use the knowledge of constraints in the form of critical, dependent and redundant roles to explicitly model different aspects of coarticulation.

3. *Redundancy modelling*: This approach uses model distributions and the knowledge of critical, dependent and redundant roles explicitly in deriving target distributions which are compact and informationally efficient from the economy of effort concept. This approach is referred as *redundancy modelling* since the redundant degrees of freedom are modelled explicitly under the assumption of efficient target planning.

4. *Anticipatory*: Temporal coarticulation on the current phone caused by the following critical phone was modelled using articulatory roles in the planning stage. Target representations were generated by applying the anticipation principle on model distributions.

5. *Carry forward*: Temporal coarticulation on the current phone due to the preceding phone was modelled using the knowledge of articulatory roles. Target distributions were derived by applying the carry forward concept on the model distributions .

Consider a triphone sequence $\{\phi_1, \phi_2, \phi_3\}$ at times $\{t_1, t_2, t_3\}$. In the *conventional* case, no role information is used and the target representations for phones $\phi_1$, $\phi_2$ and $\phi_3$ are defined using their respective phone-specific distributions $\mathcal{N}(\mu_i^{\phi_1}, \Sigma_i^{\phi_1})$, $\mathcal{N}(\mu_i^{\phi_2}, \Sigma_i^{\phi_2})$ and $\mathcal{N}(\mu_i^{\phi_3}, \Sigma_i^{\phi_3})$. Since only mean values are used by the execution stage for trajectory synthesis, the targets are represented using phone specific means, $\mu_i^{\phi_1}$, $\mu_i^{\phi_2}$ and $\mu_i^{\phi_3}$. For the rest of this section, the targets are described using the means rather than the distributions.

The model distributions estimated in the process of identification of roles capture the constraints of speech production system implicitly. Recall from Chapter 3 that the model distribution of a critical articulator is equivalent to its phone-specific distribution; the model distribution for a dependent articulator is estimated from D-step conditioned on the positions of the critical articulator(s); the model distribution of a redundant articulator remains equal to the grand distribution. The *baseline* hypothesis also called *relaxation* uses model distributions as targets. Therefore, for an articulator $i$ the target for $\phi_1$ is defined as $\hat{m}_i^{\phi_1}$ ,for $\phi_2$ as $\hat{m}_i^{\phi_2}$, and for $\phi_3$ as $\hat{m}_i^{\phi_3}$.

### Modelling temporal coarticulation

Let an articulator $i$ be critical (C) for phones $\phi_1$ and $\phi_3$ at times $t_1$ and $t_3$, and redundant (R) for phone $\phi_2$ at time $t_2$ as shown in Figure 7.2. This triphone sequence is denoted as CRC (critical-redundant-critical) in terms of articulatory roles. The articulator $i$ is redundant for phone $\phi_2$ and hence is prone to coarticulation due to the neighbouring critical phone. In the *anticipatory* case, the target position required for phone $\phi_3$ at $t_3$ can be anticipated as early as $t_2$ due to the redundant nature of $i$ for $\phi_2$. Therefore, the target position for $\phi_2$ becomes $\hat{m}_i^{\phi_3}$, and this represents the anticipation of the target position for $\phi_3$ as early as $\phi_2$. Similarly, in the *carry forward* case, the target position of $\phi_1$ is carried forward onto $\phi_2$, and the target position for $\phi_2$ is specified as $\hat{m}_i^{\phi_1}$.

The target position of a dependent articulator is controlled partially by the critical articulators with which it is correlated. The remaining degrees of freedom are redundant and prone to coarticulation. Let the articulator $i$ be critical (C) for phones $\phi_1$ and $\phi_3$ at times $t_1$ and $t_3$, and dependent (D) for phone $\phi_2$ at time $t_2$ as shown in Figure 7.3. The target distributions of $i$ for $\phi_1$, $\phi_2$ and $\phi_3$ are initialised to baseline configuration as before. The model distribution $\mathcal{N}(\hat{m}_i^{\phi_2}, \hat{S}_i^{\phi_2})$ of dependent articulator $i$ is estimated from the D-step conditioned on the positions of critical articulators for $\phi_2$ and the correlations between them. Under *anticipatory* coarticulation hypothesis, the remaining degrees of freedom of a dependent articulator are prone to coarticulation due to neighbouring critical phone $\phi_3$. The mean of $i$ for $\phi_2$ is modelled as $(\hat{m}_i^{\phi_2} + \hat{m}_i^{\phi_3})/2$, which represents a tug of war between its dependent behaviour and the coarticulation due to the neighbours. In *carry forward* case, the coarticulation due to preceding phone $\phi_1$ is modelled in a similar way resulting in the mean $(\hat{m}_i^{\phi_1} + \hat{m}_i^{\phi_3})/2$.

| left context | current | right context |
|---|---|---|
| C | D or R | C |
| D | D or R | C |
| R | D or R | C |
| C | D or R | D |
| C | D or R | R |

Table 7.1: *Role based triphone contexts (C is critical, D is dependent, R is redundant) considered for modelling redundancy, anticipatory and carry forward coarticulation effects in the planning stage.*

Table 7.1 shows different triphone contexts considered for redundancy modelling, anticipatory and carry forward coarticulation on dependent and redundant phones. Tables 7.3 and 7.2 show target distributions derived from different hypothesis for these triphone contexts for dependent and redundant articulators respectively.

### Modelling redundancy

Coarticulation due to minimisation of effort in the planning stage was modelled using linear interpolation. Under *redundancy modelling* hypothesis, the articulator $i$ tends

to minimise effort when moving from target for $\phi_1$ to $\phi_3$ via $\phi_2$. This minimal effort movement is modelled by linearly interpolating between the target positions for $\phi_1$ and $\phi_3$. For triphone contexts, the interpolated mean at time $t_2$ for phone $\phi_2$ is given as

$$m_{\mathrm{I}}^{\phi_2} = \hat{m}_i^{\phi_1} + \frac{\left(\hat{m}_i^{\phi_3} - \hat{m}_i^{\phi_1}\right)(t_2 - t_1)}{t_3 - t_1} \tag{7.1}$$

Under *redundancy modelling*, the target distribution of a redundant articulator is specified as $\mathcal{N}(m_{\mathrm{I}}^{\phi_2}, \hat{S}_i^{\phi_2})$, where the target mean position is set to the interpolated mean. For a dependent articulator, the mean position is a trade-off between the model mean and the interpolated mean, i.e., $(m_{\mathrm{I}}^{\phi_2} + \hat{m}_i^{\phi_2})/2$. Redundancy modelling hypothesis was applied to all triphone contexts in Table 7.1. The following quadphone contexts were also considered

- C D D C
- C D R C
- C R D C
- C R R C

where phones $\phi_1$ and $\phi_4$ are critical at times $t_1$ and $t_4$ and phones $\phi_2$ and $\phi_3$ are non-critical (D/R) at times $t_2$ and $t_3$ respectively. Here, the interpolated mean at time $t_2$ for phone $\phi_2$ is calculated as

$$m_{\mathrm{I}}^{\phi_2} = \hat{m}_i^{\phi_1} + \frac{\left(\hat{m}_i^{\phi_4} - \hat{m}_i^{\phi_1}\right)(t_2 - t_1)}{t_4 - t_1} \tag{7.2}$$

and at time $t_3$ for $\phi_3$ is calculated as

$$m_{\mathrm{I}}^{\phi_3} = \hat{m}_i^{\phi_1} + \frac{\left(\hat{m}_i^{\phi_4} - \hat{m}_i^{\phi_1}\right)(t_3 - t_1)}{t_4 - t_1} \tag{7.3}$$

Table 7.4 shows the resultant target distributions after applying redundancy modelling to C * * C context. As before, the target mean of a dependent articulator is the average between the model and the interpolated mean and that of redundant articulator is set to the interpolated mean.

## 7.3 Trajectory generation

The techniques used for synthesis are

- simple linear interpolation
- Blackburn and Young's model of coarticulation (Blackburn and Young, 2000)

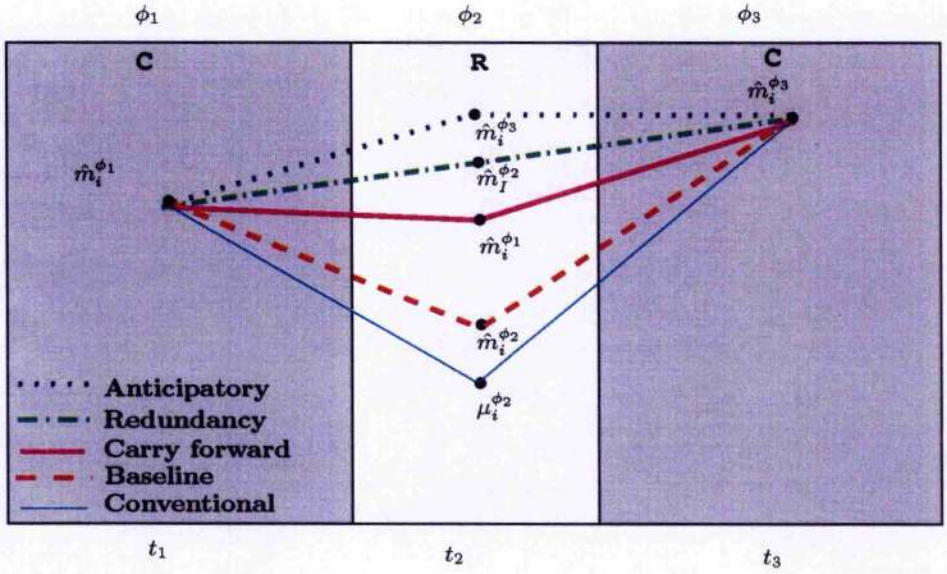



Figure 7.2: *Target specification for articulator i in critical-redundant-critical (CRC) context for phones $\phi_1$, $\phi_2$, $\phi_3$ at midpoint locations $t_1$, $t_2$ and $t_3$ generated using conventional (cyan, thin), baseline (red, dashed), redundancy modelling (green,dot dashed), anticipation (blue,dotted) and carry forward (magenta, solid) approaches.*

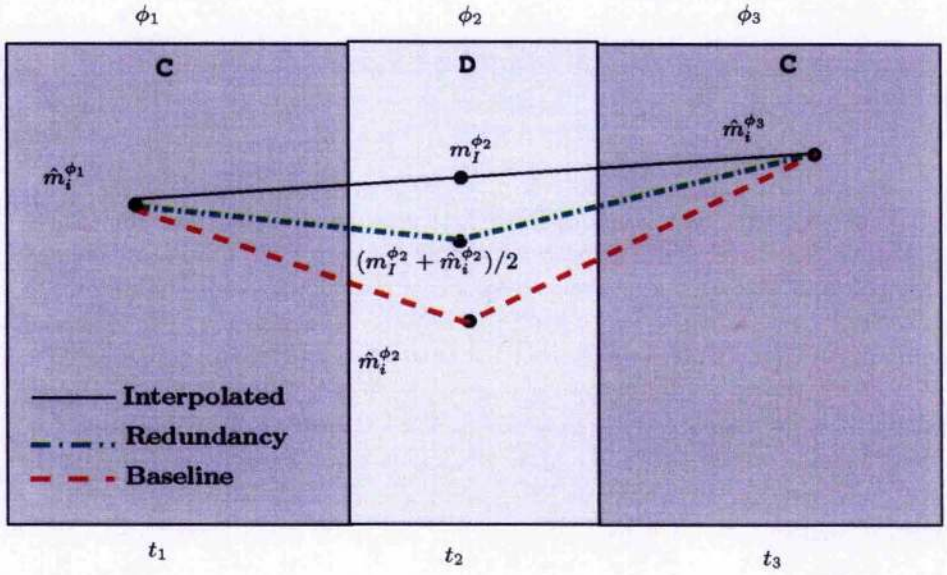



Figure 7.3: *Target specification for articulator i in critical-dependent-critical (CDC) context for phones $\phi_1$, $\phi_2$, $\phi_3$ at midpoint locations $t_1$, $t_2$ and $t_3$. Baseline (red, dashed) and redundancy modelling (green,dot dashed) approaches for dependent target specification are illustrated along with the interpolated mean $m_I^{\phi_2}$.*

Table 7.2: *Target specification for redundant articulator i in different triphone contexts where at least one of the neighbouring roles is critical.*

| Method | Phone $\phi_1$ at $t_1$ | Phone $\phi_2$ at $t_2$ | Phone $\phi_3$ at $t_3$ |
|---|---|---|---|
| Conventional | $\mu_i^{\phi_1}$ | $\mu_i^{\phi_2}$ | $\mu_i^{\phi_3}$ |
| Baseline | $\hat{m}_i^{\phi_1}$ | $\hat{m}_i^{\phi_2}$ | $\hat{m}_i^{\phi_3}$ |
| Anticipation | $\hat{m}_i^{\phi_1}$ | $\hat{m}_i^{\phi_3}$ | $\hat{m}_i^{\phi_3}$ |
| Carry-forward | $\hat{m}_i^{\phi_1}$ | $\hat{m}_i^{\phi_1}$ | $\hat{m}_i^{\phi_3}$ |
| Modelling redundancy | $\hat{m}_i^{\phi_1}$ | $m_{\mathrm{I}}^{\phi_2}$ | $\hat{m}_i^{\phi_3}$ |

Table 7.3: *Target specification for dependent articulator i in different triphone contexts where at least one of the neighbouring roles is critical.*

| Method | Phone $\phi_1$ at $t_1$ | Phone $\phi_2$ at $t_2$ | Phone $\phi_3$ at $t_3$ |
|---|---|---|---|
| Conventional | $\mu_i^{\phi_1}$ | $\mu_i^{\phi_2}$ | $\mu_i^{\phi_3}$ |
| Baseline | $\hat{m}_i^{\phi_1}$ | $\hat{m}_i^{\phi_2}$ | $\hat{m}_i^{\phi_3}$ |
| Anticipation | $\hat{m}_i^{\phi_1}$ | $(\hat{m}_i^{\phi_2} + \hat{m}_i^{\phi_3})/2$ | $\hat{m}_i^{\phi_3}$ |
| Carry-forward | $\hat{m}_i^{\phi_1}$ | $(\hat{m}_i^{\phi_1} + \hat{m}_i^{\phi_2})/2$ | $\hat{m}_i^{\phi_3}$ |
| Modelling redundancy | $\hat{m}_i^{\phi_1}$ | $(m_{\mathrm{I}}^{\phi_2} + \hat{m}_i^{\phi_2})/2$ | $\hat{m}_i^{\phi_3}$ |

Simple linear interpolation (LINT) is one of the simple and most commonly used techniques to synthesise smooth and blended movements of articulators. Successive mean positions of all phones in an utterance are linearly interpolated to generate trajectories from target distributions. The targets are treated as points (target means) for implementing linear interpolation. The target variance is ignored which is a short coming of this approach.

Blackburn and Young's coarticulation model uses articulatory positions and curvatures for modelling context sensitive effects caused by the immediate neighbours on the current phone. Articulatory curvature for the current phone is calculated as the difference between the accelerations from the previous and the following phones. Position and curvature means and variances were estimated for each phone from the target means. The methodology for this coarticulation model is given in (Blackburn and Young, 2000). New position distributions were derived for each phone in the utterance conditioned on the position and curvature statistics. Though Blackburn and Young's coarticulation model generates new position distributions by modelling context sensitive effects in the execution phase, the trajectories are generated only from the new position means. The position variance information is also ignored in trajectory synthesis. The two techniques used for trajectory generation presented above utilise the target mean information for synthesis. Nevertheless, target representations in the form of statistical distributions can be input to models which use both target mean and variance such as, MLPG algorithm (Tokuda et al., 2000), trajectory HMMs (Tokuda et al., 2007) for trajectory synthesis. The following section, Section 7.4, presents the implementation details of the planning and execution phases.

Table 7.4: *Target specification for articulator i under redundancy modelling hypothesis for a sequence of 4 phones where the role played by the articulator is critical for first and last phones.*

| Phone $\phi_1$ at $t_1$ | Phone $\phi_2$ at $t_2$ | Phone $\phi_3$ at $t_3$ | Phone $\phi_4$ at $t_4$ |
|---|---|---|---|
| C | D | D | C |
| $\hat{m}_i^{\phi_1}$ | $(m_{\mathrm{I}}^{\phi_2} + \hat{m}_i^{\phi_2})/2$ | $(m_{\mathrm{I}}^{\phi_3} + \hat{m}_i^{\phi_3})/2$ | $\hat{m}_i^{\phi_4}$ |
| C | D | R | C |
| $\hat{m}_i^{\phi_1}$ | $(m_{\mathrm{I}}^{\phi_2} + \hat{m}_i^{\phi_2})/2$ | $m_{\mathrm{I}}^{\phi_3}$ | $\hat{m}_i^{\phi_4}$ |
| C | R | D | C |
| $\hat{m}_i^{\phi_1}$ | $m_{\mathrm{I}}^{\phi_2}$ | $(m_{\mathrm{I}}^{\phi_3} + \hat{m}_i^{\phi_3})/2$ | $\hat{m}_i^{\phi_4}$ |
| C | R | R | C |
| $\hat{m}_i^{\phi_1}$ | $m_{\mathrm{I}}^{\phi_2}$ | $m_{\mathrm{I}}^{\phi_3}$ | $\hat{m}_i^{\phi_4}$ |

## 7.4 Implementation

The articulatory role information was extracted using

- IPA chart: critical, noncritical (redundant) roles
- ACIDA: critical, dependent and redundant roles
  - the IPA level of complexity, where the average number of critical dimensions per phone equals to that calculated from IPA chart
  - the 2×IPA level of complexity, where the average number of critical dimensions per phone equals twice the number of IPA critical dimensions

IPA model distributions were derived from the knowledge of critical dimensions from the IPA chart. The IPA model distributions were subject to the dependent update step (D-step) to obtain IPA+D model statistics. The ACIDA model distributions were estimated using the proposed algorithm at the IPA and the 2xIPA levels of complexity. The target distributions were derived from the knowledge of the articulatory roles and the corresponding model distributions. Trajectories were estimated from target distributions in raw (measured), PCA (PC1, PC3 , PC4, PC5 and PC7) and LDA (LD1, LD3, LD4, LD5 and LD7) based articulatory spaces using linear interpolation and Blackburn and Young's model. All synthetic trajectories were filtered using a zero phase order 10 lowpass filter at 20Hz sampling frequency. The trajectories in PCA and LDA spaces were mapped back to the articulatory space for comparison with the measured trajectories. Synthetic trajectories were generated for 459 utterances of a total of 460 sentences. The measurements from one sentence (268) were corrupt and were excluded from evaluation. Results of trajectory generation experiments for both male and female speakers are presented in the following sections.

## 7.5 Results

Three kinds of evaluation measured were used to analyse the goodness of fit of the synthetic trajectories generated using various methods to the measured trajectory

- Correlation
- Root Mean Squared Error (RMSE)
- Normalised RMSE (RMSE values normalised by grand standard deviations)

Correlation and RMSE values estimated from positions provide a way of comparing the performance of the synthetic trajectories in an objective and quantitative way. Correlation between measured and synthetic trajectories were computed and averaged across all sentences and articulators to obtain mean correlation value. Normalised RMSE values were obtained by normalising RMSE of each articulator by its grand standard deviation. Mean RMSE and normalised RMSE values were calculated by averaging across all sentences and articulators. The values of correlation, RMSE and normalised RMSE values painted a similar picture. Therefore, the rest of the analysis is presented using correlation as evaluation metric. Results from RMSE and normalised RMSE at the IPA level of complexity are presented in Tables C.47 (male), C.48 (female), C.49 (male), C.50 (female). Results from RMSE and normalised RMSE at the 2×IPA level of complexity are presented in Tables C.51 (male), C.52 (female), C.53 (male), C.54 (female).

### 7.5.1 Choice of constraints

The performance of the models obtained from critical coordinates derived from the IPA chart and the ACIDA algorithm were compared for all hypotheses. Tables 7.5 and 7.6 show the mean correlation computed for male and female speakers using linear interpolation (LINT) and Blackburn and Young's (BY) model respectively. The significance of the difference between correlation values was computed using t-test at level of significance $\alpha = 0.05$. The models derived from ACIDA at IPA level of complexity outperformed the models derived from the IPA chart for both male and female speakers. For example, under the baseline hypothesis, the correlation between the measured trajectories and the synthetic trajectories derived from the IPA chart for male speaker using linear interpolation was 0.29, where as the correlation obtained from ACIDA models at the IPA level of complexity was 0.61. The difference between the mean correlations was found to be significant from the t-test results. The results from Blackburn and Young's model shown in Table 7.6 also painted a similar picture. Constraints in the form of critical, dependent and redundant roles given by the ACIDA algorithm capture the constraints and the degrees of freedom of articulators during the production of phones more efficiently than typical critical/non-critical constraints derived from the IPA chart.

The IPA+D models were obtained by updating the IPA model distributions using dependency update step (D-step) from ACIDA algorithm (see Section 3.3.2). Using

the D-step improved the performance of the models trained using the IPA information. For example, for the female speaker, the baseline correlation improved from 0.28 to 0.56 for the LINT method. The correlation, RMSE and normalised RMSE values from the ACIDA models were better than the IPA+D models by a small margin (5%) for the male speaker. For example, the normalised RMSE for baseline hypothesis for male speaker was found to be 0.84 for the IPA+D models and 0.80 for the ACIDA models. For the female speaker, the correlations obtained from the IPA+D models were slightly better (2%- 3%) than the ACIDA models for both LINT and BY methods. However, in all cases, the difference between IPA+D and ACIDA models was found to be statistically insignificant. Therefore, the performance of IPA+D models was comparable to that of the ACIDA models at the same level of complexity.

To summarise, the results showed that target distributions derived using articulatory constraints in the form of critical, dependent and redundant roles yield better results than those estimated from constraints derived from IPA chart which show only critical/non-critical discrimination. Also incorporating the D-step improved performance of the IPA models significantly.

| | male | | | female | | |
|---|---|---|---|---|---|---|
| | **IPA** | **IPA+D** | **ACIDA$_{IPA}$** | **IPA** | **IPA+D** | **ACIDA$_{IPA}$** |
| Baseline | 0.29 | 0.57 | 0.61 | 0.28 | 0.56 | 0.55 |
| Modelling redundancy | 0.31 | 0.59 | **0.62** | 0.30 | **0.57** | 0.56 |
| Anticipatory | 0.30 | 0.59 | **0.62** | 0.30 | **0.57** | 0.55 |
| Carry forward | 0.26 | 0.56 | 0.59 | 0.25 | 0.54 | 0.52 |

Table 7.5: *Mean correlation between measured trajectories and synthetic trajectories generated in raw articulatory space from IPA, IPA+D, and ACIDA$_{ipa}$ model distributions for baseline, redundancy modelling, anticipatory, carry forward hypotheses using linear interpolation. Correlations are averaged across all sentences and articulators.*

| | male | | | female | | |
|---|---|---|---|---|---|---|
| | **IPA** | **IPA+D** | **ACIDA$_{IPA}$** | **IPA** | **IPA+D** | **ACIDA$_{IPA}$** |
| Baseline | 0.32 | 0.63 | **0.67** | 0.32 | **0.61** | 0.59 |
| Modelling redundancy | 0.32 | 0.62 | 0.66 | 0.30 | 0.59 | 0.58 |
| Anticipatory | 0.30 | 0.62 | 0.65 | 0.31 | 0.60 | 0.57 |
| Carry forward | 0.28 | 0.61 | 0.64 | 0.27 | 0.58 | 0.56 |

Table 7.6: *Mean correlation between measured trajectories and synthetic trajectories generated in raw articulatory space from IPA, IPA+D, and ACIDA$_{ipa}$ model distributions for baseline, redundancy modelling, anticipatory, carry forward hypotheses using Blackburn and Young's model. Correlations are averaged across all sentences and articulators.*

### 7.5.2 Evaluation of hypotheses

Different theories of articulatory behaviour implemented in the planning stage of speech production were evaluated using correlation, RMSE and normalised RMSE values computed between measured and synthetic trajectories. Comparisons were also made across different levels of complexity (IPA and 2×IPA) and different trajectory generation techniques. Statistical significance of the difference between the evaluation measures given by various hypotheses and methods was computed using student's t-test at level of significance ($\alpha = 0.05$). Correlations averaged across all sentences and articulators at the IPA and the 2×IPA level of complexities for both male and female speakers are shown in Figures 7.4 and 7.5 respectively. The patterns of correlation, RMSE and normalised RMSE values were similar for both speakers. Hence, the rest of the section is presented using mean correlation values.

#### IPA level of complexity

The performance of baseline, redundancy modelling, anticipatory and carry forward theories for the **linear interpolation** method at IPA level of complexity are compared first. The highest mean correlation between the measured trajectories and synthetic trajectories was given by redundancy modelling, 0.65 in LD4 feature space for the male speaker and 0.60 in LD1, LD4 and LD5 feature spaces for the female speaker. The lowest correlation value was given by carry forward hypotheses, 0.53 in PC4 and PC5 feature spaces for the male speaker and 0.51 in PC1 space for the female speaker. Across all feature spaces, the redundancy modelling hypothesis gave the best values of correlation, RMSE and NRMSE whereas the worst performance was given by the carry forward hypothesis. The trajectories generated using baseline and anticipatory hypotheses yielded similar results. Statistical significance test showed no significant difference between the performance of baseline, redundancy modelling, anticipatory and carry forward hypotheses in most of the feature spaces since the results differed from one another by a slight margin. Only one significant difference was found in LD1 feature space for both speakers between redundancy modelling (0.64/0.60-m/f) and carry forward (0.53/0.53-m/f) models. Though the performance of LDA based models was better than PCA based models, the improvement was statistically insignificant.

Using **Blackburn and Young's model** for trajectory generation improved the fit of the synthetic trajectories to the measured trajectory by 9 to 10%. The best performance was given by the baseline models, followed by the redundancy modelling. The performance of both anticipatory and carry forward models was worse than others. Spreading the target positions in forward and backward directions affects the estimation of curvature values used for updating the position distributions of the articulators. Any loss in curvature information due to target spreading could contribute to the poorer performance of temporal coarticulation models over the baseline models. Similar to the results from LINT method, no statistically significant difference in performance was found between the models from different hypotheses for most of the feature spaces. For the male speaker, the baseline performance was better than anticipatory and carry forward performances and the improvement was statistically significant only in the LD1

feature space. For the female speaker, only the improvement given by the baseline over the carry forward models was statistically significant also in the LD1 feature space.

At the IPA level of complexity, the difference between the performances of the LINT and the BY models under each hypothesis was evaluated for statistical significance. The results showed that improvements given by the BY models over the LINT models were not statistically significant in any feature space for any hypothesis. This could be due to the bias form the correlation values obtained from redundant articulators for which the curvature estimate is zero. The BY method reduces to a simple linear interpolation when there are no curvature estimates.

The average number of critical dimensions/modes per phone was 1.8 at the IPA level of complexity and the resulting distributions were very sparse. Any improvements given by redundancy modelling hypothesis and LDA based representations were found to be statistically insignificant. The best representation at the IPA level of complexity was also found to be lossy (8%/11%-m/f) when compared with the performance of the conventional models. The conventional models use phone-specific distributions for generating target representations for each articulator and are not parsimonious when compared with the models trained at the IPA level of complexity. In the next stage of analysis, the complexity of the models was increased and the performance of the resulting trajectories was evaluated in comparison with the conventional approach.

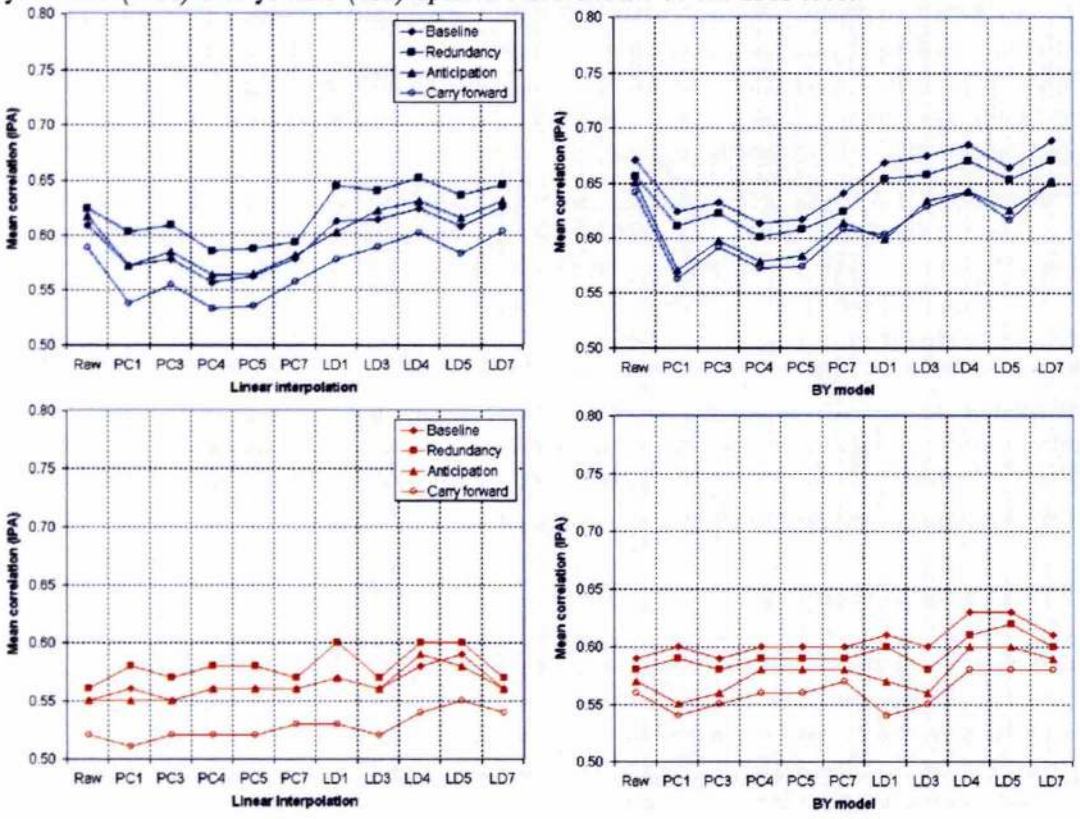
#### 2×IPA level of complexity

To capture more detail, the critical threshold was lowered to 2×IPA level of complexity where the average number of critical dimensions per phone were doubled. Increasing the complexity of the models improved the performance by 7 to 10% over the performance at the IPA level. The improvements were found to be statistically significant for most PC and LD based representations for both LINT and BY methods.

For both male and female speakers, the best performance was given by redundancy modelling followed by anticipation and baseline methods for linear interpolation. For BY model, baseline models gave the best performance, followed by redundancy modelling and anticipatory models. The carry forward approach gave the worst performance for both LINT and BY models. No statistically significant differences was found between the performances of baseline, redundancy modelling, anticipatory and carry forward approaches, since the differences were of small magnitude.

The performance of trajectories estimated using the proposed hypotheses was compared with that of **conventional model** at 2×IPA level of complexity. For the linear interpolation, comparable performances were achieved by redundancy modelling, anticipation and baseline approaches when compared with the conventional approach, any differences found were minor and statistically insignificant for both speakers. For the BY model, the performance of only the baseline approach was comparable to that of the conventional models for the male speaker. For the female speaker, the conventional model performed better than the rest of the models and the improvement was also statistically significant. Loss of articulatory curvature information in completely redundant contexts is the main reason for the poor performance of the proposed role based models when compared with the conventional models.

Figure 7.4: *Mean correlations (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using various hypothesis for male speaker for linear interpolation and Blackburn and Young (BY) model. Results for male (blue) and female (red) speakers are shown at the IPA level.*

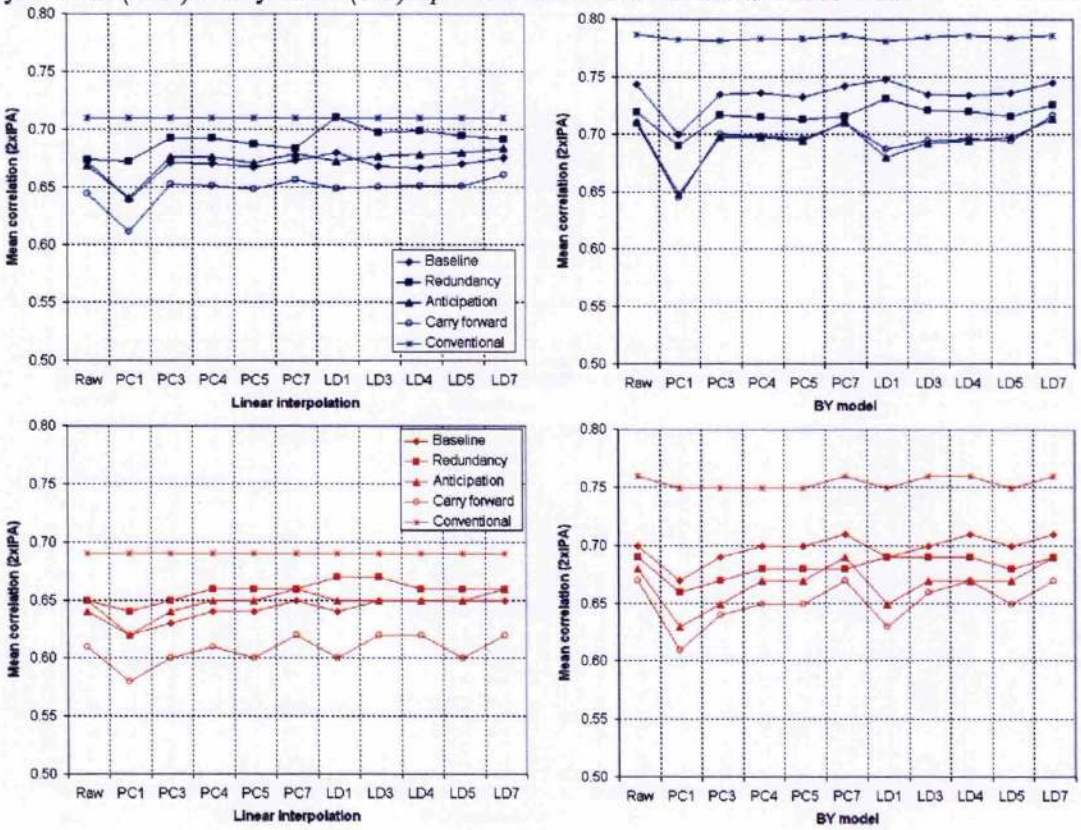



## 7.6 Discussion

Different coarticulation hypotheses were evaluated by generating trajectories from the target representations derived from the respective coarticulation models. The results obtained from trajectory generation experiments showed that the constraints in the form of critical, dependent and redundant roles are better representations than the critical/non-critical constraints derived from IPA chart. Also increasing complexity to 2×IPA level significantly improved the performance of the models. The results were inconclusive when different hypotheses and feature spaces were compared to find the best approach and feature representation respectively. Though redundancy modelling (for LINT) and baseline (for BY) representations gave the best values of correlation, RMSE and normalised RMSE, their performance was not significantly different from that given by other hypotheses.

There were instances where the proposed coarticulation models better modelled the articulatory behaviour than the conventional approach. Figures 7.6 and 7.7 show two such examples where the proposed models performed better than conventional models when velocity correlations were compared. Figure 7.6 shows the measured trajectory

Figure 7.5: *Mean correlations (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using various hypothesis for male speaker for linear interpolation and Blackburn and Young (BY) model. Results for male (blue) and female (red) speakers are shown at the 2× IPA level.*




and the synthetic trajectory generated under the redundancy modelling criterion in raw articulatory feature space using LINT for $\text{TT}_x$ for male speaker. It can be observed that the trajectory generated using redundancy modelling shows some potential in matching the path of $\text{TT}_x$ movements than the conventional trajectory. Correlation value computed in the *velocity space* between actual and synthetic trajectories for $\text{TT}_x$ for this sentence showed that the redundancy modelling (0.76) better captures the articulatory behaviour than the conventional model (0.49). Here, the improvement in the positional correlation given by the redundancy modelling over the conventional model was 19%.

Nasalisation of vowels due to context sensitivity was better captured by the anticipatory hypothesis when compared with the conventional approach for some utterances. Figure 7.7 shows one such example where nasalisation of vowels [ə] in "an" and [i] in "immediate" was better captured by anticipatory models than conventional models. Correlation in the velocity space was improved by 60% when the anticipatory models were used over the conventional models, whereas similar position correlation values were given by both anticipatory and conventional models.

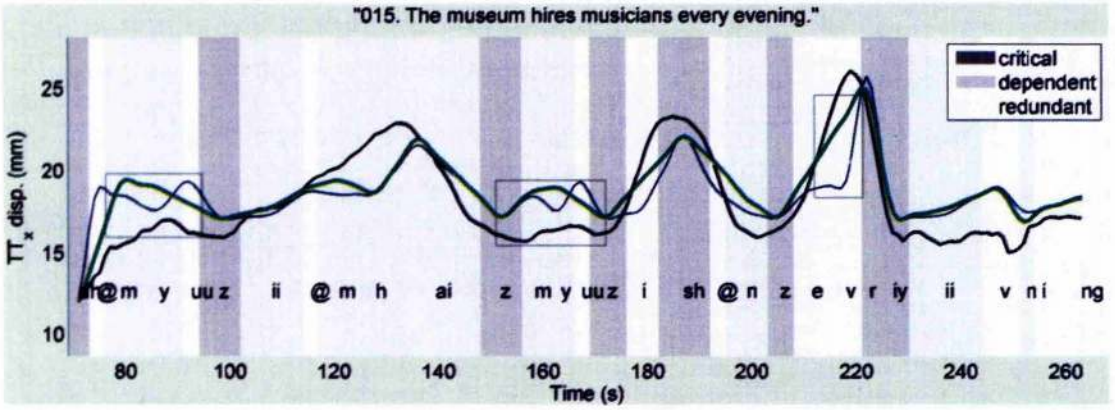



Figure 7.6: *Measured ( black) and synthetic trajectories generated from conventional (thin blue) and redundancy modelling (thick green) hypotheses for the male speaker at the 2×IPA level of complexity in the raw articulatory space.*

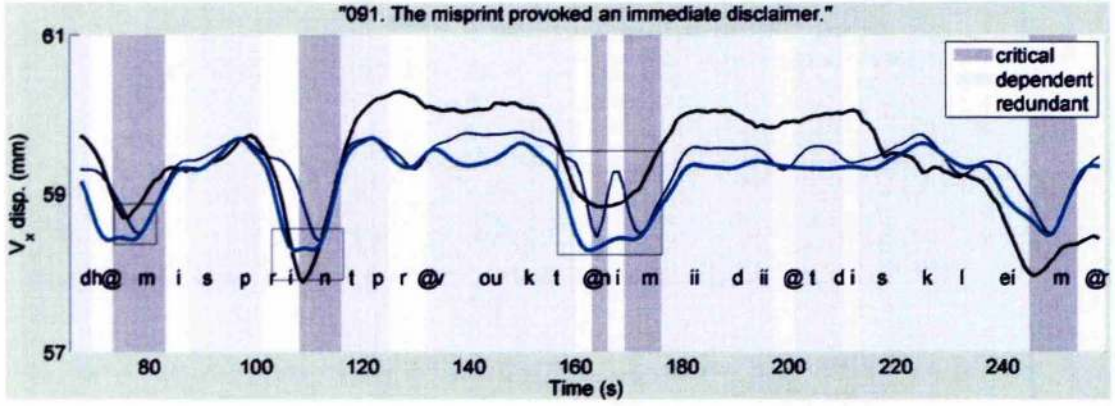



Figure 7.7: *Measured (black) and synthetic trajectories (solid) generated from conventional (thin blue) and anticipatory (thick cyan) hypotheses for the male speaker at the 2×IPA level of complexity in the raw articulatory space.*

Carry forward coarticulation models gave poor performance across all feature spaces and methods. The degradation in performance caused by carry forward models when compared with redundancy modelling (LINT) and baseline (BY) methods was statistically significant in LD1 feature space for both speakers and at both levels of critical threshold. Carry forward coarticulation is a low level phenomenon resulting from the inertia of articulators and is more relevant to the execution stage than planning stage. Therefore, models built in the planning stage have performed poorly. The performance of models could be improved by modelling carry forward coarticulation in the execution phase than in planning phase.

The trajectory generation techniques used for this analysis ignore the target variance which is a valuable source of information. There is a need for using a trajectory generation technique which utilises both target mean and variance information for synthesising the articulatory movements from target distributions to be able to fully evaluate the

potential of the coarticulation approaches. Precise spatial targets do not exist for all articulators and speech sounds. Different factors such as coarticulation, phone duration, articulatory inertia contribute to the variation in the target position. A more realistic model would require target specifications in the form of statistical windows rather than spatial points. The present approach also ignores the duration of phones when modelling coarticulation and limits the span to immediate neighbours. The models could be further improved by incorporating durational modelling and wider spans of coarticulation in the structure.

The experiments in this work were performed on articulatory data alone. The best RMSE value obtained from the experiments was 1.55mm for BY model in LD7 space for baseline hypothesis for male speaker (1.68 for LINT in LD1 for redundancy modelling). The best RMSE value reported when acoustic information was combined with articulatory information for trajectory generation was 1.40mm for female speaker (Richmond, 2007b). There are some known problems with flesh point calibration (Frankel, 2003; Richmond, 2007a) in MOCHA-TIMIT recordings. Some of the transcription errors have been corrected and corrupt measurements were ignored (Chapter 2) to improve the performance of the models and to obtain good lists of critical, dependent and redundant articulators. According to Richmond (2009), the average RMSE value dropped from 1.54mm to 0.99mm when recordings from a better EMA resource were used. Improvements could be made with further recordings of speech articulation, e.g., new capture techniques, larger corpora, multiple subjects and various speaking styles.

When comparisons were made across different feature spaces, positional correlations from LDA models were higher than PCA based models. However, no significant difference was found between LDA, PCA and raw feature spaces. But the models in some feature spaces were found to be more compact representations than the others. The following section presents analysis on compactness of models.

### 7.6.1 Compactness of the models

In the *conventional* approach to estimating model distributions, neither the knowledge of articulatory roles nor the interdependencies are considered, and model distributions are simply set equal to the phone specific distributions. Recall from Section 6.5 that if $a$ (14) represents the number of modes, the number of parameters (means and variances) needed for modelling $\varphi$ phones in the *conventional* approach is $2a\varphi$.

When the proposed algorithm (ACIDA) is used for estimating the model distributions in each feature space, the number of parameters required for estimating models is given according to Eq. 6.1 in Section 6.5 as

$$2a + \varphi\left(2\hat{k}^{\phi} + \frac{\hat{k}^{\phi}(\hat{k}^{\phi}-1)}{2}\right)$$

The reductions in number of parameters over conventional models reported at the IPA level of complexity in Section 6.5 were: 82% for PC1/LD1, 78% for PC3/LD3 and PC4/LD4, 77% for PC5/LD5 and 76% for PC7/LD7 and raw articulatory features.

At the 2×IPA level of complexity, the reduction percentage slightly lowered due to increase in model complexity when compared with that at the IPA level. Nevertheless, the models were still parsimonious when compared with conventional approach. The raw articulatory representation required 49% less parameters than conventional approach. The performance of PCA and LDA representations was as follows: 56% for PC1/LD1, 51% for PC3/LD3, 51% for PC4/LD4, 50% for PC5/LD5 and 50% for PC7/LD7 and raw articulatory features.

## 7.7 Conclusion

A statistical framework for modelling coarticulation in the planning stage was presented. Different aspects of coarticulation were modelled using the ACIDA model distributions in various feature representations and the performance of the resulting models was tested by generating trajectories. The span of coarticulation was limited to the immediate neighbours. Constraints in the planning stage derived from the IPA chart were compared with those from the proposed role identification algorithm. Synthetic trajectories were generated in raw, different PCA and LDA based representations at two levels of complexity, IPA and 2×IPA. Simple linear interpolation and Blackburn and Young's model were used for synthesis. Positional correlation and RMSE were used for evaluating the goodness of fit of the synthetic trajectories to the measured trajectories. The results showed that the constraints in the form of critical, dependent and redundant articulators are better representations than those derived from phonological knowledge which make only critical/noncritical distinction. The redundancy modelling hypothesis for trajectory generation by the linear interpolation and the baseline hypothesis for trajectory generation by the Blackburn and Young's model gave the best performance. LDA based models performed better than PCA based models. However, any small improvements found were statistically insignificant. Carry forward coarticulation, a phenomenon commonly attributed to the inertia of the articulators, when modelled in the planning stage, gave poor performance. Increasing the complexity of the models to 2×IPA complexity gave significant improvement. The redundancy modelling and the anticipatory approaches were found to be better models of articulatory behaviour than the conventional models in certain contexts. The models estimated using the ACIDA were found to be compact when compared with conventional models.

The statistical framework in the planning stage of articulation presented in this chapter provides a basis for building more complex and accurate models of speech articulation. Proposed framework can be improved by incorporating longer spans of coarticulation and durational modelling. Simple trajectory generation techniques were used for testing the proposed framework. Limitations posed by trajectory generation techniques used in this work could be addressed by using more complete models which treat targets as statistical windows rather than points.

# Chapter 8

# Conclusion

In this chapter, the work presented in the course of this thesis is summarised. Implications of the approach are discussed. Contributions made to the field of speech research are presented along with the directions for future work.

## 8.1 Summary

Coarticulation, a direct consequence of the nature of speech production, affects the performance of recognition and synthesis systems. The state-of-the-art systems model coarticulatory effects on the surface and ignore the articulatory domain which is the source of coarticulation. There is a potential for improving coarticulation models from direct descriptions of the speech articulation. Previous approaches to modelling coarticulation in planning and execution phases along the same lines were presented and analysed. Models of coarticulation relied on phonological knowledge for incorporating the knowledge of articulation where measured data was unavailable. Representations such as distinctive binary features derived from phonological knowledge are limited in many ways and are poor representations of speech articulation. Gesture priorities in the form of critical/noncritical roles fail to identify and explain the constraints due to biomechanical links of articulators. On the other hand, statistical models trained on measured articulatory data, though powerful, are merely descriptive and ignore the cause of constraints. This thesis focuses on deriving better representations of constraints on articulators and thereby, building descriptive as well as explanatory models of speech articulation.

The articulatory constraint identification algorithm (ACIDA) which identifies and captures the constraints on articulators during speech production in a statistical way from measured articulatory data was presented. Articulatory data used in this work comprised of measurements of upper lip, lower lip, lower incisor, tongue tip, blade, dorsum and velum for 460 sentences from 2 speakers (MOCHA-TIMIT (Wrench, 2001)). The algorithm identifies critical, dependent and redundant roles which explain the fully constrained, partially constrained and totally redundant degrees of freedom of articulators and also estimates the respective distributions. Identification divergence which is the

statistical difference (KL divergence) between the grand distributions and the phone-specific distributions was used for identification of critical roles. Inter-articulatory correlations were used to identify and update the distributions of dependent articulators (D-step) which are partially constrained due to their relationships with the critical articulators. The distributions of redundant articulators were set to grand distributions which characterise their unconstrained nature. The 1D and 2D versions of the ACIDA algorithm were implemented and the results were analysed. Convergence scale computed between model and phone-specific distributions for 1D and 2D cases showed that the convergence of the models to the phone specific models improves as more number of critical dimensions were identified by lowering the critical threshold (from 5 to 0.1 in steps of 0.1). Evaluation scale was computed between model distributions (1D and 2D) and actual phone distributions (full phone covariance) across a similar range of thresholds. Evaluation scale values showed that 2D models which capture the correlations between x and y movements of articulators outperformed the 1D models which assume independence between x and y movements.

Lists of expected critical coordinates derived from the IPA chart were compared with the critical coordinates identified using the proposed algorithm for consonants, vowels and diphthongs. The analysis showed that the identified critical coordinates compared well to the expected critical coordinates for consonants. Some additional critical coordinates were identified for fricatives. No critical coordinates were identified for neutral vowel [ə] and for also other central and reduced vowels. Some substitutions occurred due to existence of strong correlations amongst articulators. Some inter-speaker differences were also found. Evaluation scales computed from model distributions from expected critical dimensions fitted poorly to the actual phone distributions when compared with those from the 1D and the 2D ACIDA model distributions. The performance of the models from the expected critical coordinates improved significantly when inter-articulatory dependencies were incorporated using D-step. The proposed algorithm was evaluated against an exhaustive search, where all critical articulator combinations were tested according to minimax criterion. The evaluation scale values showed that the proposed algorithm performed as well as exhaustive search models. However, exhaustive search approach was found to be very expensive computationally when compared with the ACIDA algorithm.

A statistical framework for building models of articulation from measured data using role information was derived in two stages. In the first stage, different articulatory feature spaces were generated aimed at obtaining compact and informationally rich representations which can be related to the independently moving articulatory components. Linear orthogonal transforms were employed to obtain compact and informationally efficient articulatory feature sets and the knowledge of inter-articulatory correlations was used for establishing independence between the articulatory coordinates. Different PCA and LDA based representations were derived by grouping strongly correlated articulators together thereby eliminating weak correlations amongst articulators. The ACIDA algorithm was used to identify the constraints on the articulatory gestures indicated by the PCA and LDA mode shapes. Critical modes from more compact models such as PC1/LD1 were difficult to interpret whereas those from less compact models such as PC7/LD7 were closely related to the identified critical coordinates in the raw articulatory space. Evaluation scale and recognition performance were used to

analyse the efficiency of the representations. LDA based models which emphasise the separability between phone classes performed better than PCA based models. Linear orthogonal transforms such as PC7/LD7 derived by eliminating important correlations amongst articulators gave no improvement over raw articulatory representation. The performance of PC3/LD3 models derived by ignoring weak and insignificant correlations amongst articulators was comparable to that of PC1/LD1 where all correlations are retained.

Different aspects of coarticulation were modelled using the knowledge of articulatory roles and the model distributions in different feature spaces. In the planning phase, the target distributions were estimated for each phone in the sequence using the model distributions under conventional, baseline, redundancy modelling, anticipatory and carry forward hypotheses. The span of coarticulation was limited to the immediate neighbours. Articulatory constraints were derived from the phonological knowledge (IPA chart) and the ACIDA algorithm. In the execution phase, articulatory trajectories were generated from the target representations using linear interpolation and Blackburn and Young's model of coarticulation (Blackburn and Young, 2000). Correlation, root mean squared error (RMSE) and normalised RMSE were used to evaluate the synthetic trajectories in comparison with the measured trajectories. The results showed that the trajectories estimated from the constraints derived from the ACIDA algorithm outperformed those derived from the IPA chart. Including the interarticulatory dependencies (D-step) in the estimation of model distributions the performance of the IPA models. Increasing the complexity of the ACIDA models from the IPA level to the 2×IPA level improved the performance significantly. Modelling redundancy using the concept of effort minimising behaviour of articulators gave the best performance for the linear interpolation method. The baseline approach gave the best performance for the Blackburn and Young's model. The LDA based models performed better than the PCA models. However, the improvements in all cases were of small magnitude and were found to be statistically insignificant. Carry forward coarticulation, a phenomenon occurring in the execution phase due to inertia of articulators, when modelled in the planning stage, degraded the performance of the models. The model of phone distributions obtained using the proposed algorithm, through recognition of articulatory roles, is shown to be more compact and more informative than a conventional statistical description. There is a scope for building better models of articulation by improving the framework in the planning stage and by addressing limitations in the execution phase.

To summarise,

- the proposed algorithm identifies and captures the constraints on the articulators in the form of critical, dependent and redundant roles in an entirely statistical and data-driven way.

- Identified constraints compare well with the expected constraints derived from the phonological knowledge. Identified constraints also captured speaker dependent behaviours, physiological links and provide a transformation from phonological to phonetic domain. The fit of the models to the measured distributions was better than that of the models derived from the IPA based constraints.

- Results from the proposed algorithm not only compared well to those from the exhaustive search but also had faster computation speeds.

- Linear orthogonal transforms such as PCA and LDA provide compact and informationally rich representations when compared with the raw articulatory feature set at the expense of loss of interpretation power.

- Articulatory constraints in the form of critical, dependent and redundant roles generated better models of coarticulation than those in the form of critical/noncritical priorities from the IPA chart.

- The proposed coarticulation models not only generated compact representations but showed some potential by capturing the behaviour of articulators closely than the conventional models in some cases. However, more complete models are essential to fully evaluate the efficiency of the proposed statistical framework in modelling coarticulatory effects.

## 8.2 Contribution

The main focus of this work was to build statistical models of speech articulation which reflect the nature of speech production and have the potential to model the coarticulatory effects. The primary contribution from this thesis is the articulatory constraint identification algorithm (ACIDA). The desirable features of the proposed algorithm and contributions from the undertaken work are discussed in detail below.

### 8.2.1 Nature of the constraints

In phonology theory, the place and manner of articulation for each phone are encoded in the form of discrete binary features (Chomsky and Halle, 1968) and the IPA chart can be viewed as a short cut representation depicting intersection of different features. It is difficult to transform phonological binary features to multi-valued commands for articulators, some attempts have resulted in knowledge-driven quantised articulatory configurations (Larar et al., 1988; Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000). The quantised representations fail to incorporate the variation in the target of articulator due to factors such as coarticulation, speaking rate, style and language. The proposed algorithm identifies the constraints on the articulators during production of each speech sound in the form of critical, dependent and redundant roles. It captures the essence of speech production by differentiating between tightly constrained articulators (critical), consequent movements of linked parts of the anatomy (i.e., dependent), and redundant parts that are most susceptible to the biomechanical effects of coarticulation from targets of neighbouring phonemes. The target models estimated using role information are specified as statistical windows which allow for all possible variations due to coarticulation.

### 8.2.2 Mapping from phonemes to phones

When mapping from phonological to phonetic domain, various factors such as language, speaker, style, rate, coarticulation result in different realisations of a speech sound. Such variations are denoted using diacritics in the 'narrow' transcriptions of speech. The proposed algorithm provides a mapping from phonological to phonetic domain by capturing the characteristics of typical phones within a language from the measured articulatory data. The critical articulators identified from the algorithm not only compared well to the expected constraints derived from IPA but also provided speaker-specific constraints due to variations in speaking styles.

### 8.2.3 Articulatory dependencies

Constraints derived from phonological knowledge mostly fall into critical/noncritical categories. For example, in feature based approach (Henke, 1965), only critical features are specified for each phone whereas the noncritical features remain unspecified. Articulators are linked biomechanically and their interdependencies need to be considered when specifying constraints. Incorporation of such inter-articulatory relationships reduces the degrees of freedom of articulators. In the coproduction theory, the linkages between articulators were incorporated in the structure of gestures in the form of passive gestures which undergo changes due to their relationship with the active gestures (Saltzman and Munhall, 1989). However, the gestures are heuristically scored into active, passive and inactive categories from the knowledge of phonology. The proposed algorithm makes use of grand inter-articulatory correlations to identify the dependent articulators and updates their distributions conditioned on the distributions of critical articulators with which they share strong and significant correlations. The D-step in the proposed algorithm performs the identification and estimation of relevant distributions using statistical methods. Experiments showed that the distributions updated using the knowledge of critical, dependent and redundant roles fitted well to the measured data than the model distributions derived from IPA knowledge. Incorporating articulatory correlations using D-step improved the fit of the IPA based models significantly.

### 8.2.4 Informational efficiency

The algorithm also provides compact and informationally efficient representations when compared with the conventional approach. With complexity equivalent to the IPA descriptions, the reductions in the models' parameters were 80% and 77% for 1D and 2D cases respectively, when averaged across the two subjects. The models became less compact as the number of critical dimensions increased, e.g., reductions in parameters of 61% and 28% were achieved at the lower threshold for 1D and 2D (with average 3.6 and 5.6 critical dimensions/phone). Comparisons with exhaustive search procedure demonstrated that the proposed algorithm performed equally well but for much less computational load.

Analysis of different feature spaces demonstrated that representations more informationally efficient and compact than raw articulatory models could be derived by applying

proposed algorithm on feature spaces derived using PCA and LDA. It was also possible to identify the movements of articulators for each phone using critical PCA and LDA mode shapes. The analysis showed that the most compact representations are difficult to interpret and hence there is trade-off between compactness and power of interpretation. The results demonstrated that there is no loss in performance when small and insignificant correlations are excluded when estimating the PCA and LDA transformation matrices. However, PCA and LDA offer no benefit over raw articulatory space when important correlations between articulators are ignored when estimating transformation matrices. The LDA representation which maximises separability between phone classes gave slightly better performance over the PCA.

### 8.2.5 Coarticulation modelling

When modelling coarticulation, it is important to know the degrees of freedom of articulators. An unconstrained articulator is more prone to effects such as anticipation of following phone targets and carry forward coarticulation from previous phones. The articulatory movements default to cost minimising behaviour which leads to spatial coarticulatory effects such as target undershoot and overshoot. The knowledge of constraints in the form of articulatory roles could potentially benefit coarticulation modelling. A critical articulator is less prone to coarticulation effects due to its constrained nature but causes the maximum coarticulation effects on the neighbouring unconstrained articulators. A dependent articulator is partially constrained due to its correlation with one or more critical articulators but its remaining degrees of freedom are redundant and hence are prone to coarticulation. A redundant articulator is completely unconstrained and is more susceptible to coarticulation due to neighbouring critical articulators. The targets of dependent and redundant articulators also vary due to the effort minimisation behaviour adapted by the articulator when moving from one critical target to the next. Trajectory generation experiments were performed to test the above theories of coarticulation using constraints in the form of identified articulatory roles in comparison with the constraints from the IPA chart. The results showed that the trajectories generated from the model distributions from ACIDA algorithm were a better fit to the measured trajectories than the ones derived from IPA models. Yet again, inclusion of D-step in estimation of model distributions improved the performance of the IPA models. The results also showed that the proposed coarticulation models have the potential to capture the behaviour of articulators during speech production. For example, anticipatory lowering of velum was better modelled by the proposed anticipatory models than the conventional model in some cases, modelling redundancy based on economy of effort principle also showed some potential in modelling the x movements of articulators. The results also showed that it is not suitable to model carry forward coarticulation in the planning stage of speech production. The statistical framework for modelling coarticulation proposed in this thesis provides a basis for building more complete and accurate models of speech articulation.

### 8.2.6 Publications

The work presented in this thesis has appeared in several articles published at various stages of during the period of study of this thesis: Singampalli and Jackson (2005, 2007a,b,c, 2008); Jackson and Singampalli (2008a,b, 2009); Singampalli and Jackson (2009).

## 8.3 Potential applications

The proposed articulatory constraint identification algorithm has potential applications in the fields of speech science and technology. The algorithm can be used for linguistic studies of various languages, dialects and speakers, for instance in determining phonetic inventories. The identified articulatory constraints could supplement existing theories of coarticulation such as feature spreading (Henke, 1965; Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973) and overlap of articulatory gestures (Browman and Goldstein, 1986; Saltzman and Munhall, 1989) with objective and statistical evidence. Phenomenon such as dipping of tongue during bilabial VCV sequences (known as the trough effect (Lindblom and Sussman, 2002)) could be modelled using the relaxation hypothesis according to which an articulator when redundant relaxes momentarily before reaching the next target position.

The proposed algorithm has the potential to improve the performance of speech recognition and synthesis systems. In engineering, many ASR systems have attempted to incorporate articulatory constraints (King et al., 2007), inspired by distinctive features (Kirchhoff, 1999; Metze and Waibel, 2002; Frankel et al., 2004; Eide, 2001; Koreman et al., 1998), in the form of quantized gestural configurations (Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000), or within a hidden (pseudo-)articulatory layer via forward (Russell and Jackson, 2002; Richards and Bridle, 1999) or inverse mapping (Richmond, 2006; Frankel et al., 2000). The physiological constraints offered by human speech production have been incorporated into speech synthesis via articulatory codebooks, regression and neural-network approaches for forward mapping from articulatory to acoustic domains, as in Schroeter and Sondhi (1994). Knowledge of identified constraints could be used in such production oriented models of speech synthesis and recognition , the proposed algorithm could be used to prioritise speech gestures rather than phonetic rules.

Coarticulation in visual speech has been modelled using various rule based techniques (Beskow, 1995), theories of motor planning and speech production (Cohen and Massaro, 1993), and machine learning algorithms (Xue et al., 2006). In articulatory control models, the constraints have been incorporated using phonetic knowledge. The proposed algorithm could be applied to visual data to extract the constraints for generating smooth and convincing articulation.

## 8.4 Future work

### 8.4.1 Improvements to the data

Inconsistencies in the EMA recordings of the female speaker were reported by Richmond (2001, 2009). Several factors such as reattachment of coils, movement of head within the helmet etc. were found to cause a shift in the mean velum position across the sentences. The data was z score normalised using the underlying mean pattern to minimise such effects by Richmond (2001). Though similar effects were observed, no such normalisation was used for either speakers. Performing such preprocessing on the data would generate much clearer results.

This study assumes that the grand, phone specific and model distributions are Gaussian and unimodal in nature. When the validity of the assumption was tested using Kolmogorov Smirnov goodness of fit test, it was found that only a few distributions satisfy the condition of Gaussianity and unimodality. Transforming the data by applying logarithms or using Box Cox transformation (Box and Cox, 1964) can make the data normal and would improve the results.

Along with the above mentioned improvements, the future work would also focus on validation of the algorithm on more reliable measurements of articulation from various languages and from different speakers.

### 8.4.2 Improvements to the model

The 1D and 2D versions of the algorithm were presented in this thesis. The proposed algorithm could be extended easily to model data with dimensionality $\geq 3$ . However, the algorithm is based on the assumption that the grand, phone and model distributions are multivariate Gaussian in nature. One of the directions for future research is to extend the algorithm to suit multi-modal distributions. Opportunities exist for extension of the KL divergence metric which is used for estimating identification, convergence and evaluation scales to suit multi-modal distributions (Hershey and Olsen, 2007).

Unlike consonants, vowels do not have well defined places of articulation. The derivation of expected critical coordinates from IPA and the comparison with identified critical coordinates were not straightforward processes for vowels unlike consonant sounds. The targets for vowels are part acoustic and part articulatory. Time varying formant patterns form the acoustic cues for vowel sounds, whereas rapid formant transitions, noise bursts, aspirations etc. are the characteristics of consonant sounds. One of the interesting directions for future work would be the identification of acoustic constraints using the proposed algorithm and to analyse the findings with the expected formant characteristics of speech sounds. Comparison of acoustic and articulatory analysis of initial and final vowels of diphthongs using ACIDA algorithm would also be carried out.

### 8.4.3 Modelling coarticulation

#### Planning phase

The span of coarticulation in the proposed approach was limited to the immediate neighbours for modelling temporal coarticulation effects such as anticipation. The onset of anticipation could occur at few segments earlier than the immediate neighbour (Moll and Daniloff, 1971). The effort minimisation hypothesis was also limited to immediate neighbours where the articulator is critical for at least one phone in the triphone sequences considered. Some quadphone sequences were also considered where the articulator is critical for phones at the either ends of the sequence. Longer spans of coarticulation and a variety of contexts would be considered in future for modelling different aspects of coarticulation.

The duration of the phones was not considered when implementing different coarticulation hypotheses. In some cases, redundant articulators tend to relax completely when there is sufficient duration to turn off the underlying muscle activation before reaching for the next target. In other cases where there is not enough time to do so, the minimum effort behaviour for an articulator would be to assume the position intermediate between the previous and the next targets. It would be interesting to investigate the relationship between phone duration and the observed coarticulatory effects.

#### Execution phase

One of the main shortcomings of the methods used for generating trajectories from target distributions was the exclusion of target variance information. The targets were treated as points rather than statistical distributions. The trajectory generation was carried out to evaluate the potential of the proposed statistical framework for modelling coarticulation in the planning stage. Therefore, simple methods such as linear interpolation and Blackburn and Young's model were chosen. It would only be possible to evaluate the full potential of proposed coarticulation models only by employing trajectory generation techniques which treat targets as statistical windows. Therefore, future work would focus on using models such as, MLPG algorithm (Tokuda et al., 2000), trajectory HMMs (Tokuda et al., 2007) for generation of trajectories from the target distributions.

#### Identifying independent components

Principal components analysis and linear discriminant analysis were used generate compact and informationally rich feature representations whereas the knowledge of correlations was used to identify the independently moving articulatory components. Techniques such as independent components analysis (ICA) (Hyvärinen and Oja, 2000.) which optimise the statistical independence between the underlying components could be used on articulatory data to identify the independent articulatory groups. Linear components analysis (Kirirani et al., 1977; Maeda, 1990; Badin et al., 2002) was also used for identifying the underlying independent components of the speech production

system. Future work also focuses on investigation of the ICA and LCA techniques as feature extraction techniques for the articulatory data.

### 8.4.4 Synthesis and recognition

Opportunities exist to explore knowledge of articulatory roles in the synthesis of speech, whether explicitly, e.g., for visual/articulatory speech synthesis, or implicitly, e.g., in a join cost or smoothing function for concatenative synthesis. Future work also focuses on ways of exploiting new knowledge of articulatory constraints as conditional dependencies in probabilistic speech models for ASR. The knowledge of constraints would be applied to generate smoother and continuous trajectories in the hidden articulatory layers of models such as segmental HMMs (Russell and Jackson, 2005).

# Appendix A

# Mocha-Timit phone notation

| Mocha symbol | IPA symbol |
|---|---|
| p | p |
| b | b |
| m | m |
| t | t |
| d | d |
| n | n |
| k | k |
| g | g |
| ng | ŋ |
| f | f |
| v | v |
| th | θ |
| dh | ð |
| s | s |
| z | z |
| sh | ʃ |
| zh | ʒ |
| ch | ʧ |
| jh | ʤ |
| l | l |
| r | ɹ |
| w | w |
| y | j |
| h | h |

Table A.1: *Mocha symbols and corresponding IPA symbols for consonants in database.*

| Mocha symbol | IPA symbol |
|---|---|
| a | æ |
| e | ɛ |
| i | ɪ |
| ii | iː |
| iy | i |
| | |
| @ | ə |
| @@ | ɚ |
| uh | ʌ |
| | |
| aa | ɑ |
| o | ɒ |
| oo | ɔ |
| u | ʊ |
| uu | u |

Table A.2: *Mocha symbols and corresponding IPA symbols for front, mid and back vowels in database*

| Mocha symbol | IPA symbol |
|---|---|
| ai | aɪ |
| ei | eɪ |
| eir | ɛə |
| i@ | ɪə |
| oi | ɔɪ |
| ou | oʊ |
| ow | aʊ |

Table A.3: *Mocha symbols and corresponding IPA symbols for diphthongs in database*

# Appendix B

# Algorithms

## B.1 Generation of covariance ellipses

The procedure used for generation of the covariance ellipse is presented in this section. The covariance ellipse represents the variation within the data in terms of $\pm 2\sigma$ from the mean. The major axis of the ellipse points in the direction of the maximum variance and the minor axis points in the direction of the minimum variance in the data. The angle of orientation and the length of major and minor axes are derived from the eigenvectors and eigenvalues of the covariance matrix generated from the horizontal and vertical movements of the articulators. Let $\mathbf{A} = [x_{1_i} \quad y_{1_i}]$, $i = \{1, 2, .., n_1\}$, where $n_1$ is the number of samples, represent the traces of movement of an articulator in horizontal $x$ and vertical $y$ directions.

- Let $\mu_{x_1}$ and $\mu_{y_1}$ be the means of $x_{1_i}$ and $y_{1_i}$ respectively.
- Calculate the covariance between $x_{1_i}$ and $y_{1_i}$, $R_{xy} = \mathrm{Cov}(x_{1_i}, y_{1_i})$.
- Compute the eigenvalues, $s_1, s_2$ and eigenvectors, $\mathbf{v_1}, \mathbf{v_2}$ of the covariance, $R_{xy}$.
- Define a unit circle, $\begin{bmatrix} c_x \\ c_y \end{bmatrix} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$, where $\theta = 0 : 2\pi$.
- Stretch the circle in the $x$ direction by $s_1$ and in the $y$ direction by $s_2$, i.e., $\begin{bmatrix} c'_x \\ c'_y \end{bmatrix} = \begin{bmatrix} s_1 \cos(\theta) \\ s_2 \sin(\theta) \end{bmatrix}$.
- Rotate the ellipse in the direction given by the eigenvectors, $\begin{bmatrix} c''_x \\ c''_y \end{bmatrix} = \begin{bmatrix} s_1 \cos(\theta) \\ s_2 \sin(\theta) \end{bmatrix} [\mathbf{v_1} \ \mathbf{v_2}]$.
- Center the stretched, rotated ellipse at the mean, $(\mu_{x_1}, \mu_{y_1})$, $\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} c''_x + \mu_{x_1} \\ c''_y + \mu_{y_1} \end{bmatrix}$
- Plot $X$ and $Y$.

## B.2 Conditional distribution

Let $\mathbf{X}$ be $p$ dimensional and distributed according to $\mathrm{N}(\mu, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is nonsingular. Let $\mathbf{X}$ be partitioned into $q$ and $p-q$ component subvectors

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix}$$

and their mean be

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$$

where $\mu^{(1)}$ is $q$ dimensional and $\mu^{(1)}$ is $p-q$ dimensional. Let the covariance be

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

where the dimensionality of $\mathbf{\Sigma}_{11}$ is $q \times q$ , $\mathbf{\Sigma}_{22}$ is $(p-q) \times (p-q)$, $\mathbf{\Sigma}_{12}$ is $q \times (p-q)$ and $\mathbf{\Sigma}_{21}$ is $(p-q) \times q$.

The subvectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ can be linearly, nonsingularly transformed to two independent subvectors

$$\mathbf{Y}^{(1)} = \mathbf{X}^{(1)} + \mathbf{B}\mathbf{X}^{(2)} \tag{B.1}$$

$$\mathbf{Y}^{(2)} = \mathbf{X}^{(2)} \tag{B.2}$$

Since the vectors $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are assumed to be independent, the covariance between them is set to zero to solve for $\mathbf{B}$,

$$\mathscr{E}(\mathbf{Y}^{(1)} - \mathscr{E}\mathbf{Y}^{(1)})(\mathbf{Y}^{(2)} - \mathscr{E}\mathbf{Y}^{(2)})' = 0 \tag{B.3}$$

$$\Rightarrow \quad \mathbf{\Sigma}_{12} + \mathbf{B}\mathbf{\Sigma}_{22} = 0 \tag{B.4}$$

Therefore, $\mathbf{B} = -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$.

Hence,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{pmatrix} = \begin{pmatrix} I & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ 0 & I \end{pmatrix} \mathbf{X} \tag{B.5}$$

The vector $\mathbf{Y}$ has a normal distribution with mean

$$\mathscr{E}\begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{pmatrix} = \mathscr{E}\begin{pmatrix} I & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ 0 & I \end{pmatrix} \mathbf{X}, \tag{B.6}$$

$$= \begin{pmatrix} I & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \tag{B.7}$$

$$= \begin{pmatrix} \mu^{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mu^{(2)} \\ \mu^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{v}^{(1)} \\ \mathbf{v}^{(2)} \end{pmatrix}, \tag{B.8}$$

and covariance

$$\mathscr{C}(\mathbf{Y}) = \begin{bmatrix} \mathscr{E}(\mathbf{Y}^{(1)} - \mathbf{v}^{(1)})(\mathbf{Y}^{(1)} - \mathbf{v}^{(1)})' & \mathscr{E}(\mathbf{Y}^{(1)} - \mathbf{v}^{(1)})(\mathbf{Y}^{(2)} - \mathbf{v}^{(2)})' \\ \mathscr{E}(\mathbf{Y}^{(2)} - \mathbf{v}^{(2)})(\mathbf{Y}^{(1)} - \mathbf{v}^{(1)})' & \mathscr{E}(\mathbf{Y}^{(2)} - \mathbf{v}^{(2)})(\mathbf{Y}^{(2)} - \mathbf{v}^{(2)})' \end{bmatrix}, \quad \text{(B.9)}$$

$$= \begin{bmatrix} \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} & 0 \\ 0 & \mathbf{\Sigma}_{22} \end{bmatrix}. \quad \text{(B.10)}$$

The distributions of the subvectors $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are given as

$$f(\mathbf{y}^{(1)}) = \mathcal{N}(\mathbf{y}^{(1)}; \mu^{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mu^{(2)}, \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}), \quad \text{(B.11)}$$

$$f(\mathbf{y}^{(2)}) = \mathcal{N}(\mathbf{y}^{(2)}; \mu^{(1)}, \mathbf{\Sigma}_{22}) \quad \text{(B.12)}$$

The joint density is given as

$$f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}) = f(\mathbf{y}^{(1)}) f(\mathbf{y}^{(2)}) \quad \text{(B.13)}$$

The density of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ can be obtained from $f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$ as

$$g(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = f\left[\mathbf{y}^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}), \mathbf{y}^{(2)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})\right] J(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \quad \text{(B.14)}$$

where $\mathbf{y}^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ denotes that $\mathbf{y}^{(1)}$ is a function of both $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ given by eq.B.1. Also $J(\mathbf{x}^{(1)}$ is the Jacobian of the transformation given as

$$J(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \text{mod} \begin{vmatrix} \partial y^{(1)}/\partial x^{(1)} & \partial y^{(1)}/\partial x^{(2)} \\ \partial y^{(2)}/\partial x^{(1)} & \partial y^{(2)}/\partial x^{(2)} \end{vmatrix} \quad \text{(B.15)}$$

After substituting value of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ from eq.B.5 in eq.B.15. the value of the Jacobian is found to be 1. Therefore eq.B.14 becomes

$$g(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = f[\mathbf{y}^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}), \mathbf{y}^{(2)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})] \quad \text{(B.16)}$$

The conditional density of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)}$ is

$$g(\mathbf{x}^{(1)}|\mathbf{x}^{(2)}) = \frac{g(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{g(\mathbf{x}^{(2)})} \quad \text{(B.17)}$$

$$= f(\mathbf{y}^{(1)}) \quad \text{(B.18)}$$

$$g(\mathbf{x}^{(1)}|\mathbf{x}^{(2)}) = \mathcal{N}\left(\mathbf{x}^{(1)}; \mu^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \mu^{(2)}), \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\right). \quad \text{(B.19)}$$

The mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ of the conditional distribution are

$$\bar{\mathbf{m}} = \mu^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \mu^{(2)}) \quad \text{(B.20)}$$

$$\bar{\mathbf{S}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \quad \text{(B.21)}$$

Note that the mean $\hat{\mathbf{m}}$ is estimated using samples in $\mathbf{x}^{(2)}$ where as the covariance $\hat{\mathbf{S}}$ is independent of $\mathbf{x}^{(2)}$.

## B.3 Significance tests

Various statistical significance tests were used to analyse the articulatory data throughout this report. This section provides a small introduction to the procedure used in testing for statistical significance.

### B.3.1 A brief introduction to significance tests

There basic steps in the procedure used for hypothesis testing using the available data for determining the statistical significance of the given a some data are:

- stating the **null hypothesis**,
- assuming the **research hypothesis**,
- choosing the **level of significance** and therefore the **critical value**,
- computing the **test statistic** from the data and determining new **level of significance** by comparing it with the **critical value**,
- accepting or rejecting **null hypothesis**.

(to be completed...)

### B.3.2 Kolmogorov-Smirnov goodness-of-fit test

Let $F(x_P)$, $x_P \in \mathbf{X}_P$, be the distribution of the position population from which the samples are derived. Let $F_0(x_P)$ be the hypothesized distribution, here a univariate Gaussian. Therefore

$$F_0(x_P) = \frac{1}{2}(1 + \text{erf}(\frac{z}{\sqrt{2}})), \tag{B.22}$$

where erf(.) is the Gauss error function, and $z = \frac{x_P - \mu_{P_\phi}}{\sigma_{P_\phi}}$ is the $z$ score which is the difference between each sample and the mean divided by the standard deviation. A positive $z$ score represents that the sample is greater than the mean and viceversa.

- The null hypothesis assumes that there is no difference between the observed distribution of the samples and the hypothesised distribution i.e., $H_0 : F(x_P) = F_0(x_P)$.
- The research hypothesis states that the two distributions are significantly different, i.e., $H_1 : F(x_P) \neq F_0(x_P)$.
- Select the level of significance, $\alpha$, which is the amount of risk associated with rejecting the null hypothesis when it actually is true and is expressed in terms of a percentage. For example, a significance value of 0.05 implies that there is a 5% chance of rejecting the null hypothesis when it is actually true.

- The hypothesised cumulative distribution of the samples is calculated as given by the eq.B.22.
- The observed cumulative step function of the sample $F(x_P)$ is calculated, where $F(x) = k/n$, where $k$ is the number of samples less than or equal to $x$, given $n$ number of samples.
- The maximum absolute difference between the two distributions, $d = \max|F_0(x_P) - F(x_P)|$ is computed.
- The observed difference $d = d_{\text{val}}$ is compared with the critical value $d_{\text{crt}}$ and null hypothesis is accepted or rejected accordingly. The value of $d_{\text{crt}}$ depends on the number of samples in the group $n$ and the level of confidence $\alpha$. If $d_{\text{val}} < d_{\text{crit}}$, the null hypothesis is accepted and it is considered that the observed difference $d_{\text{val}}$ is only due to chance. Therefore, the probability of rejecting the null hypothesis by chance is higher than the level of significance $\alpha$. Otherwise, if $d_{\text{val}} > d_{\text{crit}}$, the null hypothesis is rejected and the probability of rejecting the null hypothesis by chance is less than the level of significance $\alpha$.

### B.3.3 Pearson's test of correlation

Let $R$ be the correlation between the samples of two groups A and B. Let the samples in group A be denoted as $x_A$ and in B be $x_B$.

- Null hypothesis assumes that there is no correlation between the variables, i.e., $H_0 : R = 0$
- Research hypothesis states that there is a correlation between the variables, i.e., $H_1 : R \neq 0$
- Set the level of significance $\alpha$ that determines the probability of getting a value of $R$ other than zero when the null hypothesis is true. For example, if $\alpha = 0.05$, there is one in twenty chance of rejecting the null hypothesis by chance when it actually holds true.
- Calculate the correlation coefficient given $n$ number of samples

$$R = \frac{\sum x_A x_B - \frac{(\sum x_A)(\sum x_B)}{n}}{\sqrt{(\sum x_A^2 - \frac{(\sum x_A)^2}{n})(\sum x_B^2 - \frac{(\sum x_B)^2}{n})}}. \tag{B.23}$$

- For sample sizes larger than 100, significance of the $R$ is obtained by computing the t statistic which determines the probability of getting a non zero value of $R$ by chance from a population whose correlation $R$ is zero

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \tag{B.24}$$

- The obtained t value, $t = t_{\text{val}}$ is compared with the critical value $t_{\text{crt}}$ to determine the significance of the correlation. The critical value depends on the level of significance $\alpha$ and the number of samples $n$. If $t < t_{\text{crt}}$, the null hypothesis is accepted and the probability of rejecting the null hypothesis by chance is higher than the level of significance $\alpha$. Alternatively, if $t > t_{\text{crt}}$, the null hypothesis is rejected and the probability of rejecting the null hypothesis by chance is lower than the level of significance $\alpha$.

### B.3.4 Independent samples t-test

Let $A_1 = [x_1(i)]$, $i = \{1, 2, .., n_1\}$ be the samples in group-1 where $n_1$ is the number of samples in the group. Let $A_2 = [x_2(i)]$, $i = \{1, 2, .., n_2\}$ be the data in group-2 where $n_2$ is the number of samples in the group. Let the population mean of the group-1 be $\tilde{\mu}_1$ and that of group-2 be $\tilde{\mu}_2$.

- The null hypothesis assumes that the means of the populations represented by the groups are equal, i.e., $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$.
- The research hypothesis states that there is a significant difference between the means, i.e., $H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$.
- Let $\alpha$ be the level of significance. The level of significance determines the probability of rejecting the null hypothesis by chance when it actually holds true.
- Let $\mu_1$, $\mu_2$ be the sample means and $s_1^2$, $s_2^2$ be the sample variances of the two groups respectively.
- The t-test assumes that the variances of the two sample groups are equal. To test for the same, Levene's statistic is computed, the procedure for which is given in sectionB.3.4.
- The t-statistic is computed as the difference between the observed and the hypothesised differences between the means with respect to the standard error of the difference

$$t = \frac{(\mu_1 - \mu_2) - (\tilde{\mu}_1 - \tilde{\mu}_2)}{\text{estimate of standard error}} \tag{B.25}$$

- Under null hypothesis, the difference between the population means is zero. Therefore,

$$t = \frac{\mu_1 - \mu_2}{\text{estimate of standard error}} \tag{B.26}$$

- The standard deviation of the sample mean is an estimate of the amount by which the sample mean differs from the population mean and is known as the standard error. The standard errors of the sampling distributions are given as

$$\text{SE}_1 = \frac{s_1}{\sqrt{n_1}}; \quad \text{SE}_2 = \frac{s_2}{\sqrt{n_2}}. \tag{B.27}$$

- The variance of each sampling distribution is the square of the corresponding standard error given in eq.B.27. The variance of the sampling distribution of differences is found using the variance sum law as

$$\mathrm{SE}^2_{\mathrm{grp_1-grp_2}} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \quad \text{if } n_1 = n_2 \tag{B.28}$$

$$\mathrm{SE}^2_{\mathrm{p}} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \quad \text{if } n_1 \neq n_2 \tag{B.29}$$

$$\tag{B.30}$$

- The standard error of the sampling distribution of the differences is obtained by taking the square root of the variance

$$\mathrm{SE}_{\mathrm{grp_1-grp_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{if } n_1 = n_2 \tag{B.31}$$

$$\mathrm{SE}_{\mathrm{pooled}} = \sqrt{\frac{\mathrm{SE}^2_{\mathrm{p}}}{n_1} + \frac{\mathrm{SE}^2_{\mathrm{p}}}{n_2}} \quad \text{if } n_1 \neq n_2 \tag{B.32}$$

- Substitute the estimate of the standard error given by eq.B.31 in eq.B.26 to obtain the test statistic. If the sample sizes are equal,

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]}}. \tag{B.33}$$

- If the sample sizes are unequal, the pooled variance estimate is used in the computation of the test statistic

$$t = \frac{\mu_2 - \mu_1}{\sqrt{\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{B.34}$$

- The obtained value of t, $t = t_{\mathrm{val}}$, is compared with the critical value $t_{\mathrm{crt}}$ to determine the significance of the difference between the means. The critical value depends on the significance level $\alpha$ and the number of degrees of freedom, here, $n_1 + n_2 - 2$. If $t_{\mathrm{val}} < t_{\mathrm{crt}}$, the null hypothesis is accepted and it is assumed that the observed difference between the means $t_{\mathrm{val}}$ is by chance. The probability of rejecting the null hypothesis here would be larger than the level of significance. Alternatively, if $t_{\mathrm{val}} > t_{\mathrm{crt}}$, the null hypothesis is rejected and the probability of rejecting the null hypothesis by chance is smaller than the level of significance.

### Homogeneity of variances

One of the important assumptions of the independent samples t-test is the homogeneity of variances (Field, 2005), i.e., the variances are equal. Let the number of groups be equal to $M$. The Levene's F statistic is used to determine the significance of the equality of variances of the two groups. Let the grand mean obtained by considering $n_1 + n_2$ samples be $\mu$

$$\mu = \frac{\sum_{i=1}^{n_1} x_1(i) + \sum_{i=1}^{n_2} x_2(i)}{n_1 + n_2}. \tag{B.35}$$

- The null hypothesis assumes that there is no significant difference between the variances of the populations represented by the two groups, i.e., $H_0 : \tilde{s}_1^2 = \tilde{s}_2^2$.
- The research hypothesis states that there is a significant difference between the variances, i.e., $H_1 : \tilde{s}_1^2 \neq \tilde{s}_2^2$.
- Let $\alpha$ be the level of significance. For example, if $\alpha = 0.05$, there is one in twenty chance of rejecting the null hypothesis by chance when it actually holds true.
- The weighted sum of squared difference between the mean of each group and the grand mean, i.e., the variance between groups is calculated. This value is called the model or hypothesised sum of squares denoted by

$$\mathrm{SS_H} = n_2(\mu_2 - \mu)^2 + n_1(\mu_1 - \mu)^2. \tag{B.36}$$

- The variance within the groups, i.e.,weighted sum of squared differences of each group considering the group mean is calculated. This value is called the residual sum of squares and is denoted by

$$\begin{aligned} \mathrm{SS_R} &= \sum_{i=1}^{n_2}(x_2(i) - \mu_2)^2 + \sum_{i=1}^{n_1}(x_1(i) - \mu_1)^2, && \text{(B.37)} \\ &= (n_2 - 1)s_2^2 + (n_1 - 1)s_1^2. && \text{(B.38)} \end{aligned}$$

- The total sum of squares is the sum of the hypothesised sum of squares and the residual sum of squares, $\mathrm{SS_T} = \mathrm{SS_H} + \mathrm{SS_R}$. This value represents the total variance in the samples and is computed by calculating the difference between the grand mean $\mu$ and all the samples in group-1 and group-2 and summing the squared difference

$$\mathrm{SS_T} = \sum_{i=1}^{n_1}(x_1(i) - \mu)^2 + \sum_{i=1}^{n_2}(x_2(i) - \mu)^2. \tag{B.39}$$

- The hypothesised mean square value and the residual mean square value are computed using the sum of square values

$$\begin{aligned} \mathrm{MS_H} &= \frac{\mathrm{SS_H}}{(M-1)}, && \text{(B.40)} \\ \mathrm{MS_R} &= \frac{\mathrm{SS_R}}{(n_1 - 1) + (n_2 - 1)}. && \text{(B.41)} \end{aligned}$$

- The $F$ ratio is obtained by dividing the hypothesised mean squares by the residual mean squares

$$\mathrm{F} = \frac{\mathrm{MS_H}}{\mathrm{MS_R}}. \tag{B.42}$$

- The obtained value of $\mathrm{F} = \mathrm{F_{val}}$ ratio is compared with the critical value $\mathrm{F_{crt}}$ to obtain the significance of difference between the variances. If $\mathrm{F_{val}} < \mathrm{F_{crt}}$, the null hypothesis is accepted and the observed difference in variances between the groups

is by chance. The probability of rejecting the null hypothesis by chance would then be higher than the level of significance. If $F_{val} > F_{crt}$, the null hypothesis is rejected and there exists a difference in variances between the groups which is not by chance. The probability of rejecting the null hypothesis by chance would then be lower than the level of significance.

### Dealing with unbalanced data

If the number of samples in the two groups of data, $n_1$ and $n_2$ are unequal, this imbalance effects the calculation of residual sum of squares (eq.B.37) and in turn effects the F ratio (eq.B.42) if the variance of the group with large number of samples is higher. The residual sum of squares value would be more biased towards the group which has higher number of samples and the F ratio would be conservative as a consequence. There are two approaches to reduce the impact of large sample size associated with large variance(Field, 2005):

1. Brown-Forsythe F ratio
2. Welch F ratio

In the Brown-Forsythe's approach (Field, 2005), the effect of the unbalanced data sets is reduced by considering the variances weighted by the sample sizes as a proportion of the total sample size in the computation of the residual sum of squares. Eq.B.37 can be rewritten as

$$SS_R^{BF} = \left(1 - \frac{n_2}{n_1 + n_2}\right) \sum_{i=1}^{n_2} (x_2(i) - \mu_2)^2 + \left(1 - \frac{n_1}{n_1 + n_2}\right) \sum_{i=1}^{n_1} (x_1(i) - \mu_1)^2. \quad \text{(B.43)}$$

The residual mean square value is recomputed using the Brown-Forsythe's residual sum of squares, $SS_R^{BF}$, as

$$MS_R^{BF} = \frac{SS_R^{BF}}{n_1 + n_2 - 2}. \quad \text{(B.44)}$$

The F ratio calculated using the new estimate of residual mean square value is known as Brown-Forsythe F ratio

$$F^{BF} = \frac{MS_H}{MS_R^{BF}}. \quad \text{(B.45)}$$

The Welch's approach (Field, 2005) is based on using the weighted means and variances in calculating the $F$ ratio. The weights associated with each group are the reciprocals of the squared standard errors, i.e., the variances. Therefore, more emphasis is given to the sample mean that is closer to the population mean. The weights for the two groups are given as

$$w_1 = \frac{n_1}{s_1^2}; \qquad w_2 = \frac{n_2}{s_2^2}. \quad \text{(B.46)}$$

The Welch grand mean is computed as

$$\mu^W = \frac{\sum_{i=1}^{2} w_i \mu_i}{\sum_{i=1}^{2} w_i}. \tag{B.47}$$

The model sum of squares given by eq.B.36 is modified to incorporate the weights and the adjusted grand mean

$$\mathrm{SS}_{\mathrm{H}}^{\mathrm{W}} = \sum_{i=1}^{2} w_i(\mu_i - \mu^W). \tag{B.48}$$

The modified hypothesised mean square value is computed using the Welch's sum of squares value

$$\mathrm{MS}_{\mathrm{H}}^{\mathrm{W}} = \frac{\mathrm{SS}_{\mathrm{H}}^{\mathrm{W}}}{M-1} \tag{B.49}$$

where M is the number of groups.
The residual sum of squares is modified and expressed solely interms of the weights and is denoted by $\Lambda$

$$\Lambda = 3\frac{\sum_{i=1}^{2} \frac{\left(1-\frac{w_i}{\sum_{i=1}^{2} w_i}\right)}{n_i-1}}{M^2-1}. \tag{B.50}$$

The Welch's F ratio is then given by

$$\mathrm{F}^{\mathrm{W}} = \frac{\mathrm{MS}_{\mathrm{H}}^{\mathrm{W}}}{1+\frac{2\Lambda(M-2)}{3}}. \tag{B.51}$$

The Welch's F ratio is more robust to the imbalances in the sample sizes of the groups.

# Appendix C

# Supporting plots and tables

## C.1 Identification of articulatory constraints

Table C.1: *1D dependent and redundant articulatory coordinates identified using ACIDA algorithm at IPA level of complexity for male speaker.*

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [p] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_y$ $TD_y$ $V_y$ | $TB_x$ $TD_x$ $V_x$ |
| [b] | $UL_x$ $LL_x$ $LI_y$ $TT_y$ $TB_y$ $TD_y$ $V_y$ | $LI_x$ $TT_x$ $TB_x$ $TD_x$ $V_x$ |
| [m] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_y$ | - |
| [t] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ | $UL_x$ $TT_x$ $TD_y$ $V_x$ $V_y$ |
| [d] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ | $UL_x$ $TT_x$ $V_x$ $V_y$ |
| [n] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_x$ $TB_y$ $TD_x$ $V_y$ | $UL_x$ $TD_y$ |
| [k] | $UL_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $UL_x$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ |
| [g] | $UL_y$ $LL_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $UL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ |
| [ŋ] | $UL_y$ $LL_x$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_y$ | $UL_x$ $LL_y$ $LI_x$ $LI_y$ |
| [f] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ | $V_x$ $V_y$ |
| [v] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_y$ $TD_y$ $V_y$ | $TT_x$ $TB_x$ $TD_x$ $V_x$ |
| [θ] | $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ | $UL_x$ $V_y$ |
| [ð] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ | $V_y$ |
| [s] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ | $UL_x$ |
| [z] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ | $UL_x$ $V_y$ |
| [ʃ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | - |
| [ʒ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | - |

Continued on next page...

Table C.1 continued from previous page

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [ʧ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TD_x$ $TD_y$ $V_y$ | $V_x$ |
| [ʤ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TD_x$ $TD_y$ $V_x$ $V_y$ | - |
| [l] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɹ] | $UL_y$ $LL_x$ $LL_y$ $TT_x$ $TT_y$ $TD_x$ $TD_y$ | $UL_x$ $LI_x$ $LI_y$ $TB_y$ $V_x$ $V_y$ |
| [w] | $UL_x$ $LL_x$ $LL_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_y$ | $LI_x$ $LI_y$ $V_x$ |
| [j] | $UL_y$ $LI_x$ $LI_y$ $TT_y$ $TD_y$ $V_y$ | $UL_x$ $LL_x$ $LL_y$ $TT_x$ $TB_x$ $TD_x$ $V_x$ |
| [h] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [æ] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ | $UL_y$ $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɛ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ | $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɪ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [iː] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ | $LI_x$ $V_y$ |
| [i] | $UL_y$ $LL_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_y$ | $UL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $V_x$ |
| [ə] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɚ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ | $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ʌ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɑ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TD_x$ $TD_y$ $V_y$ | $V_x$ |
| [ɒ] | $UL_y$ $LI_x$ $LI_y$ $TT_y$ $TD_y$ | $UL_x$ $LL_x$ $LL_y$ $TT_x$ $TB_x$ $TD_x$ $V_x$ $V_y$ |
| [ɔ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $TD_y$ | $V_x$ $V_y$ |
| [ʊ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [u] | $UL_y$ $LL_x$ $TT_y$ $TB_y$ $V_x$ $V_y$ | $UL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_x$ $TD_x$ |
| [aɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TT_x$ |
| [aɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ | $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [eɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_x$ | $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [eɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_y$ | $LI_x$ $V_x$ |
| [ɛə] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ | $TT_x$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɛə] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ | $TT_x$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɪə] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TB_x$ $TD_x$ $V_x$ $V_y$ | $LI_x$ $TB_y$ $TD_y$ |
| [ɪə̀] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TB_x$ | $LI_x$ $TT_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɔɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ | $V_x$ $V_y$ |

Continued on next page...

Table C.1 continued from previous page

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [ɔɪ] | $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_y$ | $V_x$ |
| [oʊ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [oʊ] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_x$ $TD_y$ $V_y$ | $TB_y$ $V_x$ |
| [aʊ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TT_x$ |
| [aʊ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TD_x$ $TD_y$ $V_y$ | $V_x$ |

Table C.2: *1D dependent and redundant articulatory coordinates identified using ACIDA algorithm at IPA level of complexity for female speaker.*

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [p] | $UL_x$ $LL_x$ $LL_y$ $LI_y$ $TT_y$ $TB_y$ | $LI_x$ $TT_x$ $TB_x$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [b] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_y$ $TD_y$ $V_x$ | $TT_x$ $TB_x$ $TD_x$ $V_y$ |
| [m] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_y$ $TD_x$ $TD_y$ $V_x$ | $TT_x$ $TB_x$ $V_y$ |
| [t] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ | $UL_x$ $TD_y$ $V_y$ |
| [d] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $V_x$ | $UL_x$ $TB_x$ $TD_y$ $V_y$ |
| [n] | $UL_x$ $UL_y$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $LL_x$ $TB_x$ $TD_y$ |
| [k] | $LL_x$ $LI_x$ $TB_y$ $V_x$ | $UL_x$ $UL_y$ $LL_y$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_x$ $V_y$ |
| [g] | $LI_x$ $TB_y$ $V_x$ $V_y$ | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_x$ |
| [ŋ] | $LL_x$ $LI_x$ $TB_y$ $V_x$ $V_y$ | $UL_x$ $UL_y$ $LL_y$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_x$ |
| [f] | $UL_x$ $UL_y$ $LI_x$ $LI_y$ $TT_y$ $TD_x$ $TD_y$ $V_x$ | $TT_x$ $TB_x$ $TB_y$ $V_y$ |
| [v] | $UL_y$ $LI_x$ $LI_y$ $TT_y$ $TD_y$ $V_x$ | $UL_x$ $TT_x$ $TB_x$ $TB_y$ $TD_x$ $V_y$ |
| [θ] | $UL_y$ $LL_y$ $LI_x$ $LI_y$ $TB_x$ $TD_x$ $TD_y$ $V_x$ $V_y$ | $UL_x$ $LL_x$ |
| [ð] | $UL_y$ $LL_y$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $UL_x$ $LL_x$ $LI_x$ $TD_y$ |
| [s] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TB_y$ $TD_x$ $V_x$ | $UL_x$ $TD_y$ $V_y$ |
| [z] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $UL_x$ $TD_y$ |
| [ʃ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TD_y$ |
| [ʒ] | $UL_x$ $UL_y$ $LI_x$ $TT_x$ $TB_x$ $TD_y$ $V_x$ $V_y$ | $LL_y$ |
| [ʧ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TD_y$ |
| [ʤ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TD_y$ |
| [l] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |

Continued on next page...

Table C.2 continued from previous page

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [ɹ] | $UL_x$ $LI_y$ $TT_y$ $TB_x$ $TD_x$ | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TB_y$ $TD_y$ $V_x$ $V_y$ |
| [w] | $UL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $LL_x$ $TT_x$ $TB_x$ |
| [j] | $UL_y$ $LI_y$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $UL_x$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TD_x$ $V_y$ |
| [h] | $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $V_x$ $V_y$ | $UL_x$ $TB_x$ $TD_x$ $TD_y$ |
| [æ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TB_y$ $V_x$ $V_y$ | $TT_x$ $TB_x$ $TD_x$ |
| [ɛ] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TD_y$ | $LI_x$ $TT_x$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ |
| [ɪ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [iː] | $UL_y$ $LI_y$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $UL_x$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TD_x$ $V_y$ |
| [i] | $UL_x$ $LL_x$ $LL_y$ $LI_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ | $UL_y$ $LI_x$ $V_y$ |
| [ə] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [ɚ] | $UL_x$ $LL_y$ $LI_x$ $LI_y$ $TB_y$ $TD_y$ $V_x$ $V_y$ | $UL_y$ $TT_x$ $TB_x$ $TD_x$ |
| [ʌ] | $LI_y$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TD_x$ $V_y$ |
| [ɑ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $LL_y$ $V_y$ |
| [ɒ] | $UL_x$ $UL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $LL_x$ $LL_y$ $V_y$ |
| [ɔ] | $UL_x$ $UL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $V_x$ $V_y$ | $LL_x$ |
| [ʊ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [u] | $UL_y$ $LI_y$ $TB_x$ $TD_y$ $V_x$ | $UL_x$ $LL_x$ $LL_y$ $LI_x$ $TT_x$ $TT_y$ $TD_x$ $V_y$ |
| [aɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TB_x$ |
| [aɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_x$ $LI_y$ $TT_y$ $TD_y$ $V_x$ | $TT_x$ $TB_x$ $TB_y$ $TD_x$ $V_y$ |
| [eɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $LI_x$ $TT_x$ $TD_x$ $V_y$ |
| [eɪ] | $UL_x$ $UL_y$ $LL_x$ $LI_y$ $TT_x$ $TT_y$ $TB_y$ $TD_x$ $V_x$ | $LI_x$ $V_y$ |
| [ɛə] | $UL_x$ $LL_x$ $LL_y$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ | $LI_x$ $TT_x$ $TD_y$ $V_y$ |
| [ɛə] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $TT_y$ $TD_y$ $V_x$ | $LI_x$ $TT_x$ $TB_x$ $TD_x$ $V_y$ |
| [ɪə] | $UL_x$ $UL_y$ $LI_y$ $TT_y$ $TB_x$ $TD_y$ $V_x$ | $LL_x$ $LI_x$ $TT_x$ $TD_x$ $V_y$ |
| [ɪə] | $UL_x$ $LL_y$ $LI_y$ $TT_y$ $TB_y$ $TD_x$ | $LL_x$ $LI_x$ $TT_x$ $TB_x$ $TD_y$ $V_x$ $V_y$ |
| [ɔɪ] | $UL_x$ $LL_x$ $LI_x$ $LI_y$ $TB_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $UL_y$ |
| [ɔɪ] | $LL_x$ $LL_y$ $LI_x$ $TB_y$ $V_x$ $V_y$ | $UL_x$ $UL_y$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TD_x$ |
| [oʊ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [oʊ] | - | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TT_y$ $TB_x$ $TB_y$ $TD_x$ $TD_y$ $V_x$ $V_y$ |
| [aʊ] | $UL_x$ $UL_y$ $LL_x$ $LL_y$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $V_x$ $V_y$ | $TB_x$ |
| [aʊ] | $UL_x$ $LI_x$ $LI_y$ $TT_x$ $TB_y$ $TD_x$ $V_x$ | $UL_y$ $LL_x$ $LL_y$ $TT_y$ $TD_y$ $V_y$ |

Table C.3: *2D dependent and redundant articulatory coordinates identified using ACIDA algorithm at IPA level of complexity for male speaker.*

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [p] | LI TT TB TD V | |
| [b] | LI TT TB TD V | |
| [m] | LI TT TB TD | |
| [t] | UL LL LI TB TD | V |
| [d] | UL LL LI TB TD | V |
| [n] | UL LL LI TB TD | |
| [k] | UL LL TT TB V | LI |
| [g] | UL LL TT TB V | LI |
| [ŋ] | UL LL TT TB | LI |
| [f] | UL LI TT TB TD V | |
| [v] | UL LI TT TB TD V | |
| [θ] | UL LI TB TD V | |
| [ð] | UL LL LI TB TD | V |
| [s] | UL LL TB TD | V |
| [z] | UL LL TB TD | V |
| [ʃ] | UL LL TB V | |
| [ʒ] | UL TB V | |
| [ʧ] | UL LL TB TD | V |
| [ʤ] | UL LL TD V | |
| [l] | | UL LL LI TT TB TD V |
| [ɹ] | UL LL LI TB TD | V |
| [w] | LL LI TB TD V | |
| [j] | UL LL LI TT TD V | |
| [h] | UL LL LI TB TD | V |
| [æ] | UL LI TT TB TD V | |
| [ɛ] | UL LI TT TB TD V | |
| [ɪ] | | UL LL LI TT TB TD V |
| [iː] | UL LL LI TD V | |
| [i] | UL LL LI TT TD V | |
| [ə] | | UL LL LI TT TB TD V |
| [ɚ] | | UL LL LI TT TB TD V |
| [ʌ] | | UL LL LI TT TB TD V |
| [ɑ] | UL LI TT TD V | |
| [ɒ] | UL LL LI TT TD | V |
| [ɔ] | UL LI TT TD V | |
| [ʊ] | | UL LL LI TT TB TD V |
| [u] | UL LL TT TB V | LI |
| [aɪ] | UL LI TT TB V | |
| [aɪ] | UL LI TT TB TD V | |
| [eɪ] | UL LI TT TB TD V | |
| [eɪ] | UL LI TB TD V | |

Continued on next page...

Table C.3 continued from previous page

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [ɛə] | UL LI TT TB TD V | |
| [ɛə] | UL LL TT TB | TD V |
| [ɪə] | UL LI TB TD V | |
| [ɪə] | UL LI TT TB TD V | |
| [ɔɪ] | UL LI TB TD V | |
| [ɔɪ] | UL LI TT TB V | |
| [oʊ] | | UL LL LI TT TB TD V |
| [oʊ] | LL LI TT TB TD V | |
| [aʊ] | UL LI TT TD V | |
| [aʊ] | UL LL LI TT TD V | |

Table C.4: *2D dependent and redundant articulatory coordinates identified using ACIDA algorithm at IPA level of complexity for female speaker.*

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [p] | LL LI TT TB TD | V |
| [b] | LI TT TB TD V | - |
| [m] | LI TT TB TD V | - |
| [t] | UL LL LI TB TD V | - |
| [d] | UL LL LI TB TD V | - |
| [n] | UL LL LI TB TD V | - |
| [k] | UL LL LI TT TB V | - |
| [g] | UL LL LI TT TB V | - |
| [ŋ] | UL LL LI TT TB V | - |
| [f] | UL LI TT TD V | TB |
| [v] | UL LI TT TD V | TB |
| [θ] | UL TD V | - |
| [ð] | UL LL LI TB TD V | - |
| [s] | UL LL TB TD V | - |
| [z] | UL LL TB TD V | - |
| [ʃ] | UL LL TB TD V | - |
| [ʒ] | UL LL TD V | - |
| [ʧ] | LL TB TD V | - |
| [ʤ] | UL LL TB TD V | - |
| [l] | - | UL LL LI TT TB TD V |
| [ɹ] | UL LL LI TB TD | V |
| [w] | LI TT TB V | - |
| [j] | UL LL LI TD V | - |
| [h] | UL LL LI TB TD V | - |
| [æ] | UL LI TT TB V | - |
| [ɛ] | UL LI TT TD | TB V |
| [ɪ] | - | UL LL LI TT TB TD V |
| [iː] | UL LL LI TT V | - |

Continued on next page...

Table C.4 continued from previous page

| Phones | Dependent coordinates | Redundant coordinates |
|---|---|---|
| [i] | UL LL LI TT V | - |
| [ə] | - | UL LL LI TT TB TD V |
| [ɚ] | UL LI TT TD V | - |
| [ʌ] | UL LI TT TD V | LL |
| [ɑ] | UL LI TT TD V | LL |
| [ɒ] | UL LI TT TD V | LL |
| [ɔ] | UL LI TT TB V | - |
| [ʊ] | - | UL LL LI TT TB TD V |
| [u] | UL LL LI TT TB V | - |
| [aɪ] | UL LL LI TB TD V | - |
| [aɪ] | UL LI TT TD V | TB |
| [eɪ] | UL LI TT TD V | - |
| [eɪ] | UL LI TT TD V | - |
| [ɛə] | UL LL TT TB V | - |
| [ɛə] | UL LI TT TD V | LL |
| [ɪə] | LL LI TD V | - |
| [ɪə] | LL LI TT TB TD | V |
| [ɔɪ] | UL LI TT TB | - |
| [ɔɪ] | UL LL LI TT TB | - |
| [oʊ] | - | UL LL LI TT TB TD V |
| [oʊ] | - | UL LL LI TT TB TD V |
| [aʊ] | UL LL LI TB V | - |
| [aʊ] | UL LI TT TD V | LL |

## C.2 Comparison with IPA

| Consonants | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|
| [p] | $UL_y$ $LL_y$ | $UL_y$ $LL_y$ | $UL_y$ |
| [b] | $UL_y$ $LL_y$ | $UL_y$ $LL_y$ | $UL_y$ $LL_y$ |
| [m] | $UL_y$ $LL_y$ $V_x$ | $UL_y$ $LL_y$ $V_x$ | $LL_y$ $UL_y$ |
| [t] | $TT_y$ $TT_x$ | $TT_y$ | $TT_y$ $LI_y$ |
| [d] | $TT_y$ $TT_x$ | $TT_y$ | $TT_y$ |
| [n] | $TT_y$ $TT_x$ $V_x$ | $TT_y$ $V_x$ | $TT_y$ |
| [k] | $TD_y$ | $TD_y$ | $TD_y$ |
| [g] | $TD_y$ | $TD_y$ | $TD_y$ |
| [ŋ] | $TD_y$ $V_x$ | $TD_y$ $V_x$ | $TD_y$ |
| [f] | $LL_y$ $LL_x$ | $LL_y$ $UL_y$ | $LL_y$ $LL_x$ |
| [v] | $LL_y$ $LL_x$ | $LL_y$ $UL_y$ | $LL_y$ $LL_x$ |
| [θ] | $TT_y$ $TT_x$ | $TT_x$ $TT_y$ $LL_y$ | $TT_y$ $TT_x$ $TB_y$ |
| [ð] | $TT_y$ $TT_x$ | $TT_x$ $TT_y$ | $TT_x$ $TT_y$ |
| [s] | $TT_y$ $TT_x$ | $LI_y$ $TT_x$ $TT_y$ | $LI_y$ $TT_y$ $TB_x$ |
| [z] | $TT_y$ $TT_x$ | $LI_y$ $TT_x$ $TT_y$ | $LI_y$ $TT_y$ $TT_x$ |
| [ʃ] | $TT_y$ $TT_x$ | $TT_y$ $TB_x$ $LI_y$ $TD_y$ | $TT_y$ $LI_y$ $TT_x$ |
| [ʒ] | $TT_y$ $TT_x$ | $LI_y$ $TT_y$ $TD_y$ $TT_x$ $LL_y$ | $TT_y$ $LI_y$ $TB_y$ $TD_x$ $LL_x$ |
| [ʧ] | $TT_y$ $TT_x$ | $LI_y$ $TT_y$ $TB_x$ $TB_y$ | $TT_y$ $LI_y$ $TT_x$ |
| [ʤ] | $TT_y$ $TT_x$ | $TT_y$ $TB_y$ $TT_x$ $LI_y$ | $LI_y$ $TT_y$ $TT_x$ |
| [l] | $TT_y$ $TT_x$ | - | - |
| [ɹ] | $TT_y$ $TT_x$ | $TB_x$ | $TT_x$ |
| [w] | $UL_x$ $LL_x$, $TD_y$ | $UL_y$ | $UL_y$ $TD_y$ $LL_y$ |
| [j] | $TB_y$ $TB_x$ | $TB_y$ | $TB_y$ |
| [h] | - | - | $TT_y$ |

Table C.5: *Expected (from IPA) and identified 1D critical coordinates for consonants for male and female speakers.*

| Vowels | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|
| [æ](near-open) | $TT_y$ | $LL_y$ | $LL_y$ $TD_y$ |
| [ɛ](open-mid) | $TT_y$ | $LL_y$ | $LL_y$ |
| [ɪ](close) | $TT_y$ | - | - |
| [iː] (close) | $TT_y$ | $TD_y$ $LL_y$ $TT_x$ | $TB_y$ |
| [i] (close) | $TT_y$ | $TD_y$ | $TD_y$ $TT_x$ |
| [ə](mid) | $TB_y$ | - | - |
| [ɚ](rhotacized) | $TB_y$ | $LL_y$ | $TT_y$ $LL_x$ |
| [ʌ](mid) | $TB_y$ | - | $TB_y$ |
| [ɑ] (open) | $TD_y$ | $LL_y$ $TB_y$ $TB_x$ | $TB_y$ $TD_x$ |
| [ɒ] (open rounded) | $TD_y$ $UL_x$ $LL_x$ | $TB_y$ | $TB_y$ $TD_x$ |
| [ɔ] (mid rounded) | $TD_y$ $UL_x$ $LL_x$ | $LL_y$ $TB_x$ $TT_y$ | $TD_x$ $LL_y$ $TD_y$ $TB_y$ |
| [ʊ] (near-close rounded) | $TD_y$ $UL_x$ $LL_x$ | - | - |
| [u] (close rounded) | $TD_y$ $UL_x$ $LL_x$ | $TD_y$ | $TB_y$ |

Table C.6: *Expected (from IPA) and identified 1D critical coordinates for front, mid and back vowels for male and female speakers.*

| Diphthongs | | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|---|
| [aɪ] | [a] (front open) | $TT_y$ | $LL_y$ $TD_y$ | $TT_y$ $LL_y$ $TD_y$ |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ | $LL_y$ |
| [eɪ] | [e] (front close) | $TT_y$ | $LL_y$ | $LL_y$ $TB_y$ |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ $TT_x$ $TD_y$ | $TD_y$ $LL_y$ $TB_x$ |
| [ɛə] | [ɛ] (front mid) | $TT_y$ | $LL_y$ | $LI_y$ $UL_y$ |
| | [ə] (center mid) | $TB_y$ | $LL_y$ | $TB_y$ $LI_y$ |
| [ɪə] | [ɪ] (front close) | $TT_y$ | $LL_y$ $TT_x$ | $TB_y$ $LL_y$ |
| | [ə] (center mid) | $TB_y$ | $LL_y$ | $UL_y$ |
| [ɔɪ] | [ɔ] (back mid rounded) | $TD_y$ $UL_x$ $LL_x$ | $TT_x$ $LL_y$ $TT_y$ | $TT_x$ $TD_y$ $TT_y$ $LL_y$ |
| | [ɪ] (front close) | $TT_y$ | $LL_y$ $TD_y$ $UL_x$ | $TD_y$ |
| [oʊ] | [o] (back mid rounded) | $TD_y$ $UL_x$ $LL_x$ | - | - |
| | [ʊ] (back close rounded) | $TD_y$ $UL_x$ $LL_x$ | $UL_y$ $LL_y$ | - |
| [aʊ] | [a](front open) | $TT_y$ | $LL_y$ $TD_y$ $TT_y$ | $TT_y$ $TD_y$ |
| | [ʊ] (back close rounded) | $TD_y$ $UL_x$ $LL_x$ | $TB_y$ $TB_x$ | $TB_x$ |

Table C.7: *Expected (from IPA) and identified 1D critical coordinates for diphthongs for male and female speakers.*

| Consonants | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|
| [p] | UL LL | UL | UL |
| [b] | UL LL | UL LL | UL LL |
| [m] | UL LL V | UL LL V | LL UL |
| [t] | TT | TT | TT |
| [d] | TT | TT | TT |
| [n] | TT V | TT V | TT |
| [k] | TD | TD | TD |
| [g] | TD | TD | TD |
| [ŋ] | TD V | TD VV | TD |
| [f] | LL | LL | LL |
| [v] | LL | LL | LL |
| [θ] | TT | TT LL | TT TB LL LI |
| [ð] | TT | TT | TT |
| [s] | TT | TT LI | LI TT |
| [z] | TT | TT LI | LI TT |
| [ʃ] | TT | TT LI TD | LI TT |
| [ʒ] | TT | LI TT TD LL | TT LI TB |
| [ʧ] | TT | TT LI | LI TT UL |
| [ʤ] | TT | TT TB LI | LI TT |
| [l] | TT | - | - |
| [ɹ] | TT | TT | TT |
| [w] | UL LL TD | UL TT | UL TD LL |
| [j] | TB | TB | TB TT |
| [h] | - | TT | TT |

Table C.8: *Expected (from IPA) and identified 2D critical coordinates for consonants for male and female speakers.*

| Vowels | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|
| [æ] ( near-open) | TT | LL | LL TD |
| [ɛ] (open-mid) | TT | LL | LL |
| [ɪ] (close) | TT | - | - |
| [iː] (close) | TT | TB TT | TB TD |
| [i] (close) | TT | TB | TB TD |
| [ə] (mid) | TB | - | - |
| [ɚ] (rhotacized) | TB | - | TB LL |
| [ʌ] (mid) | TB | - | TB |
| [ɑ] (open) | TD | LL TB | TB |
| [ɒ] (open rounded) | TD UL LL | TB | TB |
| [ɔ] (mid rounded) | TD UL LL | TB LL | TD LL |
| [ʊ] (near close rounded) | TD UL LL | - | - |
| [u] (close rounded) | TD UL LL | TD | TD |

Table C.9: *Expected (from IPA) and identified 2D critical coordinates for front, mid and back vowels for male and female speakers.*

| Diphthongs | | Expected (IPA) | Identified (male) | Identified (female) |
|---|---|---|---|---|
| [aɪ] | [a] (front open) | TT | LL TD | TT |
| | [ɪ] (front close) | TT | LL | LL |
| [eɪ] | [e] (front close) | TT | LL | LL TB |
| | [ɪ] (front close) | TT | LL TT | TB LL |
| [ɛə] | [ɛ] (front mid) | TT | LL | LI TD |
| | [ə] (centre mid) | TB | LI | TB |
| [ɪə] | [ɪ] (front close) | TT | LL TT | TB TT UL |
| | [ə] (centre mid) | TB | LL | UL |
| [ɔɪ] | [ɔ] (back mid rounded) | TD UL LL | LL TT | TD V LL |
| | [ɪ] (front close) | TT | LL TD | TD V |
| [oʊ] | [o] (back mid rounded) | TD UL LL | - | - |
| | [ʊ] (back close rounded) | TD UL LL | UL | - |
| [aʊ] | [a] (front open) | TT | LL TB | TT TD |
| | [ʊ] (back close rounded) | TD UL LL | TB | TB |

Table C.10: *Expected (from IPA) and identified 2D critical coordinates for diphthongs for male and female speakers.*

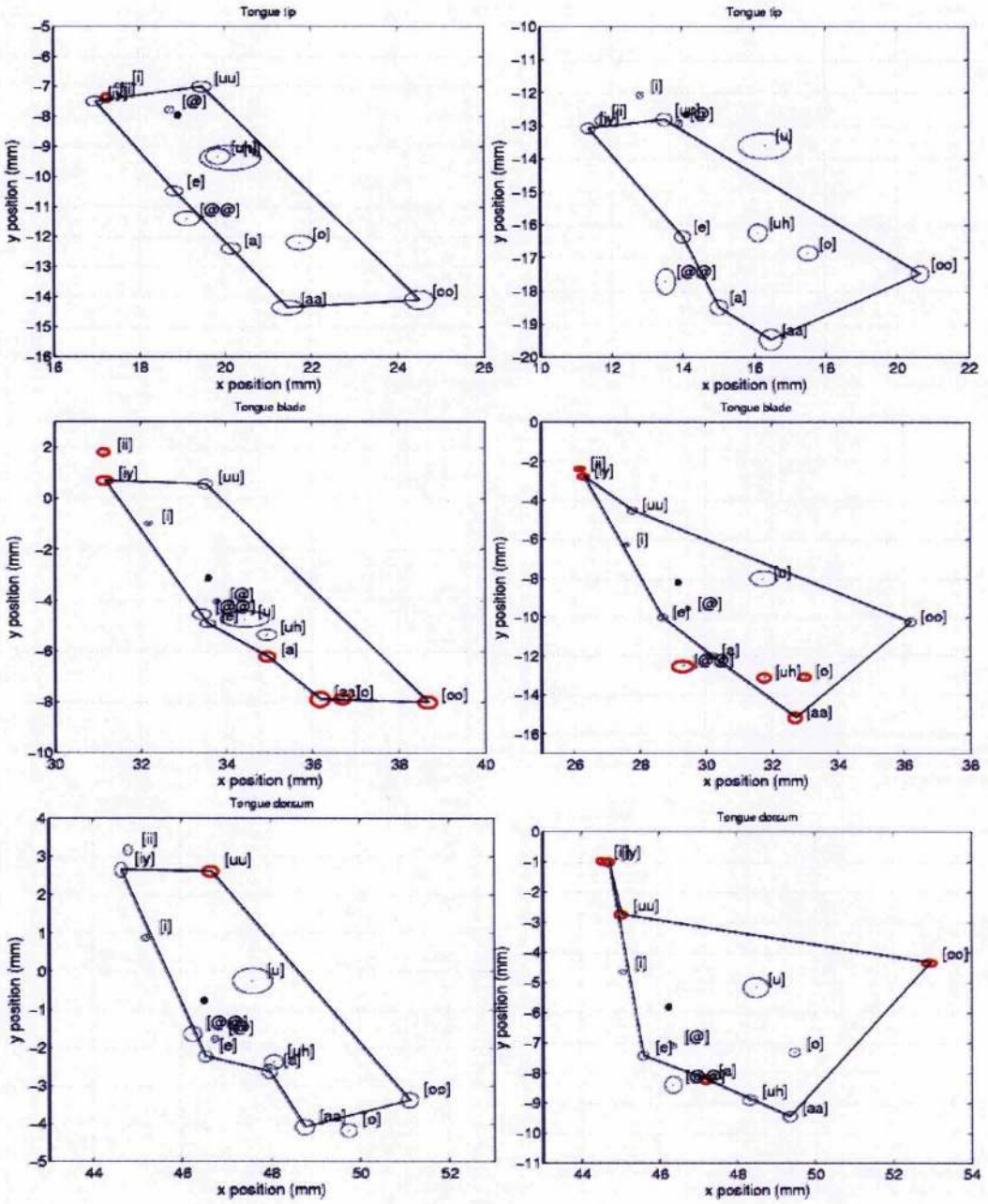



Figure C.1: *Articulatory vowel quadrilaterals for* TT *(top),* TB *(centre) and* BOTTOM *(right) obtained by joining the model means of vowels closest to the primary cardinals in the database for male (left) and female (right) speakers. The standard errors of model means are depicted using covariance ellipses for vowels for which the articulatory coordinate is critical (thick red) and not critical (thin black).*

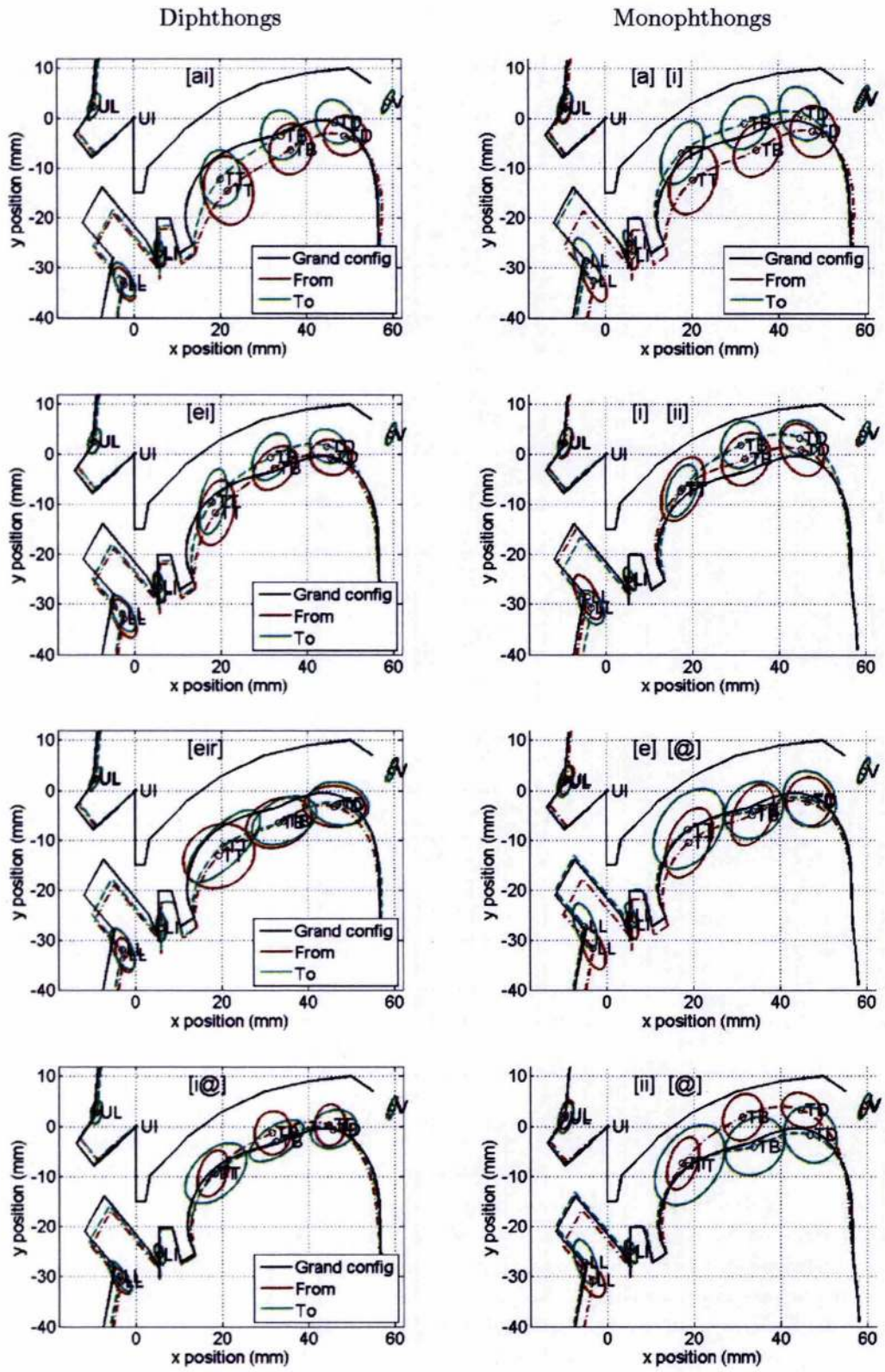



Table C.11: *Comparison of diphthong and monophthong realisations for male speaker. Diphthongs from top to bottom :*[aɪ], [eɪ], [ɛə], [ɪə].

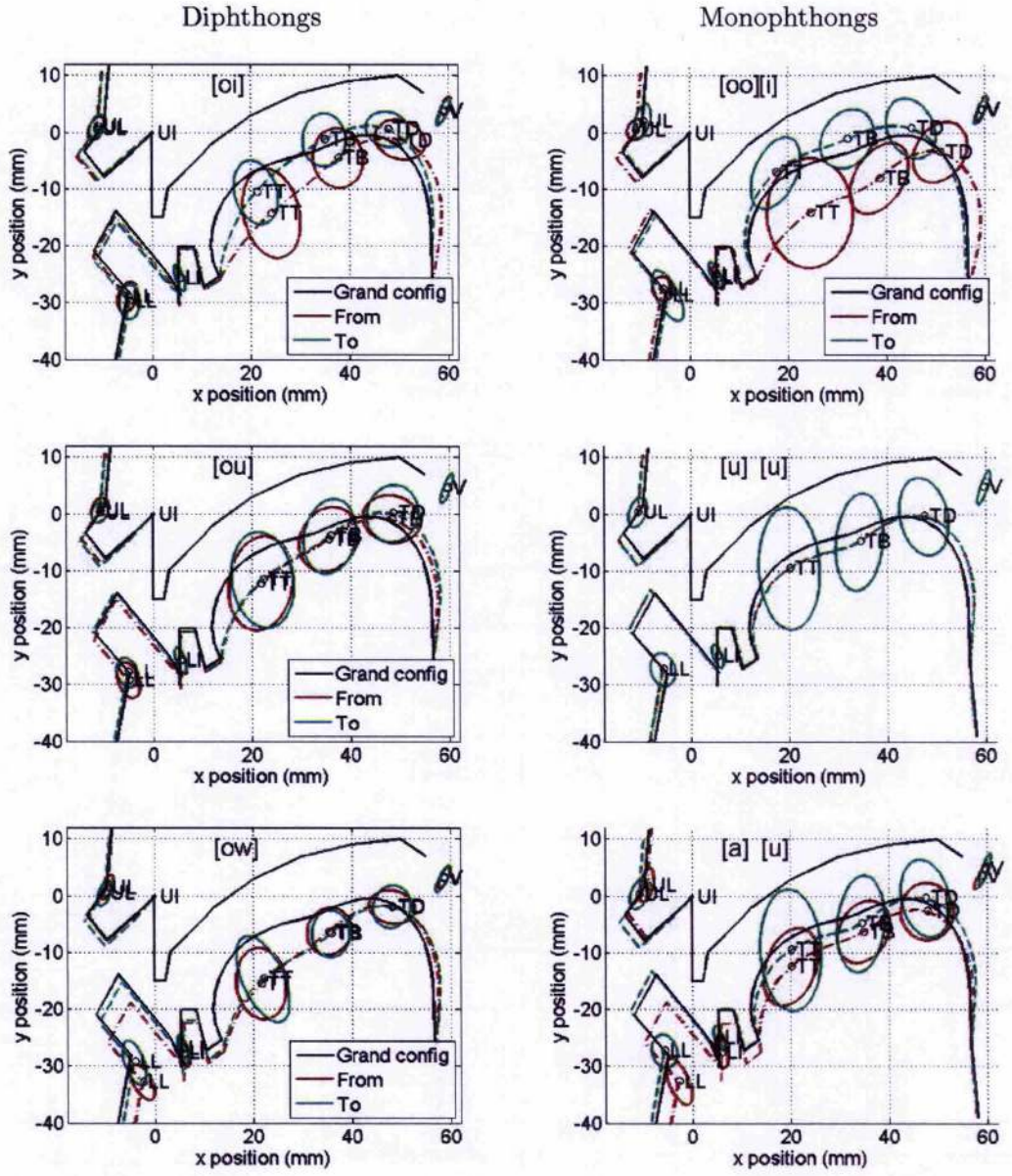



Table C.12: *Male speaker comparison plots. Diphthongs from top to bottom: /ɔɪ/, /oʊ/, /aʊ/.*

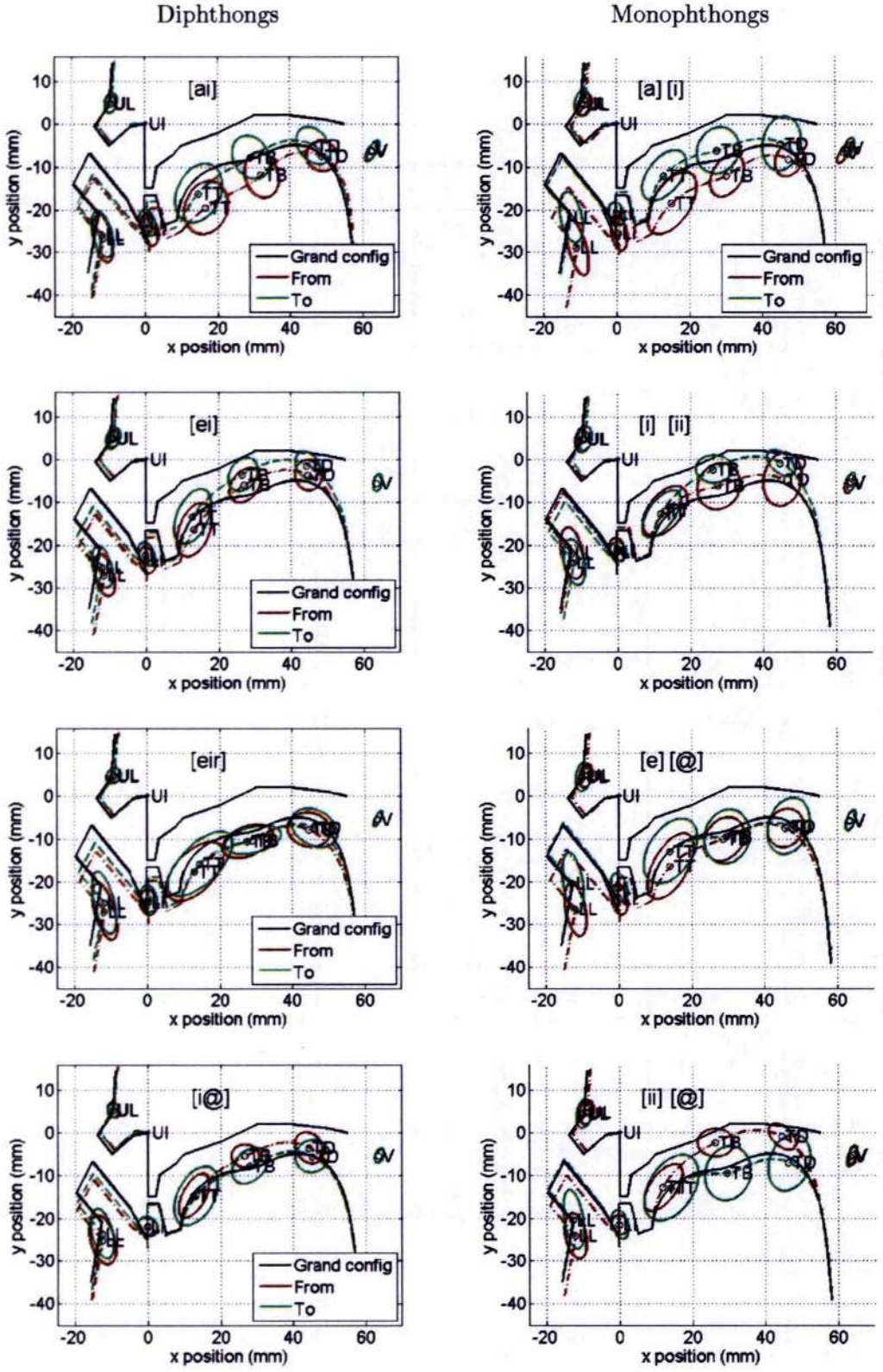



Table C.13: *Female speaker comparison plots. Diphthongs from top to bottom :[aɪ], [eɪ], [ɛə], [ɪə].*

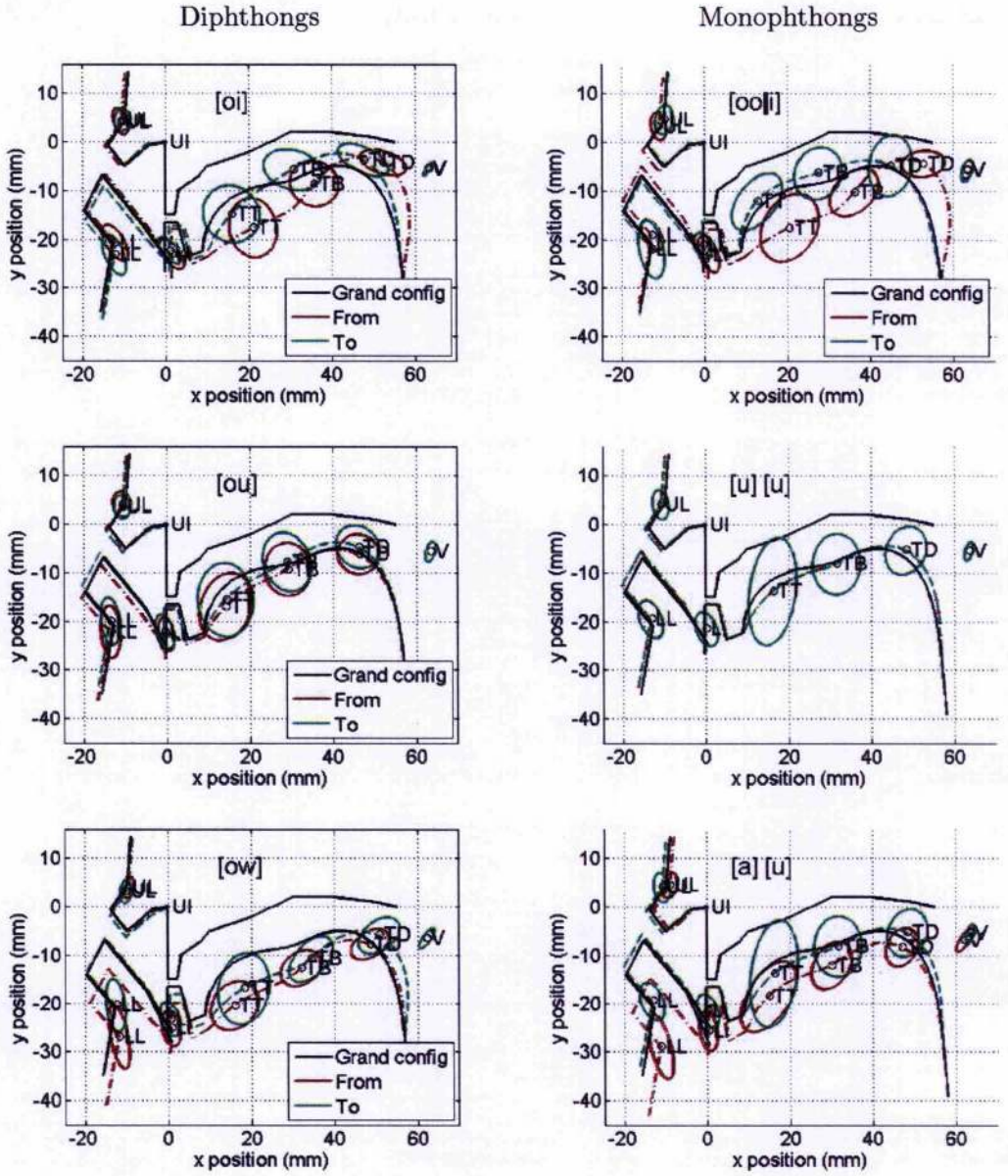



Table C.14: *Female speaker comparison plots. Diphthongs from top to bottom: [ɔɪ], [oʊ], [aʊ].*

## C.3 Comparison with exhaustive search

| Phone | Identified critical coordinates DFS | Identified critical coordinates ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | ULy LLy | ULy LLy | 39 | 39 | 0.8 | 0.8 |
| [b] | ULy LLy | ULy LLy | 32 | 32 | 0.6 | 0.6 |
| [m] | ULy LLy Vx | LLy ULy Vx | 40 | 40 | 0.6 | 0.3 |
| [t] | TTy | TTy | 49 | 49 | 0.7 | 0.7 |
| [d] | TTy | TTy | 46 | 46 | 0.9 | 0.9 |
| [n] | TTy Vx | TTy Vx | 44 | 44 | 0.8 | 0.8 |
| [k] | TDy | TDy | 33 | 33 | 1.1 | 1.1 |
| [g] | TDy | TDy | 35 | 35 | 1.2 | 1.2 |
| [ŋ] | TDy Vx | TDy Vx | 39 | 39 | 0.7 | 0.7 |
| [f] | LLy ULy | ULx LLy | 48 | 49 | 1.6 | 1.3 |
| [v] | LLy ULy | ULx LLy | 46 | 46 | 1.6 | 0.9 |
| [θ] | TTx TTy LLy | TTx LLx TTy | 44 | 45 | 0.9 | 0.8 |
| [ð] | TTx TTy | TTx TTy | 32 | 32 | 1.2 | 1.2 |
| [s] | LIy TTx TTy | TTx TTy LIy | 30 | 30 | 0.9 | 0.9 |
| [z] | LIy TTx TTy | LIy TBx TTy | 29 | 29 | 1.2 | 1.2 |
| [ʃ] | TTy TBx LIy TDy | LIy TTy TTx TBy | 25 | 26 | 1.5 | 0.7 |
| [ʒ] | LIy TTy TDy TTx LLy | LIy TBx TBy | 245 | 432 | 0.7 | 1.4 |
| [ʧ] | LIy TTy TBx TBy | LIy TTx TTy TBy | 24 | 27 | 0.8 | 0.6 |
| [ʤ] | TTy TBy TTx LIy | LIy TTx TTy TBy | 25 | 25 | 0.9 | 0.9 |
| [l] | | LIy | 45 | 43 | 1.6 | 1.0 |
| [ɹ] | TBx | TBx | 49 | 49 | 0.9 | 0.9 |
| [w] | ULy | ULy | 49 | 49 | 1.5 | 1.5 |
| [j] | TBy | TDy | 44 | 44 | 0.7 | 0.7 |
| [h] | | | 65 | 65 | 1.3 | 1.3 |
| [æ] | LLy | LIy | 47 | 48 | 1.2 | 0.8 |
| [ɛ] | LLy | LLy | 46 | 46 | 0.5 | 0.5 |
| [ɪ] | | | 47 | 47 | 0.8 | 0.8 |
| [i:] | TDy LLy TTx | LLy TBx TBy | 25 | 28 | 0.9 | 0.6 |
| [i] | TDy | TBy | 54 | 54 | 1.5 | 1.5 |
| [ə] | | | 40 | 40 | 0.3 | 0.3 |
| [ɚ] | LLy | LLy | 80 | 80 | 1.2 | 1.2 |
| [ʌ] | | | 43 | 43 | 0.9 | 0.9 |
| [ɑ] | LLy TBy TBx | LIy TDy | 34 | 45 | 0.6 | 1.2 |
| [ɒ] | TBy | TDy | 40 | 41 | 1.3 | 1.2 |
| [ɔ] | LLy TBx TTy | LLy TTx TBy | 41 | 48 | 0.7 | 0.6 |
| [ʊ] | | | 56 | 56 | 1.5 | 1.5 |
| [u] | TDy | TDy | 43 | 43 | 1.4 | 1.4 |
| [aɪ] | LLy TDy | LLy TDy | 39 | 39 | 1.2 | 1.2 |
| [aɪ] | LLy | LLy | 53 | 53 | 0.8 | 0.8 |
| [eɪ] | LLy | LLy | 58 | 58 | 1.0 | 1.0 |
| [eɪ] | LLy TTx TDy | LLy TTx TDy | 27 | 27 | 0.3 | 0.3 |
| [ɛə] | LLy | LIy | 86 | 99 | 1.5 | 1.5 |
| [ɛə] | LLy | LIy | 180 | 183 | 1.5 | 1.0 |
| [ɪə] | LLy TTx | LLy TTx | 142 | 142 | 1.2 | 1.2 |
| [ɪə] | LLy | LLx | 208 | 220 | 0.9 | 0.9 |
| [ɔɪ] | TTx LLy TTy | TTx LLx TTy | 78 | 78 | 1.2 | 1.1 |
| [ɔɪ] | LLy TDy ULx | ULx LLy TDy | 118 | 118 | 0.8 | 0.8 |
| [oʊ] | | LIx | 52 | 50 | 1.5 | 1.1 |
| [oʊ] | ULy LLy | ULy LLy | 46 | 46 | 0.8 | 0.8 |
| [aʊ] | LLy TDy TTy | LLy TBy | 59 | 70 | 1.5 | 1.4 |
| [aʊ] | TBy TBx | TTy TDx | 42 | 49 | 1.4 | 0.8 |

Table C.15: *1D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at IPA critical threshold for male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | Identified critical coordinates ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | ULy | ULy | 43 | 43 | 1.3 | 1.3 |
| [b] | ULy LLy | ULy LLy | 43 | 43 | 1.0 | 1.0 |
| [m] | LLy ULy | ULy LLy | 37 | 38 | 1.3 | 1.3 |
| [t] | TTy LIy | TTy LIy | 25 | 25 | 0.5 | 0.5 |
| [d] | TTy | TTy | 35 | 35 | 1.3 | 1.3 |
| [n] | TTy | TTy Vx | 32 | 31 | 1.5 | 0.5 |
| [k] | TDy | TDy | 25 | 25 | 0.3 | 0.3 |
| [g] | TDy | TDy | 25 | 25 | 0.3 | 0.3 |
| [ŋ] | TDy | TDy | 23 | 23 | 1.1 | 1.1 |
| [f] | LLy LLx | LLx LLy | 34 | 35 | 0.7 | 0.7 |
| [v] | LLy LLx | LLx LLy | 35 | 35 | 0.9 | 0.9 |
| [θ] | TTy TTx TBy | TBy TTx TTy | 38 | 36 | 1.3 | 1.1 |
| [ð] | TTx TTy | TTx TTy | 26 | 26 | 1.3 | 1.3 |
| [s] | LIy TTy TBx | LIy TTy TDx | 19 | 20 | 1.4 | 1.2 |
| [z] | LIy TTy TTx | LIy TTx TTy | 24 | 24 | 1.0 | 1.0 |
| [ʃ] | TTy LIy TTx | LIy TTy TBx | 39 | 39 | 1.4 | 1.4 |
| [ʒ] | TTy LIy TBy TDx LLx | TBy LIy TDx TTy LLx | 116 | 116 | 0.8 | 0.8 |
| [ʧ] | TTy LIy TTx | LLx LIy TTy TBx | 30 | 27 | 1.6 | 1.1 |
| [ʤ] | LIy TTy TTx | LIy TTx TTy | 27 | 27 | 1.2 | 1.2 |
| [l] | | | 26 | 26 | 0.9 | 0.9 |
| [ɹ] | TTx | TTx | 34 | 34 | 0.7 | 0.7 |
| [w] | ULy TDy LLy | ULy LLy TDy | 27 | 27 | 1.3 | 1.3 |
| [j] | TBy | TBy | 35 | 35 | 1.0 | 1.0 |
| [h] | TTy | LIy | 41 | 44 | 1.0 | 0.7 |
| | | | | | | |
| [æ] | LLy TDy | LLy TDy | 42 | 42 | 0.7 | 0.7 |
| [ɛ] | LLy | LIy | 42 | 43 | 1.1 | 0.8 |
| [ɪ] | | | 32 | 32 | 0.7 | 0.7 |
| [iː] | TBy | TBy | 30 | 30 | 1.5 | 1.5 |
| [i] | TDy TTx | TBx TDy | 25 | 24 | 1.3 | 1.1 |
| [ə] | | | 30 | 30 | 0.2 | 0.2 |
| [ɚ] | TTy LLx | LIy TBy | 36 | 36 | 1.5 | 1.1 |
| [ʌ] | TBy | TBy | 40 | 40 | 1.1 | 1.1 |
| [ɑ] | TBy TDx | TBy TDx | 33 | 33 | 0.9 | 0.9 |
| [ɒ] | TBy TDx | TBx TBy | 30 | 30 | 0.8 | 0.7 |
| [ɔ] | TDx LLy TDy TBy | LLy TBy TDx TDy | 16 | 16 | 0.2 | 0.2 |
| [ʊ] | | LLy | 43 | 39 | 1.6 | 0.8 |
| [u] | TBy | TBy | 32 | 32 | 1.2 | 1.2 |
| | | | | | | |
| [aɪ] | TTy LLy TDy | LIy TDy | 51 | 55 | 1.4 | 0.9 |
| [aɪ] | LLy | LLy | 40 | 40 | 0.5 | 0.5 |
| [eɪ] | LLy TBy | LLy TBy | 37 | 37 | 1.2 | 1.2 |
| [eɪ] | TDy LLy TBx | ULx TDy | 20 | 37 | 0.5 | 1.4 |
| [ɛə] | LIy ULy | LLy TBy | 79 | 68 | 1.3 | 1.1 |
| [ɛə] | TBy LIy | LLy TBy | 77 | 76 | 1.1 | 0.8 |
| [ɪə] | TBy LLy | ULx TBy | 101 | 99 | 1.6 | 1.4 |
| [ɪə] | ULy | ULy | 129 | 129 | 1.0 | 1.0 |
| [ɔɪ] | TTx TDy TTy LLy | LIy TDx TDy | 54 | 65 | 0.9 | 1.4 |
| [ɔɪ] | TDy | TDy | 60 | 60 | 1.5 | 1.5 |
| [oʊ] | | | 42 | 42 | 1.4 | 1.4 |
| [oʊ] | | | 37 | 37 | 0.7 | 0.7 |
| [aʊ] | TTy TDy | LIy TBy | 52 | 53 | 1.4 | 1.2 |
| [aʊ] | TBx | TBx | 42 | 42 | 1.1 | 1.1 |

Table C.16: *1D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at IPA critical threshold for female speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | Identified critical coordinates ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | UL LL | LL UL | 34 | 33 | 0.6 | 0.6 |
| [b] | UL LL | UL LL | 28 | 28 | 0.5 | 0.5 |
| [m] | UL LL V | LL UL V | 35 | 35 | 0.7 | 0.7 |
| [t] | TT | TT | 26 | 26 | 0.6 | 0.6 |
| [d] | TT | TT | 24 | 24 | 0.3 | 0.3 |
| [n] | TT V | TT V | 23 | 23 | 0.3 | 0.3 |
| [k] | TD | TD | 25 | 25 | 1.0 | 1.0 |
| [g] | TD | TD | 29 | 29 | 1.7 | 1.8 |
| [ŋ] | TD V | TD V | 33 | 33 | 0.8 | 0.8 |
| [f] | LL | UL LL | 42 | 39 | 2.2 | 1.7 |
| [v] | LL | LL | 40 | 40 | 2.0 | 2.0 |
| [θ] | TT LL | LL TT | 40 | 40 | 1.4 | 1.4 |
| [ð] | TT | TT | 26 | 26 | 1.2 | 1.2 |
| [s] | TT LI | LI TT | 25 | 25 | 1.0 | 1.0 |
| [z] | TT LI | TT LI | 24 | 24 | 1.7 | 1.7 |
| [ʃ] | TT LI TD | TT LI TB | 16 | 15 | 1.1 | 0.9 |
| [ʒ] | LI TT TD LL | LL TT LI TD | 173 | 173 | 0.9 | 1.0 |
| [ʧ] | TT LI | LI TT TB | 31 | 16 | 2.2 | 0.9 |
| [ʤ] | TT TB LI | LI TT TB | 15 | 15 | 1.0 | 1.0 |
| [l] | | TD | 38 | 28 | 2.3 | 1.5 |
| [ɹ] | TT | TT | 44 | 44 | 0.7 | 0.7 |
| [w] | UL TT | TT UL | 31 | 31 | 1.8 | 1.8 |
| [j] | TB | TB | 32 | 32 | 1.6 | 1.6 |
| [h] | TT | TT | 32 | 32 | 0.9 | 0.9 |
| [æ] | LL | LI | 41 | 40 | 2.2 | 1.0 |
| [ɛ] | LL | LI | 41 | 42 | 1.0 | 0.9 |
| [ɪ] | | | 40 | 40 | 1.8 | 1.8 |
| [i:] | TB TT | TT TB | 21 | 21 | 1.6 | 1.6 |
| [i] | TB | TB | 30 | 30 | 1.9 | 1.9 |
| [ə] | | | 34 | 34 | 0.3 | 0.3 |
| [ɚ] | | TT | 74 | 47 | 2.2 | 1.7 |
| [ʌ] | | | 37 | 37 | 1.8 | 1.8 |
| [ɑ] | LL TB | LL TB | 26 | 26 | 0.7 | 0.7 |
| [ɒ] | TB | TB | 26 | 26 | 1.4 | 1.4 |
| [ɔ] | TB LL | LL TB | 34 | 34 | 1.3 | 1.3 |
| [ʊ] | | | 49 | 49 | 1.9 | 1.9 |
| [u] | TD | TD | 32 | 32 | 1.8 | 1.8 |
| [aɪ] | LL TD | TB LL | 26 | 22 | 1.2 | 0.5 |
| [aɪ] | LL | LL | 46 | 46 | 1.7 | 1.7 |
| [eɪ] | LL | LL | 52 | 52 | 1.6 | 1.6 |
| [eɪ] | LL TT | TB LI | 28 | 31 | 1.9 | 1.2 |
| [ɛə] | LL | LL | 74 | 74 | 2.0 | 2.0 |
| [ɛə] | LI | UL LI | 163 | 152 | 2.1 | 1.4 |
| [ɪə] | LL TT | LL TT | 110 | 110 | 1.0 | 1.0 |
| [ɪə] | LL | LL TD | 186 | 131 | 2.0 | 1.0 |
| [ɔɪ] | LL TT | LL TT | 68 | 68 | 1.8 | 1.8 |
| [ɔɪ] | LL TD | LL TD | 75 | 75 | 1.7 | 1.7 |
| [oʊ] | | | 46 | 46 | 1.9 | 1.9 |
| [oʊ] | UL LL | UL | 40 | 45 | 0.9 | 2.1 |
| [aʊ] | LL TB | LL TB | 38 | 38 | 1.0 | 1.0 |
| [aʊ] | TB | TT | 35 | 42 | 1.7 | 1.2 |

Table C.17: *2D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at IPA critical threshold for male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | UL | UL | 38 | 38 | 1.8 | 1.8 |
| [b] | UL LL | UL LL | 36 | 36 | 1.2 | 1.2 |
| [m] | LL UL | UL LL | 31 | 31 | 1.4 | 1.4 |
| [t] | TT | TT | 18 | 18 | 1.6 | 1.6 |
| [d] | TT | TT | 22 | 22 | 1.1 | 1.1 |
| [n] | TT | TT | 22 | 22 | 1.6 | 1.6 |
| [k] | TD | TD | 22 | 22 | 0.3 | 0.3 |
| [g] | TD | TD | 23 | 23 | 0.4 | 0.4 |
| [ŋ] | TD | TD | 19 | 19 | 1.1 | 1.1 |
| [f] | LL | LL | 34 | 34 | 1.1 | 1.1 |
| [v] | LL | LL | 34 | 34 | 1.2 | 1.2 |
| [θ] | TT TB | LI LL TT | 32 | 33 | 2.0 | 1.5 |
| [ð] | TT | TT | 24 | 24 | 1.3 | 1.3 |
| [s] | LI TT | TT LI | 16 | 16 | 1.6 | 1.6 |
| [z] | LI TT | TT LI | 19 | 19 | 1.1 | 1.1 |
| [ʃ] | LI TT | LI TT | 31 | 31 | 1.8 | 1.8 |
| [ʒ] | TT LI TB | LI TB TT | 94 | 94 | 1.6 | 1.6 |
| [ʧ] | LI TT UL | UL LI TT | 21 | 21 | 1.1 | 1.1 |
| [ʤ] | LI TT | TT LI | 22 | 22 | 1.7 | 1.7 |
| [l] | | | 25 | 25 | 1.6 | 1.6 |
| [ɹ] | TT | TT | 27 | 27 | 0.7 | 0.7 |
| [w] | UL TD LL | UL LL TD | 25 | 24 | 1.0 | 1.0 |
| [j] | TB TT | TD | 26 | 32 | 1.5 | 1.5 |
| [h] | TT | TT | 30 | 30 | 1.0 | 1.0 |
| [æ] | LL TD | TD LL | 24 | 24 | 0.9 | 0.9 |
| [ɛ] | LL | LI | 40 | 40 | 1.2 | 1.0 |
| [ɪ] | | | 31 | 31 | 1.2 | 1.2 |
| [iː] | TB TD | TD | 17 | 20 | 1.4 | 1.3 |
| [i] | TB TD | TD | 18 | 22 | 1.3 | 1.6 |
| [ə] | | | 29 | 29 | 0.3 | 0.3 |
| [ɚ] | TB LL | LL TB | 27 | 27 | 1.6 | 1.6 |
| [ʌ] | TB | TB | 25 | 25 | 1.3 | 1.3 |
| [ɑ] | TB | TB | 29 | 29 | 1.2 | 1.2 |
| [ɒ] | TB | TB | 28 | 28 | 0.9 | 0.9 |
| [ɔ] | TD LL | LL TD | 18 | 18 | 1.5 | 1.5 |
| [ʊ] | | | 41 | 41 | 1.8 | 1.8 |
| [u] | TD | TB | 25 | 25 | 1.9 | 1.6 |
| [aɪ] | TT LL TD | LL TB | 29 | 31 | 0.6 | 1.0 |
| [aɪ] | LL | LL | 38 | 38 | 1.0 | 1.0 |
| [eɪ] | LL TB | LL TD | 23 | 24 | 1.4 | 1.0 |
| [eɪ] | TB LL | LL TD | 18 | 18 | 1.3 | 0.8 |
| [ɛə] | LI TD | LL TB | 60 | 59 | 1.7 | 1.2 |
| [ɛə] | TB | TB | 81 | 81 | 1.8 | 1.8 |
| [ɪə] | TB TT UL | TD LL | 43 | 75 | 0.9 | 1.8 |
| [ɪə] | UL | UL | 117 | 117 | 1.7 | 1.7 |
| [ɔɪ] | TD V LL | LL TD V | 51 | 51 | 1.8 | 1.8 |
| [ɔɪ] | TD V | LL V TD | 45 | 41 | 1.9 | 1.2 |
| [oʊ] | | | 42 | 42 | 1.6 | 1.6 |
| [oʊ] | | | 36 | 36 | 1.0 | 1.0 |
| [aʊ] | TT TD | LI TB | 30 | 33 | 1.8 | 1.4 |
| [aʊ] | TB | TB | 33 | 33 | 1.6 | 1.6 |

Table C.18: *2D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at IPA critical threshold for female speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | Identified critical coordinates ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | **ULy** **LLy** **Vx** | **ULy** **Vx** **LLy** | **36** | **37** | **0.2** | **0.2** |
| [b] | ULy LLy | ULy **Vx** LLy | 32 | 31 | 0.6 | 0.2 |
| [m] | **ULy** **LLy** **Vx** | **LLy** **ULy** **Vx** | **40** | **40** | **0.6** | **0.3** |
| [t] | TTy TTx | TTy TBx | 32 | 30 | 0.6 | 0.5 |
| [d] | TTy TTx | TTy TBx | 29 | 28 | 0.3 | 0.2 |
| [n] | **TTy** **Vx** **TBx** | **TTy** **TBx** **Vx** | **28** | **28** | **0.1** | **0.1** |
| [k] | TDy TBy | TTy TDy | 32 | 32 | 0.5 | 0.5 |
| [g] | TDy TBy Vy | TTy TDy Vy | 33 | 33 | 0.5 | 0.5 |
| [ŋ] | TDy Vx LIx | TBy TDy Vx | 38 | 35 | 0.5 | 0.3 |
| [f] | LLy ULy LIy | LIx **TBy** LLy ULy | 46 | 43 | 0.6 | 0.4 |
| [v] | LLy ULy LIy LLx **Vx** | ULx ULy LLy LIy | 41 | 41 | 0.5 | 0.5 |
| [θ] | TTx TTy LLy LIy | TBx TBy LLy TTy | 43 | 42 | 0.5 | 0.4 |
| [ð] | TTx TTy TBy | **LLy** TBy TTy TBx | 27 | 24 | 0.6 | 0.4 |
| [s] | **LIy** **TTx** **TTy** **LLy** **ULy** | **ULy** **LLy** **LIy** **TTx** **TTy** | **25** | **25** | **0.5** | **0.5** |
| [z] | **LIy** **TTx** **TTy** **LLy** | **LLy** **LIy** **TTy** **TTx** | **25** | **25** | **0.4** | **0.4** |
| [ʃ] | TTy TBx LIy TDy LLy | ULy LIy TTy TDx TBy | 22 | 27 | 0.4 | 0.4 |
| [ʒ] | LIy TTy TDy TTx LLy Vy | TDx TDy TTy LLy LIy Vy | 229 | 267 | 0.6 | 0.6 |
| [ʧ] | LIy TTy TBx TBy Vx LLy | LLy TTy TBx Vx LIy TDy | 19 | 20 | 0.3 | 0.2 |
| [ʤ] | TTy TBy TTx LIy Vx TDy | TTy TBx Vx TBy LLy LIy | 21 | 19 | 1.4 | 0.4 |
| [l] | TBy TTy LIx TBx | LIx TTx TDy TTy | 27 | 28 | 0.3 | 0.1 |
| [ɹ] | **TBx** **TTy** | **TTy** **TBx** | **45** | **45** | **0.4** | **0.4** |
| [w] | **ULy** **TTy** **LLy** **TTx** **TDy** | **ULy** **LLy** **TTx** **TTy** **TDy** | **33** | **34** | **0.2** | **0.2** |
| [j] | TBy TTy | TTy TBy **TDx** | 41 | 36 | 0.6 | 0.5 |
| [h] | LLy TTx | LIy TBx | 43 | 50 | 0.6 | 0.4 |
| [æ] | **LLy** TBy **TBx** | LLy TDy | 32 | 42 | 0.6 | 0.4 |
| [ɛ] | **LLy** | **LLy** | **46** | **46** | **0.5** | **0.5** |
| [ɪ] | TBy TTx | TBx TBy | 29 | 33 | 0.3 | 0.3 |
| [i:] | **TDy** **LLy** **TTx** **TTy** | **TTx** **TTy** **LLy** **TDy** | **20** | **21** | **0.4** | **0.2** |
| [i] | TDy TTx LLy TTy | LLx TTx TTy TDy | 26 | 28 | 0.3 | 0.2 |
| [ə] | | | **40** | **40** | **0.3** | **0.3** |
| [ɚ] | **LLy** **LIy** **TTy** **TBx** **TDy** | **LIy** **TDy** **TTy** **TBx** **LLy** | **33** | **30** | **0.4** | **0.3** |
| [ʌ] | **LLy** | **LLy** | **39** | **39** | **0.5** | **0.5** |
| [ɑ] | LLy TBy TBx | **ULy** LLy TBy TTx | 34 | 34 | 0.6 | 0.3 |
| [ɒ] | **TBy** **TDx** **LLy** | **LLy** **TBy** **TDx** | **29** | **28** | **0.2** | **0.2** |
| [ɔ] | **LLy** **TBx** **TTy** **ULy** **ULx** | **TBx** **ULy** **LLy** **ULx** **TTy** | **39** | **38** | **0.4** | **0.3** |
| [ʊ] | **LLy** **ULx** | **ULx** **LLy** | **46** | **46** | **0.4** | **0.4** |
| [u] | **TDy** **ULy** **LLy** | **ULy** **LLy** **TDy** | **37** | **37** | **0.4** | **0.4** |
| [aɪ] | LLy **TDy** TBx TTy | LLy TBx TBy | 27 | 30 | 0.4 | 0.4 |
| [aɪ] | LLy TTx TTy Vy | TDx TTy LLy Vx | 30 | 34 | 0.5 | 0.5 |
| [eɪ] | LLy TDy TTx | LLy TBy TBx | 33 | 36 | 0.4 | 0.3 |
| [eɪ] | **LLy** **TTx** **TDy** | **LLy** **TTx** **TDy** | **27** | **27** | **0.3** | **0.3** |
| [ɛə] | **LLy** **TBy** | **LLy** **TBy** | **76** | **76** | **0.4** | **0.4** |
| [ɛə] | LLy TBy | **ULx** LLy TBy | 164 | 159 | 0.5 | 0.2 |
| [ɪə] | **LLy** **TTx** **TBy** **TTy** **ULx** | **LLy** **TBy** **TTx** **TTy** **ULx** | **110** | **110** | **0.2** | **0.2** |
| [ɪə] | LLy ULx TDy | ULx LLy TBy | 187 | 183 | 0.4 | 0.3 |
| [ɔɪ] | TTx LLy TTy ULx TDy ULy | LLy TDy TTy ULx TBx ULy | 61 | 54 | 0.3 | 0.1 |
| [ɔɪ] | LLy TDy ULx TTx TTy | LLy ULx **Vy** TTy TDy TDx | 66 | 74 | 0.6 | 0.3 |
| [oʊ] | LLy ULy | ULx LLy **TBy** | 45 | 41 | 0.6 | 0.5 |
| [oʊ] | **ULy** **LLy** **TTx** | **ULy** **LLy** **TTx** | **36** | **36** | **0.4** | **0.4** |
| [aʊ] | LLy TDy TTy TBx Vy | TDx TTy Vy TDy LLy | 36 | 42 | 0.3 | 0.2 |
| [aʊ] | TBy TBx TTy Vx | LIy TTx TBy Vx | 36 | 41 | 0.4 | 0.4 |

Table C.19: *1D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at 2× IPA critical threshold for male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | Identified critical coordinates ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | ULy LLy | ULy LLy | 39 | 39 | 0.5 | 0.5 |
| [b] | ULy LLy Vx | ULx ULy Vx LLy LLx | 41 | 37 | 0.6 | 0.3 |
| [m] | LLy ULy Vx ULx | ULx LLy ULy Vx | 33 | 33 | 0.4 | 0.4 |
| [t] | TTy LIy | TTy LIy | 25 | 25 | 0.5 | 0.5 |
| [d] | TTy LIy | TTy LIy | 32 | 32 | 0.5 | 0.5 |
| [n] | TTy Vx | TTy Vx | 31 | 31 | 0.5 | 0.5 |
| [k] | TDy | TDy | 25 | 25 | 0.3 | 0.3 |
| [g] | TDy | TDy | 25 | 25 | 0.3 | 0.3 |
| [ŋ] | TDy Vx | TDy Vy | 22 | 22 | 0.5 | 0.4 |
| [f] | LLy LLx | LLx TBy LLy | 34 | 31 | 0.7 | 0.5 |
| [v] | LLy LLx ULy | ULy LLx LLy | 34 | 34 | 0.5 | 0.3 |
| [θ] | TTy TTx TBy LIy LLy | TBy LLy TTy TBx | 30 | 32 | 0.5 | 0.4 |
| [ð] | TTx TTy TBy LLy LIy | LLy TBy TTx TTy | 20 | 20 | 0.2 | 0.2 |
| [s] | LIy TTy TBx TBy LLy | LIy LLy TBy TTy TTx | 14 | 15 | 0.5 | 0.5 |
| [z] | LIy TTy TTx LLy TBy | LLy TBy LIy TBx TTy | 19 | 18 | 0.5 | 0.5 |
| [ʃ] | TTy LIy TTx LLx TBy ULy | ULy LLx LIy TTx TBy TTy | 28 | 29 | 0.5 | 0.5 |
| [ʒ] | TTy LIy TBy TDx LLx TDy | TDx LLx TTy LIy TBy LLy | 138 | 110 | 1.3 | 0.8 |
| [ʧ] | TTy LIy TTx LLx LLy Vx | LLx Vx LLy LIy TTx TTy | 23 | 23 | 0.6 | 0.6 |
| [ʤ] | LIy TTy TTx LLx LLy Vx | LIy ULy ULx TBx Vx TTy | 22 | 27 | 0.6 | 0.5 |
| [l] | TDy TDx | TBx TDy | 20 | 19 | 0.3 | 0.2 |
| [ɹ] | TTx | TTx TDy | 34 | 31 | 0.7 | 0.3 |
| [w] | ULy TDy LLy TDx TTy | LLy ULy TDy TTy TTx | 25 | 24 | 0.2 | 0.2 |
| [j] | TBy TDx TTy ULx LLy | LLx LIy TTx TBy | 24 | 25 | 0.5 | 0.5 |
| [h] | TTy ULy | ULy TTy | 39 | 39 | 0.4 | 0.4 |
| [æ] | LLy TDy TDx | LLy TBy TDx Vx | 28 | 25 | 0.7 | 0.3 |
| [ɛ] | LLy TDy | LLy TDy | 38 | 38 | 0.3 | 0.3 |
| [ɪ] | | TBy | 32 | 29 | 0.7 | 0.3 |
| [iː] | TBy LLy TDx TTy | TBy TTx TDy LLy | 16 | 17 | 0.4 | 0.4 |
| [i] | TDy TTx TBy | TTx TBy TDy LIy | 21 | 19 | 0.6 | 0.3 |
| [ə] | | | 30 | 30 | 0.2 | 0.2 |
| [ɚ] | TTy LLx LLy TDy Vx | LLx TDy LLy Vx TTy | 26 | 27 | 0.2 | 0.2 |
| [ʌ] | TBy LLy TDx | LLy TTx TBy | 26 | 33 | 0.4 | 0.4 |
| [ɑ] | TBy TDx ULy TTy | ULy TTy TBx TBy | 28 | 27 | 0.5 | 0.5 |
| [ɒ] | TBy TDx TDy | LLy TBx TBy TDy | 27 | 25 | 0.7 | 0.4 |
| [ɔ] | TDx LLy TDy TBy | LLy TBy TDx TDy | 16 | 16 | 0.2 | 0.2 |
| [ʊ] | LLy TDx | LLy TDx Vx | 29 | 27 | 0.5 | 0.3 |
| [u] | TBy LLy | LLy TBy | 29 | 29 | 0.4 | 0.4 |
| [aɪ] | TTy LLy TDy TBx | LLy TTy TBx TDy | 29 | 29 | 0.4 | 0.4 |
| [aɪ] | LLy | LLy | 40 | 40 | 0.5 | 0.5 |
| [eɪ] | LLy TBy TDx TTy | LLy TBx TDy | 22 | 26 | 0.4 | 0.5 |
| [eɪ] | TDy LLy TBx | LLy TBx TBy TDy | 20 | 18 | 0.5 | 0.2 |
| [ɛə] | LIy ULy TDy TDx TBy | TBx LIy ULy TTx TBy | 56 | 60 | 0.5 | 0.4 |
| [ɛə] | TBy LIy ULy | ULy TBy LIy | 75 | 74 | 0.3 | 0.3 |
| [ɪə] | TBy LLy TBx ULy TTy | TBy TTy TBx LLy ULy | 50 | 50 | 0.3 | 0.3 |
| [ɪə] | ULy LIy TDy | ULy LIy TTx TDy | 107 | 87 | 0.6 | 0.4 |
| [ɔɪ] | TTx TDy TTy LLy Vy | LLy TTy TBx TDy Vy | 47 | 48 | 0.4 | 0.3 |
| [ɔɪ] | TDy LLy | LLy LIx TDy | 55 | 58 | 0.7 | 0.5 |
| [oʊ] | TTy LLx | LLx TTy TBy | 38 | 35 | 0.6 | 0.4 |
| [oʊ] | LLy | LLy TDy ULy | 35 | 31 | 0.7 | 0.4 |
| [aʊ] | TTy TDy TBx LLy | LLy TTy TBx TDy | 31 | 30 | 0.4 | 0.4 |
| [aʊ] | TBx TTy TDy ULx LLy Vx | ULx ULy TBx TDy Vx | 27 | 29 | 0.3 | 0.4 |

Table C.20: *1D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale* $\Upsilon^{\phi}_{eval}$ *and maximum identification divergence* $J^{\phi}_{max}$ *values at 2×IPA critical threshold for female speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | Identified critical coordinates DFS | ES | $\Upsilon_{eval}$ DFS | $\Upsilon_{eval}$ ES | $J_{max}$ DFS | $J_{max}$ ES |
|---|---|---|---|---|---|---|
| [p] | UL LL | LL UL | 34 | 33 | 0.6 | 0.6 |
| [b] | UL LL | UL LL | 28 | 28 | 0.5 | 0.5 |
| [m] | UL LL V TT | V UL TT LL | 23 | 23 | 0.2 | 0.2 |
| [t] | TT | TT | 26 | 26 | 0.6 | 0.6 |
| [d] | TT | TT | 24 | 24 | 0.3 | 0.3 |
| [n] | TT V | TT V | 23 | 23 | 0.3 | 0.3 |
| [k] | TD TB | TB TD | 23 | 23 | 0.6 | 0.6 |
| [g] | TD V TB | TD LI V | 27 | 26 | 0.6 | 0.4 |
| [ŋ] | TD V LI UL | V LL TD | 31 | 33 | 0.3 | 0.6 |
| [f] | LL UL TT | UL TT LL | 36 | 36 | 0.6 | 0.6 |
| [v] | LL UL LI | LL LI UL | 35 | 35 | 0.7 | 0.7 |
| [θ] | TT LL LI UL | TB LL UL TT | 38 | 34 | 0.6 | 0.4 |
| [ð] | TT TD LL | TT TB | 18 | 19 | 0.7 | 0.7 |
| [s] | TT LI LL | LI LL TT | 22 | 22 | 0.6 | 0.6 |
| [z] | TT LI LL | LI TT LL | 20 | 20 | 0.5 | 0.5 |
| [ʃ] | TT LI TD LL | UL TT LI TB | 14 | 14 | 0.5 | 0.4 |
| [ʒ] | LI TT TD LL V | TT LL V LI TD | 161 | 161 | 0.7 | 0.7 |
| [ʧ] | TT LI TB V LL | LI V TT LL TD | 14 | 14 | 0.6 | 0.4 |
| [ʤ] | TT TB LI V LL | TD LI TT V LL | 13 | 14 | 0.6 | 0.5 |
| [l] | TB TT LI | LI TB TT | 20 | 20 | 0.3 | 0.3 |
| [ɹ] | TT | TT | 44 | 44 | 0.7 | 0.7 |
| [w] | UL TT LL | LL TT UL | 28 | 28 | 0.4 | 0.4 |
| [j] | TB TT UL | TT TD | 26 | 25 | 0.7 | 0.7 |
| [h] | TT LL | TT LL | 31 | 31 | 0.4 | 0.4 |
| [æ] | LL TB | LL TB | 26 | 26 | 0.5 | 0.5 |
| [ɛ] | LL TB | LL TB | 26 | 26 | 0.5 | 0.5 |
| [ɪ] | TT | TT | 24 | 24 | 0.6 | 0.6 |
| [i:] | TB TT LL | LL TT TD | 18 | 17 | 0.6 | 0.5 |
| [i] | TB TT | TB TT | 22 | 22 | 0.6 | 0.6 |
| [ə] | | | 34 | 34 | 0.3 | 0.3 |
| [ɚ] | LL TB LI TT | LL TB LI TT | 23 | 23 | 0.4 | 0.4 |
| [ʌ] | LL TB | TD LL | 24 | 25 | 0.4 | 0.4 |
| [ɑ] | LL TB | LL TB | 26 | 26 | 0.7 | 0.7 |
| [ɒ] | TB LL | LL TB | 23 | 23 | 0.6 | 0.7 |
| [ɔ] | TB LL UL TT LI | LI LL TB TD | 32 | 30 | 0.3 | 0.6 |
| [ʊ] | LL UL | UL LL | 39 | 39 | 0.5 | 0.6 |
| [u] | TD UL | UL TD | 30 | 30 | 0.7 | 0.7 |
| [aɪ] | LL TD TT | TB LL | 21 | 22 | 0.5 | 0.5 |
| [aɪ] | LL TT | LL TT | 26 | 26 | 0.6 | 0.6 |
| [eɪ] | LL TT TD | TT LL TD | 27 | 27 | 0.6 | 0.6 |
| [eɪ] | LL TT TD | TB LL TT | 20 | 22 | 0.6 | 0.5 |
| [ɛə] | LL TD UL LI | LL UL LI TB | 70 | 66 | 0.4 | 0.3 |
| [ɛə] | LI UL TT TD | LI UL TD TT | 174 | 174 | 0.4 | 0.4 |
| [ɪə] | LL TT TB UL | TT LL TD | 86 | 77 | 0.7 | 0.6 |
| [ɪə] | LL TB UL V | V TB UL LL | 131 | 131 | 0.6 | 0.6 |
| [ɔɪ] | LL TT UL TD | TT TD LL UL | 52 | 52 | 0.5 | 0.5 |
| [ɔɪ] | LL TD UL TT | LL UL TT TD | 52 | 52 | 0.6 | 0.6 |
| [oʊ] | TT LL | LL TT | 32 | 32 | 0.7 | 0.7 |
| [oʊ] | UL LL TT | UL LL TT | 30 | 30 | 0.4 | 0.4 |
| [aʊ] | LL TB TD UL | TD LL TT | 31 | 31 | 0.6 | 0.6 |
| [aʊ] | TB TT V | V TD TT | 27 | 27 | 0.5 | 0.5 |

Table C.21: *2D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale $\Upsilon^{\phi}_{eval}$ and maximum identification divergence $J^{\phi}_{max}$ values at 2× IPA critical threshold for male speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

| Phone | DFS | ES | DFS | ES | DFS | ES |
|---|---|---|---|---|---|---|
| [p] | UL LL | LL LI | 33 | 33 | 0.4 | 0.4 |
| [b] | UL LL V | UL V LL | 34 | 33 | 0.3 | 0.3 |
| [m] | LL UL V | V LL UL | 30 | 30 | 0.6 | 0.6 |
| [t] | TT LI | LI TT | 15 | 15 | 0.3 | 0.3 |
| [d] | TT LI | TT LI | 21 | 21 | 0.4 | 0.4 |
| [n] | TT V | TT V | 20 | 20 | 0.3 | 0.3 |
| [k] | TD | TD | 22 | 22 | 0.3 | 0.3 |
| [g] | TD | TD | 23 | 23 | 0.4 | 0.4 |
| [ŋ] | TD V | TD V | 17 | 17 | 0.7 | 0.7 |
| [f] | LL UL TB | UL LL TB | 23 | 23 | 0.4 | 0.4 |
| [v] | LL UL | LL UL | 32 | 32 | 0.7 | 0.7 |
| [θ] | TT TB LL LI UL | LL TT TB UL LI | 23 | 23 | 0.5 | 0.5 |
| [ð] | TT TB LL | TT TB LL | 18 | 18 | 0.7 | 0.7 |
| [s] | LI TT TB LL | TB TT LL LI | 9 | 9 | 0.5 | 0.5 |
| [z] | LI TT TB LL | LI TB TT LL | 12 | 12 | 0.5 | 0.5 |
| [ʃ] | LI TT UL TB TD | TD TB UL LI TT | 18 | 18 | 0.5 | 0.5 |
| [ʒ] | TT LI TB UL V TD | LI TB TD TT UL V | 67 | 67 | 0.6 | 0.6 |
| [ɳ] | LI TT UL V LL | V LL TT LI | 16 | 18 | 0.7 | 0.7 |
| [ɖ] | LI TT LL V TD | UL LI TT | 13 | 19 | 0.8 | 0.8 |
| [l] | TD | TD | 19 | 19 | 0.6 | 0.6 |
| [ɹ] | TT | TT | 27 | 27 | 0.7 | 0.7 |
| [w] | UL TD LL TT | UL LL TT TD | 22 | 22 | 0.3 | 0.3 |
| [j] | TB TT LL | TT LL TB | 22 | 22 | 0.7 | 0.7 |
| [h] | TT UL | LL TB | 28 | 33 | 0.8 | 0.6 |
| | | | | | | |
| [æ] | LL TD LI V | LL TT TD | 22 | 23 | 0.3 | 0.6 |
| [ɛ] | LL TD | LL TB | 26 | 25 | 0.6 | 0.5 |
| [ɪ] | TT | TT | 21 | 21 | 0.5 | 0.5 |
| [i:] | TB TD LL TT | LL TD TB TT | 12 | 12 | 0.3 | 0.3 |
| [i] | TB TD TT | TT TB TD | 14 | 14 | 0.6 | 0.6 |
| [ə] | | | 29 | 29 | 0.3 | 0.3 |
| [ɚ] | TB LL TT V | TT LL TD V | 22 | 20 | 0.4 | 0.3 |
| [ʌ] | TB LL | LL TB | 22 | 22 | 0.3 | 0.3 |
| [ɑ] | TB TT UL | TB LL TD | 23 | 21 | 0.7 | 0.6 |
| [ɒ] | TB TD LL | LI TD LL | 22 | 23 | 0.6 | 0.3 |
| [ɔ] | TD LL TT | TD TT LL | 15 | 15 | 0.4 | 0.4 |
| [ʊ] | LL TD | LL TD | 24 | 24 | 0.6 | 0.6 |
| [u] | TD LL | TD LL | 21 | 21 | 0.5 | 0.5 |
| | | | | | | |
| [aɪ] | TT LL TD | LL TD TT | 29 | 29 | 0.6 | 0.6 |
| [aɪ] | LL TD | LL TD | 26 | 26 | 0.6 | 0.6 |
| [eɪ] | LL TB TT | LL TB TT | 19 | 19 | 0.6 | 0.6 |
| [eɪ] | TB LL TD | LL TD | 15 | 18 | 0.4 | 0.8 |
| [ɛə] | LI TD UL V TB | LI V UL TD TB | 47 | 47 | 0.8 | 0.8 |
| [ɛə] | TB LI UL V TD | LL TB V | 64 | 76 | 0.3 | 0.7 |
| [ɪə] | TB TT UL LL | UL LL TB TT | 46 | 46 | 0.6 | 0.6 |
| [ɪə] | UL TB LI | UL LI TB | 68 | 68 | 0.5 | 0.5 |
| [ɔɪ] | TD V LL TT LI | V LL TT LI TD | 40 | 40 | 0.6 | 0.6 |
| [ɔɪ] | TD V LL LI | TD V LI LL | 41 | 41 | 0.7 | 0.7 |
| [oʊ] | TT LL | LL TT | 28 | 28 | 0.6 | 0.6 |
| [oʊ] | LL TD TT | UL TD | 23 | 26 | 0.3 | 0.7 |
| [aʊ] | TT TD LL V | V LL TD TT | 24 | 25 | 0.6 | 0.6 |
| [aʊ] | TB UL V TD LL | V UL TD | 21 | 25 | 0.1 | 0.7 |

Table C.22: *2D critical modes identified using proposed depth-first search algorithm (DFS) and exhaustive search (ES) along with evaluation scale* $\Upsilon^{\phi}_{eval}$ *and maximum identification divergence* $J^{\phi}_{max}$ *values at 2×IPA critical threshold for female speaker. Identical (blue background), similar (yellow background) results are shown along with substituted (pink), deleted(blue) and inserted (green) critical coordinates.*

## C.4 Articulatory modelling

Table C.23: *Critical modes identified using proposed algorithm from PC1, LD1, PC3 and LD3 features for male and female speakers.*

| Phones | PC1 | | LD1 | | PC3 | | LD3 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [p] | 5 | 3 | 2 | 2 | 2 1 | 1 2 3 | 2 1 | 2 5 |
| [b] | 4 3 | 5 | 2 | 2 5 | 1 2 | 1 2 3 4 | 1 2 | 2 5 1 |
| [m] | 4 3 | 4 3 | 2 6 5 | 2 6 5 | 1 2 | 1 2 | 1 2 13 | 2 1 5 |
| [t] | 2 | | 1 | 1 | | 9 | 7 | 7 |
| [d] | 2 | | 1 | 1 | | 9 | 7 | 7 |
| [n] | 2 | | 1 6 | 1 | 13 | 9 | 7 13 | 7 |
| [k] | 3 5 | 3 | 4 3 | 3 | 9 | 9 8 | 10 8 9 | 8 12 |
| [g] | 3 5 | 3 | 3 4 | 3 | 9 11 | 8 9 | 8 10 | 8 12 |
| [ŋ] | 3 | 3 2 | 4 3 | 3 5 | 9 | 8 9 | 8 13 | 8 12 |
| [f] | 5 | 3 4 6 10 | 7 2 1 | 2 6 7 | 1 3 2 | 1 2 | 1 3 | 1 2 5 |
| [v] | 7 5 | 4 10 6 3 | 7 2 | 2 6 7 | 1 3 2 | 1 2 3 | 1 3 | 1 2 5 6 |
| [θ] | 1 3 2 | 1 3 | 1 6 3 | 1 2 8 6 3 | 7 1 5 | 7 11 5 | 7 9 | 7 8 11 |
| [ð] | 1 4 3 2 | 1 2 | 1 3 6 | 1 3 | 7 9 8 | 7 | 7 9 8 | 7 11 8 |
| [s] | 1 2 | 1 3 7 2 | 1 | 1 | 1 7 | 7 1 2 | 7 1 | 7 1 11 2 |
| [z] | 1 2 4 | 1 7 3 2 | 1 | 1 | 1 7 | 7 1 | 7 1 | 7 1 11 |
| [ʃ] | 2 1 4 3 | 2 6 4 1 | 1 4 3 2 | 1 4 2 | 8 7 9 | 8 9 2 7 | 7 9 8 | 7 11 2 |
| [ʒ] | 2 3 1 4 10 | 2 6 4 1 | 3 1 2 4 | 1 4 | 8 9 1 7 4 | 8 9 | 7 8 1 2 9 | 7 11 8 |

Continued on next page...

Table C.23 continued from previous page

| Phones | PC1 | | LD1 | | PC3 | | LD3 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [ʧ] | 2 1 3 4 | 2 4 6 1 | 1 3 4 2 | 1 4 2 | 8 7 1 9 | 8 9 7 2 | 7 9 8 1 | 7 11 2 |
| [ʤ] | 2 1 3 4 | 2 4 6 1 | 1 3 4 2 | 1 4 2 | 8 9 7 1 | 2 7 9 8 | 7 8 9 1 | 7 11 2 |
| [l] | 4 | | 3 7 | | 9 | | 8 | 12 |
| [ɹ] | 1 | 1 | 4 | 4 | 7 | 7 | 9 | 11 |
| [w] | 4 | 4 3 1 | 2 3 | 3 2 | 2 9 1 | 9 1 | 7 1 | 2 8 |
| [j] | 2 3 | 3 1 | 3 1 | 2 1 3 | 9 8 | 8 | 8 7 | 8 7 |
| [h] | | 2 | | | | | | 11 |
| [æ] | 2 | 2 3 1 4 | 1 | 1 3 2 | 1 | 1 | 1 | 1 8 |
| [ɛ] | 2 4 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| [ɪ] | 2 | | | | | | | |
| [iː] | 3 2 | 3 1 | 3 2 1 | 2 3 1 | 9 8 1 | 8 9 7 | 8 7 1 | 8 7 |
| [i] | 3 2 | 3 1 | 2 3 | 2 3 1 4 | 9 8 | 8 7 9 | 8 7 | 8 7 |
| [ə] | | | | | | | | |
| [ɚ] | 2 | 2 | 6 1 | 1 7 | 1 | 8 9 | | 7 |
| [ʌ] | | 1 2 | | 1 | | | | 7 |
| [ɑ] | 2 4 | 1 2 | 1 2 | 1 | 1 7 8 | 8 7 | 1 7 | 7 8 |
| [ɒ] | 2 | 1 | | 2 1 | 8 | 7 | | 7 |
| [ɔ] | 2 | 1 4 3 | 2 1 | 2 1 3 | 1 8 7 | 7 1 9 | 7 1 | 7 1 |
| [ʊ] | | | | | 1 | | 1 | |
| [u] | 3 | 3 | 3 | 3 2 | 9 | | 8 | 8 |

Continued on next page...

Table C.23 continued from previous page

| Phones | PC1 | | LD1 | | PC3 | | LD3 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [aɪ] | 2 5 | 2 1 3 | 1 | 1 | 1 7 | 1 7 | 1 7 | 7 |
| [aɪ] | 5 3 | 3 | 1 2 | 2 1 | 1 | 1 | 1 | 1 |
| [eɪ] | 3 4 2 | 3 1 7 | 1 2 | 2 1 | 1 11 9 | 1 7 | 1 8 | 8 1 |
| [eɪ] | 3 4 | 3 1 | 2 1 3 | 2 3 | 1 9 | 9 7 | 8 1 | 8 |
| [ɛə] | 2 4 | 2 3 4 | 1 2 | 1 2 3 | 1 | 1 | 1 | 1 |
| [ɛə] | 2 4 | | 1 3 | 1 | 1 | | 1 7 | 1 |
| [ɪə] | 3 4 2 | 3 1 | 2 1 | 2 1 | 1 | 7 8 11 9 | 1 7 | 8 7 |
| [ɪə] | 2 | | 2 1 | 1 | 1 | | 1 | 2 |
| [ɔɪ] | 4 2 3 | 1 3 4 | 1 3 2 | 3 1 4 2 | 1 9 7 | 7 9 14 | 7 1 8 | 7 14 8 1 11 |
| [ɔɪ] | 3 4 12 | 3 | 3 1 | 3 2 | 9 1 | 9 14 | 8 | 8 14 |
| [oʊ] | 2 | | 1 | | 2 | | 7 | |
| [oʊ] | 4 | 4 | 2 | | 2 1 | | 1 | |
| [aʊ] | 2 3 | 2 1 3 | 1 | 1 | 1 8 | 8 9 7 | 1 7 | 7 |
| [aʊ] | 2 | 1 | 1 | 1 2 | 8 7 | 7 | 7 | 7 |

Table C.24: *Critical modes identified using proposed algorithm from PC4, LD4, PC5 and LD5 features for male and female speakers.*

| Phones | PC4 | | LD4 | | PC5 | | LD5 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [p] | 1 3 | 1 | 2 3 | 2 | 1 3 | 1 | 2 3 | 2 |
| [b] | 1 3 | 1 3 | 2 3 | 2 3 | 1 3 | 1 3 | 2 3 | 2 3 |
| [m] | 1 3 13 | 3 1 | 2 3 13 | 2 3 | 1 3 13 | 3 1 | 2 3 13 | 2 3 |
| [t] | | 9 | 7 | 7 | | 11 | 10 | 11 |
| [d] | | 9 | 7 | 7 | | 11 | 10 | 11 |
| [n] | 13 | 9 | 7 13 | 7 13 | 13 | 11 | 10 13 | 11 13 |
| [k] | 9 | 9 8 | 10 8 9 | 8 12 | 11 12 | 10 11 | 12 11 | 12 10 |
| [g] | 9 11 | 8 9 | 8 10 9 | 8 12 | 11 12 | 10 11 | 11 12 | 12 10 |
| [ŋ] | 9 | 8 9 | 8 13 14 | 8 12 | 11 | 10 11 | 11 12 13 | 12 |
| [f] | 3 4 1 | 3 4 | 3 2 6 | 3 4 | 3 4 1 | 3 4 | 3 2 6 | 3 4 |
| [v] | 3 4 1 | 4 3 | 3 | 4 3 | 3 4 1 | 4 3 | 3 6 2 | 4 3 |
| [θ] | 7 3 | 7 11 9 5 | 7 6 | 7 8 11 | 7 3 5 | 7 11 12 | 7 10 3 | 11 7 12 4 |
| [ð] | 7 9 8 | 7 8 9 | 7 9 8 | 7 11 8 | 7 11 10 | 7 11 | 7 10 | 7 11 |
| [s] | 3 7 | 7 3 4 | 7 3 | 7 3 11 4 8 | 3 7 | 7 3 4 | 10 3 7 | 3 7 4 11 |
| [z] | 3 7 5 | 7 3 4 | 7 3 | 7 3 11 4 | 3 7 5 | 7 3 4 | 3 7 10 | 3 7 4 11 |
| [ʃ] | 8 7 9 | 8 9 7 4 | 7 9 8 | 7 11 | 10 7 11 | 11 10 7 4 | 10 7 11 | 11 7 9 |
| [ʒ] | 8 9 3 5 7 | 8 9 | 7 8 3 9 | 7 11 8 | 10 3 11 5 7 | 10 11 7 | 10 11 3 7 | 11 7 12 |

Continued on next page...

Table C.24 continued from previous page

| Phones | PC4 | | LD4 | | PC5 | | LD5 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [ʧ] | 8 7 3 9 | 8 9 7 4 | 7 9 8 3 | 7 11 3 | 10 7 3 | 10 11 7 4 | 10 7 3 | 11 7 2 |
| [ʤ] | 8 9 7 3 | 4 9 8 7 | 7 8 9 3 | 7 11 3 | 10 7 3 11 | 4 11 7 10 | 10 11 7 3 | 11 9 3 |
| [l] | 9 | | 8 | 12 | | | 11 | |
| [ɹ] | 7 | 7 | 9 | 11 | 7 | 7 | 7 | 7 |
| [w] | 1 9 | 1 9 3 | 2 7 | 2 8 | 1 11 | 1 11 3 | 2 | 2 12 7 |
| [j] | 9 8 | 8 9 | 8 7 | 8 7 | 10 11 | 10 | 11 | 12 11 |
| [h] | | | | 11 | | | | 11 |
| [æ] | 3 | 3 | 3 | 3 8 | 3 | 3 10 | 3 | 3 12 |
| [ɛ] | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| [ɪ] | | | 7 | | | | | |
| [iː] | 9 8 3 | 8 9 7 | 8 7 3 | 8 7 | 11 10 3 | 10 11 | 11 3 | 12 11 |
| [i] | 9 8 | 8 7 9 | 8 7 | 8 7 | 11 10 | 10 11 7 | 11 | 12 7 11 |
| [ə] | | | | | | | | |
| [ɚ] | 3 | 8 9 4 | | 7 | 3 | 10 11 4 | 3 | 11 |
| [ʌ] | | | | 7 | | 10 | | 11 |
| [ɑ] | 3 7 8 | 8 7 | 3 7 | 7 8 | 3 10 | 10 7 | 3 11 | 11 7 12 |
| [ɒ] | 8 | 7 | | 7 | 10 | 10 7 | | 4 |
| [ɔ] | 3 8 7 | 7 9 3 | 7 3 | 7 3 | 3 10 7 | 7 11 3 10 | 3 10 7 | 7 11 |
| [ʊ] | 3 | | | | 3 | | 3 | |
| [u] | 9 | | 8 | 8 | 11 1 | | 11 | 12 3 |

Continued on next page...

Table C.24 continued from previous page

| Phones | PC4 | | LD4 | | PC5 | | LD5 | |
|---|---|---|---|---|---|---|---|---|
| | (m) | (f) | (m) | (f) | (m) | (f) | (m) | (f) |
| [aɪ] | 3 7 | 3 7 | 3 7 | 7 | 3 | 3 10 | 3 | 3 12 |
| [aɪ] | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| [eɪ] | 3 11 9 | 3 7 11 | 3 8 | 8 3 | 3 | 3 | 3 11 | 12 3 |
| [eɪ] | 3 9 | 9 7 | 8 3 | 8 | 3 | 11 10 7 | 11 3 7 | 12 7 |
| [ɛə] | 3 2 | 3 | 3 | 3 8 | 3 2 | 3 | 3 | 3 12 |
| [ɛə] | 3 | 3 | 3 7 | 3 | 3 | 10 | 3 | 3 |
| [ɪə] | 3 | 7 8 11 9 1 | 3 7 | 8 7 | 3 7 | 10 12 7 11 1 | 3 | 12 7 |
| [ɪə] | 3 | 1 | 3 | 2 | 3 | 1 | 3 | 2 |
| [ɔɪ] | 3 9 7 | 7 9 14 | 7 3 8 | 7 14 8 3 11 | 3 7 11 | 7 11 14 | 3 7 10 1 11 | 7 12 14 |
| [ɔɪ] | 9 3 | 9 14 | 8 | 8 14 3 | 11 12 | 11 14 | 11 | 12 14 |
| [oʊ] | | | 7 | | | | | |
| [oʊ] | 1 3 | | 2 3 | | 1 3 | | 2 3 | |
| [aʊ] | 3 8 | 8 9 7 3 | 3 7 | 7 | 3 10 | 10 11 | 3 10 | 11 3 |
| [aʊ] | 8 7 | 7 | 7 | 7 | 10 7 | 7 | 10 | 7 11 |

Table C.25: *Critical modes identified using proposed algorithm from PC7 and LD7 features for male and female speakers.*

| Phones | PC7 | | LD7 | |
|---|---|---|---|---|
| | (m) | (f) | (m) | (f) |
| [p] | 1 3 | 1 | 2 4 | 2 |
| [b] | 1 3 | 1 3 | 2 4 | 2 4 |
| [m] | 1 3 13 | 3 1 | 2 4 13 | 4 2 |
| [t] | 7 | 5 | 8 | 8 |
| [d] | 7 | 5 | 8 | 8 |
| [n] | 7 13 | 7 | 8 13 | 8 |
| [k] | 11 | 11 | 12 | 12 |
| [g] | 11 | 11 | 12 | 12 11 |
| [ŋ] | 11 | 11 | 12 13 | 12 11 |
| [f] | 3 1 | 3 4 | 4 2 8 | 4 3 |
| [v] | 3 4 1 | 3 4 | 4 2 | 4 3 |
| [θ] | 8 3 7 | 8 9 7 | 8 7 4 | 8 7 10 4 6 |
| [ð] | 8 7 | 8 7 | 8 7 | 8 7 |
| [s] | 5 8 7 | 5 7 8 | 8 6 7 | 6 8 7 |
| [z] | 5 8 7 3 | 5 8 7 | 8 6 7 | 6 8 7 |
| [ʃ] | 7 5 8 | 5 7 8 | 8 9 6 | 6 8 7 |
| [ʒ] | 5 7 9 8 3 | 5 10 4 8 13 | 6 8 7 10 | 8 9 6 3 |

Continued on next page...

Table C.25 continued from previous page

| Phones | PC7 | | LD7 | |
|---|---|---|---|---|
| | (m) | (f) | (m) | (f) |
| [ʧ] | 5 7 10 9 | 5 7 8 4 | 6 9 8 | 6 8 7 |
| [ʤ] | 7 5 8 9 | 5 7 8 | 8 9 6 10 | 6 8 7 |
| [l] | | | | |
| [ɹ] | 10 | 12 | 7 | 7 |
| [w] | 1 | 1 11 3 | 2 8 4 | 2 4 11 |
| [j] | 9 | 9 | 10 | 10 8 |
| [h] | | 7 | 7 | |
| [æ] | 3 9 | 3 9 | 4 10 | 4 12 |
| [ɛ] | 3 | 3 | 4 | 4 |
| [ɪ] | | | | |
| [iː] | 9 3 8 | 9 11 | 10 4 | 10 8 |
| [i] | 9 | 9 11 | 10 8 | 10 11 |
| [ə] | | | | |
| [ɚ] | 3 | 9 3 | 4 | 10 4 |
| [ʌ] | | 9 | | 10 |
| [ɑ] | 3 11 | 9 | 4 12 | 10 |
| [ɒ] | 9 | 9 | 10 | 10 |
| [ɔ] | 9 3 | 8 3 11 | 10 4 | 11 4 10 |
| [ʊ] | 3 | 3 | | 4 |
| [u] | 9 1 | 11 | 12 | 10 |

Continued on next page...

Table C.25 continued from previous page

| Phones | PC7 | | LD7 | |
|---|---|---|---|---|
| | (m) | (f) | (m) | (f) |
| [aɪ] | 3 11 | 3 9 | 4 12 | 8 4 |
| [aɪ] | 3 | 3 | 4 | 4 |
| [eɪ] | 3 | 3 9 | 4 | 4 10 |
| [eɪ] | 3 8 9 | 9 3 | 4 10 | 10 4 11 |
| [ɛə] | 3 2 | 5 1 11 9 | 4 | 6 |
| [ɛə] | 3 9 | 5 | 8 4 | 6 |
| [ɪə] | 3 8 | 9 8 3 | 4 8 | 10 8 |
| [ɪə] | 9 3 | 1 9 | 4 10 | 10 |
| [ɔɪ] | 3 8 7 2 | 11 8 14 3 | 8 4 1 | 11 14 4 8 |
| [ɔɪ] | 3 | 11 14 | 4 | 14 4 |
| [oʊ] | 3 | | 8 | |
| [oʊ] | 1 3 | | 2 4 | |
| [aʊ] | 3 9 | 9 3 | 4 8 | 8 12 |
| [aʊ] | 9 | 9 11 | 10 | 10 11 |

Table C.26: *Significance test results from combined recognition accuracy values of male and female speakers. Statistically significant difference from the t-test is denoted by 't' and from the f-test is denoted by 'f'. The empty cells denote that there is no statistically significant difference between the two recognition results, whereas × denotes the pairs for which any of the tests could not be applied.*

| | LD1 | PC1 | LD3 | PC3 | LD4 | PC4 | LD5 | PC5 | LD7 | PC7 | EMA | MFCC | MFCC-PC | MFCC-LD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LD1 | x | | | f | | f | | f | f | f | f | f | f | f |
| PC1 | | x | | | | | | | f | f | f | f | f | f |
| LD3 | | | x | | | | | f | f | f | f | f | f | f |
| PC3 | t | | | x | | | | | | | f | f | f | f |
| LD4 | | | | | x | | | | | | f | f | f | f |
| PC4 | t | | | | | x | | | | | | f | f | f |
| LD5 | | | | | | | x | | | | f | f | f | f |
| PC5 | t | | t | | | | | x | | | | f | f | f |
| LD7 | t | t | t | | | | | | x | | | f | f | f |
| PC7 | t | t | t | | | | | | | x | | f | f | f |
| EMA | t | t | t | t | t | | t | | | | x | f | f | f |
| MFCC | t | t | t | t | t | t | t | t | t | t | t | x | | |
| MFCC-PC | t | t | t | t | t | t | t | t | t | t | t | | x | f |
| MFCC-LD | t | t | t | t | t | t | t | t | t | t | t | t | | x |

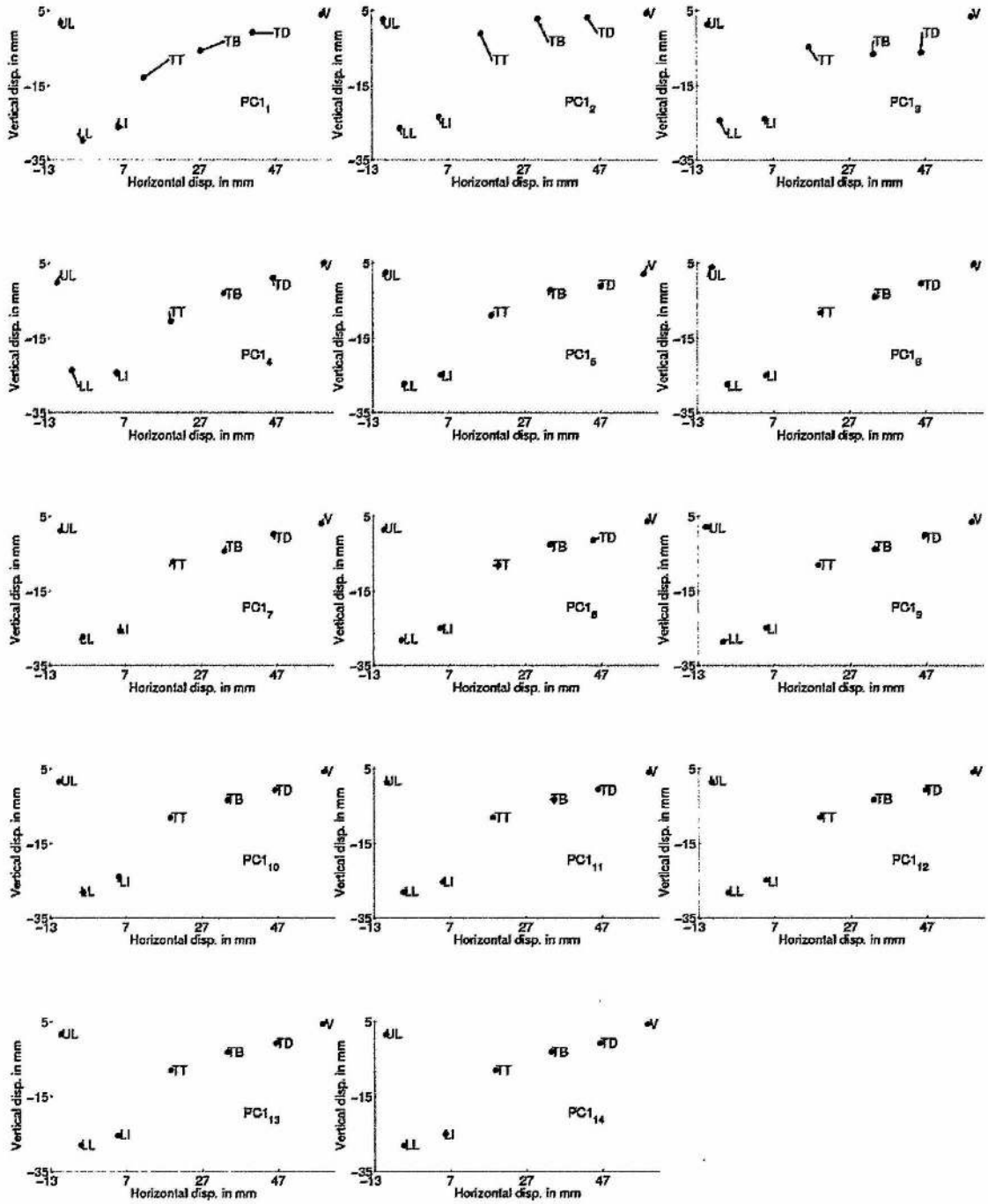



Table C.27: *PC1 mode shapes depcting the articulatory coordinate movements for male speaker.*

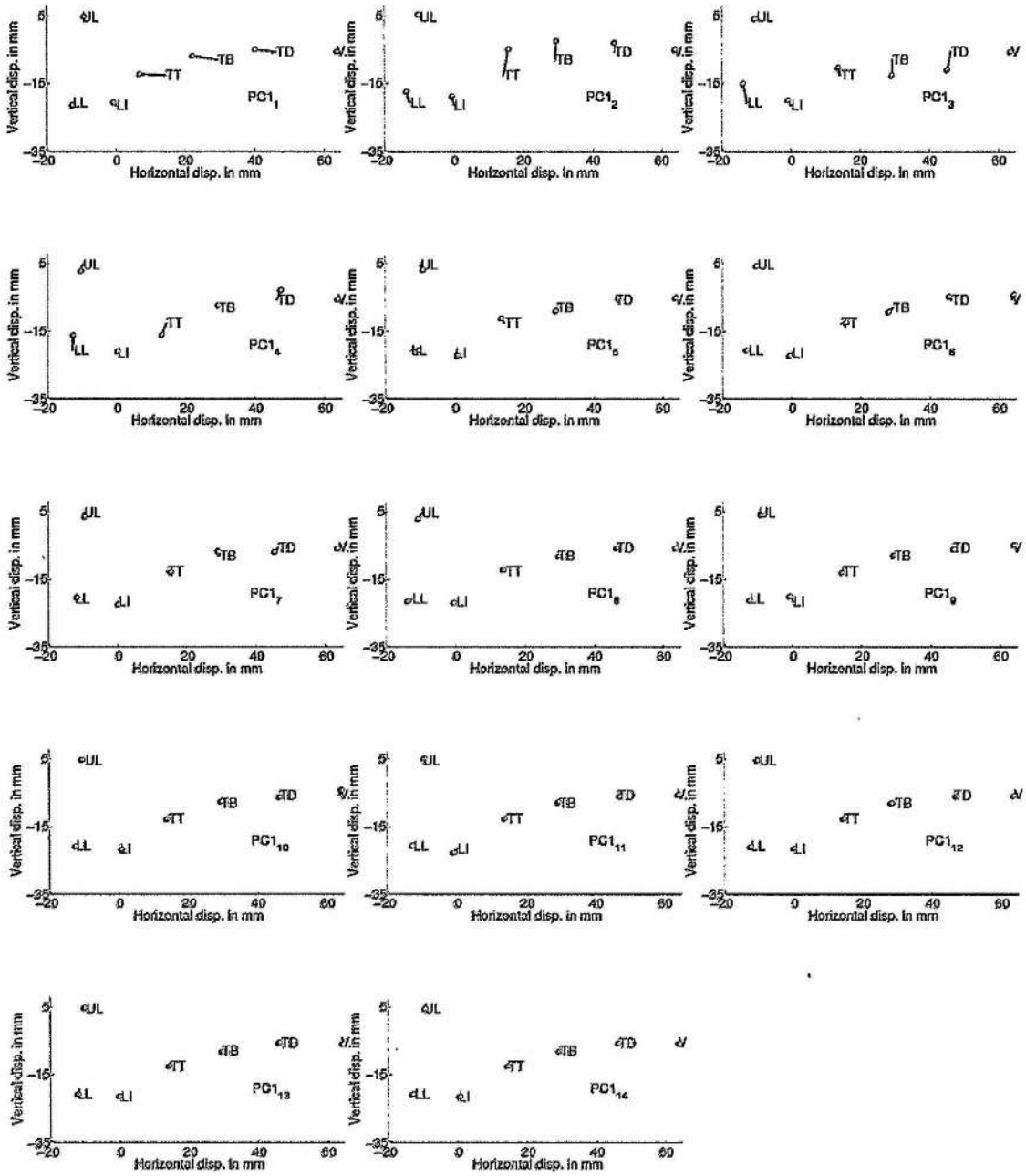



Table C.28: *PC1 mode shapes depcting the articulatory coordinate movements for female speaker.*

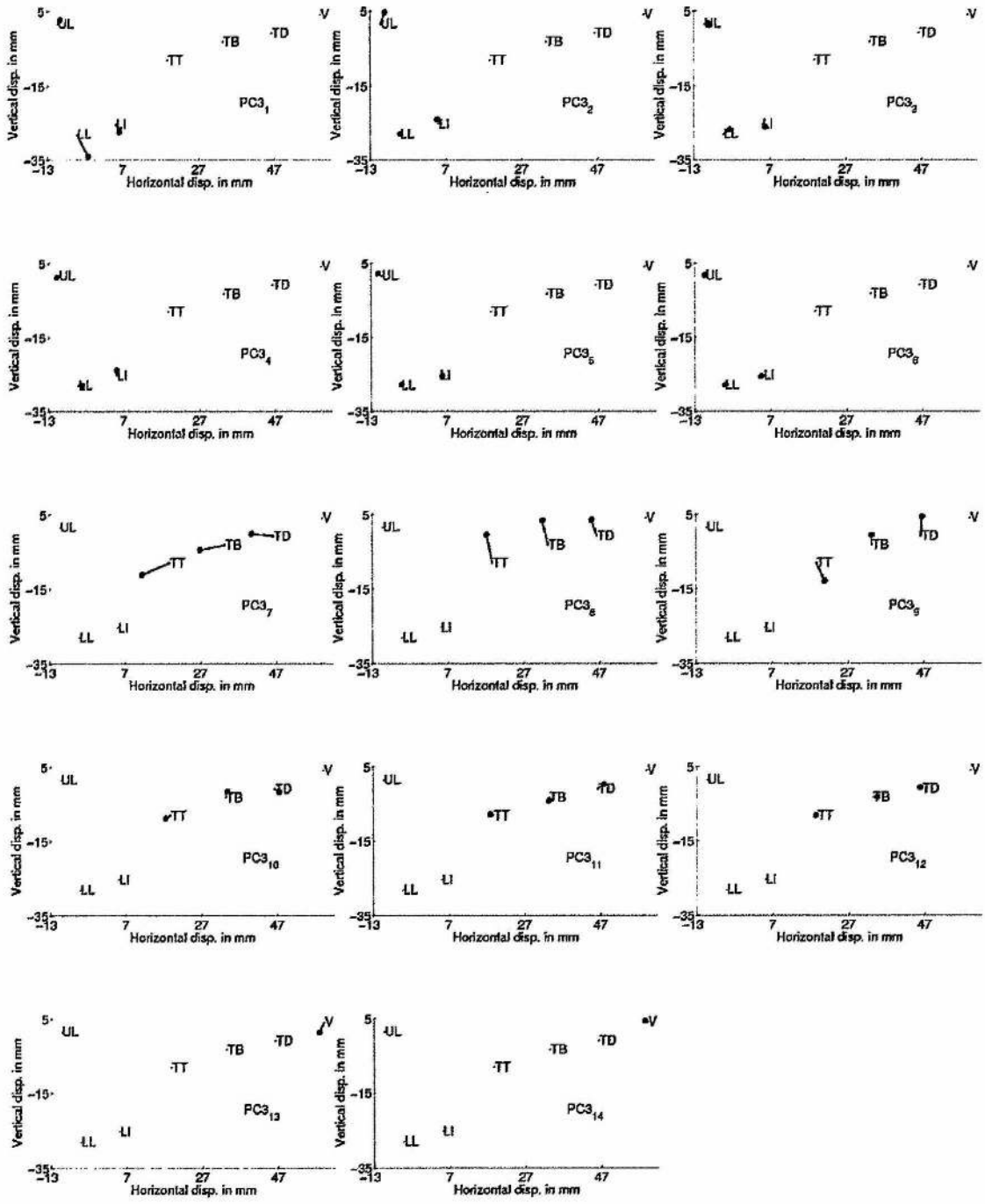



Table C.29: *PC3 mode shapes depcting the articulatory coordinate movements for male speaker.*

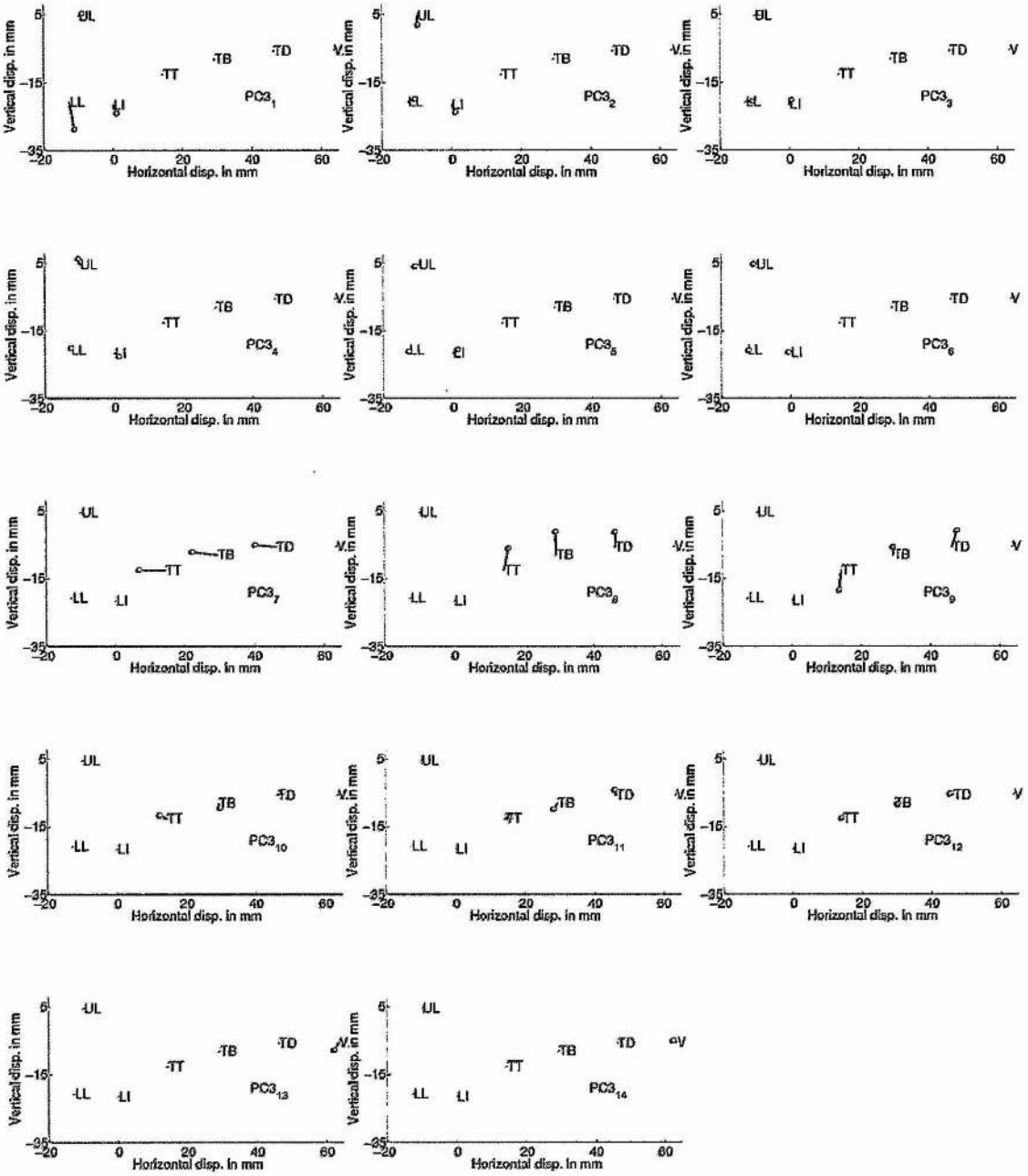



Table C.30: *PC3 mode shapes depcting the articulatory coordinate movements for female speaker.*

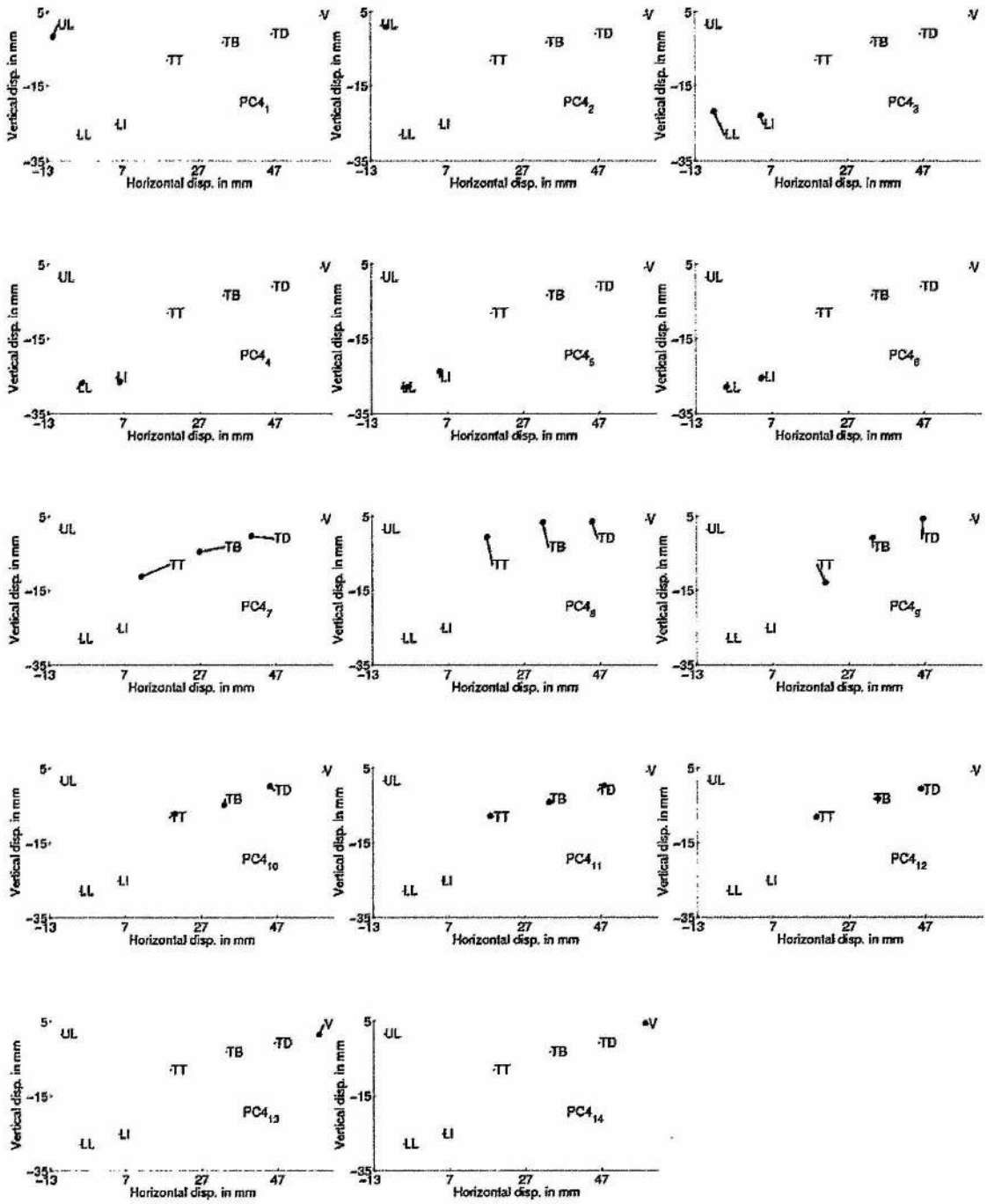



Table C.31: *PC4 mode shapes depcting the articulatory coordinate movements for male speaker.*

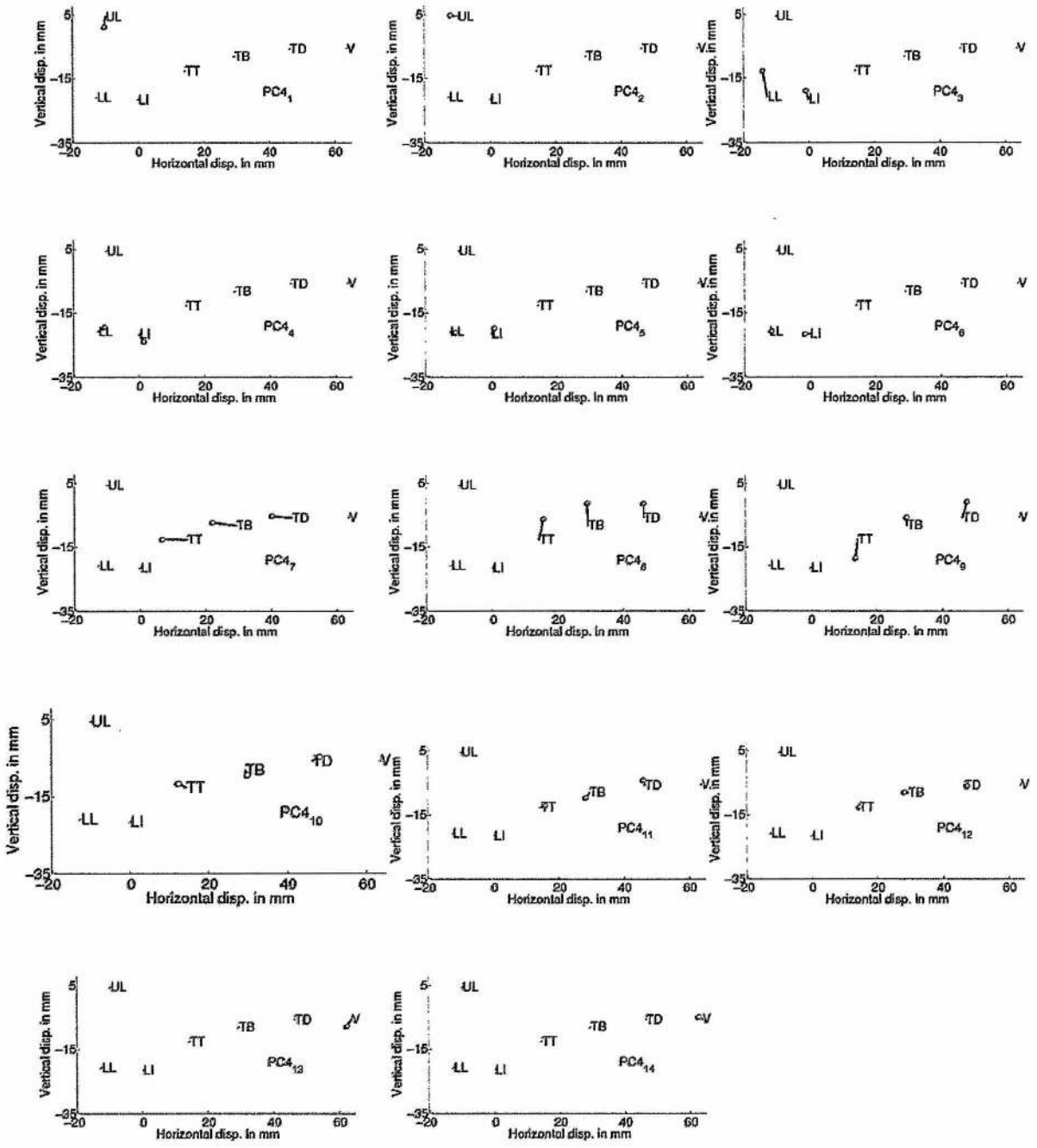



Table C.32: *PC4 mode shapes depcting the articulatory coordinate movements for female speaker.*

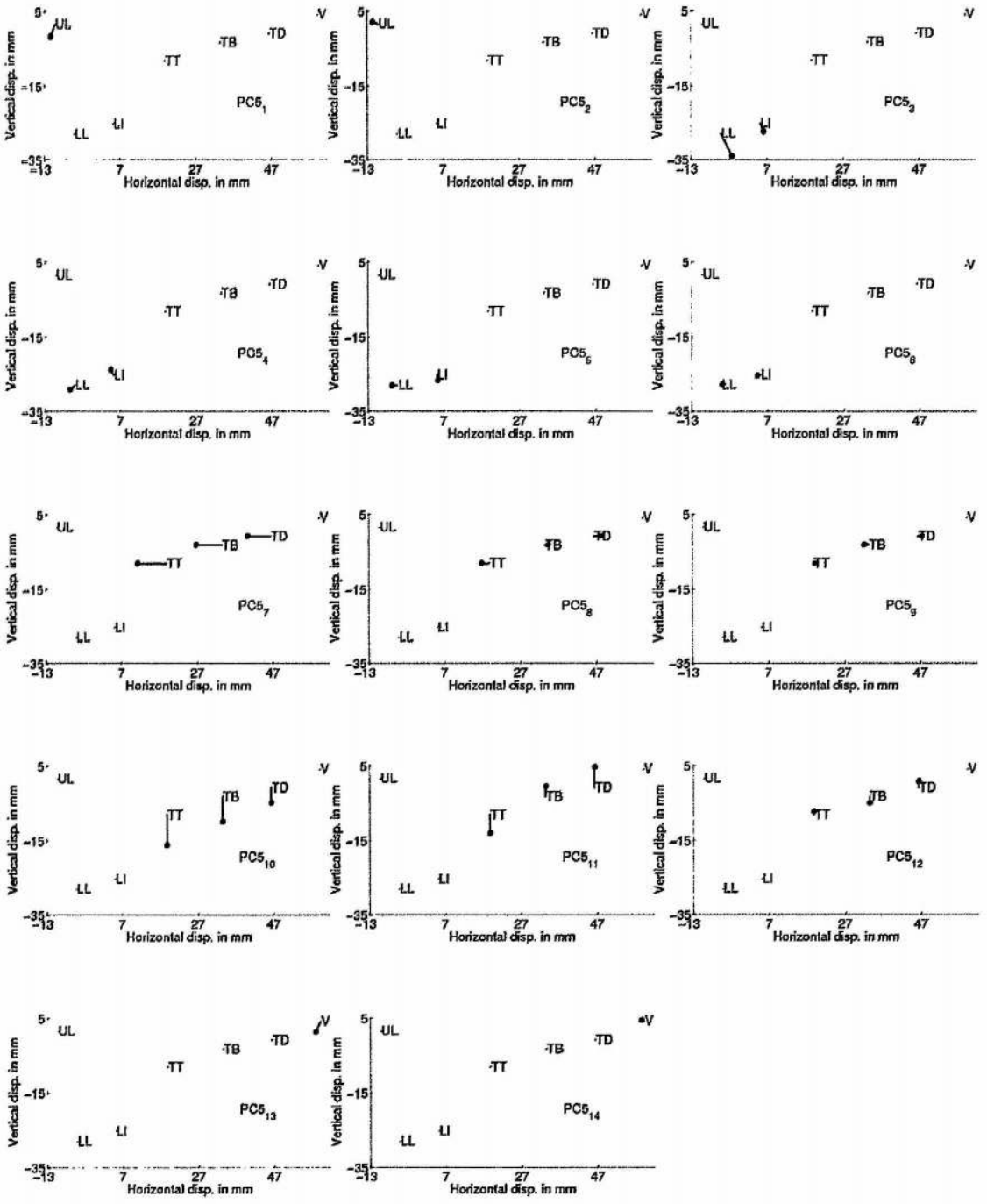



Table C.33: *PC5 mode shapes depcting the articulatory coordinate movements for male speaker.*

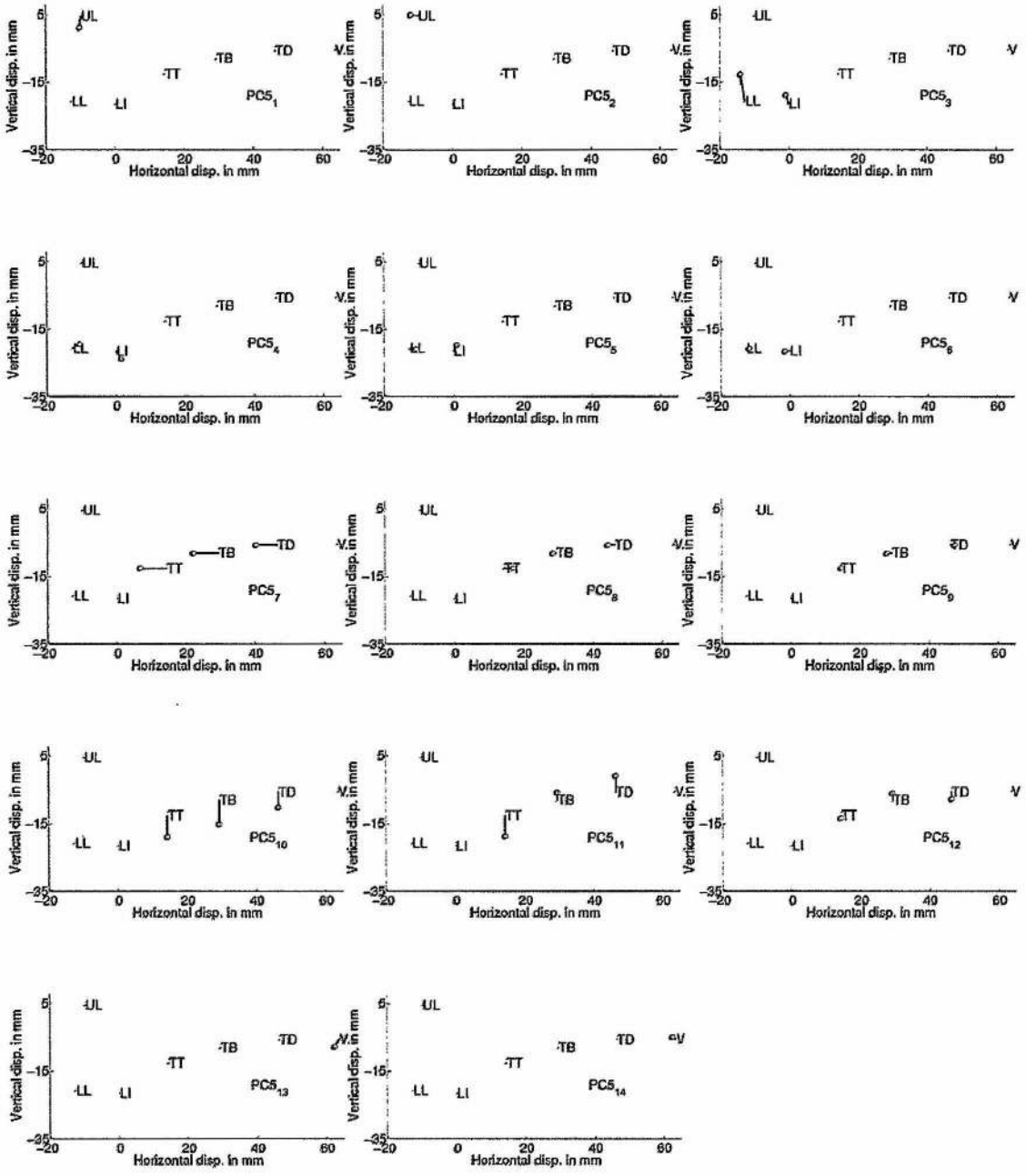



Table C.34: *PC5 mode shapes depcting the articulatory coordinate movements for female speaker.*

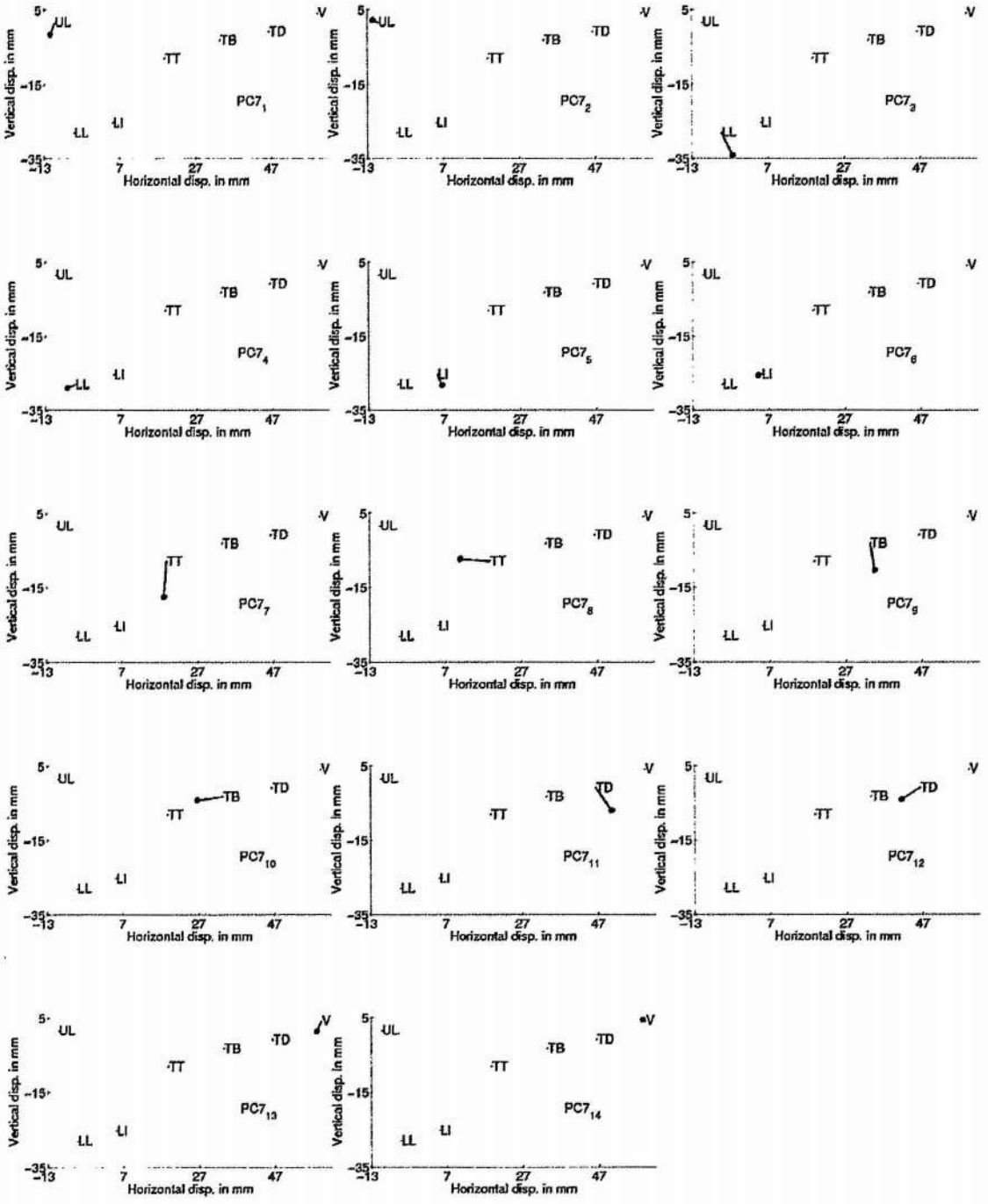



Table C.35: *PC7 mode shapes depcting the articulatory coordinate movements for male speaker.*

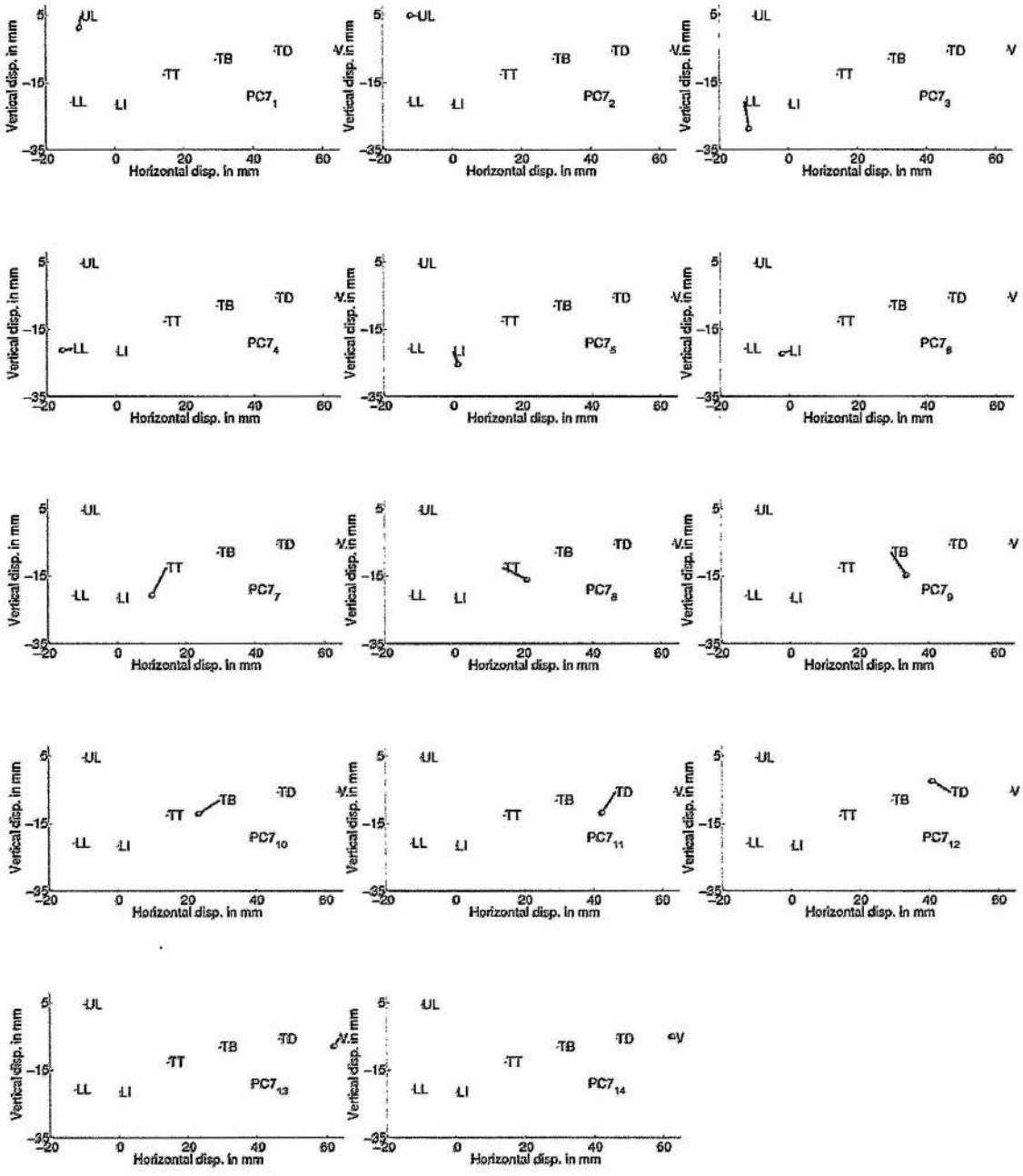



Table C.36: *PC7 mode shapes depcting the articulatory coordinate movements for female speaker.*

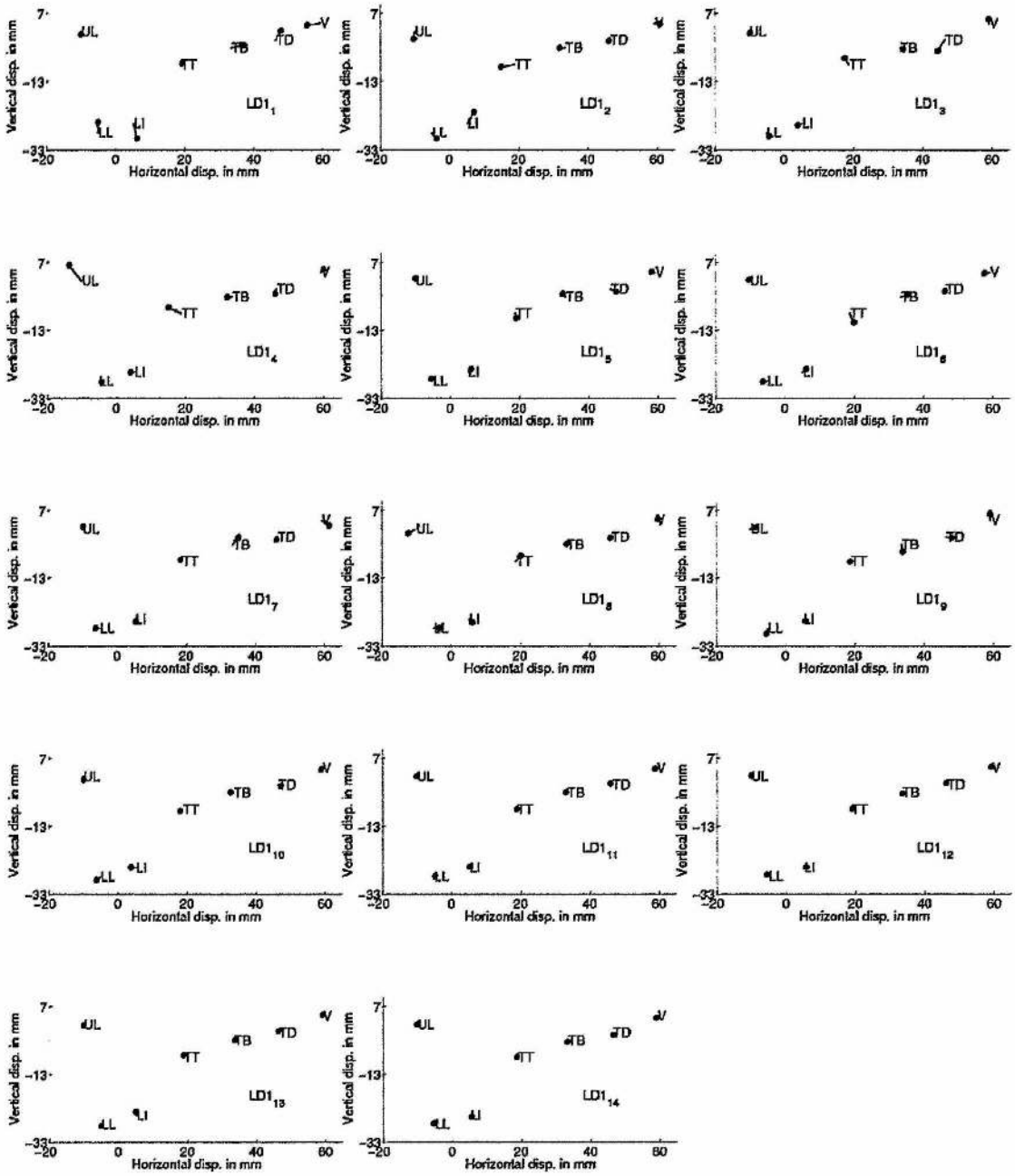



Table C.37: *LD1 mode shapes depcting the articulatory coordinate movements for male speaker.*

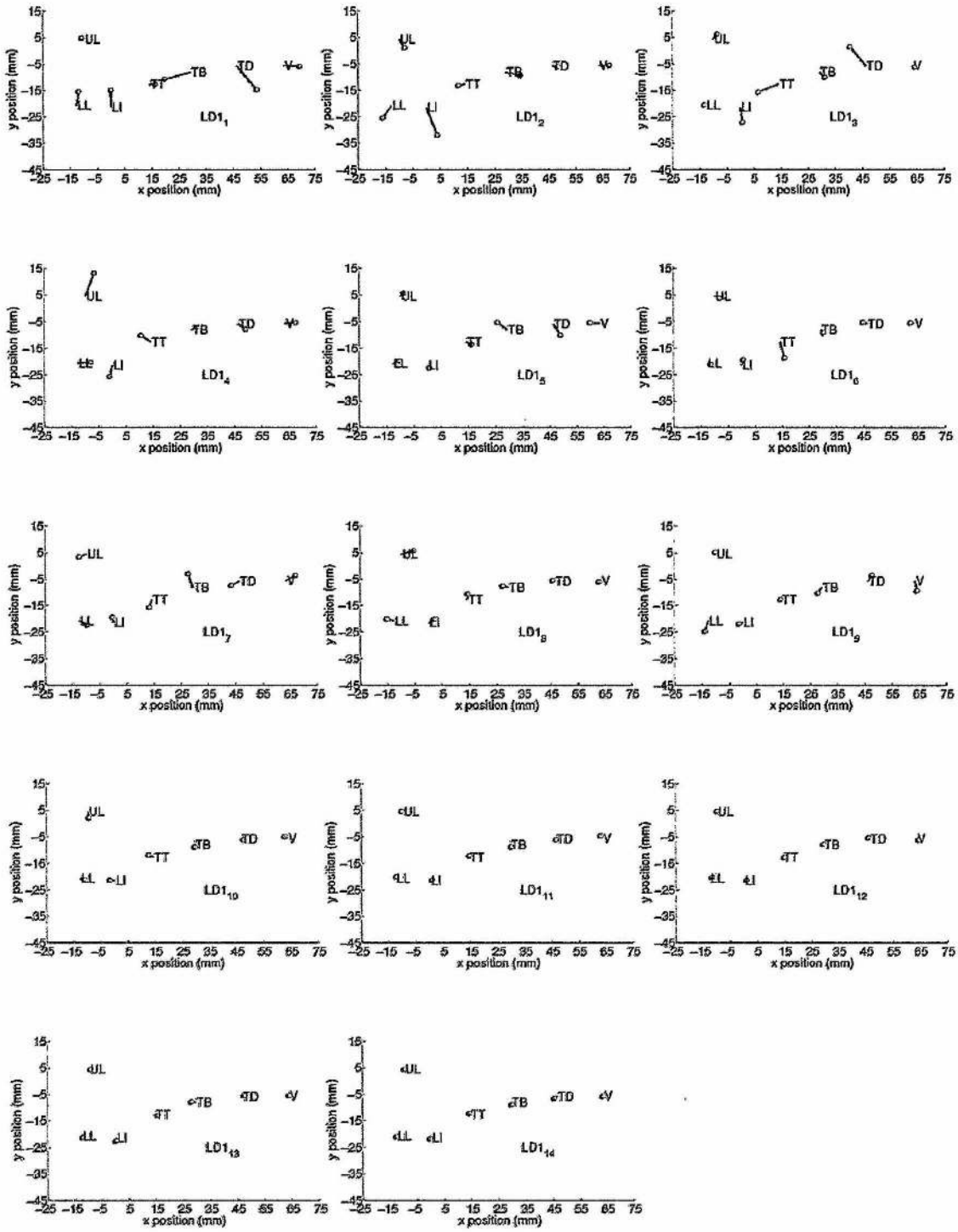



Table C.38: *LD1 mode shapes depcting the articulatory coordinate movements for female speaker.*

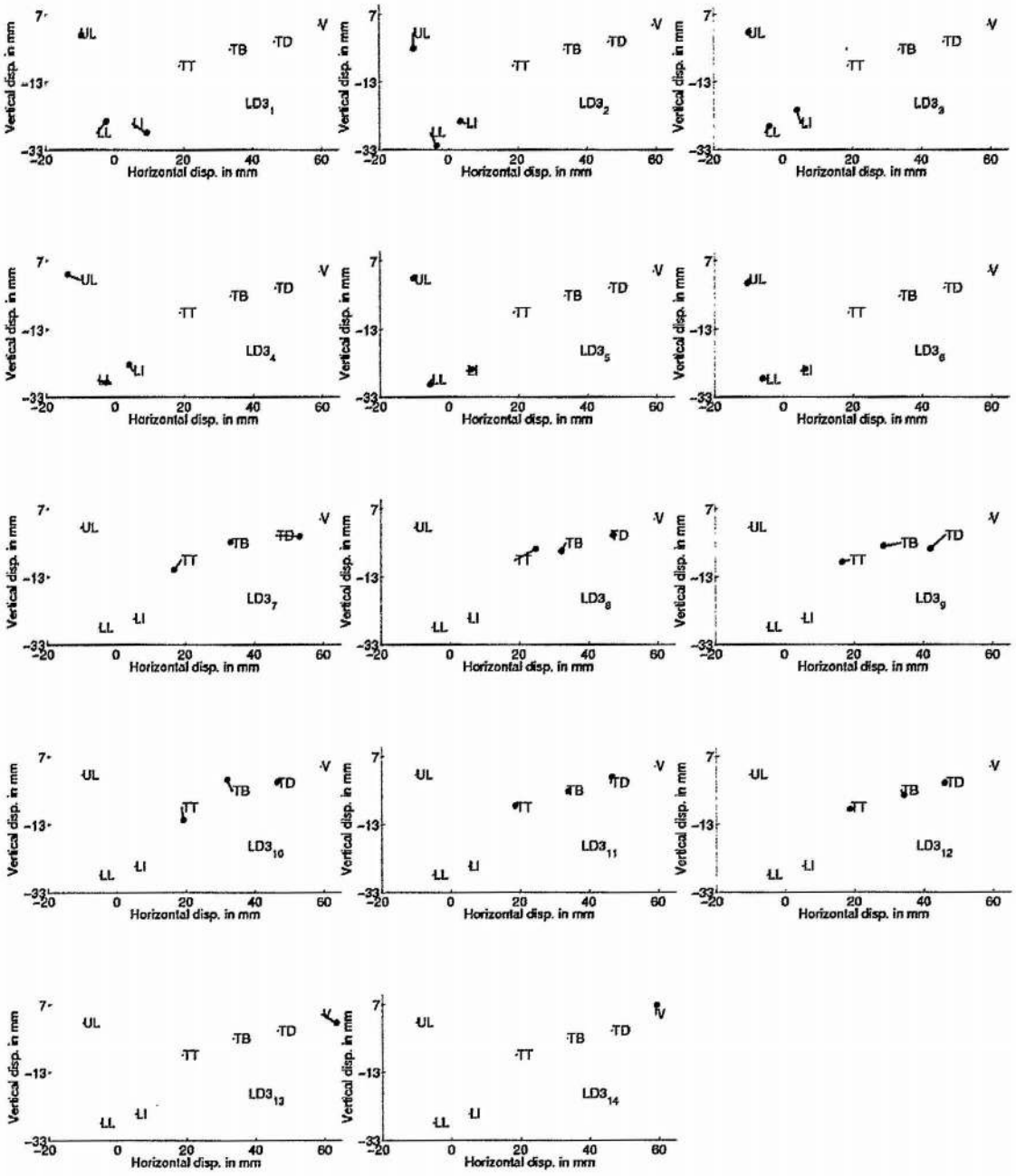



Table C.39: *LD3 mode shapes depcting the articulatory coordinate movements for male speaker.*

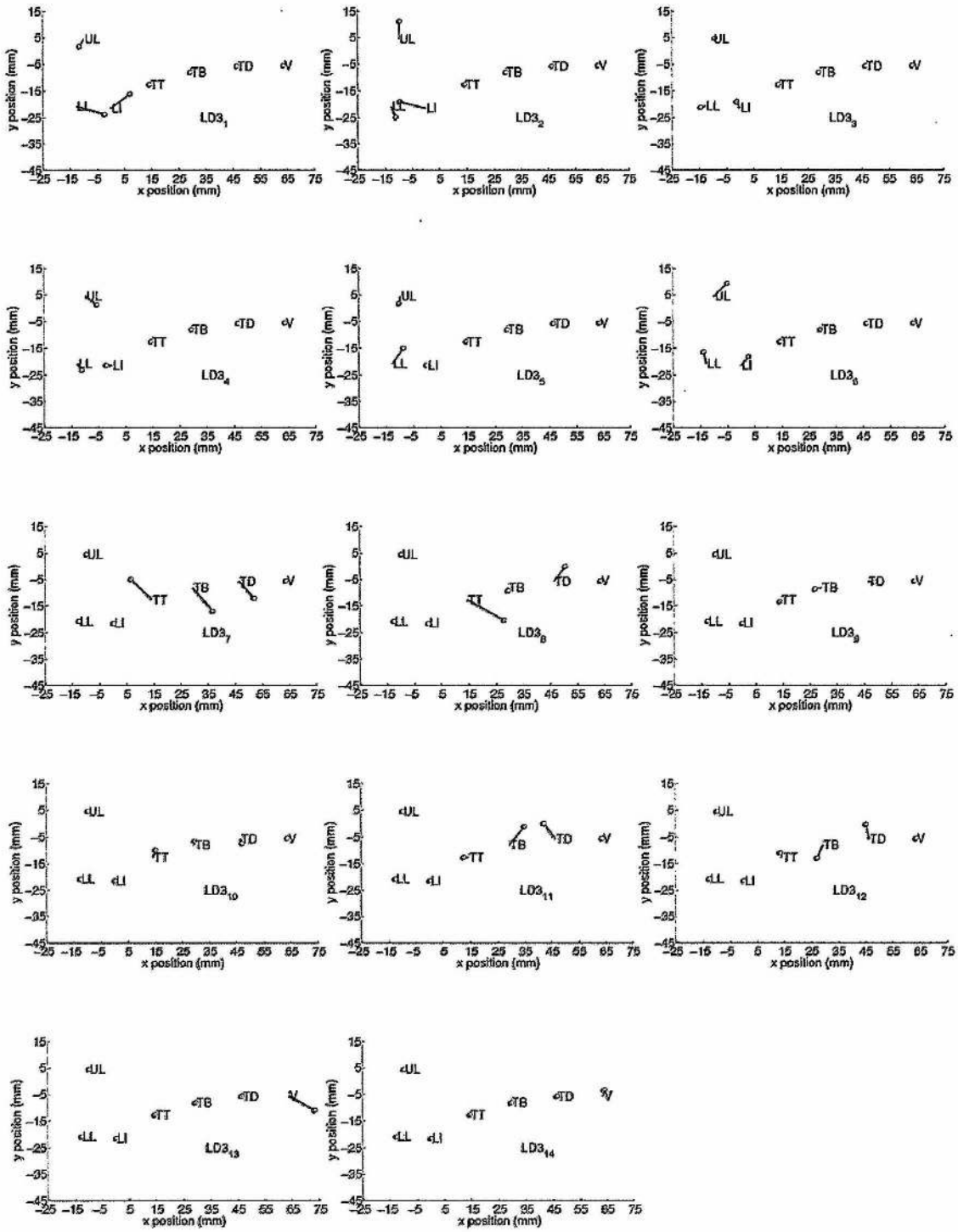



Table C.40: *LD3 mode shapes depcting the articulatory coordinate movements for female speaker.*

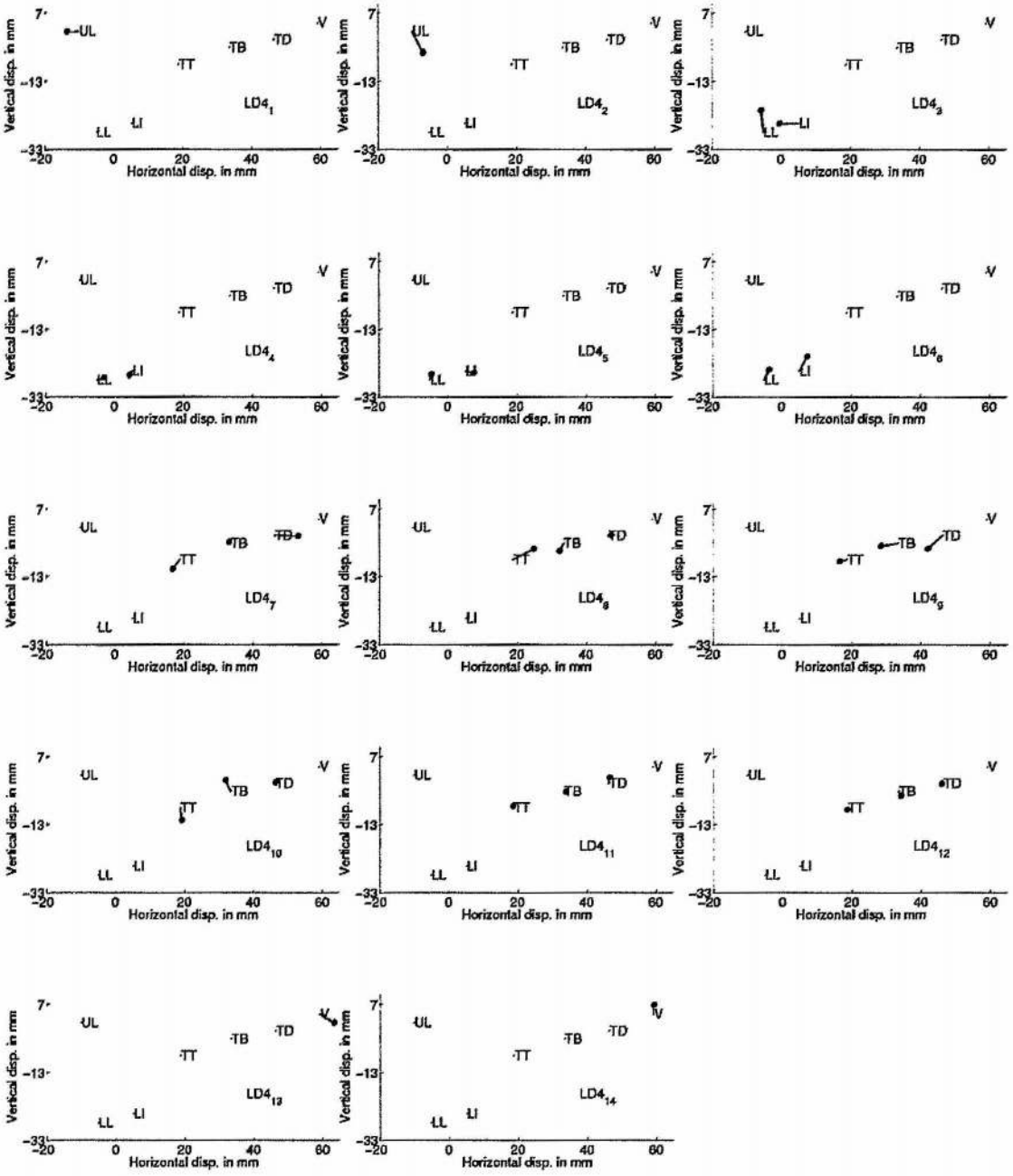



Table C.41: *LD4 mode shapes depcting the articulatory coordinate movements for male speaker.*

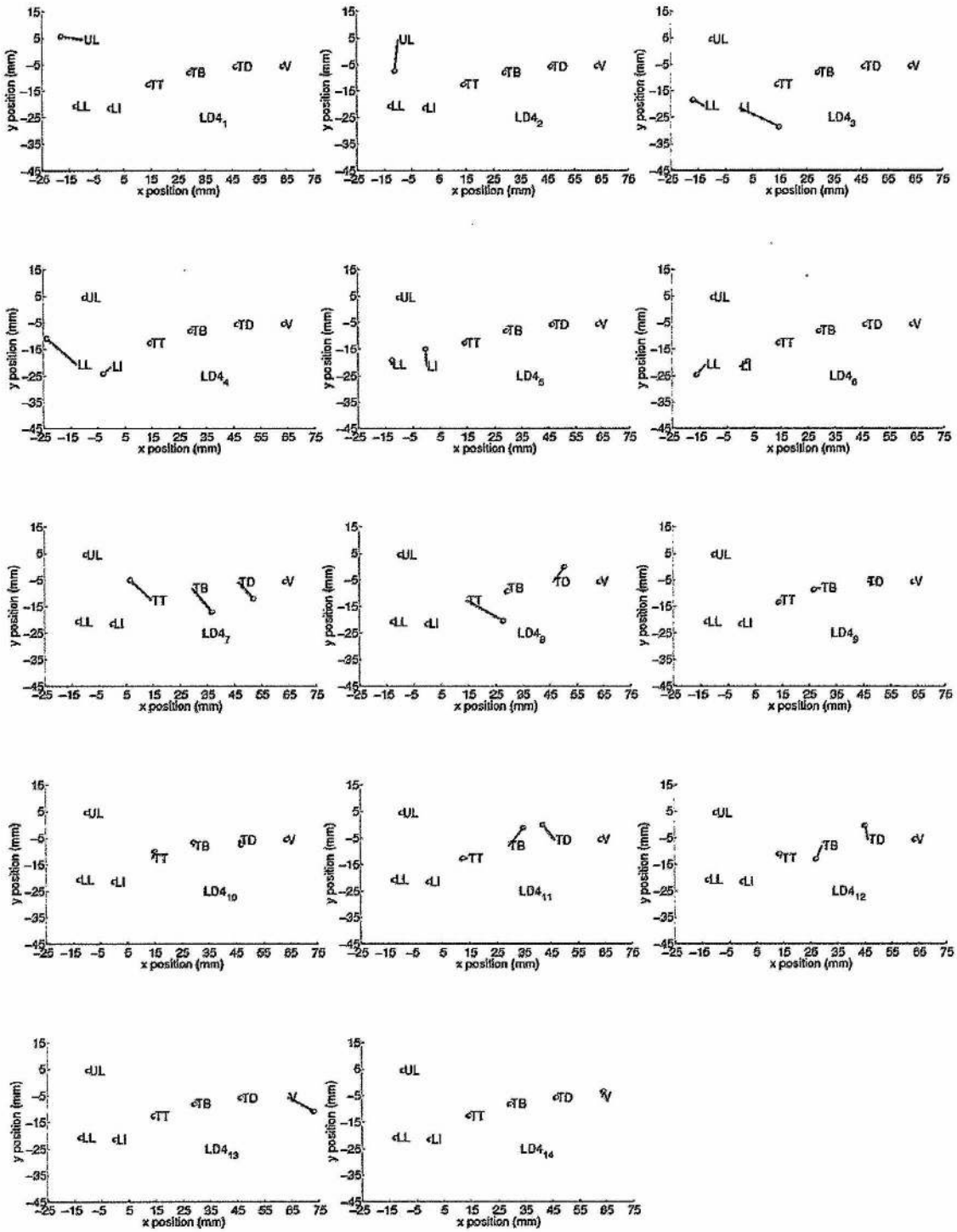



Table C.42: *LD4 mode shapes depcting the articulatory coordinate movements for female speaker.*

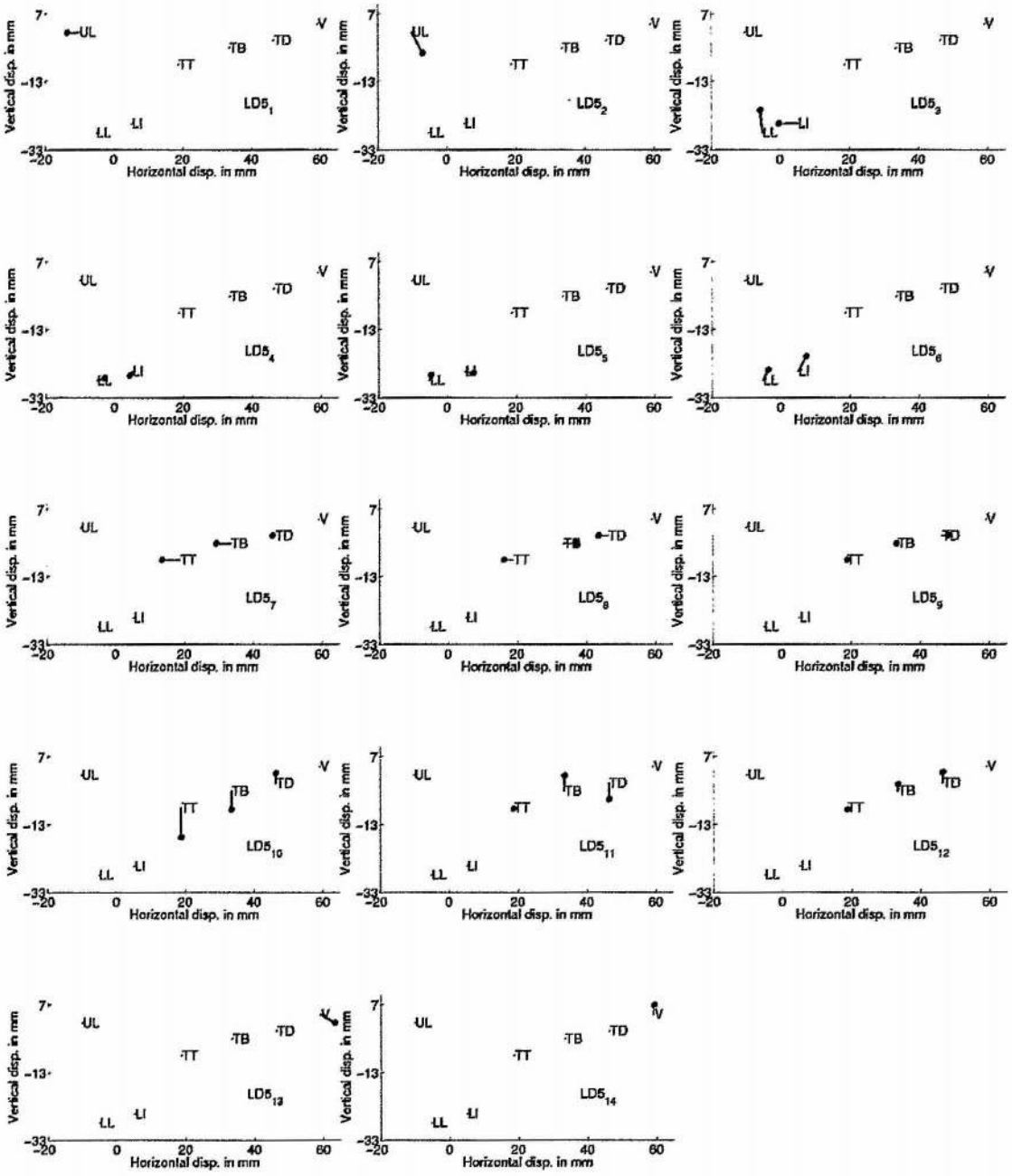



Table C.43: *LD5 mode shapes depcting the articulatory coordinate movements for male speaker.*

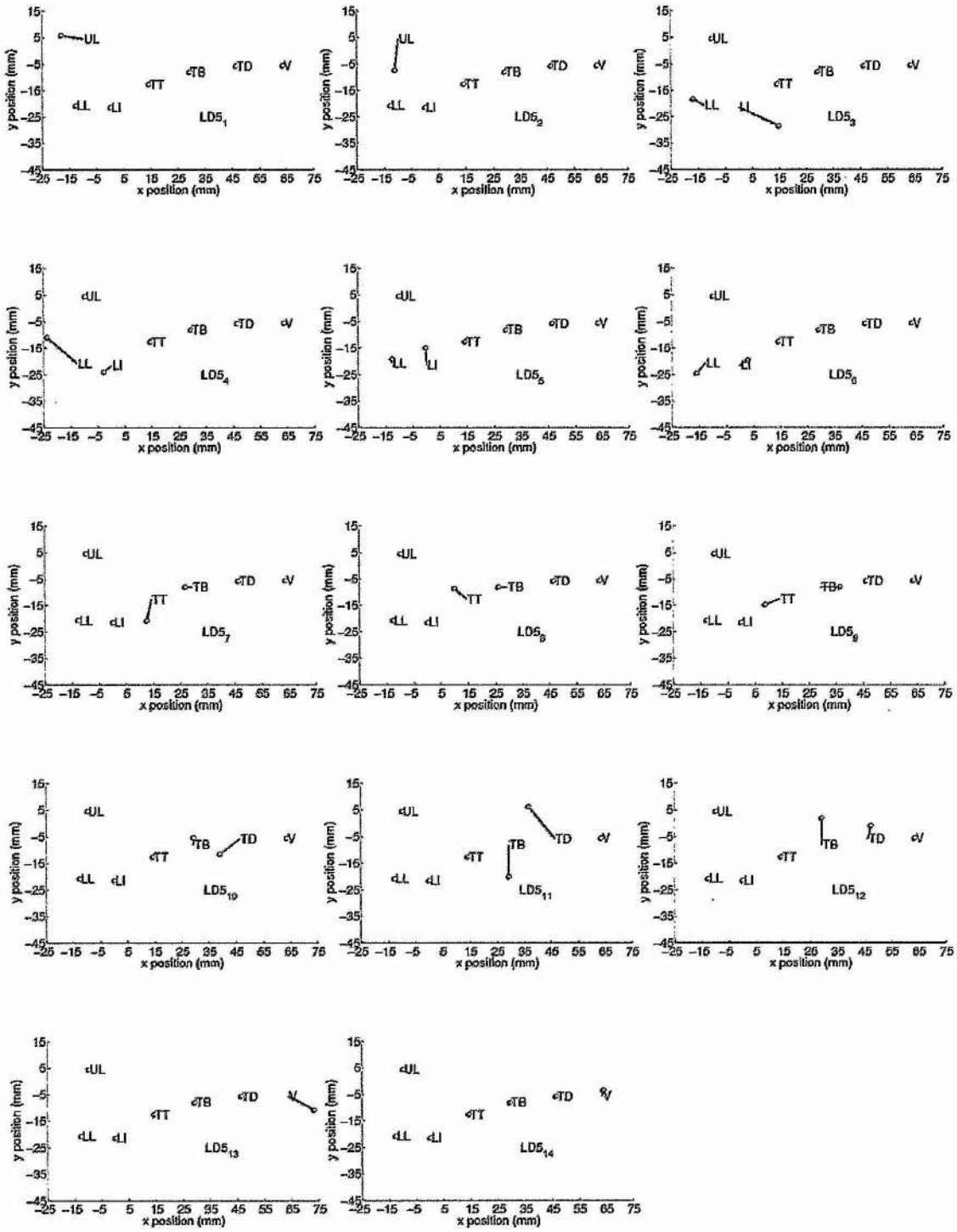



Table C.44: *LD5 mode shapes depcting the articulatory coordinate movements for female speaker.*

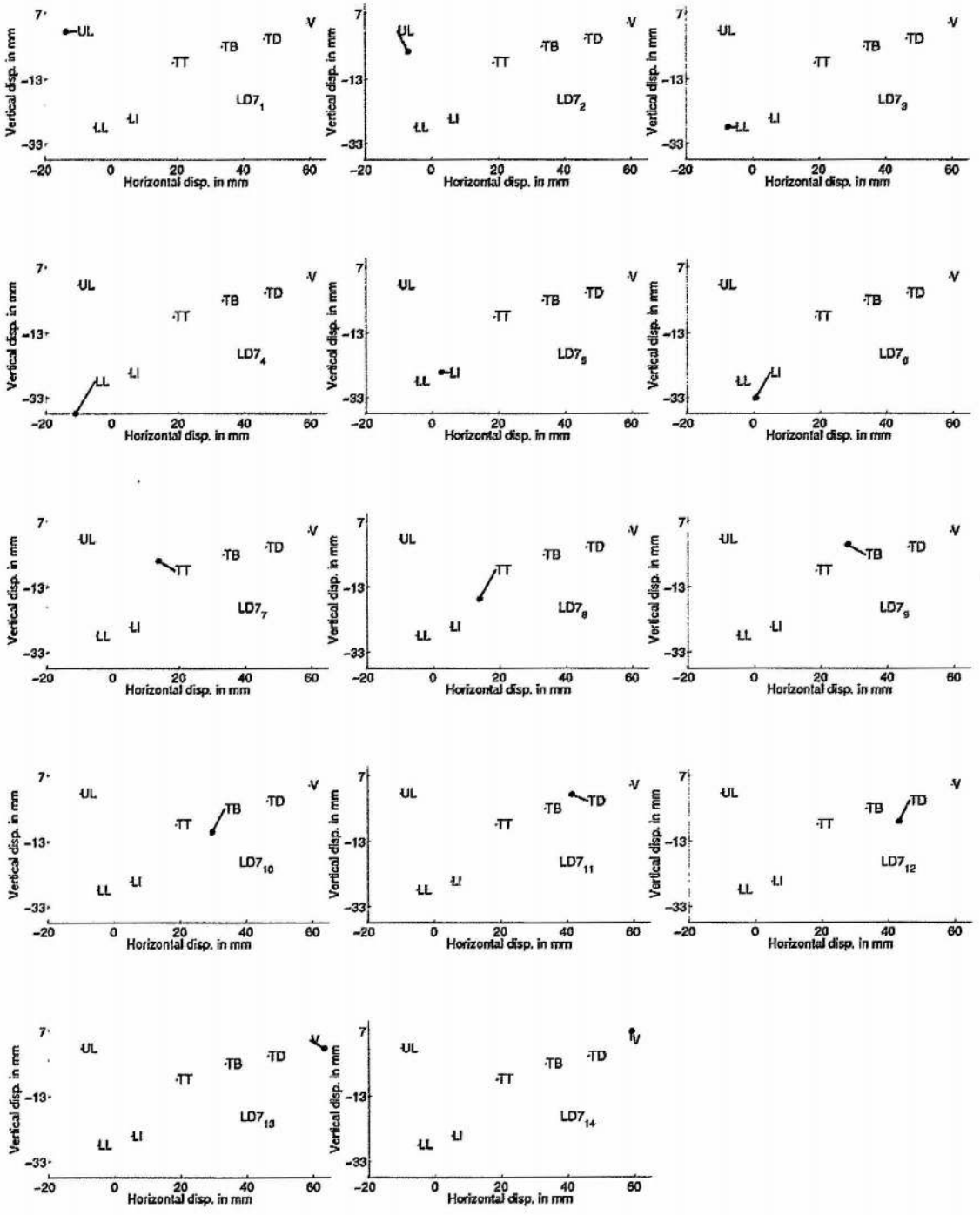



Table C.45: *LD7 mode shapes depcting the articulatory coordinate movements for male speaker.*

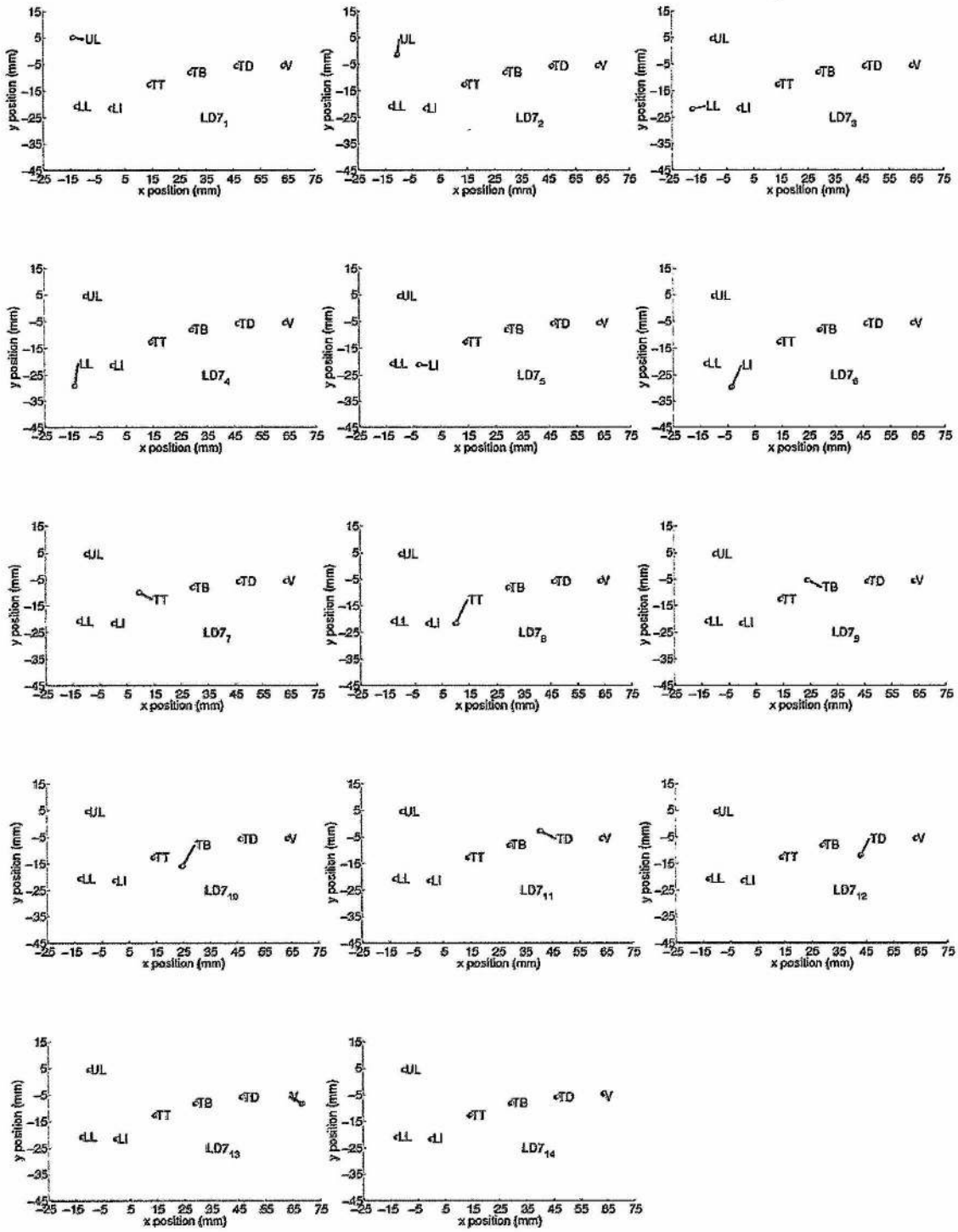



Table C.46: *LD7 mode shapes depcting the articulatory coordinate movements for female speaker.*

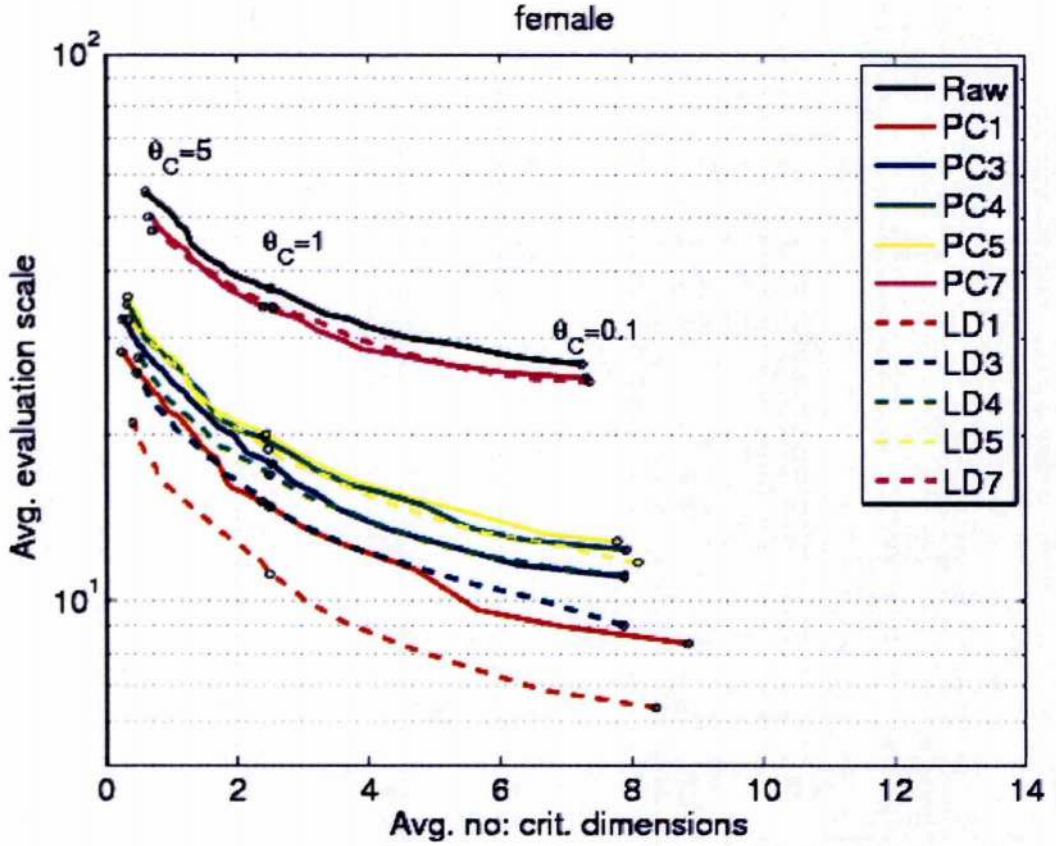



Figure C.2: *Evaluation scale $\Upsilon_{eval}$ averaged across all phones on the y axis and the average number of critical dimensions per phone on the x axis. Evaluation scale was computed from PC1, PC3, PC4, PC5, PC7 and LD1, LD3, LD4, LD5, LD7 features at various critical thresholds, $0.1 \leq \theta_C \leq 5$ for the female speaker.*

| Features | LINT | | | | BY | | | |
|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | BS | EM | ANT | CF |
| **IPA** | 2.22 | 2.19 | 2.20 | 2.26 | 2.12 | 2.18 | 2.20 | 2.22 |
| **IPA+D** | 2.04 | 2.01 | 2.01 | 2.05 | 1.87 | 1.92 | 1.91 | 1.92 |
| **Raw** | 1.94 | 1.92 | 1.92 | 1.98 | 1.76 | 1.83 | 1.84 | 1.86 |
| **PC1** | 1.95 | 1.87 | 1.91 | 2.00 | 1.78 | 1.84 | 1.91 | 1.95 |
| **PC3** | 1.97 | 1.88 | 1.91 | 1.99 | 1.81 | 1.85 | 1.89 | 1.92 |
| **PC4** | 1.98 | 1.90 | 1.92 | 2.00 | 1.82 | 1.86 | 1.90 | 1.93 |
| **PC5** | 1.97 | 1.90 | 1.93 | 2.00 | 1.82 | 1.85 | 1.89 | 1.93 |
| **PC7** | 1.94 | 1.91 | 1.91 | 1.98 | 1.76 | 1.83 | 1.84 | 1.87 |
| **LD1** | 1.93 | 1.85 | 1.90 | 1.98 | 1.79 | 1.83 | 1.91 | 1.92 |
| **LD3** | 1.90 | 1.83 | 1.85 | 1.93 | 1.72 | 1.79 | 1.82 | 1.84 |
| **LD4** | 1.89 | 1.81 | 1.84 | 1.91 | 1.69 | 1.76 | 1.80 | 1.81 |
| **LD5** | 1.93 | 1.87 | 1.89 | 1.96 | 1.76 | 1.83 | 1.86 | 1.89 |
| **LD7** | 1.88 | 1.85 | 1.85 | 1.92 | 1.69 | 1.76 | 1.78 | 1.79 |

Table C.47: *Mean RMSE in mm (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF) hypotheses for male speaker for simple linear interpolation (LINT) and Blackburn and Young (BY) models at IPA level of complexity.*

| Features | LINT | | | | BY | | | |
|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | BS | EM | ANT | CF |
| **IPA** | 0.28 | 0.30 | 0.30 | 0.25 | 0.32 | 0.30 | 0.31 | 0.27 |
| **IPA+D** | 0.56 | 0.57 | 0.57 | 0.54 | 0.61 | 0.59 | 0.60 | 0.58 |
| **Raw** | 0.55 | 0.56 | 0.55 | 0.52 | 0.59 | 0.58 | 0.57 | 0.56 |
| **PC1** | 0.56 | 0.58 | 0.55 | 0.51 | 0.60 | 0.59 | 0.55 | 0.54 |
| **PC3** | 0.55 | 0.57 | 0.55 | 0.52 | 0.59 | 0.58 | 0.56 | 0.55 |
| **PC4** | 0.56 | 0.58 | 0.56 | 0.52 | 0.60 | 0.59 | 0.58 | 0.56 |
| **PC5** | 0.56 | 0.58 | 0.56 | 0.52 | 0.60 | 0.59 | 0.58 | 0.56 |
| **PC7** | 0.56 | 0.57 | 0.56 | 0.53 | 0.60 | 0.59 | 0.58 | 0.57 |
| **LD1** | 0.57 | 0.60 | 0.57 | 0.53 | 0.61 | 0.60 | 0.57 | 0.54 |
| **LD3** | 0.56 | 0.57 | 0.56 | 0.52 | 0.60 | 0.58 | 0.56 | 0.55 |
| **LD4** | 0.58 | 0.60 | 0.59 | 0.54 | 0.63 | 0.61 | 0.60 | 0.58 |
| **LD5** | 0.59 | 0.60 | 0.58 | 0.55 | 0.63 | 0.62 | 0.60 | 0.58 |
| **LD7** | 0.56 | 0.57 | 0.56 | 0.54 | 0.61 | 0.60 | 0.59 | 0.58 |

Table C.48: *Mean RMSE in mm (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF) hypotheses for female speaker for simple linear interpolation (LINT) and Blackburn and Young (BY) models at IPA level of complexity.*

| Features | LINT | | | | BY | | | |
|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | BS | EM | ANT | CF |
| **IPA** | 0.91 | 0.90 | 0.91 | 0.92 | 0.88 | 0.90 | 0.91 | 0.91 |
| **IPA+D** | 0.84 | 0.83 | 0.83 | 0.84 | 0.78 | 0.79 | 0.79 | 0.80 |
| **Raw** | 0.80 | 0.79 | 0.79 | 0.81 | 0.73 | 0.76 | 0.76 | 0.77 |
| **PC1** | 0.81 | 0.78 | 0.80 | 0.83 | 0.76 | 0.78 | 0.80 | 0.81 |
| **PC3** | 0.81 | 0.79 | 0.80 | 0.82 | 0.76 | 0.77 | 0.79 | 0.79 |
| **PC4** | 0.82 | 0.80 | 0.81 | 0.83 | 0.77 | 0.78 | 0.80 | 0.80 |
| **PC5** | 0.82 | 0.79 | 0.81 | 0.83 | 0.76 | 0.78 | 0.79 | 0.80 |
| **PC7** | 0.81 | 0.79 | 0.80 | 0.82 | 0.74 | 0.77 | 0.77 | 0.78 |
| **LD1** | 0.80 | 0.77 | 0.79 | 0.81 | 0.75 | 0.76 | 0.79 | 0.79 |
| **LD3** | 0.79 | 0.77 | 0.78 | 0.80 | 0.73 | 0.75 | 0.77 | 0.77 |
| **LD4** | 0.79 | 0.76 | 0.77 | 0.79 | 0.72 | 0.74 | 0.76 | 0.76 |
| **LD5** | 0.80 | 0.78 | 0.79 | 0.81 | 0.74 | 0.76 | 0.78 | 0.78 |
| **LD7** | 0.78 | 0.77 | 0.78 | 0.80 | 0.71 | 0.74 | 0.75 | 0.75 |

Table C.49: *Mean normalised RMSE (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF) hypotheses for male speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at IPA level of complexity.*

| Features | LINT | | | | BY | | | |
|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | BS | EM | ANT | CF |
| **IPA** | 0.92 | 0.92 | 0.91 | 0.94 | 0.90 | 0.91 | 0.92 | 0.93 |
| **IPA+D** | 0.86 | 0.85 | 0.85 | 0.86 | 0.81 | 0.82 | 0.82 | 0.83 |
| **Raw** | 0.82 | 0.82 | 0.82 | 0.84 | 0.77 | 0.79 | 0.80 | 0.80 |
| **PC1** | 0.82 | 0.80 | 0.81 | 0.85 | 0.78 | 0.80 | 0.82 | 0.83 |
| **PC3** | 0.82 | 0.81 | 0.81 | 0.84 | 0.78 | 0.79 | 0.81 | 0.82 |
| **PC4** | 0.82 | 0.80 | 0.81 | 0.84 | 0.78 | 0.79 | 0.80 | 0.81 |
| **PC5** | 0.82 | 0.80 | 0.81 | 0.84 | 0.78 | 0.79 | 0.80 | 0.81 |
| **PC7** | 0.82 | 0.81 | 0.81 | 0.84 | 0.77 | 0.79 | 0.80 | 0.81 |
| **LD1** | 0.83 | 0.80 | 0.82 | 0.85 | 0.79 | 0.80 | 0.82 | 0.83 |
| **LD3** | 0.82 | 0.81 | 0.81 | 0.84 | 0.78 | 0.80 | 0.81 | 0.82 |
| **LD4** | 0.82 | 0.80 | 0.81 | 0.84 | 0.77 | 0.79 | 0.80 | 0.81 |
| **LD5** | 0.82 | 0.81 | 0.82 | 0.84 | 0.77 | 0.80 | 0.81 | 0.82 |
| **LD7** | 0.82 | 0.82 | 0.82 | 0.84 | 0.78 | 0.80 | 0.80 | 0.81 |

Table C.50: *Mean normalised RMSE (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF) hypotheses for female speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at IPA level of complexity.*

| Features | LINT | | | | | BY | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | CONV | BS | EM | ANT | CF | CONV |
| **Raw** | 1.80 | 1.79 | 1.79 | 1.85 | 1.73 | 1.57 | 1.67 | 1.69 | 1.69 | 1.47 |
| **PC1** | 1.83 | 1.73 | 1.77 | 1.87 | 1.73 | 1.64 | 1.67 | 1.75 | 1.79 | 1.49 |
| **PC3** | 1.80 | 1.73 | 1.75 | 1.81 | 1.73 | 1.59 | 1.65 | 1.70 | 1.69 | 1.49 |
| **PC4** | 1.81 | 1.73 | 1.75 | 1.82 | 1.73 | 1.59 | 1.66 | 1.70 | 1.70 | 1.49 |
| **PC5** | 1.81 | 1.75 | 1.76 | 1.83 | 1.73 | 1.60 | 1.67 | 1.70 | 1.71 | 1.49 |
| **PC7** | 1.79 | 1.79 | 1.77 | 1.82 | 1.73 | 1.57 | 1.67 | 1.67 | 1.67 | 1.47 |
| **LD1** | 1.79 | 1.68 | 1.74 | 1.82 | 1.73 | 1.59 | 1.62 | 1.72 | 1.72 | 1.50 |
| **LD3** | 1.80 | 1.70 | 1.73 | 1.80 | 1.73 | 1.58 | 1.63 | 1.69 | 1.69 | 1.47 |
| **LD4** | 1.80 | 1.71 | 1.73 | 1.81 | 1.73 | 1.58 | 1.63 | 1.68 | 1.69 | 1.47 |
| **LD5** | 1.81 | 1.73 | 1.74 | 1.82 | 1.73 | 1.58 | 1.65 | 1.68 | 1.70 | 1.48 |
| **LD7** | 1.78 | 1.75 | 1.74 | 1.80 | 1.73 | 1.55 | 1.63 | 1.64 | 1.64 | 1.47 |

Table C.51: *Mean RMSE in mm (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF), conventional (CONV) hypotheses for male speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at 2×IPA level of complexity.*

| Features | LINT | | | | | BY | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | CONV | BS | EM | ANT | CF | CONV |
| **Raw** | 1.97 | 1.97 | 1.97 | 2.05 | 1.88 | 1.79 | 1.87 | 1.89 | 1.91 | 1.68 |
| **PC1** | 1.98 | 1.91 | 1.94 | 2.05 | 1.88 | 1.84 | 1.87 | 1.94 | 1.97 | 1.70 |
| **PC3** | 1.97 | 1.92 | 1.93 | 2.02 | 1.88 | 1.82 | 1.88 | 1.91 | 1.93 | 1.70 |
| **PC4** | 1.96 | 1.92 | 1.92 | 2.01 | 1.88 | 1.81 | 1.86 | 1.89 | 1.91 | 1.70 |
| **PC5** | 1.97 | 1.92 | 1.92 | 2.02 | 1.88 | 1.81 | 1.87 | 1.90 | 1.92 | 1.70 |
| **PC7** | 1.96 | 1.97 | 1.94 | 2.02 | 1.88 | 1.78 | 1.88 | 1.87 | 1.89 | 1.68 |
| **LD1** | 1.97 | 1.90 | 1.94 | 2.05 | 1.88 | 1.83 | 1.87 | 1.93 | 1.97 | 1.71 |
| **LD3** | 1.98 | 1.92 | 1.94 | 2.01 | 1.88 | 1.83 | 1.88 | 1.91 | 1.92 | 1.70 |
| **LD4** | 1.98 | 1.93 | 1.94 | 2.02 | 1.88 | 1.82 | 1.88 | 1.91 | 1.92 | 1.69 |
| **LD5** | 1.99 | 1.96 | 1.96 | 2.06 | 1.88 | 1.84 | 1.91 | 1.94 | 1.97 | 1.70 |
| **LD7** | 1.97 | 1.97 | 1.94 | 2.03 | 1.88 | 1.80 | 1.88 | 1.87 | 1.91 | 1.69 |

Table C.52: *Mean RMSE in mm (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF), conventional (CONV) hypotheses for female speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at 2×IPA level of complexity.*

| Features | LINT | | | | | BY | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | CONV | BS | EM | ANT | CF | CONV |
| **Raw** | 0.75 | 0.75 | 0.75 | 0.77 | 0.72 | 0.67 | 0.70 | 0.71 | 0.71 | 0.62 |
| **PC1** | 0.77 | 0.73 | 0.75 | 0.78 | 0.72 | 0.70 | 0.71 | 0.74 | 0.75 | 0.63 |
| **PC3** | 0.75 | 0.73 | 0.74 | 0.75 | 0.72 | 0.67 | 0.70 | 0.71 | 0.71 | 0.63 |
| **PC4** | 0.75 | 0.73 | 0.74 | 0.76 | 0.72 | 0.68 | 0.70 | 0.71 | 0.71 | 0.63 |
| **PC5** | 0.76 | 0.74 | 0.74 | 0.76 | 0.72 | 0.68 | 0.70 | 0.72 | 0.72 | 0.63 |
| **PC7** | 0.75 | 0.74 | 0.74 | 0.76 | 0.72 | 0.67 | 0.70 | 0.71 | 0.71 | 0.63 |
| **LD1** | 0.75 | 0.71 | 0.73 | 0.75 | 0.72 | 0.68 | 0.69 | 0.73 | 0.72 | 0.64 |
| **LD3** | 0.75 | 0.72 | 0.73 | 0.75 | 0.72 | 0.68 | 0.69 | 0.71 | 0.71 | 0.63 |
| **LD4** | 0.75 | 0.72 | 0.73 | 0.76 | 0.72 | 0.67 | 0.69 | 0.71 | 0.71 | 0.63 |
| **LD5** | 0.75 | 0.73 | 0.73 | 0.76 | 0.72 | 0.67 | 0.70 | 0.71 | 0.72 | 0.63 |
| **LD7** | 0.75 | 0.73 | 0.73 | 0.75 | 0.72 | 0.66 | 0.69 | 0.70 | 0.69 | 0.63 |

Table C.53: *Mean normalised RMSE (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF), conventional (CONV) hypotheses for male speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at 2×IPA level of complexity.*

| Features | LINT | | | | | BY | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BS | EM | ANT | CF | CONV | BS | EM | ANT | CF | CONV |
| **Raw** | 0.78 | 0.78 | 0.78 | 0.80 | 0.75 | 0.72 | 0.74 | 0.74 | 0.75 | 0.67 |
| **PC1** | 0.79 | 0.76 | 0.77 | 0.81 | 0.75 | 0.74 | 0.75 | 0.77 | 0.78 | 0.68 |
| **PC3** | 0.79 | 0.77 | 0.77 | 0.80 | 0.75 | 0.73 | 0.75 | 0.76 | 0.77 | 0.68 |
| **PC4** | 0.78 | 0.77 | 0.77 | 0.80 | 0.75 | 0.73 | 0.74 | 0.75 | 0.76 | 0.68 |
| **PC5** | 0.78 | 0.76 | 0.77 | 0.80 | 0.75 | 0.72 | 0.74 | 0.75 | 0.76 | 0.68 |
| **PC7** | 0.78 | 0.78 | 0.77 | 0.80 | 0.75 | 0.72 | 0.75 | 0.75 | 0.75 | 0.67 |
| **LD1** | 0.78 | 0.75 | 0.77 | 0.80 | 0.75 | 0.73 | 0.74 | 0.76 | 0.77 | 0.68 |
| **LD3** | 0.78 | 0.76 | 0.77 | 0.79 | 0.75 | 0.73 | 0.74 | 0.76 | 0.76 | 0.68 |
| **LD4** | 0.78 | 0.76 | 0.77 | 0.79 | 0.75 | 0.73 | 0.74 | 0.75 | 0.76 | 0.68 |
| **LD5** | 0.79 | 0.77 | 0.78 | 0.80 | 0.75 | 0.73 | 0.76 | 0.76 | 0.77 | 0.68 |
| **LD7** | 0.78 | 0.78 | 0.77 | 0.80 | 0.75 | 0.72 | 0.75 | 0.75 | 0.76 | 0.68 |

Table C.54: *Mean normalised RMSE (averaged across all sentences and articulators) between measured trajectory and synthetic trajectories generated using baseline (BS), effort minimisation (EM), anticipatory (ANT), carry forward (CF), conventional (CONV) hypotheses for female speaker for linear interpolation (LINT) and Blackburn and Young (BY) models at 2×IPA level of complexity.*

# Bibliography


Ananthakrishnan, G., Engwall, O., 2008. Important regions in the articulator trajectory. Proc. Int. Sem. on Spch. Prod. (ISSP'08)*Strasbourg*, 305–308.

Ananthakrishnan, G., Neiberg, D., Engwall, O., 2009. In search of non-uniqueness in the acoustic to articulatory mapping. Proc. Interspeech, *Brighton*, 2799–2802.

Anderson, T., 1984. An introduction to multivariate statistical analysis, 2nd Edition. Wiley, New York.

Atal, B., J., C., Mathews, M., Tukey, J., 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. J. Acoust. Soc. Am. 63 (5), 1535–1556.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C., 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. J. Phon. 30 (3), 533–553.

Badin, P., Serrurier, A., 2006. Three-dimensional linear modeling of tongue: Articulatory data and models. In: Laboissière, R. (Ed.), Proc. Int. Sem. Spch. Prod. (ISSP'06). Ubatuba, Brazil, pp. 395–402.

Bailly, G., Abry, C., Boe, L., Laboissiere, R., Perrier, P., Schwartz, J., 1992. Inversion and speech recognition. In: Proc. EUSIPCO. Vol. 1. pp. 159–164.

Bakis, R., 1991. Coarticulation modeling with continuous-state HMMs. Proc. IEEE Workshop Automatic Speech Recognition, Harriman, New York, 20–21.

Beskow, J., 1995. Rule-based visual speech synthesis. In: In Proceedings of Eurospeech '95. pp. 299–302.

Blackburn, S., Young, S., March 2000. A self-learning predictive model of articulator movements during speech production. J. Acoust. Soc. Am. 103 (3), 1659–70.

Bladon, R. A. W., Al-Bamerni, A., 1976. Coarticulation resistance in English /l/. J. Phon. 4, 135–50.

Box, G. E. P., Cox, D. R., 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B (Methodological) 26 (2), 211–252.

Browman, C. P., Goldstein, L., 1986. Towards an articulatory phonology. Phonology 3, 219–52.

Butterworth, B., 1980. Some constraints on models of language production. In: Butterworth, B. (Ed.), Language production: Speech and Talk. Vol. 1. Academic Press, London.

Chang, S., Greenberg, S., Wester, M., 2001. An elitist approach to articulatory-acoustic feature classification. In: Proc. Eurospeech, Aalborg, Denmark. pp. 1725–1728.

Chen, J.-Y., Hershey, J., Olsen, P., Yashchin, E., 31 2008-April 4 2008. Accelerated monte carlo for kullback-leibler divergence between gaussian mixture models. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. pp. 4553–4556.

Chomsky, N., Halle, M., 1968. The sound pattern of English. Harper & Row, New York.

Cohen, M., Grossberg, S., Stork, D., 1988. Speech perception and production by a self-organizing neural network. In: Lee, Y. (Ed.), Evolution, learning, cognition, and advanced architectures. World Scientific Publishers, Hong Kong.

Cohen, M. M., Massaro, D. W., 1993. Modeling coarticulation in synthetic visual speech. In: Models and Techniques in Computer Animation. Springer-Verlag, pp. 139–156.

Coker, C. H., 1976. A model of articulatory dynamics and control. Proc. IEEE 64 (4), 452–460.

Collins, M. J., Krishnamurthy, A. K., Ahalt, S. C., 1999. Generating gestural scores from articulatory data using temporal decomposition. IEEE Transactions on Speech and Audio Processing 7, 230–233.

Coppin, B., 2004. Artificial Intelligence Illuminated, 1st Edition. Jones & Bartlett, ISBN 0763732303.

Dang, J., Honda, K., 2004. Construction and control of a physiological articulatory model. J. Acoust. Soc. Am. 115 (2), 853–870.

Dang, J., Honda, M., Honda, K., 2004. Investigation of coarticulation in continuous speech of Japanese. Acoust. Sci. & Tech. 25 (5), 318 – 329.

Dang, J., Tiede, M., Yuan, J., 2009. Comparison of vowel structures of Japanese and English in articulatory and auditory spaces. Proc. Interspeech, *Brighton*, 2815–2818.

Dang, J., Wei, J., Suzuki, T., Perrier, P., 2005. Investigation and modelling of coarticulation during speech. Proc. Interspeech, *Lisbon*, 1025–1028.

Daniloff, R., Hammarberg, R., 1973. On defining coarticulation. J. Phon. 1, 239–248.

Deng, L., Ma, J., 2000. Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics. J. Acoust. Soc. Am. 108 (6), 3036–3048.

Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. scomm 22, 93–111.

Deng, L., Sun, D. X., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Am. 95 (5), 2702–2719.

Dirven, R., Verspoor, M. (Eds.), 2004. Cognitive Exploration of Language and Linguistics. John Benjamins Publishing Co.

Dressler, W. U., Prinzhorn, M., Rennison., J. R. (Eds.), 1992. Phonologica. Rosenberg & Sellier (Torino).

Eide, E., 2001. Distinctive features for use in an automatic speech recognition system. Proc. Eurospeech '01, *Aalborg, Denmark*, 1613–1313.

Erler, K., Freeman, G. H., 1996. An HMM-based speech recognizer using overlapping articulatory features. J. Acoust. Soc. Am. 100 (4), 2500–2513.

Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue musles: a TMS study. European Journal of Neuroscience 15, 399–402.

Fant, G., 1969. Distinctive features and phonetic dimensions. Applications of Linguistics, Cambridge, UK.

Field, A., 2005. Discovering statistics using SPSS, 2nd Edition. SAGE Pub.

Fowler, C. A., Saltzman, E., 1993. Coordination and coarticulation in speech production. Language and Speech 36, 171–195.

Frankel, J., 2003. Linear dynamic models for automatic speech recognition. Ph.D. thesis, CSTR, Univ. of Edinburgh.

Frankel, J., King, S., 2001. ASR-Articulatory Speech Recognition. Proc. Eurospeech '01, *Aalborg, Denmark*, 599–602.

Frankel, J., Richmond, K., King, S., Taylor, P., 2000. An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. Proc. Int. Conf. on Spoken Lang. Proc., *Beijing* 4, 254–257.

Frankel, J., Wester, M., King, S., 2004. Articulatory feature recognition using dynamic Bayesian networks. Proc. Int. Conf. on Spoken Lang. Proc., *Jeju, Korea*, 1477–1480.

Frankel, J., Wester, M., King, S., Oct. 2007. Articulatory feature recognition using dynamic Bayesian networks. Computer Speech & Language 21 (4), 620–640.

Garrett, M., 1980. Levels of processing in sentence production. In: Butterworth, B. (Ed.), Language production: Speech and Talk. Vol. 1. Academic Press, London.

Ghosh, P., Narayanan, S., Divenyi, P., Goldstein, L., Saltzman, E., 2009. Estimation of articulatory gesture patterns from speech acoustics. Proc. Interspeech, *Brighton*, 2803–2806.

Guenther, F., 1994. A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics 72, 43–53.

Guenther, F. H., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review 102, 594–621.

Henke, W. L., 1965. Dynamic articulatory model of speech production using computer simulation. Ph.D. thesis, MIT, Cambridge, MA.

Hershey, J. R., Olsen, P. A., 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. Proc. IEEE-ICASSP 4, 317–320.

Hiroya, S., Mochida, T., 2005. Multi-speaker articulatory reconstruction based on eigen articulatory HMM. Proc. IEEE-ICASSP 1, 909– 912.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., Saltzman, E., 1996. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. J. Acoust. Soc. Am. 100 (3), 1819–1834.

Holmes, W., Russell, M., 1995. Experimental evaluation of segmental HMMs. Proc. IEEE-ICASSP, 536.

Holmes, W., Russell, M., 1996. Modelling speech variability with segmental HMMs. Proc. IEEE-ICASSP, 447.

Holmes, W., Russell, M., 1997. Linear dynamic segmental HMMs: Variability representation and training procedure. Proc. IEEE-ICASSP, 1399.

Hyvärinen, A., Oja, E., 2000. Independent Component Analysis: Algorithms and Applications. Neural Networks 13 (4-5), 411–430.

International Phonetic Association, 2003. Handbook of the International Phonetic Association: A guide to the use of International Phonetic Alphabet. Cambridge University Press.

Izenman, A., 1991. Recent developments in nonparametric density estimation. Journal of the American Statistical Association 86 (413), 205–224.

Jackson, E. J., 1991. A user's guide to principal components. Wiley.

Jackson, P., Lo, B., Russell, M. J., 2002. Data-driven, non-linear, formant-to-acoustic mapping for ASR. Electronics Letters 38, 667.

Jackson, P. J., Singampalli, V. D., 2009. Statistical identification of critical articulators in the production of speech. scomm 51, 695–710.

Jackson, P. J. B., Singampalli, V., Shiga, Y., Russell, M. J., 2004. Dansa project: Statistical models to relate speech gestures to meaning. CVSSP, Univ. of Surrey, Guildford, UK, EPSRC GR/S85511/01 [http://www.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/].

Jackson, P. J. B., Singampalli, V. D., Dec. 2008a. Coarticulatory constraints determined by automatic identification from articulograph data. In: Proc. Int. Sem. on Spch. Prod. (ISSP'08)*Strasbourg*. Strasbourg, France, pp. 377–380.

Jackson, P. J. B., Singampalli, V. D., 2008b. Statistical identification of critical, dependent and redundant articulators. In: J. Acoust. Soc. Am. Vol. 123. p. 3321, Presented at Acoustics'08, Paris.

Johnson, R. A., Wichern, D. W., 1998. Applied multivariate statistical analysis, 4th Edition. Prentice Hall, New Jersey.

Jung, T.-P., Krishnamurthy, A., Ahalt, S., Beckman, M., Lee, S., 1996. Deriving gestural score from articulator-movement records using weighted temporal decomposition. IEEE Transactions on Speech and Audio Processing 4, 2–18.

Kaburagi, T., Honda, M., 2001. Dynamic articulatory model based on multidimensional invariant-feature task representation. J. Acoust. Soc. Am. 110 (1), 441–452. URL `http://link.aip.org/link/?JAS/110/441/1`

Kaburagi, T., Kim, J., 2007. Generation of the vocal tract spectrum from the underlying articulatory mechanism. J. Acoust. Soc. Am. 121 (1), 456–68.

Keating, P. A., 1988. The window model of coarticulation: articulatory evidence. UCLA Working papers in Phonetics 69, 3–29.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., February 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. 121 (2), 723–742.

Kirchhoff, K., 1999. Robust speech recognition using articulatory information. Ph.D. thesis, Univ. of Bielefeld.

Kirirani, S., Sekimoto, S., Imagawa, H., Fujisaki, H., 1977. Parameter descriptions of the tongue movements for vowel. Contribution papers of the 9th ICA 1, 419.

Koreman, J., Andreeva, B., Barry, W. J., 1998. Do phonetic features help to improve consonant identification in ASR? Proc. Int. Conf. on Spoken Lang. Proc., *Sidney, Australia* 3, 1035–1038.

Krňoul, Z., Železný, M., Müller, L., Kanis, J., 2006. Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis. Proc. Interspeech, *Pittsburgh, PA* 1, 585–588.

Kullback, S., 1968. Information theory and statistics, 1st Edition. Dover Pub., New York.

Ladefoged, P., 1975. A course in phonetics, 1st Edition. Harcourt Brace Jovanovich.

Ladefoged, P., 2005. Vowels and consonants, 2nd Edition. Wiley-Blackwell.

Larar, J. N., Schroeter, J., Sondhi, M., 1988. Vector quantisation of the articulatory space. IEEE Transactions oon Acoustics, Speech and Signal Processing 36 (12), 1812–1818.

Levelt, W., 1989. Speaking: From Intention to Articulation. The MIT Press.

Liberman, A. M., 1970. The grammars of speech and language. Cog. Psych. 1, 301–23.

Liese, F., Vajda, I., 2006. On divergences and informations in statistics and information theory. IEEE transactions on Information Theory 52 (10), 4394–4412.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773–81.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H & H theory. In: Hardcastle, W., Marchal, A. (Eds.), Speech production and Speech Modeling. Kluwer Academic Publishers, pp. 403–439.

Lindblom, B., Sussman, H., 2002. Principal components analysis of tongue shapes in symmetrical VCV utterances. Speech, music and hearing 44.

Livescu, K., Glass, J., Bilmes, J., 2003. Hidden feature modelling for speech recognition using dynamic Bayesian networks. In: Proc. Eurospeech, Geneva, Switzerland. Vol. 4. pp. 2529–2532.

Löfqvist, A., 1990. Speech as audible gestures. In: Hardcastle, W., Marchal, A. (Eds.), Speech production and Speech Modeling. Kluwer Academic Publishers, pp. 289–322.

Ma, J., Deng, L., Jan 2004. Target-directed mixture linear dynamical models for spontaneous speech recognition. IEEE Transactions on Speech and Audio Processing 12, 47.

MacNeilage, P. F., 1970. Motor control of serial ordering of speech. Psychol. Rev. 77, 182–196.

Maeda, S., 1990. Speech Production and Modelling. Kluwer, Ch. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model.

Markov, K., Dang, J., Nakamura, S., 2006. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. Speech Comm. 48, 161–175.

Massey, Jr., F. J., March 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46 (253).

McDermott, E., Nakamura, A., 2006. Production-oriented models forl speech recognition. IEICE Trans. Inf. & Syst. 89, 1006–1014.

McLachlan, G. M., 2004. Discriminant analysis and statistical pattern recognition. WileyBlackwell.

Meister, I. C., Wilson, S. M., Deblieck, C., Wu, Allan D. amd Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Current Biology 17 (19), 1692–1696.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am. 53 (4), 1072–1082.

Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features. Proc. ICSLP, *Denver, CO*, 2133–2136.

Moll, K., Daniloff, R., 1971. Investigation of the timing of velar movements during speech. J. Acoust. Soc. Am. 50 (2), 678–84.

Moon, S. J., Lindblom, B., 1994. Interaction between duration, context ,and speaking style in English stressed vowels. J. Acoust. Soc. Am. 96, 40–55.

Morrison, D. F., 1990. Multivariate statistical methods, 3rd Edition. McGraw-Hill.

Öhman, S. E. G., 1966. Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am. 39 (1), 151–68.

Öhman, S. E. G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41 (2), 310–20.

Okadome, T., Honda, M., July 2001. Generation of articulatory movements by using a kinematic triphone model. J. Acoust. Soc. Am. 110 (1), 453–463.

Okadome, T., Suzuki, S., Honda, M., 2000. Recovery of articulatory information from acoustics with phonemic information. In: Proc. 5th Seminar on Speech Production, Kloster Seeon, Bavaria. pp. 229–232.

Ostendorf, M., 1999. Moving beyond the "beads-on-a-stringmodel of speech. In: Proc. IEEE Automatic Speech recognition and Understanding Workshop. Vol. 1. Keystone, Colorado, USA, pp. 79–88.

Ostry, D. J., Gribble, P. L., Gracco, V. L., 1996. Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? The Journal of Neuroscience 16 (4), 1570–1579.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., Levy, S., 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. J. Acoust. Soc. Am. 92 (2), 688–700.

Parthasarathy, V., Prince, J. L., Stone, M., Murano, E. Z., Nessaiver, M., 2007. Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. J. Acoust. Soc. Am. 121 (1), 491–504.

Perkell, J., 1980. Phonetic features and physiology of speech production. In: Butterworth, B. (Ed.), Language production: Speech and Talk. Vol. 1. Academic Press, London, pp. 337–372.

Qin, C., Carreira-Perpiñán, M., Richmond, K., Wrench, A., Renals, S., 2008. Predicting tongue shapes from a few landmark locations. In: Proc. Interspeech. Brisbane, Australia, pp. 2306–2309.

Recasens, D., Pallarés, D., 1999. A study of /ɾ/ and /r/ in the light of the 'DAC' coarticulation model. J. Phon., 143 – 170.

Recasens, D., Pallarés, D. M., Fontdevilla, J., 1997. A model of lingual coarticulation based on articulatory constraints. J. Acoust. Soc. Am. 102 (1), 544–561.

Richards, H. B., Bridle, J. S., 1999. The HDM: A segmental hidden dynamical model of coarticulation. Int. Conf. on Acoustics, Speech, and Signal Processing, *Phoenix, Arizona, USA* 1, 357–360.

Richardson, M., Blimes, J., Diorio, C., 2000. Hidden-articulatory Markov models for speech recognition. Proc. Int. Conf. on Spoken Lang. Proc., *Beijing* 3, 131–134.

Richmond, K., 2001. Estimating the articulatory parameters from the acoustic speech signal. Ph.D. thesis, Univ. of Edinburgh.

Richmond, K., 2006. A trajectory mixture density network for the acoustic-articulatory inversion mapping. Proc. Interspeech, *Pittsburgh, PA.*

Richmond, K., 2007a. A multitask learning perspective on acoustic-articulatory inversion. In: Proc. Interspeech. Antwerp, Belgium, pp. 2465–2468.

Richmond, K., 2007b. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion.,. In: Chetouani, M., Hussain, A., Gas, B., Milgram, M., Zarade, Z. (Eds.), International Conference on Non-Linear Speech Processing. Springer-Verlag Berlin Heidelberg, pp. 263–272.

Richmond, K., 2009. Preliminary inversion mapping results with a new EMA corpus. Proc. Interspeech, *Brighton*, 2835–2838.

Rose, R. C., Schroeter, J., Sondhi, M. M., 1996. The potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Am. 3 (99), 1699–1709.

Rosner, B. S., Pickering, J., 1994. Vowel perception and production. OUP Oxford.

Roweis, S., 1999. Data driven production models for speech processing. Ph.D. thesis, California Institute of Technology.

Rubin, P., Vatikiotis-Bateson, E., 1998. Animal Acoustic Communication. Springer-Verlag, Ch. Measuring and modelling speech production, pp. 251–290.

Russell, M., Jackson, P., 2005. A multiple-level linear/linear segmental hmm with a formant-based intermediate layer. Computer Speech and Language 19, 205–225.

Russell, M. J., Holmes, W., 1997. Linear trajectory segmental HMMs. IEEE Sig. Proc. Letters 4, 72–74.

Russell, M. J., Jackson, P. J. B., 2002. Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations. Proc. ICSLP, *Denver, CO*, 1253–1256.

Saltzman, E. L., Munhall, K., 1989. A dynamic approach to gestural patterning in speech production. Ecology Psychology 1 (4), 333–82.

Schroeter, J., Sondhi, M. M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. SAP 2 (1), 133–150.

Schwartz, R., Chow, Y. L., Kimball, O., Roucos, S., Krasner, M., Makhoul, J., 1985. Context-dependent modelling for acoustic recognition in continuous speech. In: icaspp. pp. 1205–1208.

Shadle, C. H., 1985. The acoustics of fricative consonants. Ph.D. thesis, MIT, Cambridge, MA.

Sheskin, D. J., 2000. Handbook of parametric and nonparametric statistical procedures, 2nd Edition. Chapman and Hall/CRC, USA.

Shiga, Y., Jackson, P., 2008. Start-node and end-node pruning for efficient segmental-HMM decoding. Electronics Letters 40 (1), 60–61.

Singampalli, V. D., 2006. Statistical models to relate speech gestures to meaning. Mphil transfer report, Univ. of Surrey.

Singampalli, V. D., Jackson, P. J., 2005. Statistical models to relate speech gestures to meaning. In: Proc. One-day meeting on Trajectory models for speech processing. Univ. of Edinburgh.

Singampalli, V. D., Jackson, P. J., 2007a. Coarticulatory relations in a compact model of articulatory dynamics. In: Proc. One-day meeting on Unified Models for Speech Recognition and Synthesis. Univ. of Birmingham.

Singampalli, V. D., Jackson, P. J., 2007b. A statistical technique for identifying articulatory roles in speech production. In: Proc. One-day meeting for Young Speech Researchers.

Singampalli, V. D., Jackson, P. J., 2008. Towards deriving compact and meaningful articulatory representations: an analysis of feature extraction techniques. In: Proc. One day speech meeting for Young speech researchers. Univ. of Surrey, Guildford.

Singampalli, V. D., Jackson, P. J., 2009. Roles in articulation for speech animation. In: One day BMVA symposium. Univ. of Edinburgh.

Singampalli, V. D., Jackson, P. J. B., 2007c. Statistical identification of critical, dependent and redundant articulators. Proc. Interspeech, *Antwerp*, 70–73.

Soquet, A., Saerens, M., Lecuit, V., 1999. Complementary cues for speech recognition. Proc. ICPhS, *San Francisco, CA*, 1645–1648.

Sun, D. X., 1997. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. IEEE International Conference on Acoustics, Speech, and Signal Processing 3, 1751 – 1754.

Tokuda, K., Yoshimura, T., Masuko, T., Kobiyashi, T., Kitamura, T., June 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. ICASSP. Vol. 3. Istanbul, Turkey, pp. 1315–1318.

Tokuda, K., Zen, H., Kitamura, T., 2007. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic vector feature sequences. Comp. Speech & Lang. 21 (1), 153–73.

Toth, A. R., Black, A. W., 2005. Cross-speaker articulatory position data for phonetic feature prediction. Proc. Interspeech, *Lisbon*, 2973–2976.

Uraga, E., Hain, T., 2006. Automatic speech recognition experiments with articulatory data. Proc. Interspeech, *Pittsburgh, PA*, 353–356.

Westbury, J. R., Turner, G., Dembowski, J., 1994. X-ray microbeam speech production database user's handbook. Univ. of Wisconsin, Madison, WI.

Wilson, S. M., Saygun, A. P., Sereno, M. I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7, 701–702.

Wrench, A. A., 2001. A new resource for production modelling in speech technology. Proc. Inst. of Acoust., *Stratford-upon-Avon, UK* 23 (3), 207–217.

Wrench, A. A.and Hardcastle, W., 2000. A multichannel articulatory speech database and its application for Automatic Speech Recognition. In: Proc. 5th seminar on speech production:Models and Data.

Xue, J., Borgstrom, J., Jiang, J., Bernstein, L., Alwan, A., 2006. Acoustically driven talking face synthesis using dynamic bayesian networks. Multimedia and Expo, IEEE International Conference on, 1165–1168.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK book. `http://htk.eng.cam.ac.uk/`.

Zlokarnik, I., 1993. Experiments with articulatory speech recognizer. Proc. Eurospeech '93, *Berlin, Germany*, 2215–2218.