Statistical Estimation of Articulatory Trajectories from the Speech Signal Using Dynamical and Phonological Constraints

by

Sorin Vasile Dusan

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2000

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

In speech science and technology, the acoustic-to-articulatory mapping is known as a difficult problem due to its non-linear and one-to-many characteristics. Over the years, different optimization techniques have been proposed to solve this problem. One of these methods is based on the extended Kalman filtering and smoothing. Although the application of this technique to vowels was promising, its extension to all classes of speech sounds has not been successful. This thesis focuses on developing and improving a statistical method of estimating the articulatory trajectories from the speech signal based on the extended Kalman filtering and smoothing.

In this study, we proposed a new way of constraining the acoustic-to-articulatory inversion by imposing high-level phonological constraints in addition to the dynamical ones. These phonological constraints were imposed by constructing different dynamical models with separate acoustic observation functions for each coproduction unit of speech consisting of two consecutive phones. Each observation sub-function was approximated in small regions by piecewise linear functions using articulatory-acoustic look-up tables. The estimation of the model parameters was based on a direct maximum-likelihood method using training articulatory-acoustic trajectories. An integrated method has been proposed in this study for the recognition of coproduction units and segmentation of the speech signal based on maximum-likelihood of the acoustic observations given different coproduction models. The likelihood of the acoustic observations given every phonological coproduction model was computed using the innovation sequences from the extended Kalman filter. The smoothed articulatory states of the corresponding model with the highest likelihood were used as the best estimate of the articulatory trajectories in every segment. Good estimation results for all classes of speech sounds have been obtained in different experiments using both synthesized and real human data.

Acknowledgment

I am very grateful to Professor Li Deng for his supervision and support during my Ph.D. studies at the University of Waterloo in the wonderful area of speech processing.

I have special thanks to Professor Amir Khandani, Professor Li Deng, Professor Andrew Wong, Professor George Freeman and Professor Paul Fieguth who taught the graduate courses I took at the University of Waterloo.

Also, I have special gratitude to Dr. Shinji Maeda for his permission to use in this research the articulatory and acoustic models developed by him, and to Dr. Jinawu Dang, Dr. Rafael Laboissière and Dr. Arturo Galván for their contributions to these models.

I am grateful to NSERC Canada, Government of Ontario, Institute of Computer Research and University of Waterloo for the scholarships and awards offered to me during my Ph.D. studies and to Nortel Telecom and Department of Electrical and Computer Engineering of University of Waterloo for the research and teaching assistantships provided.

I am also grateful to the University of Wisconsin, for making available to me the X-ray Microbeam Database which was supported by research grant number R 01 DC 00820 from the National Institute of Deafness and Other Communicative Disorders, U. S. National Institute of Health.

Last but not least, I would like to thank my Ph.D. committee members Dr. Alex Acero from Microsoft Research U.S.A., and Professor Amir Khandani, Professor Xuemin Shen, Professor Andrew Wong, Professor Li Deng from University of Waterloo for their suggestions and comments to this thesis.

Contents

List of Tables

List of Illustrations

Chapter 1

Introduction

This introductory chapter presents the subject, purpose and motivation of this study and an overview of the methods used for acoustic-to-articulatory inversion. At the end of this chapter the scope and organization of the thesis are also presented.

1.1 Subject, Purpose and Motivation

The use of articulatory representation of speech has a few attractive advantages over the acoustic representation. A low-dimensional, slow-varying articulatory description of speech is considered by many scientists and researchers as an appropriate representation with potential applications to different areas in speech science and technology. However, the articulatory representation is mostly used as a 'laboratory' description of speech, mainly because of the difficulties in acquiring the vocal-tract profiles and articulators' trajectories from humans during the speech production. Even with all the recent progress in the medical imaging techniques, e.g., magnetic resonance imaging (MRI), the application of such methods are im-

practical or not possible for many of the speech science and technology areas. Thus, the recovering of the articulators' positions and motions from the speech acoustics, commonly called *speech inversion*, brought a new hope and perspective to the problem of obtaining the articulatory representation of speech. Over more than 30 years, various approaches to the estimation of the articulatory parameters and vocal-tract shapes from the speech signal have been proposed. One of the main difficulties of this speech inverse transformation consists in its non-unique characteristics. Although substantial progresses have been achieved in this area of speech processing in all theoretical, experimental and computational domains of research, there is no robust technique known to be successfully applied to all classes of speech sounds and different speakers.

Most of the recent approaches to the speech inversion have shifted from the objective of finding analytical, static solutions to finding dynamically constrained trajectories of the articulators from pre-recorded databases of articulatory and acoustic speech measurements. In this context, an elegant and promising technique of estimating the hidden articulatory trajectories from the speech signal was applied successfully to vowels, for the first time more than 20 years ago, based on the extended Kalman filtering and smoothing. Later, other approaches have attempted to extend this technique to other classes of speech sounds, but the results were not very successful. On the other hand, recent experimental studies have shown that the articulators' positions can be accurately recovered using human articulatory and acoustic speech measurements. Yet, a practical, generalized solution to all different sounds and speakers has not been obtained.

The subject of this thesis is the estimation of the articulatory trajectories and the recognition of some combination of articulatory gestures from the speech acoustic signal. The main purpose of this study consists in developing a generalized

speech inversion method, applicable to all classes of speech sounds and different speakers. This study focused on finding new solutions and improvements capable to overcome the limitations and drawbacks revealed by the previous studies of speech inversion. Thus, this research is more experimental and computational rather than theoretical, although some theoretical aspects have been discussed. The objective of this study was also to evaluate the developed method on real speech data, containing a limited number of combinations of speech sounds from different classes. A long-time goal, following this study, would be the application of the speech inversion method to different areas, like speech coding and recognition, and teaching speech production to hearing and speaking impaired.

This study was motivated by two factors. First, it was motivated by the potential of using the articulatory representations in different areas of speech science and technology. Second, it was motivated by the relative un-success of the previous approaches to generalize the method of estimating the articulatory trajectories from the speech signal to all classes of sounds. Relative to the second factor, this research was motivated by the divergent and somehow controversial outcomes of three different studies concerning the estimation of articulatory trajectories using Kalman filtering, which we found in literature (Shirai and Honda [89]; Wilhelms *et al.* [105]; Ramsay and Deng [75]). While the first study, [89], showed relatively accurate results in estimating articulatory trajectories for vowels, the second study, [105], did not succeed in obtaining the same accuracy for both vowels and consonants. Thus, for some consonants, like plosives and nasals, the estimated articulatory trajectories were not accurate, showing instabilities of the method in the intervals close to the vocal-tract constrictions. Moreover, the third study, [75], approached the estimation of articulatory trajectories using Kalman filtering as an internal process of an automatic speech recognition system, without explicitly addressing the drawbacks

revealed by the second study, related to consonantal sounds. If the generalization of Kalman filtering estimation approach to all classes of speech sounds were successful, we would wonder why this method has not been yet successfully applied into different areas in speech processing, like automatic speech recognition, speech coding, teaching the deaf to speak, etc. But, as many researchers stated, we believe that the success of the application of the speech inversion method to those areas depends fundamentally on the accuracy of the method of estimating the articulatory trajectories. In other words, efforts should be made first to develop accurate speech inversion methods, before these methods could be successfully applied to different areas of speech. We quote from a paper presented at the 1994 Meeting of the Acoustical Society of America:

> In order to perform automatic speech recognition based on the movements of the articulators, there must be a reliable mechanism for estimating these articulatory positions directly from speech. Section IV provides several reasons why this is a formidable problem. None of the articulatory models that have been applied as speech synthesizers or speech mimics have been successfully applied to ASR, and it is unlikely that they will be until better techniques are found for acoustic to articulatory inversion. (R. C. Rose, J. Schroeter and M. M. Sondhi 1996, [80])

We are in agreement with the above statement of those authors and we hope that the contributions of this research will encourage and support other studies in achieving the ultimate goal of applying the recovered articulatory information to different areas of speech science and technology.

1.2 Overview of Acoustic-to-Articulatory Inversion and Recognition of Articulatory Gestures

Speech represents the most common and natural way of communication for people and is a part of language like writing and sign language. As a communication process, speech is a way of transmitting information or thoughts from one person to other persons. This information, structured in words, is coded into acoustic signals by the speaker vocal system and decoded from acoustic signals into words by listeners' ears and brains. The natural speech is always produced by an articulatory and vocal-tract system. It is not clear whether or not, in recognizing the speech sounds, people recover in their minds some temporal information about the state of the articulatory system of the speaker who produced it. As the linguistic-acoustic relationship does not necessarily rely on the modeling of dynamical articulatory system, the identity of a sound is not directly associated to the state of the articulators which produced it, but that state might theoretically be inferred from the acoustic signal.

One benefit of analyzing the speech process in a domain closer to the source, in the transmitting chain, is the lower redundancy of articulatory signals compared to that of final coded acoustic signals. At a rate of about 10 phonemes per second, and an information of about 5 bits/phoneme the speech has an average information of about 50 bits/sec, which is about three orders of magnitude less than the average information transmitted through a communication channel by sampling the speech signal 8000 times per second and coding each sample with 8 bits, resulting in 64 Kbits/sec. It is certain that not all the information transmitted this way is useful and most of it is redundant.

A second benefit of this would be the reduced variability of speech in the ar-

ticulatory domain. One of the most difficult problems in processing and analyzing the speech is the great variability of the speech signal. This huge variability encountered in the speech signal has different origins: differences among speakers' anatomical and physiological vocal-tract structures including those due to gender and age, differences among speakers' ways of speaking and coordinating the numerous articulators' gestures, differences in ways of speaking of one speaker due to physiological, psychological and anatomical variations, differences in speech sounds of one speaker due to coarticulation, articulatory compensation and prosodic variations and differences in the environment or speech transmission channels. Based on the assumption that speech is produced by an articulatory and vocal-tract system, a possible way of reducing this large variability is to analyze the speech not in the acoustic domain but in the articulatory domain. This variability reduction is based on the fact that for a language, the speakers produce a specific speech sound using less variable articulatory gestures in a vocal-tract geometric task domain, and the remaining variability is mainly due to coarticulation, articulatory compensation and prosody.

A third benefit of analyzing the speech in the articulatory domain could arise from analyzing slow movements of the articulators which produced the speech comparing to the quick changes in the acoustic speech signal.

The recovery of articulatory state and motion from speech signal, the so called *speech inversion problem*, could have both theoretical and practical applications. It could help the motor theory of speech perception (Liberman *et al.*, [48]; Liberman and Mattingly, [49]), the articulatory phonology (Browman and Goldstein, [4], [6]) and have applications in speech recognition (Zlokarnik, [107]; Ramsay and Deng, [75]), speaker recognition, speech synthesis (Wilhelms *et al.*, [105]), speech coding (Gupta and Schroeter, [33]; Schroeter and Shondi, [85]) and teaching deaf people

to speak (Zahorian and Venkat, [106]).

An analytical formulation of this inverse transformation based on the solutions of wave equation and boundary conditions is very laborious and in fact this function is represented by a chain of transformations, most of them nonlinear. First, the articulatory parameters have to be transformed into vocal-tract acoustic tubes described by area and length functions. Then using the boundary conditions the acoustic wave equation has to be solved in order to obtain the transfer function of the vocal-tract. The convolution of the impulse response of this transfer function with an excitation signal finally gives the speech signal which is further applied to some feature extraction transformation in order to be represented by acoustic feature parameters like formants, FFT, LPC or cepstrum parameters. A practical approach consists of using a mapping between acoustic and articulatory domains instead of an analytical function. This mapping can be done between pairs of corresponding points in the two domains, acoustic and articulatory. A codebook of many pairs of vectors in the acoustic and articulatory domains can be used for the implementation of such mapping. These codebooks can be obtained using articulatory and vocal-tract acoustic models or direct measurement of simultaneous acoustic and articulatory speech data. The static solution for the acoustic-to-articulatory inverse mapping problem suffers of non-uniqueness because of the one-to-many characteristics of this nonlinear inverse transformation. A dynamically constrained approach, based on articulatory or acoustic dynamic modeling, could help searching for a unique solution but it still represents a difficult nonlinear optimization problem. The characteristics of the application for which the inverse mapping solution is being sought can determine the accuracy and the limits of the estimated articulatory parameters and motion. For some applications like speech coding for example, it is not so important the accuracy of the estimated articulatory trajectories but the

accuracy of the re-synthesized acoustic signal. For other applications like automatic speech recognition or teaching deaf people to speak, the accuracy of the estimated articulatory positions and trajectory might be crucial.

After more than thirty years of various attempts, the speech researchers still regard the acoustic-to-articulatory inversion as an open and very challenging problem. Moreover, there is no general method known for recovering the vocal tract shapes from the speech signal for all classes of speech sounds and robust enough to be applied to solve and help a practical problem.

There were many attempts of estimating the vocal tract shapes from the formant frequencies of the speech signal, but these parameters are not representative for all classes of speech sounds, hence these methods cannot be generalized for the speech inversion problem. The non-unique solution of the acoustic-to-articulatory mapping motivated researchers to employ nonlinear optimization methods and find optimal articulatory trajectories and tract shapes by using dynamic constraints on more than one frame of speech.

The multiple-to-one nature of the forward transformation, from articulators to acoustics, has been proven by modeling speech production using multiple acoustic tubes (Flanagan, [26]) and by some bite-block experiments of articulatory compensation, in which a subject bites down on a small block and tries to produce a natural speech sound with an unnatural position of the mandible (Lindblom *et al.*, [50]). Another example of multiple-to-one characteristics of the transformation is the speech produced by ventriloquists, who are able to produce some speech sounds with completely different articulator positions from those of normal speech. It appears that because of the multiple-to-one nature of the forward transformation found in articulatory compensation experiments, speech production modeling and ventriloquist' way of speaking, it should not be possible to recover accurately the

articulator positions from speech acoustics. However it is not fully understood to what extent this articulatory compensation phenomenon is used during naturally produced speech. On the other hand, the one-to-many nature of the inverse mapping based on vocal-tract acoustic modeling is very sensitive to the assumptions underlying the models, like the nature of losses. In the articulatory compensation experiments there are still perceptual differences between sounds produced with bite-blocks and those normally produced (Flege *et al.*, [27]).

Among the first researchers who approached the speech inversion problem were Mermelstein and Schroeder who proposed a method of estimating a smoothed area function from formant frequencies in 1965, [61]. Each of them extended this first study later. Schroeder proposed the use of measured spectra and acoustic impedance measurements at the lips in order to constrain the area function estimated (Schroeder, [58]). Using Fourier series of the logarithmic area function he obtained unique solutions for the area function for the assumption of small perturbations. Also he proved that there is a unique relation between the impedance function at the lips and the resonance frequencies of the vocal-tract. Mermelstein determined by computer simulation the area function using the first six admittance poles and zeroes for some Russian vowels for which X-ray data were published, [59]. Gopinath and Sondhi reviewed the theory of determination of vocal-tract shape from acoustic measurements and input impedance, [29]. A new method of estimating the area function of the vocal-tract employing the inverse filtering of the acoustic speech waveforms has been suggested by Wakita, [103]. He proved a direct relation between the inverse filter model and the acoustic tube model of speech.

Shirai and Honda studied the estimation of articulatory motion using an articulatory dynamical model and nonlinear filtering, [89]. They modeled the nonlinear observation function relating the formant frequencies to articulatory parameters by

third-order multi-variable polynomials and used the Kalman filtering technique to estimate unique articulatory trajectories from formant frequencies.

A numerical approach for studying the relationship between the vocal-tract shape and its corresponding acoustic output has been done by Atal *et al.*, [1], using a computer sorting technique. They constructed a codebook of 30720 articulatory and acoustic pairs of vectors by sampling the whole space of an articulatory model and storing the articulatory and acoustic data. They called all the vectors which mapped into the same acoustic vector, an articulatory fiber.

A study of generating vocal-tract shapes from formant frequencies has been published by Ladefoged *et al.*, [46]. In that research the authors used a factor analysis method, called PARAFAC, to represent the tongue shape by two components: a front raising and a back raising component. The weights of these two components and of a third one representing the distance between the lips were determined using a stepwise multiple regression technique from three formant frequencies of 50 vowels. The recovered vocal-tract shapes were compared with X-ray diagrams of speakers of British English, American English and Russian.

An extension of Kalman filtering technique for estimation of articulatory trajectories for vowels and consonants has been done by Wilhelms *et al.*, [105]. They used as acoustic features large vectors of short time spectra computed from estimated ARMA coefficients for vowels and some consonants. The Jacobian matrix of the observation function has been computed using numerical approximation. This study revealed difficulties in estimating the articulatory parameters for consonants.

Schroeter *et al.*, [84] proposed a method of estimating the articulatory parameters using a vocal tract/cord model and an articulatory-acoustic codebook. In this study they have used the LPC parameters as acoustic vectors and sampled the

articulatory space between pairs of root shapes. They extended later their work to a multi-frame approach. In 1989 Schroeter and Sondhi, [86] presented a method based on dynamic programming to search the articulatory codebooks. They have used the LPC derived cepstral coefficients as acoustic feature and introduced a lifter in computation of the acoustic distance and a dynamic cost in making a transition from a vocal tract shape to another one.

Methods of speech inversion based on neural networks have been published by Shirai and Kobayashi, [93], Shirai, [88] and Papcun *et al.*, [68]. In the first two papers the authors used an articulatory model and trained neural networks to approximate the nonlinear relationship between articulatory parameters of the model and the cepstrum coefficients as acoustic parameters. The third paper presents a method of training the acoustic-to-articulatory network on X-ray microbeam data.

A study of inverse mapping in speech based on an articulatory model for robot speech synthesis has been done by Laboissière, [44]. For this task, the target articulatory parameters of a robot containing an articulatory speech synthesizer have to be estimated from a real speech in order to teach the robot to produce speech.

An optimization method based on conditional minimum of work has been used by Sorokin, [95] and Sorokin and Trushkin, [97], for determination of vocal-tract shape for vowels from formant frequencies and by Sorokin, [96] for fricatives.

A study for recovering the vocal-tract area function for vowels and fricative consonants has been published by Beautemps *et al.*, [2]. They proposed an extension of the $\alpha\beta$ model for computing area function along the vocal-tract and used the formant frequencies as acoustic parameters for inversion.

Ramsay and Deng, [75], proposed a stochastic target model for articulatory speech recognition, and the estimation of articulatory state was based on extended

Kalman filtering technique. The novelty of their approach consist in modeling the articulatory state by a Markov chain, proposing an articulatory target model and using the Estimate-Maximize (EM) algorithm for estimating model parameters. The whole nonlinear observation function has been approximated by a codebook containing acoustic and articulatory parameters and the Jacobian matrices for each region of linearization. This codebook has been structured as a binary decision tree.

Another work based on dynamic programming search of an articulatory codebook search has been presented by Richards *et al.*, [78]. They attempted to estimate the articulatory representation of speech using the cepstral coefficients and a large codebook containing 160000 entries derived using the Distinctive Regions Model.

A study of inverse mapping problem using real human articulatory and acoustic data has proven that, for some classes of speech sounds, the articulator positions can be accurately recovered (Hogden *et al.*, [34]). In this study, the authors used a look-up table method for mapping from the acoustic space represented by smoothed spectra to the articulatory space represented by X and Y coordinates of some receiver coils placed on articulators. They used electromagnetic midsagittal articulography to record the data. Another study tried to recover the articulatory dynamics from speech acoustics using a genetic algorithm and the information contained in the formant frequencies (McGowan and Lee [56]).

Computational models for speech production have been proposed by Saltzman and Munhall [82], Kaburagi and Honda [38]. These studies modeled the human articulatory system and used additional constraints to determine a unique trajectory of the articulators.

Recent results in speech inversion based on human data recorded using the electromagnetic midsagittal articulography have been published by Suzuki *et al.*, [99].

They constructed a large codebook of 222894 articulatory-acoustic pair data and used dynamic constraints in the search of this codebook. They applied this method to continuous speech utterances containing vowels and consonants in Japanese. Their best results of inversion were based on a segment interval of the search of 160ms. Root mean squared (RMS) errors of about 2 mm have been obtained in estimating articulatory trajectories with this searching method.

In an early stage of our study we experimented, [20], an extension of the Kalman filtering approach of Shirai and Honda, using formant frequencies as acoustic parameters and the linearization of the observation function on small regions based on a codebook. Another extension in our early work was the estimation of vocal-tract shapes from more general acoustic features using Kalman filtering, [21]. We used the mel-frequency cepstrum coefficients, and created a large codebook of 235000 pairs of articulatory and acoustic vectors using Maeda's articulatory model, [55]. The linearization of the observation functions has been done on 10000 small regions using a clustering method. The estimation of model parameters has been applied using the EM algorithm like in [75]. In this study the inversion method has been applied for vowels only.

An experimental study of recognizing articulatory gestures has been published by Papcun *et al.*, [68]. These authors used a neural network trained on x-ray microbeam data to estimate the articulatory trajectories from the speech acoustics. They constructed articulatory templates for the recognition of release gestures of three articulators — lower lip, tongue tip and tongue dorsum — in the production of 6 English stop consonants: /p/, /b/, /t/, /d/, /k/ and /g/. For a small corpus of 90 gestures from three different speakers, the gestures were recognized correctly from 94.4% to 98.9%.

During the last three decades researchers have tried different approaches to the

complicated and challenging problem of acoustic-to-articulatory inversion. Both analytical and computational frameworks of these approaches have some advantages and disadvantages, depending on the application and generalization of methods. Despite the large variety of all these methods used to solve the speech inverse problem, some of them just mentioned above, there is no practical, generalized method known for different speakers and all classes of speech sounds.

The state-of-the-art in this area of speech inversion is probably still in its early stage. The very recent attempts of providing general methods of inversion for all classes of speech sounds, like in (Suzuki *et al.*, [99]), it is worth to be mentioned here. This study not only introduces a generalized method based on human data but also provides a qualitative and quantitative error evaluation of estimated articulatory parameters. Although their direct approach is based on exhaustive search of an articulatory-acoustic database acquired using electromagnetic articulographic measurements, it has been successfully applied to all classes of speech sounds for a single speaker.

In this context, the study of inverse mapping in speech presented in this thesis is trying to provide a generalized method of speech inversion to all classes of speech sounds of a language. In this thesis we describe our study of recovering vocal-tract shape and its dynamics based on a new approach of Kalman filtering by applying new phonological constraints. This method has proven to be a general and robust approach for speech inversion for all classes of speech sounds and can be applied using either articulatory models or human articulatory-acoustic measurements.

1.3 Scope and Organization of the Document

Chapter 2 describes a background of speech production modeling by giving examples of vocal-tract acoustic models, articulatory models and of modeling the articulatory dynamics. The background of Kalman filtering state estimation method is not included in this chapter, but in the main chapter dedicated to the acoustic-to-articulatory inversion.

In Chapter 3 we present some experiments of articulatory and acoustic vowel classification which support the usefulness of articulatory analysis of speech and the potential application of the speech inversion method to speech recognition. Although the topic of this chapter is not directly related to the speech inversion problem, we included this chapter in this thesis as a preliminary work on articulatory analysis of speech.

Chapter 4 represents the main part of this thesis and presents the general method of inverting the articulatory-to-acoustic transformation. In this chapter, we describe the coproduction segments and models, the modeling of the articulatory-acoustic function by using codebooks, the model parameter estimation using a direct maximum-likelihood method, the estimation of articulatory trajectories based on extended Kalman filtering and smoothing and the new way of applying phonological constraints to the speech inversion as an integrated approach of recognizing the coproduction models and estimating articulatory trajectories.

Chapter 5 presents experimental results for acoustic-to-articulatory inversion based on three sets of experiments, using synthesized and real speech data. These experimental results are based on speech segments containing a number of different classes of speech sounds. Different coproduction segments have been used for both synthesized and real speech data. The first set of experiments is based on

synthesized speech with an articulatory-acoustic model. The second set of experiments is based on real speech data acquired with an electromagnetic midsagittal articulograph. The third set of experiments presents examples based on real speech acquired with an X-ray microbeam system.

In Chapter 6 we present two potential applications of the speech inversion method. First, a new method of displaying the dynamics of the vocal-tract over time is presented as an application for general speech research and as an aid in teaching the speaking and hearing impaired to speak or teaching foreign languages. Second, applications of the speech inversion method to automatic speech recognition are suggested.

Finally, in Chapter 7 we conclude this dissertation by presenting a summary of this thesis and the contributions of this research. At the end of this chapter an outline of the future work is also presented.

Chapter 2

A Background of Speech Production

This chapter presents a background of speech production and examples of vocal-tract acoustic models, articulatory models and models of articulatory dynamics. These examples contain solutions of the wave equation in the vocal-tract, a statistic articulatory model and dynamic modeling approaches of articulatory system.

2.1 Vocal-Tract Acoustic Models

The theories of speech production have been developed by scientists and researchers over the years. A first step in understanding the production of speech sounds in the vocal-tract was the theory of the wave propagation proposed by Chiba and Kajiyama, [12]. According to this theory the speech sound is produced by a planar acoustic wave which propagates between glottis and lips and nostrils in the vocal and respectively nasal tract. The acoustic theory of speech production was

published by Fant in 1960, [25]. He proposed a source-filter theory of speech production according to which the speech sound is produced by a source, e.g., glottis in the case of vowels, and filtered by the vocal-tract. Another step in understanding the speech production was the application of the perturbation theory to study the acoustic effect of small variations in the area function of the vocal-tract (Schroeder, [58]; Mermelstein, [60]). The quantal theory of speech was proposed by Stevens, [98]. According to this theory large variations in the acoustic domain can be produced by small variations in the articulatory domain and large variation in the articulatory domain can have little effect in the acoustic domain. A more detailed overview of these theories can be found in [9]. Further developments in the field of speech production were carried out. Two reference books of speech processing have been published by Flanagan, [26] and Rabiner, [74].

In this section, an example of providing solutions for the wave propagation equations in time domain is reviewed.

The speech production system contains the vocal tract which begins at glottis and ends at the lips, the nasal tract which begins at the velum and ends at the nostrils and the source of excitation which is in the glottis for voiced sounds or somewhere between glottis and lips for unvoiced sounds. The vocal tract consists of the pharyngeal cavity and oral cavity and has for an average male the total length of about 17.5 cm. The nasal tract has at at the beginning a common part which continues further with two separate, parallel, nasal tracts which end at the nostrils.

The wave propagation in the vocal-tract is based on some laws of physics. These laws which describe the generation and propagation of sound in the vocal system are: the fundamental laws of conservation of mass, conservation of momentum, conservation of energy and the laws of thermodynamics and fluid mechanics. The

complete acoustic theory of speech production must consider the following effects: 1) excitation of sound in the vocal tract 2) variation in time of the shape of vocal tract 3) nasal coupling 4) radiation of sound at the lips and nostrils 5) losses due to viscous friction and heat conduction 6) softness and vibration of the vocal and nasal tract walls.

Using the laws governing the generation and propagation of sound in the vocal system, a set of partial differential equations can be derived. Taking into account all the effects which appear in the vocal system the formulation and solutions of this set of differential equations is very difficult, therefore some simple assumptions have to be taken. For a simple configuration of the vocal system, the vocal tract is modeled as a nonuniform tube.

Portnoff, [73], derived the following pair of equations for an acoustic tube approximating the vocal-tract

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t}, \tag{2.1}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (p/A)}{\partial t} + \frac{\partial A}{\partial t}, \tag{2.2}$$

where

$p = p(x,t)$	is the sound pressure in the tube
$u = u(x,t)$	is the volume velocity in the tube
$A = A(x,t)$	is the area function of the tube (cross-sectional area)
ρ	is the density of air in the tube
c	is the velocity of sound
x	is the distance from glottis
t	is the time

The solutions for these equations are very complicated, except for some simplifications and approximations of the model.

A time-domain simulation of the vocal-tract system, which we used in this study, has been developed by Maeda in 1982, [54]. This approach is based on the acoustic transmission line of the vocal-tract and a direct solving in time-domain the equations governing the sound generation and propagation in the vocal-tract. In this simulation the vocal system consists of the pharyngeal, nasal and oral cavities. The glottis is connected to a lung pressure source. The planar propagation assumption of the acoustic waves in the vocal-tract is considered. Three differential equations were used to describe the evolution of pressure $p(x,t)$ and volume velocity $u(x,t)$ inside an acoustic tube with non-rigid walls. These were the equation of motion (EQM), equation of continuity (EQC) and equation of wall vibration (EQW)

$$\frac{\partial p}{\partial x} + \frac{\partial}{\partial t}\frac{\rho_0 u}{A_0} + \frac{ru}{A_0} = 0, \tag{2.3}$$

$$\frac{\partial u}{\partial x} + \frac{\partial}{\partial t}\frac{A_0 p}{\rho_0 c^2} + \frac{\partial A_0}{\partial t} + \frac{\partial S_0 y}{\partial t} = 0, \tag{2.4}$$

$$m\frac{\partial^2 y}{\partial t^2} + b\frac{\partial y}{\partial t} + ky = S_0 p. \tag{2.5}$$

where ρ_0 is the density of air at equilibrium, c is the sound velocity, r is the flow resistance, y is the amplitude of the yielding of walls due to the pressure inside the tube, A_0 and S_0 are the given area and perimeter, and m, b, k represent the mass, mechanical resistance and stiffness of the wall per unit length. The equation describing the evolution of area due to wall yielding is

$$A(x,t) = A_0(x,t) + y(x,t)S_0(x,t). \tag{2.6}$$

In this approach the losses due to heat conduction and viscous friction at the walls have not been taken into account, since they only produce a small increase in the bandwidth of the formants. The boundary condition at the glottis is

$$P_{sub}(t) = p(x_0,t), \tag{2.7}$$

where $P_{sub}(t)$ represents the sub-glottal pressure value of the source. The conditions at the nasal coupling point x_k, are defined by the equations

$$u(x_k^-, t) = u(x_k^+, t) + u'(0, t), \tag{2.8}$$

$$p(x_k^-, t) = p(x_k^+, t) = p'(0, t), \tag{2.9}$$

where '$-$', '$+$' and '$'$' indicate respectively the pharyngeal cavity end, the beginning of the oral cavity and the nasal tract. The boundary conditions at the mouth opening and nostrils are approximated by a radiation load in a form of a parallel circuit consisting of a conductance and a susceptance, both independent of frequency.

Numeric solutions of the above continuous-time equations for $p(x, t)$ and $u(x, t)$ can be obtained by discretizing in both time and space domains. The discretization of $u(x, t)$ in the space domain, along the vocal-tract, can be done at some points x_j using variable sampling intervals X_i

$$x_j = \sum_{i=0}^{j} X_i, \tag{2.10}$$

where $j = 0, 1, 2, ..., M$. The pressure $p(x, t)$ will be sampled at the middle point between x_{j-1} and x_j. The discretization in time of the $p(x, t)$ and $u(x, t)$ variables can be done at sampled points $t = nT$, for $n = 0, 1, 2, ...$, where T represents a fixed sampling interval. In Figure 2.1 the sampling in the two dimensional space is represented. For the discretization in space, the integrations can be approximated by the 'midpoint' rule and by the 'rectangular' rule. After these approximations, the discrete space equations of motion (EQM), continuity (EQC) and wall vibration (EQW) become

$$P_{j-1} - P_j = \frac{d}{dt}\frac{\rho_0 X_{j-1}}{2A_{j-1}}U_j + \frac{X_{j-1}r_{j-1}}{2A_{j-1}}U_j + \frac{d}{dt}\frac{\rho_0 X_j}{2A_j}U_j + \frac{X_j r_j}{2A_j}U_j, \tag{2.11}$$

$$U_j - U_{j+1} = \frac{d}{dt}\frac{X_j A_j}{\rho_0 c^2}P_j + + \frac{d}{dt}X_j A_j + \frac{d}{dt}X_j S_j y_j, \tag{2.12}$$

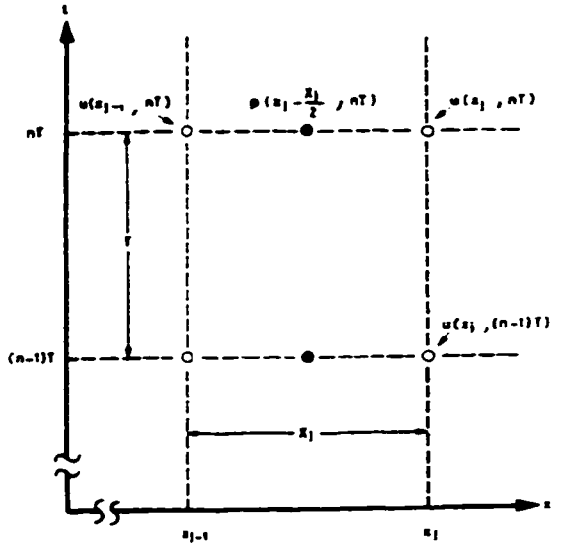

Figure 2.1: Sampling of the pressure and volume velocity in the two dimensional space (from Maeda, [54])

$$S_j P_j = m\frac{d^2}{dt^2}y_j + b\frac{d}{dt}y_j + ky_j. \tag{2.13}$$

Eliminating y_j from EQC and EQW, they can be written as

$$U_j - U_{j+1} = u_1 + u_2 + u_3, \tag{2.14}$$

$$P_j = \frac{d}{dt}\frac{m}{X_j S_j^2}u_3 + \frac{b}{X_j S_j^2}u_3 + \int_0^t \frac{k}{X_j S_j^2}u_3 dt, \tag{2.15}$$

where

$$u_1 = \frac{d}{dt}\frac{X_j A_j}{\rho_0 c^2}P_j, \tag{2.16}$$

$$u_2 = \frac{d}{dt}X_j A_j, \tag{2.17}$$

$$u_3 = \frac{d}{dt}X_j S_j y_j. \tag{2.18}$$

From these equations the analogy of the acoustic transmission line with the electric transmission line can be observed. The velocity-pressure variables play the role of

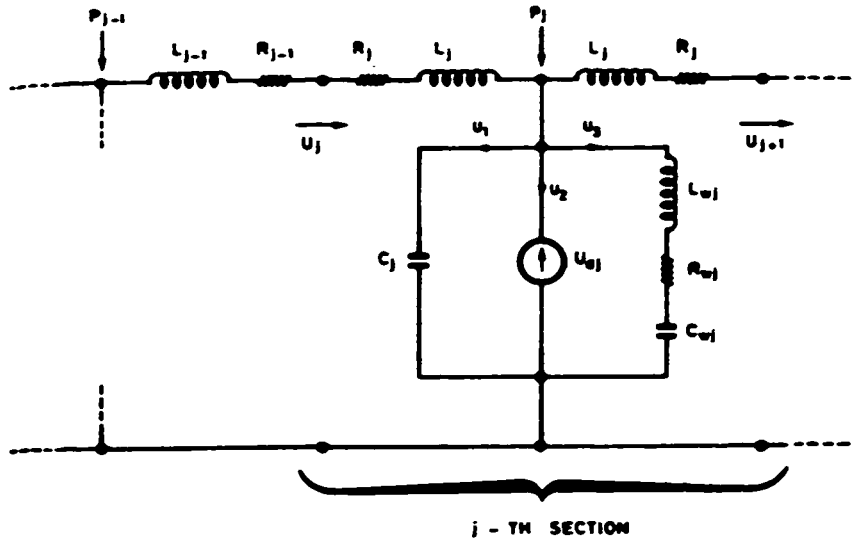

Figure 2.2: The electric equivalent circuit of a rectangular section of the vocal-tract tube (from Maeda, [54])

the current-voltage variables from the electric counterpart. The equivalent electric circuit corresponding to one section of the acoustic tube approximating the vocal-tract is presented in Figure 2.2. The circuit elements can be identified as

$$L_j = \rho_0 X_j / 2A_j, \tag{2.19}$$

$$R_j = 4\pi\mu X_j / A_j, \tag{2.20}$$

$$C_j = X_j A_j / (\rho_0 c^2), \tag{2.21}$$

$$Ud_j = -\frac{d}{dt}(X_j A_j), \tag{2.22}$$

$$Lw_j = m/(X_j S_j^2), \tag{2.23}$$

$$Rw_j = b/(X_j S_j^2), \tag{2.24}$$

$$Cw_j = (X_j S_j^2)/k. \tag{2.25}$$

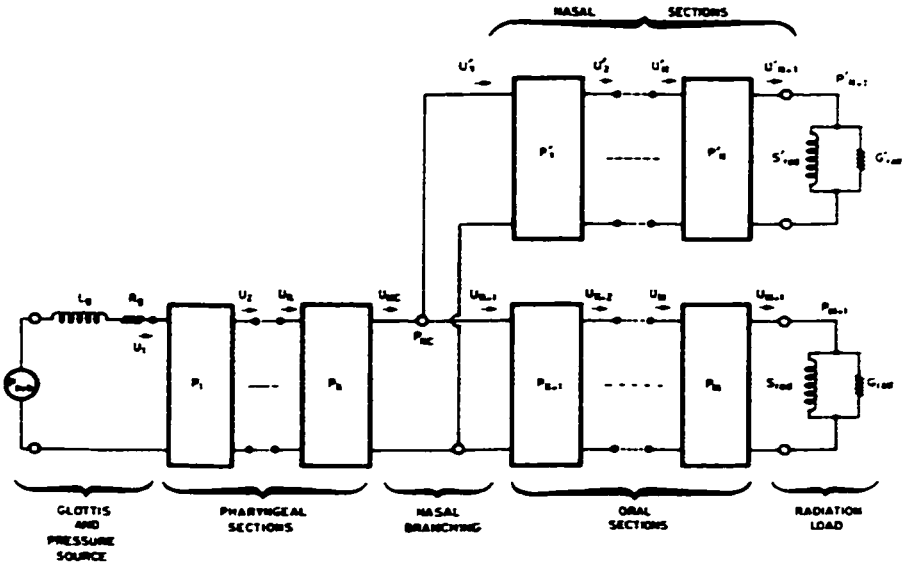

Figure 2.3: The transmission line representation of the whole vocal-tract (from Maeda, [54])

Using these notations, EQM and EQC can be rewritten

$$P_{j-1} - P_j = \frac{d}{dt}(L_{j-1} + L_j)U_j + (R_{j-1} + R_j)U_j, \tag{2.26}$$

$$U_j - U_{j+1} = u_1 + u_2 + u_3, \tag{2.27}$$

where

$$u_1 = \frac{d}{dt}C_j P_j, \tag{2.28}$$

$$u_2 = -Ud_j, \tag{2.29}$$

$$P_j = \frac{d}{dt}Lw_j u_3 + Rw_j u_3 + \int_0^t \frac{u_3}{Cw_j}dt, \tag{2.30}$$

The transmission-line representation of the vocal-tract is presented in Figure 2.3. From this picture, the boundary conditions at the nasal branch point which occurs between the section K and the section $K+1$ are

$$P_K - P_{NC} = \frac{d}{dt}L_K U_{NC} + R_K U_{NC}, \tag{2.31}$$

$$P_{NC} - P_{K+1} = \frac{d}{dt} L_{K+1} U_{K+1} + R_{K+1} U_{K+1}. \tag{2.32}$$

$$P_{NC} - P_1' = \frac{d}{dt} L_1' U_1' + R_1' U_1'. \tag{2.33}$$

$$U_{NC} = U_{K+1} + U_1'. \tag{2.34}$$

The discretization in time can be obtained by applying the 'trapezoid' rule for computing an integral as

$$\int_{(n-1)T}^{nT} y(t)dt = T[y(n) + y(n-1)]/2. \tag{2.35}$$

For the case $y(t) = dx/dt$, this formula has the form

$$[x(n) - x(n-1)]/T = [y(n) + y(n-1)]/2. \tag{2.36}$$

which is called *central difference with averaging*. In the discrete space equations there are three kinds of terms

$$y_1(t) = c_1(t)x(t). \tag{2.37}$$

$$y_2(t) = \frac{d}{dt} c_2(t)x(t). \tag{2.38}$$

$$y_3(t) = \int_0^t c_3(t)x(t)dt. \tag{2.39}$$

where $c_i(t)$ are coefficients and $x(t)$ is $P(t)$ or $U(t)$. The first kind of terms can be digitized at time $t = nT$ directly as

$$y_1(n) = c_1(n)x(n). \tag{2.40}$$

The second term can be approximated using a recursive formula

$$y_2(n) = (2/T)c_2(n)x(n) - Q(n-1). \tag{2.41}$$

where the recursion formula is

$$Q(n-1) = (1/T)c_2(n-1)x(n-1) - Q(n-2). \tag{2.42}$$

Similarly, the third term can be approximated by the equation

$$y_3(n) = (T/2)c_3(n)x(n) + V(n-1). \tag{2.43}$$

where the recursion formula is

$$V(n-1) = Tc_3(n-1)x(n-1) + V(n-2). \tag{2.44}$$

Applying this rules, the discrete space equations described above can be transformed into linear algebraic equations in which $U_j(n)$, $P_j(n)$, $U_{NC}(n)$ and $P_{NC}(n)$ are their solutions. These algebraic equations can be solved recursively in time starting from the initial conditions of the vocal-tract for which $Q(0) = 0$ and $V(0) = 0$. Three sets of equations are derived by eliminating the $P_j(n)$

$$\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_K \\ F_{NC} \end{bmatrix} = \begin{bmatrix} H_1 & b_1 & & & & \\ b_1 & H_2 & b_2 & & & \\ & & \ddots & & & \\ & & & b_{K-1} & H_K & b_K \\ & & & & b_K & H_{NC-1} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_K \\ U_{NC} \\ P_{NC} \end{bmatrix}. \tag{2.45}$$

for the pharyngeal tract,

$$\begin{bmatrix} F_{K+1} \\ F_{K+2} \\ \vdots \\ F_M \\ F_{M+1} \end{bmatrix} = \begin{bmatrix} 1 & H_{K+1} & b_{K+1} & & & & \\ & b_{K+1} & H_{K+2} & b_{K+2} & & & \\ & & & \ddots & & & \\ & & & & b_{M-1} & H_M & b_M \\ & & & & & b_M & H_{M+1} \end{bmatrix} \begin{bmatrix} P_{NC} \\ U_{K+1} \\ U_{K+2} \\ \vdots \\ U_M \\ U_{M+1} \end{bmatrix}. \tag{2.46}$$

for the oral tract and

$$\begin{bmatrix} F_1' \\ F_2' \\ \vdots \\ F_N' \\ F_{N+1}' \end{bmatrix} = \begin{bmatrix} 1 & H_1' & b_1' & & & & \\ & b_1' & H_2' & b_2' & & & \\ & & & \ddots & & & \\ & & & & b_{N-1}' & H_N' & b_N' \\ & & & & & b_N' & H_{N+1}' \end{bmatrix} \begin{bmatrix} P_{NC} \\ U_1' \\ U_2' \\ \vdots \\ U_N' \\ U_{N+1}' \end{bmatrix}. \tag{2.47}$$

for the nasal tract. The elements in these three equations are given by

$$b_j(n) = 1/[2C_j(n)/T + Yw_j(n)], \tag{2.48}$$

$$H_j(n) = -2[L_{j-1}(n) + L_j(n)]/T - R_{j-1}(n) - R_j(n) - b_{j-1}(n) - b_j(n). \tag{2.49}$$

and

$$F_j(n) = b_{j-1}(n)[Ud_{j-1}(n) - V_{j-1}(n-1)] - b_j(n)[Ud_j(n) - V_j(n-1)] - Q_j(n-1), \tag{2.50}$$

where

$$Yw_j(n) = 1/[2Lw_j(n)/T - Rw_j(n) + T/2Cw_j(n)], \tag{2.51}$$

$$Ud_j(n) = [A_j(n)X_j(n) - A_j(n-1)X_j(n-1)]/T, \tag{2.52}$$

$$Q_j(n-1) = (4/T)[L_{j-1}(n-1) + L_j(n-1)]U_j(n-1) - Q_j(n-2). \tag{2.53}$$

$$V_j(n-1) = Vc_j(n-1) - Yw_j(n)[Qwl_j(n-1) - Qwc_j(n-1)], \tag{2.54}$$

$$Vc_j(n-1) = (4/T)C_j(n-1)P_j(n-1) - Vc_j(n-2), \tag{2.55}$$

$$Qwl_j(n-1) = (4/T)Lw_j(n-1)u_3(n-1) - Qwl_j(n-2). \tag{2.56}$$

$$Qwc_j(n-1) = [T/Cw_j(n-1)]u_3(n-1) + Qwc_j(n-2). \tag{2.57}$$

and

$$u_3(n) = Yw_j(n)[P_j(n) + Qwl_j(n-1) - Qwc_j(n-1)]. \quad (2.58)$$

The three matrix equations can be solved using an elimination-substitution procedure. A detailed derivation of these transmission line equations and a computer program for solving them has been provided by Maeda, [54]. After solving these equations and the values of $U_{NC}(n)$, $U_j(n)$ and $U'_j(n)$ are obtained for all j, the pressure $P_j(n)$ and $P'_j(n)$ are computed using

$$P_j(n) = b_j(n)[U_j(n) - U_{j-1}(n)Ud_j(n) + V_j(n-1)]. \quad (2.59)$$

The simulated speech signal can be obtained from the sampled values of pressure $P_M(n)$ and $P_N(n)$. Digital simulation of speech utterances using this method and program gives a high quality speech signal. We studied the acoustic-to-articulatory inversion using two kinds of data: simulated data and human direct measured acoustic-articulatory data. For the experiments based on synthesized data we have used this method and program with the permission of the author, Dr. Maeda.

2.2 Static Articulatory Models

In this section, a statistic articulatory model originally developed by Maeda, [53], [55], which was extensively used in this study with the permission of the author, is presented. This is a static articulatory model, that is, it transforms a set of articulatory parameters into a vocal-tract shape and the corresponding area function.

Over the years, various articulatory models and articulatory speech synthesizers have been developed (Coker and Fujimura, [13]; Mermelstein, [60]; Maeda, [53], [55]; Rubin *et al.*, [81]; Meyer *et al.*, [62]; Shondi and Schroeter, [94]; Kohler and

Lacroix, [41]). These models provided vocal-tract shapes by specifying a number of articulatory parameters which controlled the position of the lips, jaw, tongue tip, tongue body, tongue dorsum, larynx, velum and hyoid bone. From vectors of articulatory parameters static vocal-tract shapes and corresponding area functions can be obtained for different speech sounds, thus these models can be classified as static articulatory models. Usually these models provide the vocal-tract shapes in the two dimensional space represented by the midsagittal plane. Some other articulatory models provide 3D shapes of the vocal-tract (Dang and Honda [14], Engwall [24]).

A class of articulatory models of great interest represents those models derived by statistical analysis of lateral X-ray images of persons producing speech. Such an articulatory model was originally developed by Maeda, [53], [55] and was later extended by Laboissière and Galván, [45]. The midsagittal vocal-tract shapes of a female speaker were extracted from X-ray films recorded with a rate of 50 frames/sec. A total of 519 frames were used from ten sentences in French. A semi-polar coordinate system with a fix relation to the hard palate was used, as depicted in Figure 2.4. The contour of the vocal-tract was sampled at the intersections with the semi-polar coordinate grid lines. Using this representation the tongue shape is specified by a vector z_t, of variables corresponding to the distances of the tongue contour along the grid lines 31/30/29/28 to 7. Another variable, the jaw opening, represented by the distance between the central upper and lower incisors, is measured for each frame. The tongue shape can be described as a weighted sum of some linear components, using a procedure called general linear component model (Overall, [67]). Each linear component represents the effect of a specific articulator upon the shape and these components are mutually orthogonal. The vector z_t

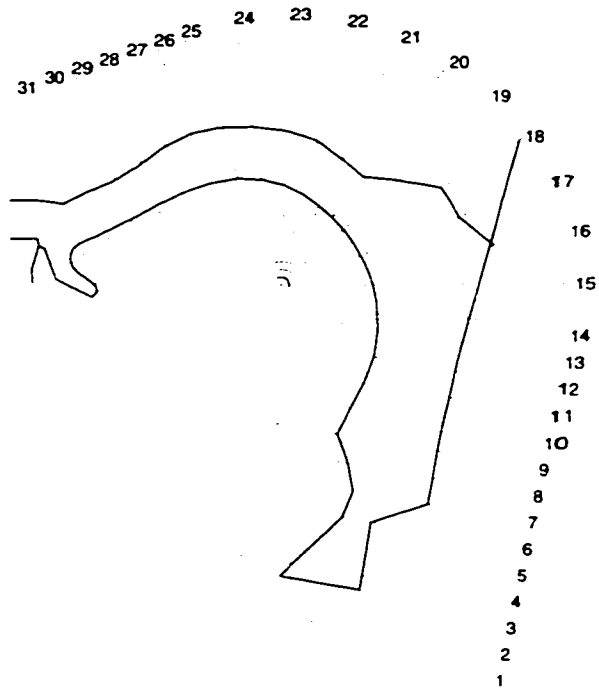

Figure 2.4: The semi-polar coordinate system for representing vocal-tract shapes

representing the tongue contour can be specified by a linear equation

$$z_t = A_t p + B_t \tag{2.60}$$

where $p^T = [j\ b\ d\ t_x\ t_y]$ is a vector of parameters representing the jaw, tongue body, tongue dorsum, tongue tip x and tongue tip y variables, $\boldsymbol{A}_t$ is a matrix of weights or loading coefficients and B_t is a vector representing the mean tongue position. A_t and B_t are computed from X-ray data by linear regression. First, the loadings A_{tj} corresponding to the jaw (the first column of the A_t matrix), are determined from data z_t of all the frames, knowing the jaw opening, applying a linear regression

$$z_t = A_{tj} j + B_t \tag{2.61}$$

where A_{tj} is a vector of loadings corresponding to the jaw and j is the jaw opening parameter (measured from X-ray data). To ensure the orthogonality of the jaw component with the other components of the vector p the influence of jaw parameter j is subtracted from the data

$$z_t^{new} = z_t - A_{tj}j + B_t. \tag{2.62}$$

After the subtraction of the jaw component, the loadings A_{tb} of a tongue body component corresponding to a grid direction in the pharyngeal region are computed from the remaining data z_t^{new}, by another linear regression. Then the influence of the tongue body component is subtracted from data. The loadings A_{td} of the tongue dorsal component corresponding to a grid direction in the dorsal region are computed similarly. Finally the loadings of tongue tip x and y components are computed. The first three elements of the vector p are accountable for about 94% of the tongue data variances. Now for any value of the tongue parameter vector p, a vector z_t representing the distances of the tongue contour can be computed using the above formula. Similarly, a linear equation can be applied to describe the z_v vector representing the velum distances along the grid lines 16 to 19

$$z_v = A_v \times d_v + B_v \tag{2.63}$$

where d_v represents the distance between velum and the throat wall, and A_v and B_v are matrices of coefficients computed by linear regression. Three other linear independent parameters are added to specify the lip protrusion, lip aperture and pharynx hight: l_x, l_y and p_h. Thus the variable parts of the vocal-tract shape can be specified by a vector of nine parameters of the model: j, b, d, t_x, t_y, d_v, l_x, l_y and p_h. All the parameters are normalized to their standard deviations. We added a fixed nasal-tract connected to the velum port for the production of nasal sounds.

In order to compute the equivalent area function of the vocal-tract, the two-dimensional articulatory model is converted to a concatenation of polygons separated by the grid lines of the model and further these polygons are replaced by rectangles with the same area, obtaining a straight model. The hight of each rectangle represents the midsagittal distance of the vocal-tract for the corresponding segment and the length of each rectangle represent the length of that segment. An $\alpha\beta$ model can transform the midsagittal distance d into area of the section A using the formula

$$A = \alpha d^{\beta} \tag{2.64}$$

where α and β are some coefficients depending on the position x of the section along the vocal-tract. An improved $\alpha\beta$ model has been used to obtain the area function from the midsagittal distances, (Perrier *et al.*, [70]).

The concatenation of the above articulatory model controlled by nine parameters with the $\alpha\beta$ model of computing the area function and with the vocal-tract acoustic model described in this chapter produced an articulatory speech synthesizer capable to synthesize high quality speech sounds. We used this articulatory synthesizer extensively to produce continuous voiced and unvoiced speech.

2.3 Dynamic Articulatory Modeling

In this section a few methods of modeling the dynamics of the articulators are presented. These modeling approaches take into account the kinematics and dynamics which govern the vocal system. Some of the most common approaches use the mass-spring system to describe the motion of articulators based on the dynamic parameters mass, damping and stiffness (Saltzman and Munhall [82], McGown and Lee [56]). Other approaches do not use explicitly the mass of the articulators and

are based on a quantitative formulation of a dynamic model described by second order critically damped equation (Browman and Goldstein [6], Kröger *et al.*, [42]).

The movement of the articulators is very complex and determined by the coordinated action of many muscles involved in producing articulatory gestures. These individual gestures of the articulators are usually overlapped in time. One important effect of this overlapping is the co-articulation phenomenon. The rotation and translation movements of the articulators have to be taken into account in order to produce a phonemic gesture defined in a vocal-tract task space. These complex movements are usually approximated by simple dynamic equations. Thus, un articulatory dynamic model can be approximated by a set of second order differential equations which describe the motion of each articulator based on its equivalent spring constant, damping constant and mass. If x_i defines the position of the i-th articulator, its motion can be approximated by the following equation

$$m\ddot{x}_i + b\dot{x}_i + kx_i = f, \tag{2.65}$$

where m, b and k represent the mass coefficient, damping coefficient and spring coefficient respectively of the x_i articulator and f is a driving force. During the production of speech these coefficients and force are functions of time. In a matrix form the equation which describes the motions of all articulators is

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{B}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} \tag{2.66}$$

where $\mathbf{M}$, $\mathbf{B}$ and $\mathbf{K}$ are matrices specifying the masses, damping coefficients and spring coefficients of the articulators. If these matrices are set diagonal the articulators are considered independent of each other, but this is a crude assumption because, in reality the motions of most of the articulators are correlated. The vector $\mathbf{f}$ is the input of the articulatory system and is related to the phonemic target or

the rest position of the articulators. This equation can be written in the following form

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{B}\dot{\mathbf{x}} + \mathbf{K}(\mathbf{x} - \mathbf{x}_0) = \mathbf{0} \tag{2.67}$$

where $\mathbf{x}_0$ is a vector representing the phonemic target or the rest position of the articulators (Saltzman and Munhall, [82]; (McGown and Lee, [56]).

These equations have been applied to a tract-variable dynamical model (Saltzman and Munhall, [82]) and to a state-variable dynamical model (Kaburagi and Honda, [38]). In the former study, the tract-variables are defined as the major task for articulators to create local constrictions in different regions of the vocal tract. The linear dynamical equation describing the motion of tract-variables $\mathbf{z}$ is

$$\ddot{\mathbf{z}} = \mathbf{M}^{-1}[-\mathbf{B}\dot{\mathbf{z}} - \mathbf{K}(\mathbf{z} - \mathbf{z}_0)] \tag{2.68}$$

where $\mathbf{z}_0$ is the target or rest position for the tract variables. These tract-variables defined as location and constriction degree of tongue tip and tongue dorsum, lips protrusion and aperture, velum and glottal states, are functions of model articulator variables which represent the actual position of the articulators. The nonlinear relationships between the tract variables $\mathbf{z}$ and corresponding model articulator variables ϕ are defined by the equations

$$\mathbf{z} = \mathbf{z}(\phi) \tag{2.69}$$

$$\dot{\mathbf{z}} = \mathbf{J}(\phi)\dot{\phi} \tag{2.70}$$

$$\ddot{\mathbf{z}} = \mathbf{J}(\phi)\ddot{\phi} + \dot{\mathbf{J}}(\phi, \dot{\phi})\dot{\phi} \tag{2.71}$$

where z is an $m \times 1$ tract variable vector and ϕ is an $n \times 1$ current articulator position vector. The $\mathbf{J}$ matrix is an $m \times n$ Jacobian matrix which has its elements defined by partial derivatives $\partial z_i / \partial \phi_j$ evaluated at the current ϕ.

The equation of motion of the model articulators is

$$\mathbf{M}(\mathbf{J}\ddot{\phi} + \dot{\mathbf{J}}\dot{\phi}) + \mathbf{B}\mathbf{J}\dot{\phi} + \mathbf{K}[\mathbf{z}(\phi) - \mathbf{z}_0(\phi_0)] = \mathbf{0}. \tag{2.72}$$

This equation can be written in the following form

$$\ddot{\phi} = \mathbf{J}^*\{(\mathbf{M}^{-1}[-\mathbf{B}\mathbf{J}\dot{\phi} - \mathbf{K}[\mathbf{z}(\phi) - \mathbf{z}_0(\phi_0)]]) - \dot{\mathbf{J}}\dot{\phi}\} \tag{2.73}$$

where the pseudo-inverse Jacobian $\mathbf{J}^* = \mathbf{W}^{-1}\mathbf{J}^T(\mathbf{J}\mathbf{W}^{-1}\mathbf{J}^T)^{-1}$ and W is a diagonal weighting matrix. This last equation describes the movements to tract-variable targets $\mathbf{z}_0$ of the model-articulator variables ϕ.

In the latter study, [38], Kaburagi and Honda tried to solve a linear equation relating some state-variable, $\mathbf{x}$, to some tract variable, $\mathbf{z}$, defined by

$$\mathbf{z} = \mathbf{E}\mathbf{x} \tag{2.74}$$

where $\mathbf{E}$ is a conversion matrix. They applied the second-order discrete linear equations to the state-variables $\mathbf{x}$ representing the relative positions of the articulators to a jaw-based coordinate system as follows

$$\mathbf{x}_i(n) - 2\tau\mathbf{x}_i(n-1) + \tau^2\mathbf{x}_i(n-2) = (1-\tau)^2\mathbf{f}_i(n) \tag{2.75}$$

where n denotes time sample, τ is a constant and $\mathbf{f}_i$ represents an input force for the i-th state variable articulator which can be the jaw, upper lip, lower lip or tongue. The vector $\mathbf{y}$ defining the absolute positions of the articulators are obtained by linear transformation from state-variable vector $\mathbf{x}$, and the tract variable vector $\mathbf{z}$ is also obtained from the absolute position vector $\mathbf{y}$ by a linear transformation. Thus the tract variables are linear transformations of the state variable as shown in Equation 2.74. When a motor task is given at a moment by the tract-variable target, the state variables $\mathbf{x}$ are determined by solving these simultaneous linear equations.

If the dimension of **z** is smaller than that of **x**, these equations restrict the values to only a subspace of **x**, and the inverse mapping, from **z** to **x** becomes one-to-many. To solve this problem a cost function is used. This cost function is defined as a sum of quadratic forms of changes in output movements and input forces of the system. Given the motor task, a unique solution of these equations is determined by minimizing the cost function. This is an optimal control problem with linear dynamics and quadratic criteria and can be solved by dynamic programming.

Browman and Goldstein, in their work on articulatory phonology. ([4]. [6]. [7]). defined gestures as dynamic articulatory structures. The speech is modeled as 'constellations' of articulatory gestures. The gestures are specified by a set of tract variable (e.g., lip protrusion LP. lip aperture LA. tongue body constrict location TBCL. etc.). In an early study. [8]. they modeled the dynamics of the articulators by fitting sinusoids to the articulatory trajectories. based on the motion of an undamped mass-spring system. They found a very good match between the fitted sinusoids and the real articulatory trajectories. Then they extended the dynamic modeling to a damped mass-spring model, which they used in articulatory phonology. A gesture was defined as the dynamic patterns of the articulators in moving from an arbitrary rest position toward a target and back to the rest position. This represents a complete cycle of a gesture. The elementary gestures of different tract variable can overlap in time. From this work, we have been inspired to use the segments of speech defined by the onset interval of the gestures, that is, the dynamic patterns in moving the articulators from an arbitrary rest position (defined by a particular phoneme), toward a target position (defined by a different phoneme).

Chapter 3

Articulatory and Acoustic Representations

The potential of using the articulatory representation for automatic speech recognition is emphasized by comparing two simple phonetic classification experiments based on articulatory features, respectively, acoustic features. These experiments are first performed in the original vector spaces, and then in some transformed sub-spaces, called *task spaces*. The articulatory and acoustic data used in these experiments have been simultaneously recorded with an electromagnetic midsagittal articulograph.

3.1 Articulatory and Acoustic Representations in Original Spaces

The articulatory representations of speech could be useful for phonemic classification or speech recognition if a reliable method of acoustic-to-articulatory mapping

would be used. Experiments of phonemic classification based on direct articulatory measurements have shown a promising potential of using articulatory features. In this section, to emphasize the potential of using articulatory features for speech recognition, a phonemic classification experiment based on articulatory representations of speech sounds is presented in comparison with the classification based on the corresponding acoustic features. In articulatory phonetics the speech sounds are analyzed using the configurations of the vocal-tract and some of the articulators. The fact that each speech sound has a distinct combination of place and manner or articulation and voicing status suggests that this information can be useful in recognizing and classifying speech sounds. Considering this distinct articulatory-phonetic mapping, it is expected that an automatic classification of the speech units could be possible if some appropriate articulatory features would be available.

For human speech recognition, the motor theory of speech perception [49], states that humans use some knowledge of their internal articulation mode in the process of perception and classification of sounds. However, to what extent and how much articulatory information is inferred in recognizing speech sounds and perception of speech is not known. It is known that, before transformed into acoustic signal, the speech information is encoded by the speaker into articulatory gestures and it is expected that in this articulatory stage the linguistic information should be present at least as much as into the speech acoustic signal.

There is a great interest in extracting the articulatory information from the speech acoustic signal and using this information to improve the recognition rate of automatic speech recognizers. Experiments of automatic speech recognition and phonetic classification have been done based on articulatory measurements added to the acoustic speech data to show how much the articulatory data can improve the recognition rate. Such preliminary results have shown that by appending articula-

tory data to the acoustic data the speech recognition error rate decreased (Petajan [71], Zlokarnik [107], [108]). A different approach of articulatory based automatic speech recognition tried to use some prior articulatory knowledge and the acoustic features of speech (Deng and Sun [16], [17]). In this experiment the authors imposed a mapping between the states of a Hidden Markov Model (HMM) and the articulatory features. Although this approach has provided promising results, the knowledge about articulation has not been inferred from data and additional sources of prior articulatory knowledge need to be used.

In this section, an acoustic-phonetic and an articulatory-phonetic relationship are analyzed in a form of phonetic classification experiments based on simultaneous articulatory and acoustic data. A small amount of articulatory data, simultaneously recorded with the acoustic data, have been acquired for this experiment using an alternating magnetic field method (Schonle *et al.*, [83], Tuller *et al.*, [101]). We used the technique of Electromagnetic Midsagittal Articulography (EMA), (Perkell et. al., [69]) employing a device built by Carstens Medizinelektronik GmbH, Gottingen, Germany, [10].

The phoneme-specific vocal-tract shapes are considered those invariant features representing the phonemic targets during speech production and are called motor tasks or dynamic tasks (Kaburugi and Honda, [38]; McGowan and Lee, [56]). A description of these phonemic targets can be made using some reduced subspaces, called *task spaces*, in both articulatory and acoustic domains (Honda and Kaburagi, [35]). The phonemic target represents some points in the multidimensional articulatory or acoustic spaces to be reached during the production of a particular phoneme. Depending on speech rate and stress, these points can or cannot be reached but always represent the target towards which the articulatory and acoustic vectors are driven during production of each phoneme. It is difficult to define exactly the tar-

gets therefore for each phoneme we consider a target region representing a subspace of the original space defined by the invariant features for a specific degree of articulation. For this description there are an infinity of targets defined by the invariant units or gestural primitives for each degree and way of articulation. The articulators are driven by forces specifying the targets for each phoneme-specific gesture during speech production. The analysis of speech patterns can provide information about phonemic targets if the articulatory and acoustic vectors are analyzed at those points in time at which the forces for the next phonemic target are applied. These points can be found at the minimum velocity of the articulatory or acoustic vectors. Multivariate statistical methods can be used for analysis of these task spaces (Morrison, [64]). One of such methods is the principal component analysis (PCA). Using the generalized eigenvectors one can find phoneme-specific subspaces of the original multidimensional spaces in which the tasks have the maximum concentration (Honda and Kaburagi. [35]). This means that in these subspaces. the task vectors have minimum variance along each direction. The task space can be defined as having a number of directions for which the variance ratio of all vectors to that of phoneme-specific vectors, represents most of the total variance of this ratio. Usually 2 or 3 dimensions are enough to represent these task spaces. The phonemic target can be approximated by the center of gravity of these task spaces. To see how well these task spaces can represent the phonemes or the phonemic targets some experiments of analysis and classification of articulatory and acoustic patterns have been done in both original spaces and task spaces. If the results are comparable we can say that even though with drastic reduction of the dimensionality, the task spaces can still represent well the phonemic targets or those invariant features of each phoneme.

A probabilistic pattern analysis and classification method, such as *APACS*,

(Chan and Wong, [11]), can be applied to both articulatory and acoustic patterns of speech. We used this classification method in both acoustic and articulatory spaces, as presented in [23]. This method handles uncertainty that comes from inconsistent, incorrect or missing information in the training examples and can be classified as an inductive learning method, or a method of classification based on learning from examples. The set of training examples contains a number of objects and each object belongs to a class represented by a phoneme. Each object is described by a number of attributes. The whole algorithm has three phases.

In the first phase, the detection of underlying patterns in the training examples, is done by computing the contingency table for each attribute. Each element of the contingency table represents the number of vectors from the training set that belong to a specific class and have a specific attribute value. For each attribute we can compute the expected table in which each element represents the expected number of objects that have a specific attribute value and belong to a specific class. For detecting the underlying patterns we can detect the relevant features for classification. These relevant features are those attribute values that are important for the characterization of a certain class of objects. This detection of relevant features is based on the difference between the probability of an object to belong to a specific class and the probability of the same object to belong to the same class given a specific value of the attribute. If this difference is significant then that attribute value is a relevant feature.

The second phase of the algorithm is to construct rules for classification based on the detected patterns. The classification rules describe each class of the training examples probabilistically.

The third phase of the algorithm is the prediction of class membership of new objects not used in the training set. This prediction can be achieved by evaluating

each attribute of the new object if it is relevant and in this case we can compute the weight of evidences for that object having that attribute value for belonging to different classes. After we evaluate all attributes of the new vector we can have more than one class predicted for that object. We choose the class that has the greatest sum of weight of evidences, to be the class with the greatest probability for the new object to be included in.

We have done some phonetic classification experiments using this statistical analysis and classification method for both acoustic and articulatory speech patterns. These experiments have used first the features from the original acoustic and articulatory spaces and then the features from the task spaces, in order to determine how well the acoustic and articulatory task spaces can represent the phonemic targets speech. These experiments were carried out for 5 American English vowels, produced in consonantal contexts with relative slow movements of the articulators by a single male subject, in order to increase the probability of the articulators to reach the phonemic target for each vowel. For the orthographic representation of these vowels and consonants we used the TIMIT convention, [65]. The five vowels used were /ah/, /eh/, /iy/, /ao/ and /uh/. The utterances recorded were of the form VCV (vowel-consonant-vowel). For each of the five vowels, 17 VCV tokens were produced by selecting one of the following consonants: /b/, /d/, /f/, /g/, /h/, /zh/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /sh/, /t/, /v/ and /z/. Each such VCV token has been repeated three times. In all these tokens the first and second vowels were the same and thus each of the 5 vowels has been produced 102 times. In these simple experiments we did not actually split the whole data in training and test data, since the comparative results are of interest and not the absolut ones.

The articulatory and acoustic data representing the middle vowel positions in the articulatory space are plotted in Figure 3.1 and in the acoustic space are plotted

in Figure 3.2. In the acoustic domain, we used a single graph to represent the plots in both F1-F2 plane and F3-F4 plane. The articulatory data, obtained with an electromagnetic midsagittal articulometer (Carstens GmbH, [10]), consist of vectors containing the x and y coordinates of 3 sensors placed on the lower lip, tongue body and tongue dorsum respectively (6 dimensions) in the midsagittal plane. Other two sensors, placed on the nose and upper teeth, were used as references and for correcting the movements of the head related to the helmet. The acoustic data consist of vectors containing the first 4 formant frequencies derived from a 14-th order LPC analysis. Each ellipsis from these figures is drawn using two standard deviations in each dimension, for each vowel. In the articulatory domain, the plots corresponding to different vowels are quite overlapped. In the acoustic domain, the plots in F1-F2 plane are quite disjoint for these vowels, but they are overlapped in F3-F4 plane.

The results of vowel classification in the articulatory and acoustic spaces are presented in Tables 3.1 and 3.2. In each classification table the elements of the main diagonal represent the number of correct classified vectors whereas all other elements represents the number of misclassified vectors. As can be seen in both original spaces the results are similar. The lower classification results in the articulatory space for /ao/ and /uh/ may suggest that the x and y positions of the three coils placed on the articulators are a little less relevant for these vowels. In the acoustic domain, the classification results are better, probably due to the greater separability of the acoustic patterns in the F1-F2 plane. Thus, a single /uh/ vector has been misclassified as /ao/.

This small experiment shows that the information encoded into articulatory domain is relevant to the phonetic identity of the vowels, and thus, the articulatory features might be potential features for automatic speech recognition.

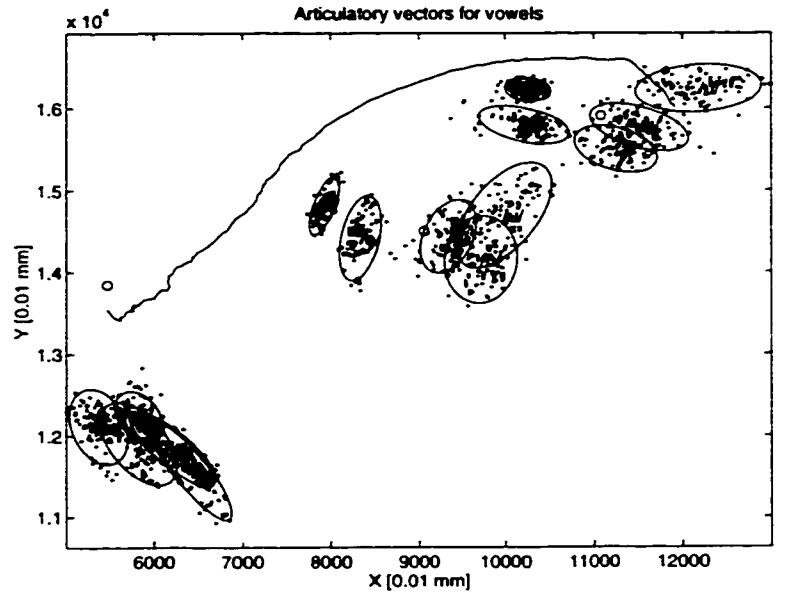

Figure 3.1: Scatter plot of the articulatory vectors

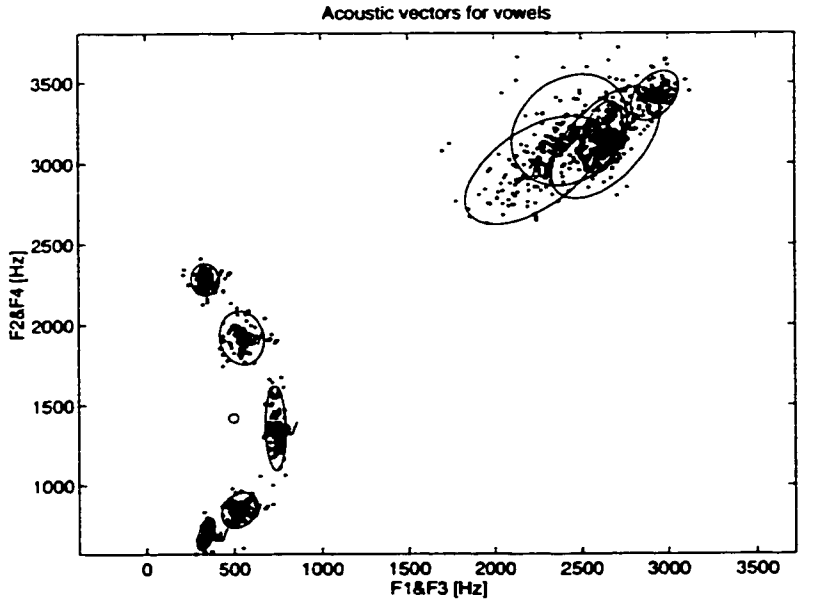

Figure 3.2: Scatter plot of the acoustic vectors

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	101	1	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	98	4
/uh/	0	0	0	5	97

Table 3.1: Confusion matrix for classification using articulatory features

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	102	0
/uh/	0	0	0	1	101

Table 3.2: Confusion matrix for classification using acoustic features

3.2 Articulatory and Acoustic Representations in Task Spaces

A statistical analysis method has been used as in [35] in both articulatory and acoustic spaces to study the degree of concentration of the target distributions in each space. The phonemic target is defined as a subspace linearly transformed from the original space so that the distribution of the achieved tasks is maximally concentrated for that phoneme. This phoneme specific subspace, called task space, is linearly transformed from the original space by finding the generalized eigenvectors for the covariance matrix of all vectors, $\mathbf{C}_a$ and covariance matrix of each phoneme vectors, $\mathbf{C}_p$, in each space. This has been done by solving the matrix equation

$$\mathbf{C}_a\mathbf{F} = \mathbf{C}_p\mathbf{F}\mathbf{\Lambda} \tag{3.1}$$

where $\mathbf{F}$ is a matrix containing in each column an eigenvector and $\mathbf{\Lambda}$ is a diagonal matrix containing the generalized eigenvalues which are equal to the variance ratio of all vectors to the phoneme specific vectors for each dimension. Figure 3.3 presents the degrees of concentration of target distribution along the first dimension for each of the 5 vowels in both articulatory and acoustic spaces. This degree of concentration is represented by the ratio of standard deviation of the entire distribution to that of each vowel distribution in the task space. From this figure one can observe a higher concentration for vowels /eh/ and /iy/ in the articulatory task space and of vowels /iy/, /ao/ and /uh/ in the acoustic task space. The sum of the variances ratios for the first two dimensions of maximum concentration represents more than 90 percent of the whole variance ratio for each vowel in each space. Each dimension of the two dimensional task spaces represents a linear transformation of all dimensions of the original space. In a matrix form the task space, $\mathbf{z}$, can be

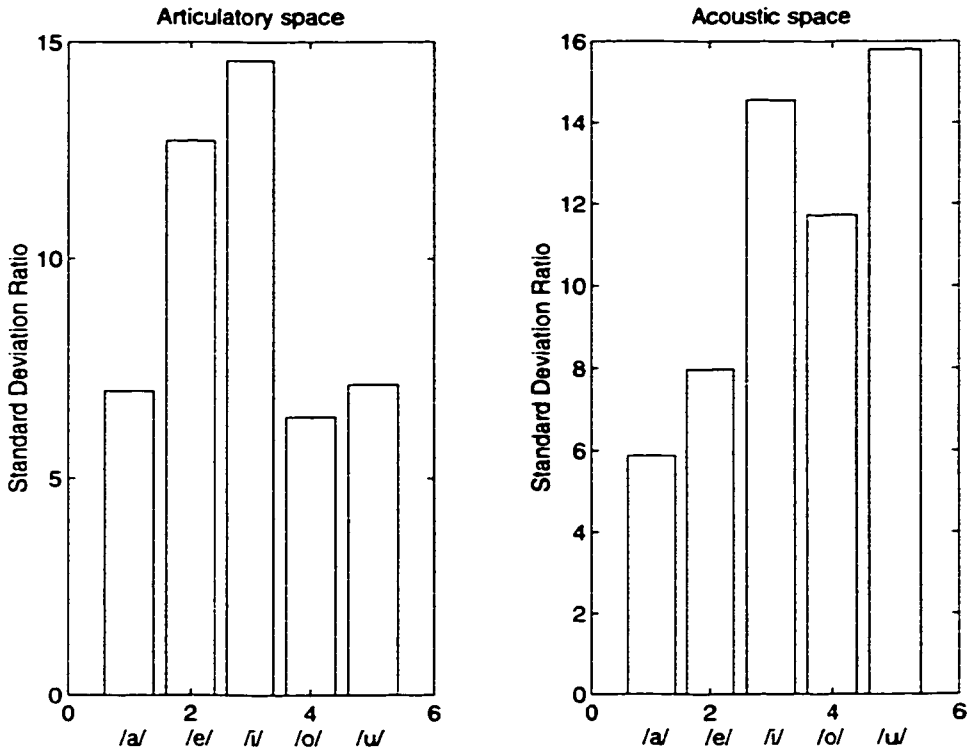

Figure 3.3: Degrees of concentration of target distributions in articulatory and acoustic spaces.

represented by a linear transformation of the original space, $\mathbf{x}$, as follows:

$$\mathbf{z} = \mathbf{E}\mathbf{x} \tag{3.2}$$

where $\mathbf{E}$ is a matrix with coefficients of transformation obtained from matrix $\mathbf{F}$ of Equation 3.1 (In this case $\mathbf{E}^T$ represents the first two columns of matrix $\mathbf{F}$). Because $\mathbf{z}$ contains the first two principal components of the ratio of covariance matrix of all vectors to the covariance matrix of one phoneme vectors , its dimensions are perpendicular each other for different positive eigenvalues. The articulatory and acoustic task spaces are presented for all vowels studied in the following figures. In all these figures the abscissa is represented by the first dimension of maximum

concentration and the ordinate is represented by the second dimension of the maximum concentration in each task space. The two principal axes of each ellipse are equal to two times the standard deviation along each axis.

In order to reveal the properties of the task spaces, the experiments of classification of articulatory and acoustic patterns of speech were repeated for each of the task spaces using the same APACS method (Chan and Wong, [11]).

The classification results for the articulatory and acoustic task spaces are presented in Table 3.3 and 3.4 for the vowel /ah/, Table 3.5 and 3.6 for the vowel /eh/, Table 3.7 and 3.8 for the vowel /iy/, Table 3.9 and 3.10 for the vowel /ao/, Table 3.11 and 3.12 for the vowel /uh/. For each of the vowels the classification results are similar in the articulatory and acoustic task spaces, as in the case in the original spaces. One finding of these classification experiments is that in the task space of each vowel in both articulatory and acoustic domains, that specific vowel for which the task space was constructed was better classified than in the original space. We drawn the conclusion that the task spaces represent some kind of filters for better viewing the articulatory or acoustic vectors of that particular sound.

We observed a kind of similarity of vector distributions between the articulatory task space for /uh/ and the acoustic task space for /ah/. With some simple linear transformations the two task spaces can be superimposed as presented in Figure 3.14. In the two task spaces, inverted and rotated, one can observe the similarity of vowel distributions of the scattered vectors with the well known *vowel triangle* from the F1-F2 plane (F1 representing the first format frequency on abscissa and F2 representing the second format frequency on ordinate).

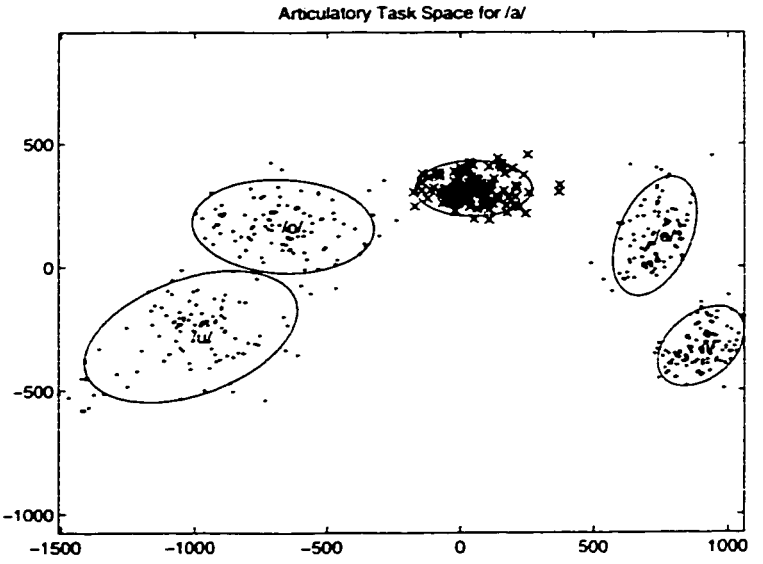

Figure 3.4: Articulatory task space of the vowel /ah/

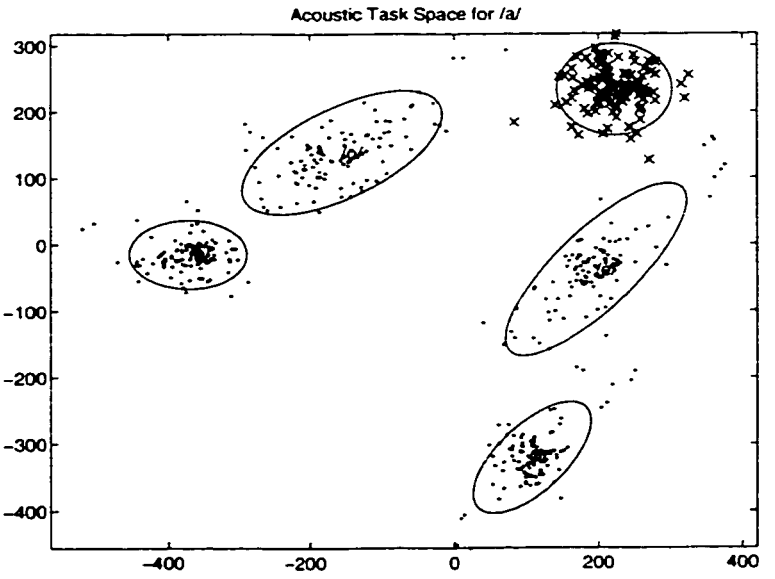

Figure 3.5: Acoustic task space of the vowel /ah/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	102	0
/uh/	0	0	0	4	98

Table 3.3: Confusion matrix for classification in the articulatory task space of /ah/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	95	1	6	0
/iy/	0	3	99	0	0
/ao/	1	0	0	99	2
/uh/	0	0	0	4	102

Table 3.4: Confusion matrix for classification in the acoustic task space of /ah/

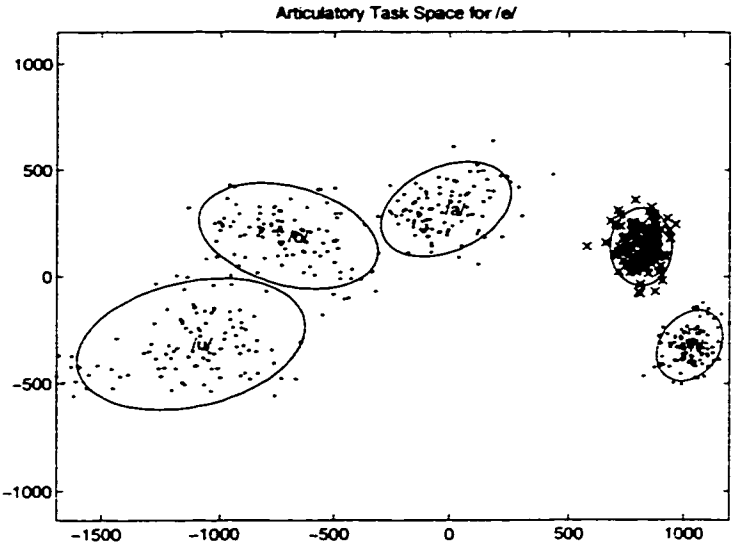

Figure 3.6: Articulatory task space of the vowel /eh/

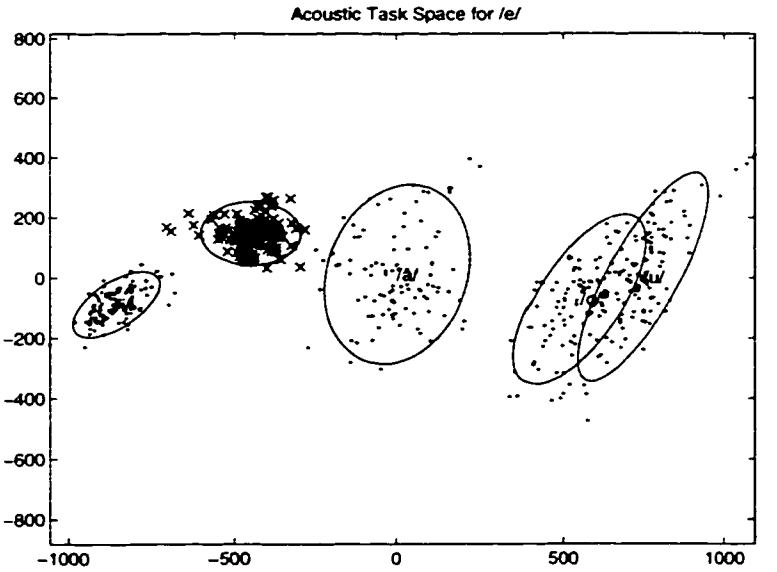

Figure 3.7: Acoustic task space of the vowel /eh/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	100	0	0	2	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	101	1
/uh/	0	0	0	3	99

Table 3.5: Confusion matrix for classification in the articulatory task space of /eh/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	99	2	0	1	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	91	11
/uh/	0	0	0	15	87

Table 3.6: Confusion matrix for classification in the acoustic task space of /eh/

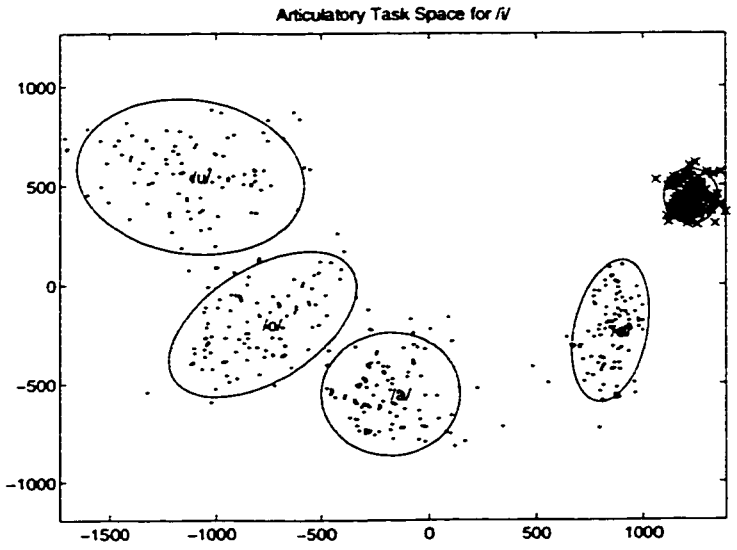

Figure 3.8: Articulatory task space of the vowel /iy/

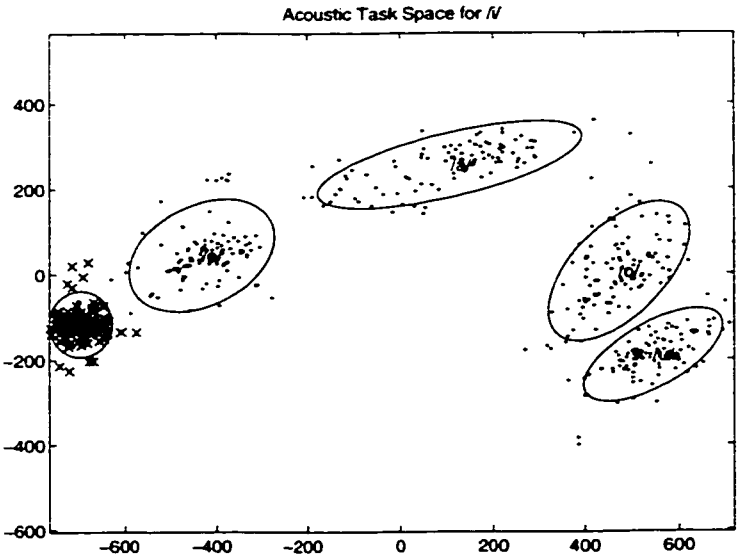

Figure 3.9: Acoustic task space of the vowel /iy/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	4	0	0	97	1
/uh/	0	0	0	6	96

Table 3.7: Confusion matrix for classification in the articulatory task space of /iy/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	102	0	0	0	0
/eh/	0	99	3	0	0
/iy/	0	0	102	0	0
/ao/	1	0	0	100	1
/uh/	0	0	0	2	100

Table 3.8: Confusion matrix for classification in the acoustic task space of /iy/

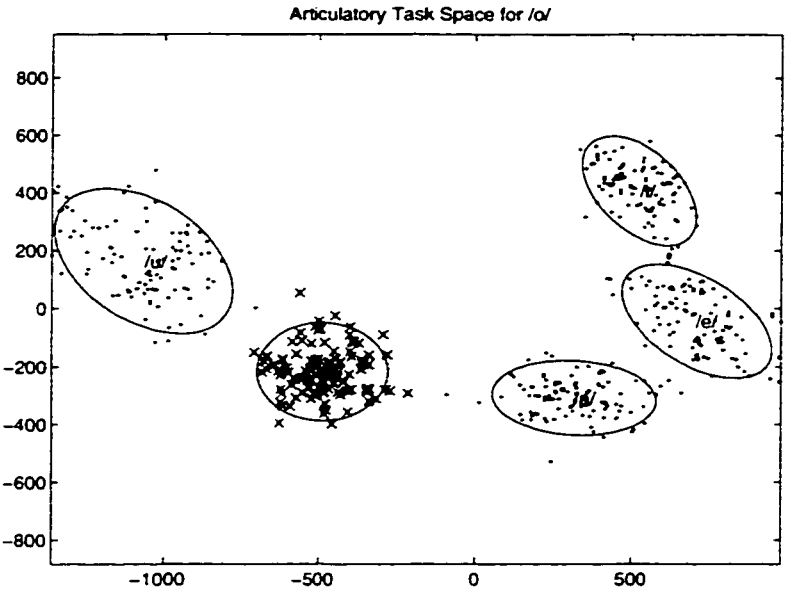

Figure 3.10: Articulatory task space of the vowel /ao/

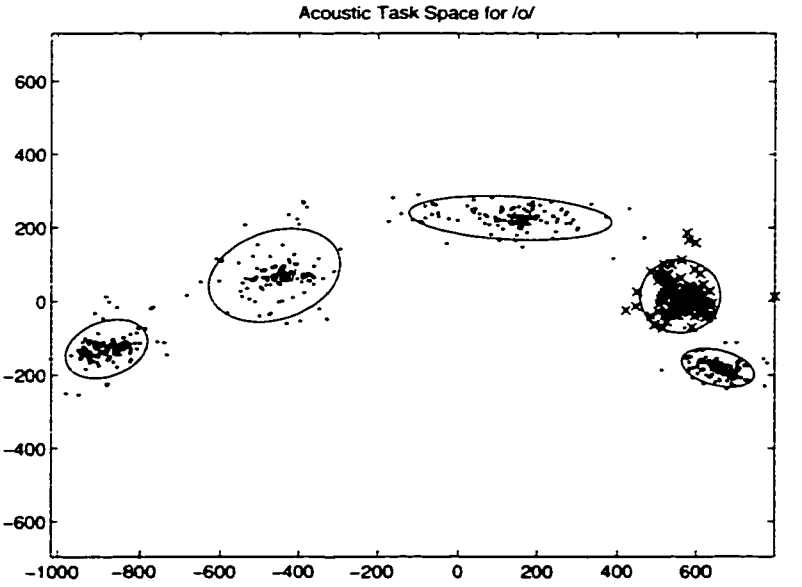

Figure 3.11: Acoustic task space of the vowel /ao/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	96	6	0	0	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	0	0	0	102	0
/uh/	0	0	0	1	101

Table 3.9: Confusion matrix for classification in the articulatory task space of /ao/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	101	0	0	1	0
/eh/	1	98	3	0	0
/iy/	0	0	102	0	0
/ao/	1	0	0	101	0
/uh/	0	0	0	1	101

Table 3.10: Confusion matrix for classification in the acoustic task space of /ao/

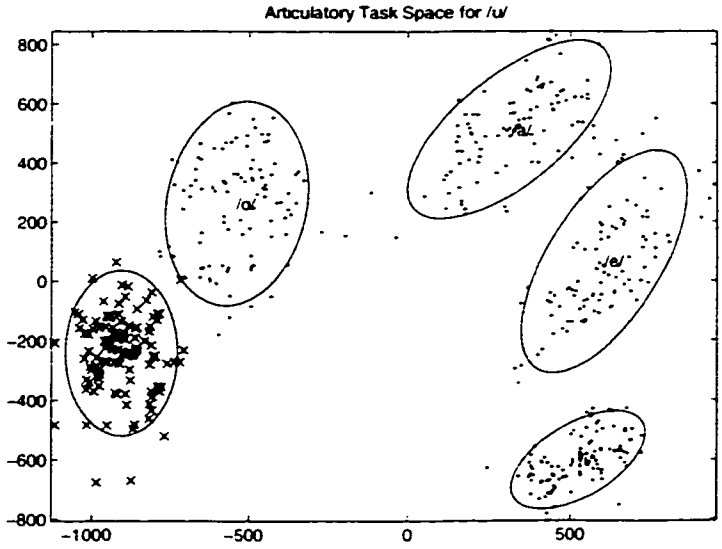

Figure 3.12: Articulatory task space of the vowel /uh/

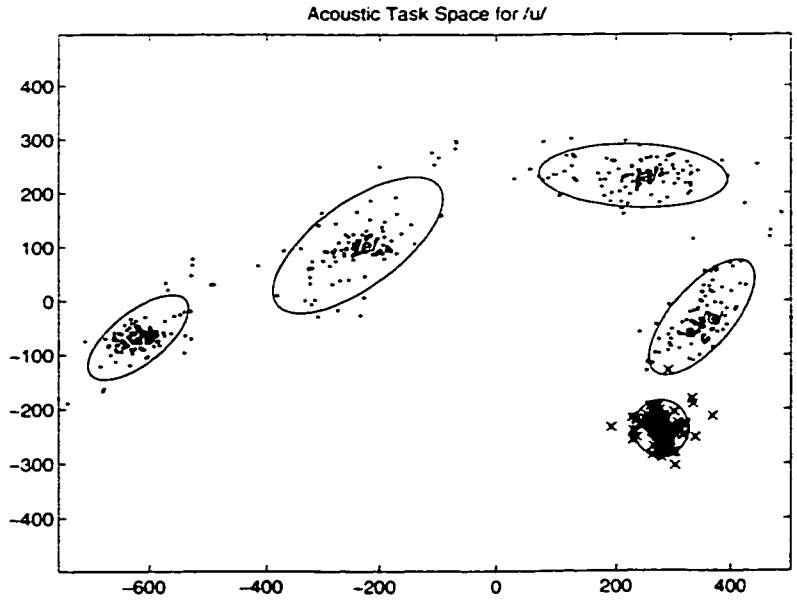

Figure 3.13: Acoustic task space of the vowel /uh/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	99	3	0	0	0
/eh/	2	99	1	0	0
/iy/	0	0	102	0	0
/ao/	0	1	0	100	1
/uh/	0	0	0	0	102

Table 3.11: Confusion matrix for classification in the articulatory task space of /uh/

	/ah/	/eh/	/iy/	/ao/	/uh/
/ah/	100	0	0	2	0
/eh/	0	102	0	0	0
/iy/	0	0	102	0	0
/ao/	1	0	0	101	0
/uh/	0	0	0	0	102

Table 3.12: Confusion matrix for classification in the acoustic task space of /uh/

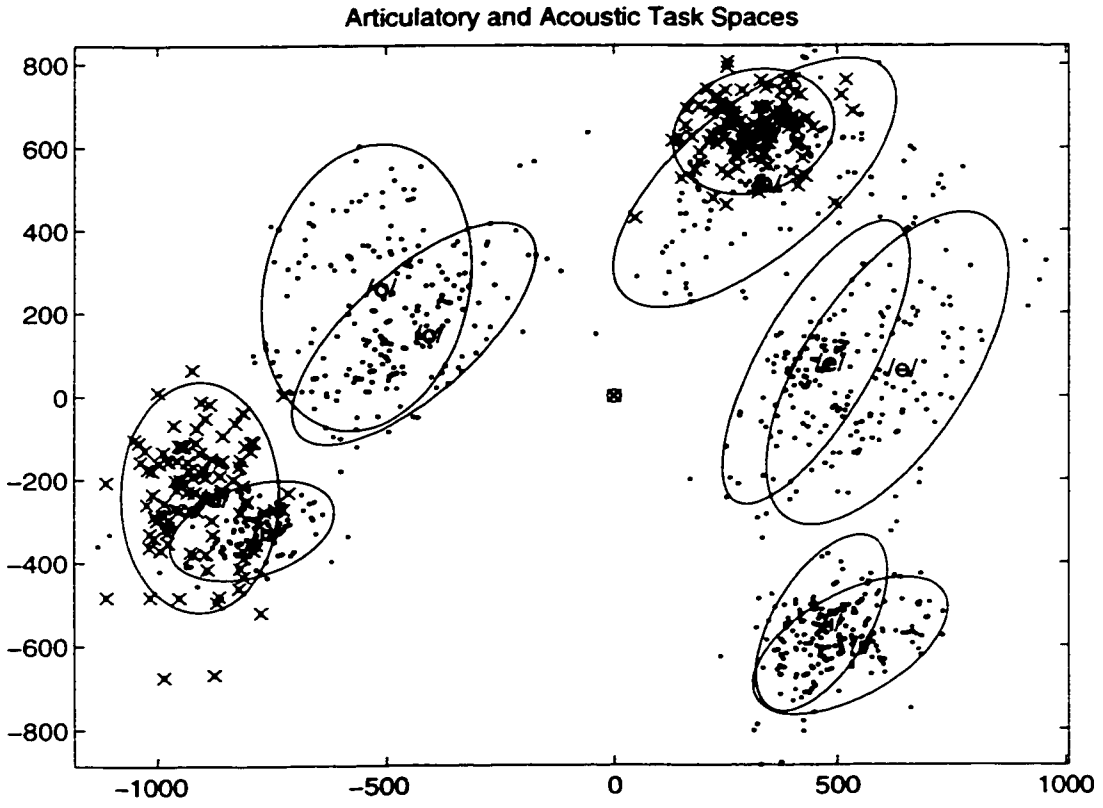

Figure 3.14: Superposition of acoustic task space of the vowel /ah/ and articulatory task space of the vowel /uh/

Chapter 4

Acoustic-to-Articulatory Inversion

This main chapter of this dissertation presents the new phonological, gestural-based speech inversion method. This speech inversion approach accounts for coarticulation phenomena and, together with a vocal-tract length normalization, can be used to systematically account for interspeaker variabilities. The presentation of the speech inversion method starts with the formulation of the phonological coproduction segments and models of speech, continues with the method of approximating the non-linear observation function, the statistical method for estimating the parameters of the models, the estimation of articulatory state based on Kalman filtering technique and concludes with the segmentation and recognition of the phonological coproduction models and estimation of the most likely articulatory trajectories.

The acoustic-to-articulatory inversion method presented in this thesis consists of two main parts: the training, which is described by the first three sections of this chapter and the estimation of articulatory trajectories by Kalman filtering and recognition of models, which are described by the last two sections.

As the natural speech represents the output of the dynamic articulatory system, the movements of this system produce continuous changes in the shapes of the vocal tract as it goes from one phoneme to another. Thus, the speech acoustic signal is not merely a simple concatenation of stationary segments representing phonemes. It contains the underlying dynamic constraints imposed by the articulatory system and thus, it is a natural way of approaching the speech inverse problem from a dynamic point of view. The statistical method for sound-to-gesture inversion presented in this thesis is based on dynamical system modeling of speech production and has at its heart the statistical versions of the state-space model representing the dynamic vocal system. The two equations defining the state-space model of a dynamic system are the state equation, which imposes dynamic constraints on the evolution in time of the system, and the output or observation equation, which models the direct relationship between the hidden state variable and the measurements or the observation variable. These equations contain stochastic components in a form of random variables whose only known parameters are the type of distribution and the first and second order statistics. The well developed statistical methods of filtering, predicting and smoothing used in linear dynamical modeling and their extension to the non-linear dynamic systems can be applied to estimate the evolution in time of the hidden state variable of the system. System identification techniques like the Maximum-Likelihood (ML) method can be applied to the estimation of model parameters from training observations of the system. A ML version of model parameter estimation is the Expectation-Maximization (EM) algorithm (Dempster et. al., 1977, [15]), which iteratively estimates the model parameters from data with unobserved components and has been extensively used for parameter estimation in the speech recognition field.

One of the main innovations proposed by this study in the area of acoustic-

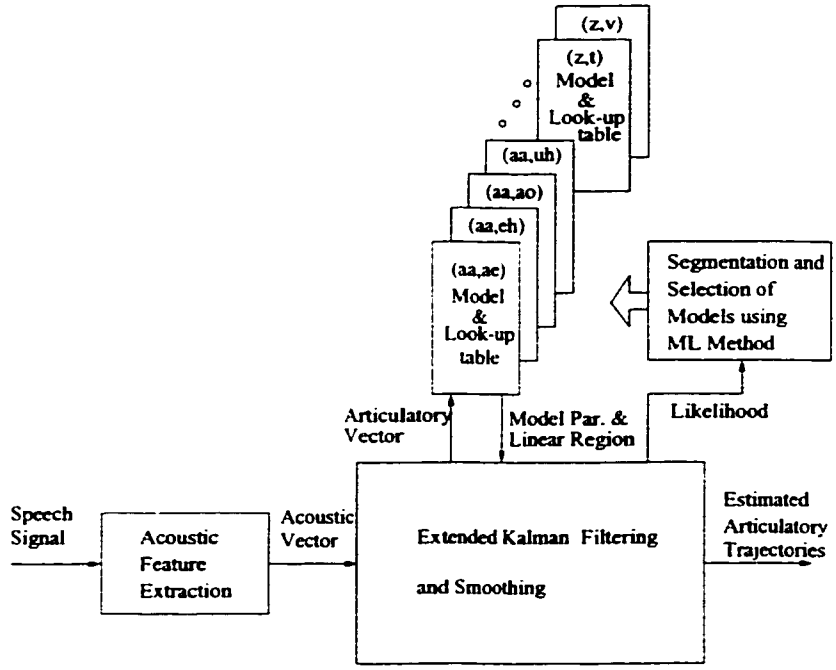

Figure 4.1: General Block Diagram of the Speech Inversion Method based on Phonological and Dynamical Constraints

to-articulatory inversion consists in the new way of imposing high-level phonological constraints to this inversion. We imposed these phonological constraints by building different coproduction models and associating to each model a particular articulatory-acoustic sub-function.

The general block diagram of the acoustic-to-articulatory inversion method developed in this study is presented in Figure 4.1. The speech signal is pre-processed and the acoustic feature extracted as a sequence of acoustic vectors. These acoustic parameters are applied to the extended Kalman filtering and smoothing block. For a short segment of speech, the extended Kalman filter is applied repetitively using each time the parameters of a different phonological dynamical model. A likelihood measure is computed for each model and, based on this measure, the model with

the maximum-likelihood is selected. A method of segmenting the speech signal is also used based on the likelihood measure. Then, for each recognized segment of speech, the Kalman smoother is applied using the corresponding model parameters. The estimated articulatory trajectories are obtained from the smoother. Details regarding each phase of the speech inversion method are given in the next sections.

4.1 Coproduction Segments and Models of Speech

The speech inversion method developed in this study is based on modeling the speech production system using linear dynamic models with non-linear observation functions. Studies of phonemic speech recognition based on modeling the speech dynamics, considered the parameters of speech to be non-stationary and, thus, functions of the corresponding phonetic unit to which the speech segment was affiliated (Shirai and Kobayashi [91] and [92]; Ostendorf and Roukos [66]; Digalakis *et al.*, [18] and [19]; Ramsay and Deng [75] and [76]).

Adopting the technique used in classical approaches of automatic speech recognition, these authors employed a different dynamic model for each phoneme or phonemic transition [92], [75]. Even though such an approach is motivated primarily by the goal of classification and recognition of the phonemes, we adopted this method in our study of acoustic-to-articulatory inversion in order to improve the accuracy of the estimated articulatory trajectories. This approach is supported by the conclusions of some studies (Shirai and Kobayashi [91] and [92]; Browman and Goldstein [8]; Kröger *et al.*, [42]), which have shown that the articulatory dynamic parameters change during the speech production and this is related to the linguistic information and phonetic identity of the speech sounds.

The main innovation in imposing these new phonological constraints consists

in dividing the articulatory-acoustic function into sub-functions corresponding to phonological coproduction models. Thus, not only that we use different dynamic parameters to model different speech segments (coproduction of gestures), but we employed segment specific articulatory-to-acoustic mappings. This last characteristics has probably the most important effect on the accuracy of estimated articulatory trajectories. In addition, our experiments have shown that averaging the parameters of all models corresponding to different phonological units will decrease the accuracy in estimating the articulatory trajectories. This method of speech inversion based on constructing a different dynamical model with a different observation function for each phonological coproduction model does not represent just a generalized mathematical formulation adopted from speech recognition but is one of our important findings revealed by our experiments of inversion for different classes of speech sounds, based on synthesized and real speech data.

The previous studies dedicated to inverting the articulatory-to-acoustic transformation did not impose any phonological or linguistic constraint on this inversion, so they considered the speech as the output of a single dynamic model with time invariant parameters. On the other hand, preliminary studies and experiments of articulatory based speech recognition which, subsequently, tried to estimate the articulatory state, using different models for different phonetic/phonologic units, did not actually divide the observation function (Shirai [87]; Shirai and Kobayashi [90] and [91]; Ramsay and Deng [75] and [76]; Zlokarnik [107] and [108]; Krstulovic [43]; King and Wrench [40]).

A question of whether or not one needs to know the phonologic affiliation of a speech segment in order to accurately recover the articulatory gestures from the speech acoustics may arise. This, in our opinion, represents an open question which needs more debates and experimental evidences from speech researchers and

scientists. However, our experimental findings encourage us to support the idea that, knowing the phonological or linguistical affiliation of a speech acoustic segment helps in increasing the accuracy of estimating the articulatory trajectories.

Having established this new characteristics of the speech inversion method as developed here, we will define now the phonological coproduction segments (or units) and models of speech. The main idea in defining the phonological coproduction units to be associated to the dynamical models relies on the concept of gesture from articulatory phonology. These gestures in the area of articulatory phonology represent a new way of defining the articulatory structures (Browman and Goldstein [8], [4], [5], [6] [7]). These phonological gestures are patterns of articulatory movement corresponding to segments of speech, unlike in the traditional linguistic phonetic research where the phonetic units are represented by static physical parameters or vocal-tract shapes. The original definition of gestures in articulatory phonology was based on specifying the geometric tract variables which produce the gestures (e.g., bilabial gestures). In this thesis we adopt the concept of gestures as phonologic structures and extend it from the tract variables to the more general articulatory variables which can be represented by the parameters of an articulatory model or by the coordinates of some pellets placed on different articulators. Thus, we define the *phonological coproduction segment* as the articulatory realizations of a *constellation* of elementary, possible overlapped gestures, as defined by Browman and Goldstein [7], in moving toward a phonetic target and starting from an initial state represented by a different phonetic unit. An entire utterance can thus be modeled by a concatenation of such coproduction segments or units. We associate a dynamical model to each coproduction segment. The models describe the movements from an initial articulatory configuration corresponding to a phonetic unit α, to a target articulatory configuration, corresponding to a different phonetic unit β.

Such an (α, β) model can be associated with the sequence of articulatory patterns in producing any phonologically possible combination of two consecutive phonetic units.

Even though the articulatory gestures represent the articulatory movements necessary to produce the phonetic target β, we do not associate this phonological unit with the phonetic unit β because the dynamic patterns of this motion are to a great extent dependent on the initial articulatory state, represented by the phonetic unit α. Figure 4.2 presents the trajectories of a tongue body articulatory parameter recorded with an electromagnetic midsagittal articulograph (EMA), for two segments of speech — /ah s/ and /eh s/, from a male speaker. The articulatory parameter represents the X position of a sensor placed on tongue body (T2-X). Although both segments have the same target phoneme /s/, the trajectories are different. In this figure one can see that the 'T2-X' parameter has a positive slope for the /eh s/ segment and a negative slope for the /ah s/ segment. Not only the trajectories are different, but also the articulator's position of the target phoneme /s/. Table 4.1 presents the dynamic parameters Φ_1 and Φ_1, for the two segments of speech, of a second order linear system described by the equation

$$x(k+1) = \Phi_1 x(k) + \Phi_2 x(k-1) + w(k), \tag{4.1}$$

where k is the discrete time, x is the articulatory variable (T2-X), and w is a driving variable. The dynamic model parameters were estimated using a Maximum-Likelihood method from the articulatory trajectories of the two different segments, /ah s/ and /eh s/. The details of the ML model parameter estimation used, will be presented in one of the next sections.

The human vocal apparatus, as a dynamic system, contains different articulators, excitation sources, sound radiators and some fixed parts of the vocal and nasal

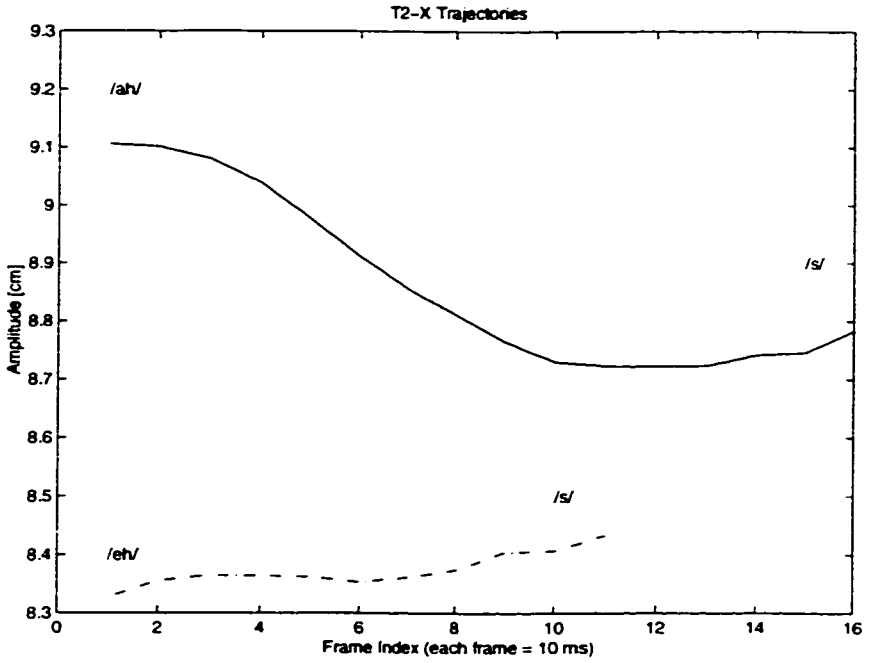

Figure 4.2: Examples of T2-X parameter trajectories for /ah s/ and /eh s/ segments of speech. (EMA recordings from a male speaker)

	Φ_1	Φ_2
/ah s/	1.9748	-0.9746
/eh s/	1.0872	-0.0862

Table 4.1: Examples of different dynamic model parameters for T2-X

tracts. The speech, representing at a higher level a sequence of phonological units, is the result of overlapped articulatory gestures in which the co-articulation of consecutive sounds can significantly change their patterns. These patterns can also be altered by changing the rate of the speech due to the dynamic constraints imposed by the articulators. Thus, the speech, as the output of the dynamic vocal system is produced by the dynamic co-operation of the articulators. The theory of speech production is modeling speech as having two main components: the source and the filter. Both parts have dynamic behavior but in encoding the phonologic information into speech acoustics the filter component has the main dynamic role. The dynamics of the source is also an important factor because changes in the state and location of the source can also impose phonologic information into speech acoustics. Starting with the muscles acting upon the lungs to produce an air flow necessary to excite and pass the vocal and nasal tracts, continuing with those acting upon the glottis, larynx, tongue, velum and jaw, and finishing with those moving the lips and controlling the head position for speech direction, all these muscles have a co-ordinated activity controlled by the brain using nervous impulses transmitted through the spinal cord, spinal nerves and nerve fibers. The dynamics of the vocal apparatus is thus very complex and determined by the co-ordinated movements of its parts and their interaction in the three dimensional space. Each articulator or dynamic part of the vocal system is controlled by different muscles, some of them having antagonist behaviors. It is very difficult to model accurately the movement of all these dynamic parts due to the complexity of the system and their complicated interaction in producing speech gestures. In addition to the dynamic constraints the motion of the vocal system is also determined by the various physiological constraints due to the limited space of action and the interaction of its articulators.

In this study we were concerned about modeling the motion of those articulators which are the main vocal system parts responsible for encoding phonologic-linguistic identity into the speech acoustic signal. Thus, the jaw, lips, tongue, larynx and velum are of interest in this study and not those like lungs and neck which do not impose any phonetic information upon speech. Because the motion of the articulators is influenced by their mass, damping coefficients and stiffness coefficients it is appealing to use the second order differential equations to model the time evolution of the articulators. This is the motivation of choosing second-order equation in modeling the motion of the articulators. In this study however, the modeling has been done using directly the discrete-time difference equations instead of their equivalent continuous-time differential equations. Thus, we employed the linear second-order difference equation in which the state variables are driven by some inputs towards phonetic targets. We preferred to use this direct discrete time description of articulatory motion instead of modeling the articulators using continuous-time differential equation of spring-mass-damping systems, which would need explicitly a description of mass, damping and stiffness of each articulator and then converting the continuous-time model to a discrete-time model. The continuous-time description of each articulator would be difficult to obtain for the case of a statistical articulatory model, like that we are using in this study, or for modeling the articulatory system by direct measurements of some pellets placed on articulators. It would be more difficult to assess for the statistical linear components of the model or for those small points on articulators their true mass, damping and stiffness parameters and then to convert the continuous time equations into the equivalent discrete time equations.

A simple, second order model, capable to account for coarticulation phenomena in producing the articulatory gestures is formulated in the following. This model

corresponds to a coproduction segment as defined in the beginning of this section. The two discrete-time equations describing the state-space model of the dynamic articulatory system are the state equation and respectively the observation equation

$$\mathbf{x}(k+1) = \mathbf{\Phi}_1^{(\alpha,\beta)}\mathbf{x}(k) + \mathbf{\Phi}_2^{(\alpha,\beta)}\mathbf{x}(k-1) + \mathbf{w}(k) \tag{4.2}$$

$$\mathbf{y}(k) = \mathbf{h}^{(\alpha,\beta)}[\mathbf{x}(k)] + \mathbf{v}(k) \tag{4.3}$$

where k represents the discrete-time index, $\mathbf{x}$ is a $n \times 1$ vector representing the state variable of the articulators, $\mathbf{\Phi}_1^{(\alpha,\beta)}$ and $\mathbf{\Phi}_2^{(\alpha,\beta)}$ are $n \times n$ transition matrices, $\mathbf{y}$ is a $m \times 1$ vector representing the measurements or the observation variable, $\mathbf{h}^{(\alpha,\beta)}(\mathbf{x})$ is a $m \times 1$ non-linear observation function, $\mathbf{w}$ and $\mathbf{v}$ are $n \times 1$ respectively $m \times 1$ uncorrelated Gaussian white noise vectors with zero means and $n \times n$ respectively $m \times m$ covariance matrices $\mathbf{Q}^{(\alpha,\beta)}$ and $\mathbf{R}^{(\alpha,\beta)}$ defined as follows

$$E\{\mathbf{w}(i)\mathbf{w}^T(j)\} = \mathbf{Q}^{(\alpha,\beta)}(i)\delta_{ij} \tag{4.4}$$

$$E\{\mathbf{v}(i)\mathbf{v}^T(j)\} = \mathbf{R}^{(\alpha,\beta)}(i)\delta_{ij} \tag{4.5}$$

$$E\{\mathbf{w}(i)\mathbf{v}^T(j)\} = \mathbf{0} \tag{4.6}$$

where E is the expectation sign, δ_{ij} is the Kronecker delta and T is the transpose sign. These covariance matrices are symmetric and positive definite. The initial state $\mathbf{x}(0)$ is assumed to be independent of the random vectors $\mathbf{w}$ and $\mathbf{v}$ and to have a Normal distribution for a given model with the $n \times 1$ mean vector $\mu^{(\alpha,\beta)}$ and the $n \times n$ covariance matrix $\mathbf{\Sigma}^{(\alpha,\beta)}$. The main advantage of using the state-space model defined by the state and observation equations consists in accounting separately for modeling disturbances of the state and respectively the observation. This is done by using the two independent random variables $\mathbf{w}$ and $\mathbf{v}$.

In the above discrete state-space model equations the state variable $\mathbf{x}$ can represent the parameters which control an articulatory model or can represent the

position of some points or pellets situated on articulators and used to track their movements by electromagnetic articulography, cine-radiography or magnetic resonance imaging techniques. The acoustic-to-articulatory inversion method presented in this chapter can be applied to any data consisting of simultaneous articulatory and acoustic vectors and is not restricted to using an articulatory model to generate these data. Thus, pairs of articulatory and acoustic vectors generated by an articulatory-acoustic model or acoustic vectors recorded simultaneously with articulatory measurements from EMA, MRI or cine-radiography systems can be used. We consider that the state vector $\mathbf{x}$ and the noise vector $\mathbf{w}$ have the same dimension, n. Also the acoustic observation vector $\mathbf{y}$ and the corresponding noise vector $\mathbf{v}$ have the same dimension, m.

The linear state equation is modeling the dynamic constraints imposed to the evolution of the state of the articulators. These constraints are due to the equivalent mass inertia of the articulators and their damping and stiffness coefficients. In this linear equation the noise $\mathbf{w}$ is driving the articulatory state and it also models the errors in the evolution of the state variable. The linear state equation represents a simple approximation of the real, complicated motion of the articulators, yet powerful enough to model the main mechanical processes which take place into the human vocal system.

The second equation of the state-space model of the articulatory system, represents the non-linear measurement equation or the observation equation. The measurement variable $\mathbf{y}$ can be a parametric representation of the acoustic speech signal and it is considered the surface observation of the dynamic vocal system. The non-linear function $\mathbf{h}^{(\alpha,\beta)}(\mathbf{x})$ directly relates the observation variable $\mathbf{y}$ to the hidden articulatory state variable $\mathbf{x}$. If using an articulatory model this function approximates a chain of transformations. First the articulatory variables are trans-

formed into vocal-tract midsagittal distances, then these distances are transformed into cross-sectional areas and then these areas are used to solve the wave equations into the vocal and nasal tracts. Finally the pressure and volume velocity of the air flow are computed at the lips and nostrils and the acoustic pressure signal is then processed and the speech acoustic parameters are extracted using different methods like LPC or MFCC computation, formant extraction etc. In practice, the computation of the non-linear observation function is rather complicated and time consuming even though some simplifications and approximations are assumed. The non-linearity of the observation equation makes the distribution of the observation variable non-Gaussian, even though the noise variable $\mathbf{v}$ and the state variable are considered Gaussian. In order to apply the estimation methods based on extended Kalman filtering, a method of approximating the non-linear function $\mathbf{h}^{(\alpha,\beta)}(\mathbf{x})$, on piecewise linear regions is needed. This function $\mathbf{h}^{(\alpha,\beta)}(\mathbf{x})$ and its Jacobian matrix can be computed using numerical approximation methods. It can be linearized on small regions based on the Taylor series decomposition of a function. The Jacobian matrix is needed in the Kalman filtering techniques for computing the error covariance matrix and the Kalman gain. Using the first order terms of the expansion of the $\mathbf{h}^{(\alpha,\beta)}[\mathbf{x}(k)]$ function in a Taylor series about a reference value $\mathbf{x}_*(k)$ of the state variable we can write the output equation as follows

$$\mathbf{y}(k) = \mathbf{h}^{(\alpha,\beta)}(\mathbf{x}_*) + \mathbf{H}^{(\alpha,\beta)}(\mathbf{x}_*)[\mathbf{x}(k) - \mathbf{x}_*] + \mathbf{v}(k) \tag{4.7}$$

where $\mathbf{H}^{(\alpha,\beta)}(\mathbf{x}_*)$ is the Jacobian matrix of $\mathbf{h}^{(\alpha,\beta)}$ having the elements defined as the partial derivative of $\mathbf{h}^{(\alpha,\beta)}$ with respect to $\mathbf{x}$

$$\frac{\partial h_i^{(\alpha,\beta)}}{\partial x_j} = \frac{\partial h_i^{(\alpha,\beta)}[\mathbf{x}(t)]}{\partial x_j(t)}|_{\mathbf{x}(t)=\mathbf{x}_*(t)}, \tag{4.8}$$

where $h_i^{(\alpha,\beta)}$ is the ith element of vector $\mathbf{h}^{(\alpha,\beta)}$ and x_j is the jth element of vector $\mathbf{x}$. The speech inversion results of the previous studies based on Kalman filtering

were obtained using formant frequencies [89] and power spectrum components [105] as measurements or acoustic observation variable $\mathbf{y}$. In a preliminary stage of this study we used the formant frequencies as acoustic observation vectors [20], and then we extended the study by using a more general acoustic feature, the Mel-Frequency Cepstrum Coefficients (MFCCs) [21].

In the state equation, the transition matrices $\mathbf{\Phi}_1^{(\alpha,\beta)}$ and $\mathbf{\Phi}_2^{(\alpha,\beta)}$ are functions of the matrices describing the masses $\mathbf{M}$, damping coefficients $\mathbf{B}$ and stiffness coefficients $\mathbf{K}$, of the articulators as presented in the counterpart continuous differential equation from Chapter 2. The assumption used by other authors, [89], that the articulatory system is critically damped is not considered here, so we do not impose any relationship between $\mathbf{\Phi}_1^{(\alpha,\beta)}$ and $\mathbf{\Phi}_2^{(\alpha,\beta)}$ matrices.

The second-order state equation can be augmented to the first-order equation as follows

$$\begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{x}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_1^{(\alpha,\beta)} & \mathbf{\Phi}_2^{(\alpha,\beta)} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{x}(k-1) \end{bmatrix} + \begin{bmatrix} \mathbf{w}(k) \\ \mathbf{0} \end{bmatrix}. \tag{4.9}$$

The augmented state variable denoted by $\mathbf{z}$ becomes

$$\mathbf{z}(k) = \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{x}(k-1) \end{bmatrix}, \tag{4.10}$$

the augmented transition matrix $\mathbf{\Phi}^{(\alpha,\beta)}$ becomes

$$\mathbf{\Phi}^{(\alpha,\beta)} = \begin{bmatrix} \mathbf{\Phi}_1^{(\alpha,\beta)} & \mathbf{\Phi}_2^{(\alpha,\beta)} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{4.11}$$

and the augmented noise vector $\mathbf{w}_z$ becomes

$$\mathbf{w}_z(k) = \begin{bmatrix} \mathbf{w}(k) \\ \mathbf{0} \end{bmatrix}. \tag{4.12}$$

The augmented first-order state equation can be written in the simple form

$$\mathbf{z}(k+1) = \mathbf{\Phi}^{(\alpha,\beta)}\mathbf{z}(k) + \mathbf{w}_z(k). \tag{4.13}$$

where the augmented noise vector $\mathbf{w}_z$ has the covariance matrix $\mathbf{Q}_z^{(\alpha,\beta)}$ defined by the equation

$$\mathbf{Q}_z^{(\alpha,\beta)} = \begin{bmatrix} \mathbf{Q}^{(\alpha,\beta)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.14}$$

The dimension of the augmented state vector $\mathbf{z}$ is two times larger than that of the original state vector $\mathbf{x}$.

Using the augmented state variable $\mathbf{z}$, the corresponding observation equation becomes

$$\mathbf{y}(k) = \mathbf{g}^{(\alpha,\beta)}[\mathbf{z}(k)] + \mathbf{v}(k), \tag{4.15}$$

where $\mathbf{g}^{(\alpha,\beta)}[\mathbf{z}(k)]$ is the augmented non-linear observation function derived directly from the original observation function $\mathbf{h}^{(\alpha,\beta)}[\mathbf{x}(k)]$ through a simple transformation. It can be observed from this last equation that by augmenting the state variable the dimensions of the new function $\mathbf{g}^{(\alpha,\beta)}[\mathbf{z}]$, observation variable $\mathbf{y}$ and observation noise $\mathbf{v}$ remain the same. The corresponding augmented observation equation linearized on small regions using the Taylor series expansion becomes

$$\mathbf{y}(k) = \mathbf{g}^{(\alpha,\beta)}(\mathbf{z}_*) + \mathbf{G}^{(\alpha,\beta)}(\mathbf{z}_*)[\mathbf{z}(k) - \mathbf{z}_*] + \mathbf{v}(k) \tag{4.16}$$

where $\mathbf{G}^{(\alpha,\beta)}(\mathbf{z}_*)$ is the Jacobian matrix of $\mathbf{g}^{(\alpha,\beta)}[\mathbf{z}]$ computed at some reference points $\mathbf{z}_*$.

This last equation together with the augmented state equation describe the final dynamic model for which statistical methods of model parameter estimation and state estimation based on extended Kalman filtering and smoothing have been applied.

In this study we implemented and analyzed a second version of the dynamical model, which is based on piecewise constant targets. This target model was first applied to estimation of articulatory trajectories by Shirai and Honda [89], and later extended to a stochastic target model in estimating articulatory trajectories for automatic speech recognition by Ramsay and Deng [75]. We implemented this version in order to evaluate it and we found that it is not very accurate in estimating articulatory trajectories, even though, for automatic speech recognition purposes where an accurate estimation of articulatory trajectories is not very important, this target model might be useful.

For this target model, the second-order state equation becomes

$$\mathbf{x}(k+1) = \mathbf{\Phi}_1^{(\alpha,\beta)}\mathbf{x}(k) + \mathbf{\Phi}_2^{(\alpha,\beta)}\mathbf{x}(k-1) + \mathbf{\Psi}^{(\alpha,\beta)}\mathbf{u}(k) + \mathbf{w}(k). \tag{4.17}$$

where $\mathbf{u}$ is the target variable, constant over a certain segment of speech, as described by the equation

$$\mathbf{u}(k+1) = \mathbf{u}(k) \tag{4.18}$$

for a speech segment from $k = l_i$ to $k = l_j$ defined by the onset of an articulatory gesture. The control, or target transformation matrix $\mathbf{\Psi}$, has to be constrained by the equilibrium equation at which the state reaches the target, that is

$$\mathbf{x}(k+1) = \mathbf{x}(k) = \mathbf{x}(k-1) = \mathbf{u}(k) \tag{4.19}$$

From this condition the constraint on the $\mathbf{\Psi}^{(\alpha,\beta)}$ matrix can be easily obtained as

$$\mathbf{\Psi}^{(\alpha,\beta)} = \mathbf{I} - \mathbf{\Phi}_1^{(\alpha,\beta)} - \mathbf{\Phi}_2^{(\alpha,\beta)}. \tag{4.20}$$

where $\mathbf{I}$ is the identity matrix.

Like the state variable $\mathbf{x}(k)$, the control input $\mathbf{u}(k)$ is unobservable. Imposing different constraints on the evolution of this control input can affect the accuracy

of state estimation due to underlying assumptions regarding the control input or target, $\mathbf{u}(k)$. The assumption that this control input or target is constant over an interval corresponding to a phonologic gesture, as used by Shirai and Honda [89], and Ramsay and Deng [75], leads to constraining the approximation of the articulatory trajectories by exponential functions, which asymptotically approach the targets. Studies of articulatory motion and modeling have shown that the exponential time function cannot fit accurately the movements of the articulators which, instead, are better approximated by sinusoidal functions (Browman and Goldstein [8]; Kröger *et al.*, [42]). On the other hand, the actual control input of the articulators is affected by the complex neuro-chemical processes which take place in the human body. Thus, it is difficult to approximate the real shape of the control input. However, it is not certain that the shape of the control input could be estimated accurately from the speech acoustic signal alone. Our experiments have shown that imposing this constraint of step target affects the accuracy of articulatory trajectory estimation. If, for example, in applications like articulatory based speech recognition, such a step target could be meaningful and useful, it could be determined by other simpler methods from the estimated articulatory trajectory. Another drawback of the target model is that these targets segment the speech globally into units in which the overlapping of articulatory gestures is not allowed. That is, there is no single pair of boundaries delimiting a speech unit common to all articulatory variables. In reality, each articulatory variable has its distinct set of boundaries corresponding to an activation interval of a gesture. In a Kalman filtering approach of estimating articulatory trajectories, the individual boundaries corresponding to each articulatory variable are difficult to obtain. Theoretically, these individual boundaries could be obtained if a number of separate Kalman filters and smoothers would be used, each of them corresponding to an articulatory

variable. Unfortunately, this approach is not practical.

The second-order state equation including the target can be transformed into a first order equation by augmenting the state variable as follows

$$\begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{x}(k) \\ \mathbf{u}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_1^{(\alpha,\beta)} & \mathbf{\Phi}_2^{(\alpha,\beta)} & \mathbf{\Psi}^{(\alpha,\beta)} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{x}(k-1) \\ \mathbf{u}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{w}(k) \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad (4.21)$$

where $\mathbf{I}$ is an $n \times n$ identity matrix and $\mathbf{0}$ is an $n \times n$ matrix, or an n dimensional vector in the last column, with all elements equal to zero.

Denoting the augmented state variable by $\mathbf{z}$

$$\mathbf{z}(k) = \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{x}(k-1) \\ \mathbf{u}(k) \end{bmatrix}, \quad (4.22)$$

the augmented transition matrix by $\mathbf{\Phi}^{(\alpha,\beta)}$

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1^{(\alpha,\beta)} & \mathbf{\Phi}_2^{(\alpha,\beta)} & \mathbf{\Psi}^{(\alpha,\beta)} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}. \quad (4.23)$$

and the augmented noise vector by $\mathbf{w}_z$

$$\mathbf{w}_z(k) = \begin{bmatrix} \mathbf{w}(k) \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (4.24)$$

we can write the state equation in a simple augmented form as follows

$$\mathbf{z}(k+1) = \mathbf{\Phi}^{(\alpha,\beta)}\mathbf{z}(k) + \mathbf{w}_z(k). \quad (4.25)$$

The augmented noise vector has its covariance matrix $\mathbf{Q}_z^{(\alpha,\beta)}$ defined by the equation

$$\mathbf{Q}_z^{(\alpha,\beta)} = \begin{bmatrix} \mathbf{Q}^{(\alpha,\beta)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.26}$$

The dimension of the augmented state vector $\mathbf{z}$ is three times larger than that of the original state vector $\mathbf{x}$, and thus the whole estimation method is computationally more expensive.

4.2 Articulatory-Acoustic Function

The acoustic-to-articulatory inversion described in this thesis is based on training pairs of articulatory and acoustic trajectories. These training data can be either synthesized or measured from real subjects. The articulatory-acoustic nonlinear sub-functions are approximated from the corresponding training utterances of each phonological coproduction model. In order to prepare the data for these approximations, the continuous training utterances have to be first segmented and labeled. This can be carried out manually or by employing an automatic segmentation method based on the acoustic parameters (Ljolje and Riley [52]). After segmentation and labeling, all the tokens corresponding to each coproduction phonological unit were used to piecewise linearly approximate the corresponding sub-function.

The articulatory synthesis function, which relates the speech acoustic parameters to the corresponding articulatory parameters, requires a large computational cost. In addition, the numerical computation of the Jacobian matrix of this function decreases the speed of the overall speech inversion method. Hence, methods of approximating the articulatory-acoustic function and numerical computation of the

Jacobian matrix are desirable. In this section, a method used for approximating the articulatory-acoustic non-linear functions is presented. This method was experimented in order to carry out this approximation with piecewise linear functions in small regions. For each small linear region the mean acoustic and articulatory vectors and the Jacobian matrix which linearly relates the two spaces are needed in the process of extended Kalman filtering and smoothing.

The direct transformation from the articulatory space to the acoustic space represents the process of speech production or the synthesis and can be, in general, described as a non-linear multivariate function $\mathbf{g}$ of a multivariate argument

$$\mathbf{y} = \mathbf{g}(\mathbf{z}), \tag{4.27}$$

where $\mathbf{y}$ represents the acoustic vector and $\mathbf{z}$ represents the articulatory vector. These vectors can have various dimensions according to what exactly they represent in the acoustic and articulatory spaces. To represent the speech acoustics one can use different parameters derived from the speech acoustic signals, e.g. formant frequencies, FFT coefficients, LPC coefficients, MFCC parameters etc. The articulatory vectors can also be represented by different parameters, e.g. articulatory model parameters, measurements of pellets on different articulators, area functions of the vocal-tract, etc.

This speech production or synthesis function described by Equation 4.27 involves a high computational cost even though a very simplified production model is employed. It is more convenient therefore, to approximate this non-linear function with piecewise linear functions on small regions on which the linearity assumption holds. In addition, in our approach of speech inversion, this linearization is necessary in order to transform the dynamical model with non-linear observations into a linear perturbation model with linear observation matrix (Jacobian matrix). The

linearization of the measurement or observation function can be done using large numbers of pairs of articulatory and acoustic vectors. The creation of the collection of $(\mathbf{z}, \mathbf{y})$ pairs of vectors for many different articulatory configurations can be made by sampling the articulatory space of an articulatory model (Atal *et al.*, [1]), or by acquiring real human articulatory and acoustic data (Hogden *et al.*, [34]; Suzuki *et al.*, [99]). The former method has the advantage of simplicity but it is very difficult to synthesize accurate articulatory trajectories close to real ones. The latter method has the advantage that it does not contain the inherent approximations introduced by the use of articulatory-acoustic models. A disadvantage of this method is that it is difficult in practice to obtain large articulatory and acoustic database for all classes of speech sounds and many speakers. In this study of speech inversion we experimented the approximation of the articulatory-to-acoustic function using both methods of generating and acquiring articulatory and acoustic data.

The linearization of the articulatory-to-acoustic function in small regions has been used first by Shirai and Honda [89] and Atal *et al.*, [1] for the purpose of inversion of this transformation. Thus, for small regions around some reference points defined by $\mathbf{z}_*$ and $\mathbf{y}_* = \mathbf{g}(\mathbf{z}_*)$ in the combined articulatory-acoustic space they approximated $\mathbf{y}$ as follows

$$\mathbf{y} \simeq \mathbf{y}_* + \mathbf{G}(\mathbf{z} - \mathbf{z}_*), \tag{4.28}$$

where $\mathbf{G}$ was the matrix of partial derivatives of $\mathbf{g}(\mathbf{z})$. This equation represents only the first order approximation of the Taylor series expansion of the non-linear function because it neglects the high order terms. Because the articulatory-to-acoustic transformation is computationally very expensive, in practice, the elements of the Jacobian matrix $\mathbf{G}$ are computed numerically by replacing the partial derivatives with partial differences.

One of the most important innovations proposed in this study consists in dividing the whole observation function into a number of sub-functions, each corresponding to a phonological coproduction model, as defined in the beginning of this chapter. In this research we propose a method to divide the whole non-linear observation function into a number of $\mathcal{N}_C$ coproduction non-linear sub-functions, where $\mathcal{N}_C \simeq \mathcal{N}_P \times \mathcal{N}_P$ and $\mathcal{N}_P$ represents the number of phonetic units of a language. This total number of coproduction models $\mathcal{N}_C$ is not perfectly equal to the square of the number of phonetic units $\mathcal{N}_P^2$ because of the linguistic and phonological constraints of the language. Each sub-division of the whole observation function belongs to a dynamic model (α, β), indexed from $1 to \mathcal{N}_C$. Thus, in the above linearized equation, the parameters become dependent of the model (α, β)

$$\mathbf{y} \simeq \mathbf{g}_{\bullet}^{(\alpha,\beta)} + \mathbf{G}_{\bullet}^{(\alpha,\beta)}[\mathbf{z} - \mathbf{z}_{\bullet}^{(\alpha,\beta)}], \tag{4.29}$$

where $\mathbf{g}_{\bullet}^{(\alpha,\beta)} = \mathbf{g}^{(\alpha,\beta)}(\mathbf{z}_{\bullet}^{(\alpha,\beta)})$ and $\mathbf{G}_{\bullet}^{(\alpha,\beta)} = \mathbf{G}^{(\alpha,\beta)}(\mathbf{z}_{\bullet}^{(\alpha,\beta)})$.

The method of approximating the non-linear function, experimented in this study, was based on articulatory-acoustic codebooks. An alternative method of approximating the non-linear observation function is presented in Appendix A, and is based on neural networks. This alternative method is more universal and is suitable to approximate any nonlinear articulatory-acoustic sub-function corresponding to a phonological coproduction model.

4.2.1 Approximating the Articulatory-Acoustic Function by Codebooks

The codebooks or look-up tables were first used for approximating the articulatory-acoustic function by Atal *et al.*, [1] and latter have been adopted by several authors

(Schroeter *et al.*, [84]; Schroeter and Sondhi [86] and [85]; Ramsay and Deng [75]; Hogden *et al.*, [34]; Dusan and Deng [20]). Because the exhaustive search of large articulatory-acoustic codebooks is time consuming, it can be simplified if some vector quantization methods are employed (Gray [31]; Linde *et al.*, [51]; Larar *et al.*, [47]).

By linearly approximating the non-linear observation sub-function, corresponding to phonological coproduction models (α, β), we constructed codebooks $\mathcal{C}^{(\alpha,\beta)}$ containing a number $\mathcal{N}_c^{(\alpha,\beta)}$ of sets of linear parameters $\mathcal{S}_i^{(\alpha,\beta)}[\mathbf{z}_\bullet^{(\alpha,\beta)}, \mathbf{g}_\bullet^{(\alpha,\beta)}, \mathbf{G}_\bullet^{(\alpha,\beta)}]$, representing the mean articulatory vector, mean acoustic vector and respectively the Jacobian matrix of each small linear region $i = 1, 2, ..., \mathcal{N}_c^{(\alpha,\beta)}$, included in the model. This kind of sets of triple articulatory-acoustic parameters has been proposed by Ramsay and Deng [75], but in that study they were not associated to phonological models but to binary-derived or 'cubic' regions without any phonologic relationship.

In this subsection, a method of creating the codebooks $\mathcal{C}^{(\alpha,\beta)}$ from training pairs of articulatory and acoustic vectors from a reference speaker is presented. The clustering of data into small regions is carried out by using self organized maps neural networks (SOM-NN). The linearization of these regions is accomplished by a multiple linear regression method. First, we assume that we have a collection of labeled sequences of pairs of articulatory-acoustic vectors for each phonological coproduction model (α, β). These training sequences can either be synthesized or acquired from a reference speaker and should contain realizations of the corresponding coproduction units using different speed and stress parameters. Thus, the articulatory-acoustic data collected from a single speaker (or reference model) can only have a variability due to some dynamical and linguistic factors like the manner of articulation and the stress and speed parameters. Eliminating the coarticula-

tion variability the phonological coproduction model represents different transition prototypes of producing the $/\alpha\beta/$ phonologic sequence. An example of such a collection of trajectories for a model $(\alpha, \beta) = (eh, iy)$ containing 20 trajectories generated with the Maeda's articulatory-acoustic model is represented in figure 4.3. For clarity, the trajectories are displayed by scattered small dots rather than continuous plots. The non-linearity of the observation function $\mathbf{g}^{(\alpha,\beta)}$ can be seen in many of the sub-plots. However, for this model the non-linearity is not very large. The time dimension has been eliminated by displaying in a separate sub-plot the $y_i - z_j$ elements of the acoustic vectors $\mathbf{y}$, represented by 10 MFCC parameters, and articulatory vectors $\mathbf{z}$, represented by 8 Maeda's model parameters.

In creating a codebook $C^{(\alpha,\beta)}$, a method of clustering the articulatory-acoustic data corresponding to the (α, β) model has to be used. The clustering method divides this data into small regions on which the linearity assumption holds. In our experiments we varied the number $\mathcal{N}_c^{(\alpha,\beta)}$ of the piecewise linear regions included in a codebook $C^{(\alpha,\beta)}$ between 3 and 50. A basic number of $\mathcal{N}_c^{(\alpha\beta)} = 10$ has been successfully used in most of our experiments.

We developed an efficient method of clustering the articulatory-acoustic data of each model using neural networks of the self-organizing map (SOM) type. These neural networks are competitive layers and are often used for classification. By applying the SOM technique a predefined number of clusters are determined from the data and for each cluster the weights (or centers) of the regions in the multidimensional acoustic-articulatory space are computed. In clustering the space of a model we used the combined space of articulatory and acoustic dimensions. For the data displayed in figure 4.3, corresponding to a model (α, β) for the /eh iy/ segment, a clustering example using $\mathcal{N}_c^{(\alpha,\beta)} = 7$ is presented in this figure by marking the centers of the clusters with small circles. The centers of these circles are the pro-

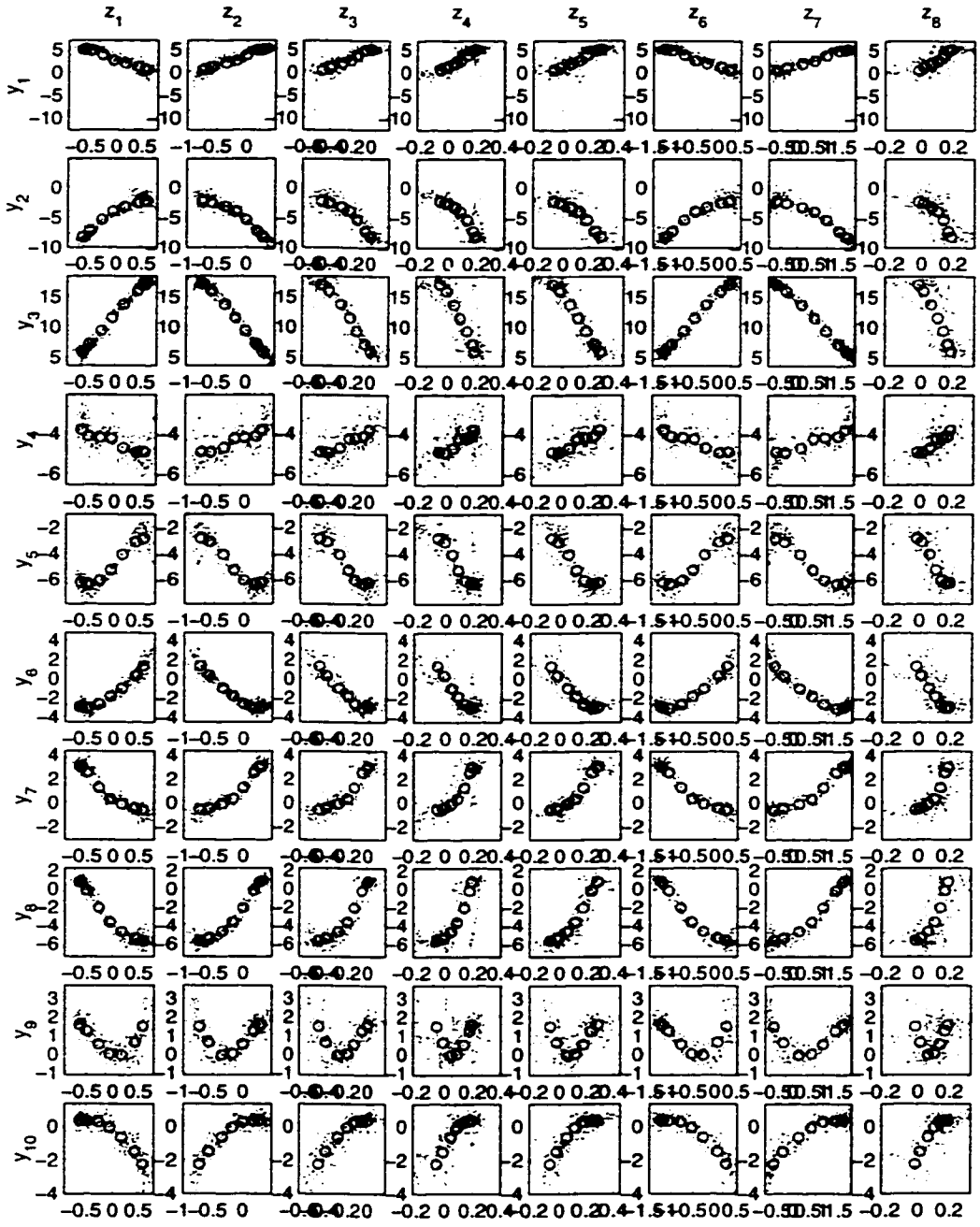

Figure 4.3: Scatter plots of 20 /eh iy/ sequences synthesized with the Maeda's models

jections of the mean cluster vectors $\mathbf{z}_{\bullet}^{(\alpha,\beta)}$ and $\mathbf{g}_{\bullet}^{(\alpha,\beta)}$ on each of the two dimensional subspaces $y_i - z_j$, where $i = 1, 2, ..., 10$ and $j = 1, 2, ..., 8$. After computing the centers or the mean values of each cluster a method of linearization based on multiple linear regression using least squares has been used. The goal of linearization in each cluster consists in finding a matrix which approximates the Jacobian matrix of the $\mathbf{g}^{(\alpha,\beta)}$ non-linear function computed at the point represented by the center of the cluster. In evaluating the accuracy of the linear approximation of the $\mathbf{g}^{(\alpha,\beta)}$ function for each cluster the errors of approximation are computed for all the data samples of a model. If the average error for all the data samples is greater than a predefined threshold then a higher number of clusters will be used, until the average error is lower than the threshold. The whole algorithm of clustering and linearization has been implemented iteratively using the MATLAB's statistics tool-box. Finally, for each phonological coproduction model (α, β), a codebook $C^{(\alpha,\beta)}$ containing a list or a look-up table of $\mathcal{N}_c^{(\alpha,\beta)}$ sets of parameters $\mathcal{S}_i^{(\alpha,\beta)}[\mathbf{z}_{\bullet}^{(\alpha,\beta)}, \mathbf{g}_{\bullet}^{(\alpha,\beta)}, \mathbf{G}_{\bullet}^{(\alpha,\beta)}]$, where $i = 1, 2, ..., \mathcal{N}_c^{(\alpha,\beta)}$, is created.

In the extended Kalman filtering, for finding the closest linear region corresponding to an arbitrary articulatory vector $\hat{\mathbf{z}}(k|k-1)$, a search based on the Euclidean distances between this new vector and the articulatory mean of each cluster has been used. Thus, the closest cluster, or linear region, will provide the most accurate approximation of the non-linear function $\mathbf{g}^{(\alpha,\beta)}$ at the point $\hat{\mathbf{z}}(k|k-1)$.

The sequence of articulatory-acoustic pairs of vectors can be either synthesized or acquired from human measurements. However, if the training articulatory-acoustic data of a model (α, β) contains a very small number of sequences, the above method is not accurate in approximating the non-linear function $\mathbf{g}^{(\alpha,\beta)}$ in small regions. This is due to the limitations of the multiple regression method employed. For this case, we developed a method of generating random sequences

of articulatory-acoustic training data from as little as a single original sequence corresponding to a model. This method is based on approximating the original training sequence (or sequences) with polynomial functions on all the sub-spaces represented by the combinations of any element of **z** with any element of **y**. This approximation is needed because the acoustic parameters are affected by noise. An example of approximating a single sequence, /aa sh/, is illustrated in Figure 4.4. The display is in a form of articulatory-acoustic sub-spaces, like in the previous figure. The original /aa sh/ sequence has been approximated on each subspace with a third-order polynomial function which is displayed with continuous lines, whereas the original sequence is displayed with dotted lines. After computing all the polynomial function coefficients, based on the method of polynomial fitting, a number of random sequences can be generated artificially from these polynomials by adding a Gaussian noise to the main polynomial trajectories. Figure 4.5 displays the generated data of 20 /aa sh/ trajectories, plotted with small dots, and the polynomial functions, plotted with continuous lines.

The random noise used to generate these trajectories should not be large enough to change the phonetic identity of each of the α and β phonemes. Then these new sequences can be used together with the original one (or ones) to do the clustering and linearization of the corresponding model. Our experiments have shown that this method could be successfully used as a solution when very few training sequences for a model are available.

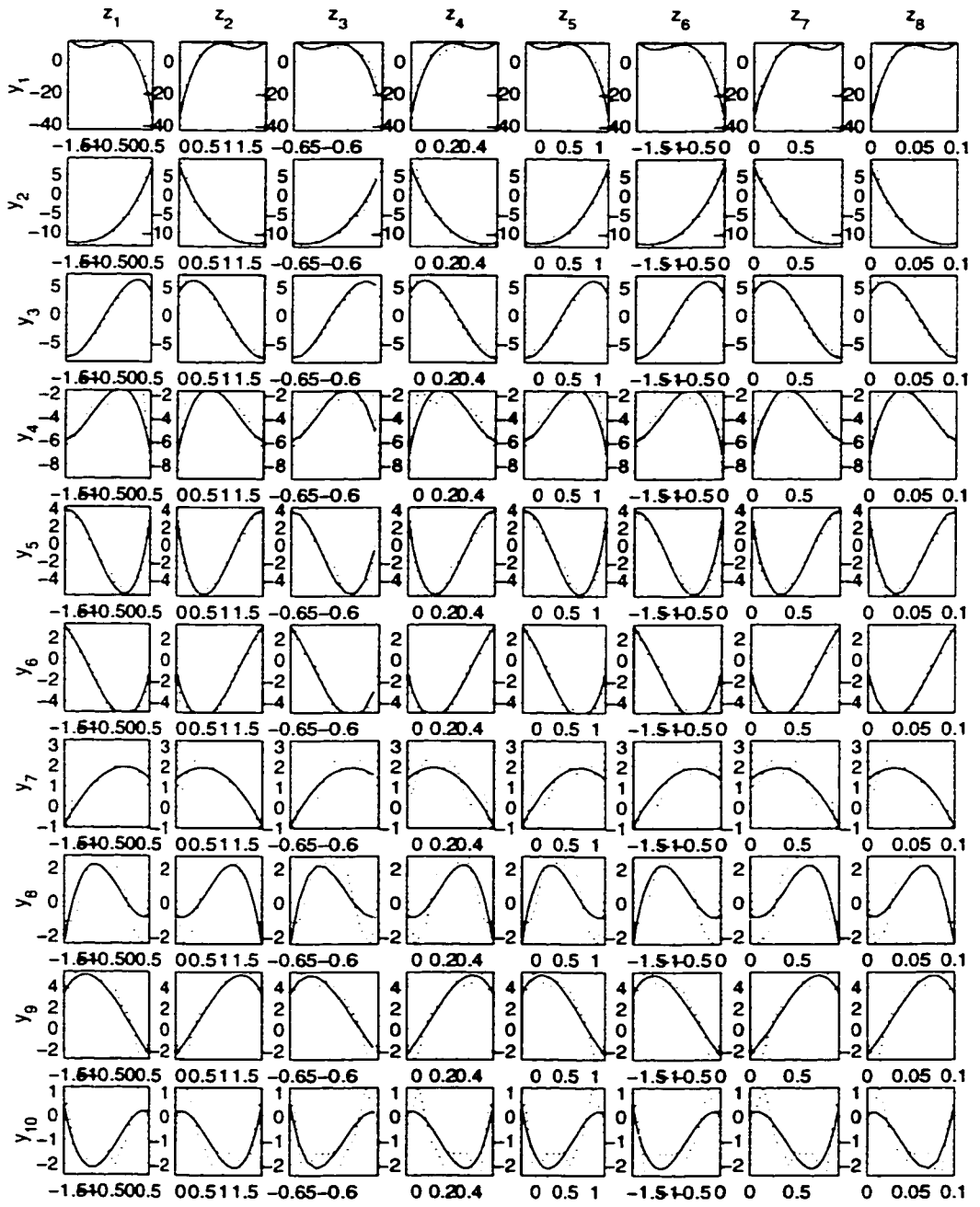

Figure 4.4: Approximation of a single /aa sh/ sequence with third-order polynomial functions (y_i = MFCC parameter, z_j = articulatory parameter)

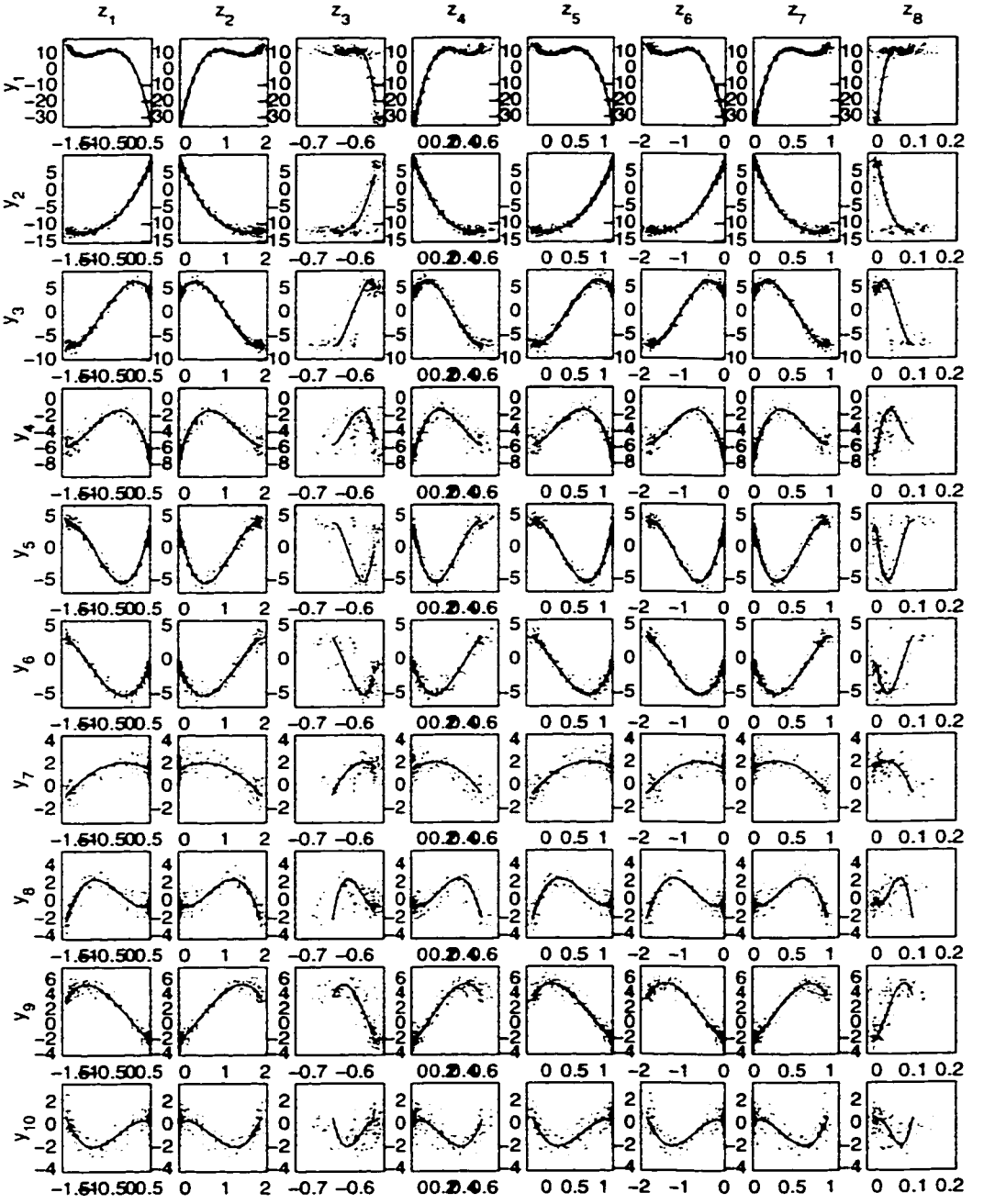

Figure 4.5: Generated data for 20 /aa sh/ sequences using third-order polynomial functions (y_i = MFCC parameter, z_j = articulatory parameter)

4.3 Maximum-Likelihood Estimation of Model Parameters using Articulatory-Acoustic Training Data

In this study we implemented a simple method of estimation of model parameters using the maximum-likelihood method and based on the training utterances. Since the estimation of articulatory trajectories used in this study is based on continuous speech training utterances, consisting of pairs of articulatory and acoustic vectors, the estimation of model parameters can be considered a supervised training process. The articulatory-acoustic training data has to be pre-segmented and labeled before the model parameter estimation starts. This segmentation and labeling can be made manually or, an automatic segmentation and labeling of the speech utterances can be employed. Such an automatic segmentation method based on the acoustic parameters alone can provide the phones and their boundaries using a phone recognition system (Ljolje and Riley [52]). From the phone transcription and boundaries, the diphone units and their boundaries can be easily obtained. The Maximum-Likelihood (ML) estimation is a statistical method of estimation based on finding an optimum parameter set which maximize a likelihood objective function. When dealing with exponential distributions, e.g., Normal distribution, the logarithm of the likelihood function is often used as the objective function. A derivation of the maximum-likelihood estimation of the parameters for a multivariate Normal distribution can be found in [30].

In this section, for the clarity of the equations, we dropped the index (α, β) of all the parameters corresponding to a coproduction model.

The estimation of the dynamic articulatory system model parameters $\theta =$

$\{\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R}\}$ has been done in this study using the a direct form of the ML method. We also implemented an alternative method of model parameter estimation, from acoustic parameters alone, using the Expectation-Maximization algorithm, [15].

In the following of this section, we derive the ML parameter estimation method and provide details of our implementation. For a particular model, described by the state space equations, the ML estimation is based on maximizing an objective function consisting of the likelihood of the complete articulatory and acoustic data $L(\mathbf{Z}, \mathbf{Y})$.

For a state-space model, corresponding to a particular phonologic articulatory gesture onset interval as defined by Browman and Goldstein in [6] and Kröger *et al.*, in [42], we can consider the system to be time-invariant. This means that, for a given interval specifying the onset of an articulatory gesture, we have a model described by the state-space equations in which the parameter set θ do not change its value and we want to estimate it from the sequences of observations, $\mathbf{Y} = [\mathbf{y}(1)\mathbf{y}(2)...\mathbf{y}(N)]$ and $\mathbf{Z} = [\mathbf{z}(1)\mathbf{z}(2)...\mathbf{z}(N)]$ corresponding to a coproduction segment. In the state-space model equations of the dynamical system, the two noise processes $\mathbf{w}_z(k)$ and $\mathbf{v}(k)$ have Normal distributions with assumed zero means and covariance matrices $\mathbf{Q}_z$ and $\mathbf{R}$

$$p(\mathbf{w}_z) \sim N(\mathbf{0}, \mathbf{Q}_z), \tag{4.30}$$

$$p(\mathbf{v}) \sim N(\mathbf{0}, \mathbf{R}). \tag{4.31}$$

Because $\mathbf{w}_z(k)$ and $\mathbf{v}(k)$ are assumed independent each other, their joint likelihood equals their probability product

$$L[\mathbf{w}_z, \mathbf{v}] = \prod_{k=1}^{N} p[\mathbf{w}_z(k)] \prod_{k=1}^{N} p[\mathbf{v}(k)], \tag{4.32}$$

for a sequence of N independent identically distributed (i.i.d.) samples of these

distributions. The logarithm of the joint likelihood function L is

$$\begin{aligned} logL[\mathbf{w}_z, \mathbf{v}] &= log\{\prod_{k=1}^{N} p[\mathbf{w}_z(k)] \prod_{k=1}^{N} p[\mathbf{v}(k)]\} \\ &= log \prod_{k=1}^{N} p[\mathbf{w}_z(k)] + log \prod_{k=1}^{N} p[\mathbf{v}(k)] \\ &= -\frac{N}{2} log|\mathbf{Q}_z| - \frac{1}{2}\sum_{k=1}^{N}[\mathbf{w}_z(k)]^T \mathbf{Q}_z^{-1}[\mathbf{w}_z(k)] \\ &\quad -\frac{N}{2} log|\mathbf{R}| - \frac{1}{2}\sum_{k=1}^{N}[\mathbf{v}(k)]^T \mathbf{R}^{-1}[\mathbf{v}(k)] \\ &\quad +constant. \end{aligned} \tag{4.33}$$

In order to estimate the parameter set $\theta = \{\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R}\}$ of a phonological coproduction model from an observation sequence using the maximum likelihood method we need the objective function defined by the joint log-likelihood of the complete data. This objective function $J(\mathbf{Z}, \mathbf{Y}, \theta)$ is defined by the log-likelihood of the state and observation variables and can be obtained using the above equation by substituting the two random variables $\mathbf{w}_z(k)$ and $\mathbf{v}(k)$ using the augmented state and observation equations

$$\begin{aligned} J(\mathbf{Z}, \mathbf{Y}, \theta) &= log\{L(\mathbf{Z}, \mathbf{Y}), \theta\} \\ &= -\frac{N}{2} log|\mathbf{Q}_z| - \frac{1}{2}\sum_{k=1}^{N}[\mathbf{z}(k) - \mathbf{\Phi}\mathbf{z}(k-1)]^T \mathbf{Q}_z^{-1}[\mathbf{z}(k) - \mathbf{\Phi}\mathbf{z}(k-1)] \\ &\quad -\frac{N}{2} log|\mathbf{R}| - \frac{1}{2}\sum_{k=1}^{N}[\mathbf{y}(k) - \mathbf{g}[\mathbf{z}(k)]]^T \mathbf{R}^{-1}[\mathbf{y}(k) - \mathbf{g}[\mathbf{z}(k)]], \end{aligned} \tag{4.34}$$

where we omitted the constant term from the previous equation. This objective function can be further written as follows

$$\begin{aligned} J(\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R}) &= logL(\mathbf{Z}, \mathbf{Y}, \theta) \\ &= -\frac{1}{2} tr\{\mathbf{Q}^{-1}[\mathbf{S}_3 - \mathbf{S}_2\mathbf{\Phi}^T - \mathbf{\Phi}\mathbf{S}_2^T + \mathbf{\Phi}\mathbf{S}_1\mathbf{\Phi}^T]\} \end{aligned}$$

$$-\frac{1}{2}tr\{\mathbf{R}^{-1}[\mathbf{S}_4 - \mathbf{S}_5 - \mathbf{S}_5^T + \mathbf{S}_6]\}$$
$$-\frac{N}{2}log|\mathbf{Q}_z| - \frac{N}{2}log|\mathbf{R}|, \tag{4.35}$$

where tr represents the **trace** (or the **spur**) of the matrices and $\mathbf{S}_1$, ..., $\mathbf{S}_6$ are the sufficient statistics defined by the following equations

$$\mathbf{S}_1 = \sum_{k=1}^{N} \mathbf{z}(k-1)\mathbf{z}(k-1)^T, \tag{4.36}$$

$$\mathbf{S}_2 = \sum_{k=1}^{N} \mathbf{z}(k)\mathbf{z}(k-1)^T, \tag{4.37}$$

$$\mathbf{S}_3 = \sum_{k=1}^{N} \mathbf{z}(k)\mathbf{z}(k)^T, \tag{4.38}$$

$$\mathbf{S}_4 = \sum_{k=1}^{N} \mathbf{y}(k)\mathbf{y}(k)^T, \tag{4.39}$$

$$\mathbf{S}_5 = \sum_{k=1}^{N} \mathbf{y}(k)\mathbf{g}(\mathbf{z}(k))^T, \tag{4.40}$$

$$\mathbf{S}_6 = \sum_{k=1}^{N} \mathbf{g}(\mathbf{z}(k))\mathbf{g}(\mathbf{z}(k))^T, \tag{4.41}$$

where $\mathbf{g}(\mathbf{z}(k))$ are computed using the articulatory-acoustic codebook or the neural network for the corresponding model sub-function.

In order to estimate the model parameters, we have to compute the derivative of the objective function with respect to each of the model parameters, make the derivative equal to zero, and solve these equations

$$\frac{\partial J(\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R})}{\partial \mathbf{P}} = 0 \tag{4.42}$$

where $\mathbf{P}$ is a matrix and represents each of the model parameters $\mathbf{\Phi}$, $\mathbf{Q}_z$, $\mathbf{R}$. In taking the derivatives of the scalar function $J(\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R})$ with respect to the matrix parameters $\mathbf{P}$ we used some rules of matrix algebra, as presented in [30]. Applying

this maximization technique the set of model parameters can be estimated as

$$\hat{\mathbf{\Phi}} = \mathbf{S}_2\mathbf{S}_1^{-1}, \tag{4.43}$$

$$\hat{\mathbf{Q}}_z = \frac{1}{N}(\mathbf{S}_3 - \mathbf{S}_2\mathbf{S}_1^{-1}\mathbf{S}_2^T), \tag{4.44}$$

$$\hat{\mathbf{R}} = \frac{1}{N}(\mathbf{S}_4 - \mathbf{S}_5 - \mathbf{S}_5^T + \mathbf{S}_6), \tag{4.45}$$

where N represents the number of samples and $\mathbf{S}_1$, ..., $\mathbf{S}_6$ are computed using the acoustic and articulatory samples from the training utterances.

In addition to the set of parameters $\theta = \{\mathbf{\Phi}, \mathbf{Q}_z, \mathbf{R}\}$ estimated above, the extended Kalman filter needs other two parameters for initialization — the mean and the covariance matrix of the initial state for each model. The initial state $\mathbf{z}(0)$ of any observation sequence corresponding to a model is assumed to have also a Normal distribution with mean μ and covariance matrix $\mathbf{\Sigma}$

$$p[\mathbf{z}(0)] \sim N(\mu, \mathbf{\Sigma}). \tag{4.46}$$

For a single observation sequence with 1 to N observations, the mean and covariance matrix of the initial state can be simply chosen as

$$\hat{\mu} = \mathbf{z}(1), \tag{4.47}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}(1), \tag{4.48}$$

where $\mathbf{z}(1)$ is the first articulatory state in the training sequence and $\mathbf{\Sigma}(1)$ is a covariance matrix which can only be experimentally chosen.

If we estimate the parameters of a model from multiple observation sequences, the extension of the above single sequence algorithm to this case is straightforward and can be done by extending the summations in the computation of the statistics $\mathbf{S}_1$, ..., $\mathbf{S}_6$ to all the observation sequences. Thus, in the case of M training observation sequences, each of them having a particular length $N_{(m)}$ the corresponding

statistics become

$$\mathbf{S}_1 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{z}(k-1)\mathbf{z}(k-1)^T, \tag{4.49}$$

$$\mathbf{S}_2 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{z}(k)\mathbf{z}(k-1)^T, \tag{4.50}$$

$$\mathbf{S}_3 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{z}(k)\mathbf{z}(k)^T, \tag{4.51}$$

$$\mathbf{S}_4 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{y}(k)\mathbf{y}(k)^T, \tag{4.52}$$

$$\mathbf{S}_5 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{y}(k)\mathbf{g}(\mathbf{z}(k))^T, \tag{4.53}$$

$$\mathbf{S}_6 = \sum_{m=1}^{M} \sum_{k=1}^{N_{(m)}} \mathbf{g}(\mathbf{z}(k))\mathbf{g}(\mathbf{z}(k))^T. \tag{4.54}$$

The estimated model parameters computed for the case of multiple training observation sequences are

$$\hat{\mathbf{\Phi}} = \mathbf{S}_2\mathbf{S}_1^{-1}, \tag{4.55}$$

$$\hat{\mathbf{Q}}_z = \frac{1}{\sum_{m=1}^{M} N_{(m)}}(\mathbf{S}_3 - \mathbf{S}_2\mathbf{S}_1^{-1}\mathbf{S}_2^T), \tag{4.56}$$

$$\hat{\mathbf{R}} = \frac{1}{\sum_{m=1}^{M} N_{(m)}}(\mathbf{S}_4 - \mathbf{S}_5 - \mathbf{S}_5^T + \mathbf{S}_6). \tag{4.57}$$

For the case of multiple observation sequences, the mean and covariance matrix of the initial state can be computed as

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{z}_{(m)}(1), \tag{4.58}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{M-1}\sum_{m=1}^{M} [\mathbf{z}_{(m)}(1) - \hat{\mu}][\mathbf{z}_{(m)}(1) - \hat{\mu}]^T, \tag{4.59}$$

where $\mathbf{z}_{(m)}(1)$ is the first articulatory state in the training sequence number m.

We successfully implemented this method of model parameter estimation based on the articulatory and acoustic training utterances. We also implemented the

Expectation-Maximization algorithm for ML parameter estimation which uses the acoustic observation alone. As expected, the estimated parameters using the EM algorithm are not as accurate as those obtained from the direct ML method, but they are close.

4.4 Articulatory State Estimation Using the Extended Kalman Filtering and Smoothing

The statistical method of state estimation, which is at the core of the acoustic-to-articulatory inversion used in this study, is based on the extension of Kalman filtering and smoothing to non-linear systems. Before the presentation of the extended Kalman filtering (EKF) and smoothing methods, which will be done in detail later in this section, the basics of the Kalman filtering and smoothing applied to a generic linear dynamical model will be described in the following. In this section, for the convenience of a simplified notation, we dropped the index (α, β) of the model for all the parameters corresponding to a phonological coproduction model.

The Kalman filtering, predicting and smoothing are mean-squared estimation methods. The mean-squared estimation is a method in which an estimator $\hat{\theta}$ of a random parameter θ is determined by minimizing the mean-squared error objective function J, that is

$$J[\tilde{\theta}_{MS}(k)] = E\{[\theta - \hat{\theta}_{MS}(k)]^T[\theta - \hat{\theta}_{MS}(k)]\} \tag{4.60}$$

where $\tilde{\theta}_{MS}(k) = \theta - \hat{\theta}_{MS}(k)$ is the estimation error, and the estimation is based on the observation data $\mathbf{y}(1)$, ..., $\mathbf{y}(k)$. According to the fundamental theorem of estimation theory, [57], minimizing the mean-squared error is equivalent to minimizing

the conditional mean-squared error

$$J_c[\tilde{\theta}_{MS}(k)] = E\{[\theta - \hat{\theta}_{MS}(k)]^T[\theta - \hat{\theta}_{MS}(k)]|\mathbf{y}(1), ...\mathbf{y}(k)\} \tag{4.61}$$

and the solution of this estimator is

$$\hat{\theta}_{MS}(k) = E\{\theta|\mathbf{Y}(k)\} \tag{4.62}$$

where $\mathbf{Y}(k)$ is the composed vector representing the observations $\mathbf{y}(1)$, ..., $\mathbf{y}(k)$.

Filtering, predicting and smoothing are three different methods of estimating the state of the systems by minimizing the mean-squared estimation error of the state. Kalman filtering is the method for estimating the current state of a dynamic system based on the past and current observations or measurements of that systems. The Kalman filtering estimate $\hat{\mathbf{z}}(k|k)$ of the state $\mathbf{z}$ at time k is defined by the expectation of state at time k given the observations from time 1 to time k

$$\hat{\mathbf{z}}(k|k) = E[\mathbf{z}(k)|\mathbf{y}(1), ..., \mathbf{y}(k)]. \tag{4.63}$$

Similarly, the first order (one step) predicted estimate $\hat{\mathbf{z}}(k|k-1)$ of the state $\mathbf{z}$ at time k is

$$\hat{\mathbf{z}}(k|k-1) = E[\mathbf{z}(k)|\mathbf{y}(1), ..., \mathbf{y}(k-1)] \tag{4.64}$$

and the first order smoothed estimate $\hat{\mathbf{z}}(k|k+1)$ of the state $\mathbf{z}$ at time k is

$$\hat{\mathbf{z}}(k|k+1) = E[\mathbf{z}(k)|\mathbf{y}(1), ..., \mathbf{y}(k+1)]. \tag{4.65}$$

The definition of the higher order predicted and smoothed estimates is straightforward. The predicted and filtered estimates of the state are coupled each other in the so called Kalman filtering technique, which is a recursive mean-squared method of obtaining the filtered estimate from the previous value of predicted estimate and vice-versa.

In the following the forward recursions of the Kalman filtering (Kalman, [39]), are presented for the application of the method to a basic linear dynamical system model as described by the linear state and observation equations for the non-stationary case

$$\mathbf{z}(k+1) = \mathbf{\Phi}(k+1,k)\mathbf{z}(k) + \mathbf{\Psi}(k+1,k)\mathbf{u}(k) + \mathbf{\Gamma}(k+1,k)\mathbf{w}(k), \tag{4.66}$$

$$\mathbf{y}(k) = \mathbf{H}(k)\mathbf{z}(k) + \mathbf{v}(k). \tag{4.67}$$

The initial state vector, $\mathbf{z}(0)$, is a multivariate Gaussian vector with mean μ and covariance matrix $\mathbf{\Sigma}$. $\mathbf{\Phi}(k+1,k)$ is the state transition matrix. $\mathbf{H}(k)$ is the observation matrix, $\mathbf{\Psi}(k+1,k)$ is the transformation matrix corresponding to the input signal $\mathbf{u}(k)$ and $\mathbf{\Gamma}(k+1,k)$ is the transformation matrix corresponding to the noise $\mathbf{w}(k)$. The two independent white noises $\mathbf{w}(k)$ and $\mathbf{v}(k)$ have the covariance matrices $\mathbf{Q}(k)$ and respectively $\mathbf{R}(k)$.

For this linear dynamic model the Kalman filter state estimation is given by the following recursive equation

$$\hat{\mathbf{z}}(k|k) = \hat{\mathbf{z}}(k|k-1) + \mathbf{K}(k)\tilde{\mathbf{y}}(k|k-1) \tag{4.68}$$

for $k = 1, 2, ..., N$, where $\mathbf{K}(k)$ is the Kalman gain matrix and $\tilde{\mathbf{y}}(k|k-1)$ is the innovation process computed as

$$\tilde{\mathbf{y}}(k|k-1) = \mathbf{y}(k) - \mathbf{H}(k)\hat{\mathbf{z}}(k|k-1). \tag{4.69}$$

The filtering equation from above consists of two terms, describing the prediction and respectively the correction. The predicted state estimation is given by the prediction equation

$$\hat{\mathbf{z}}(k|k-1) = \mathbf{\Phi}(k,k-1)\hat{\mathbf{z}}(k-1|k-1) + \mathbf{\Psi}(k,k-1)\mathbf{u}(k-1). \tag{4.70}$$

The filtering error $\tilde{\mathbf{z}}(k|k)$ defined by the equation

$$\tilde{\mathbf{z}}(k|k) = \mathbf{z}(k) - \hat{\mathbf{z}}(k|k), \tag{4.71}$$

is a zero-mean Gauss-Markov sequence with the covariance matrix defined by

$$\mathbf{P}(k|k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}(k)]\mathbf{P}(k|k-1) \tag{4.72}$$

where the predicted error covariance matrix is

$$\begin{aligned}\mathbf{P}(k|k-1) = &\mathbf{\Phi}(k,k-1)\mathbf{P}(k-1|k-1)\mathbf{\Phi}^T(k,k-1) \\ &+\mathbf{\Gamma}(k,k-1)\mathbf{Q}(k-1)\mathbf{\Gamma}^T(k,k-1),\end{aligned} \tag{4.73}$$

and the error cross-covariance matrix is

$$\mathbf{P}(k,k-1|k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}(k)]\mathbf{\Phi}(k,k-1)\mathbf{P}(k-1|k-1). \tag{4.74}$$

The Kalman gain matrix can be computed from the equation

$$\mathbf{K}(k) = \mathbf{P}(k|k-1)\mathbf{H}^T(k)[\mathbf{H}(k)\mathbf{P}(k|k-1)\mathbf{H}^T(k) + \mathbf{R}(k)]^{-1} \tag{4.75}$$

The initial conditions at time zero are: $\hat{\mathbf{z}}(0|0) = \mu$ and $\mathbf{P}(0|0) = \mathbf{\Sigma}$.

For the same basic linear dynamical model the fixed interval smoothed estimate of the state is computed using some backward recursions from $k = N$ to $k = 1$. These backward recursions (Rauch, [77]), are described by the equations

$$\hat{\mathbf{z}}(k-1|N) = \hat{\mathbf{z}}(k-1|k-1) + \mathbf{A}(k-1)[\hat{\mathbf{z}}(k|N) - \mathbf{\Phi}(k,k-1)\hat{\mathbf{z}}(k-1|k-1)], \tag{4.76}$$

$$\mathbf{A}(k-1) = \mathbf{P}(k-1|k-1)\mathbf{\Phi}(k,k-1)^T\mathbf{P}(k|k-1)^{-1}, \tag{4.77}$$

$$\mathbf{P}(k-1|N) = \mathbf{P}(k-1|k-1) + \mathbf{A}(k-1)[\mathbf{P}(k|N) - \mathbf{P}(k|k-1)]\mathbf{A}(k-1)^T, \tag{4.78}$$

$$\mathbf{P}(k,k-1|N) = \mathbf{P}(k,k-1|k) + [\mathbf{P}(k|N) - \mathbf{P}(k|k)]\mathbf{P}(k|k)^{-1}\mathbf{P}(k,k-1|k). \tag{4.79}$$

The application of Kalman filtering to the linear dynamical articulatory model with non-linear observation function can be done by approximating the non-linear function with piecewise linear functions on small regions and using the extended Kalman filtering method.

The extended Kalman filtering represents an extension of the linear Kalman filtering method to the non-linear systems in which one or both of the state-space equations are non-linear. This extension is based on the theory of small perturbations in which the non-linear system is approximated by a linear one on a small region, and the original non-linear model becomes a linear perturbation model.

The linearization of the non-linear observation function can be accomplished by expanding it in a Taylor series and, for example, restricting the expansion to the first order terms. By linearization of the non-linear equations of the state-space model on small regions some of the linear Kalman filtering equations will be changed by transforming the original non-linear model into a perturbation linear model. A variant of the extended Kalman filter has also been developed in which the state estimation is accomplished iteratively a number of times and this is called iterated extended Kalman filtering (IEKF). A complete derivation of the recursion formulas, for the general case with both equations of the state-space model non-linear, can be found in [57]. In our case, the articulatory dynamical model has only the observation equation non-linear, whereas the state equation is a linear equation. This simplifies somehow the transformation of the linear Kalman filter for the articulatory non-linear observation model.

Now, assuming that the parameters of the state-space model has been estimated from a set of training observations corresponding to a phonological coproduction unit, this model can be used for estimating the hidden articulatory state from new test data consisting of only acoustic observations. As described in the first section

of this chapter the augmented articulatory state-space model is described by the following linear state equation and non-linear observation equation

$$\mathbf{z}(k+1) = \mathbf{\Phi}\mathbf{z}(k) + \mathbf{w}_z(k), \tag{4.80}$$

$$\mathbf{y}(k) = \mathbf{g}[\mathbf{z}(k)] + \mathbf{v}(k) \tag{4.81}$$

where $\mathbf{g}[\mathbf{z}(k)]$ is a nonlinear function relating the augmenting articulatory state vector $\mathbf{z}(k)$ to the acoustic measurements or observation vector $\mathbf{y}(k)$. By linearizing the observation function on small regions around some nominal points $\mathbf{z}_*(k)$ the non-linear equation can be transformed into a linear perturbation equation

$$\delta\mathbf{y}(k) = \mathbf{G}[\mathbf{z}_*(k)]\delta\mathbf{z}(k) + \mathbf{v}(k), \tag{4.82}$$

where $\delta\mathbf{z}(k) = \mathbf{z}(k) - \mathbf{z}_*(k)$, $\delta\mathbf{y}(k) = \mathbf{y}(k) - \mathbf{g}[\mathbf{z}_*(k)]$ and $\mathbf{G}[\mathbf{z}_*(k)]$ is the Jacobian matrix of the multivariate function $\mathbf{g}[\mathbf{z}(k)]$ computed at the nominal value $\mathbf{z}_*(k)$. This perturbation equation can be written in the following form

$$\mathbf{y}(k) = \mathbf{g}(\mathbf{z}_*) + \mathbf{G}(\mathbf{z}_*)[\mathbf{z}(k) - \mathbf{z}_*] + \mathbf{v}(k), \tag{4.83}$$

which represents the first order approximation by the Taylor series expansion of the original non-linear observation equation. Because of this approximation, the extended Kalman filter and smoother will not be the optimal estimators of the state $\mathbf{z}(k)$, but the first order approximations of $\mathbf{E}[\mathbf{z}(k)|\mathbf{Y}(1,...,k)]$, respectively $\mathbf{E}[\mathbf{z}(k)|\mathbf{Y}(1,...,N)]$.

The Jacobian matrix $\mathbf{G}[\mathbf{z}_*(k)]$ has the dimension $m \times n$ and its elements are defined by the partial derivatives

$$\frac{\partial g_i}{\partial z_j} = \frac{\partial g_i[\mathbf{z}(k),k]}{\partial z_j(k)}|_{\mathbf{z}(t)=\mathbf{z}_*(t)}, \tag{4.84}$$

where $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. The nominal value $\mathbf{z}_*(k)$ can be chosen to be at each time frame k the predicted estimate of the state

$$\mathbf{z}_*(k) = \hat{\mathbf{z}}(k|k-1). \tag{4.85}$$

Thus, by obtaining at each time frame k the nominal value $\mathbf{z}_*(k) = \hat{\mathbf{z}}(k|k-1)$ from Kalman predictor and the corresponding $\mathbf{g}[\hat{\mathbf{z}}(k|k-1)]$ and $\mathbf{G}[\hat{\mathbf{z}}(k|k-1)]$ from linearization, the forward and backward recursions of the extended Kalman filtering and smoothing can be computed using these terms as follows

a) **Forward Recursions (Filtering)**

$$\hat{\mathbf{z}}(k|k-1) = \mathbf{\Phi}\hat{\mathbf{z}}(k-1|k-1). \tag{4.86}$$

$$\hat{\mathbf{z}}(k|k) = \hat{\mathbf{z}}(k|k-1) + \mathbf{K}(k;*)\{\mathbf{y}(k) - \mathbf{g}(\mathbf{z}_*) - \mathbf{G}(\mathbf{z}_*)[\hat{\mathbf{z}}(k|k-1) - \mathbf{z}_*]\}. \tag{4.87}$$

$$\mathbf{K}(k;*) = \mathbf{P}(k|k-1;*)\mathbf{G}^T(k;*)[\mathbf{G}(k;*)\mathbf{P}(k|k-1;*)\mathbf{G}^T(k;*) + \mathbf{R}]^{-1}. \tag{4.88}$$

$$\mathbf{P}(k|k-1;*) = \mathbf{\Phi}\mathbf{P}(k-1|k-1;*)\mathbf{\Phi}^T + \mathbf{Q}_z. \tag{4.89}$$

$$\mathbf{P}(k|k;*) = [\mathbf{I} - \mathbf{K}(k;*)\mathbf{G}(k;*)]\mathbf{P}(k|k-1;*), \tag{4.90}$$

$$\mathbf{P}(k,k-1|k;*) = [\mathbf{I} - \mathbf{K}(k;*)\mathbf{G}(k;*)]\mathbf{\Phi}\mathbf{P}(k-1|k-1;*). \tag{4.91}$$

where $*$ denotes the use of $\hat{\mathbf{z}}(k|k-1)$. These forward recursions of the extended Kalman filter are executed for $k = 1, 2, ..., N$. The a priori estimated mean of initial state is used for $\hat{\mathbf{z}}(0|0)$. The Equation 4.86 is called the prediction equation because it predicts the state at time k from the state at time $k-1$. The Equation 4.87 is called the correction equation because it corrects the predicted state for time k with the information obtained from the observation vector at time k. An iterative version of the extended Kalman filter (IEKF) has been proposed (Jazwinski, [37]). In this method, after running the above equations of the extended Kalman filter using the

linearization around $\mathbf{z}_*(k) = \hat{\mathbf{z}}(k|k-1)$, the model is iteratively re-linearized around previously estimated state $\mathbf{z}_*(k) = \hat{\mathbf{z}}(k|k)$.

The smoothed estimate of the state can be obtained by the following recursions of the Kalman smoother.

b) Backward Recursions (Smoothing)

$$\hat{\mathbf{z}}(k-1|N) = \hat{\mathbf{z}}(k-1|k-1) + \mathbf{A}(k-1)[\hat{\mathbf{z}}(k|N) - \mathbf{\Phi}\hat{\mathbf{z}}(k-1|k-1)], \quad (4.92)$$

$$\mathbf{A}(k-1) = \mathbf{P}(k-1|k-1)\mathbf{\Phi}^T\mathbf{P}(k|k-1)^{-1}, \quad (4.93)$$

$$\mathbf{P}(k-1|N) = \mathbf{P}(k-1|k-1) + \mathbf{A}(k-1)[\mathbf{P}(k|N) - \mathbf{P}(k|k-1)]\mathbf{A}(k-1)^T, \quad (4.94)$$

$$\mathbf{P}(k,k-1|N) = \mathbf{P}(k,k-1|k) + [\mathbf{P}(k|N) - \mathbf{P}(k|k)]\mathbf{P}(k|k)^{-1}\mathbf{P}(k,k-1|k). \quad (4.95)$$

These Kalman smoother recursions are executed for $k = N, N-1, \ldots, 1$ and use the first and second order statistics computed in the forward pass of the extended Kalman filter.

In the above forward and backward recursions the parameters are constant over segments of speech corresponding to each phonological coproduction model.

We implemented the extended Kalman filtering program in two versions, in C and MATLAB. Due to the intensive matrix computations for which MATLAB has been optimized, the difference in time execution between the two versions was relatively small.

4.5 Segmentation, Recognition of Models and Estimation of Articulatory Trajectories

The method of estimating the articulatory trajectories from the speech signal presented in this thesis is based on a new way of constraining the speech inversion, by using high-level, gestural phonological constraints. In this section, we address the issues why the recognition of phonological coproduction models is important for the estimation of articulatory trajectories and how this recognition can be implemented and integrated into the global speech inversion method.

The phonological constraints are imposed by recognizing the coproduction model (α, β), which has the highest conditional probability $p(\alpha, \beta|\mathbf{Y})$, given the observation sequence $\mathbf{Y}$, and then, by estimating the trajectories using the recognized model's parameters for the Kalman smoothing. In other words, we filter iteratively a speech segment using the parameters of different models, e.g., the dynamical model parameters and the articulatory-acoustic sub-function, and find the model which best fits the data, using a likelihood measure based on innovation sequences of the extended Kalman filtering. The whole process can be seen as an integrated method, since both model recognition and trajectory estimation are based on the same statistical method of extended Kalman filtering.

Mathematically, the recognition of coproduction models (or units) is based on the Bayes' rule, according to which the probability of a model (α, β), given an observation sequence $\mathbf{Y}$, can be expressed as a function of the probability of observations given the model. This rule can be described by the equation

$$p(\alpha, \beta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\alpha, \beta)p(\alpha, \beta)}{p(\mathbf{Y})}. \tag{4.96}$$

In this study we did not make use of the *a priori* probability of the coproduction

units or models $p(\alpha, \beta)$, so we supposed that all the models had the same *a priori* probability. Also, the probability of the observation sequence $p(\mathbf{Y})$ has the same value across all the models and, consequently, we can ignore it. Thus, for the recognition of phonological coproduction units, instead of using the probability of a model given the observation sequence we can use the probability of the observation sequence given the model

$$p(\alpha, \beta|\mathbf{Y}) \propto p(\mathbf{Y}|\alpha, \beta). \tag{4.97}$$

We modeled the above probabilities by likelihood measures. The likelihood measure used in recognizing the coproduction models can be based on the innovation sequences from the extended Kalman filtering, as in [19], and can be computed using the formula

$$logL(\mathbf{Y}|\alpha, \beta) = -\frac{1}{2}\sum_{k=1}^{N_S}\{log|\mathbf{\Sigma}_{e_k}(\alpha, \beta)| + \mathbf{e}_k^T(\alpha, \beta)\mathbf{\Sigma}_{e_k}^{-1}(\alpha, \beta)\mathbf{e}_k(\alpha, \beta)\} \tag{4.98}$$

where N_S is the number of frames in the observation sequence corresponding to a coproduction model, and the innovations process given the model $\mathbf{e}_k(\alpha, \beta)$ and its covariance matrix given the model $\mathbf{\Sigma}_{e_k}(\alpha, \beta)$ are computed from the extended Kalman filtering equations of the previous section, as follows

$$\mathbf{e}_k(\alpha, \beta) = \mathbf{y}(k) - \mathbf{g}(\mathbf{z}_*) - \mathbf{G}(\mathbf{z}_*)[\hat{\mathbf{z}}(k|k-1) - \mathbf{z}_*] \tag{4.99}$$

$$\mathbf{\Sigma}_{e_k}(\alpha, \beta) = \mathbf{G}(k; *)\mathbf{P}(k|k-1; *)\mathbf{G}^T(k; *) + \mathbf{R}. \tag{4.100}$$

For the case when the segmentation of the utterance is known and the model boundaries are given, the recognition of gestures implies simply the application of the extended Kalman filtering using different (α, β) models. After recognizing the model with the highest likelihood function, the extended Kalman smoothing is finally applied using the parameters of this model in order to estimate the articulatory trajectories. The target phoneme β of the recognized model becomes then the

base phoneme α of the next model. For the complete recognition of the next model, it is necessary to do the filtering using only the parameters of the $\alpha*$ models, where $*$ stands for any possible phoneme which is allowed to succeed the known phoneme α. Thus, to start the estimation of articulatory trajectories, the only information needed is the position of the central point of the phonemes (segmentation) and the label of the first phoneme from the utterance. If the label of the first phoneme is not known, a solution can be obtained by applying only for the first segment the filtering using all coproduction models and recognizing the one with the highest likelihood measure. Then, the algorithm can be applied as described until the end of the utterance. Except for the case of laboratory studies, where the segmentation can be carried out a priori in a manual way, the segmentation of the speech signal is not available. The automatic segmentation of the speech utterances is a rather complicated operation and sometimes requires a whole process of automatic speech recognition.

In applying the whole method of estimating the articulatory trajectories, a method of segmenting the speech signal and estimating the phoneme boundaries can be employed from the field of automatic speech recognition. Such a segmentation method and estimation of the sequence of states can be based on the popular Viterbi algorithm [102], later described in [28]. Other methods of automatic segmentation of the speech signal into phonetic units, which are not based on automatic speech recognition, can also be used (Svendsen and Soong [100]). The coproduction segment boundaries are not the same with the phoneme boundaries, usually computed by the current automatic speech recognition systems, like those based on Hidden Markov Models (HMM). In the HMM recognition methods, a search algorithm, such as Viterbi search, estimates the most likely sequence of states and the phoneme boundaries. These boundaries are delimiting the most stationary part

of the phonemes. For each phoneme, the two boundaries approximate its beginning and end, respectively. In our framework, the phoneme boundaries determined as above are not suitable for delimiting the segments corresponding to the coproduction models, because, in general, the phoneme boundaries will be placed somewhere between the true coproduction unit boundaries (usually close to the middle point of the coproduction units). Thus, for our framework, the segmentation should be based on finding the model boundaries which are placed approximately in the middle of the phonemes. The first boundary of a coproduction segment is placed approximately in the middle point of the starting phoneme α, of the model, whereas the second one is placed approximately in the middle point of the target phoneme β. The accurate estimation of the middle points of the phonemes is not critical due to the fact that we did not use piecewise constant target inputs in the state space models. This flexibility in approximating the boundaries of the coproduction segments makes a model not completely disjoint from others with which it shares one of the α or β phonemes (e.g., the coproduction models /aa eh/ and /eh s/ have an overlapping region around the middle point of the phoneme /eh/). A simple, although not very accurate, solution to estimate the model boundaries from the phoneme boundaries is to choose the median frame between these phoneme boundaries. However, such a method of segmentation based on the Viterbi search algorithm cannot be carried out without training the HMM models and, hence, this segmentation method would involve a complete automatic speech recognition task as a preprocessing phase of speech inversion. Due to the complexity of the automatic speech recognition process, we would like to refrain ourself of stipulating that condition as a prerequisite phase of the speech inversion method. An alternative solution for estimating the gestural boundaries have been sought.

Although it was not the goal of this study to develop a new search and segmen-

tation algorithm, for the purpose of demonstrating the speech inversion method, we experimentally developed an integrated method of segmentation. This segmentation method is integrated into the whole method of estimating the articulatory trajectories and uses the extended Kalman filtering results. The objective of this method is to approximate the centers of the phonemes in the speech utterance using the trained coproduction models and some cost functions based on the likelihood computation. In Figure 4.6 we present the general flow diagram of the integrated procedure of recognizing the gestures and estimating the articulatory trajectories. Given a speech signal, represented by a sequence $\mathbf{y}_1, ..., \mathbf{y}_{END}$ of acoustic vectors, e.g. MFCC parameters, first, we apply a simple method of localization and classification of the first phoneme of the utterance. This classification of the first phoneme may not be extremely accurate, because a re-classification of this phoneme will be carried out further. The localization of the first phoneme is based on energy and zero-crossing rate computations. Then the classification method is applied based on the minimum Euclidean distance. Such a distance is computed using the MFCC acoustic parameters from the frames of the first sound of the utterance and some stored reference frames for each of the phonemes. The phoneme with the minimum distance is recognized as the first phoneme p_1. We set the base of the coproduction model $\alpha = p_1$ and the beginning of the segment $B = 1$, that is, the initial frame of the first phoneme. Then the extended Kalman filtering is applied for all the (α, β) models in which $\alpha = p_1$ is known. The interval for this filtering is chosen as $[B, B + L]$, where L represents a number of frames, large enough to include at least the second phoneme in that interval. If $B + L \geq END$, we use the interval $[B, END]$. Then, some cost functions $C(\mathbf{y}(k)|\alpha\beta)$ based on the negative of the conditional likelihood of the acoustic observations given each model and for each

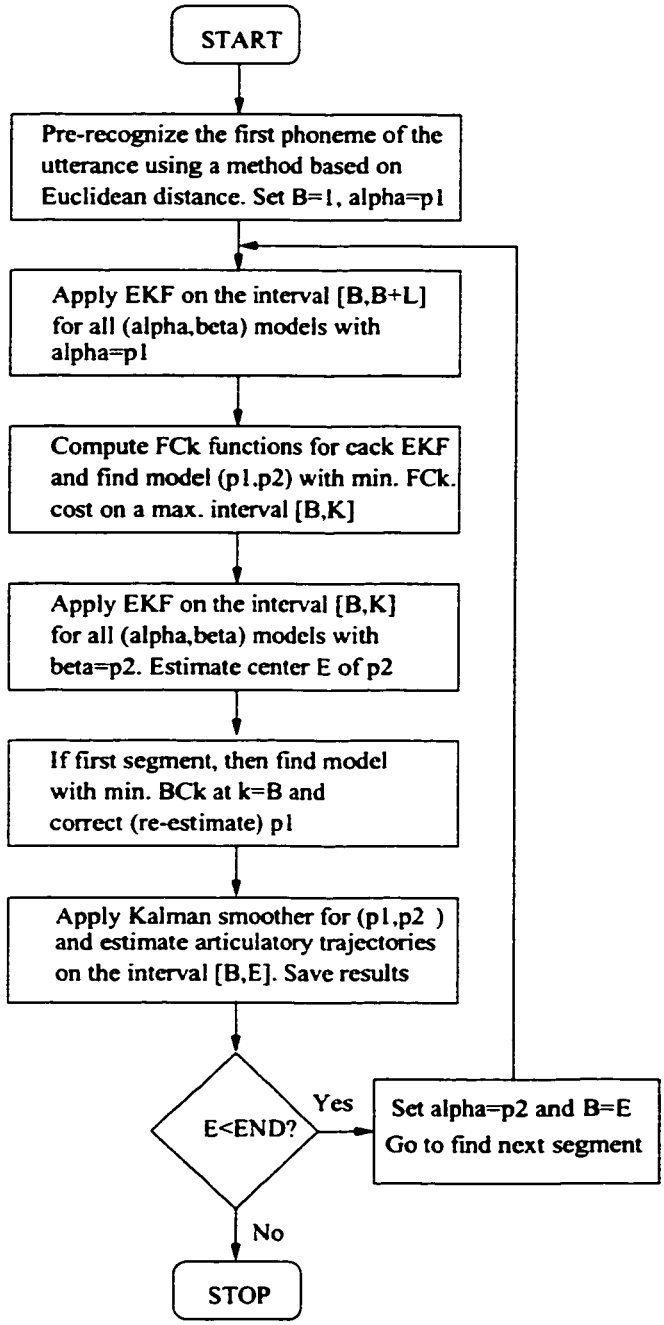

Figure 4.6: Flow diagram of the procedure of recognizing gestures and estimating articulatory trajectories

frame are computed using the innovation sequences

$$C(\mathbf{y}(k)|\alpha,\beta) = log|\mathbf{\Sigma}_{e_k}(\alpha,\beta)| + \mathbf{e}_k^T(\alpha,\beta)\mathbf{\Sigma}_{e_k}^{-1}(\alpha,\beta)\mathbf{e}_k(\alpha,\beta) \tag{4.101}$$

By summing these cost functions we compute the forward-accumulated cost functions (or distances) at each frame k

$$FC_k(\alpha,\beta) = \sum_{n=B}^{k}[\log|\mathbf{\Sigma}_{e_n}(\alpha,\beta)| + \mathbf{e}_n^T(\alpha,\beta)\mathbf{\Sigma}_{e_n}^{-1}(\alpha,\beta)\mathbf{e}_n(\alpha,\beta)], \tag{4.102}$$

for $k = B, ..., B+L$. We find the model p_1p_2 with the minimum accumulated cost on a maximum interval $[B,K]$, where $K \leq B+L$. Then we apply the extended Kalman filtering on the interval $[B,K]$, using all $(*,\beta)$ models in which $\beta = p_2$, and compute the average cost functions (Av. Co.) over all these models, on that interval. This average cost is then smoothed over a number of frames. The position of the minimum of the smoothed, averaged cost function will approximate the center E of the p_2 phoneme. If this is the first model, than from the previous filtering with fixed $\beta = p_2$ and variable $\alpha = *$, we compute some backward-accumulated cost functions

$$BC_k(\alpha,\beta) = \sum_{n=k}^{E}[\log|\mathbf{\Sigma}_{e_n}(\alpha,\beta)| + \mathbf{e}_n^T(\alpha,\beta)\mathbf{\Sigma}_{e_n}^{-1}(\alpha,\beta)\mathbf{e}_n(\alpha,\beta)], \tag{4.103}$$

for $k = B, ..., E$. Finding the model for which the backward-accumulated cost function is minimum at $k = B$, we obtain the re-classification of the first phoneme of the utterance. Once the $(\alpha,\beta) = (p_1,p_2)$ model is known, we apply the Kalman smoother for the final estimation of the articulatory trajectories on the estimated interval B,E, and we save the model name and boundaries together with the estimated articulatory trajectories. If the end of the model segment E is less than the end of the utterance END, then we set $\alpha = p_2$ and $B = E$ and go to find and estimate the next segment of speech by iterating the operations from the last five

boxes. This procedure terminates when the end of utterances is reached and the Kalman smoother was applied in the last segment.

For the purpose of illustration of the overall method, in the following we present an example. Figure 4.7 illustrates examples of the cost functions representing the negative of the likelihood for the first part of a slow utterance /aa zh aa b aa/, for five (α, β) models in which the base phoneme α is /aa/. The frame interval for this utterance was 10 ms. The five models have the target phonemes /b/, /d/, /p/, /zh/ and /eh/. The sixth plot represents the average cost over all the models in which the base phoneme α was /aa/.

A low value in the cost functions, respectively a high likelihood, means a good fit of the acoustic observations with the corresponding model. Conversely, a high value in the cost functions shows that is unlikely that the observation frame corresponds to that model. As can be seen in the 4th plot corresponding to the model /aa zh/, the cost function has the lowest cost. Then, we apply the extended Kalman filtering for all the models for which the target phoneme β is /zh/ and compute the average cost over all these models. Figure 4.8 illustrates the iterative steps of the algorithm. First, we pre-recognize the first phoneme of the utterance, $\alpha = /aa/$. The plots from each column represents the filtering phases within a pass of the block diagram, corresponding to a single segment. The sub-plots from the first raw represent the forward-accumulated cost functions computed for each segment, with known α. The forward-accumulated cost function with the lowest final value corresponds to the recognized target phoneme $\beta = p_2$. The sub-plots from the second raw represent the average cost functions (dotted line), computed by filtering with models with different α and fixed $\beta = p_2$. The minimum of the smoothed average cost functions (solid line), provides an estimate of the center of each target phoneme $\beta = p_2$. The sub-plots from the third raw represent the backward-accumulated cost functions for

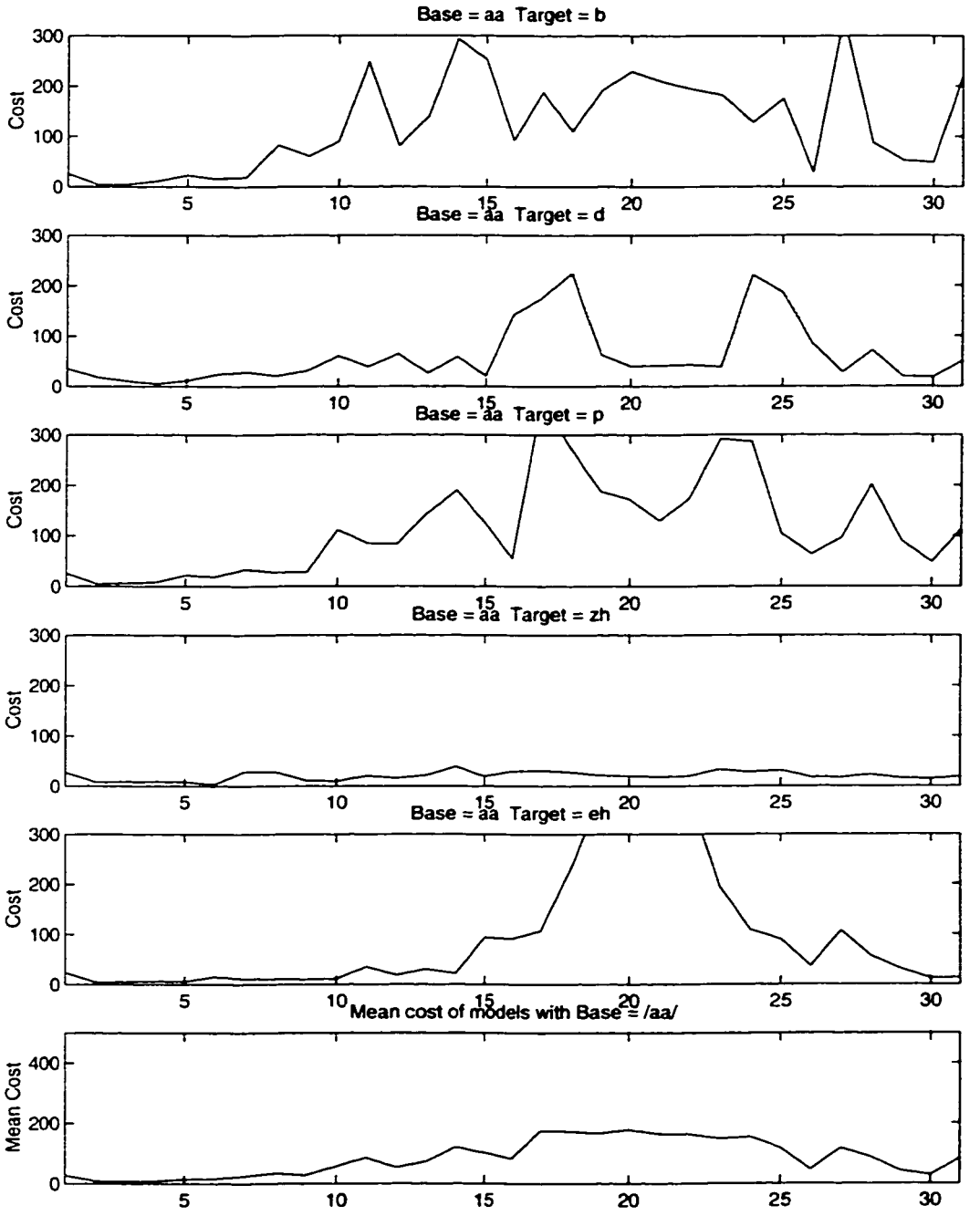

Figure 4.7: Plots of cost functions (- log-Likelihood) for 5 (α, β) models ($\alpha = /aa/$)

the re-estimation of the first phoneme $\alpha = p_1$. Finally, the sub-plot from the bottom of the figure represents the actual articulatory trajectories and the estimated model boundaries for this example. Now we describe in details each phase of the method displayed in this figure. Once we finished the second phase of filtering, for all models with $\beta = /zh/$, we compute the smoothed average cost over all these models and run an algorithm of finding an approximate median point of minimum smoothed average cost value (first sub-plot from the second raw). In this sub-plot the dotted line represents the average cost and the solid line represents the smoothed average cost. This minimum value corresponds to the 'center' of the second phoneme and is stored. As in Figure 4.8, this point was found at frame number 22, as printed in the first sub-plot of the third raw. After the recognition of the second phoneme, the reclassification of the first phoneme of the utterance is carried out by filtering the $k = 1, \ldots, 22$ interval using all the models with the target phoneme $\beta = /zh/$, and choosing the model with the minimum backward-accumulated cost, $\alpha = /aa/$, as shown in the first sub-plot from the third raw. The base phoneme $\alpha = /aa/$ of this model will be the finally recognized first phoneme of the utterance. Then we started the recognition/estimation in the second segment. In the cost functions of $\alpha = /zh/$ models we select the longest interval starting from frame 22 and having the lowest accumulated cost, as we did for the first segment. We found that this corresponds to the model (zh, aa). We repeat these steps until the end of the utterance and we found the 'central' points of the other phonemes from the utterance. In this example, the original 'central' points of the phonemes were at the frames 1, 21, 41, 61 and 81. The estimated 'central' points by using this method were 1, 22, 42, 61 and 80, as the segmentation is presented in the bottom sub-plot (dash-dot line). Once the boundaries are obtained for each coproduction segment, the estimation of articulatory trajectories is simply achieved by applying the extended

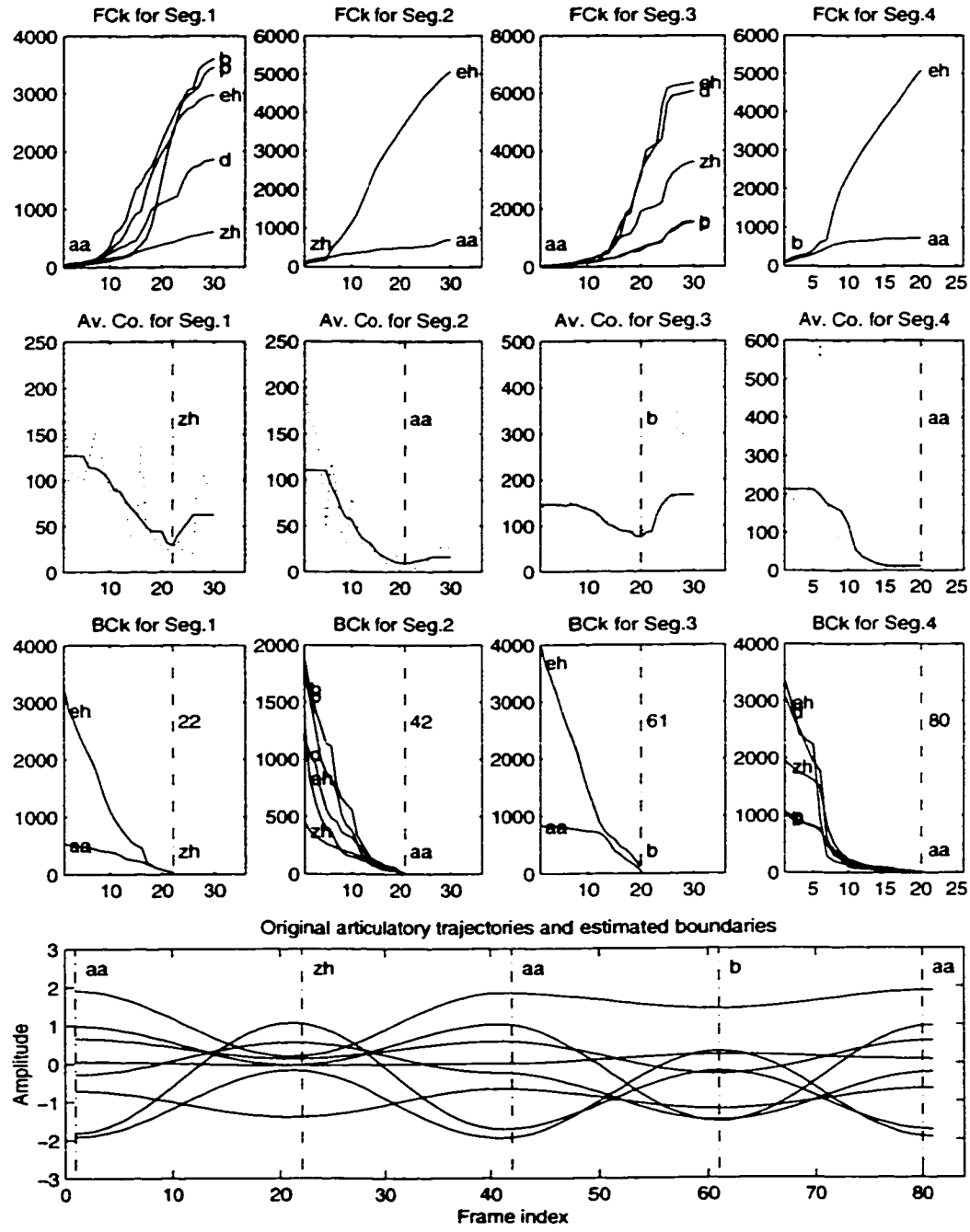

Figure 4.8: Automatic segmentation of /aa zh aa b aa/ using a maximum-likelihood method

Kalman smoother for each segment using the corresponding model parameters. The smoothed articulatory trajectories for this example of a synthesized utterance /aa zh aa b aa/ are presented in Figure 4.9. In this figure, each sub-plot represents the estimated and actual trajectories of an articulatory parameter of Maeda's model. Thus, j is the jaw, b is the tongue body, d is the tongue dorsum, tx is the tongue tip X position, ty is the tongue tip Y position, lx is the lip protrusion, ly is the lip aperture and ph is the pharynx hight. At the beginning of each segment the new smoother will use as the initial state and covariance matrix the corresponding values from the last frame of the previous smoothing interval. At the bottom of this figure, the actual (solid line) and estimated (dotted line) vocal-tract midsagittal shapes are displayed. Each shape represents the configuration of the vocal-tract at a certain frame of speech (here a frame is 10 ms), where a center of a phoneme was found. In the proximity of each vocal-tract shape the corresponding frame number and phonemic transcription are printed. For this example of synthesized speech, the actual and estimated articulatory trajectories are very close.

The method of segmentation and recognition of coproduction models is integrated into the general process of estimating articulatory trajectories. This model recognition is important not only because it provides a phonetic transcription of the speech utterance, but also because it constraints the search of the articulatory states in the process of estimating the trajectories. However, we do not expect that the phonemic transcription generated by this method to be perfect and comparable to the one provided by a state-of-the-art automatic speech recognition system based on HMM, as our goal is an accurate estimation of the articulatory trajectories and not the phonemic recognition.

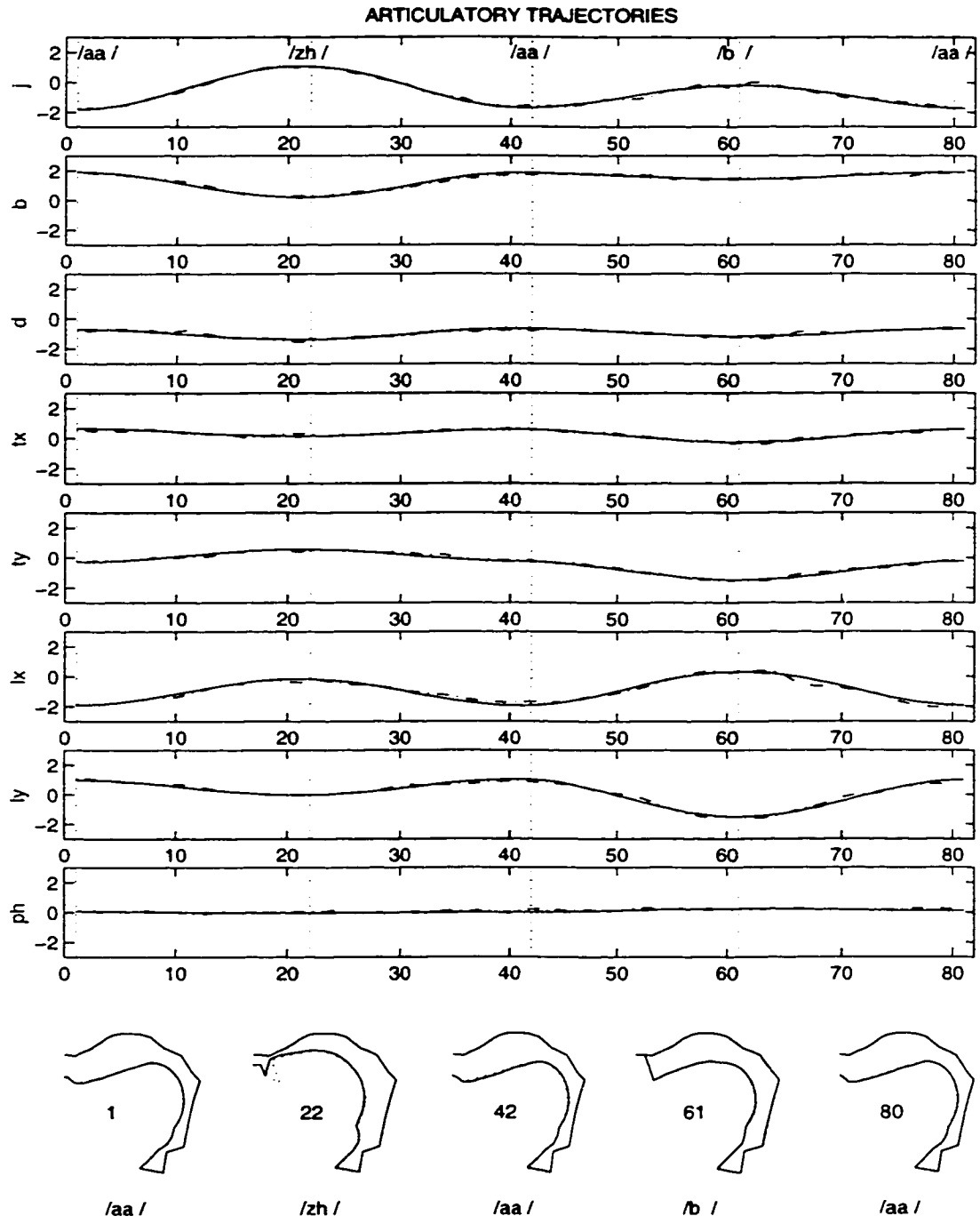

Figure 4.9: Actual (solid) and estimated (dashdot) articulatory trajectories for /aa zh aa b aa/

Chapter 5

Experimental Results

In this chapter we present results from three sets of experiments, based on synthesized and real speech data. The estimated articulatory trajectories are compared to the actual articulatory trajectories simultaneously recorded with the speech signals. The first section describes experiments based on speech data synthesized using an articulatory-acoustic model. The second section describes results of estimating articulatory trajectories based on real speech data, recorded with an Electromagnetic Midsagittal Articulograph, (Carstens [10]). The third section presents some results of estimation based on real speech data from the X-ray Microbeam Speech Production Database of University of Wisconsin, [104].

5.1 Experimental Results based on Synthesized Speech Data

We begin this section by presenting a few experimental results of estimating articulatory trajectories from our preliminary studies based on Kalman filtering tech-

nique and using articulatory-acoustic codebooks. In these preliminary studies we implemented the Kalman filtering and smoothing algorithms and the maximum-likelihood model parameter estimation method. We used two different acoustic features: formant frequencies and MFCC parameters. We created the articulatory-acoustic codebooks using the Maeda's articulatory and acoustic models [54], [55]. We did not have in these preliminary experiments real articulatory trajectories to be compared to the estimated ones, since we used for testing acoustic data alone from the TIMIT database, [65]. Also we did not use any speaker normalization technique here. We used the acoustic prototypes of the 10 vowels which were published by Peterson and Barney [72], for 61 male and female American English speakers. We used the following American English vowels: /aa/, /ae/, /ah/, /ao/, /eh/, /ey/, /ih/, /iy/, /uh/ and /uw/. For each vowel produced by each speaker, 1,000 articulatory prototypes were obtained from the Maeda's articulatory model and using the Metropolis algorithm. This synthesized articulatory-acoustic database was obtained and clustered in our laboratory in a previous project by Dr. Arturo Galván and Jeff Ma. The exact acoustic formants corresponding to these articulatory prototypes were recalculated using a frequency-domain vocal-tract acoustic model. A total of 610,000 articulatory-acoustic pairs were obtained for the whole vowel codebook. Some of the articulatory-acoustic pair vectors were identical or very close to other pair vectors in this database. A pruning was carried out and a final codebook of about 331,000 different pairs of vectors was created. Based on this second codebook, we tested the estimation method using the first three formant frequencies from vowel tokens extracted from the TIMIT database. Actual articulatory trajectories were not available. In Figure 5.1 we present the Maeda's model articulatory trajectories estimated from a seven frame token of /aa/ spoken by a female speaker. As can be seen, the estimated articulatory trajectories are

quite smooth. In Figure 5.2, the corresponding original formant frequencies and the synthesized ones using the estimated articulatory trajectories are presented. The original and synthesized formants are almost identical. This does not mean that the estimated articulatory trajectories were close to the real ones. However, the extended Kalman filtering and smoothing were working well providing smooth articulatory trajectories and well fitted acoustic trajectories, reconstructed from the estimated articulatory parameters.

In a second preliminary experiment we used a similar codebook derived from the 610,000 articulatory prototypes of vowels described above. We computed from these articulatory prototypes the vocal-tract transfer functions and the corresponding MFCC parameters. After pruning, this new codebook contained 235,000 different articulatory prototypes and the corresponding MFCC parameters as acoustic features. We obtained the MFCC parameters from the articulatory prototypes by using first a frequency-domain vocal-tract acoustic model, to obtain the transfer function of the vocal-tract, and then we developed a method to obtain the MFCC parameters from the vocal-tract transfer functions. In Figure 5.3, the estimation results are presented for a token /aa/ of a female speaker from the TIMIT database. The estimated articulatory trajectories are smooth and the area functions and profiles of the vocal-tract seem to be reasonable. Again there is no comparison with real articulatory trajectories for data obtained from TIMIT. In Figure 5.4, the original MFCC trajectories are compared to the reconstructed MFCC trajectories using the estimated articulatory parameters. As can be observed, the acoustic fit is quite good. We now present an estimation result based on the same framework as in the above figures, but using test acoustic features from the continuous French sentence "Ma chemise est roussie," spoken by a French female speaker. Figure 5.5, show the estimation results for the segment /m i z e/, from the above sentence, even though

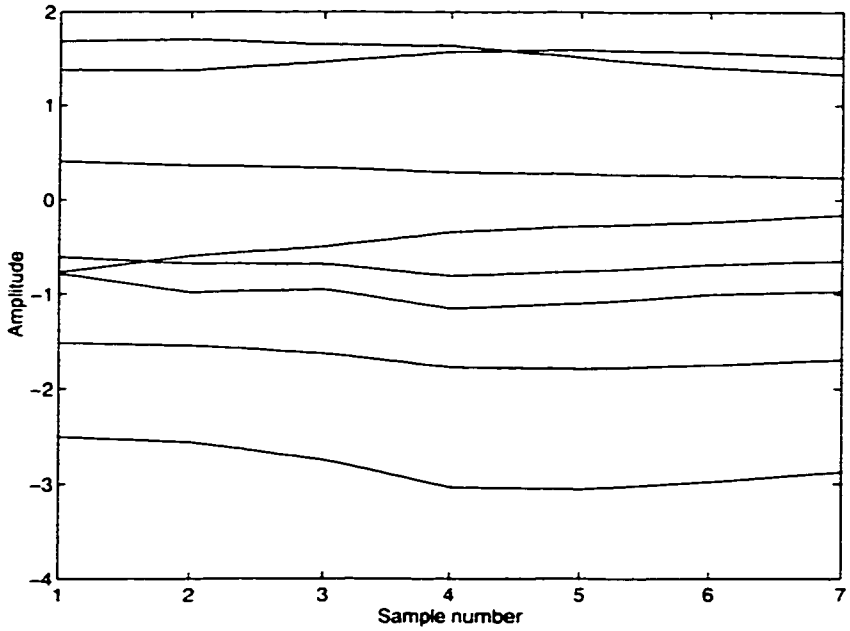

Figure 5.1: Estimated articulatory trajectories for an /aa/ token from TIMIT

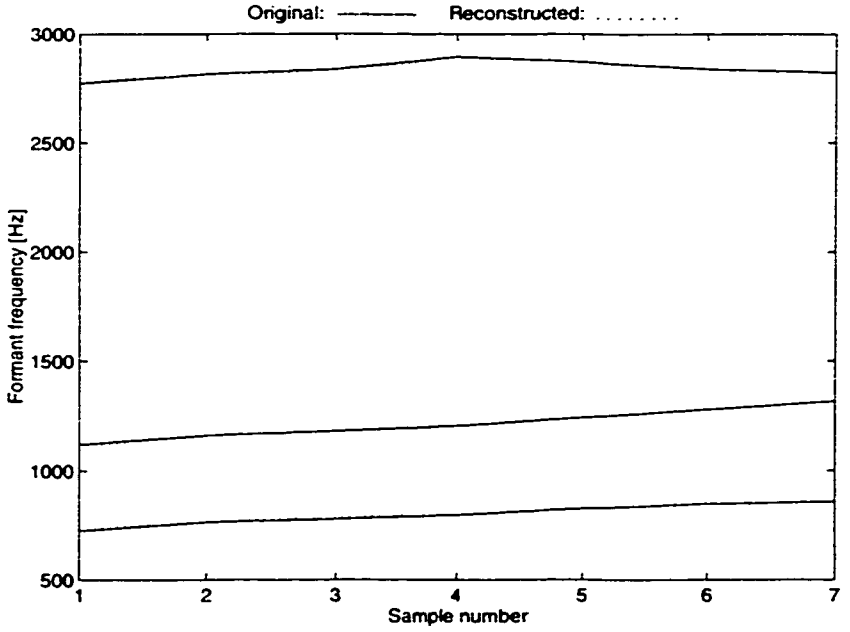

Figure 5.2: Actual and reconstructed formants for the /aa/ token

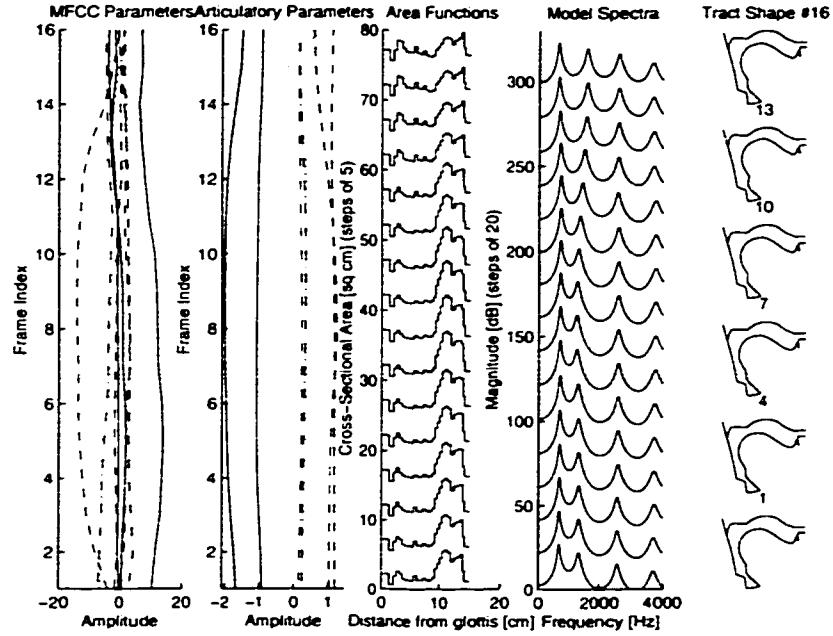

Figure 5.3: Estimated articulatory trajectories for /aa/ (TIMIT)

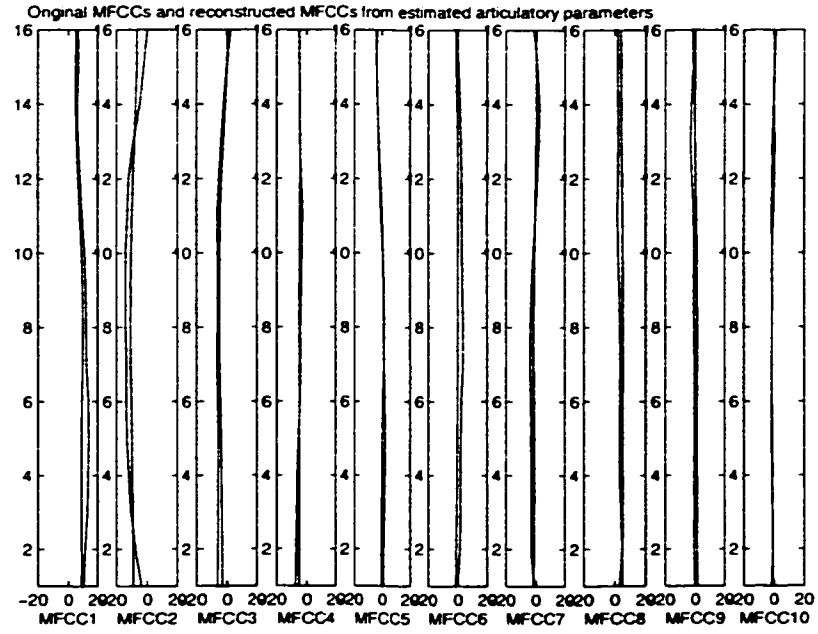

Figure 5.4: Actual and reconstructed MFCC trajectories for /aa/ (TIMIT)

in the codebook we did not include vocal-tract shapes and MFCC parameters for consonants. Again, the acoustic match is very good, as can be seen in Figure 5.6.

In these preliminary studies our goal was mainly to develop the general framework of the estimation method based on Kalman filtering, and to test its capabilities of estimating articulatory trajectories from acoustic features alone. In these experiments we did not provide or implement important innovative solutions to overcome the difficulties revealed by the previous studies of speech inversion based on Kalman filtering technique. However, we tested the articulatory-acoustic codebook approach. Because the vowel codebooks used were created by using a means of random sampling of the articulatory space, many unrealistic vocal-tract shapes were included in these codebooks. In the latter experiment, presented in Figure 5.5 and Figure 5.6, we compared the estimated articulatory trajectories to the real ones, which were extracted from X-ray data of the French female speaker from which the statistical articulatory model was built, and they were quite different. This discordance cannot be explained only by the acoustic differences in the sounds of English and French. We believe that the inclusion of unrealistic vocal-tract shapes into the codebook, by randomly sampling the articulatory space, was responsible for these differences in articulatory trajectories.

After these preliminary experiments, we will present now some experimental results obtained by our method in which we used both dynamical and phonological constraints. We synthesized speech using the Maeda's articulatory model and time-domain acoustic model. The articulatory prototypes of some vowels were extracted from the x-ray data of a French female speaker, from whom the statistical articulatory model was built (Maeda [55]). Prototypes for other phonemes and some vowels were manually adjusted in order to obtain a perceptually better acoustic match with the American English phonemes. We obtained the trajectories between

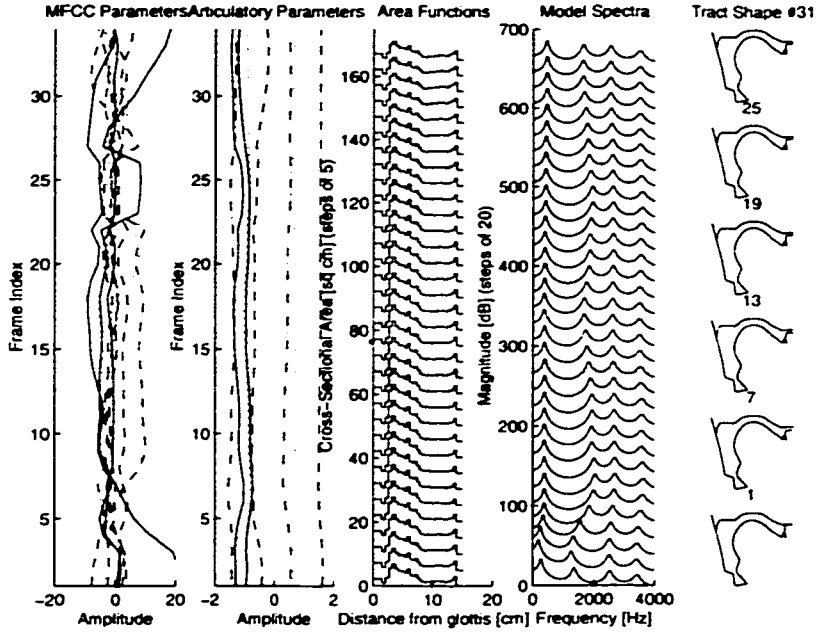

Figure 5.5: Estimated articulatory trajectories for /m i z e/

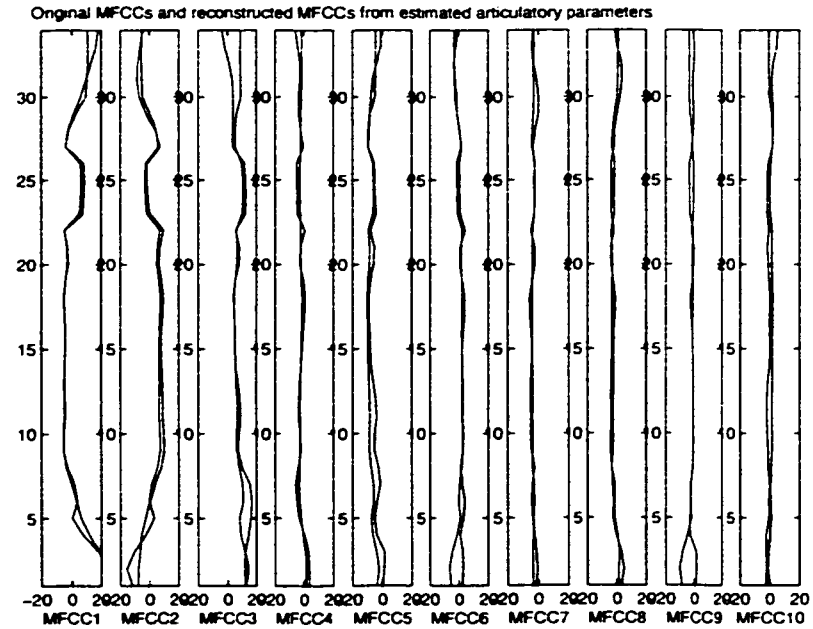

Figure 5.6: Actual and reconstructed MFCC trajectories for /m i z e/

each pair of articulatory prototypes by interpolation with a cosine function. We used different interpolation function and finally chose the cosine function because the trajectories were better matched with those obtained with an electromagnetic midsagittal articulograph. The articulatory trajectories were sampled at 10 ms. In these experiments we used only 8 articulatory model parameters, even though we added to the Maeda's articulatory model a nasal tract and another parameter that controls the velum position. The acoustic parameters used in this experiment were the MFCCs. We obtained these acoustic parameters using two different tools: the HTK tool and a local DSP tool previously developed in our group. First the speech signal was pre-emphasized and windowed using a Hamming window of 32 ms, and a frame step of 10 ms. Then the FFT was computed and the log power spectrum was applied to 25 triangular filters distributed on a mel-frequency scale. From the output of the filters the MFCC parameters were computed using the classical formula. We used the first 10 MFCC parameters after excluding the first coefficient corresponding to energy.

We generated articulatory and acoustic trajectories for different coproduction segments (pairs of phonemes, or diphones). In order to construct coproduction models for each such pair, a number of different tokens were needed. We used a random number generation method to synthesize different tokens by adding Gaussian noise to an initial trajectory. In Figure 5.7 we represented the articulatory-acoustic vectors, corresponding to 20 synthesized /eh ih/ trajectories. Each sub-plot represents a particular projection of the data on a two-dimensional space consisting of an articulatory dimension and an acoustic dimension. As can be seen in this figure, the articulatory-acoustic sub-functions are in general nonlinear. From these data a model for /eh ih/ was created using look-up tables of codebooks. Then we synthesized test data for /eh ih/ different from the training data. An example of

estimated articulatory trajectories from a test utterance for /eh ih/ is presented in Figure 5.8. At the bottom of this figure the actual and estimated vocal-tract shapes are superimposed. Similar synthesized data were obtained for /b ih/, /s ih/ and /d ih/, and presented in Figures 5.9, 5.10 and 5.11. We did not used in this experiment nasals, but other coproduction segments were synthesized as presented in Table 5.1. These examples contain transitions to three different vowels — /ih/,

	/ih/	/ah/	/uh/
/eh/	eh ih	eh ah	eh uh
/b/	b ih	b ah	b uh
/s/	s ih	s ah	s uh
/d/	d ih	d ah	d uh

Table 5.1: Table of synthesized coproduction segments

/ah/ and /uh/, from the vowel /eh/, the stop consonants /b/ and /d/, and the fricative /s/. Examples of estimated trajectories and vocal-tract shapes for /eh ah/, /b ah/, /s ah/ and /d ah/ are presented in Figures 5.12 to 5.15, and for /eh uh/, /b uh/, /s uh/ and /d uh/ are presented in Figures 5.16 to 5.19.

In a different experiment we tested other coproduction models. Examples for /aa b/, /aa d/ and /aa sh/ are presented in Figures 5.20 to 5.22. Other coproduction segments were also synthesized and experiments of estimating articulatory trajectories were carried out with similar results for these segments. The accuracy of estimated articulatory trajectories was very high in this experiment, especially due to the fact that the training and test data were not very different; that is, the Gaussian noise used to synthesize new trajectories was relatively small.

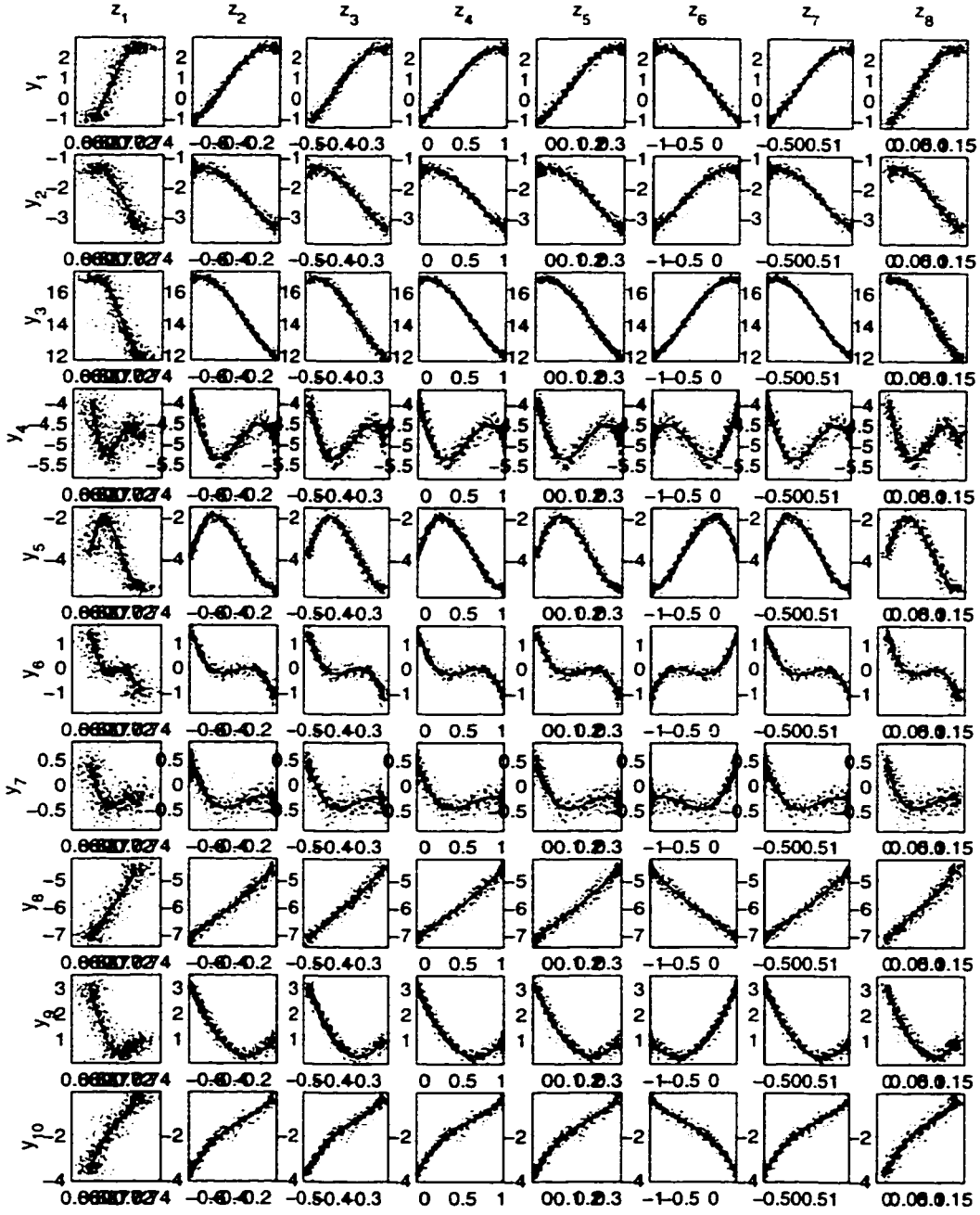

Figure 5.7: Articulatory and acoustic training data for /eh ih/ consisting of 20 synthesized segments

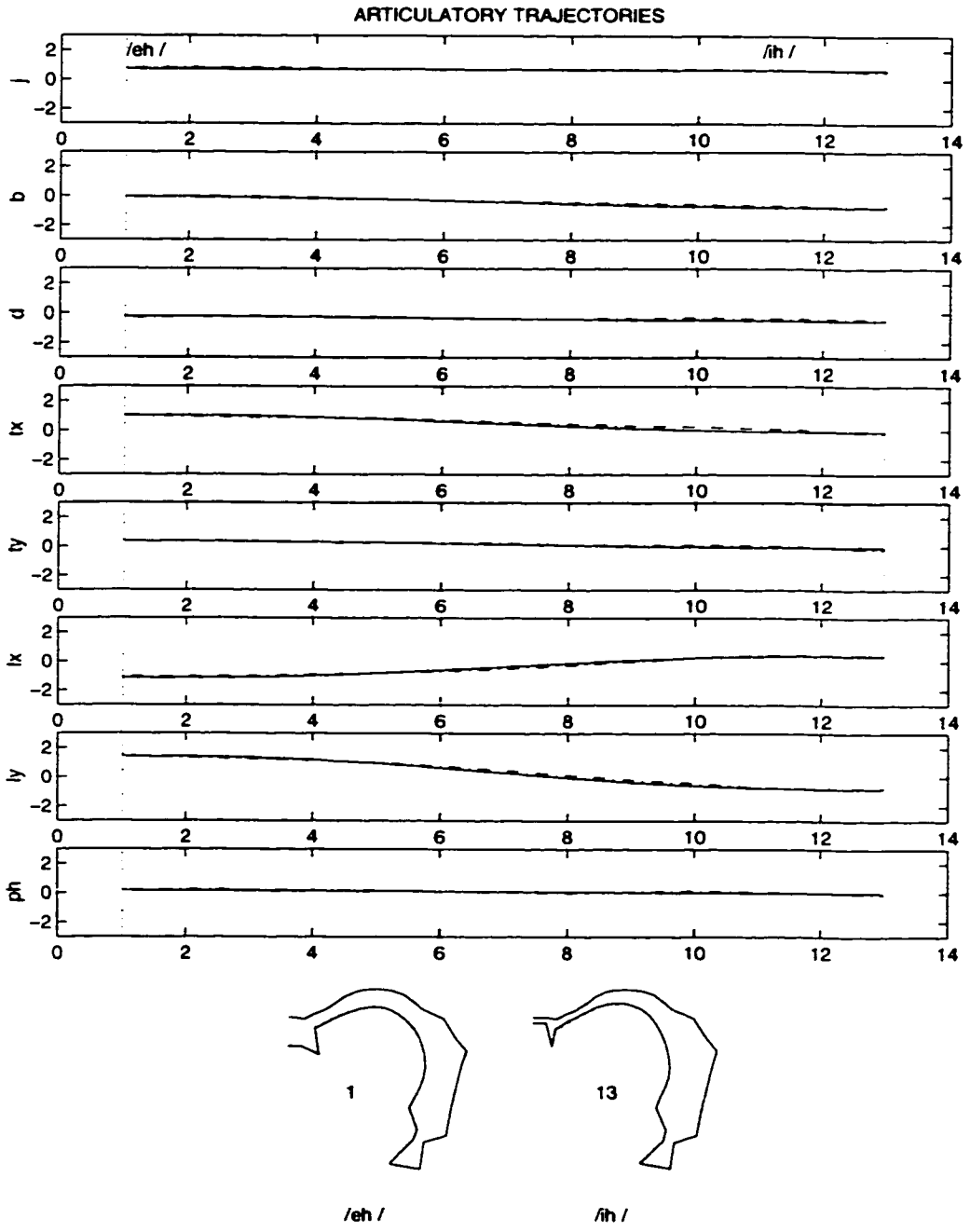

Figure 5.8: Actual and estimated articulatory trajectories for a segment /eh ih/

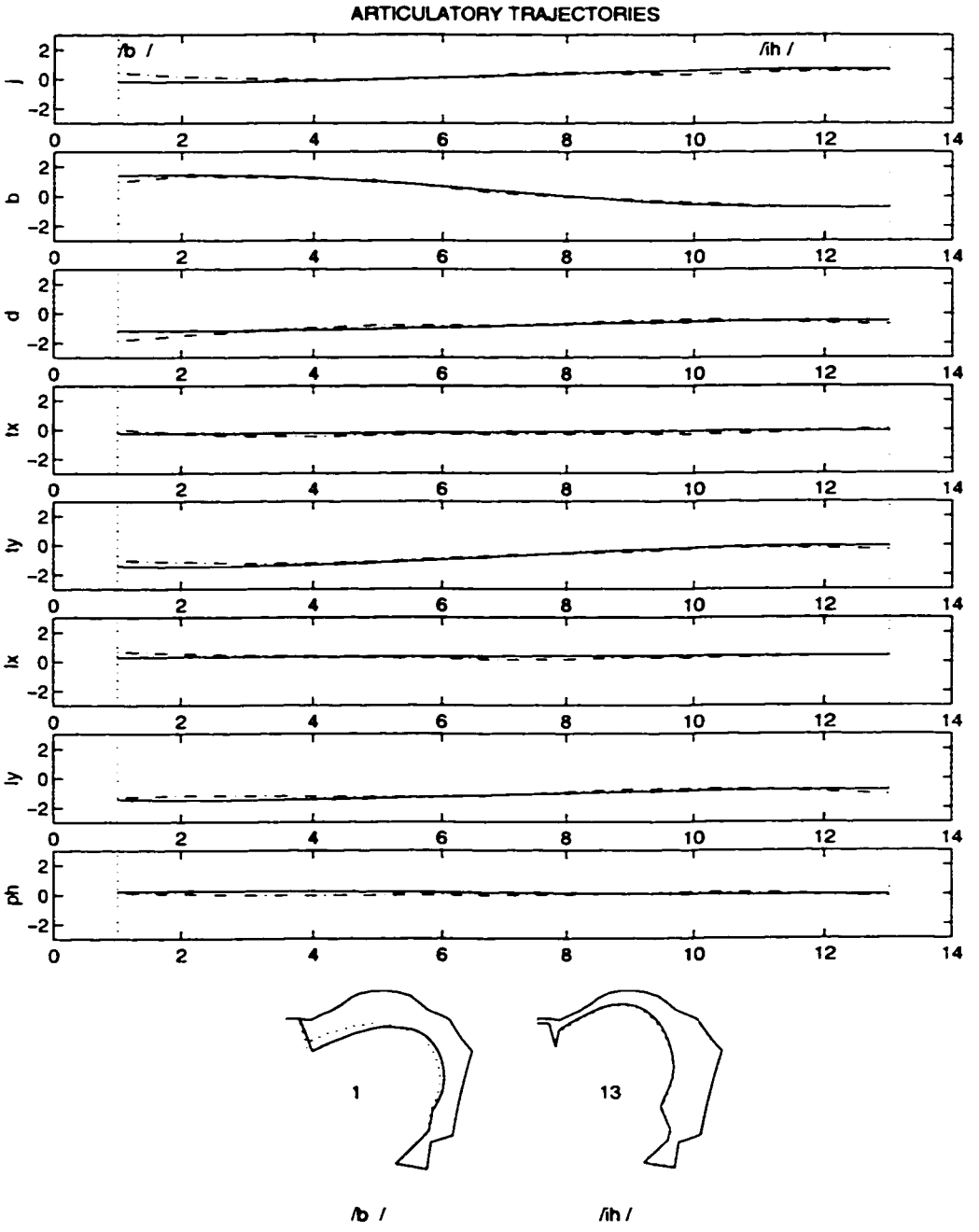

Figure 5.9: Actual and estimated articulatory trajectories for a segment /b ih/

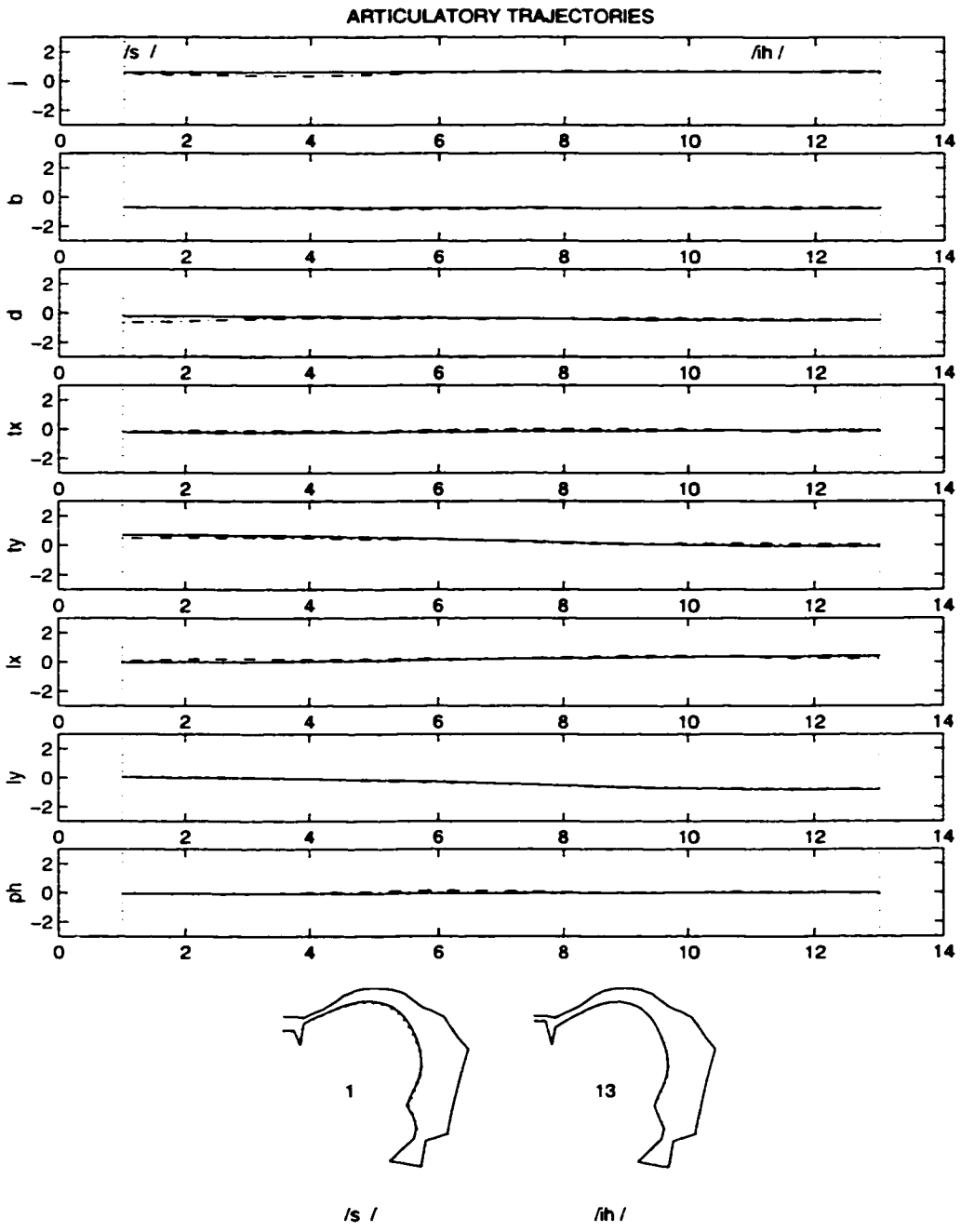

Figure 5.10: Actual and estimated articulatory trajectories for a segment /s ih/

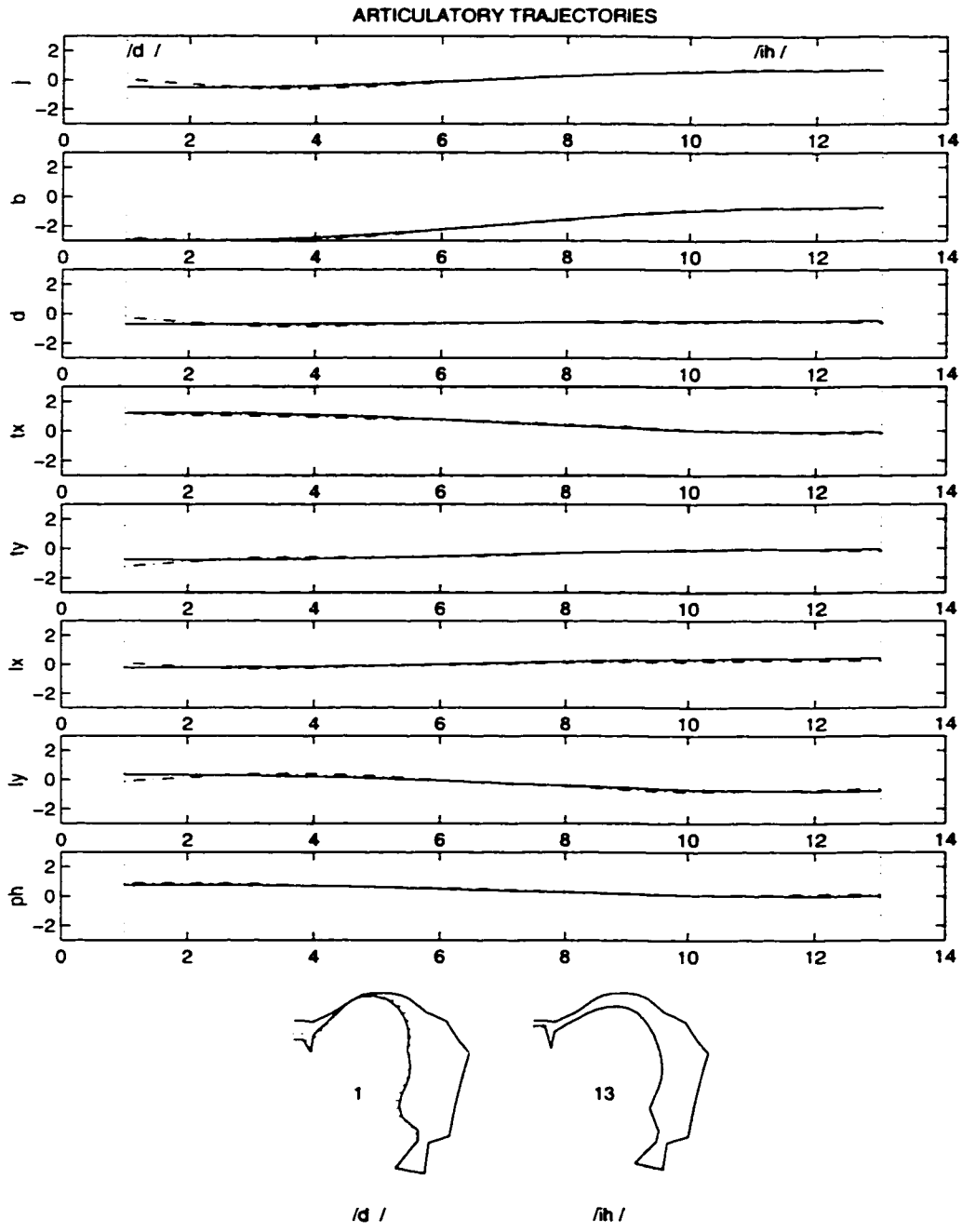

Figure 5.11: Actual and estimated articulatory trajectories for a segment /d ih/

We did not try to synthesize speech data with large variances in articulatory domain because the resulting vocal-tract shapes could change the initial phonetic identity of the sounds, and also because there was no guarantee that the resulted vocal-tract shapes were realistic or physically possible. However, this experiment based on synthesized speech data, had the main objective to develop and test the estimation method, and not to obtain real estimation results.

Experimental results from continous speech are also presented here, since the automatic segmentation method was developed and tested first on synthesized speech data. This experiment based on continuous synthesized speech contained the following coproduction segments, as presented in Table 5.2

	/aa/	/eh/	/b/	/d/	/zh/	/sh/	/p/
/aa/		aa eh	aa b	aa d	aa zh	aa sh	aa p
/eh/	eh aa		eh b	eh d	eh zh	eh sh	eh p
/b/	b aa	b eh					
/d/	d aa	d eh					
/zh/	zh aa	zh eh					
/sh/	sh aa	sh eh					
/p/	p aa	p eh					

Table 5.2: Table of synthesized segments for continuous speech experiments

Thus, Figures 5.23 and 5.24 present results of automatic segmentation and estimation for the synthesized continuous utterance /aa p aa b aa/. Other results of estimating articulatory trajectories for utterances /aa b aa p aa/ and /aa zh aa b aa/ are presented in Figures 5.25 and Figures 5.26.

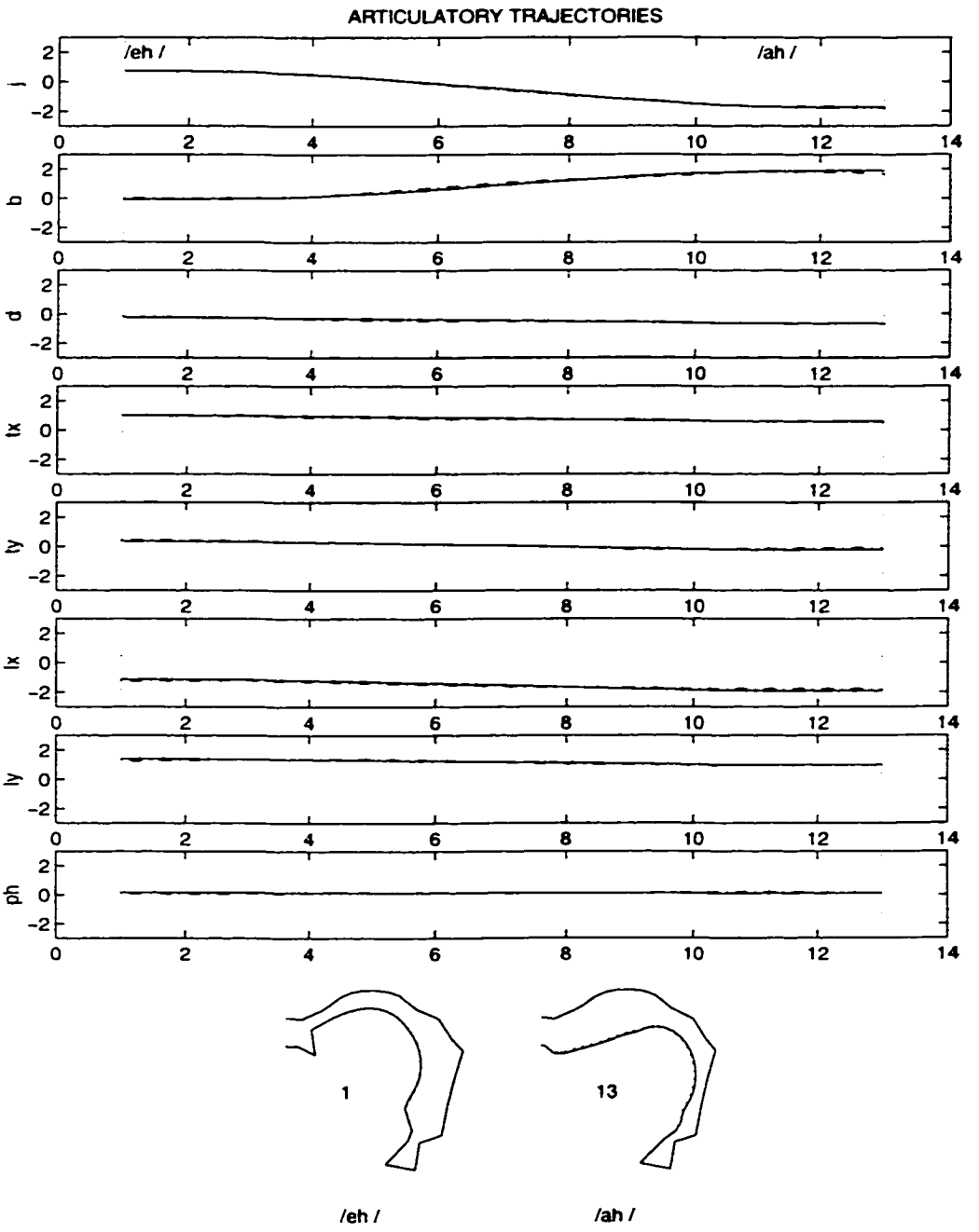

Figure 5.12: Actual and estimated articulatory trajectories for a segment /eh ah/

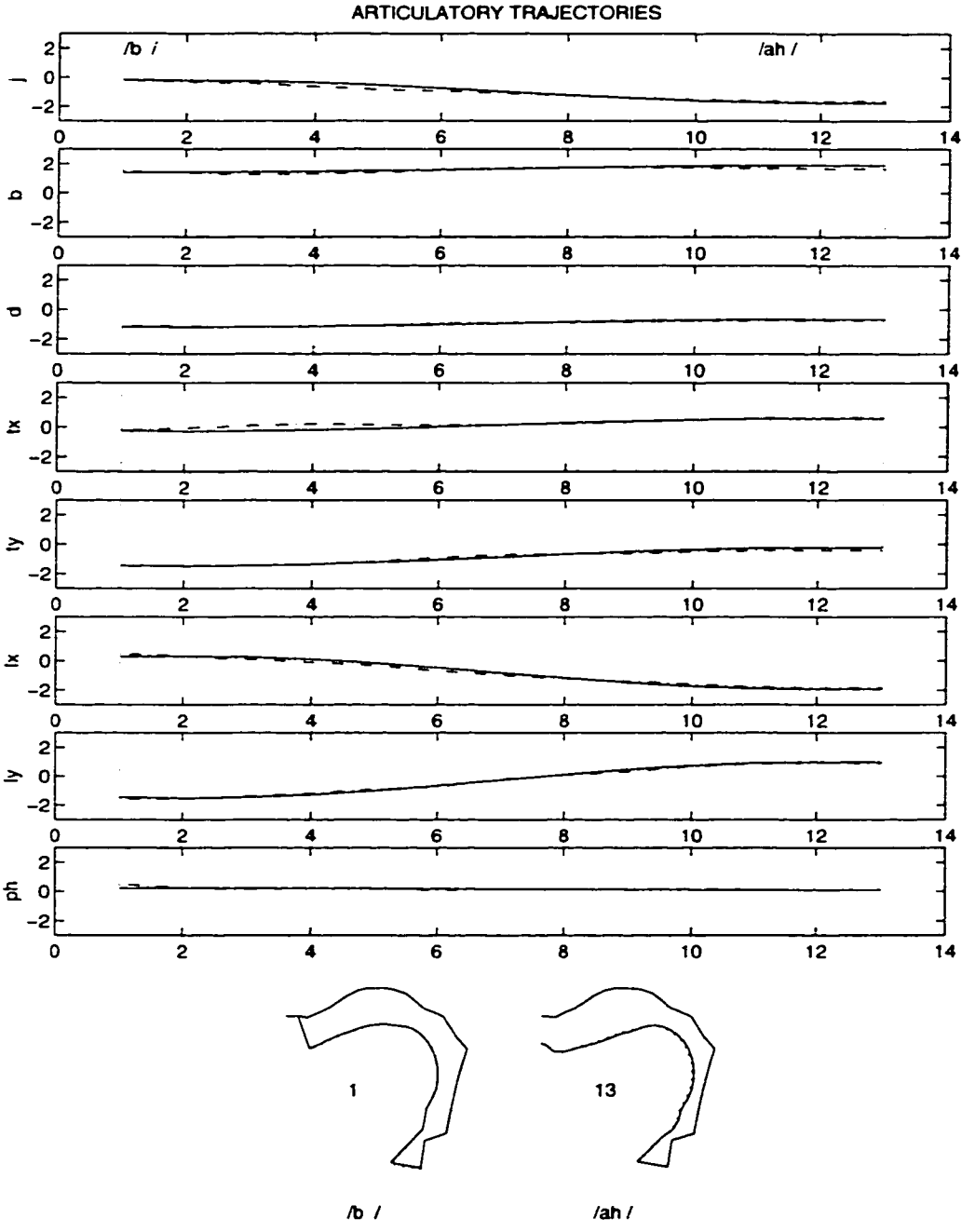

Figure 5.13: Actual and estimated articulatory trajectories for a segment /b ah/

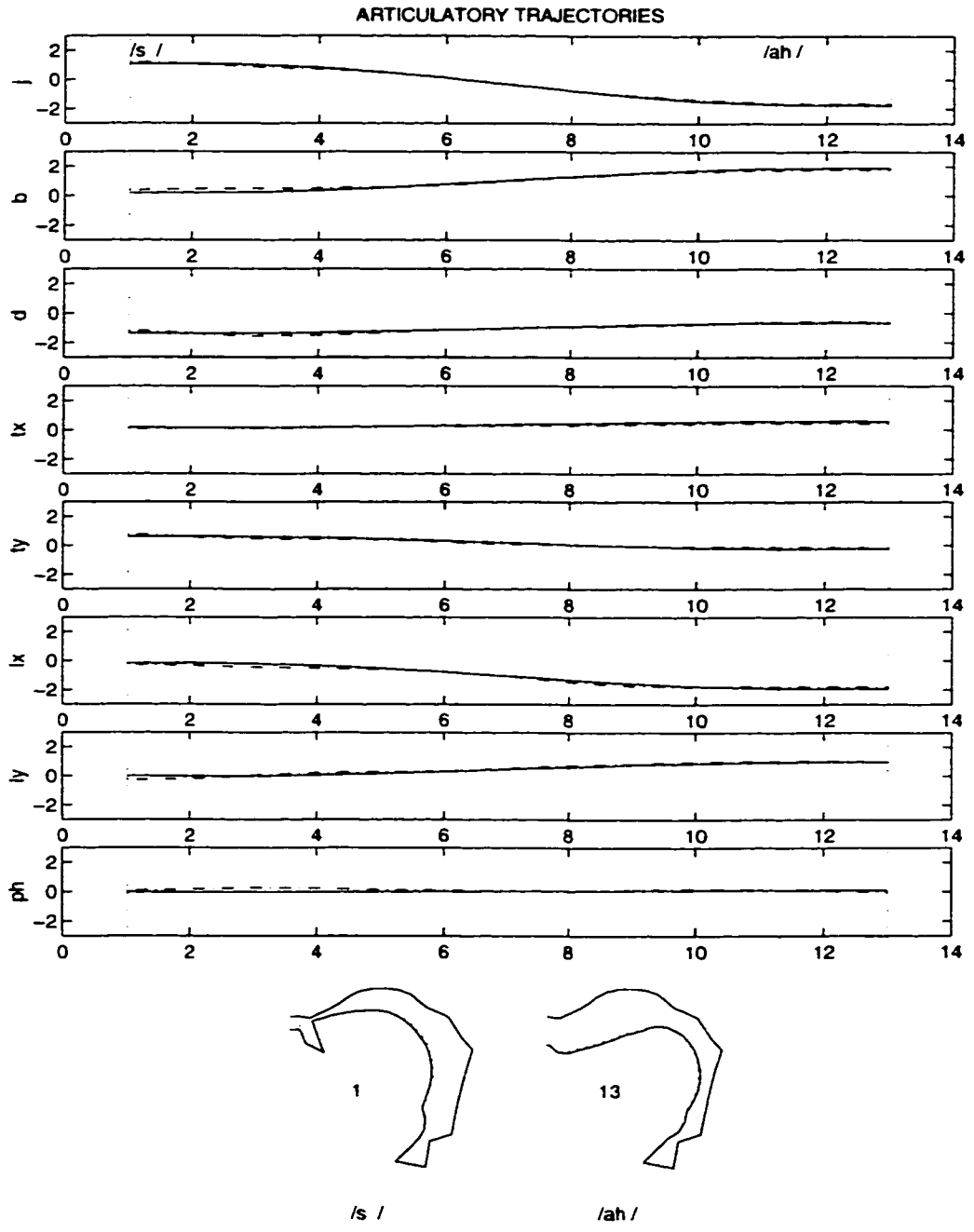

Figure 5.14: Actual and estimated articulatory trajectories for a segment /s ah/

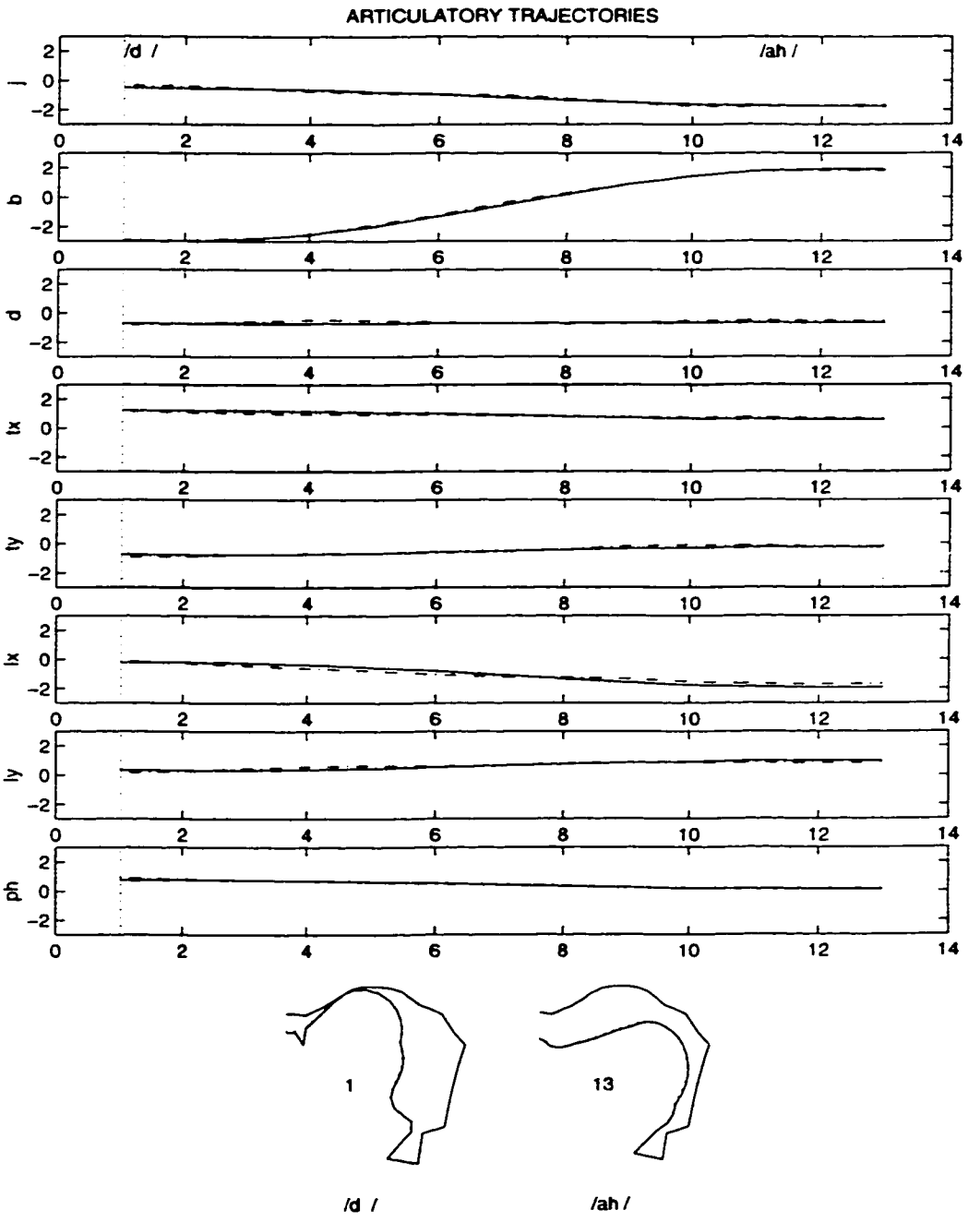

Figure 5.15: Actual and estimated articulatory trajectories for a segment /d ah/

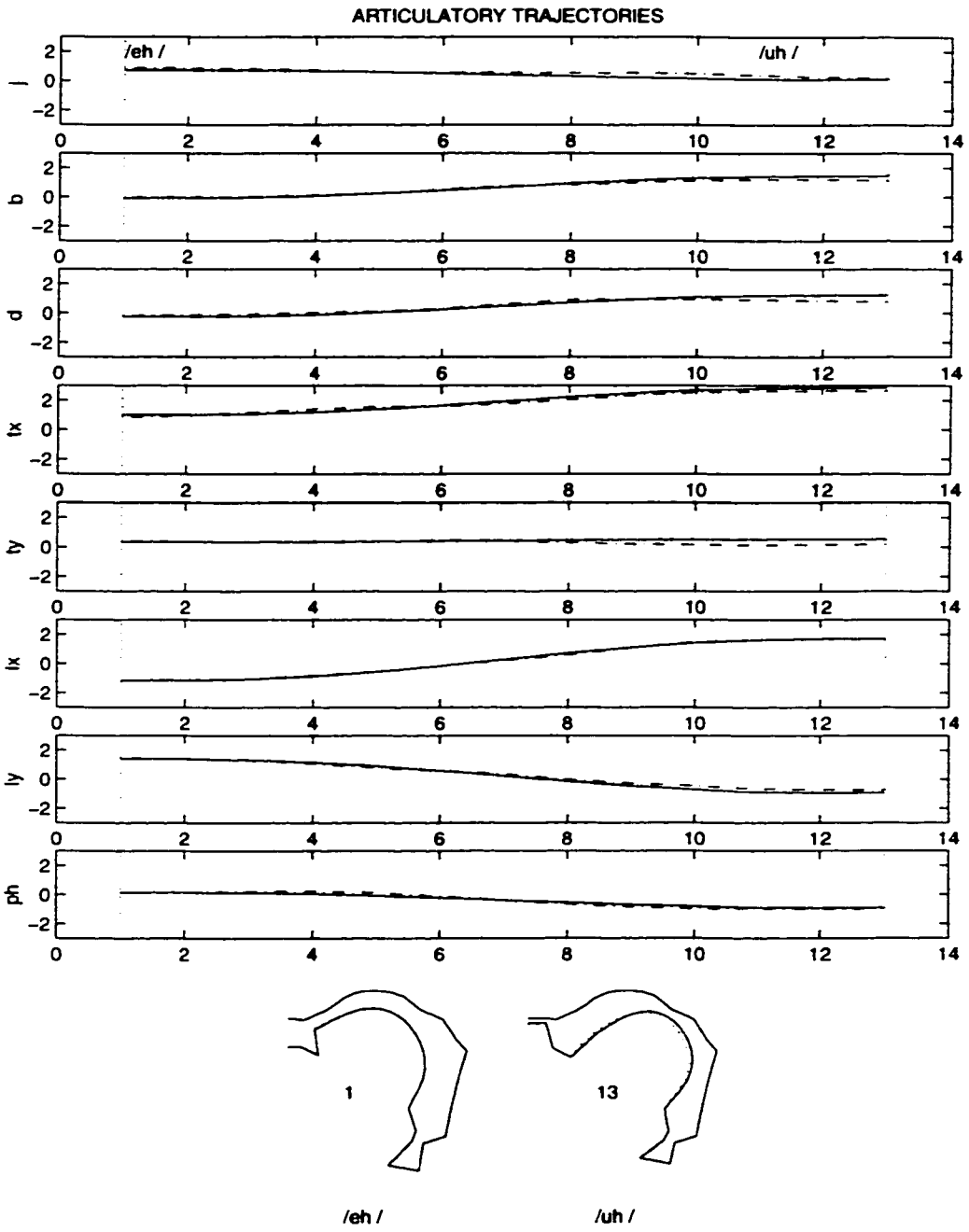

Figure 5.16: Actual and estimated articulatory trajectories for a segment /eh uh/

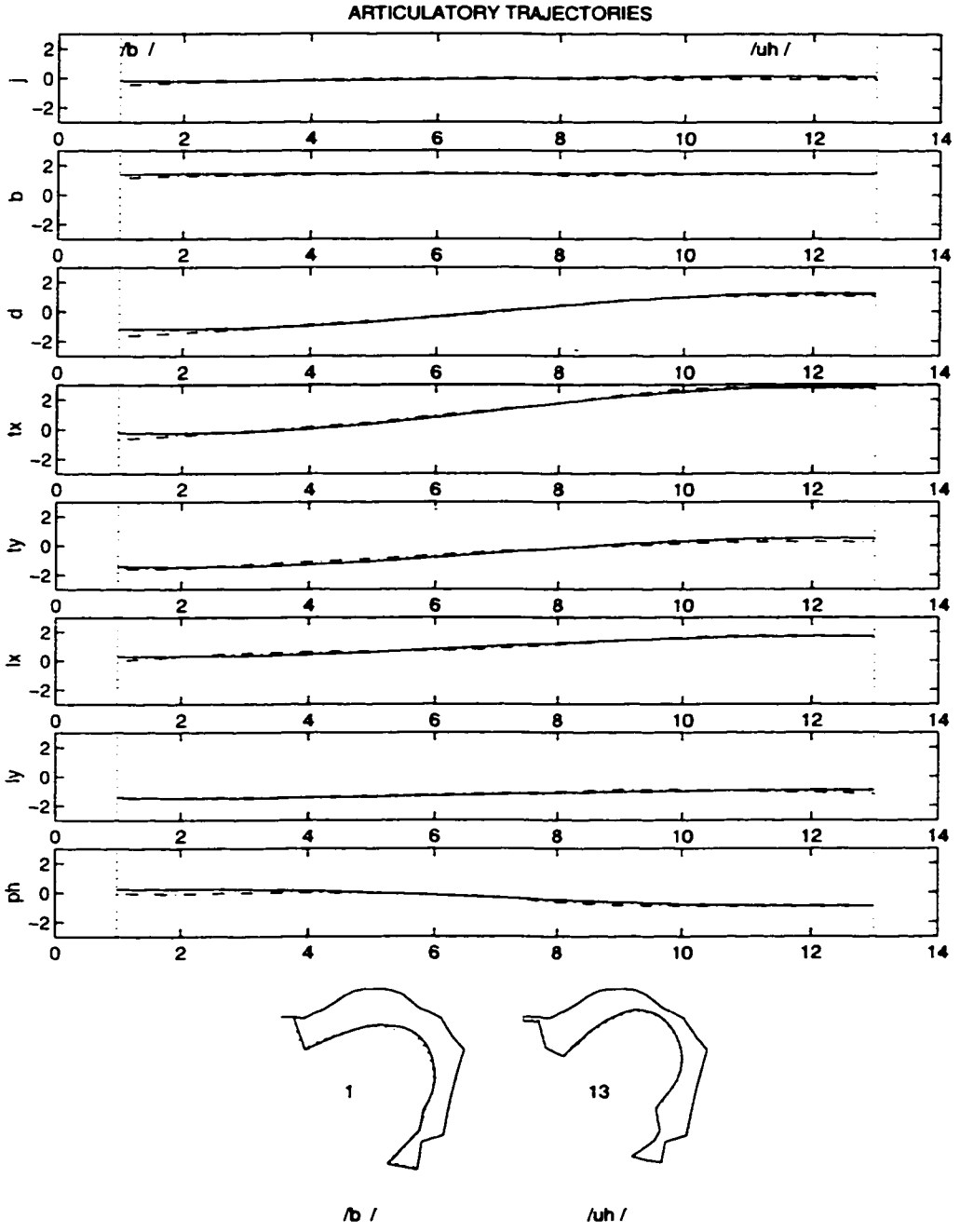

Figure 5.17: Actual and estimated articulatory trajectories for a segment /b uh/

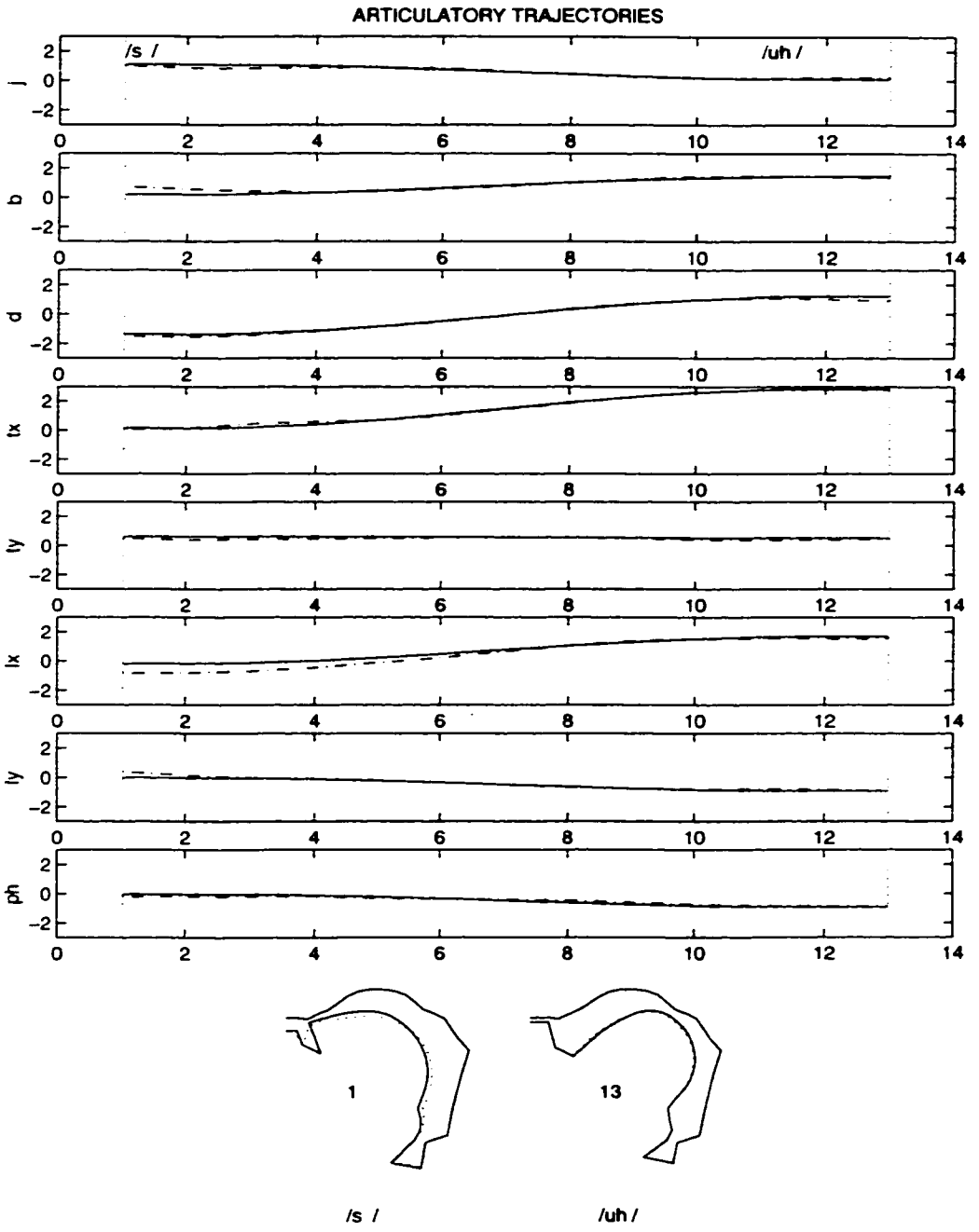

Figure 5.18: Actual and estimated articulatory trajectories for a segment /s uh/

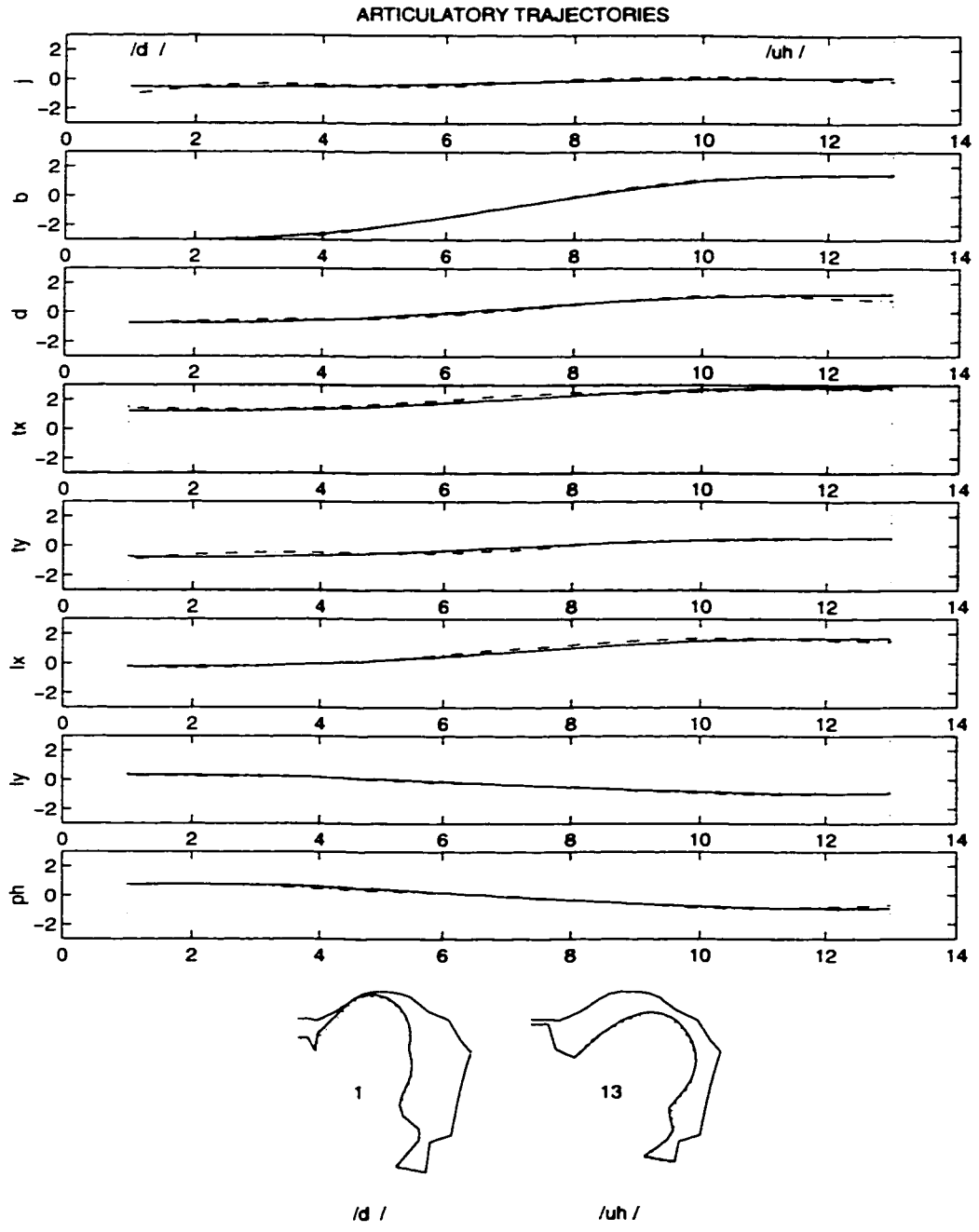

Figure 5.19: Actual and estimated articulatory trajectories for a segment /d uh/

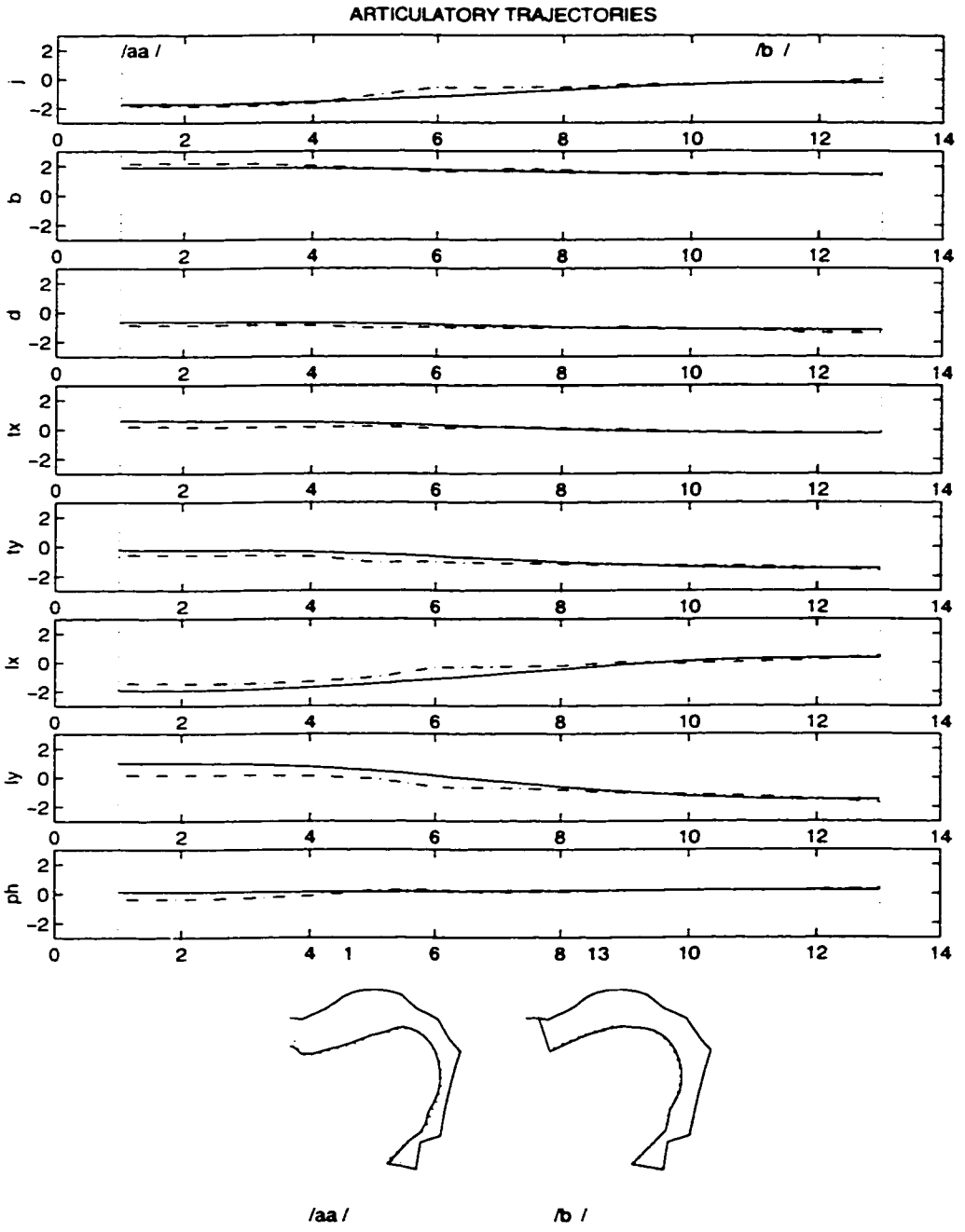

Figure 5.20: Actual and estimated articulatory trajectories for a segment /aa b/

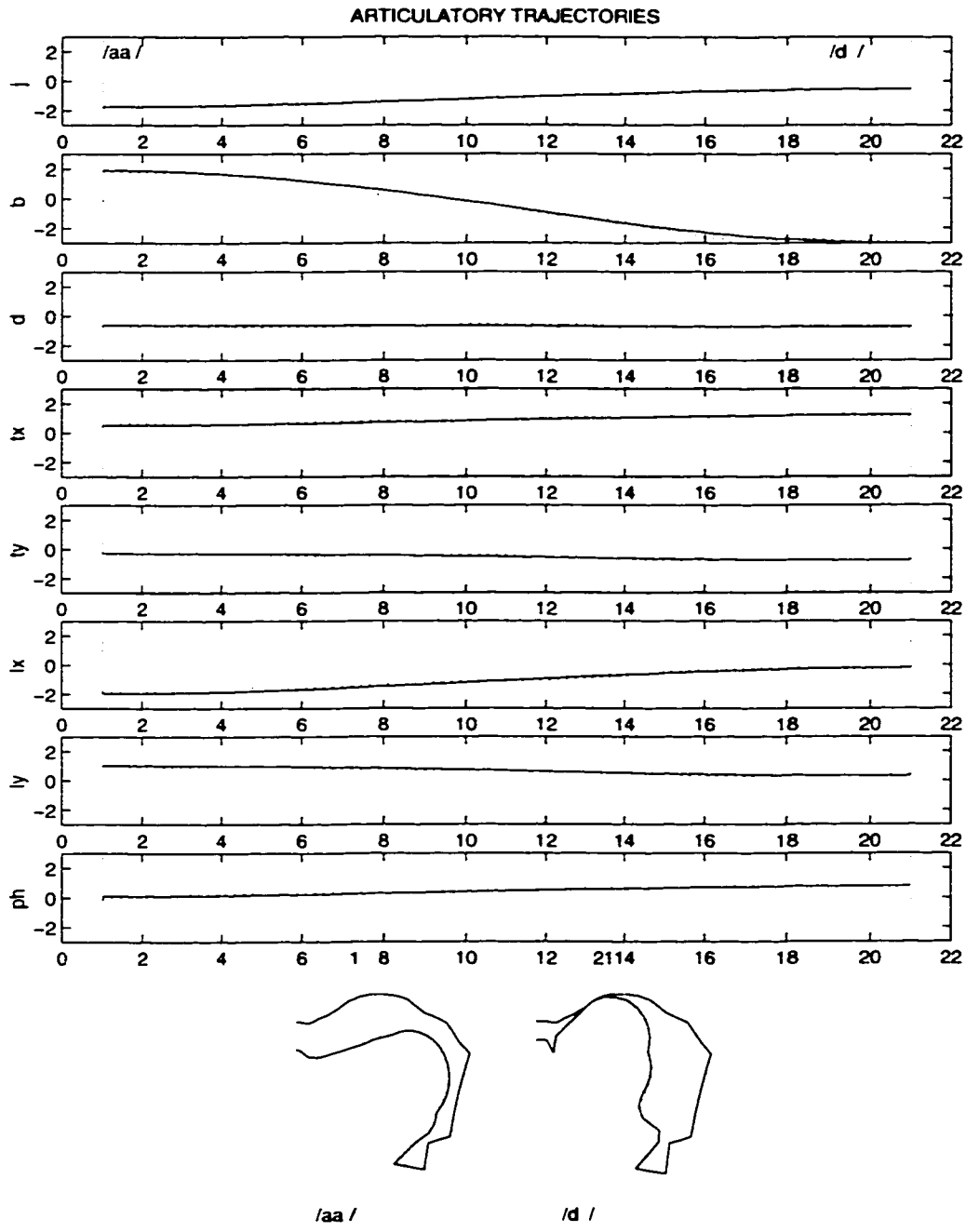

Figure 5.21: Actual and estimated articulatory trajectories for a segment /aa d/

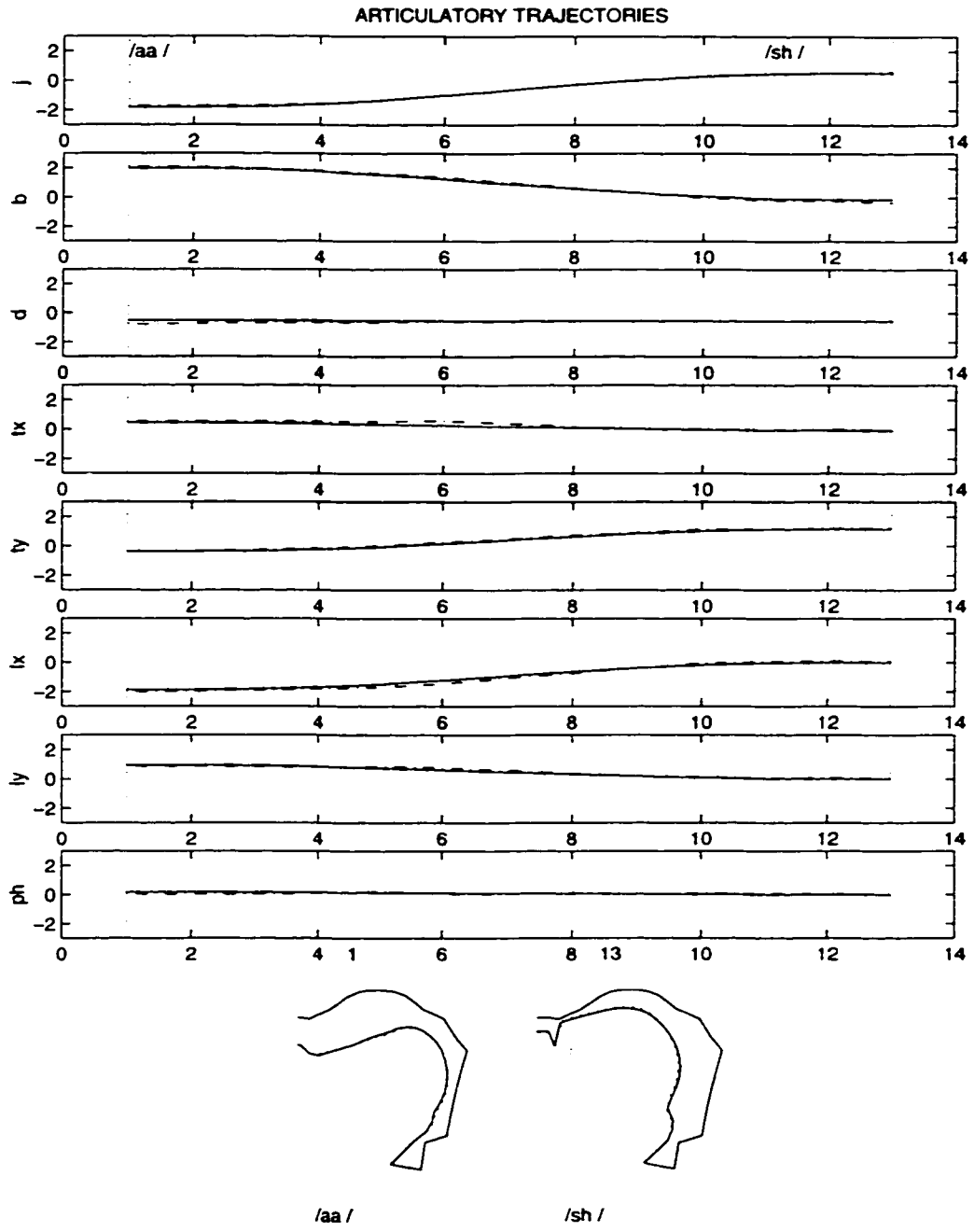

Figure 5.22: Actual and estimated articulatory trajectories for a segment /aa sh/

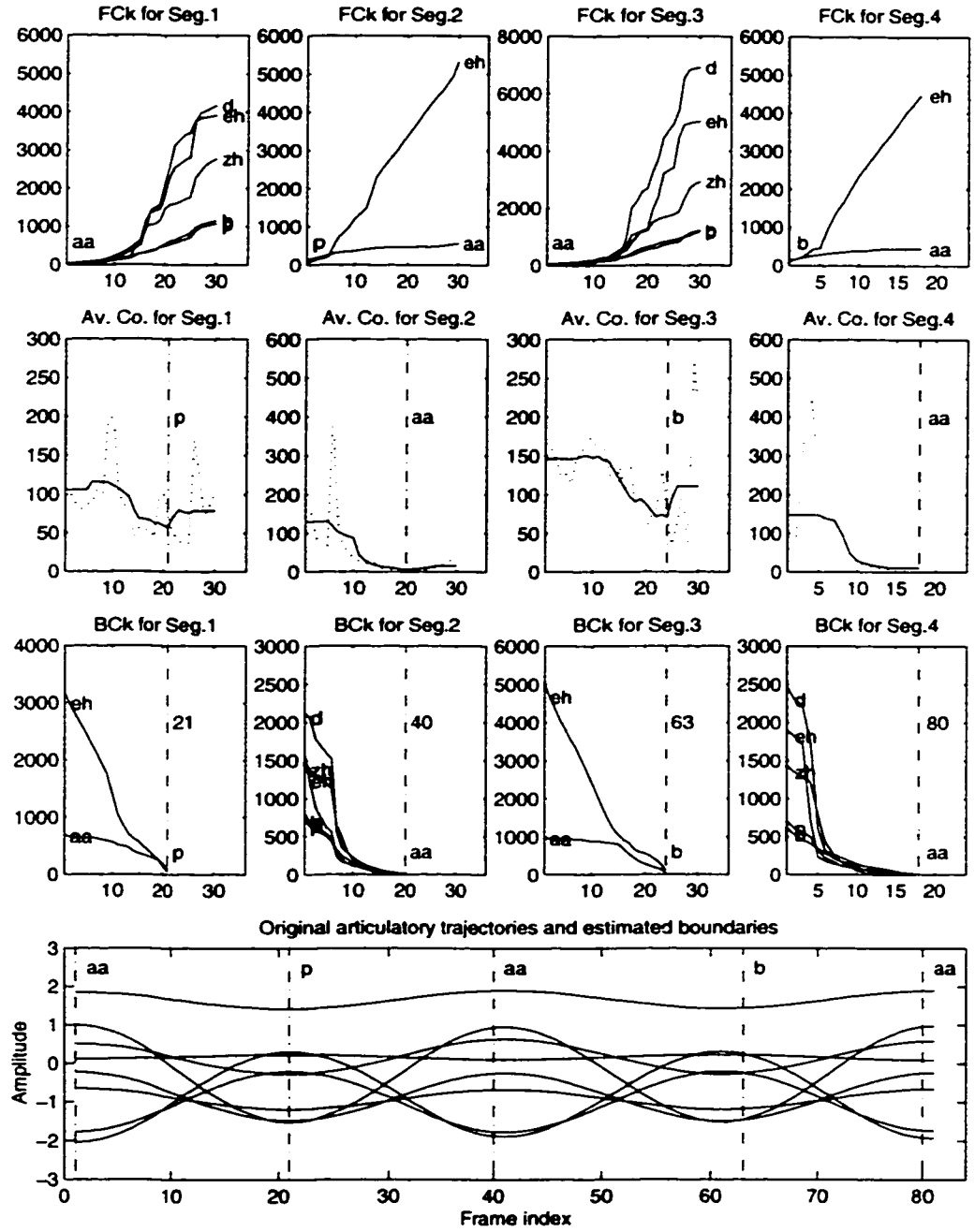

Figure 5.23: Automatic segmentation and recognition of models for an utterance /aa p aa b aa/

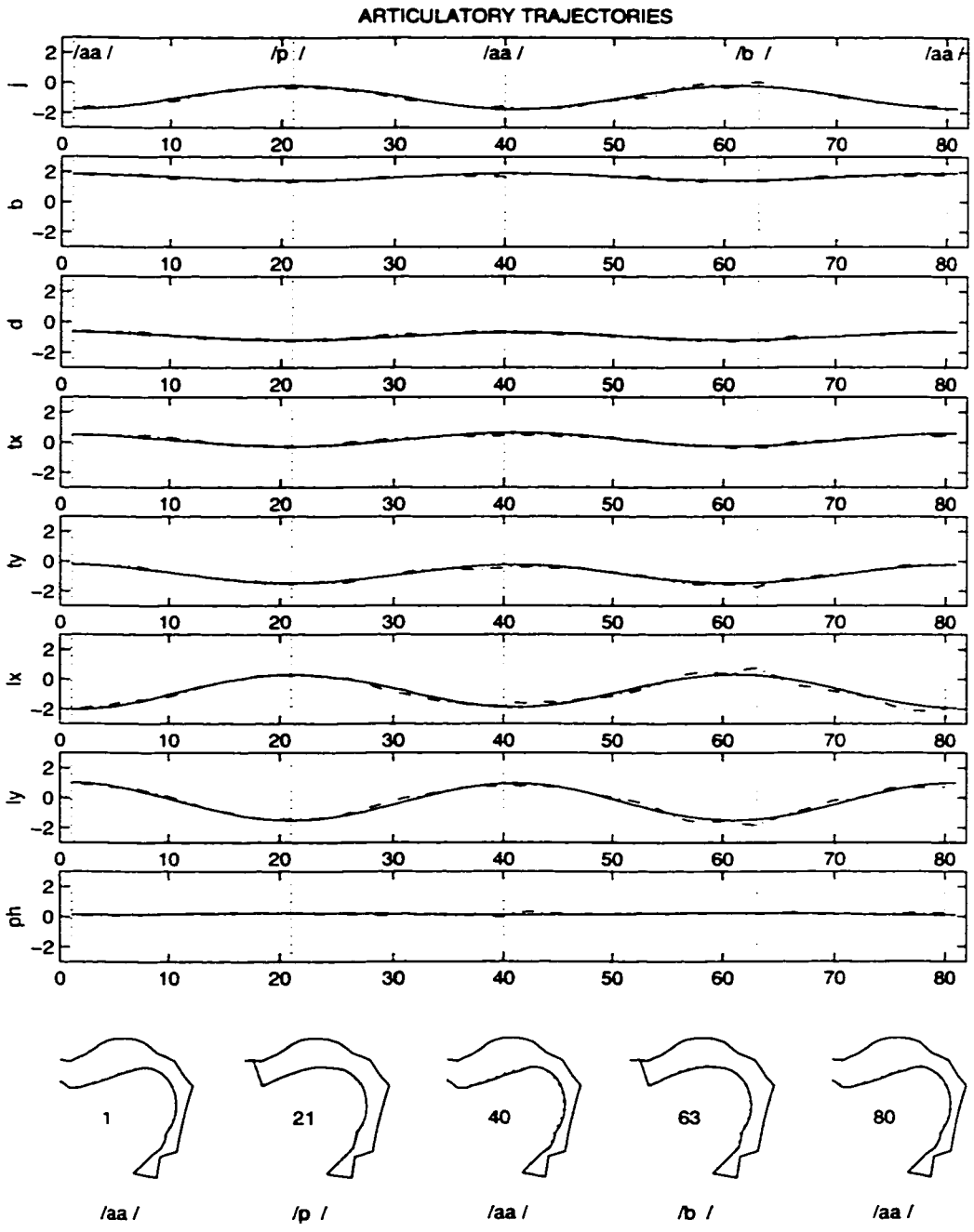

Figure 5.24: Actual and estimated articulatory trajectories for an utterance /aa p aa b aa/

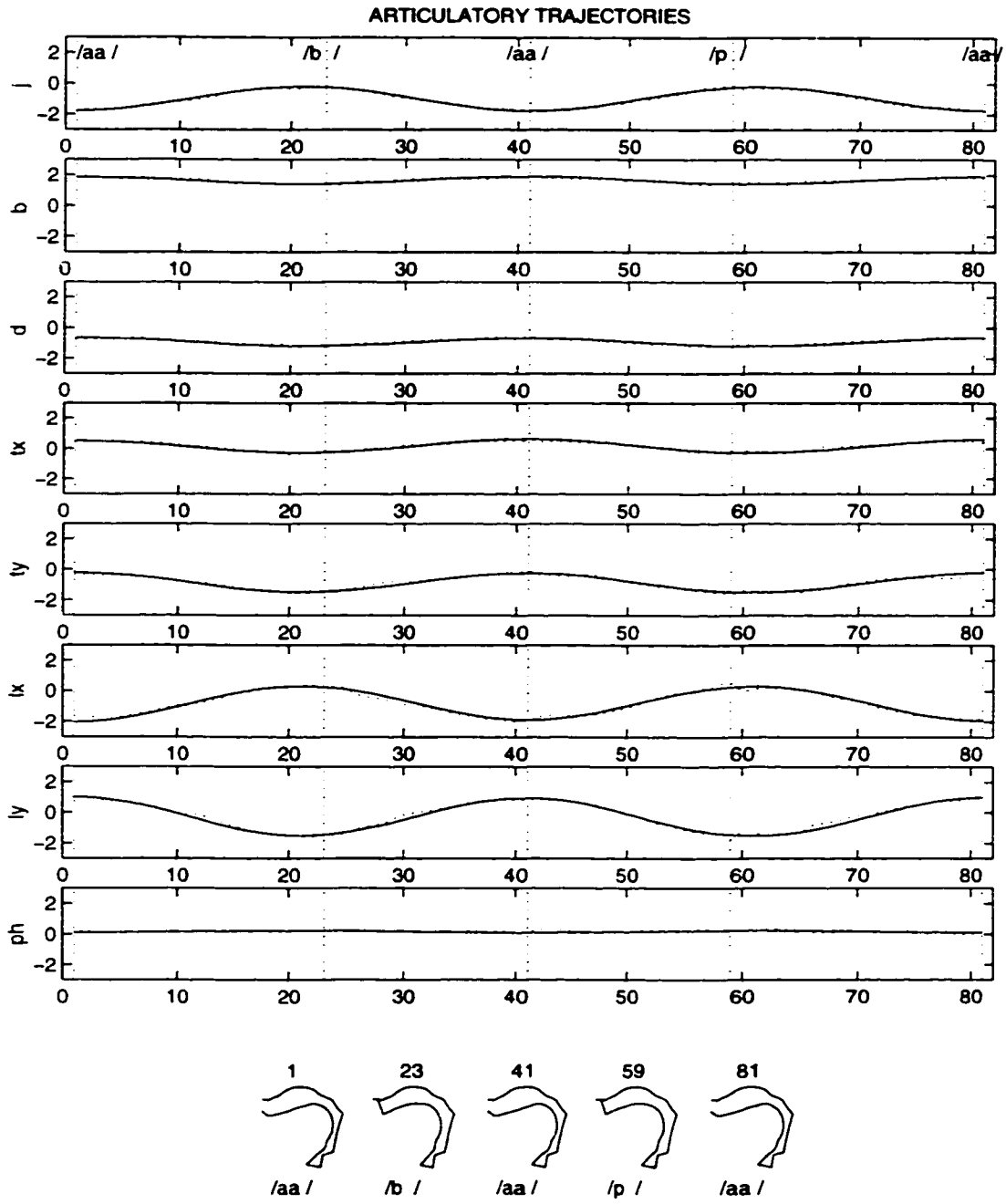

Figure 5.25: Actual and estimated articulatory trajectories for an utterance /aa b aa p aa/

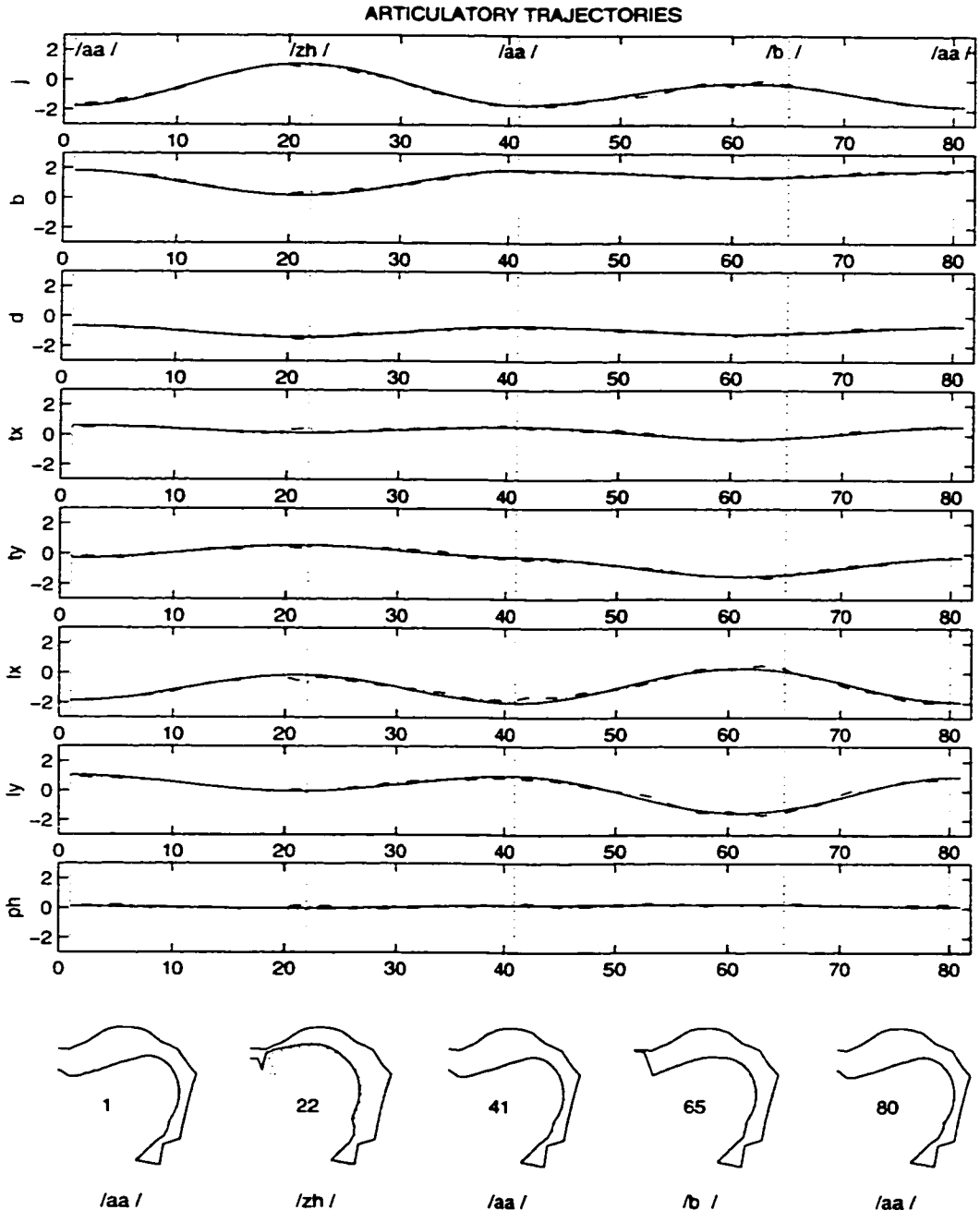

Figure 5.26: Actual and estimated articulatory trajectories for an utterance /aa zh aa b aa/

5.2 Experimental Results based on EMA Speech Data

In this section, we present experimental results obtained using Electromagnetic Midsagittal Articulography (EMA) data. The subject from whom the speech data were recorded was the author of this thesis. The articulatory-acoustic speech data recorded with the articulograph (Carstens [10]), consisted of VCV utterances containing the combinations of five English vowels — /ah/, /eh/, /iy/, /ao/ and /uh/, and 17 consonants — /b/, /d/, /f/, /g/, /h/, /zh/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /sh/, /t/, /v/ and /z/. Each VCV utterance was repeated three times. In each of the VCV utterances the first and second vowel was the same. We used three coils to trace the movement of the articulators, and they were placed on the lower lip (LL), about 2.5 cm back from the tongue tip (TT2) and about 4.5 cm back from the tongue tip (TT3). For each coil the X and Y coordinates were recorded. A total of 6 articulatory parameters were used in the articulatory vectors. The articulatory data were sampled at a frequency of 125 Hz by the articulograph. The acoustic speech signal was recorded with a sampling rate of 16,000 Hz. Like in the previous experiment, the first 10 MFCC parameters were computed from the speech signal from a window of 32 ms and a frame step of 10 ms. To synchronize the articulatory parameters with the acoustic parameters a re-sampling of the articulatory trajectories was carried out with a frequency of 100 Hz. Thus, both articulatory and acoustic parameters were finally available for frames separated at 10 ms.

Figure 5.27 presents the articulatory and acoustic data for the recorded utterance /ah b ah ah b ah ah b ah/. In Figure 5.28 the automatic recognition of the model is presented. A model was created using the first segment /ah b/ from this utterance and the random generation of new segments from the original one

using polynomial functions, as described in Section 4.2.1. Examples of estimated articulatory trajectories from a segment /ah b/, included in the training data are displayed in Figure 5.29. In the 7th sub-plot of this figure the root-mean-squared (RMS) error between the actual and estimated articulatory states is plotted. Figure 5.30 presents details from the bottom of the previous figure, of the positions of the three sensors in the vocal-tract midsagittal plane. In Figure 5.31, the corresponding actual and reconstructed MFCC trajectories are presented. In the following we present the set of model parameters estimated using ML method from the segment /ah b/ included in training data

$$\mathbf{\Phi}_1^{(ah,b)} = diag \begin{bmatrix} 1.220 \\ 1.862 \\ 0.895 \\ 1.444 \\ 1.310 \\ 1.013 \end{bmatrix}, \quad \mathbf{\Phi}_2^{(ah,b)} = diag \begin{bmatrix} -0.235 \\ -0.861 \\ 0.106 \\ -0.445 \\ -0.308 \\ -0.014 \end{bmatrix},$$

$$\mathbf{Q}^{(ah,b)} = diag \begin{bmatrix} 0.0014 \\ 0.0025 \\ 0.0002 \\ 0.0002 \\ 0.0001 \\ 0.0004 \end{bmatrix}, \quad \mathbf{R}^{(ah,b)} = diag \begin{bmatrix} 1.365 \\ 0.144 \\ 0.237 \\ 0.439 \\ 0.761 \\ 0.178 \\ 0.686 \\ 0.207 \\ 0.415 \\ 0.199 \end{bmatrix}.$$

A similar example, for a segment /ah b/ not included in training data is presented in Figure 5.32, Figure 5.33 and Figure 5.34. For this case the RMS errors are also around the value of 1 mm. In general, the RMS errors are larger for the case of segments not included in training data, as expected. This effect is mainly caused by the coproduction model observation sub-function, rather than the dynamical model parameters. For this case, the details regarding the positions of the three sensors are displayed in Figure 5.33. As in Figure 5.30, the upper curves in these plots represent the shape of the palate of the speaker. These curves start at the incisors (left), and end at the soft palate (right). The last parts of these curves (about 1 cm at the right) represent a portion of the soft palate, which is usually not fixed as depicted by these plots.

The articulatory and acoustic data recorded for the /ah m ah ah m ah ah m ah/ utterance are presented in Figure 5.35, and the estimation results for a segment /ah m/, included in training data are presented in Figure 5.36. The RMS errors are about 0.5 mm. The estimation results for a segment /ah m/, not included in training data are presented in Figure 5.37. The corresponding RMS errors are higher in this case and they are about 2.3 mm.

Other estimation results, for the segments /ah t/ and /ah g/ not included in training data, and for the segments /ah s/ and /ah l/ included in training data, are presented in Figures 5.38 to 5.41. As expected, the RMS errors are in general larger for the cases of sequences not included in training data of the corresponding models. However, these RMS errors are reasonably small, of the order of 2.0 mm.

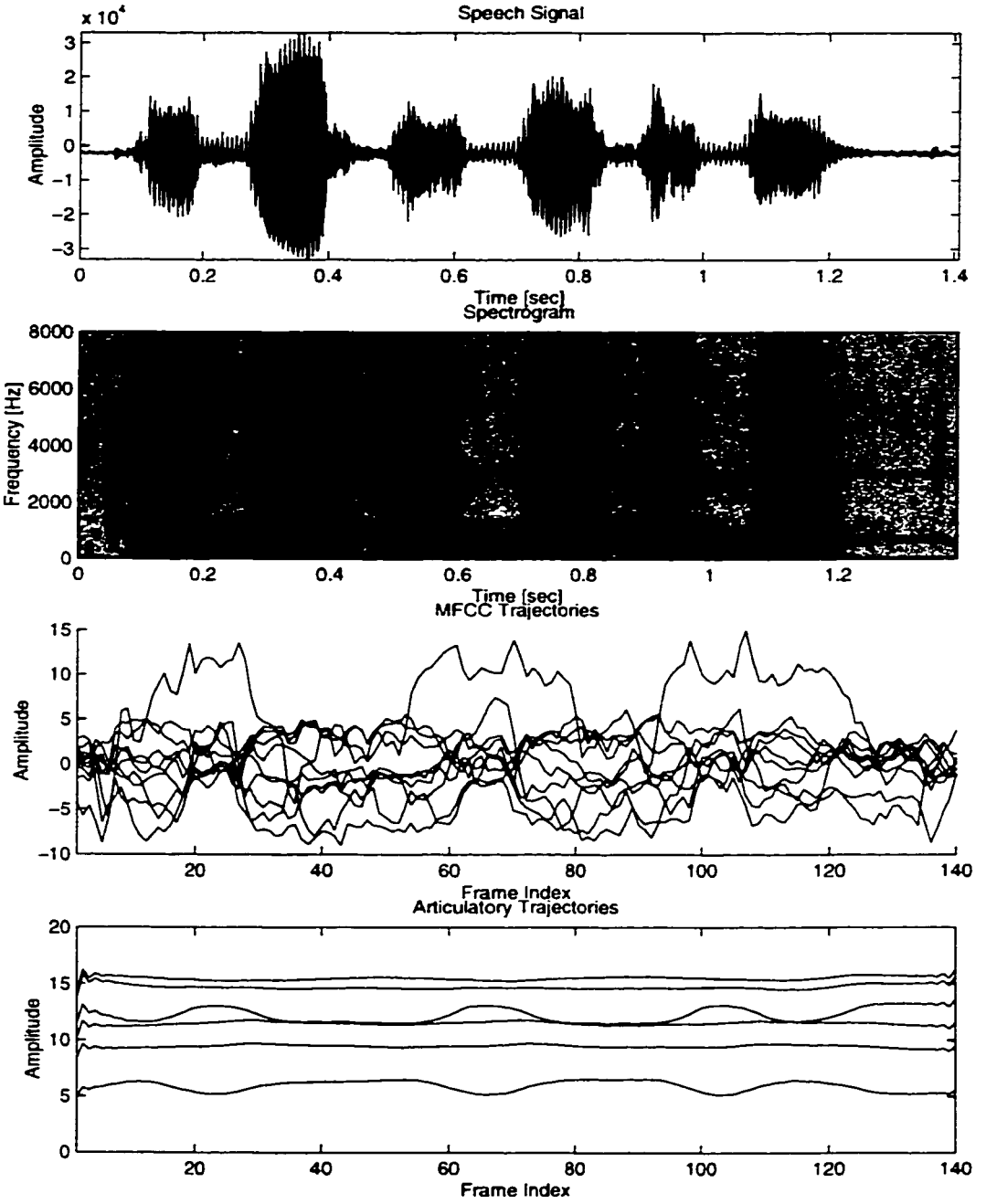

Figure 5.27: MFCC and articulatory trajectories for the /ah b ah ah b ah ah b ah/ utterance obtained from the Articulograph AG100

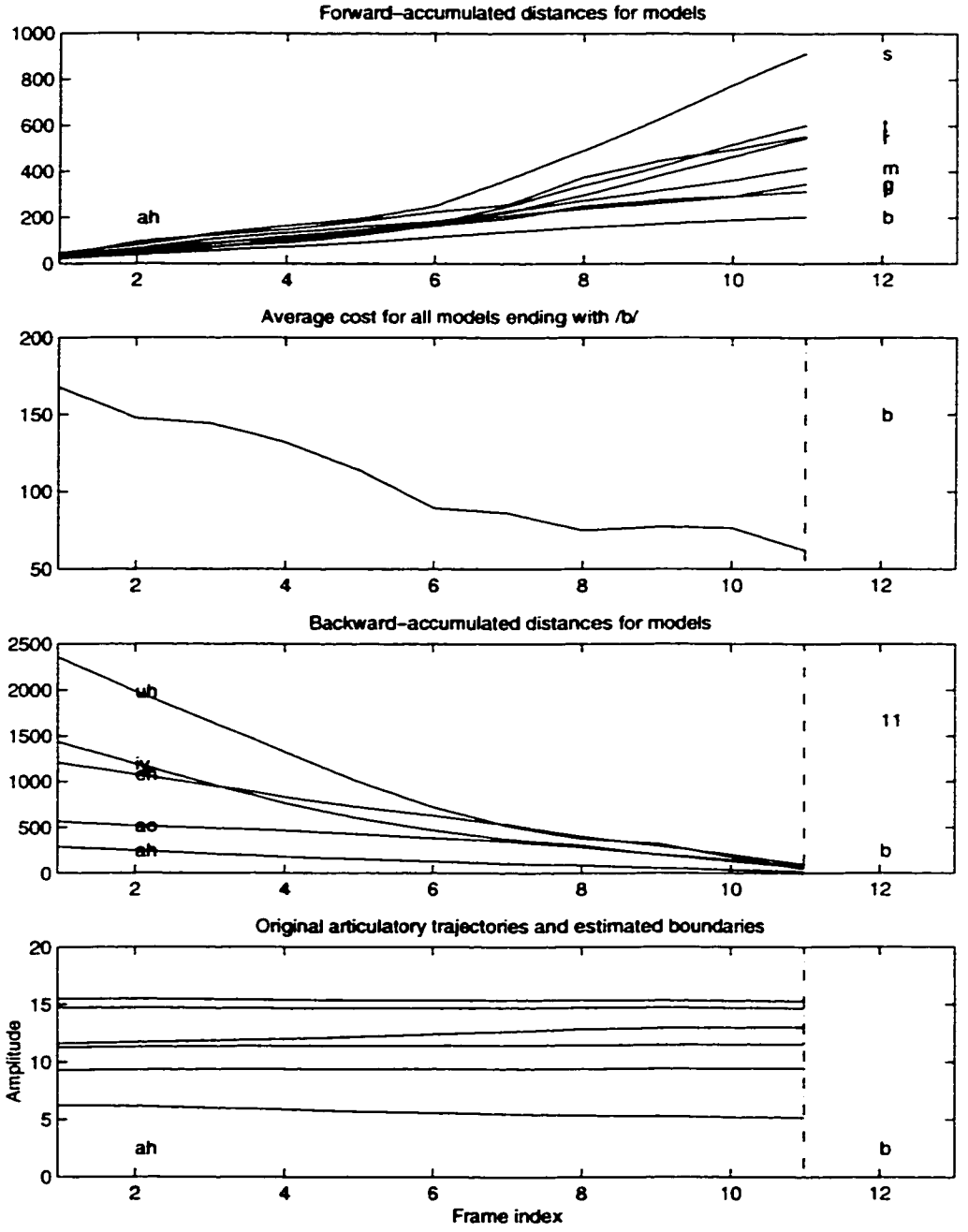

Figure 5.28: Recognition of (ah,b) model from an isolated segment /ah b/. included in training data

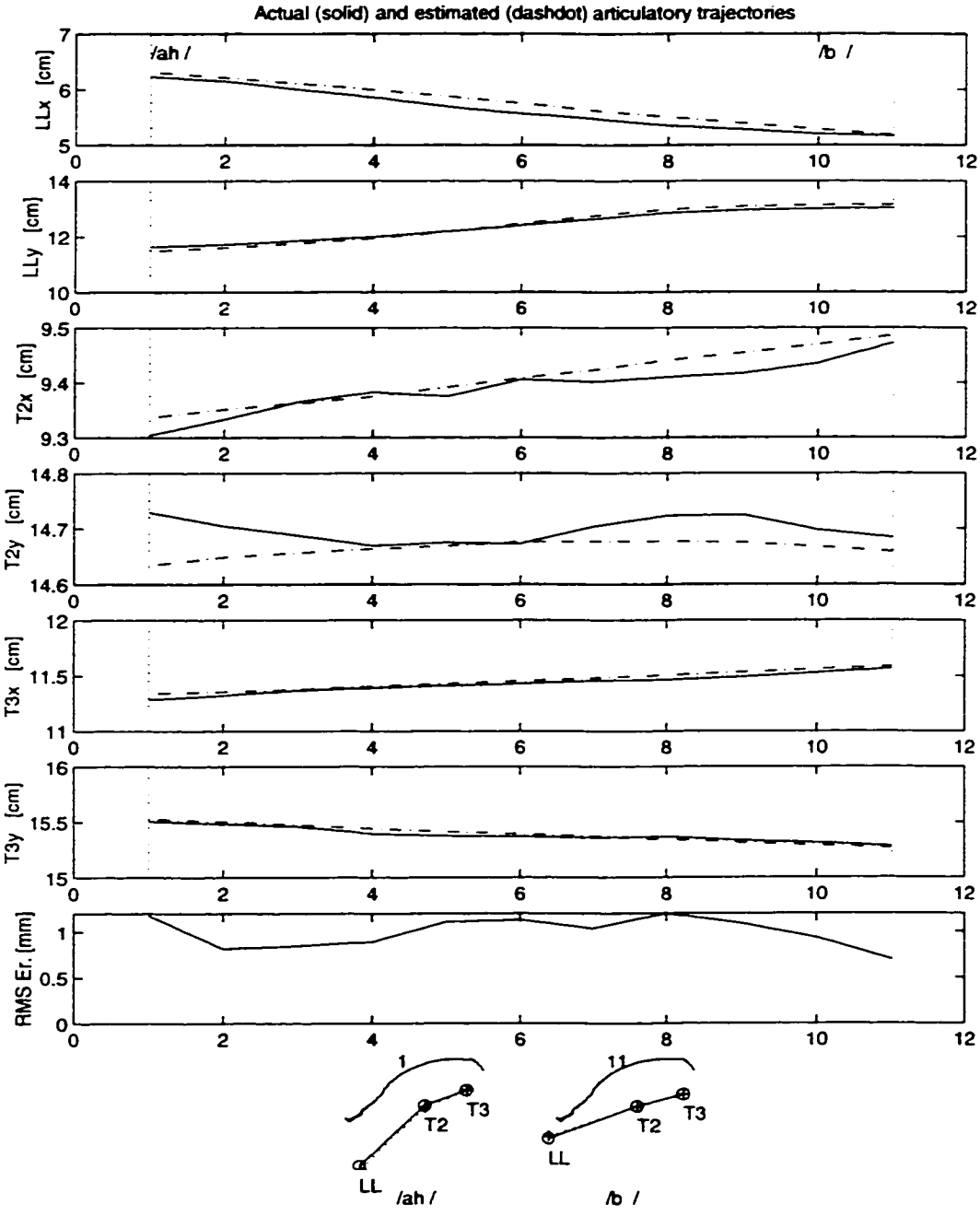

Figure 5.29: Actual and estimated articulatory trajectories for a segment /ah b/. included in training data

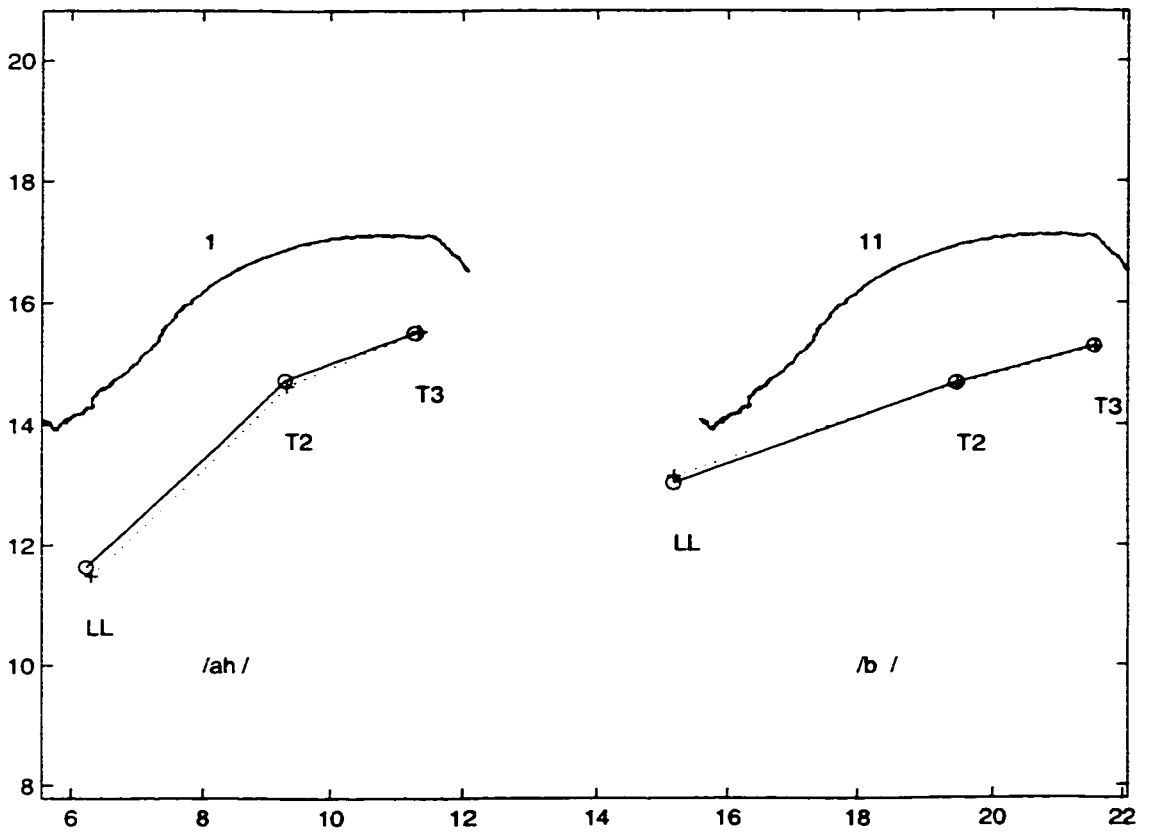

Figure 5.30: Actual and estimated VT profiles for a segment /ah b/, included in training data (detail from previous figure, axes in cm)

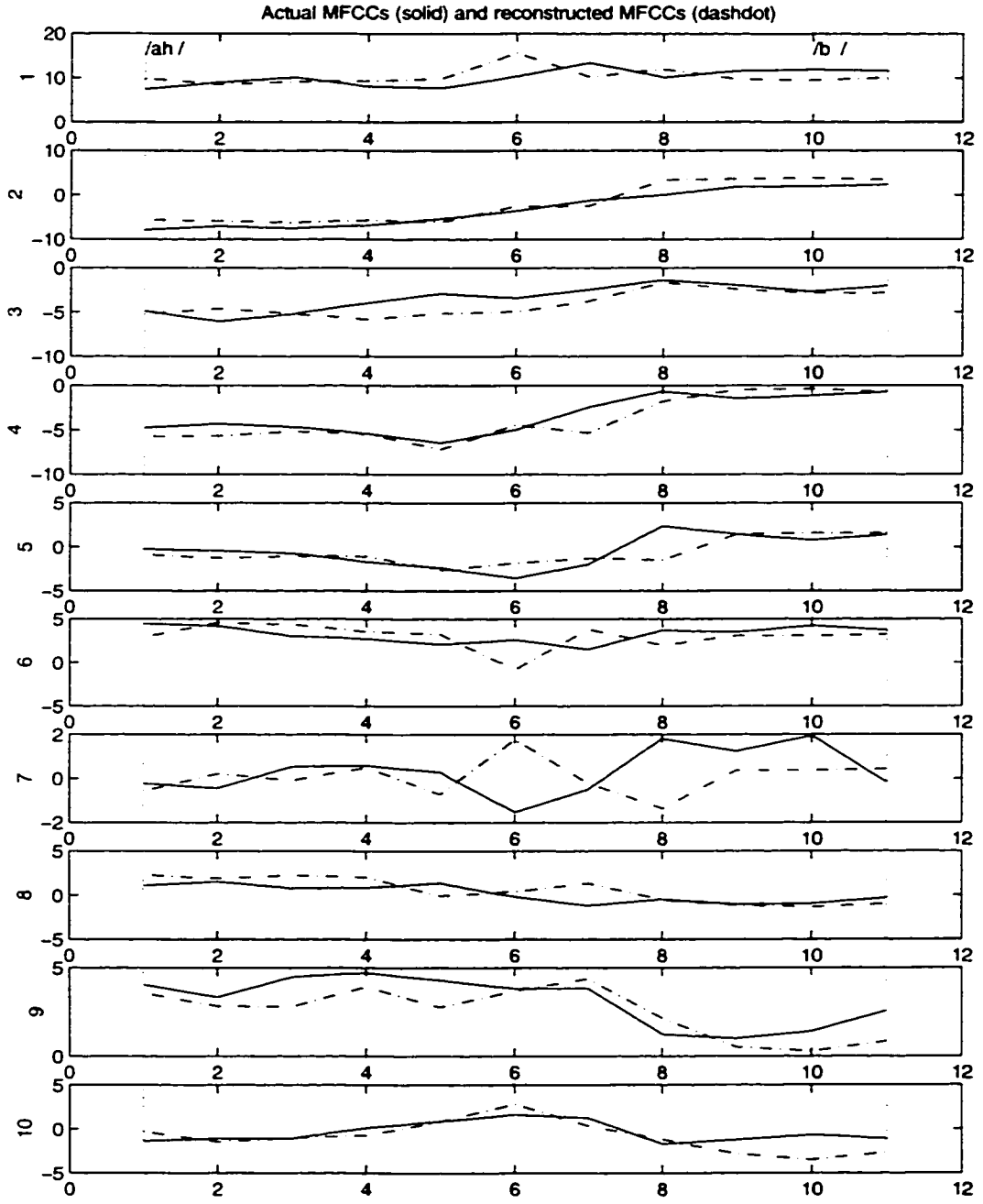

Figure 5.31: Actual and reconstructed MFCC trajectories for a segment /ah b/. included in training data

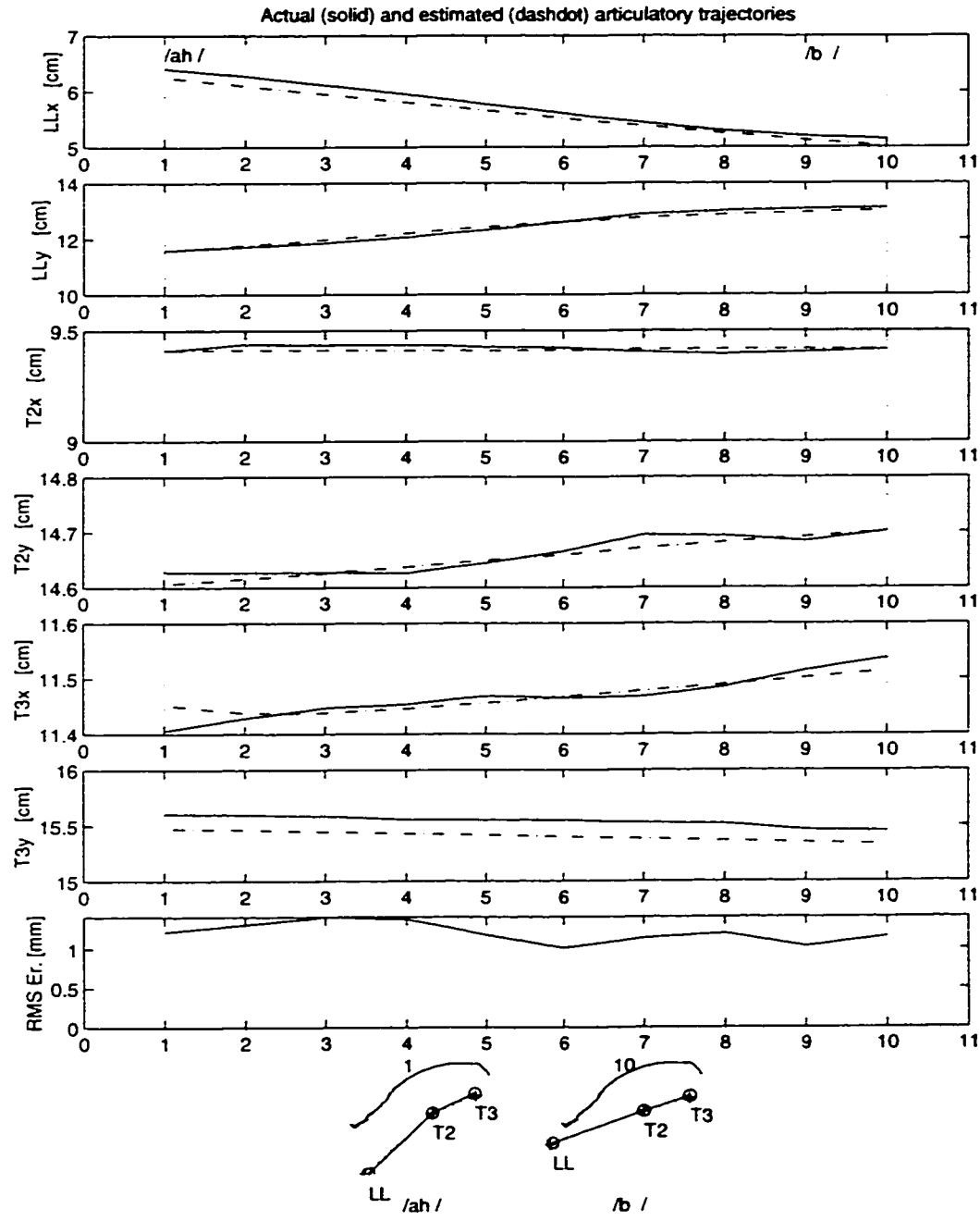

Figure 5.32: Actual and estimated articulatory trajectories for a segment /ah b/, not included in training data

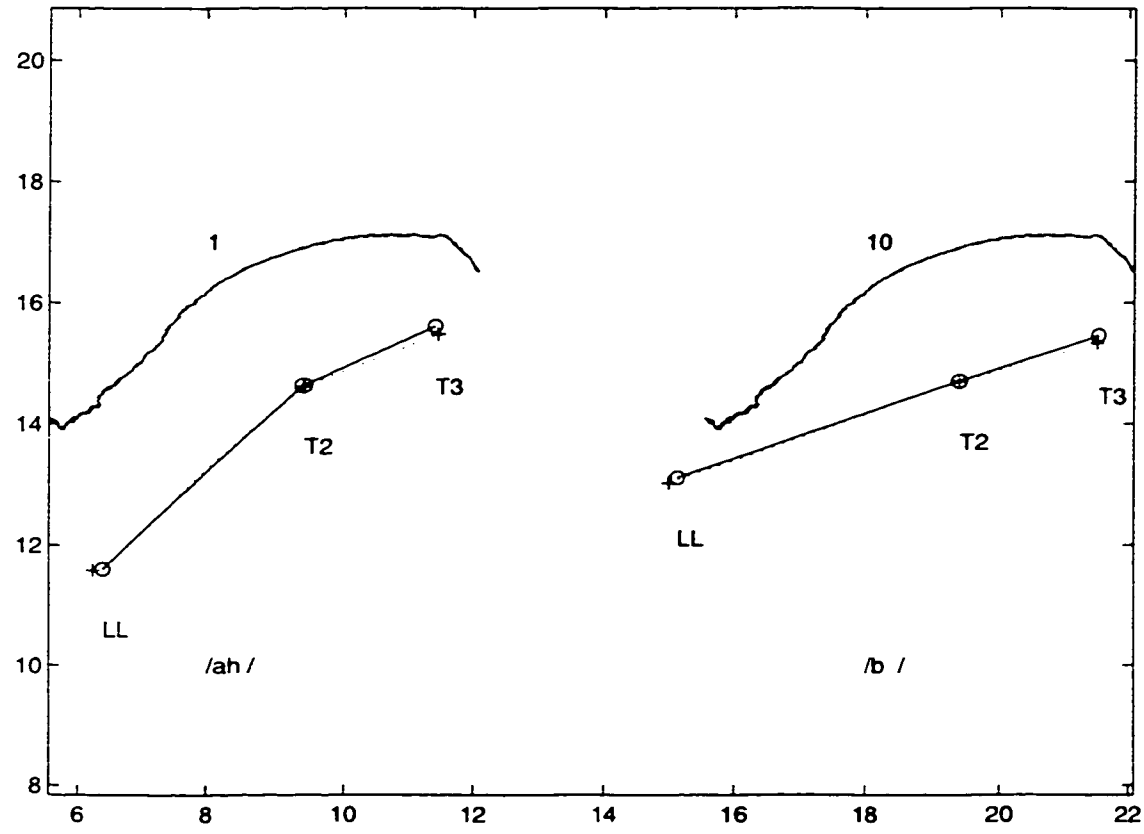

Figure 5.33: Actual and estimated VT profiles for a segment /ah b/, not included in training data (detail from previous figure, axes in cm)

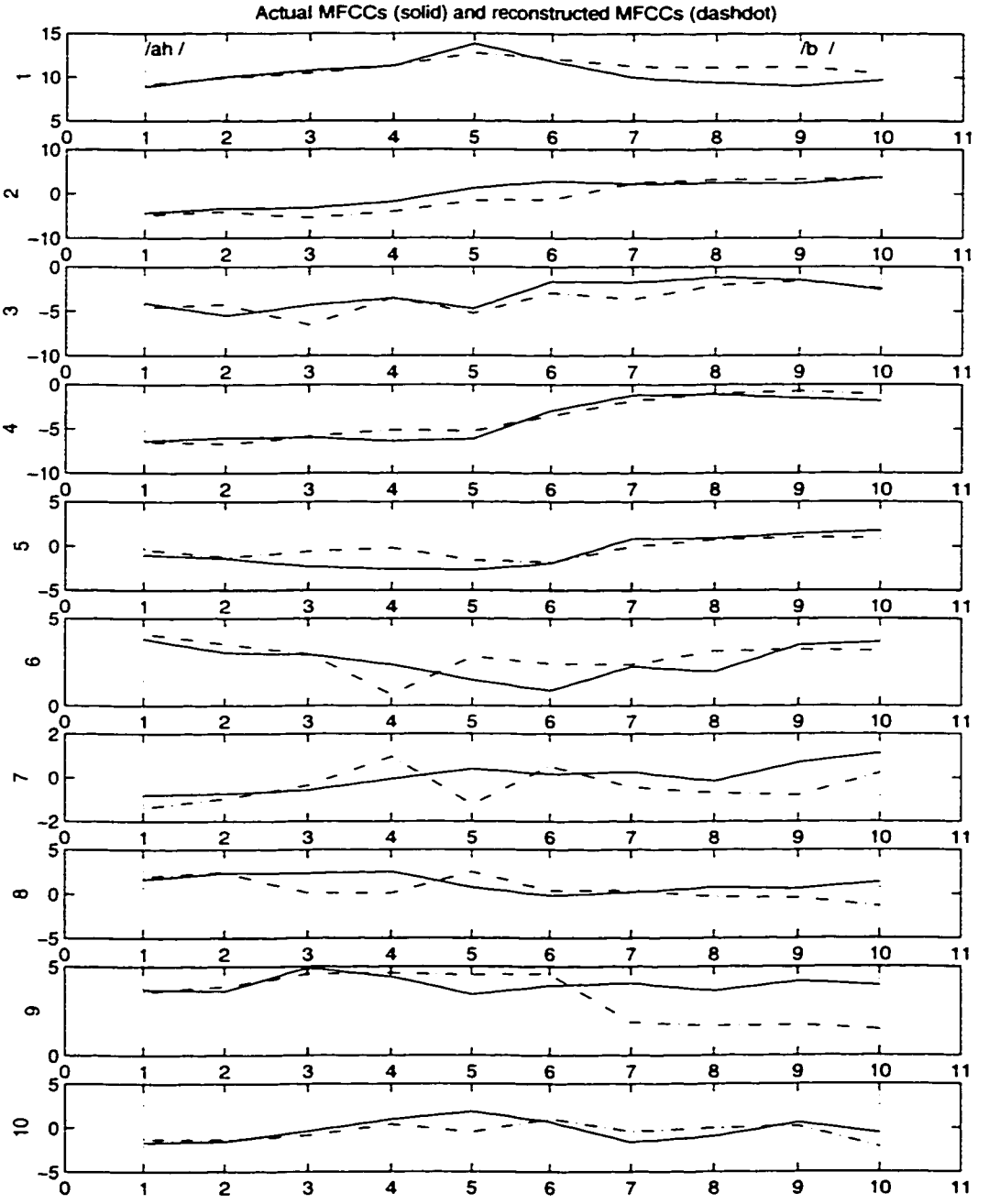

Figure 5.34: Actual and reconstructed MFCC trajectories for a segment /ah b/, not included in training data

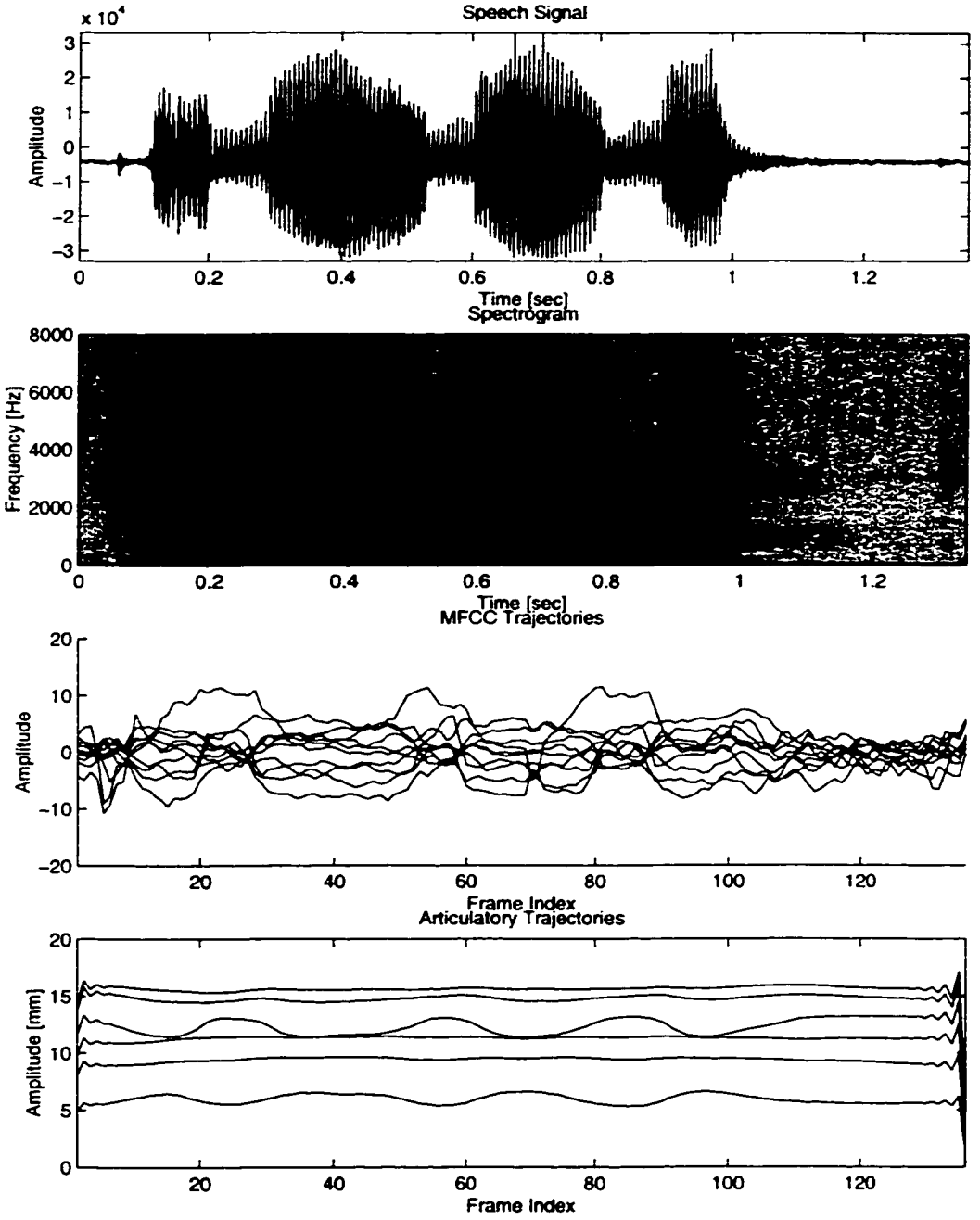

Figure 5.35: MFCC and articulatory trajectories for the /ah m ah ah m ah ah m ah/ utterance obtained from the Articulograph AG100

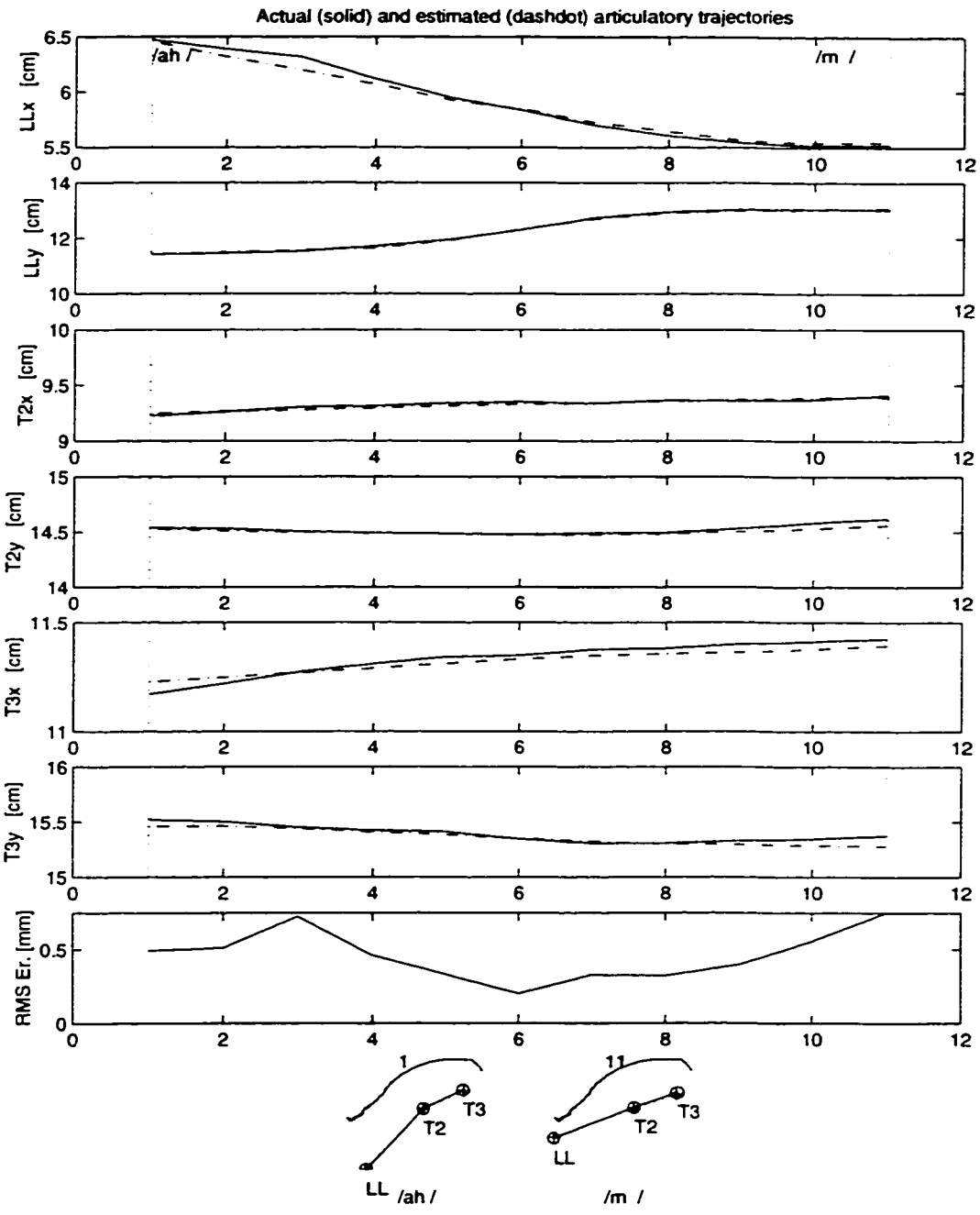

Figure 5.36: Actual and estimated articulatory trajectories for a segment /ah m/, included in training data

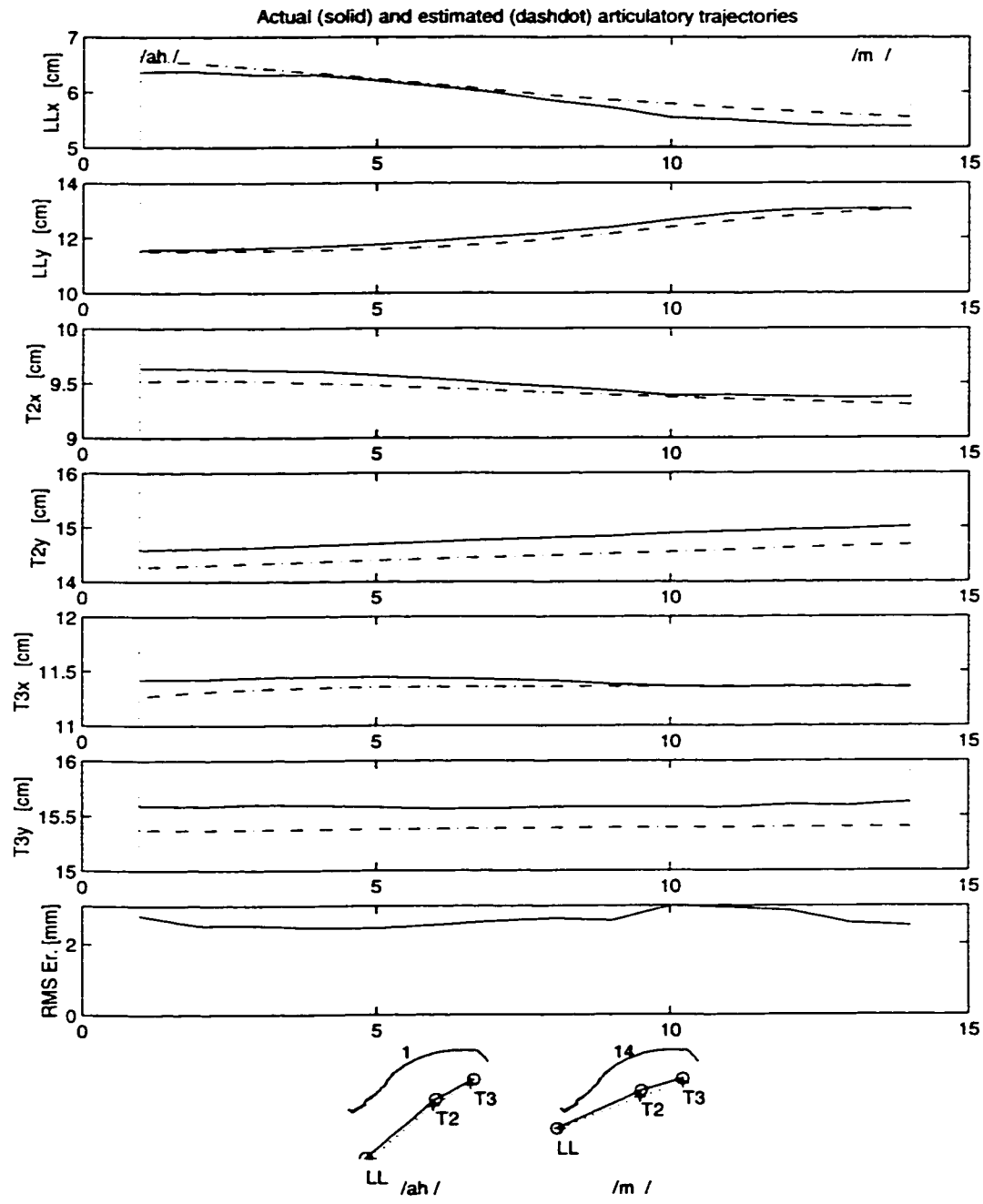

Figure 5.37: Actual and estimated articulatory trajectories for a segment /ah m/. not included in training data

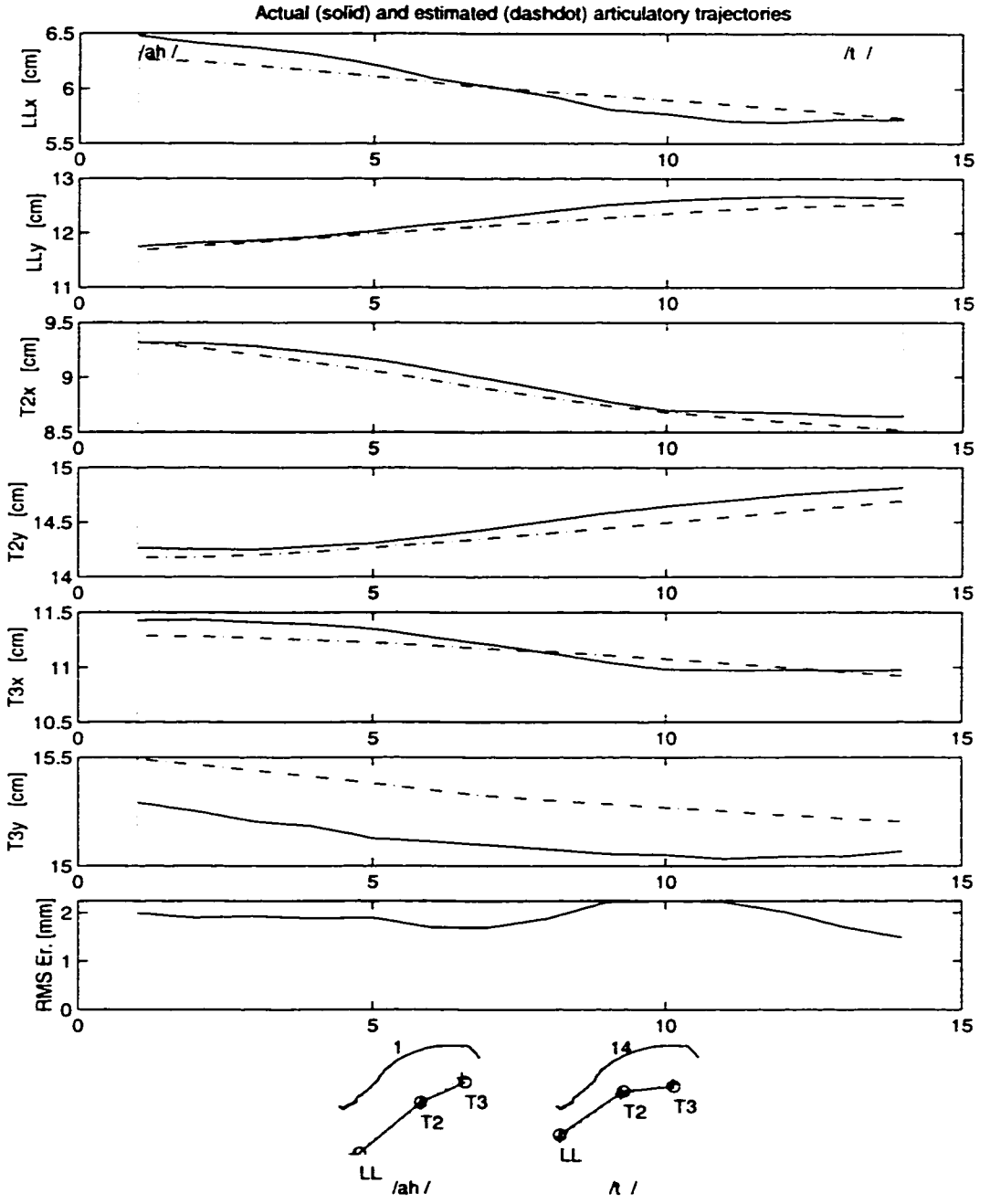

Figure 5.38: Actual and estimated articulatory trajectories for a segment /ah t/. not included in training data

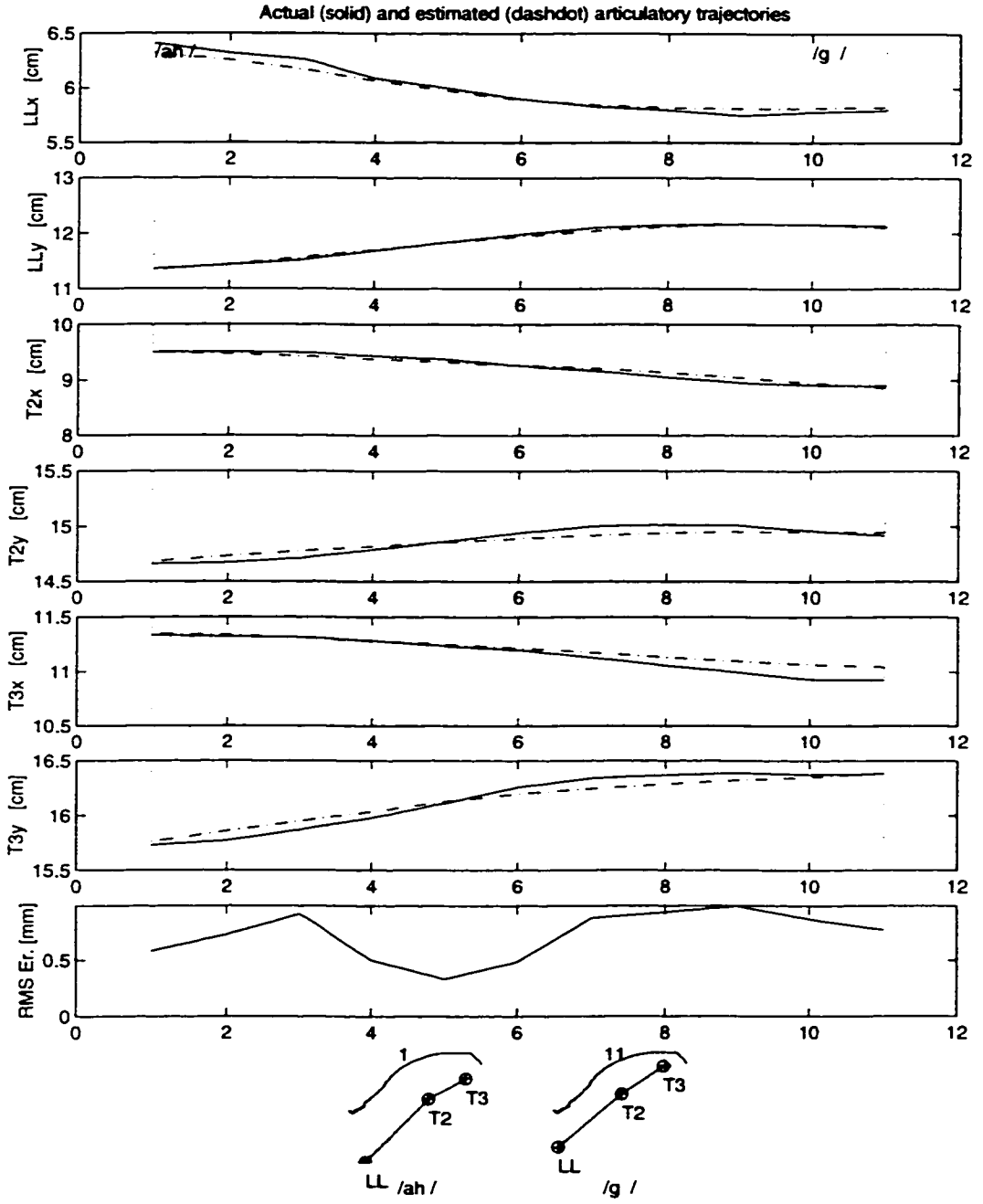

Figure 5.39: Actual and estimated articulatory trajectories for a segment /ah g/. not included in training data

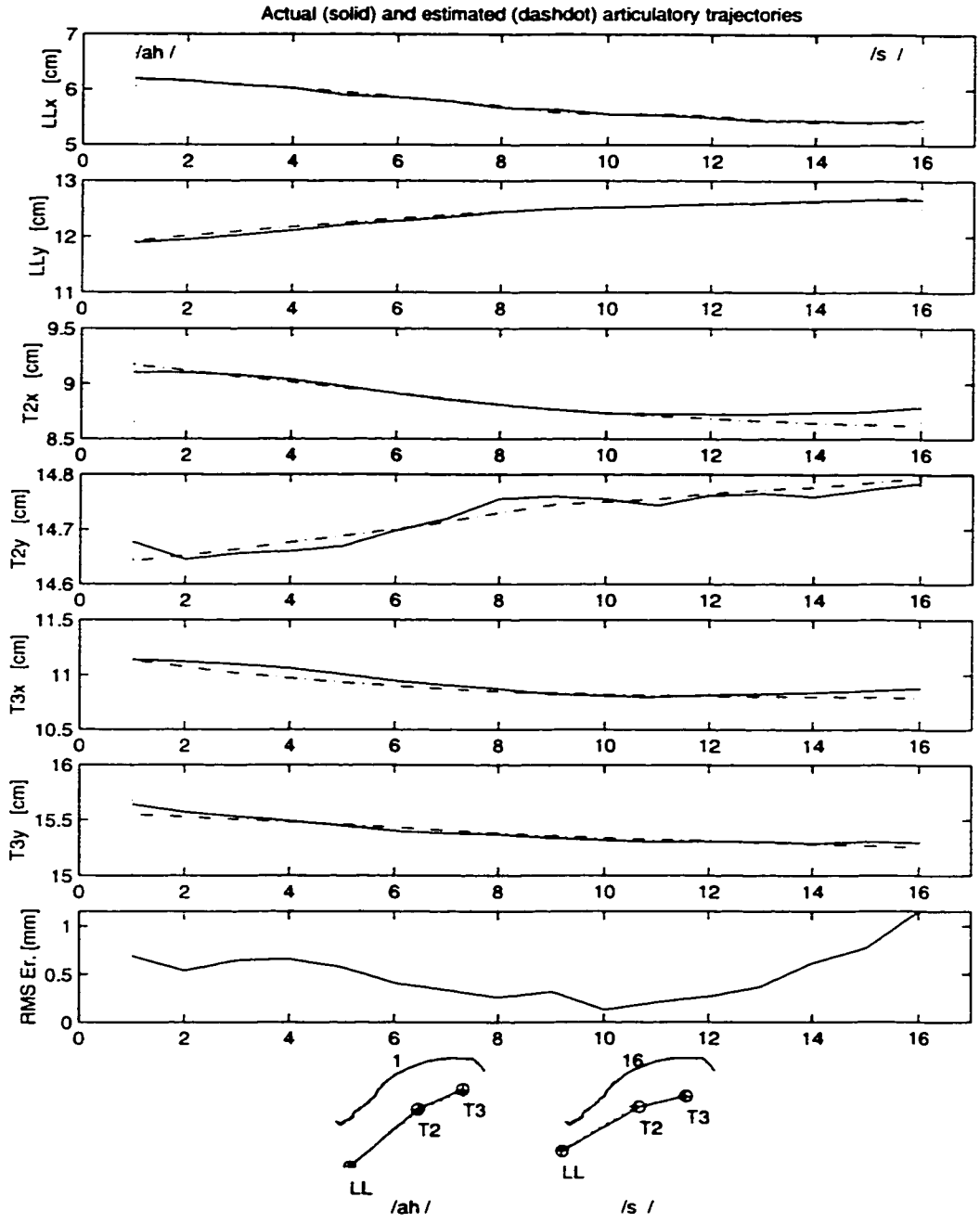

Figure 5.40: Actual and estimated articulatory trajectories for a segment /ah s/, included in training data

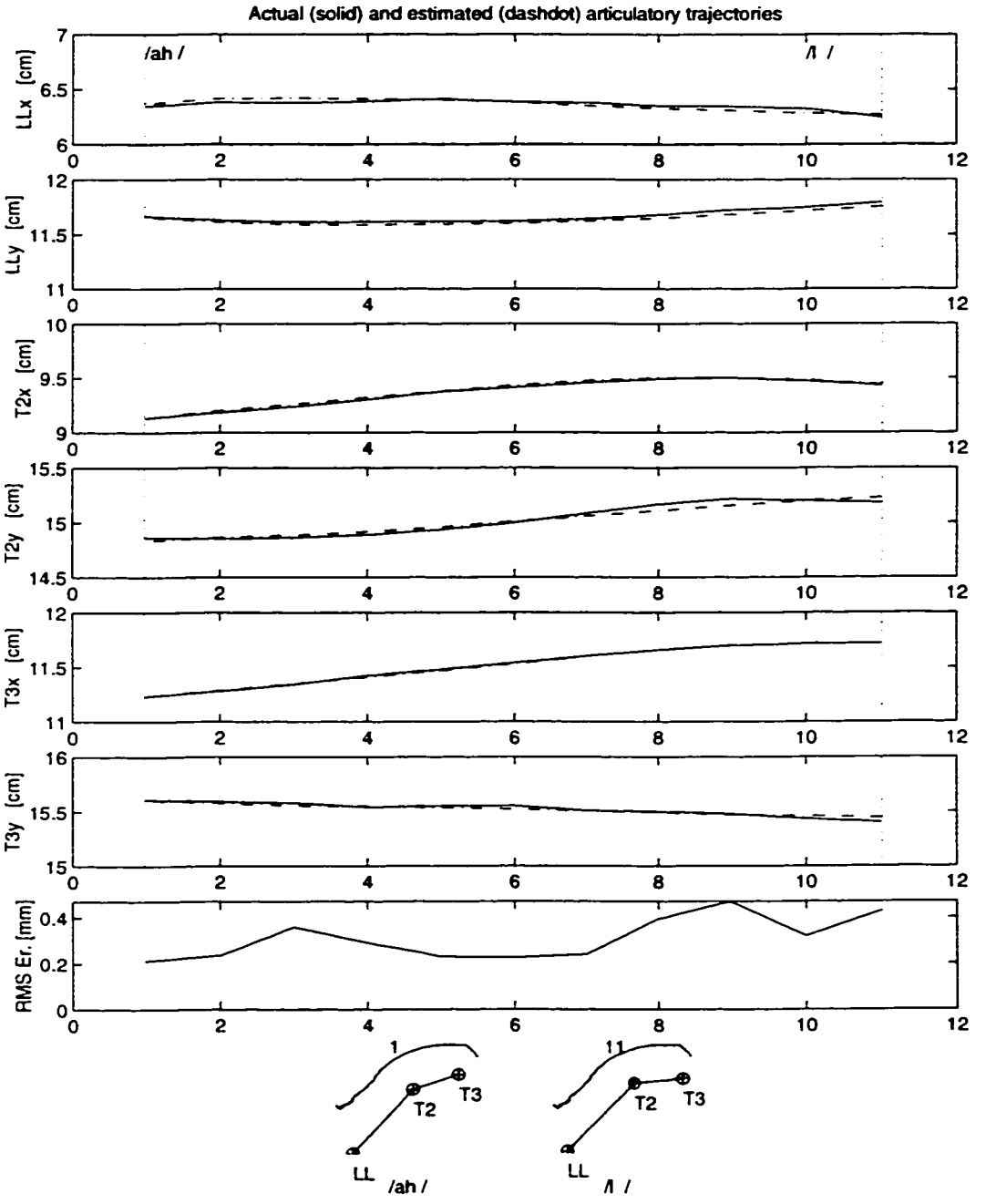

Figure 5.41: Actual and estimated articulatory trajectories for a segment /ah l/. included in training data

In the following, we present some experimental results of estimation based on continuous VCV utterances. In the next figures representing the automatic segmentation and recognition of models we used the same kind of sub-plots like in Figure 5.28, but because of the limited space we omitted the title of each sub-plot.

In Figure 5.42 we present the automatic segmentation and recognition of models for an utterance /ah b ah/ included in training data of the corresponding models. The sub-plots from the first raw represent the forward-accumulated cost functions (distances) for different models starting with /ah/ and /b/, respectively. The sub-plots from the second raw represent the average of the cost functions (distances) over all the models ending with /b/ and /ah/, respectively (dotted lines), and the smoothed average cost functions (solid lines). We used the minimum of the smoothed average distance to find the center of the second phoneme in each coproduction segment. This smoothed distance provided a better approximation of the position of the center of the second phoneme than the un-smoothed distance. The central frames of the second phoneme of each model were estimated very well. The sub-plots from the third raw represent the backward-accumulated distances of the models which end with /b/ and /ah/, respectively. We used these functions to reestimate the first phoneme of the utterance and also to check the forward estimation of the second phoneme of each coproduction segment. The sub-plot from the fourth raw represents the actual articulatory trajectories and the positions of the estimated boundaries. We used this kind of plot to check visually the accuracy of the estimated boundaries, since for both training and test data we had available the actual articulatory trajectories. The actual and estimated articulatory trajectories for the utterance /ah b ah/ are presented in Figure 5.43. An enlarged view of the vocal-tract profiles from the bottom of this figure is presented in Figure 5.44. The actual and reconstructed MFCC parameters, using the codebooks, are displayed in

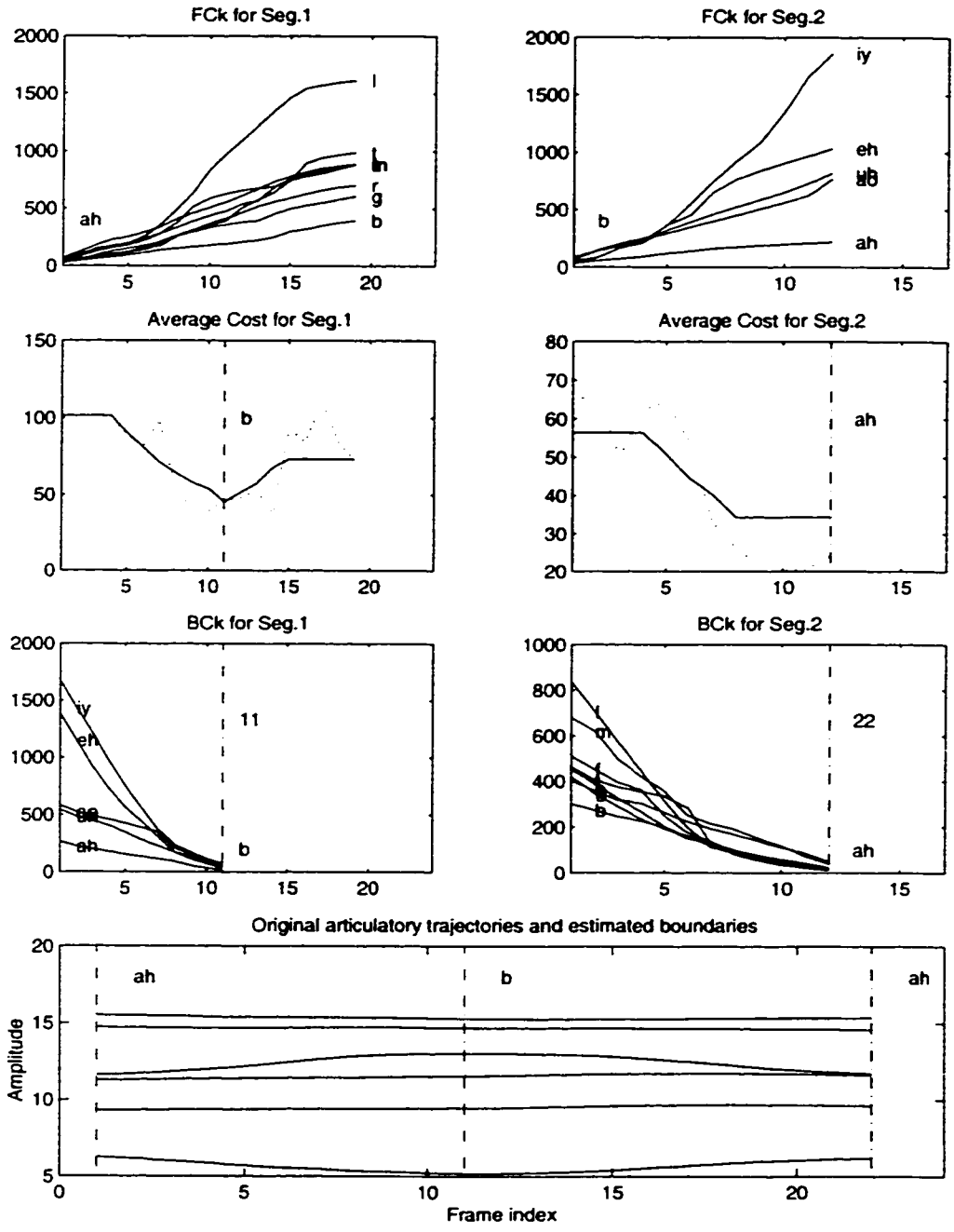

Figure 5.42: Automatic segmentation and recognition of models for an utterance /ah b ah/, included in the training data

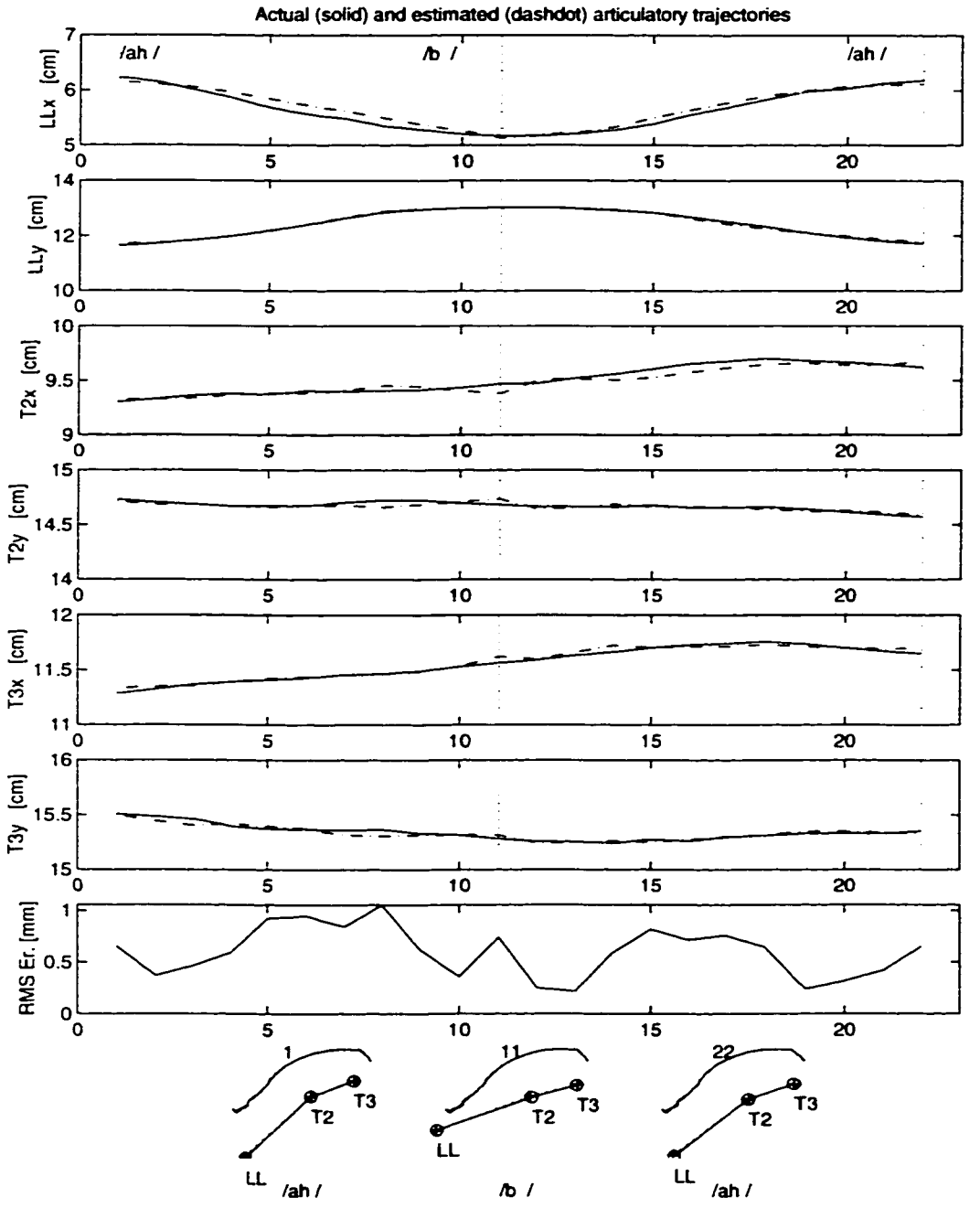

Figure 5.43: Actual and estimated articulatory trajectories for an utterance /ah b ah/. included in the training data

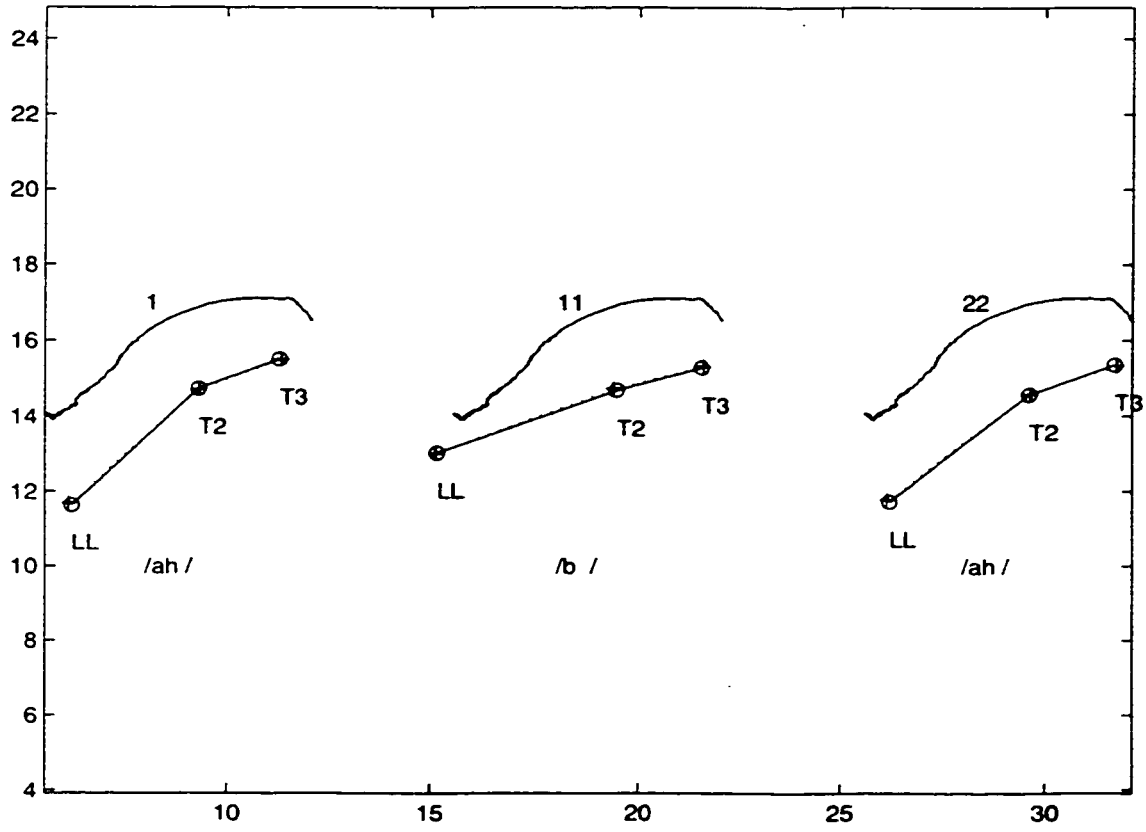

Figure 5.44: Actual and estimated VT profiles for an utterance /ah b ah/, included in training data (detail from previous figure)

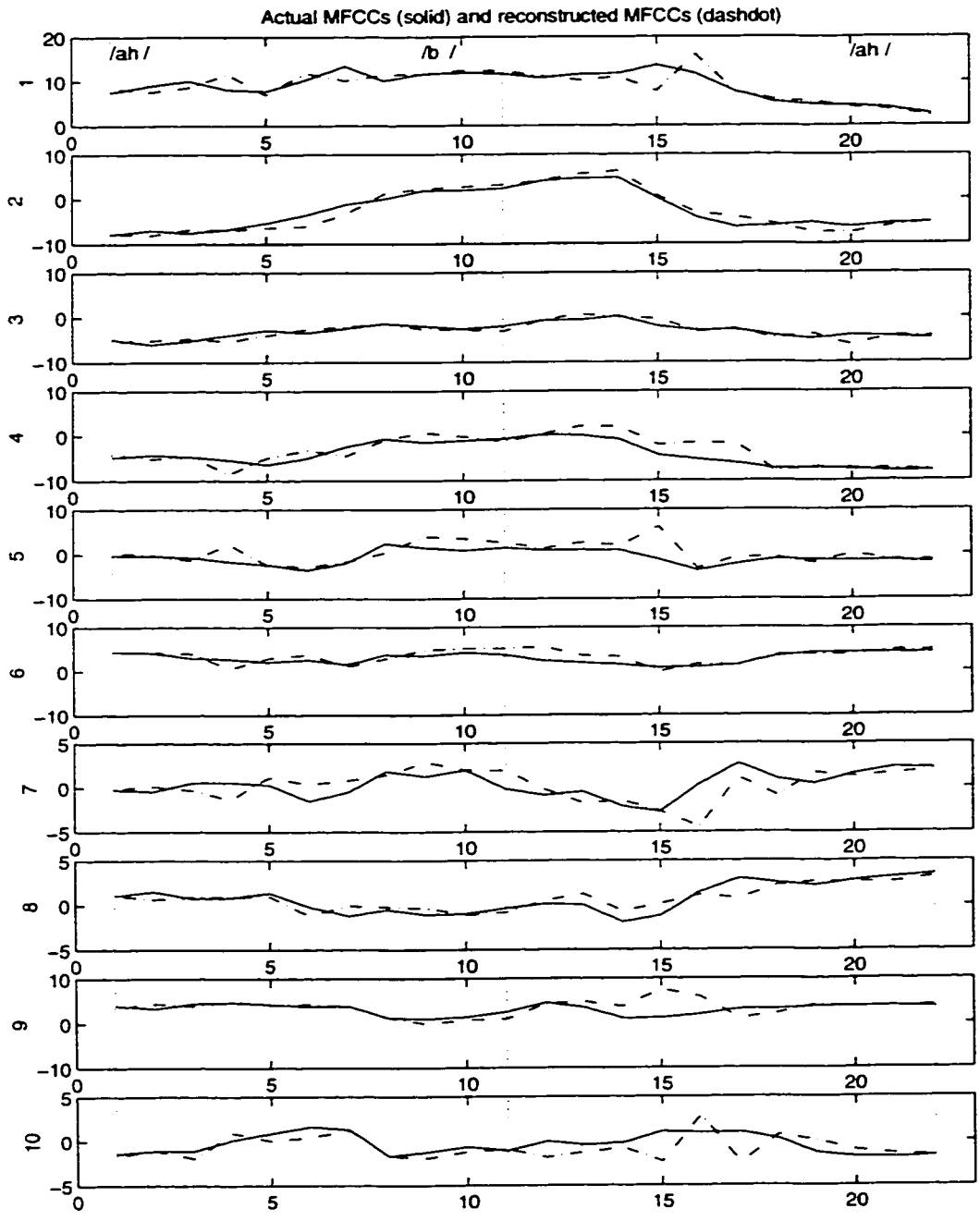

Figure 5.45: Actual and reconstructed MFCC trajectories for an utterance /ah b ah/, included in the training data

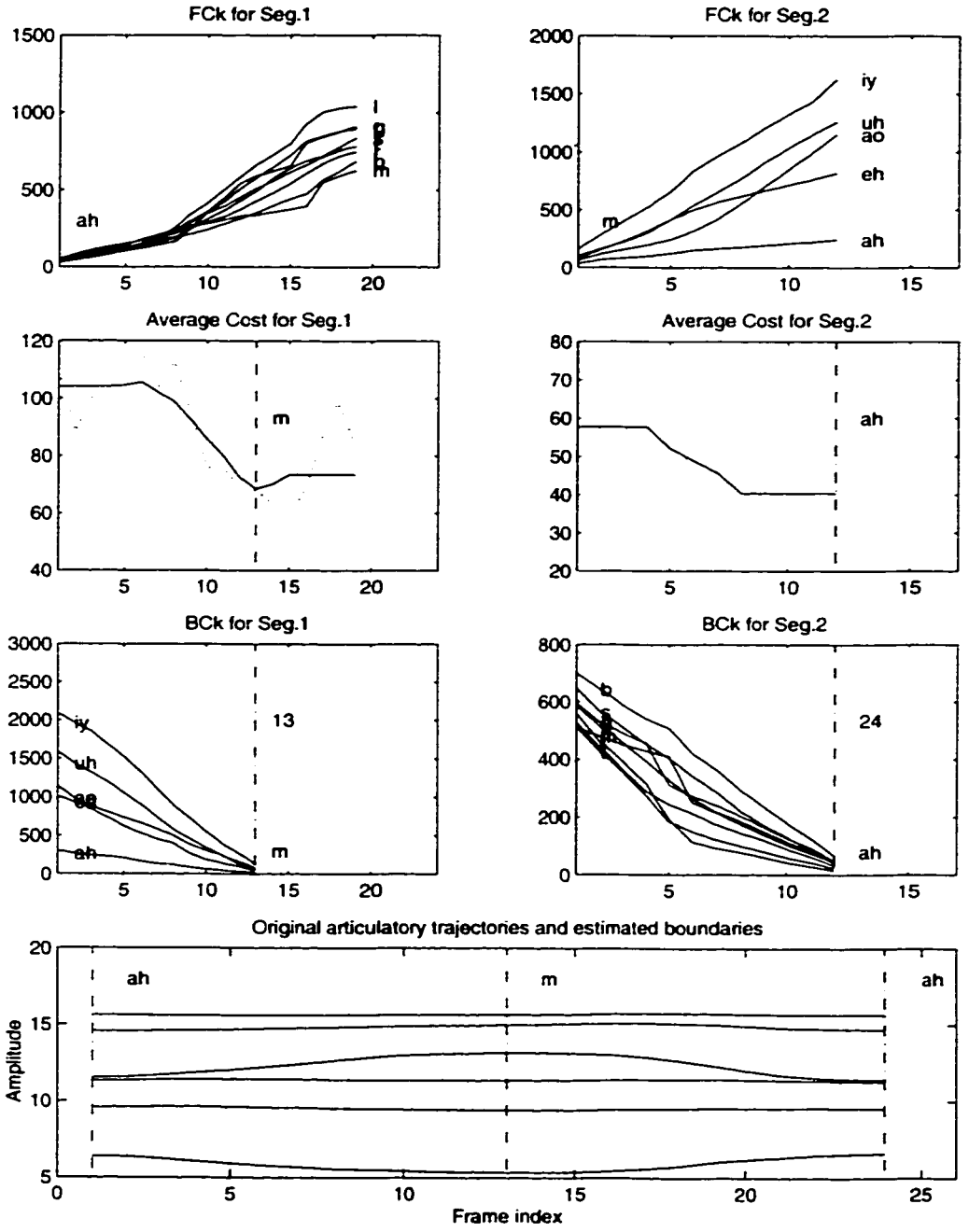

Figure 5.46: Automatic segmentation and recognition of models for an utterance /ah m ah/, not included in training data

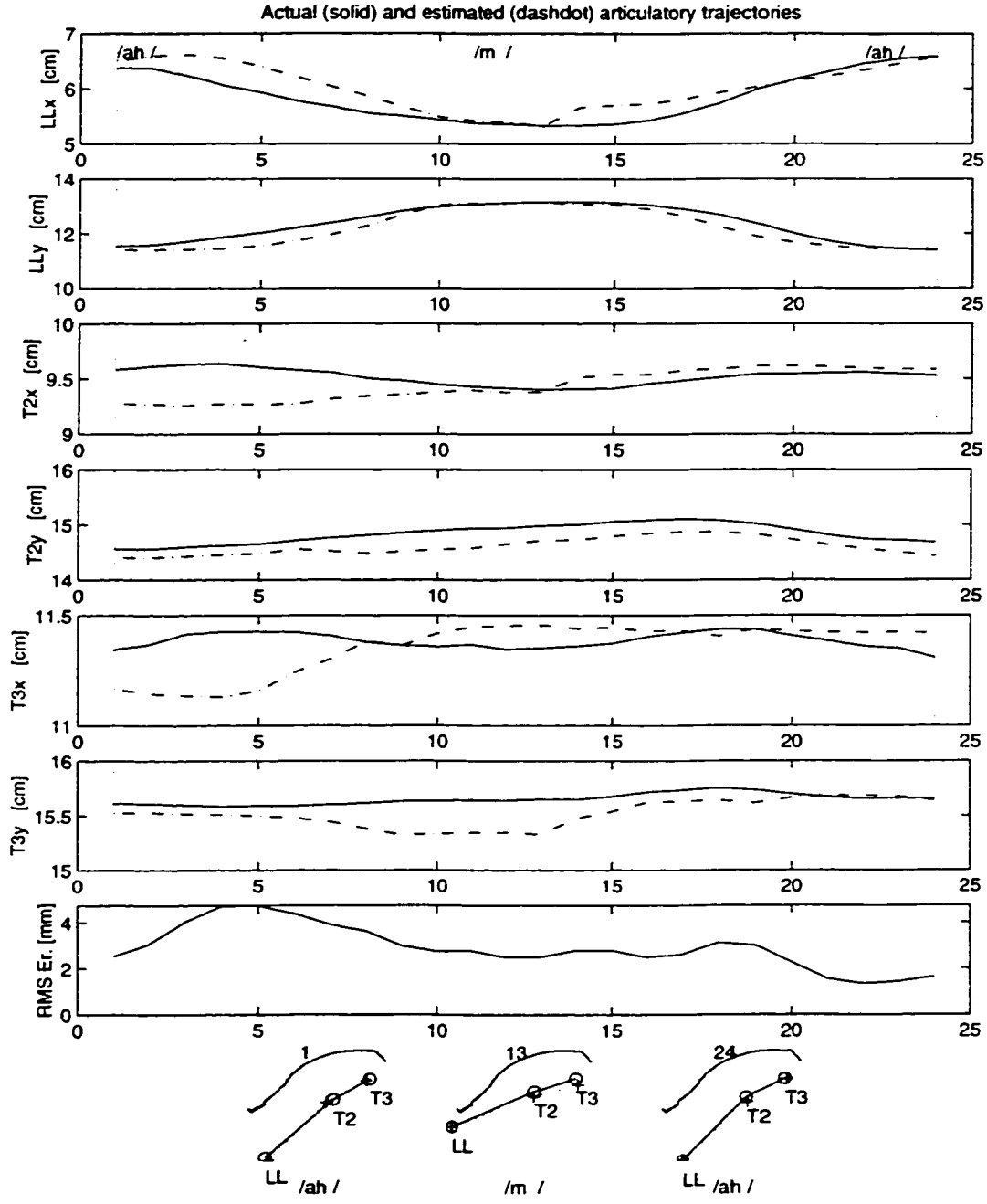

Figure 5.47: Actual and estimated articulatory trajectories for an utterance /ah m ah/, not included in training data

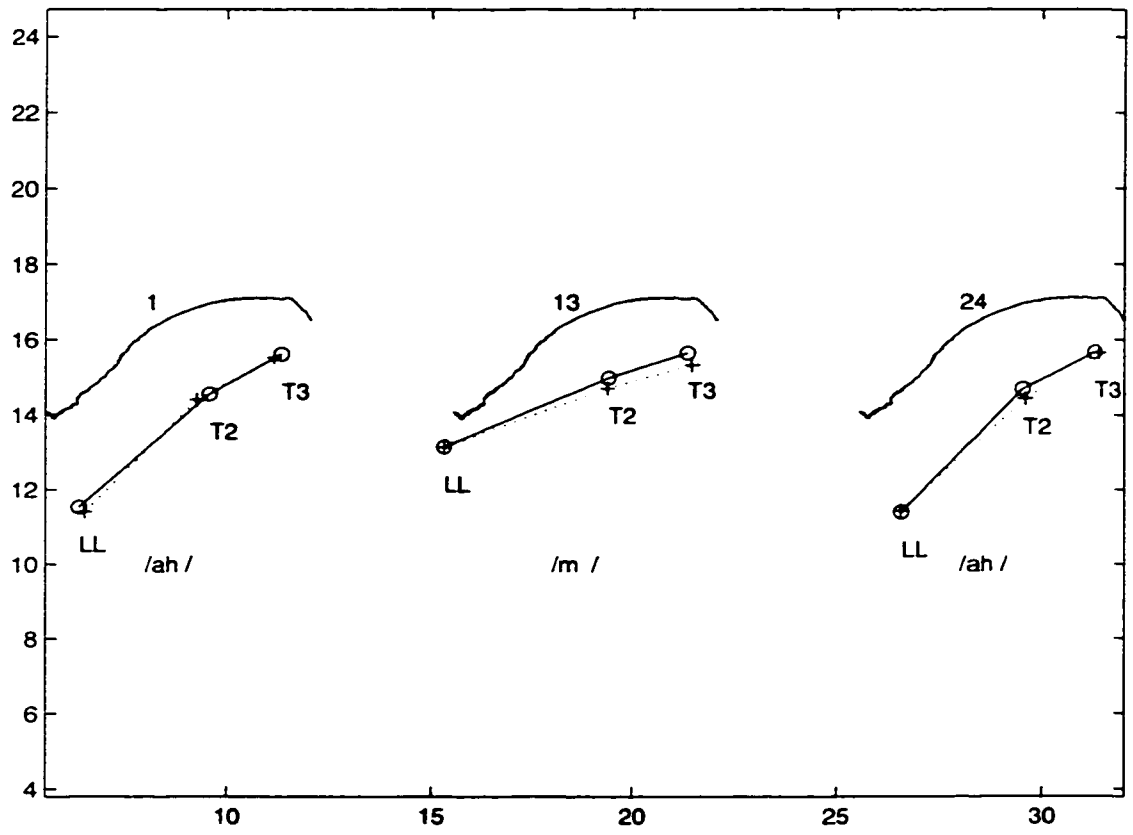

Figure 5.48: Actual and estimated VT profiles for an utterance /ah m ah/. not included in training data (detail from previous figure)

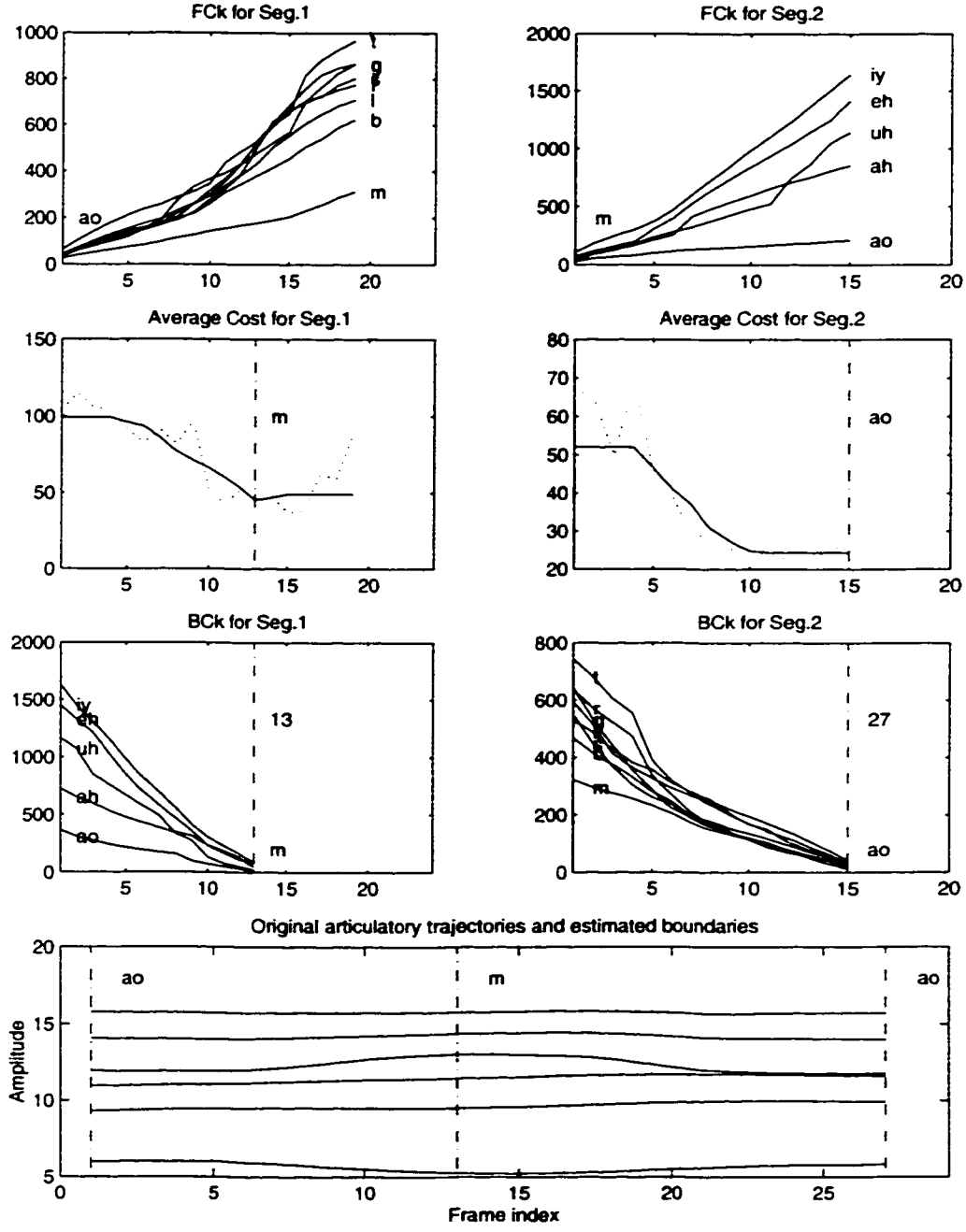

Figure 5.49: Automatic segmentation and recognition of models for an utterance /ao m ao/, included in training data

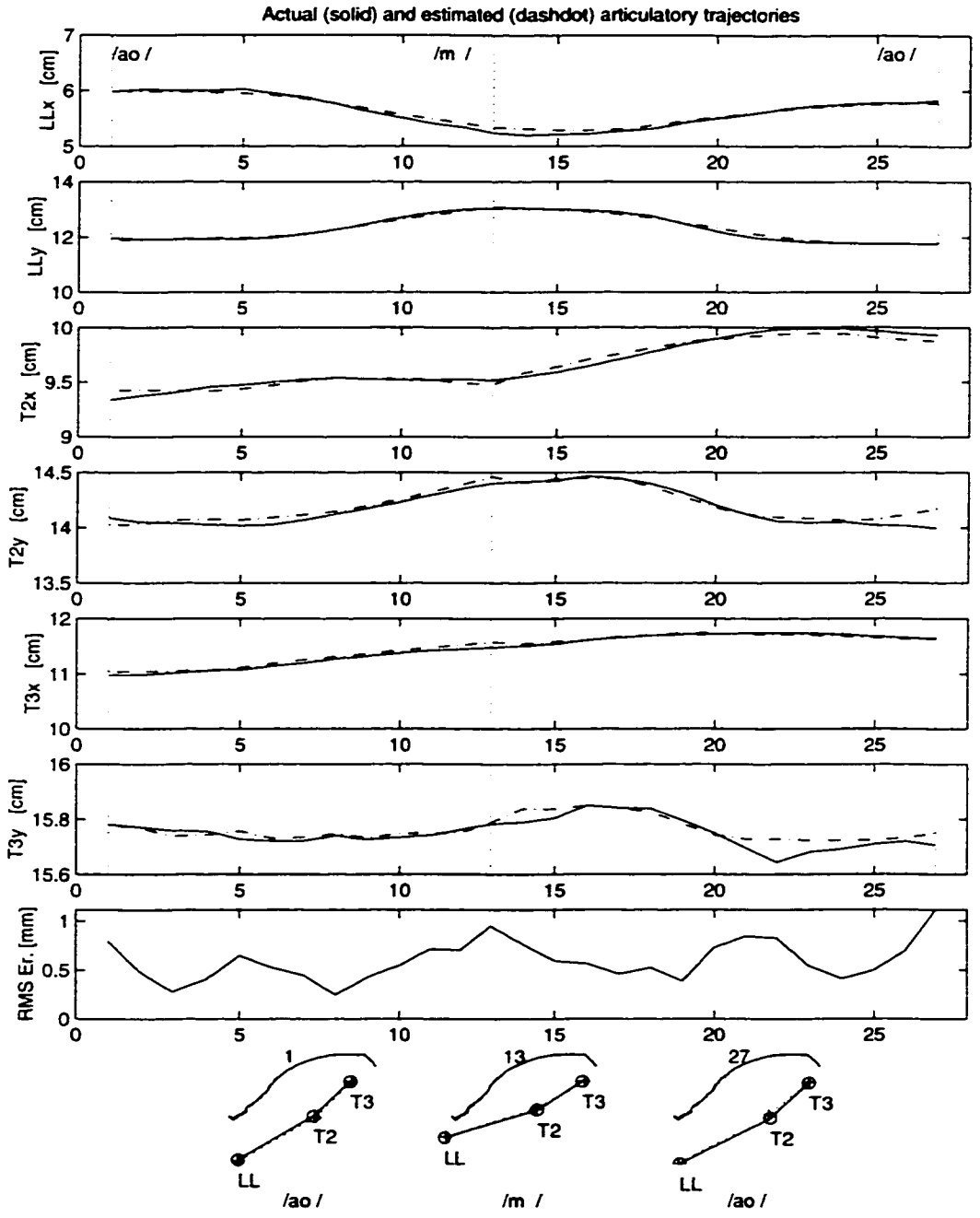

Figure 5.50: Actual and estimated articulatory trajectories for an utterance /ao m ao/. included in training data

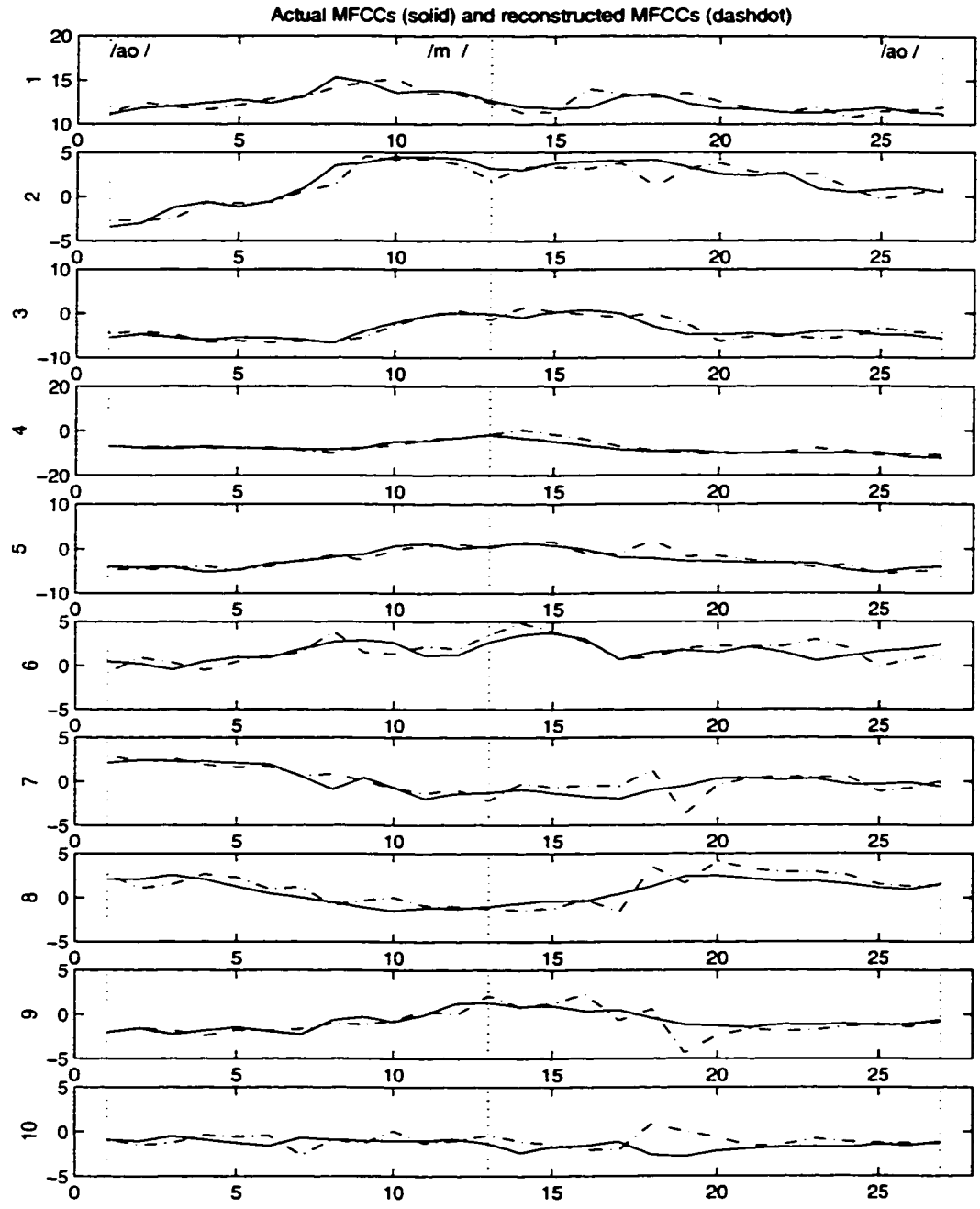

Figure 5.51: Actual and reconstructed MFCC trajectories for an utterance /ao m ao/, included in training data

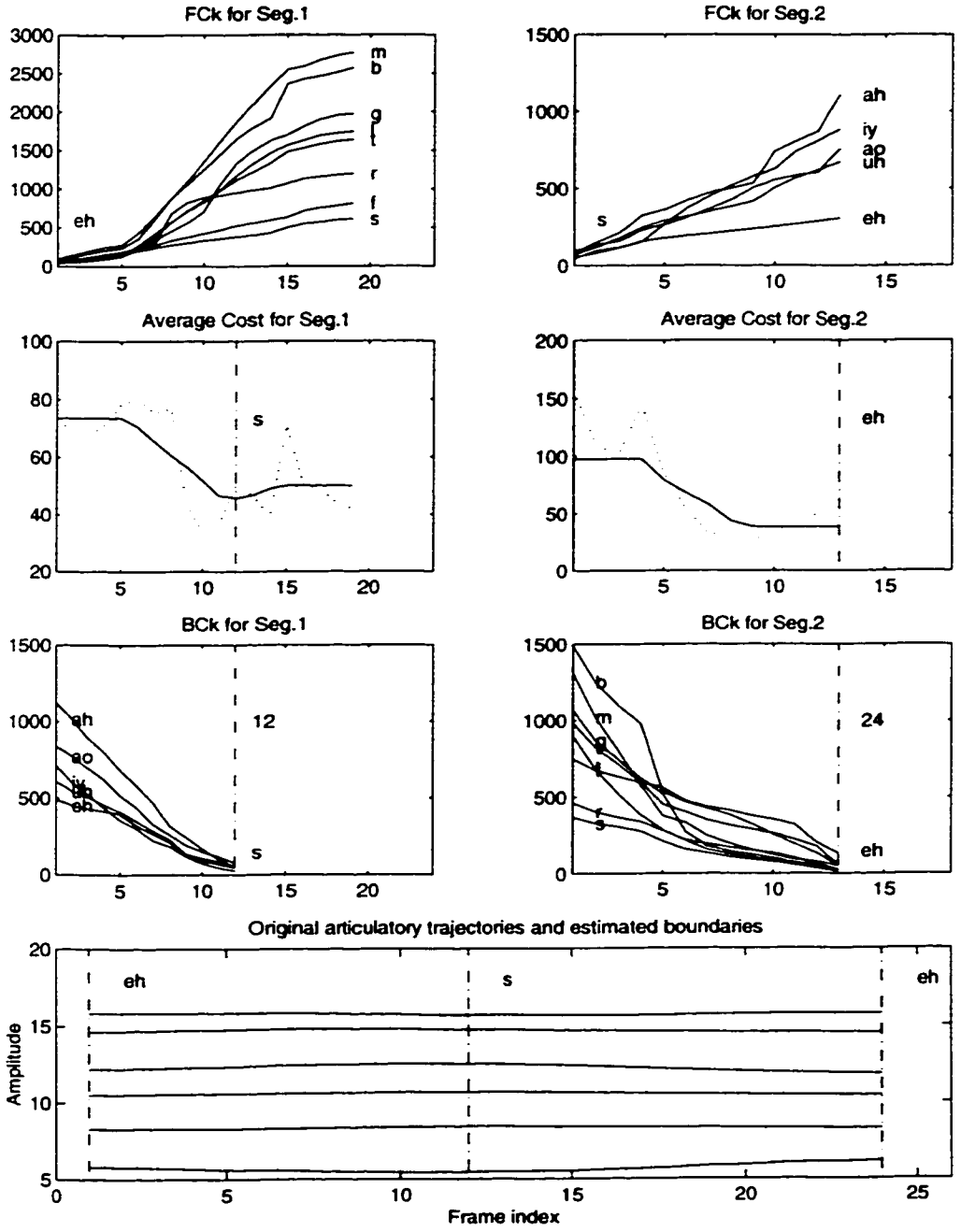

Figure 5.52: Automatic segmentation and recognition of models for an utterance /eh s eh/, not included in training data

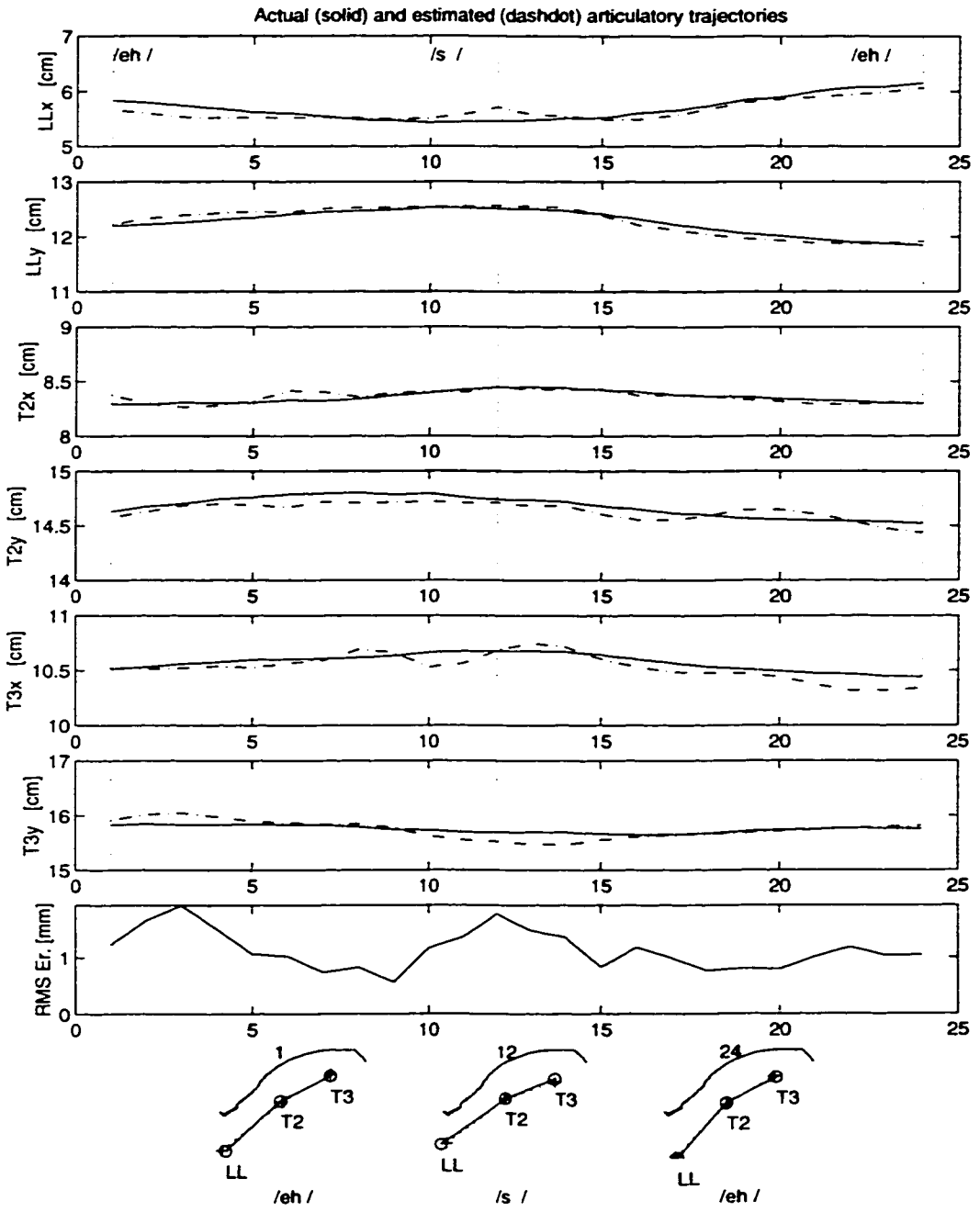

Figure 5.53: Actual and estimated articulatory trajectories for an utterance /eh s eh/, not included in training data

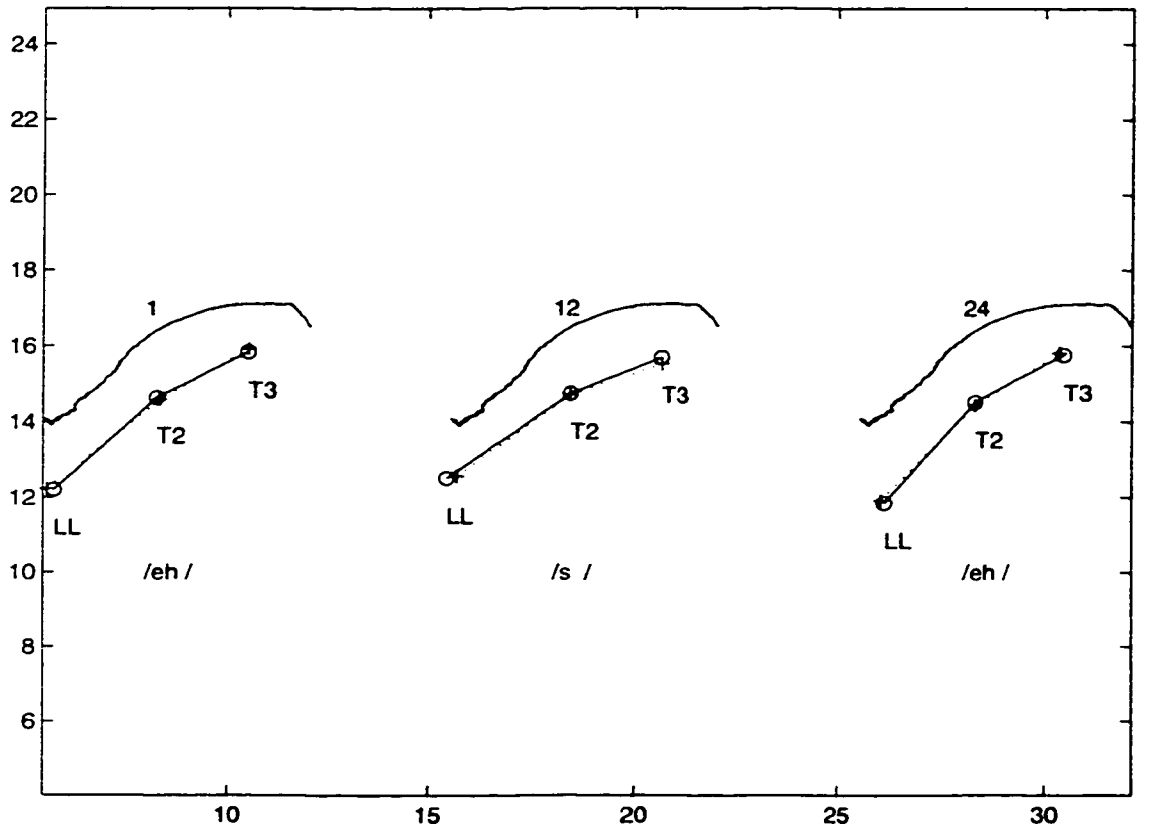

Figure 5.54: Actual and estimated VT profiles for an utterance /eh s eh/, not included in training data (detail from previous figure)

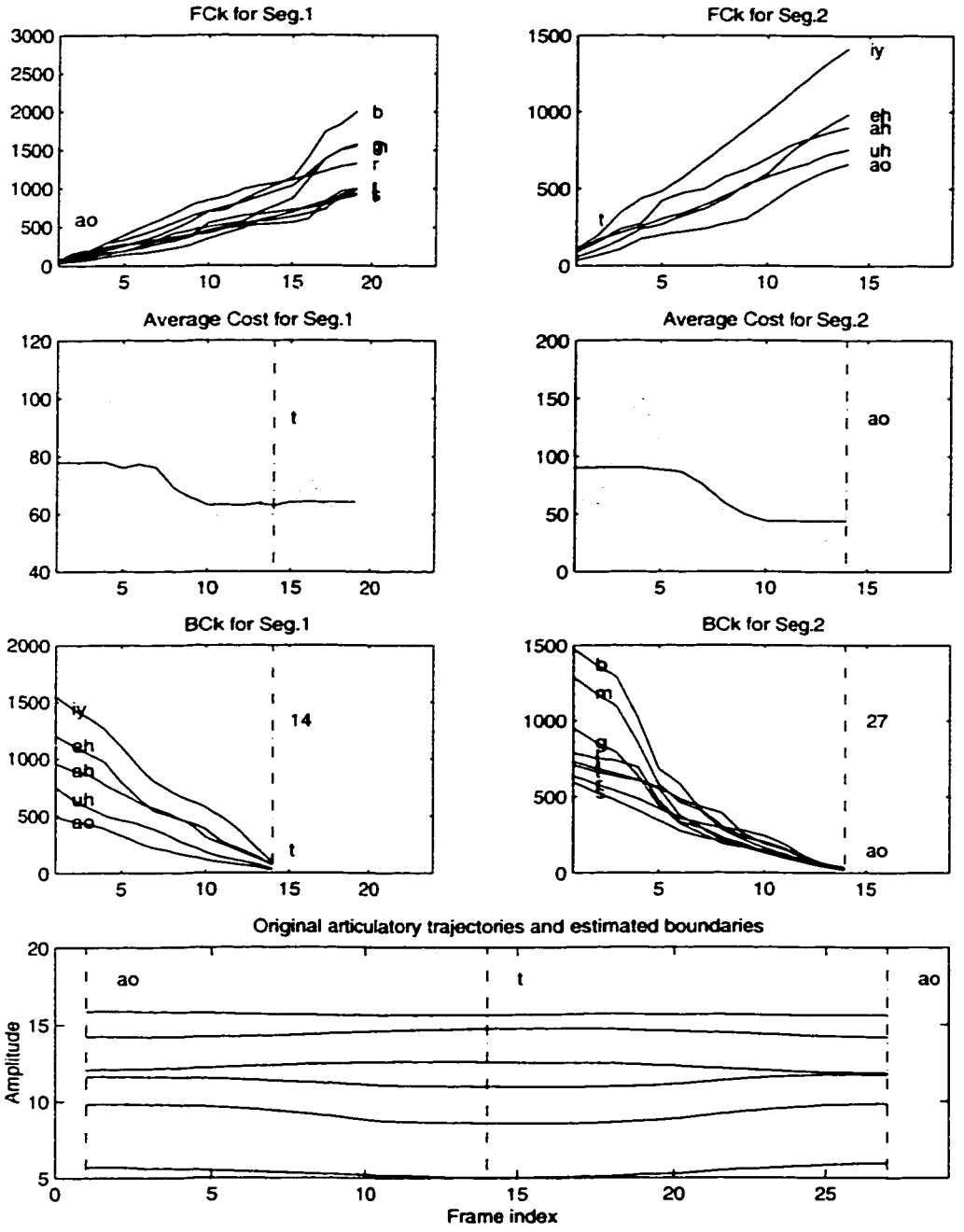

Figure 5.55: Automatic segmentation and recognition of models for an utterance /ao t ao/, not included in training data

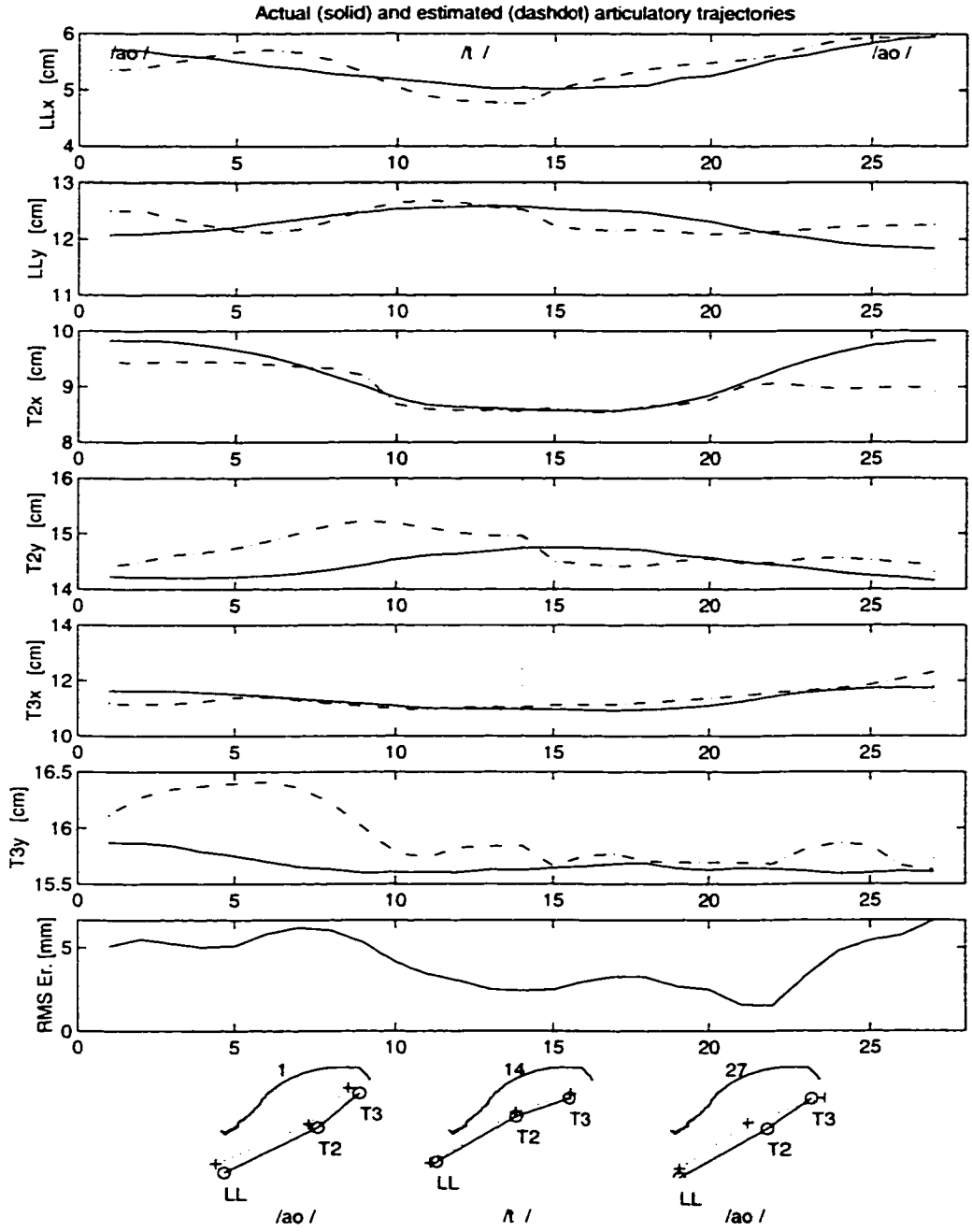

Figure 5.56: Actual and estimated articulatory trajectories for an utterance /ao t ao/. not included in training data

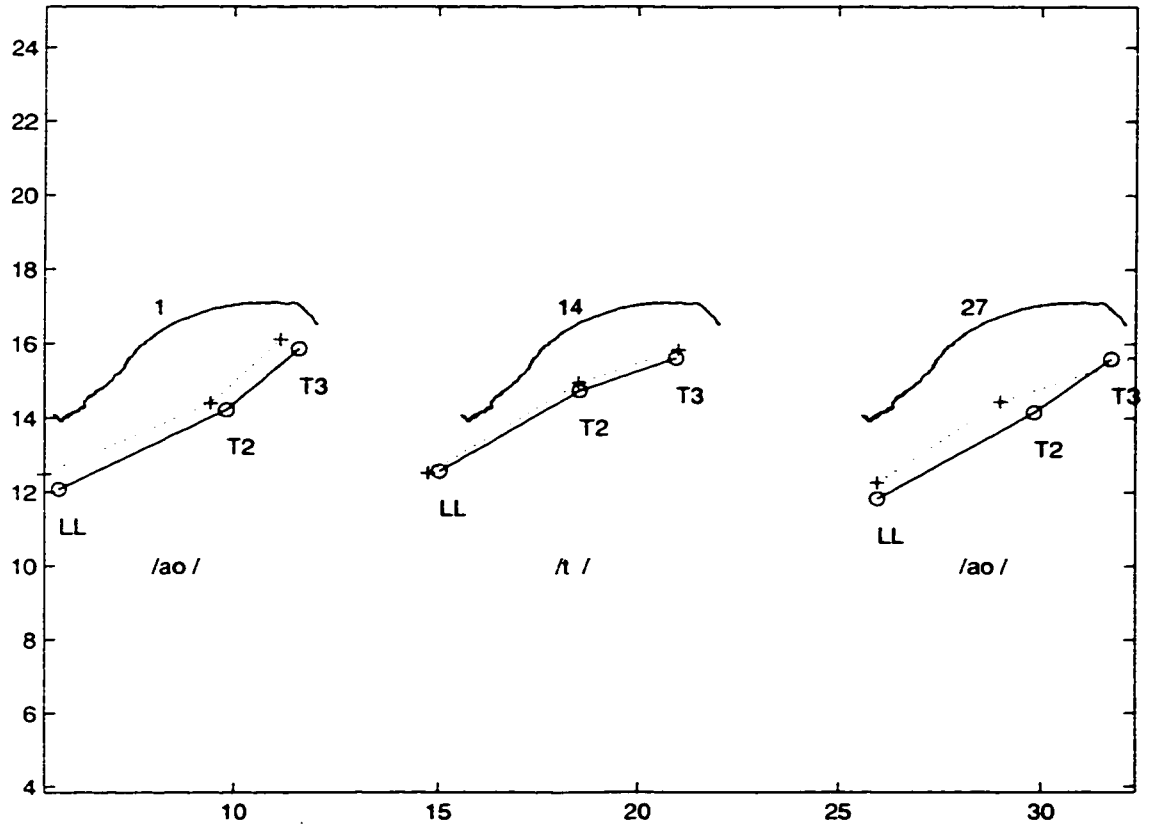

Figure 5.57: Actual and estimated VT profiles for an utterance /ao t ao/. not included in training data (detail from previous figure)

Figure 5.45.

An example of estimation from an utterance /ah m ah/, not included in training data is presented in Figures 5.46 to 5.48. The mean RMS error was about 3.0 mm for this case. The central positions of the second phoneme of each model were estimated very well.

An example of estimation from an utterance /ao m ao/ included in training data is presented in Figures 5.49 to 5.51. The RMS errors for this case were very small, because the utterance was included in training data of the models. The central positions of the second phoneme of each model were also estimated very well.

Other examples, for utterances not included in training data are presented in Figures 5.52 to 5.54 for an utterance /eh s eh/, and in Figures 5.55 to 5.57 for an utterance /ao t ao/. For both cases the central positions of the second phoneme of each model were estimated very well. However, we especially presented these examples because of the relative large differences in the estimating RMS errors of the two cases. In the first example, an average RMS error of about 1.2 mm was found. In the second example, the average RMS error was about 3.9 mm. This large error is probably due to the differences between the training and testing cases, in vocal-tract shapes and trajectories in producing the /ao t ao/ utterances.

5.3 Experimental Results based on X-ray Microbeam Speech Data

We recently received a copy of the X-ray Microbeam Speech Production Data of University of Wisconsin. In this section, we present a few experimental results obtained using the X-ray Microbeam Data of University of Wisconsin, [104]. This database consists of articulatory and acoustic recordings from 48 speakers using American English. The articulatory data was sampled every 6.866 ms (sampled at approx. 146 Hz). The speech acoustic signal was sampled at 21.739 Hz. We selected the first speaker from this database, coded JW11. From this speaker, we selected two utterances — /s ah s ah ... s ah/ and /p uh p uh ... p uh/, coded Tp105 and Tp102, respectively. The articulatory parameters we used consisted of the X and Y coordinates of four pellets placed on the lower lip (LL) and on tongue (TT1, TT2, TT3). Details regarding these recordings can be found in [104]. From the speech signal we computed the first 10 MFCC parameters, excluding the energy parameter, every 10 ms, from Hamming windows of 32 ms. We re-sampled the articulatory trajectories with a frequency of 100 Hz, corresponding to 10 ms frame intervals.

Figure 5.58 presents the data of the /s ah s ah ... s ah/ utterance of the speaker JW11. We constructed a model for the /s ah/ unit using 8 of the /s ah/ segments from this utterance. The training data for this model are presented in Figure 5.59. Figures 5.60 and 5.61 presents the estimation results for a /s ah/ segment not included in the training data.

The articulatory and acoustic data for the utterance /p uh p uh ... p uh/ are presented in Figure 5.62. The estimation results for a segment /p uh/ not included in training data are presented in Figures 5.63 to 5.65.

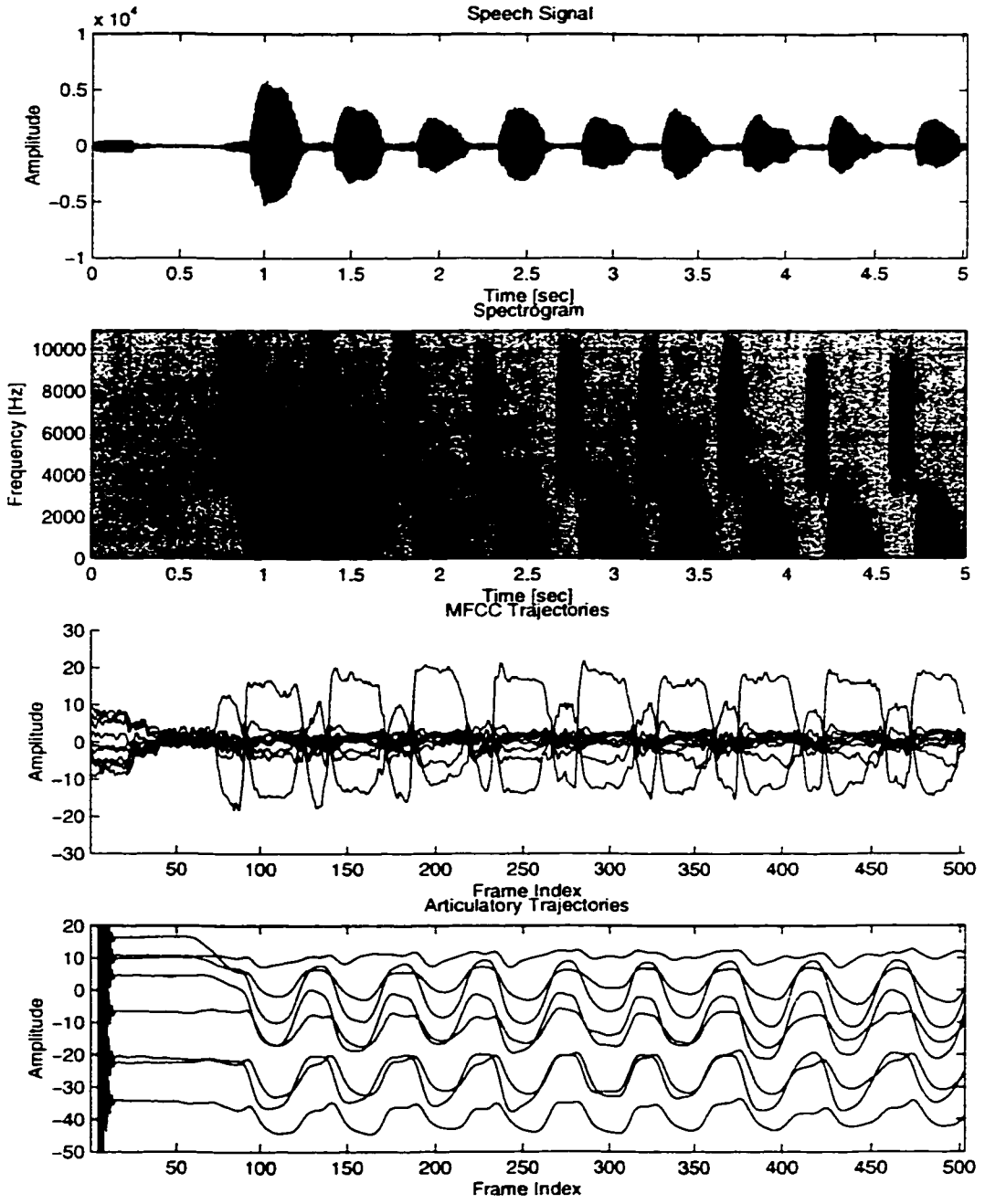

Figure 5.58: MFCC and articulatory trajectories for the /s ah s ah ... s ah/ utterance coded 'Tp105' of speaker JW11 from Wisconsin X-ray Microbeam Database

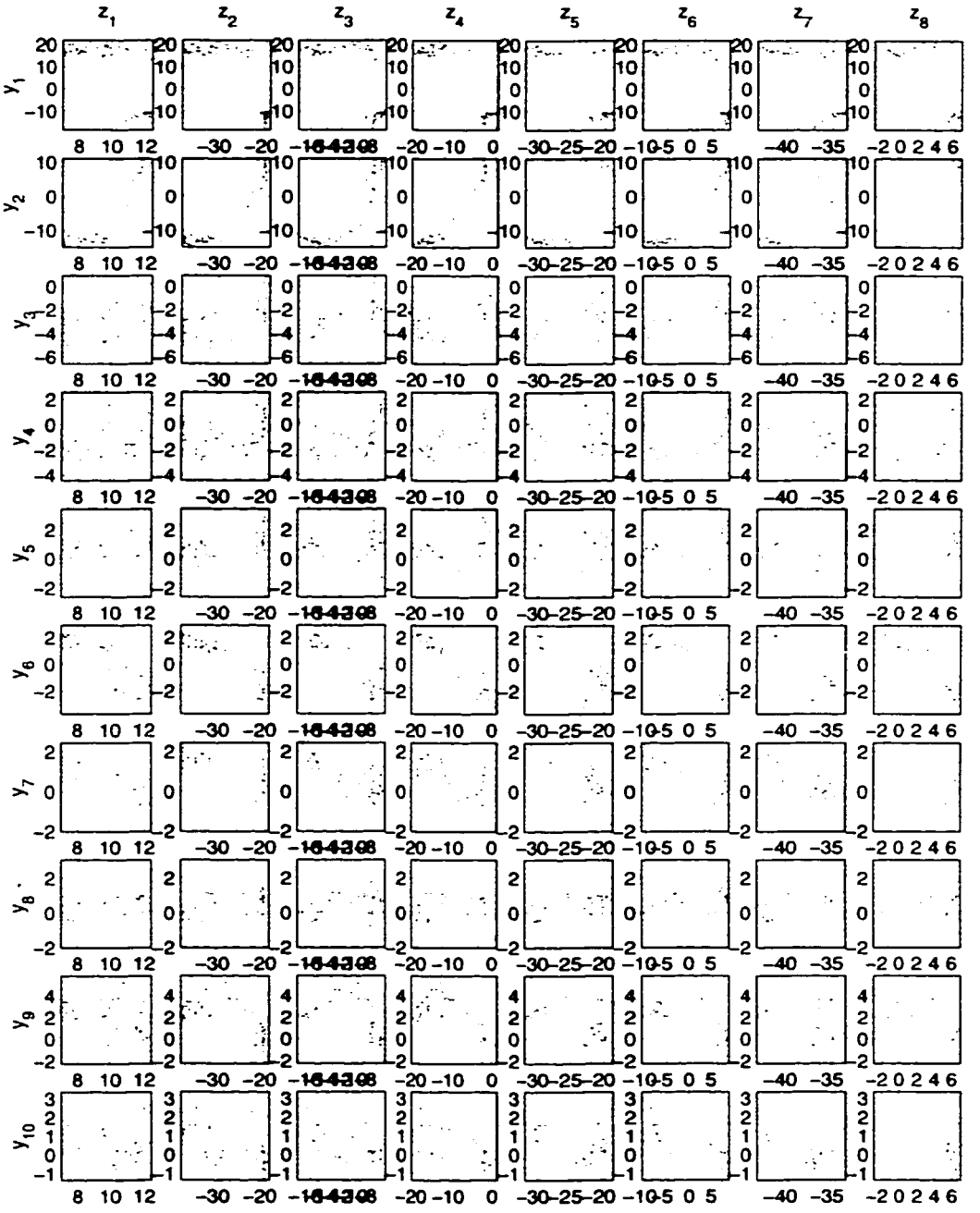

Figure 5.59: Articulatory and acoustic training data for /s ah/ consisting of 8 segments

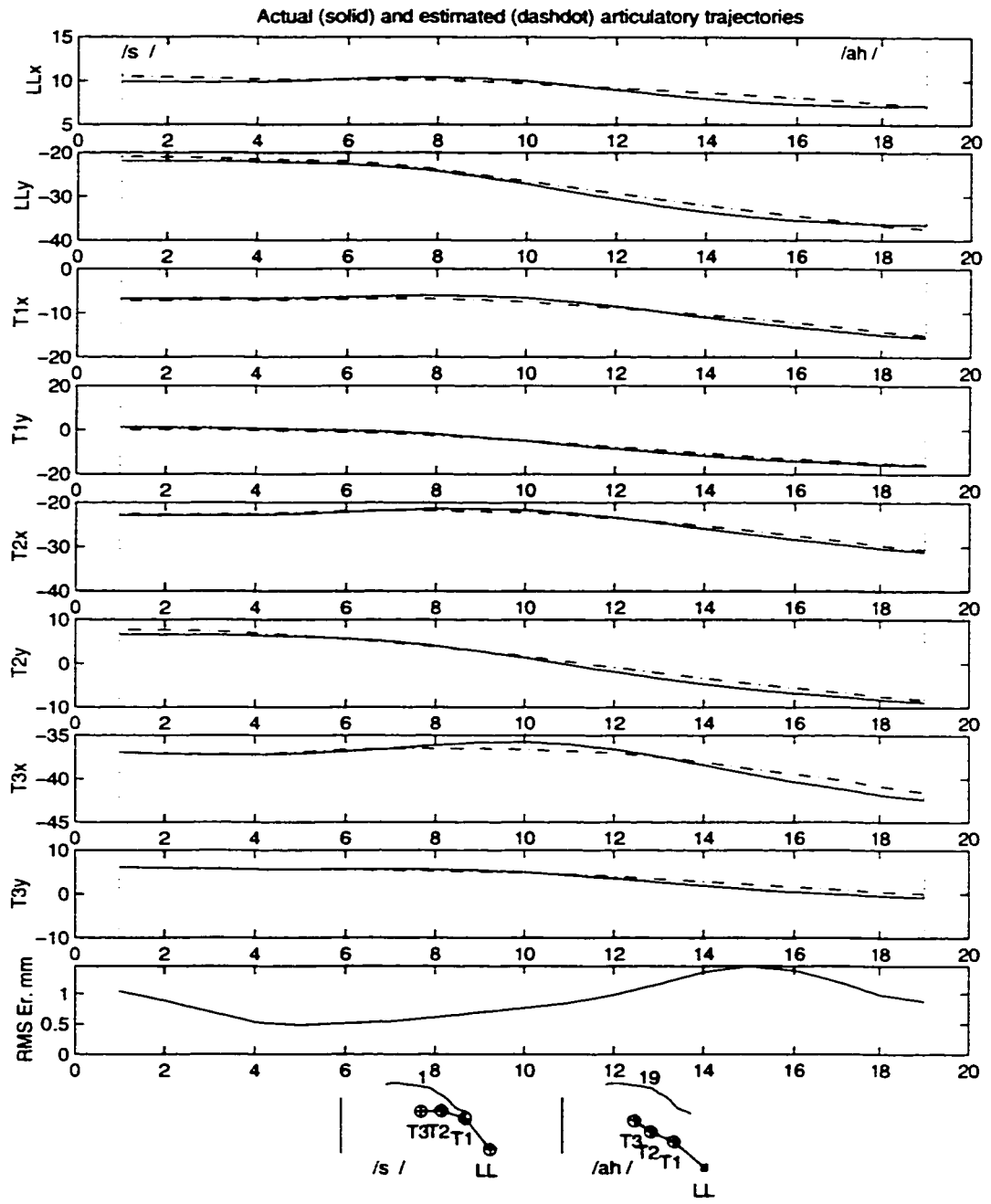

Figure 5.60: Actual and estimated articulatory trajectories for a segment /s ah/. not included in training data

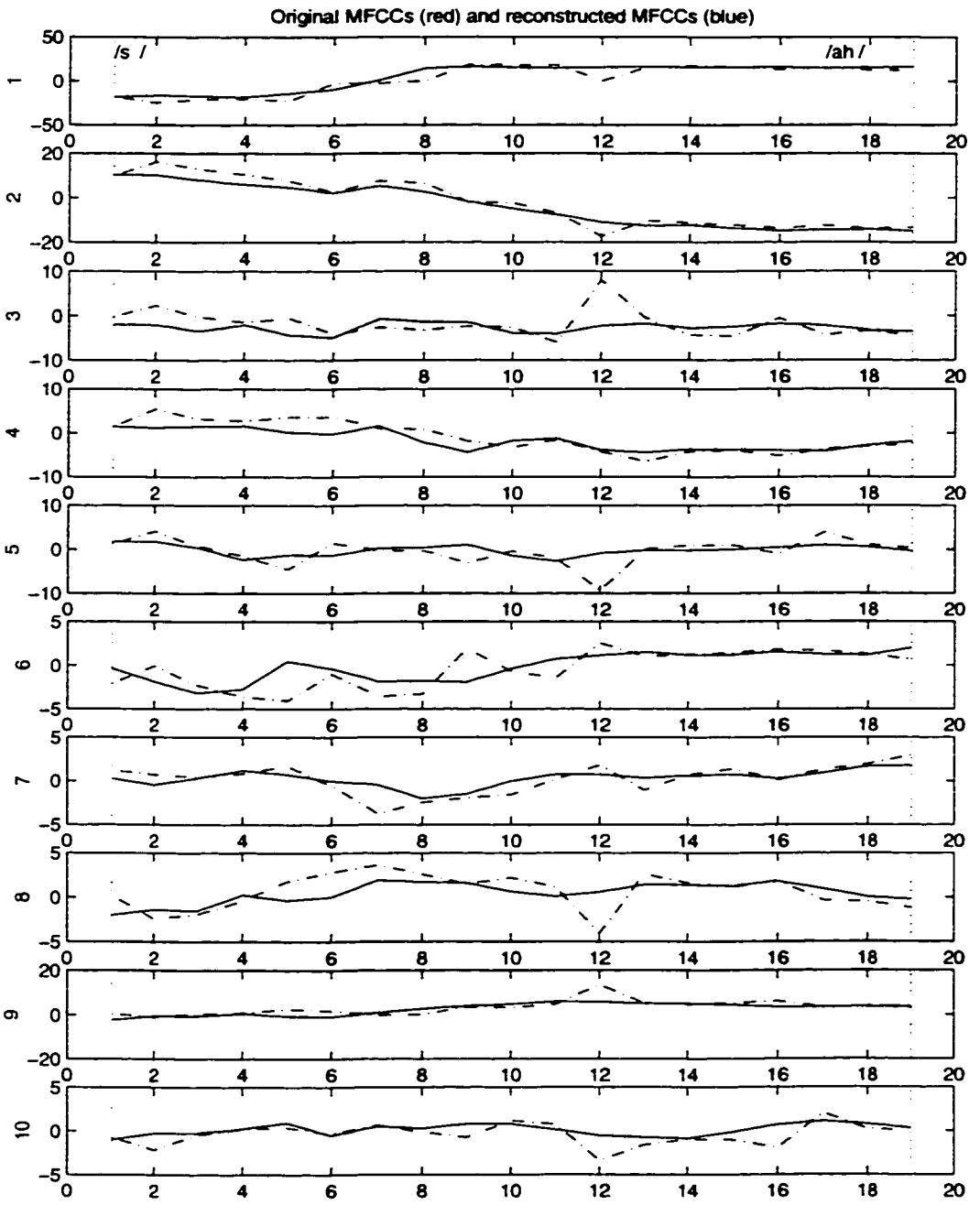

Figure 5.61: Actual and reconstructed MFCC trajectories for a segment /s ah/, not included in training data

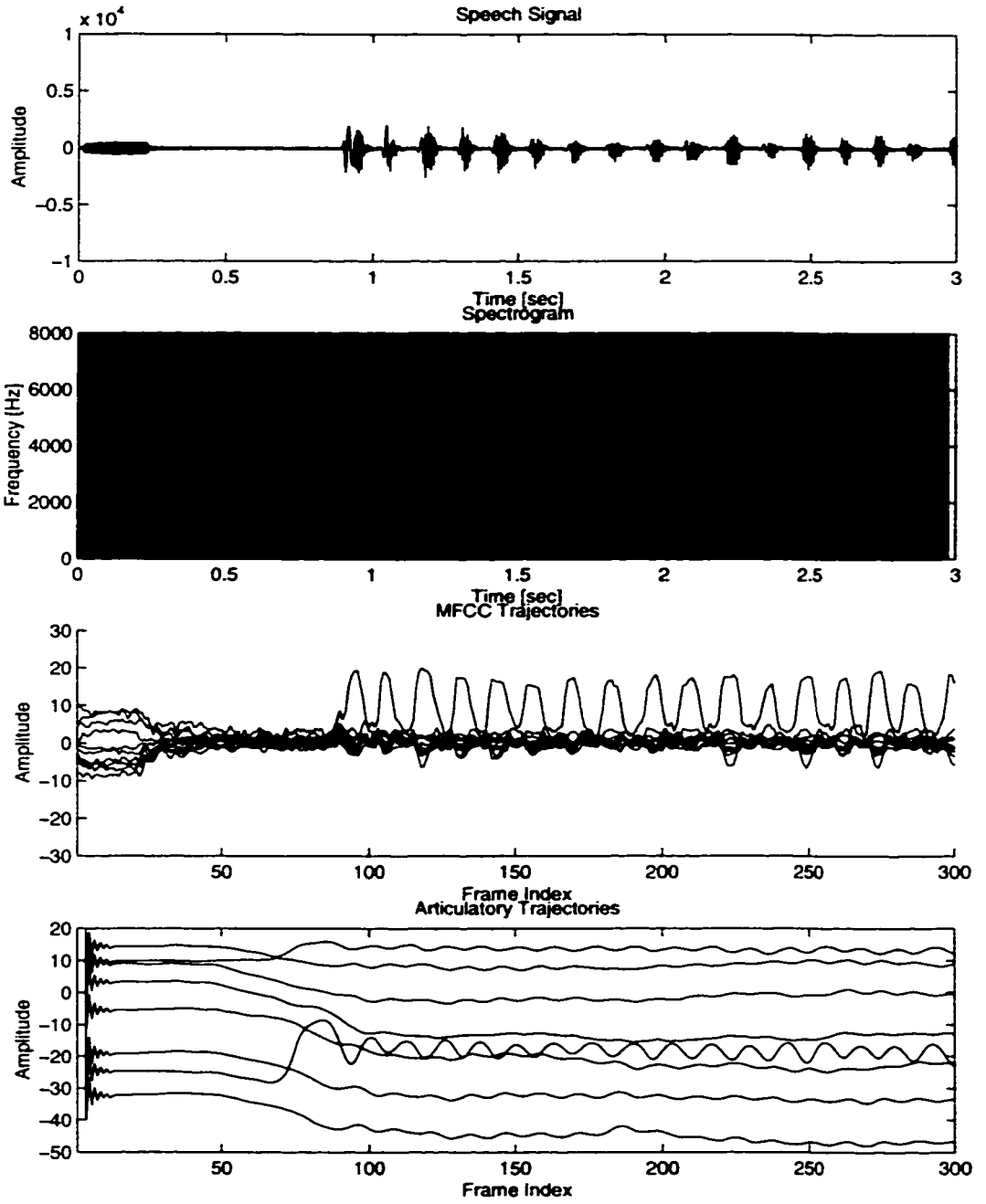

Figure 5.62: MFCC and articulatory trajectories for the /p uh p uh ... p uh/ utterance coded 'Tp102' of speaker JW11 from Wisconsin X-ray Microbeam Database

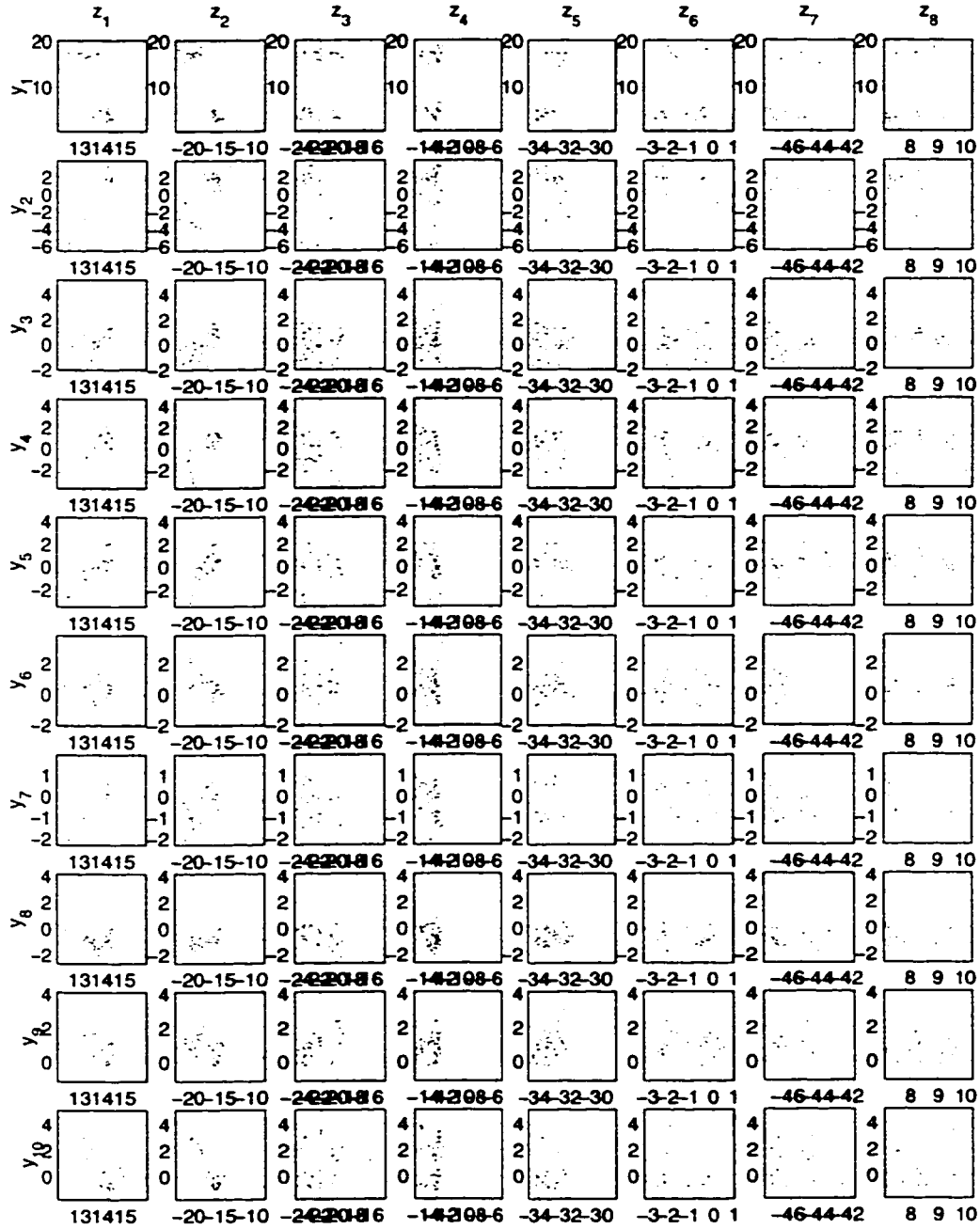

Figure 5.63: Articulatory and acoustic training data for /p uh/ consisting of 16 segments

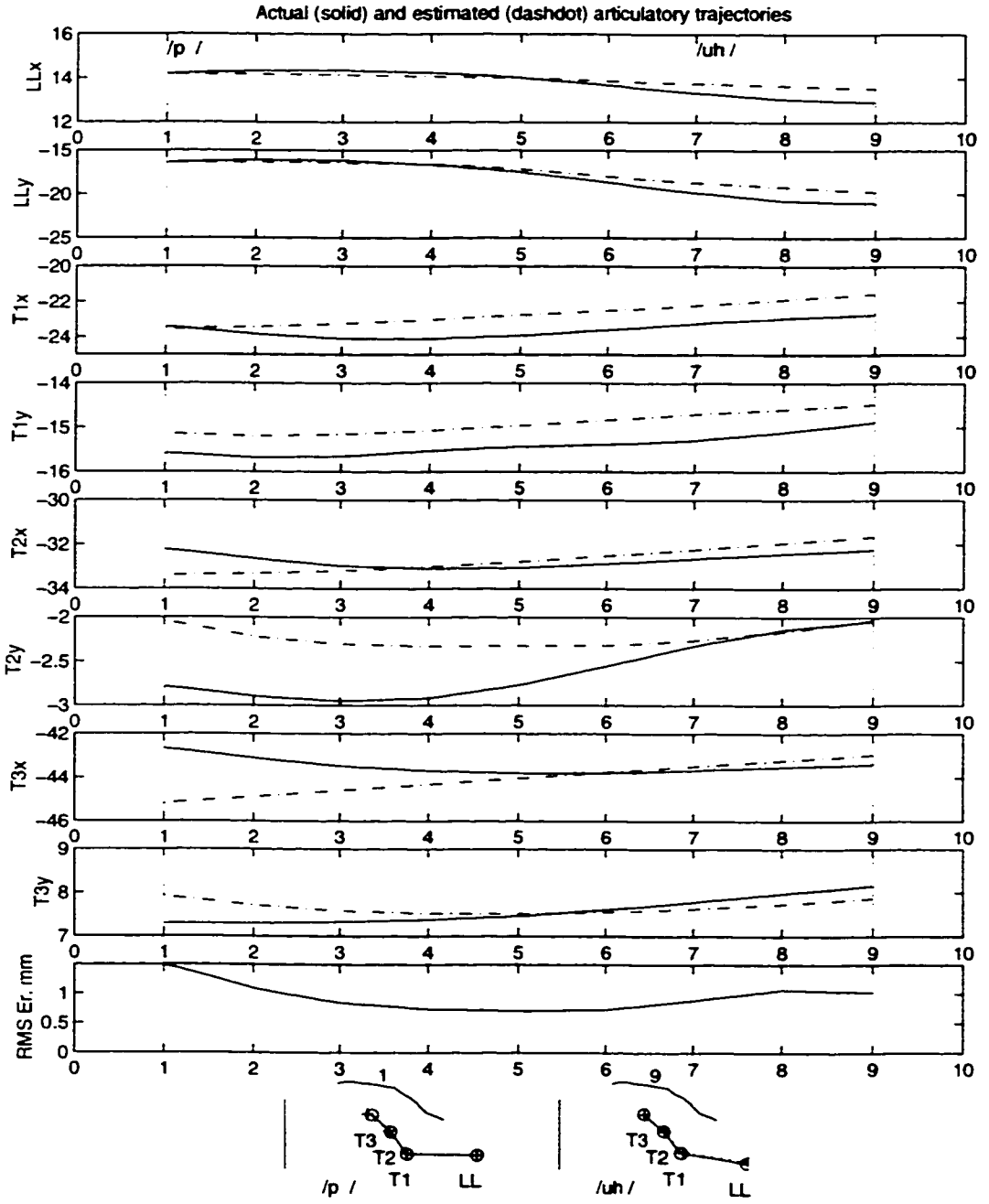

Figure 5.64: Actual and estimated articulatory trajectories for a segment /p uh/, not included in training data

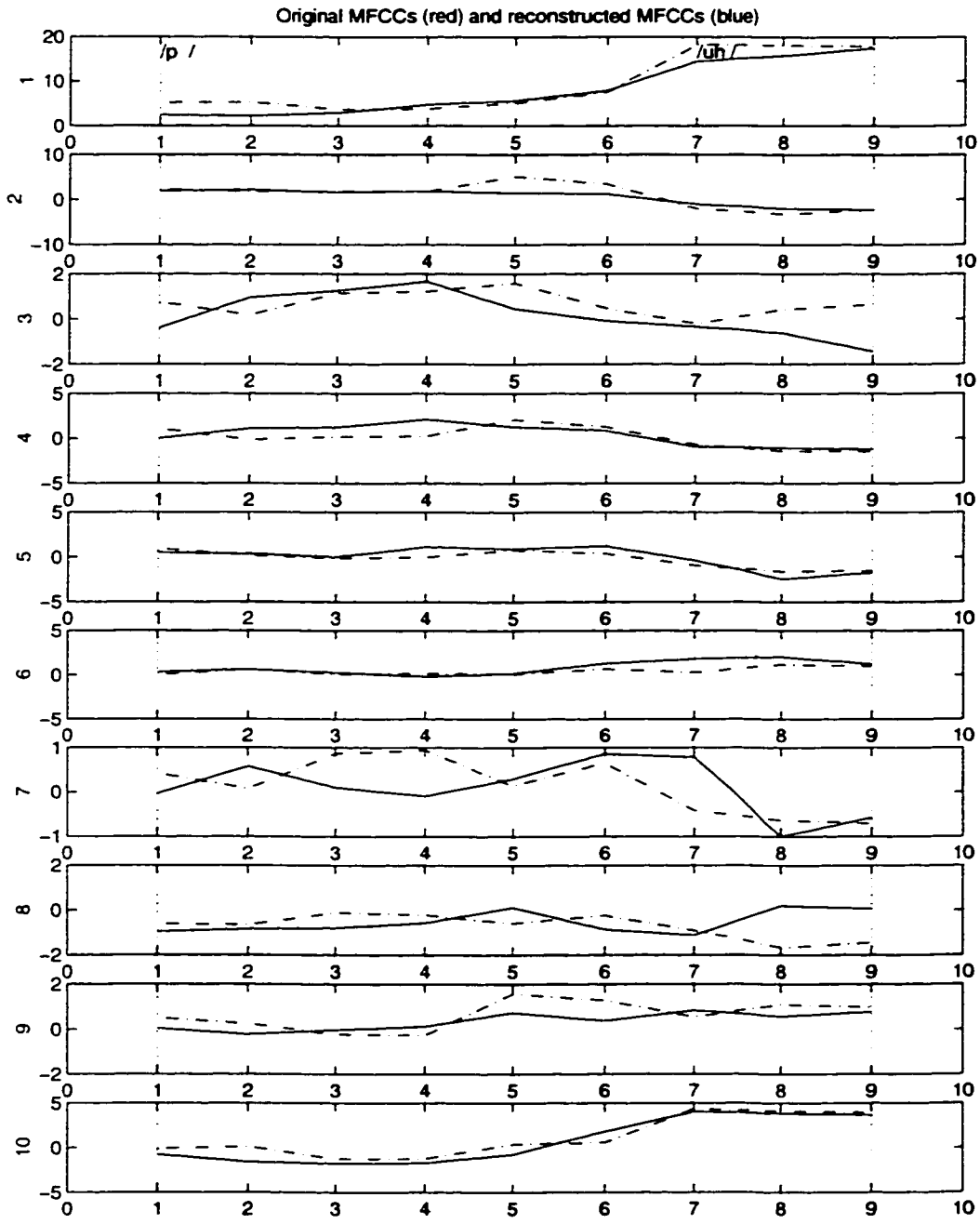

Figure 5.65: Actual and reconstructed MFCC trajectories for a segment /p uh/, not included in training data

Chapter 6

Applications

In this chapter, two potential applications of this speech inversion method to two different areas are presented. A new graphic representation method for displaying vocal-tract area function evolution over time is proposed as a general application in speech research and as an aid in teaching the hearing or speaking impaired to speak or in teaching foreign languages. A potential application of this speech inversion method to automatic speech recognition is also presented.

6.1 Displaying the Dynamics of the Vocal-Tract

In articulatory speech research the graphic representations of articulatory trajectories and vocal-tract shapes are very useful. The most common articulatory representation in speech is in a form of a sagittal contour of the vocal tract. This kind of graphic representation can be obtained from cine-radiographic images, X-rays, Magnetic Resonance Imaging systems (MRI) or electromagnetic articulography and represents the shape of the vocal tract at certain moments in time. Using this rep-

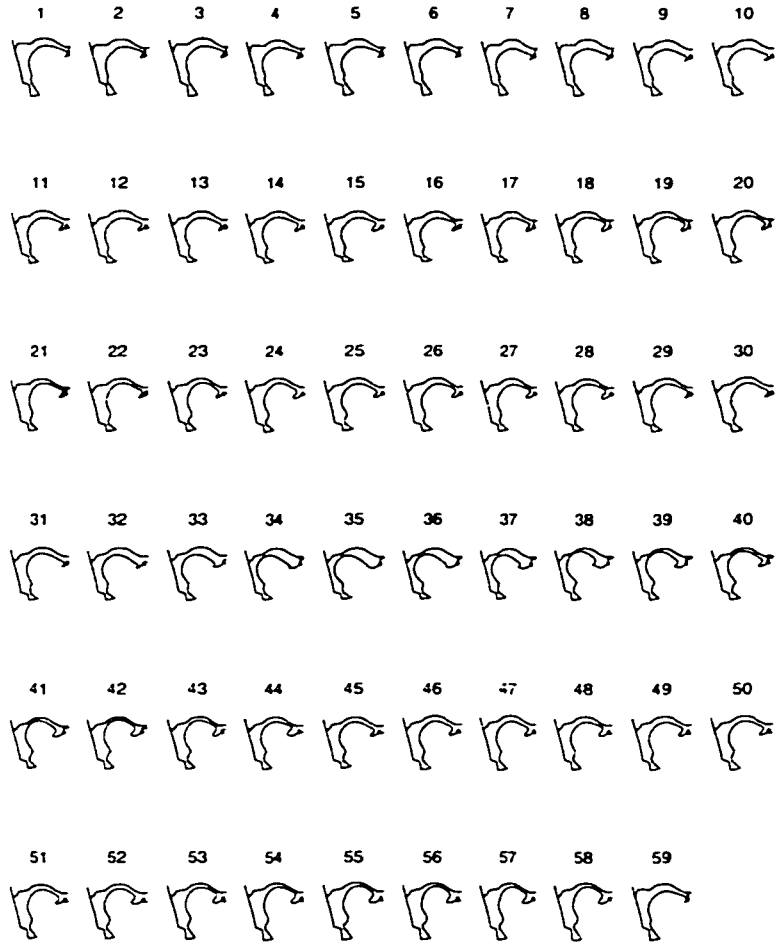

Figure 6.1: Vocal-tract shapes for the French sentence 'Ma chemise est roussie'

resentation the evolution in time of the vocal-tract can be portrayed by displaying consecutive shapes at different time points, as presented in Fig. 6.1 for a French sentence 'Ma chemise est roussie' produced by a female speaker. The shapes are extracted from X-ray images recorded at a rate of 50 frames/sec. The position of the main articulators can be observed in each vocal-tract section, but the continuity of vocal-tract shape evolution in time cannot be easily observed on these discrete shape plots. This representation of the vocal apparatus is very common in

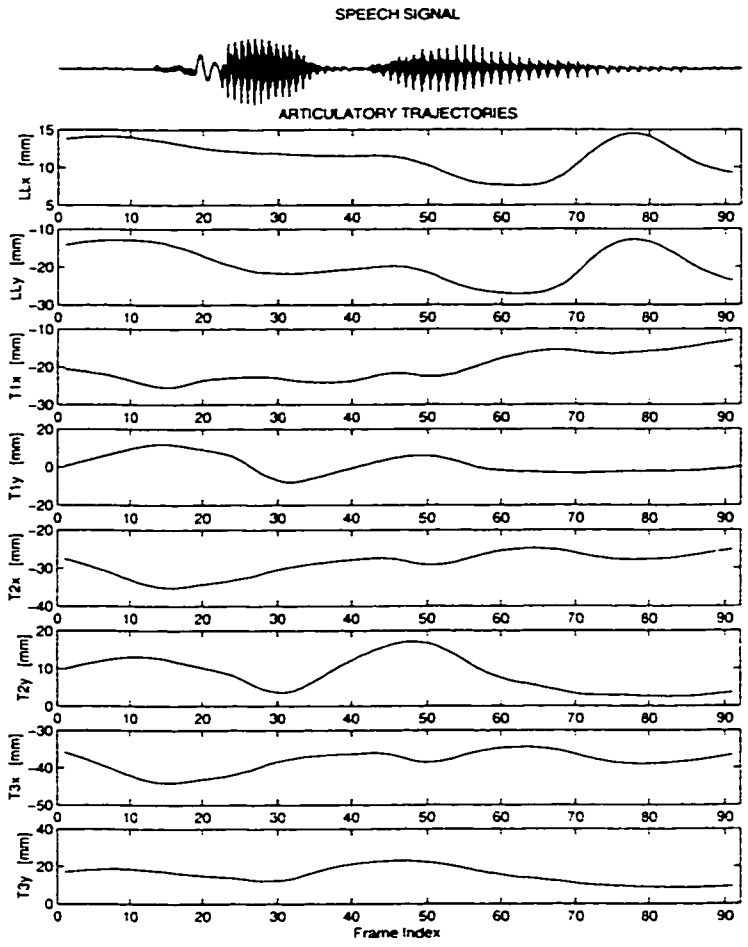

Figure 6.2: Articulatory trajectories for the word 'program'

articulatory phonetics.

A dynamic representation of the articulators in a form of articulatory trajectories is presented in Fig. 6.2. for an utterance of the word 'program.' In the first sub-plot. the speech signal is plotted. Then. in the next sub-plots. eight trajectories are displayed. representing the X and Y coordinates of 4 pellets, one placed on the lower lip and three on the tongue. These articulatory trajectories are recorded with an X-ray microbeam system. [104]. In this figure the horizontal axis represents the

time as the frame index, whereas the vertical axis represents the amplitude of the articulatory parameters in mm. When the articulators' trajectories are available from recordings like those of electromagnetic articulography. these representations can show the continuity of articulators' movements between adjacent sounds. The main advantages of these trajectory representations consist in providing the continuous time evolution of the articulators.

A new continuous time representation of the vocal-tract is proposed here using the cross-sectional area function of the vocal-tract. This kind of representation is similar to the speech spectrogram and consists of a three-dimensional display of area function. Because the main parameters represented in this kind of display are the vocal-tract cross sectional *areas*, we call this graphic representation *areogram*, by analogy with the *spectrogram* where the *spectra* are displayed as a function of time. At any time instance, both area function and spectra are two dimensional functions. The area function is an amplitude versus vocal-tract length function and the spectra represent a magnitude versus frequency function.

A static vocal-tract shape can be approximated by a concatenation of uniform tubes having different cross sectional areas. These cross sectional areas can be displayed as an amplitude plot or using a brightness plot. A hypothetic vocal-tract area function with a complete closure at the glottis an a linear area function approximated by steps of 0.5 cm is presented in Fig. 6.3. The left plot displays the amplitude versus distance from glottis and the right one displays the same area function encoded using a grey scale. The brightness of each section of the vocal-tract is proportional to the amplitude of the corresponding cross sectional area of the section. The darkest section represents the portion of the vocal-tract with a complete constriction. A logarithmic scale is used to encode the area amplitude into different grey scale levels. The grey scale is compressed towards the darkest

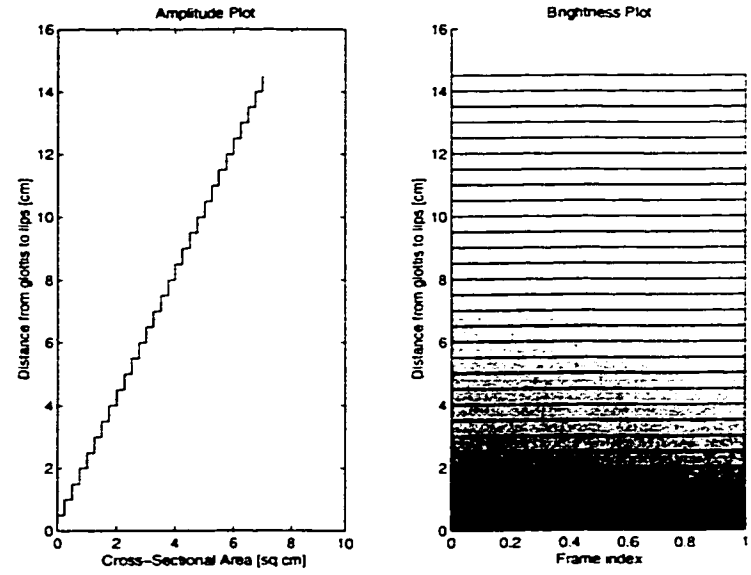

Figure 6.3: Amplitude and brightness area function representations for a hypothetic vocal-tract shape having a piecewise linear area function

levels in order to better represent constrictions with small cross sectional area.

In Fig. 6.4 and Fig. 6.5 the graphic area function representations are presented for the corresponding vocal-tract shapes of the vowels /aa/ and /iy/. The large areas from the front mouth cavity for vowel /aa/ and back cavity for vowel /iy/ are represented with the brightest grey levels.

In Fig. 6.6 and Fig. 6.7 the representations for the corresponding vocal-tract shapes of the consonants /b/ and /t/ are presented. The complete constrictions at the lips for the consonant /b/ and at the alveolar region for the consonant /t/ are represented with the darkest grey levels (black).

Concatenating the consecutive frames, a continuous time representation of the evolution of vocal-tract can be obtained. The *areogram* of the estimated vocal-tract, using the speech inversion method presented in this thesis, for an utterance

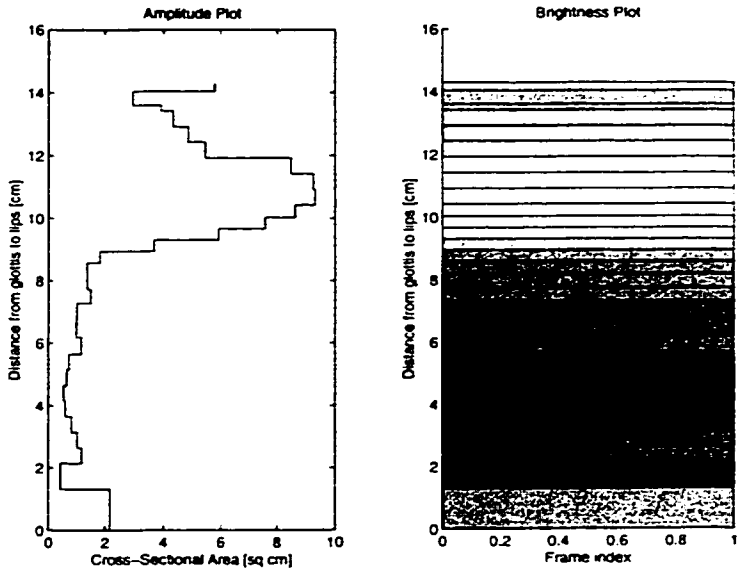

Figure 6.4: Amplitude and brightness area function representations for a vocal-tract shape of the vowel /aa/

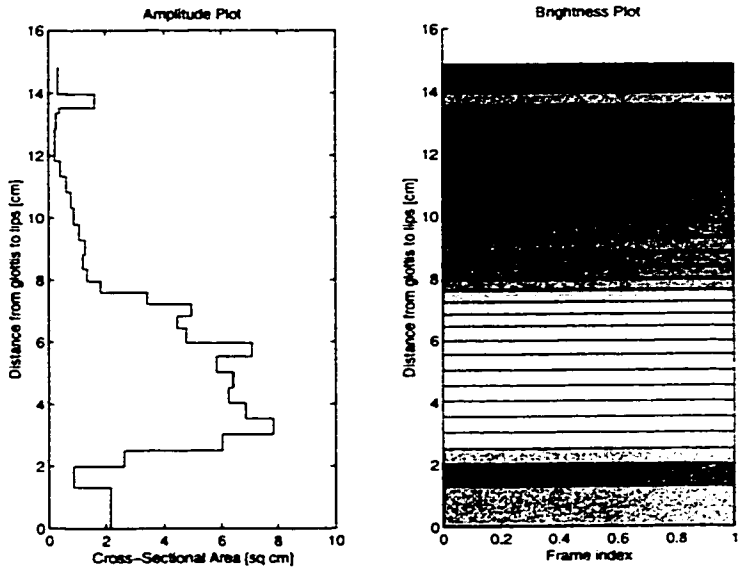

Figure 6.5: Amplitude and brightness area function representations for a vocal-tract shape of the vowel /iy/

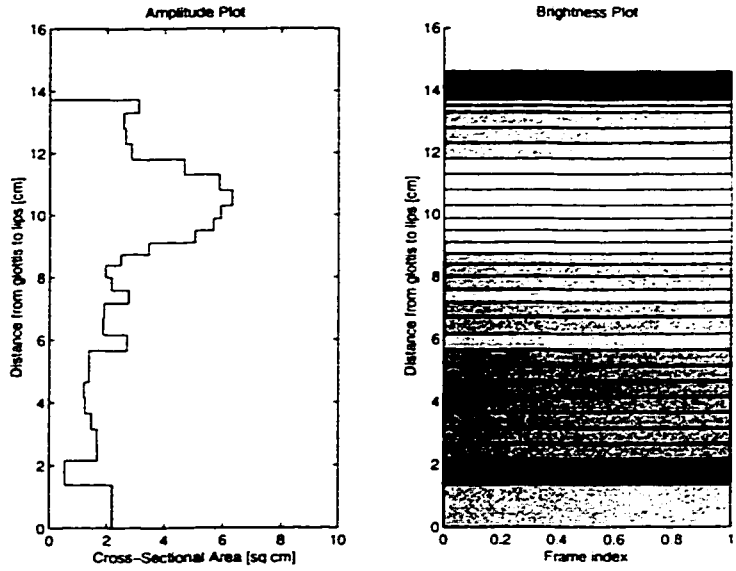

Figure 6.6: Amplitude and brightness area function representations for a vocal-tract shape of the consonant /b/

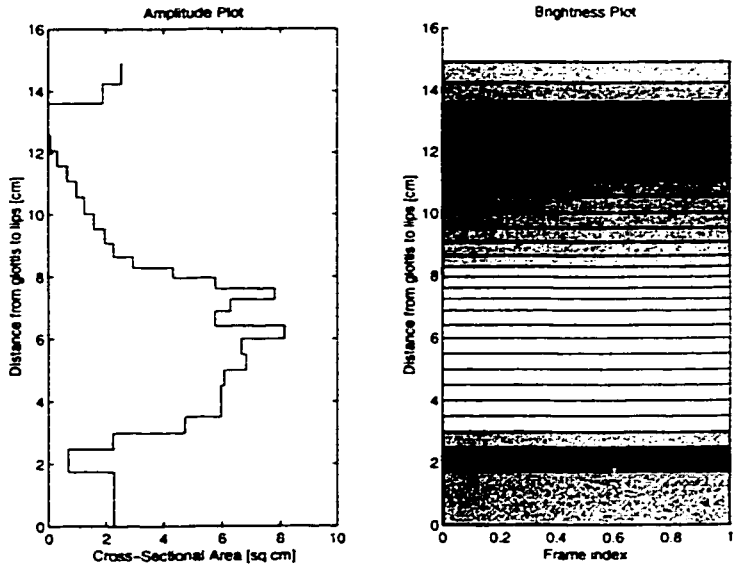

Figure 6.7: Amplitude and brightness area function representations for a vocal-tract shape of the consonant /t/

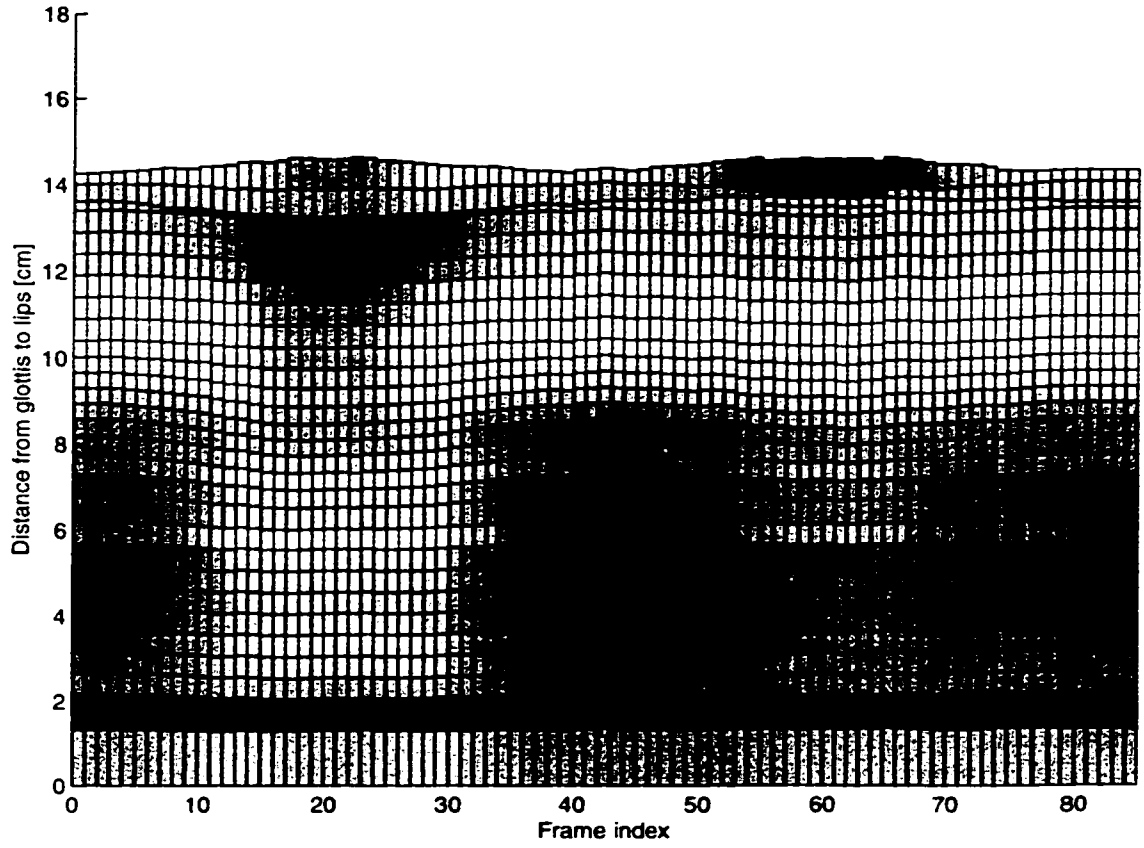

Figure 6.8: Areogram representation for the utterance /aa zh aa p aa/

/aa zh aa p aa/ is displayed in Fig. 6.8. For comparison, the spectrogram of the originally synthesized speech signal using the Maeda's models is presented in Fig. 6.9. Each small rectangle in the *areogram* plot represents a uniform section of the vocal-tract and has a uniform gray level. The abscissa represents the frame index in time. The time sampling of vocal-tract shapes is 10 ms/frame. Unlike in the corresponding spectrogram, in this *areogram* the position of the complete constriction (black level), here at the lips, can be immediately observed for the /p/ portion of the utterance. From the spectrogram the place of the articulation can only be inferred by trained persons from the formant trajectories before and after

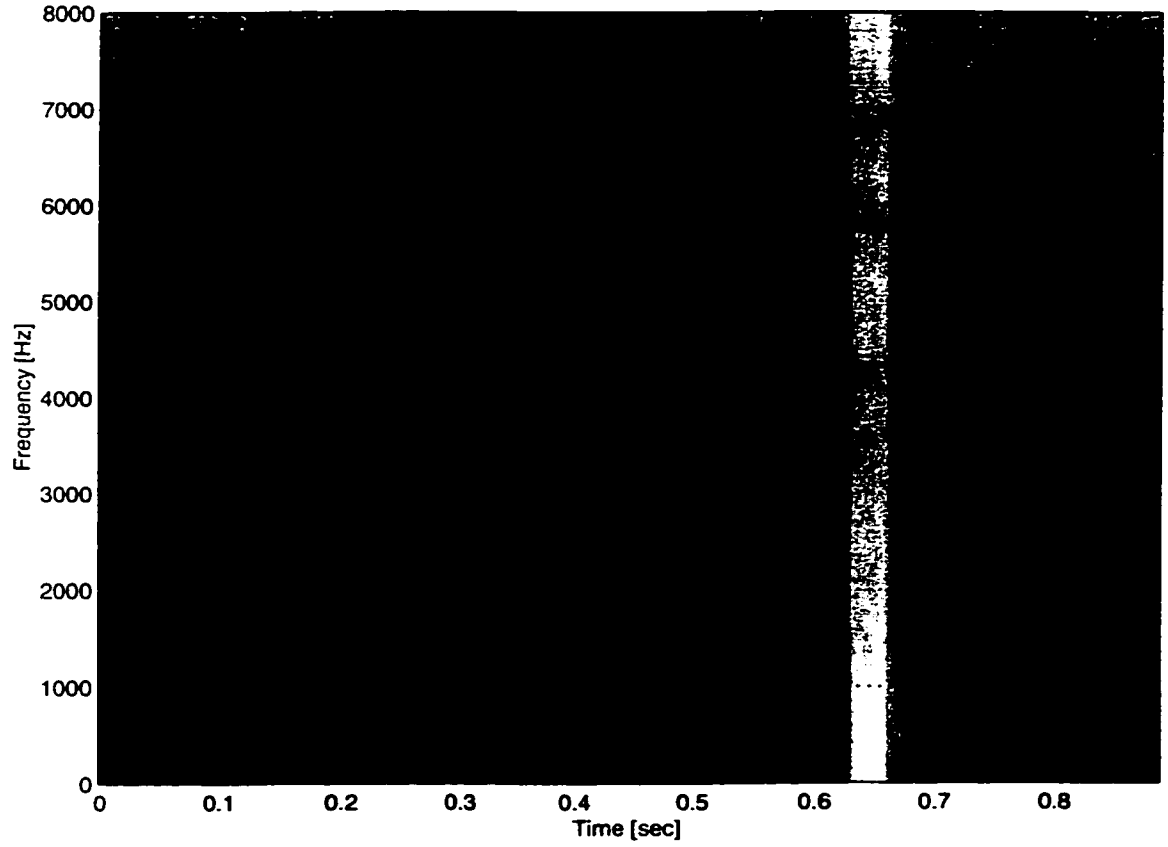

Figure 6.9: Spectrogram representation for the utterance /aa zh aa p aa/

the /p/ constriction.

This *areogram* representation of the evolution of the vocal-tract can be applied to any type of the articulatory model, as far as an area function can be computed or estimated. Also this graphic representation can be applied to any vocal-tract for which an area function has been obtained, e.g., from MRI images. The number of sections with which the vocal-tract shape is approximated can be arbitrarily large and it depends on the model used. The advantages of using this new graphic representation consist in providing a continuous time and global evolution of the vocal-tract and that this representation is based on absolute values of cross sectional areas along the tract. Once the area function is computed or estimated the graphical representation is independent of the model used to determine the area function. Similar graphic representations have been proposed by Hoole, 1993.[36], and Greisbach and Esser, 1993,[32], but using instead of amplitude of the area function the position of some points on the tongue. However, unlike the area function, the absolute position of the tongue does not represent globally the vocal-tract.

The application of this graphic representation is based on the estimated articulatory trajectories of a reference articulatory model, such as Maeda's, from the speech signal alone, as presented in this thesis. From the estimated trajectories, representing the articulatory model parameters, the vocal-tract area function can be easily computed and used in constructing the *areogram*.

In addition to the general applications of this displaying method in speech science research and phonetics, the potential applications of this *areogram* representation include teaching the speaking impaired or the deaf people to speak and teaching foreign languages.

6.2 Automatic Speech Recognition

The use of articulatory information to improve the performance of the automatic speech recognition (ASR) systems is now a long-lasting desiderate. Unfortunately, there is no practical approach of successfully recovering and applying the articulatory features to ASR. On the other hand, there are still debates regarding the way of application of articulatory information to ASR. These debates are on the questions of whether the articulatory information needs to be accurately recovered in a form of articulatory trajectories (Zlokarnik [107]), or it is enough to constrain the recognition process using this information as an a priori knowledge (Deng and Sun [17])? We quote from a critique [63], of a paper [80], presented at the 1994 Meeting of the Acoustical Society of America in the Special Session on 'Speech Recognition and Perception from an Articulatory Point of View.':

> . . . does a recognition process (natural *or* artificial) need knowledge about articulation (a) to "recover" that information or (b) to "constrain" the recognition (search) process? In other words, does the process of recognition have to "derive" a particular articulatory trajectory or is it enough for it to be "constrained" by allowable trajectories? (Roger K. Moore 1996, [63])

The speech inversion method presented in this study performs both recognition of phonological coproduction models and estimation of articulatory trajectories, simultaneously. Given the structure of the coproduction models which account for the coarticulation of two successive phonemes, by recognizing the particular succession of these models in a speech utterance, the phonemic recognition from these models is a straightforward simple task. The recognition of coproduction models is

based on the Bayes' rule, as presented in Section 4.5. The processes of model recognition and articulatory trajectory estimation are unified. The recognition of models uses the likelihood measure computed as a result of trajectory estimation (extended Kalman filtering), whereas the final trajectory estimation process benefits from the recognition of the coproduction models. Because the likelihood measure needed to make recognition decisions can be computed from the extended Kalman filtering forward recursions, without running the backward recursions of the smoother, the recognition of models can be considered as one step of time ahead of the final trajectory estimation process, which is based on the extended Kalman smoothing. However, although is not a computationally efficient solution, these two processes can be carried out simultaneously if the extended Kalman smoother is accomplished after each filtering in the model recognition phase, and the estimated trajectories are stored for all the smoothings. In this case the final filtering and smoothing are no longer necessary and the smoothed articulatory trajectories corresponding to the highest likelihood measure can be selected as the final estimates. The straightforward application of this speech inversion method to automatic speech recognition is thus an intrinsic part of this speech inversion approach.

Another application to automatic speech recognition is possible, based on the experimental evidence of improving the automatic speech recognition accuracy by adding articulatory features to the acoustic features (Petajan [71], Zlokarnik [107] and [108]). Those experiments have shown that by adding the articulatory features which were recorded simultaneously with the acoustic features, the error recognition rate has been reduced substantially. When the articulatory features were estimated from the acoustic features alone the error recognition rate was only slightly reduced. This suggests that those estimation methods were not accurate enough, and better estimation methods could eventually reduce the error rate significantly. Our ex-

periments have shown a good accuracy in estimating the articulatory trajectories using phonological and dynamical constraints. We believe that this good accuracy will provide a basis for improving the performance of automatic speech recognition systems based on acoustic and estimated articulatory features.

Chapter 7

Conclusion and Future Work

This chapter concludes this dissertation by summarizing the research described and by presenting the main contributions of this thesis. The future work is also discussed by presenting a few directions of some potential interest.

This thesis concerned the developing of a generalized method of inverting the articulatory-to-acoustic transformation. This speech research area is usually called speech inversion. It was not the objective of this study to apply the developed speech inversion method to all possible coproduction models, based on all combinations of the speech sounds, but to develop a generalizable inversion method and to evaluate it on different coproduction models, containing speech sounds from different classes (e.g., vowels, fricatives, stops, nasals). In this research, we proposed a new way of applying high-level linguistic constraints to the speech inversion. We developed a method of speech inversion based on dynamical system modeling in which we applied additional constraints — the phonological constraints — by modeling a different articulatory-acoustic function for each coproduction model consisting of any phonologically possible combination of two consecutive phonemes.

Initially, we intended to develop a generalized method of speech inversion based on some linearly transformed task spaces, with a reduced dimensionality. In order to apply this transformation method we carried out a simple experiment designed to answer to two questions: do articulatory features preserve enough information about the phonetic affiliation of speech sounds? and, is there a quasi-linear relationship between the articulatory and acoustic features of speech in transformed, task spaces, which have a drastically reduced dimensionality? In Chapter 3, we presented this preliminary analysis experiment. This experiment consisted of an articulatory-acoustic feature space transformation and a classification of vowels in both original (full) and task (reduced) spaces. Our simple experiment, provided an affirmative answer to the first question and a negative one to the second. Due to the negative answer, we did not try to implement the speech inversion method on a reduced task space, taking into account that we cannot benefit from a linear observation function on these task spaces while we may loose some important information, by drastically reducing the dimensionality of the original spaces.

The main chapter of this thesis, Chapter 4, described the general speech inversion method and each of its components. Thus, in the first section, we defined the coproduction segments and models of speech we were using. Then we presented a method of modeling the direct articulatory-acoustic function for each such coproduction model. This modeling is based on codebooks. Then, a method of estimating the model parameters was presented. This is a simple Maximum-Likelihood estimation method and is based on the articulatory and acoustic training data. The method of estimating the articulatory trajectories, based on the extended Kalman filtering and smoothing, was then described. At the end of that chapter, the general method of segmenting the speech, recognizing the units or models and finally estimating the articulatory trajectories was presented.

The speech inversion method described in this thesis was based on training speech data acquired from a single reference speaker or model. The application of this method to other speakers, can be carried out by employing a speaker normalization method or a vocal-tract length normalization method. In [22], we proposed an original method of estimating the overall vocal-tract length of other speakers, based on neural networks, and a method of normalization of the acoustic parameters, based on estimated overall vocal-tract length.

After the presentation of the main speech inversion method, in Chapter 5 we presented the experimental results. We carried out three different sets of experiments, to show the potential of the developed method of speech inversion, based on training articulatory and acoustic continuous speech data, and using dynamical and phonological constraints. The first set of experiments was designed to develop and implement the method, and was based on speech data synthesized with an articulatory-acoustic model. In a preliminary experiment of this kind, we evaluated the extended Kalman filtering state estimation method using formant frequencies as acoustic parameters. The articulatory-acoustic nonlinear function was approximated by a large codebook, generated randomly within the limits of the articulatory model used. Using the formant frequencies, this preliminary experiment, was limited to 10 American English vowels. The acoustic fit of the actual and reconstructed formants from the estimated articulatory trajectories was very good, although, there was no guarantee that the estimated articulatory parameters were realistic. Then we extended this preliminary experiment, by using a more general acoustic feature, the MFCC parameters. Again, the acoustic fit was very good. We applied in this second preliminary experiment, the inversion method to some speech data for which we had available the corresponding real articulatory trajectories, obtained from X-ray films. We found a relatively large difference

between the estimated trajectories and the real ones. We explained this fact by the numerous unrealistic vocal-tract shapes included into the articulatory-acoustic codebook, due to the random sampling. After these preliminary experiments, we approached the inversion method using our initial idea of constructing a different coproduction model for each phonologically possible combination of two consecutive phonemes. We successfully developed and applied these new constraints using continuous speech data, synthesized with the Maeda's articulatory and vocal-tract acoustic models. We applied this generalized method of inversion to combinations of speech sounds from different classes, including vowels, fricatives and stop consonants. We did not applied here the method to nasal sounds. However, the method has shown its potential on these synthesized speech data, and good results of trajectory estimation have been obtained.

The second set of experiments was based on real speech data, recorded with an Electromagnetic Midsagittal Articulograph, using the author of this thesis as a subject. The speech utterances were in a form of simple VCVs, in which the first and second vowels were the same, drown from five English vowels, and the consonant was each from a set of 17 English consonants used. Different coproduction models have been tested, including vowels, fricatives, nasals and stop consonants. The experimental results have shown good accuracies in estimating articulatory trajectories, based on real articulatory-acoustic training data. Average RMS errors of about 2 to 3 mm have been obtained for utterances not included into the training data. These results are comparable to those of a state-of-the-art study, presented in [99].

The third set of experiments was also based on real speech data, from the X-ray Microbeam Speech Production Database of University of Wisconsin, which we very recently have received. As in the previous set of experiments, also based on

real speech training data, the experimental results have shown a good accuracy in estimating articulatory trajectories. Average RMS errors of about 2 mm have been obtained for utterances not included into the training data.

Two different applications of the speech inversion method developed in this research are presented in Chapter 6. In the first application, we developed a new method of displaying the dynamics of the vocal-tract during continuous speech and using an articulatory model. This method of graphic representation is based on estimated articulatory parameters from which an area function can be easily computed as a function of time. We called this kind of representation *areogram*, by analogy to the spectrogram, in which the spectra are displayed as a function of time. This displaying method can have, in addition to the general application in speech research, a practical application in teaching the hearing or speaking impaired to speak, or in teaching foreign sounds and languages. The second application of the speech inversion method, discussed in that chapter, is that to automatic speech recognition. The direct application to automatic speech recognition is obvious, since the inversion method has an intrinsic part of segmenting the speech and recognizing the sequence of coproduction models. From these coproduction models, the phonetic classification and recognition is a straightforward step. However, we suggested another method of application to automatic speech recognition, by adding the estimated articulatory features to the acoustic ones in order to perform a speech recognition based on both acoustic and articulatory features.

7.1 Contributions

The main contribution of this thesis consists in generalizing a previously proposed speech inversion method based on Kalman filtering technique, [89], to all classes of

speech sounds. This generalization has been achieved by applying a few innovations to the initial method of articulatory state estimation based on the extended Kalman filtering and smoothing. These innovations were also implemented in order to overcome the difficulties related to consonants, revealed by a second study of speech inversion, based on Kalman filtering, [105]. The principal innovation consists in applying high-level linguistic constraints to the speech inversion method. This has been achieved by using different phonological models for any possible combination of two consecutive phonemes. Thus, the constraints are phonological and not phonetic, due to the phonological rules applied in construction of the coproduction models. The construction of the coproduction models was based on two parts: first, a different set of dynamical model parameters had to be estimated for each model, as proposed first in [91], [92] and latter applied in [75]; second, a different sub-function representing the articulatory-acoustic transformation had to be implemented for each model. This second part has not been applied before to the speech inversion problem, although, a speech inversion approach has used a simple division of the articulatory-acoustic codebook into two different parts, corresponding to voiced and unvoiced sounds, [79]. The whole method of inverting the articulatory-to-acoustic transformation, as developed in this research, represents an integrated approach of articulatory trajectory estimation and segmentation/phonetic-classification. As described in this report, we used the method of extended Kalman filtering combined with a Maximum-Likelihood method for performing the automatic segmentation and classification of coproduction models. The extended Kalman filtering was then used also as a forward step, followed by the backward step of smoothing, to obtain the estimated articulatory trajectories. The speech inversion method developed in this study, has overcome the difficulties and problems revealed by the previous approaches related to consonantal sounds, [105]. Our approach did not show any

problem in modeling these sounds, including stop, nasal and fricative consonants.

Another contribution of this research consists in proposing and successfully applying the direct method of model parameter estimation from training data using the Maximum-Likelihood method. This direct method has not been applied before to the speech inversion problem, although an extension of it, based on the EM algorithm has been proposed in [75]. A method based on codebooks, has been efficiently implemented to approximate the nonlinear articulatory-acoustic sub-function for each coproduction model.

Another contribution of this thesis consists in proposing a new method of displaying the dynamics of the vocal-tract over time, based on estimated articulatory trajectories and using an articulatory model to obtain the area functions. This method depends on the accuracy of the estimated articulatory trajectories and can only be applied in conjunction with an articulatory model.

7.2 Future Work

Future work, following this study, can be focused on the evaluation of the generalized speech inversion method on real speech data produced with different prosodic information from different speakers. Thus, a larger number of tokens, with different speed and emphasizing stress is needed for each coproduction model, in order to evaluate the method for data with large variabilities. Also, tokens of the same coproduction segment taken from different context words could show a larger variability in both articulatory and acoustic domains.

Another direction for future work is the evaluation of a variant of the inversion method based on tying the models with the same target phoneme (β) into a single

model called β. In this way, the number of models will be drastically reduced to about 40 to 50, the number of phonemes in a language. The models would become simple phonetic models. The set of dynamical model parameters will have in this case an averaged value over all the coproduction models tied. Such an approach would be more practical, due to the small number of models. However, we did a preliminary experiment of tying such models and the accuracy of the estimated articulatory trajectories has been affected.

A future experiment could be focused on an extensive evaluation of the method of approximating the articulatory-acoustic functions with multi-layer neural networks. It is known that such neural networks are able to approximate any nonlinear function. We believe that this method of linearizing the articulatory-acoustic function on small regions, will provide more accurate results than the method based on articulatory-acoustic codebooks.

The evaluation of the speech inversion method with some other acoustic feature, e.g., LPC parameters, and different frame intervals would also be an interesting direction. Even though, in this study we used two acoustic features — the formant frequencies and MFCC parameters — all these experiments were based on a 10 ms frame interval. It would also be interesting to see how the accuracy of the estimated trajectories will be affected by the dimension of the acoustic vectors. In this work we only used three formant frequencies and 10 MFCC coefficients. Increasing the dimensionality of the acoustic vectors, and decreasing the time of the analysis frame are expected to increase the accuracy of the estimated articulatory trajectories.

It would be interesting to evaluate the speech inversion method for a simple application of speech coding, using an articulatory model to synthesize the training data. A very low transmission rate is expected to be obtained by coding the slow varying estimated articulatory trajectories. However, probably the most important

application of this speech inversion method, would be to automatic speech recognition. For this application, though, a practical solution of tying the models would probably be needed. Although other speech processing areas, e.g., diphone synthesis, have to deal with the same problem of the large number of models (diphones), the construction of about 2000 such models for a language represents a real practical problem. In addition, we have to deal with the scarcity of the speech data from the training database. It is possible to have only one utterance for a model or even none. For the first case, we provided an approximate solution to construct a model from a single token. For the case when no token is available in the training data for a model, of course, the model cannot be constructed. Therefore, by tying the coproduction models into phonetic models, this problem can be overcome.

Appendix A

Approximating $\mathbf{g}^{(\alpha,\beta)}[\mathbf{z}]$ by Neural Networks

An alternative method, which we propose for the future work, for approximation and linearization of the articulatory-acoustic function could be based on artificial neural networks (NNs). Such a method uses the properties of NNs of general function approximation. For each coproduction model, a neural network can be created and trained with the same pairs of articulatory and acoustic vectors used to construct the codebooks. The neural network inputs correspond to the articulatory vectors whereas its outputs correspond to the acoustic vectors (e.g., MFCC parameters). The articulatory-acoustic mapping is non-linear, but is not one-to-many like the inverse mapping. An advantage of using the neural networks in linearizing the articulatory-acoustic function consists in the simple way of computing the Jacobian matrix needed for linearization. This method of computation is based on the backpropagation of the errors in calculation of the derivatives of the network outputs with respect to the network inputs. In order to be able to approximate the

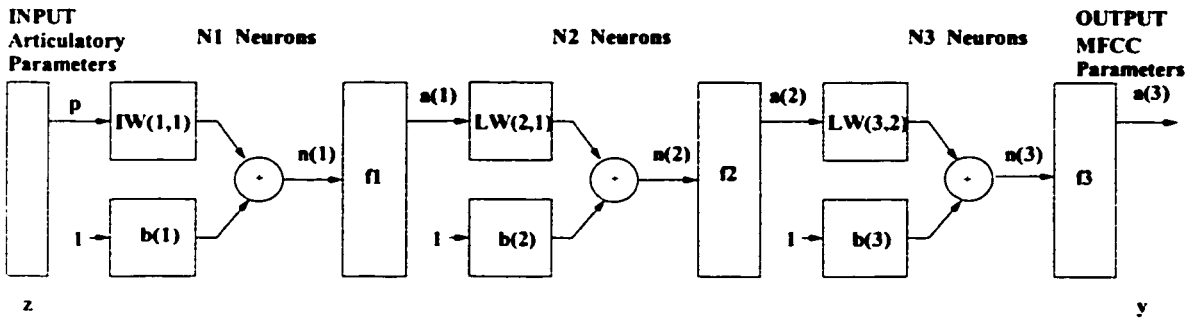

Figure A.1: Neural network of three layers for approximating an articulatory-acoustic sub-function

non-linear articulatory-acoustic function, the neural network has to be trained with pairs of simultaneously acquired articulatory and acoustic vectors. After training, the network is capable of approximating the non-linear articulatory-acoustic function for any new articulatory vector applied at its inputs. The non-linear mapping, represented by the articulatory synthesis function $\mathbf{g}^{(\alpha,\beta)}$, can be described by the equation

$$\mathbf{y} = \mathbf{g}^{(\alpha,\beta)}[\mathbf{z}], \tag{A.1}$$

where $\mathbf{z}$ is the articulatory vector and $\mathbf{y}$ is the acoustic vector. A suitable neural network consists of multiple layers perceptrons (MLP).

The topology of an MLP neural network, consisting of three layers is presented in a block diagram in Figure A.1. In this figure, the input vector of the network is denoted by $\mathbf{z}$, the weight matrices by $\mathbf{IW}$ and $\mathbf{LW}$, the bias vectors by $\mathbf{b}$, the input vectors of the transfer functions by $\mathbf{n}$, the output vectors of the transfer functions by $\mathbf{a}$ and the output vector of the network by $\mathbf{y}$. The first layer, is connected to the articulatory inputs and has N_1 neurons with sigmoid transfer functions.. The second layer has also N_2 neurons with sigmoid transfer functions. The third layer has N_3 neurons with purely linear transfer functions and its outputs correspond to

the acoustic parameters.

A method of computing the Jacobian matrix of the non-linear function, based on the backpropagation of the errors in the neural networks was presented by Bishop [3]. The Jacobian matrix $\mathbf{G}^{(\alpha,\beta)}(\mathbf{z})$ of the non-linear function, computed at the input vector $\mathbf{z} = \mathbf{z}_*$, has its elements given by the partial derivatives of each of the outputs with respect to each of the inputs

$$G_{ki}^{(\alpha,\beta)} = \frac{\partial g_k^{(\alpha,\beta)}}{\partial z_i} = \frac{\partial g_k^{(\alpha,\beta)}[\mathbf{z}(t)]}{\partial z_i(t)}|_{\mathbf{z}(t)=\mathbf{z}_*(t)}. \tag{A.2}$$

For clarity, we will omit further the model index and the time index. Because the output of the network represents the acoustic parameters, we will replace the g_k by y_k, and use the notation corresponding to Figure A.1. The elements of the Jacobian matrix can be derived as follows

$$G_{ki} = \frac{\partial y_k}{\partial z_i} = \sum_{j=1}^{20} \frac{\partial y_k}{\partial n_j(1)} \frac{\partial n_j(1)}{\partial z_i} = \sum_{j=1}^{20} \frac{\partial y_k}{\partial n_j(1)} IW_{ji}(1,1), \tag{A.3}$$

where

$$\frac{\partial y_k}{\partial n_j(1)} = \sum_{l=1}^{20} \frac{\partial y_k}{\partial n_l(2)} \frac{\partial n_l(2)}{\partial n_j(1)} = \sum_{l=1}^{20} \frac{\partial y_k}{\partial n_l(2)} LW_{lj}(2,1) f_1'(n_j(1)), \tag{A.4}$$

and

$$\frac{\partial y_k}{\partial n_l(2)} = \frac{\partial y_k}{\partial n_k(3)} \frac{\partial n_k(3)}{\partial n_l(2)} = f_2'(n_l(2)) LW_{kl}(3,2) f_3'(n_k(3)). \tag{A.5}$$

In order to find the elements of the Jacobian matrix, one needs to apply an articulatory vector, around which the matrix is needed, at the inputs of the network and compute all the activations vectors in the network. Then, by using the backpropagation approach, the elements of the Jacobian matrix can be computed using the equation

$$G_{ki} = \sum_{j=1}^{20} IW_{ji}(1,1) f_1'(n_j(1)) \sum_{l=1}^{20} LW_{lj}(2,1) f_2'(n_l(2)) LW_{kl}(3,2) f_3'(n_k(3)), \tag{A.6}$$

where $f_1'(n_j(1))$, $f_2'(n_l(2))$ and $f_3'(n_k(3))$ are the derivatives of the transfer functions of each layer in the neural network. The derivatives of the sigmoid functions can be computed directly using the original sigmoid transfer functions.

A direct way of computing the Jacobian matrix is by applying the recursive formula from above for each element of the matrix. Another way is by employing the matrix multiplication functions, and can be done by using the matrix equation

$$\begin{aligned}\mathbf{G} = \mathbf{IW}^T(1,1) * [\mathbf{f}_1'(\mathbf{n}(1)) * \mathbf{U}] .* \\ \mathbf{LW}^T(2,1) * [\mathbf{f}_2'(\mathbf{n}(2)) * \mathbf{U}] .* \\ \mathbf{LW}^T(3,2) * [\mathbf{f}_3'(\mathbf{n}(3)) * \mathbf{U}],\end{aligned} \tag{A.7}$$

where $\mathbf{U} = [11111...1]$ is a N_3 dimensional raw vector of ones and $.*$ represents the element-by-element multiplication of matrices.

The whole linearization of the articulatory-acoustic function can be accomplished by the trained neural network. For each reference articulatory input vector applied to the network, the implemented algorithm can compute the corresponding acoustic output vector (MFCC parameters) and the Jacobian matrix at the point represented by the reference vector.

An advantage of the neural network approximation method is that eliminates the need of clustering the data and the storing of the triple sets of parameters, corresponding to each of the piecewise linear region of the model. It is also an universal method of approximating any nonlinear function, such as the articulatory-acoustic one.

Bibliography

[1] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. "Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer-Sorting Technique". *Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978.

[2] D. Beautemps, P. Badin, and R. Laboissière. "Deriving Vocal-Tract Area Functions from Midsagittal Profiles and Formant Frequencies: A New Model for Vowels and Fricative Consonants Based on Experimental Data". *Speech Communication*, 16:27–47, 1995.

[3] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[4] C. Browman and L. Goldstein. "Towards an Articulatory Phonology". *Phonology Yearbook*, 3:219–252, 1987.

[5] C. Browman and L. Goldstein. "Articulatory Gestures as Phonological Units". *Phonology*, 6:201–251, 1989.

[6] C. Browman and L. Goldstein. "Tiers in Articulatory Phonology, with Some Implications for Casual Speech". In J. Kingson and M. Beckman, editors,

Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech, pages 341–376. 1990.

[7] C. Browman and L. Goldstein. "Articulatory Phonology: An Overview". *Phonetica*, 49:155–180, 1992.

[8] C. P. Browman and L. Goldstein. "Dynamic Modeling of Phonetic Structure". In V. Fromkin, editor, *Phonetic linguistics*, pages 35–53. 1985.

[9] R. Carré and M. Mrayati. "Articulatory-Acoustic-Phonetic Relations and Modeling, Regions and Modes". In Kluwer Academic Publishers, editor, *Speech Production and Speech Modelling*, pages 211–240. 1990.

[10] Carstens Medizinelektronik GmbH. *AG100 Technical Manual*, 1992.

[11] K. Chan and A. Wong. "APACS: A System for Automatic Analysis and Classification of Conceptual Patterns". *Computational Intelligence*, 6:119–131, 1990.

[12] T. Chiba and M. Kajiyama. *The Vowel. Its Nature and Structure*. Tokyo, 1941.

[13] C. Coker and O. Fujimura. "Model for Specification of the Vocal-Tract Area Function". *Journal of the Acoustical Society of America*, 40:1271, 1966.

[14] J. Dang and K. Honda. "A Physiological Model of the Dynamic Vocal Tract for Speech Production". *Technical Report of ATR*, TR-H-247, 1998.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[16] L. Deng and D. Sun. "Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features". In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1635–1638, 1993.

[17] L. Deng and D. Sun. "A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features". *Journal of the Acoustical Society of America*, 95(5 pt. 1):2702–2719, 1994.

[18] V. Digalakis, J. R. Rohlicek, and Ostendorf. M. "A Dynamical System Approach to Continuous Speech Recognition". In *International Conference on Acoustics, Speech and Signal Processing*, pages 289–292, 1991.

[19] V. Digalakis, J. R. Rohlicek, and Ostendorf. M. "ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition". *IEEE Transactions on Speech and Audio Processing*, SAP-1(2):431–442, 1993.

[20] S. Dusan and L. Deng. "Estimation of Articulatory Parameters from Speech Acoustics by Kalman Filtering". In *Proceedings of CITO Researcher Retreat, Hamilton, Canada*, 1998.

[21] S. Dusan and L. Deng. "Recovering Vocal-Tract Shapes from MFCC Parameters". In *Proceedings of International Conference on Spoken Language Processing*, pages 2251–2254, 1998.

[22] S. Dusan and L. Deng. "Vocal-Tract Length Normalization for Acoustic-to-Articulatory Mapping Using Neural Networks". *Journal of the Acoustical Society of America*, 106(4 pt. 2):2pSC5, 1999.

[23] S. Dusan, L. Deng, and A. Wong. "Classification of Articulatory and Acoustic Patterns of Speech in the Task Spaces". In *Proceedings of TRIO/ICRT Researcher Retreat, Kingston, Canada*, 1997.

[24] O. Engwall. "Modeling of the Vocal Tract in Tree Dimensions". In *Proceedings of the EUROSPEECH'99*, volume 1, pages 113–116, 1999.

[25] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[26] J. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, New York:, 1972.

[27] J. Flege, S. Fletcher, and A. Homiedan. "Compensating for a Bite Block in /s/ and /t/ Production: Palatographic, Acoustic, and Perceptual Data". *Journal of the Acoustical Society of America*, 83(1):212–228, 1988.

[28] G. D. Forney. "Convolutional Codes II: Maximum-Likelihood Decoding". *Information and Control*, 25(3):222–266, 1974.

[29] B. Gopinath and M. M. Sondhi. "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements". *The Bell System Technical Journal*, July-August:1195–1214, 1970.

[30] A. Graham. *Kronecker Products and Matrix Calculus: with Applications*. Ellis Horwood Limited, Chichester, England, 1981.

[31] R. Gray. "Vector Quantization". *IEEE ASSP Magazine*, pages 4–28, 1984.

[32] R. Greisbach and O. Esser. "The POLARGRAM Display of Tongue Movements Measured by Articulography". In Institut fur Phonetik und Sprachliche Kommunikation der Universitat Munchen, editor, *IPSKUM Forschungsberichte*, pages 169–179. 1993.

[33] S. Gupta and J. Schroeter. "Pitch-Syncronous Frame-by-Frame and Segment-Based Articulatory Analysis by Synthesis". *Journal of the Acoustical Society of America*, 94(5):2517–2530, 1993.

[34] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. "Accurate Recovery of Articulator Positions from Acoustics - New Conclusions Based on Human Data". *Journal of the Acoustical Society of America*, 100(3):1819–1834, 1996.

[35] M. Honda and T. Kaburagi. "Statistical Analysis of the Phonemic Target in Articulatory Movements". In *ASA and ASJ Third Joint Meeting*, pages 821–824. Honolulu, Hawaii, 1996.

[36] P. Hoole. "Methodological Considerations in the Use of Electromagnetic Articulography in Phonetic Research". In Institut fur Phonetik und Sprachliche Kommunikation der Universitat Munchen, editor, *IPSKUM Forschungsberichte*, pages 43–64. 1993.

[37] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York:, 1970.

[38] T. Kaburagi and M. Honda. "A Model of Articulator Trajectory Formation Based on the Motor Task of Vocal-tract Shapes". *Journal of the Acoustical Society of America*, 99(5):3154–3170, 1996.

[39] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". *Transaction ASME Journal of Basic Engineering*, 8:35–45, 1960.

[40] S. King and A. Wrench. "Dynamical System Modeling of Articulator Movement". In *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.

[41] P. Kohler and A. Lacroix. "Speech Synthesis on the Basis of Acoustical Tube Models for Vocal and Nasal Tract". *Signal Processing*, V:1143–1146, 1990.

[42] B. J. Krőger, G. Schrőder, and C. Opgen-Rhein. "A Gesture-Based Dynamic Model Describing Articulatory Movement Data". *Journal of the Acoustical Society of America*, 98(4):1878–1889, 1995.

[43] S. Krstulovic. "LPC-based Inversion of the DRM Articulatory Model". In *Proceedings of the European Conference on Speech Communication and Technology*, pages 125–128, 1999.

[44] R. Laboissière. *"Préliminaires pour une robotique de la communication parlée: inversion et contrôle d'un modèle articulatoire du conduit vocal"*. PhD thesis. INP, Grenoble, 1992.

[45] R. Laboissière and A. Galván. "Inferring the Commands of an Articulatory Model from Acoustical Specifications of Stop/Vowel Sequences". In *Proceedings of the XIII-th International Congress of Phonetic Sciences*, volume 1, pages 358–361, 1995.

[46] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice. "Generating Vocal Tract Shapes from Formant Frequencies". *Journal of the Acoustical Society of America*, 64(4):1027–1035, 1978.

[47] J. Larar, J. Schroeter, and M. Sondhi. "Vector Quantization of the Articulatory Space". *IEEE Transactions on ASSP*, 36(12):1812–1818, 1988.

[48] A. Liberman, F. Cooper, D. Shankweiler, and M. Studdert-Kennedy. "Perception of the Speech Code". *Psychological Review*, 74(6):431–461, 1967.

[49] A. Liberman and I. Mattingly. "The Motor Theory of Speech Perception Revised". *Cognition*, 21:1–36, 1985.

[50] B. Lindblom, J. Lubker, and T. Gay. "Formant Frequencies of Some Fixed-Mandible Vowels and a Model of Speech Motor Programming by Predictive Simulation". *Journal of Phonetics*, 7:146–161, 1979.

[51] Y. Linde, A. Buzo, and R. Gray. "An Algorithm for Vector Quantizer Design". *IEEE Transactions on Communications*, COM-28(1):84–95, 1980.

[52] A. Ljolje and M. D. Reley. "Automatic Segmentation and Labeling of Speach". In *International Conference on Acoustic , Speech and Signal Processing*, pages 473–476, 1991.

[53] S. Maeda. "An Articulatory Model of the Tongue Based on a Statistical Analysis". *Journal of the Acoustical Society of America*, 65(S22), 1979.

[54] S. Maeda. "A Digital Simulation Method of the Vocal-Tract System". *Speech Communication*, 1:199–229, 1982.

[55] S. Maeda. "Improved Articulatory Model". *Journal of the Acoustical Society of America*, 84:S146, 1988.

[56] R. McGowan and M. Lee. "Task Dynamic and Articulatory Recovery of Lip and Velar Approximations under Model Mismatch Conditions". *Journal of the Acoustical Society of America*, 99(1):595–608, 1996.

[57] J. M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control.* Prentice Hall Ptr, Englewood Clifs, New Jersey, 1995.

[58] P. Mermelstein. "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements". *Journal of the Acoustical Society of America*, 41(4 pt. 2):1002–1010, 1967.

[59] P. Mermelstein. "Determination of the Vocal-Tract Shape from Measured Formant Frequencies". *Journal of the Acoustical Society of America*, 41(5):1283–1294, 1967.

[60] P. Mermelstein. "Articulatory Model for the Study of Speech Production". *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[61] P. Mermelstein and M. R. Schroeder. "Determination of Smoothed Cross-Sectional Area Functions of the Vocal Tract from Formant Frequencies". In *Proceedings of the Fifth International Congress on Acoustics*, volume 1a., 1965.

[62] P. Meyer, R. Wilhelms, and H. Strube. "An Efficient Vocal Tract Model Running in Real Time". *Signal Processing*, III:377–380, 1986.

[63] R. K. Moore. "Critique: The Potential Role of Speech Production Models in Automatic Speech Recognition". *Journal of the Acoustical Society of America*, 99(3):1710–1713, 1996.

[64] D. Morrison. *Multivariate Statistical Methods*. McGraw Hill, 1967.

[65] National Institute of Standards and Technology (NIST). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Oct. 1990.

[66] M. Ostendorf and S. Roukos. "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-37(12):1857–1869, 1989.

[67] J. E. Overall. "Orthogonal Factors and Uncorrelated Factors Scores". *Psychological Report*, 10:651–662, 1962.

[68] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy. "Inferring Articulation and Recognizing Gestures from Acoustics with a Neural Network Trained on X-ray Microbeam Data". *Journal of the Acoustical Society of America*, 92(2):688–700, 1992.

[69] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson. "Electromagnetic Midsagittal Articulometer System for Transducing Speech Articulatory Movements". *Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

[70] P. Perrier, L. Boë, and R. Stock. "Vocal-Tract Function Estimation from Midsagittal Dimensions with CT Scans and a Vocal-Tract Cast: Modeling the Transition with Two Sets of Coefficients". *Journal of Speech and Hearing Research*, 35:53–67, 1992.

[71] E. D. Petajan. "Automatic Lipreading to Enhance Speech Recognition". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 40–47, 1985.

[72] G. E. Peterson and H. L. Barney. "Control Methods Used in a Study of Vowels". *Journal of the Acoustical Society of America*, 24:175–184, 1952.

[73] M. R. Portnoff. "A Quasi-One-Dimensional Digital Simulation for the Time-Varying Vocal Tract,". Master's thesis, Dept. of Elect. Engr., MIT, Cambridge, Mass., June 1973.

[74] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[75] G. Ramsay and L. Deng. "Maximum-Likelihood Estimation for Articulatory Speech Recognition using a Stochastic Target Model". In *Proceedings of the EUROSPEECH'95*, pages 1401–1404, 1995.

[76] G. Ramsay and L. Deng. "Optimal Filtering and Smoothing for Speech Recognition Using a Stochastic Target Model". In *International Conference on Spoken Language Processing*, pages 1113–1116, 1996.

[77] H. E. Rauch. "Solutions to the Linear Smoothing Problem". *IEEE Transactions on Automatic Control*, 8:371–372, 1963.

[78] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle. "Deriving Articulatory Representation of Speech". In *Proceedings of the EUROSPEECH'95*, pages 761–764, 1995.

[79] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle. "Deriving Articulatory Representation of Speech with Various Excitation Modes". In *Proceedings of International Conference on Spoken Language Processing*, pages 1233–1236, 1996.

[80] R. C. Rose, J. Schroeter, and M. M. Sondhi. "The Potential Role of Speech Production Models in Automatic Speech Recognition". *Journal of the Acoustical Society of America*, 99(3):1699–1709, 1996.

[81] P. Rubin, T. Baer, and P. Mermelstein. "An Articulatory Synthesizer for Perceptual Research". *Journal of the Acoustical Society of America*, 70(2):321–328, 1981.

[82] E. L. Saltzman and K. G. Munhall. "A Dynamical Approach to Gestural Patterning in Speech Production". *Ecological Psychology*, 1(4):333–382, 1989.

[83] P. Schonle, K. Grabe, P. Wenig, J. Hohne, J. Schrader, and B. Conrad. "Electromagnetic Articulography: Use of Alternating Magnetic Fields for Tracking Movements of Multiple Points Inside and Outside the Vocal-Tract". *Brain Language*, 31:26–35, 1987.

[84] J. Schroeter, J. N. Larar, and M. M. Sondhi. "Speech Parameter Estimation Using a Vocal Tract/Cord Model". In *International Conference on Acoustics. Speech and Signal Processing*, pages 308–311. 1987.

[85] J. Schroeter and M. Shondi. "Technics for Estimating Vocal-Tract Shapes from the Speech Signal". *IEEE Transactions on Speech and Audio Processing*. SAP-2(1):133–150, 1994.

[86] J. Schroeter and M. M. Sondhi. "Dynamic Programming Search of Articulatory Codebooks". In *International Conference on Acoustics. Speech and Signal Processing*, pages 588–591, 1989.

[87] K. Shirai. "Vowel Identification in Continuous Speech Using Articulatory Parameters". In *International Conference on Acoustics. Speech and Signal Processing*, pages 1172–1175, 1981.

[88] K. Shirai. "Estimation and Generation of Articulatory Motion Using Neural Networks". *Speech Communication*, 13:45–51, 1993.

[89] K. Shirai and M. Honda. "Estimation of Articulatory Motion". In Tokyo University Press, editor, *Dynamic Aspects of Speech Production*, pages 279–302. 1976.

[90] K. Shirai and T. Kobayashi. "Recognition of Semivowels and Consonants in Continuous Speech Using Articulatory Parameters". In *International Conference on Acoustics, Speech and Signal Processing*, pages 2004–2007. 1982.

[91] K. Shirai and T. Kobayashi. "Considerations on Articulatory Dynamics for Continuous Speech Recognition". In *International Conference on Acoustics, Speech and Signal Processing*, pages 324–327, 1983.

[92] K. Shirai and T. Kobayashi. "Estimating Articulatory Motion from Seech Wave". *Speech Communication*, 5:159–170, 1986.

[93] K. Shirai and T. Kobayasi. "Estimation of Articulatory Motion Using Neural Networks". *Journal of Phonetics*, 19:379–385, 1991.

[94] M. Shondi and J. Schroeter. "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(7):955–967, 1987.

[95] V. Sorokin. "Determination of Vocal Tract Shape for Vowels". *Speech Communication*, 11(1):71–85, 1992.

[96] V. Sorokin. "Inverse Problem for Fricatives". *Speech Communication*, 14:249–262, 1994.

[97] V. Sorokin and A. Trushkin. "Articulatory-to-Acoustic Mapping for Inverse Problem". *Speech Communication*, 19:105–118, 1996.

[98] K. N. Stevens. "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data". In McGraw-Hill, editor, *Human Communication: a Unified View*, pages 51–66. 1972.

[99] S. Suzuki, T. Okadome, and M. Honda. "Determination of Articulatory Positions from Speech Acoustics by Applying Dynamic Articulatory Constraints". In *Proceedings of International Conference on Spoken Language Processing*, pages 2251–2254, 1998.

[100] T. Svendsen and F. K. Soong. "On the Automatic Segmentation of Speech Signal". In *International Conference on Acoustic , Speech and Signal Processing*, pages 77–80, 1987.

[101] B. Tuller, S. Shao, and J. A. S. Kelso. "An Evaluation of an Alternating Magnetic Field Device for Monitoring Tongue Movements". *Journal of the Acoustical Society of America*, 88(2):674–679, 1990.

[102] A. J. Viterbi. "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm". *IEEE Transactions on Information Theory*, 1967.

[103] H. Wakita. "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms". *IEEE Trans. Audio Electroacoust.*, AU-21:417–427, 1973.

[104] J. R. Westbury. *X-ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin, Madison, WI, 1994.

[105] R. Wilhelms, P. Meyer, and H. W. Strube. "Estimation of Articulatory Trajectory by Kalman Filter". In I.T. Young et al., editor, *Signal Processing III: Theories and Applications*, pages 477–480. Elsevier Science Publishers B.V.(North-Holland), 1986.

[106] S. A. Zahorian and S. Venkat. "Vowel Articulation Training Aid for the Deaf". In *International Conference on Acoustics, Speech and Signal Processing*, pages 1121–1124, 1990.

[107] I. Zlokarnik. "Experiments with an Articulatory Speech Recognizer". In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2215–2218, 1993.

[108] I. Zlokarnik. "Adding Articulatory Features to Acoustic Features for Automatic Speech Recognition". *Journal of the Acoustical Society of America*, 97(5 pt. 2):3246, 1995.