DYNAMIC SYSTEM MODELING AND STATE ESTIMATION FOR SPEECH SIGNAL

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

İ. YÜCEL ÖZBEK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENG.

APRIL 2010

Approval of the thesis:

DYNAMIC SYSTEM MODELING AND STATE ESTIMATION FOR SPEECH SIGNAL

submitted by **İ. YÜCEL ÖZBEK** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Eng. Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen _____
Head of Department, **Electrical and Electronics Eng.**

Prof. Dr. Mübeccel Demirekler _____
Supervisor, **Electrical and Electronics Eng. Dept.**

Examining Committee Members:

Prof. Dr. Kemal Leblebicioğlu
Electrical and Electronics Engineering Dept., METU _____

Prof. Dr. Mübeccel Demirekler
Electrical and Electronics Engineering Dept., METU _____

Assoc. Prof. Dr. Tolga Çiloğlu
Electrical and Electronics Engineering Dept., METU _____

Assist. Prof. Dr. Afşar Saranlı
Electrical and Electronics Engineering Dept., METU _____

Prof. Dr. Salim Kayhan
Electrical and Electronics Engineering Dept., HU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: İ. YÜCEL ÖZBEK

Signature :

ABSTRACT

DYNAMIC SYSTEM MODELING AND STATE ESTIMATION FOR SPEECH SIGNAL

Özbek, İ. Yücel
Ph.D., Department of Electrical and Electronics Eng.
Supervisor : Prof. Dr. Mübeccel Demirekler

April 2010, 128 pages

This thesis presents an all-inclusive framework on how the current formant tracking and audio (and/or visual)-to-articulatory inversion algorithms can be improved. The possible improvements are summarized as follows:

The first part of the thesis investigates the problem of the formant frequency estimation when the number of formants to be estimated fixed or variable respectively.

The fixed number of formant tracking method is based on the assumption that the number of formant frequencies is fixed along the speech utterance. The proposed algorithm is based on the combination of a dynamic programming algorithm and Kalman filtering/smoothing. In this method, the speech signal is divided into voiced and unvoiced segments, and the formant candidates are associated via dynamic programming algorithm for each voiced and unvoiced part separately. Individual adaptive Kalman filtering/smoothing is used to perform the formant frequency estimation. The performance of the proposed algorithm is compared with some algorithms given in the literature.

The variable number of formant tracking method considers those formant frequencies which are visible in the spectrogram. Therefore, the number of formant frequencies are not fixed and they can change along the speech waveform. In that case, it is also necessary to estimate the number of formants to track. For this purpose, the proposed algorithm uses extra logic (formant track start/end decision unit). The measurement update of each individual formant trajectories is handled via Kalman filters. The performance of the proposed algorithm is illustrated by some examples

The second part of this thesis is concerned with improving audiovisual to articulatory inversion performance. The related studies can be examined in two parts; Gaussian mixture model (GMM) regression based inversion and Jump Markov Linear System (JMLS) based inversion.

GMM regression based inversion method involves modeling audio (and /or visual) and articulatory data as a joint Gaussian mixture model. The conditional expectation of this distribution gives the desired articulatory estimate. In this method, we examine the usefulness of the combination of various acoustic features and effectiveness of various types of fusion techniques in combination with audiovisual features. Also, we propose dynamic smoothing methods to smooth articulatory trajectories. The performance of the proposed algorithm is illustrated and compared with conventional algorithms.

JMLS inversion involves tying the acoustic (and/or visual) spaces and articulatory space via multiple state space representations. In this way, the articulatory inversion problem is converted into the state estimation problem where the audiovisual data are considered as measurements and articulatory positions are state variables. The proposed inversion method first learns the parameter set of the state space model via an expectation maximization (EM) based algorithm and the state estimation is handled via interactive multiple model (IMM) filter/smoother.

Keywords: Formant Tracking, Audiovisual-to-Articulatory Inversion, Dynamic

System Modelling, State Estimation, Kalman Filtering

ÖZ

KONUŞMA İŞARETİ İÇİN DİNAMİK SİSTEM MODELLEME VE DURUM KESTİRİMİ

Özbek, İ. Yücel
Doktora, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Mübeccel Demirekler

Eylül 2009, 128 sayfa

Bu tez çalışması formant frekanslarının izlenmesi ve akustikden (ve/veya görselden) artikülatörlere evirme algoritmalarının performanslarının iyileştirilmesi için kapsamlı bir çerçeve sunmaktadır. Olası iyileştirmeler özet olarak aşağıda sunulmaktadır.

Bu tezin ilk bölümü sabit ve değişken sayıdaki formant frekanslarının kestirimlerinin nasıl yapılması gerektigini araştırmaktadır.

Sabit sayıdaki formant frekanslarının izlenmesi, sabit sayıda formant frekansının konuşma süresi boyunca var olduğu varsayımına dayanmaktadır. Önerilen yöntem dinamik programlama ve Kalman süzgeci/düzgünleştiricisinin birleştirilmesi (birlikte kullanılması) ilkesine dayanmaktadır. Bu yöntemde konuşma işareti ötümlü veya ötümsüz olarak bölütlendirilmekte ve her bir bölüt için formant adayları ile formant izleri, dinamik programlama yardımıyla eşleştirilmektedir. Eşleştirme işleminden sonra her bir formant frekansının kestirimi Kalman süzgeci /düzgünleştiricisi ile yapılmaktadır. Önerilen bu algoritmanın performansı literatürde varolan diğer algoritmalar ile karşılaştırılmıştır.

Değişken sayıdaki formant frekanslarının izlenmesi esnasında yalnızca spektrogramda görülen formant frekansları dikkate alınmaktadır. Bu nedenle izlenmesi gereken formant frekans sayısı zamanla değişebilmektedir. Bu durumda izlenecek formant frekanslarının sayısınında kestirilmesi gerekmektedir. Bu amaçla önerilen yöntem bazı algoritmalar (formant başlatma/bitirme karar mekanizması gibi) kullanmaktadır. Herbir formant frekans izinin gelen ölçümle beslenmesi Kalman süzgeci ile yapılmaktadır. Bu yöntemin başarısı çeşitli örneklerle gösterilmiştir.

Bu tezin ikinci bölümünde akustikden (ve/veya görselden) artikülatörlere evirme algoritmalarının performansları iyileştirilmiştir. Bu konuda yapılan çalışmalar iki kategoride incelenmektedir: Gaussian karışım modellere (GKM) dayalı evirme ve doğrusal atlamalı Markov sistemlere (DAMS) dayalı evirme.

GKM yöntemine dayalı evirmede artikülatörlerin hareketleri (pozisyonları) ve akustik (ve/veya görsel) veriler ortak dağılımlı Gaussian karışımı olarak modellenir. Bu dağılımın şartlı ortalaması, istenilen kestirim fonksiyonudur. Önerilen bu yöntemde, farklı akustik özniteliklerin birleştirilmesinin faydaları ve farklı füzyon yöntemlerinin akustik ve görsel verilerin birleştirilmesindeki etkinlikleri incelenmiştir. Ayrıca kestirilen artikülatörsel izlerin düzgünleştirilmesi için farklı dinamik düzgünleştirme yöntemleri önerilmiştir. Önerilen yöntemlerin performansı literatürde var olan diğer algoritmalar ile karşılaştırılmıştır.

DAMS yöntemine dayalı evirmede akustik uzay ile artikülatörsel uzayı birbirine çoklu sayıda durum-uzay gösterimleri ile bağlanmaktadır. Bu yöntemle artikülatörsel evirme problemi, ölçümleri akustik (ve/veya görsel) veriler olan, durum vektörünün ise artikülatörlerin pozisyonlarından oluşan durum kestirimi problemine dönüştürülmektedir. Önerilen evirme yöntemi öncelikle durum-uzay modellerinin parametrelerini beklenti enbüyültülmesi (BE) yöntemi ile öğrenir ve durum etkileşimli çoklu model (EÇM) süzgeci/düzgünleştirici yardımı ile kestirilir.

Anahtar Kelimeler: Formant İzleme, Akustik ve Görselden Artikülatörlere Evrime,

Dinamik Sistem Modelleme, Durum Kestirimi, Kalman Süzgeci

To my wife, Mehlika and my lovely daughter, Zeynep.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisor, Prof. Dr. Mübeccel Demirekler, for her guidance, patience and encouragement throughout my Ph.D. studies at the METU. I greatly appreciate her share in every step taken in the development of the thesis.

I am also grateful to Assoc. Prof. Dr. Mark Hasegawa-Johnson, for his guidance, invaluable comments and suggestions have contributed to the thesis during my studies at the University of Illinois at Urbana-Champaign.

I would also like to thank the member of my thesis committee, Assoc. Prof. Dr. Tolga Çiloğlu, Assist. Prof. Dr. Afşar Saranlı and Prof. Dr. Salim Kayhan for their support and suggestions which improved the quality of the thesis.

I owe so much to my dear friend Umut Orguner (Assist. Prof. Dr at the LIU in Sweden) who has always been there to help me whenever I needed. He is the only one that contributes to the thesis with every aspect: invaluable discussion, constant encouragement and support, useful feedback and suggestion, and true friendship. I can never thank him enough for his support and friendship. Without him I could never finish this thesis.

I am also very grateful to Evren Ekmekçi who is my roommate for his friendship and company, who made my times in the department enjoyable and worthy. He never hesitated to lend a hand whenever I needed.

I also would like to thank to my fellows: Emre Özkan, Eren Akdemir, Evren İmre, Turgay Koç, İsmail Tirtom, Alper Koz and Sebahatin Topal for the ideas they shared with me, their support and friendship through these years.

Lastly, I would like to thank to my wife, Mehlika, for her unlimited support and patience, and my parents, Yalçın and Cevriye Özbek, who made me feel their complete trust and support at all moments in my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Speech processing is and will probably always be an important ingredient of ever speeding artificial intelligence quest of mankind. A tremendous amount of research effort has been devoted into specific sub-areas of speech processing in the last several decades. Although being exhaustive about these sub-areas is impossible, we can count them roughly as

- Speech recognition;
- Speaker recognition;
- Text to speech synthesis;
- Speech coding.

Speech processing itself can be seen as a sub-field of signal processing and hence the so called transfer function (input-output) based methodologies have dominated the field for a long time after Fant's work [2]. On the other hand, in control theory, which is a quite related field to signal processing, after the seminal work of Kalman [3], the state space based methods have been the leading force behind the studies. There have been many work applying such state space methods to speech processing too. For example, state space models have been used in speech enhancement [4, 5], speech coding [6, 7], speech recognition [8, 9, 10], formant tracking [11, 12, 13] and articulatory inversion [14], etc.

This thesis is related to the (further) improvement of some speech processing sub-blocks using the state space methodology which would in turn be expected to lead to improved speech processing performance. The thesis can be divided

roughly into two parts:

- Improved formant frequency estimation using state space methodologies
- Improved acoustic(and/or visual) to articulatory inversion using state space methodologies

In the first part of this thesis, we improve current formant frequency estimation methods by using state space methodologies. The formant frequencies are closely related to natural frequencies of vocal tract. The air path of speech production mechanism (the vocal tract) is composed of various articulatory cavities (oral, pharyngeal, nasal, sinus etc.). Each cavity has particular natural frequencies at which the contained air naturally tends to vibrate. If the air inside the vocal tract is vibrated at natural frequencies, the vibrations are reinforced and the vocal tract resonates. That is, the vocal tract from glottis to lips acts as an acoustic resonator during speech production [15, 16]. The resonance frequencies, also known as formants, can be observed as the peaks of the magnitude spectrum of the speech signal. Since formants are a rich source of information about uttered speech and the speaker, the reliable formant estimation is critical for a wide range of applications, such as speech synthesis [15, 16, 17, 18, 19, 20], speech recognition [21, 22], voice conversion [23], vocal tract normalization [24, 25], measuring vocal tract length [26], accent classification [27] and speech enhancement [11].

The estimation of the formant frequencies is a difficult problem and must be solved efficiently and effectively before a good performance speech processing applications can be built. There is a numerous amount of research about formant tracking in the literature and they can be broadly categorized into three. The first category includes the filter bank based methods where the frequency content of the speech signal is extracted via filter banks. In [28, 29], AM-FM modulation model and inverse-filter control method are applied to formant tracking by using fixed-center-frequency band-pass filters. Time-varying adaptive filter bank methods, [30, 31, 32, 33], are later introduced to overcome the difficulties encountered in the extraction of the formant frequencies via fixed-centered filter banks. Estimating the correct center frequencies of the filter

banks and suppressing the leakage signals from closely separated filters still remain as open problems of these methods. The spectrum representation based methods, which relate the formants to the peaks of the spectrum, comprise the second category. In this category of methods, the formant estimation is done in two stages. In the first stage, formants (candidates) are estimated in frame-base via peak-picking [34] or root-solving [35, 36, 37, 38, 39]. It is also possible to represent the spectrum as a Gaussian mixture model (GMM) [40, 41] in which the mixture components comprise the formant candidates. In the second stage, some continuity and consistency constraints are imposed on the formant candidates in order to find the actual formant frequencies. Different methods including dynamic programming [36, 37, 38], Hidden Markov Model (HMM) [35, 42] or other function minimization techniques [39] are utilized for the consistency check. The main difficulties of this category of methods are setting the continuity constraints along the speech utterance and eliminating the spurious peaks in the spectrum. The final category contains the techniques that cast the problem into a state-space estimation framework. The formant frequencies are modeled as time-varying hidden states of a dynamical system and the estimation is done through the observations. It is possible to use the speech signal itself as the observation [12, 43, 44]. If the dynamical system is non-linear, Extended Kalman Filter (EKF) [12, 43] or EKF with interacting multiple models (IMM) filter [44] can be used for state estimation. In [13], observations are chosen as Mel frequency cepstral coefficients (MFCC) and the nonlinear relation between the state and the observations are modeled by Multi-Layer Perceptron (MLP) neural networks. In [45, 1, 46, 47, 48], the use of Linear predictive cepstral coefficient (LPCC) is preferred instead of MFCC to reduce the degree of non-linearity. In these works, nonlinear state estimation is done via particle filters [48], or by linearizing the observation equation [1, 46, 47], or by using Kalman filter/smoother with quantized observations [45]. There are also some studies which are the combination of spectrum representation and dynamic system modeling. As an example, in [49] a Kalman filtering technique is combined with an HMM.

The proposed formant tracking methods in this thesis are based on linear predictive coding (LPC). For each frame, the roots of denominator of LPC filter are considered as measurements (formant candidates) and the aim is to estimate the format trajectories from these measurements. From this point of view, the formant estimation problem using a state space framework is actually what is called as "target tracking" problem [50, 51, 3] in control theory. Target tracking is a mature area of research and the methodologies applied vary according to whether the number of targets tracked is fixed or variable. In accordance with this, we propose two methods for formant tracking, one for fixed and the other for variable number of formants.

The fixed number of formant tracking method is based on the assumption that a fixed number of formant frequencies exist in the speech utterance even if some of the resonance frequencies are not visible in the spectrogram. We propose a systematic framework for tracking fixed number of formant frequencies [52, 53, 54] . The difficulties involved in estimating the format frequencies change according to whether the part of the speech utterance that is under consideration is voiced or unvoiced. The spectrum of a voiced speech segment is accurately represented by the LPC spectrum and therefore, the poles of the LPC filter give relatively accurate formant candidate frequencies. On the other hand, the spectrum of an unvoiced speech segment is (relatively) poorly represented by the LPC spectrum and therefore, the formant candidate obtained from an unvoiced speech segment will be noisy and relatively inaccurate. Considering all these facts, in the proposed fixed number of format tracking method, the speech signal is segmented into voiced and unvoiced parts. For each part, the data association problem (which formant candidate belongs to which formant trajectory (track)) is solved via a dynamic programming algorithm separately. After the data association stage, the individual formant frequencies are tracked via Kalman filtering/smoothing. The parameter set for the state-space model that Kalman filter uses is changed according to the type of speech segment.

The variable number of formant tracking method proposed in this thesis consider those formant frequencies which are visible in the spectrogram. Therefore, the number of formant frequencies are not fixed and they can change from one

speech segment to another. For this purpose, we propose a formant tracker (called Visible Vocal Tract Resonance (VVTR) tracker) [55] which is also based on linear predictive coding (LPC). The only predefined parameter is the order of the LPC filter (there is no need to define a pre-defined number of formants to track). The algorithm has a formant track start/end decision unit and new trajectories start if there are consistent measurements in the consecutive frames and old trajectories end if no more consistent formant candidates are available. If the suitable formant candidates are available, a Kalman filter tracks the formants and forms trajectories (tracks).

The second part of this thesis is concerned with improving acoustic (and/or visual) to articulatory inversion performance by using the state space methodologies. Articulatory inversion is involved in advanced text to speech systems as a crucial block and by increasing its performance, one can increase the performance of such systems considerably. The problem of the articulatory inversion involves the estimation of the movements (position) of articulatory organs such as, lip, jaw, tongue, etc. from given speech utterance and visual data. Electromagnetic Articulography (EMA) trajectories provide the movement of certain articulators during a speech utterance. Similar to formant frequencies, the movement of the articulators also shows slowly varying dynamic properties and the estimation of these movements can provide useful information about the speech production and this information can be used in a variety of speech processing applications including speech recognition and synthesis.

The early studies in the literature about acoustic-to-articulatory inversion use an analytical function between acoustic and articulatory space [56, 57, 58]. The analytical formulation is based on solutions of the wave equations and boundary conditions, which are highly nonlinear. Similar to these studies, the articulatory inversion process involves finding the mapping from acoustic to vocal tract area [59, 60, 16, 61]. After the development of electromagnetic articulography (EMA), the acoustic and articulatory data is recorded simultaneously [62, 63]. Therefore, the articulatory inversion process turns into finding the mapping from acoustic (and/or visual) to articulatory positions. Recently, various methods have been proposed to find reliable estimates of the articulatory trajectories.

Some examples of different methods and features used in this area are: Neural networks and mixture density networks in [64, 65, 66]; Gaussian mixture model (GMM) regression in [67]; HMMs in [68]; Support vector machine (SVM) regression in [69, 70] and a combination of acoustic and visual features in [71, 72].

In this thesis, articulatory inversion studies are mainly based on two parts: Gaussian mixture model (GMM) regression based inversion and Jump Markov Linear System (JMLS) based inversion.

GMM regression based inversion method involves the modeling audio (and /or visual) and articulatory data as a joint Gaussian mixture model. The conditional expectation of this distribution gives the desired articulatory estimate. The parameter set of the joint distribution is estimated using Expectation Maximization (EM) algorithm. In this method, we show that the formant trajectories and their energies are useful for GMM regression based inversion method [73]. In addition to this, we show that different audio-visual fusion strategies also significantly improve the performance of the articulatory inversion process. A very important contribution here is the Kalman smoother based smoothing stage for the output of the GMM inversion which improves post processing performance significantly. This type of processing also enables the incorporation of auxiliary information (phonetic transcription of the test data) into the smoothing process boosting the performance of the articulatory inversion.

JMLS based inversion method involves the state space representation of the complete inverse problem (and hence not only the smoothing part) which is posed as a state estimation problem. In this method, audio features are considered as measurements (observable quantities), and the articulatory positions are then represented by the state vector (the hidden quantity). The parameter set of the JMLS is estimated using a training database. In the training stage, we propose a learning algorithm, which is a generalization of the Hidden Markov Model (HMM) based training given in [68] to estimate the unknown parameter set. The estimation of the state and its smoothed version is handled via jump Markov linear systems [74, 75].

1.1 Detailed Organization of The Thesis

Chapter 2 investigates the problem of formant frequency estimation when the number of formants to estimate is fixed. Although, the number of formant frequencies is known, the so called data association problem still exists and has to be taken care of since there are multiple formant frequencies. In this chapter, we propose a dynamic programming based approach to solve this problem and individual adaptive Kalman filters are used to perform the formant frequency estimation. The process noise terms of the individual Kalman filters are adjusted according to the voiced and unvoiced parts of the speech waveforms.

Chapter 3 investigates the problem of formant frequency estimation when the number of formants to estimate is variable. The case where the number of formant frequencies involved is unknown is more challenging and necessitates also the estimation of this number along with formant frequencies using extra logic. We, in in this chapter, resort to the area of multi-target tracking and apply a nearest neighbor methodology which is equipped with special target initiation and deletion logic to the problem.

Chapter 4 deals with improving the current acoustic to articulatory inversion systems by employing state space modeling and estimation. We use Kalman smoothers in the smoothing phase of the current acoustic to articulatory inversion algorithms. We show that, by using such dynamic smoothers, we are able to not only get better performance but also incorporate auxiliary information into the smoothing process more easily. Also the fusion of audio and visual information becomes more structured enabling the invention of modified versions of the well-known fusion alternatives.

Chapter 5 presents an articulatory inversion method which is based on JMLSs. In this chapter, we replace the whole acoustic to articulatory inversion system (not only the smoothing phase) by a dynamic state estimator. However, in order to be able to get sufficient modality in our system, we use the modeling framework of jump Markov linear systems (JMLS).

1.2 Contributions of The Thesis

Although some of the merits of this thesis have been mentioned above, we list the major contributions of it below for the sake of clarity:

- Contributions in format tracking
  - Combining dynamic programming algorithm with state space modeling in the tracking of fixed-number of formant frequencies (Chapter 2).
  - The derivation of the state space representation of the concatenated tube model for explaining the existence of variable number of formants in the spectrogram with the controllability concept (Chapter 3).
  - A multi-target tracking based algorithm for tracking variable number of formants (Chapter 3).
- Contributions in acoustic to articulatory inversion
  - Examination of the effects of incorporating various types of acoustic (and visual) features into the articulatory inversion process (Chapter 4).
  - Proposal of the formant frequencies as an effective feature for acoustic-to-articulatory inversion.
  - Unlike the currently used low-pass filter based smoothers, the proposal of dynamic (model-based) smoothing in order to make use of as much information from the GMM inversion (both the mean and covariances) as possible (Chapter 4).
  - Effective incorporation of extra information such as phonetic transcription (if any) into the smoothing framework (Chapter 4).
  - A novel audio-visual fusion method (called as "modified fusion algorithm") that is shown to give superior performance than the conventional methods (Chapter 4).

- A novel approach of using Jump Markov linear system model (JMLS) for articulatory inversion is presented (Chapter 5).
- Effective and efficient JMLS training algorithms based on ML and MAP criteria for JMLS is introduced (Chapter 5).
- An efficient smoothing algorithm for JMLS is introduced (Chapter 5).

Most of the material presented has been submitted to academic journals and were already presented in peer-reviewed conferences. The following is a list of the resulting publications:

- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler,"Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing" Submitted to *IEEE Trans. Audio Speech Lang. Process.,*
- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler,"On Improving Dynamic State Space Approaches to Articulatory Inversion with MAP based Parameter Estimation" To be submitted to *IEEE Trans. Audio Speech Lang. Process.,*
- Emre Özkan and **İ. Y. Özbek** and M. Demirekler, "Dynamic Speech Spectrum Representation and Tracking Variable Number of Vocal Tract Resonance Frequencies With Time-Varying Dirichlet Process Mixture Models" *IEEE Trans. Audio Speech Lang. Process.*, vol.17, no.8, pp. 1518-1532, Nov. 2009
- **İ. Y. Özbek** and M. Demirekler,"ML vs. MAP Parameter Estimation of Linear Dynamic System for Acoustic-to-Articulatory Inversion: A Comparative Study" Submitted to *in Proc. EUSIPCO*, 2010
- **İ. Y. Özbek** and M. Demirekler,"Audiovisual Articulatorty Inversion Based on Gaussian Mixture Model (GMM)" in *in Proc. SİU*, 2010
- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler,"Formant Trajectories for Acoustic-to-Articulatory Inversion" in *in Proc. Interspeech*, 2009
- V. Mitra and **İ. Y. Özbek** and H. Nam and X. Zhou and C. Espy-Wilson,"From Acoustic to Vocal Tract Time Functions" in *in Proc. ICASSP*, 2009
- **İ. Y. Özbek** and M. Demirekler,"Vocal Tract Resonances Tracking Based on Voiced and Unvoiced Speech Classification Using Dynamic Programming and Fixed Interval Kalman Smoother" in *in Proc. ICASSP*, 2008
- **İ. Y. Özbek** and M. Demirekler,"Tracking of vocal tract resonances based on dynamic programming and Kalman filtering" in *in Proc. SİU*, 2008

- **İ. Y. Özbek** and M. Demirekler,"Tracking of Visible Vocal Tract Resonances (VVTR) Based on Kalman Filtering" in *in Proc. Interspeech*, 2006
- **İ. Y. Özbek** and M. Demirekler,"Tracking of Speech Formant Frequencies" in *in Proc. SİU*, 2006

CHAPTER 2

TRACKING FIXED NUMBER OF VOCAL TRACT RESONANCE FREQUENCIES

2.1 Introduction

Vocal tract resonances (VTRs) contain very useful information about uttered speech and speaker. They are used in many speech applications (i.e. speech recognition, synthesis, accent classification etc). Hence, reliable estimation of formants is important in order to improve performance of these applications. Recently, numerous methods are proposed to track formants that use Kalman filtering (KF) [1, 46, 55, 47], dynamic programming (DP) [36, 38, 37], HMM [42], GMM [76] or combination of them [49, 77]. In this work, we propose a new Kalman filtering/smoothing and dynamic programming combination algorithm to track and estimate formant frequencies accurately. Doing this combination, we consider formant tracking process as a kind of multi-target tracking process. In multi-target tracking applications, there are two important issues; data association (that is, which measurement belongs to which target), and position estimation. Using a similar idea, we consider formant candidates form LPC analysis stage as measurements from targets that correspond to formant frequencies. DP is considered as a data association stage, in which labeling of the formant candidates are handled. Estimation of formant location is done in KF stage. The proposed method is explained in Fig.2.2 in detail. From our point of view, without using the KF stage, the tracker has lack of main estimation stage. Fig.2.2 indicates that the formant tracking procedure applied to voiced and unvoiced parts of the speech are not the same. Indeed this is one of the factors that

improve the performance of the system. The reason for this differentiation is the basic observation that for voiced regions formant candidates given by LPC is much more reliable compared to the unvoiced regions. The direct implication of this observation is the differentiation of parameters of trackers for two cases. For the voiced speech, nominal formant frequencies (independent of phone) are used as additional information in DP part with relatively low importance. Furthermore, the formant measurements (output of the DP stage) contain 'low noise' so the model generated for KF part has a small measurement noise covariance. For the unvoiced speech, the line connecting formants (similar to [38, 76, 33]) of the proceeding and succeeding voiced regions are used as nominal formant frequencies which are called 'estimated nominal VTRs'. They are quite effective in DP stage where LPC outputs are not reliable. KF parameters are selected according to a re-examination of the voicing decision. The measurement covariance parameter in KF is relatively high for unvoiced part due to less reliable LPC outputs

2.2 Baseline And The Proposed Method

The baseline method is well-known formant tracking system based on dynamic programming [36, 38, 37]. The general block diagram of the baseline system can be seen in Fig.2.1. The sub-blocks of the baseline system are explained in Sec. 2.2.2 and Sec.2.2.3 The general scheme of proposed VTR estimation procedure can be seen in Fig.2.2. The sub-blocks are explained as follows.

2.2.1 Unsupervised Speech Segmentation And Segment Based Classification: Voiced vs. Unvoiced

In this work, we use Level building dynamic programming (LBDP) algorithm in order to segment speech signal into homogenous units. LBDP is first introduced by Rabiner [78] as an algorithm to solve the connected word recognition problem and it is used in automatic speech segmentation task with some modifications [79]. This method is suitable for fulfilling segmentation task without

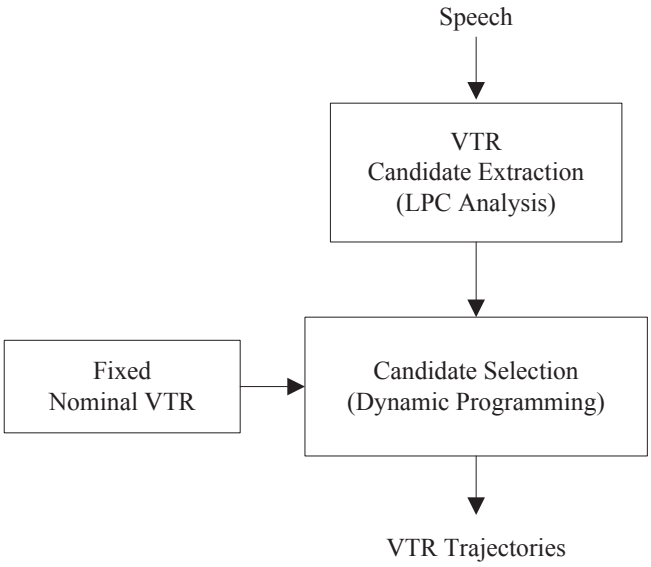


Figure 2.1: General scheme of baseline formant estimation procedure.

using training data. The algorithm only requires the number of segments for a given utterance to be segmented. An LBDP-based method optimally locates the segment boundaries by minimizing a distortion metric by using a dynamic programming based method. The details of the method can be found in [79]. The number of segments, L is proportional to the duration of the speech utterance, $\mathcal{T}$ (sec) . The number of segments can be chosen as $L = 40\mathcal{T}$. One of the drawbacks of LBDP-based algorithm is its computational complexity. To overcome this difficulty, it is necessary to use maximum and minimum duration of the segment units to prune the search space. In this work, the range of the search space of each level is limited to 0.025-0.5 secs. After segmentation phase, assuming that we obtained L speech segments $S = \{S_0, S_1, \ldots, S_L\}$, in classification stage, we should make a decision on whether the given segment is voiced or unvoiced. For this purpose we use two energy measurements and corresponding threshold values. First one is the average energy of the segment in dB and corresponding threshold is denoted by T_E. The second measurement is the energy ratio of the low frequency band (100 - 900 Hz) and high frequency band (3700 - 5000 Hz) in dB and its corresponding threshold is the T_{LH}. Using these two thresholds, voiced and unvoiced segments are determined by the

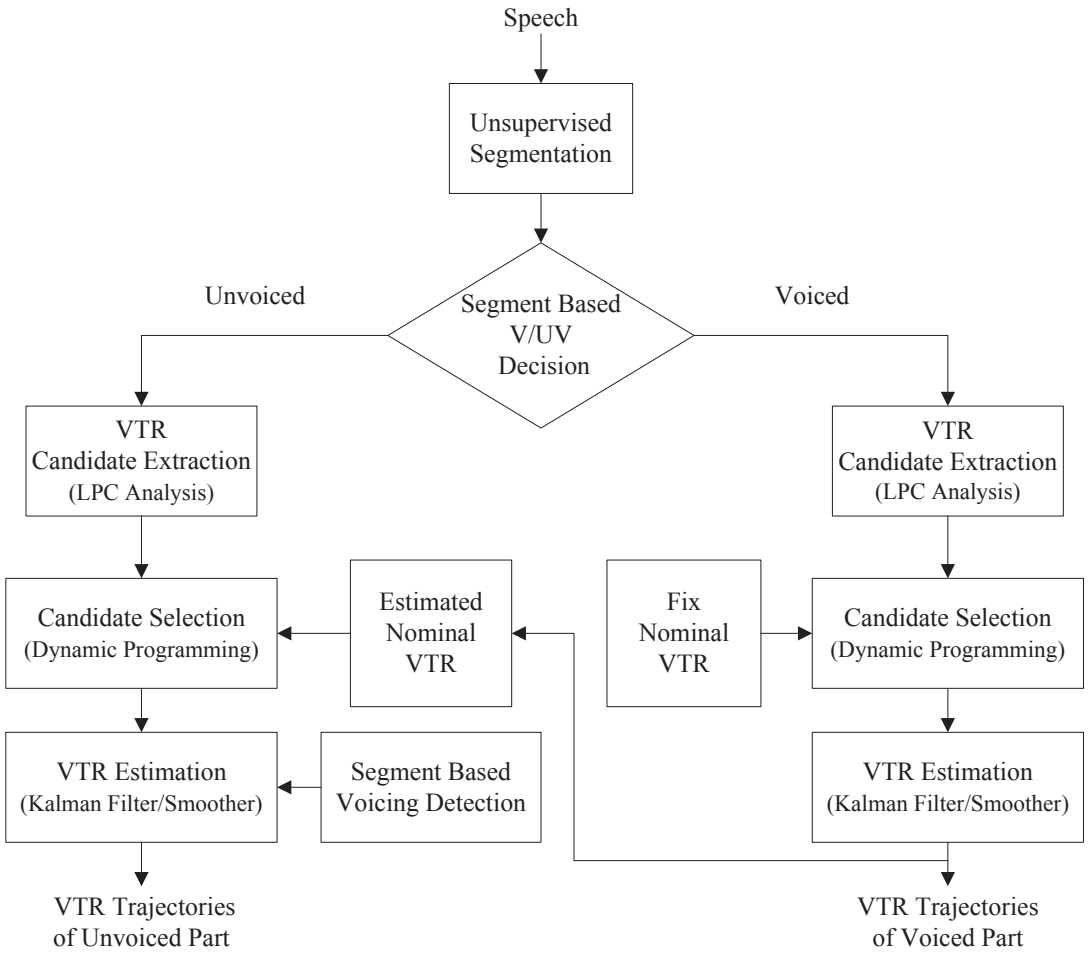


Figure 2.2: General scheme of the proposed VTR estimation procedure.

following two sets A and B.

$$A = \{S_i | \text{ Average Energy of } S_i > T_E \text{ for all i}\} \tag{2.1}$$

$$B = \{S_j | \text{ Energy Ratio of Low to High Band of } S_j > T_{LH} \text{ for all j}\} \tag{2.2}$$

The voiced segments are the set of $V = A \cap B$ and unvoiced segments are the set of $UV = S - (A \cap B)$. Fig.2.3 denotes an example for unsupervised speech segmentation.

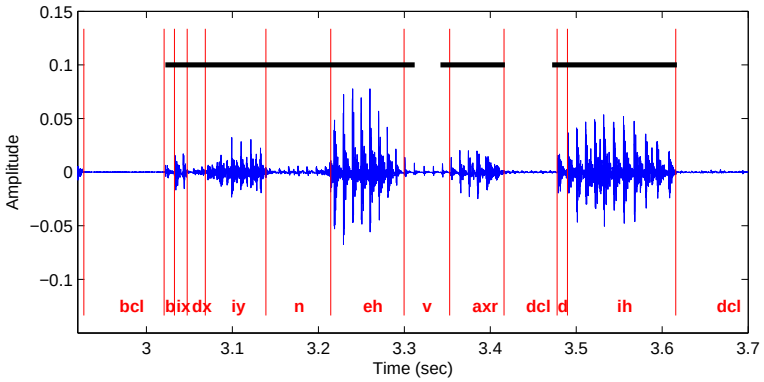


Figure 2.3: The speech utterance segmented into voiced and unvoiced part. The solid black lines show the voiced segment. The utterance is taken from VTR database [1]

2.2.2 Vocal Tract Resonance Candidates Based on LPC

The aim of this work is to track four lower vocal tract resonance frequencies. For this purpose, the sampling frequency of the speech signal is chosen as 10 KHz. Voiced speech signals have a natural spectral tilt, with the lower frequencies (below 1 kHz) having greater energy than the higher frequencies. The lower frequencies have more energy because they contain the glottal waveform and the radiation load from the lips. Therefore, it is desirable to remove this spectral tilt by using a pre-emphasis filter in order to improve formant tracker's performance. A common method of pre-emphasis is to filter the speech signal using a high-pass filter that attenuates the low frequencies and the energy in the speech signal is re-distributed to be approximately equal in all frequency regions [80]. The transfer function of the high-pass filter $H_p(z)$ used in this work is given as follows

$$H_p(z) = 1 - 0.97z^{-1}$$

There are several methods for formant candidate extraction such as peak picking [34] and linear prediction (LP) model pole extraction [35, 36, 37, 38, 39], etc. In this work we choose LP method to find formant candidates due to its simplicity of the method. The LP model is a linear prediction model where it is assumed that the signal $s(k)$ is predictable from a limited number of its past p values

[81]:

$$s(k) = a_1 s(k-1) +, ..., + a_p s(k-p) + e(k)$$
$$s(k) = \sum_{i=1}^{p} a_i s(k-i) + e(k) \tag{2.3}$$

where, a_i is the ith linear prediction coefficient , $e(k)$ is the prediction error and $s(k)$ is the speech signal. In the z-domain, (2.3) can be written as follows

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + E(z) \tag{2.4}$$

And forming Eq. (2.4) into transfer function form;

$$\begin{aligned} H(z) &= \frac{S(z)}{E(Z)} \\ &= \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \end{aligned} \tag{2.5}$$

This filter is the well-known all pole filter [80]. The input of the filter is the error function which can be interpreted as the unpredictable (innovation) part of the signal which derives the all-pole system. In this work to extract the formant candidates, first the speech signal is divided into overlapping frames of length 40 ms via Hamming window. These segments have 6 ms overlap with each other. For each frame the linear prediction coefficients are estimated by autocorrelation method. The primary LP order is set to 12 so that there usually will be six complex-pole pairs which introduce the resonances of the system. The real valued poles are eliminated and the poles are then sorted regarding to their frequencies. To find formant candidates, $A(z)$ is written as;

$$\begin{aligned} A(z) &= 1 - \sum_{k=1}^{p} a_k z^{-k} \\ &= \sum_{k=1}^{p/2} (1 - c_k z^{-1})(1 - c_k^* z^{-1}) \end{aligned} \tag{2.6}$$

where,

$$c_k = |c_k| e^{j\omega_k} \text{ is the } k\text{th complex root of } A(z)$$

The formant candidate frequencies R_k and bandwidths B_k are estimated as follows

$$R_k = \frac{\omega_k}{2\pi T_s} \tag{2.7}$$

$$B_k = -\frac{\ln(|c_k|)}{\pi T_s} \tag{2.8}$$

where, T_s is the corresponding sampling frequency.

2.2.3 Estimation of Vocal Tract Resonances

In this work, the estimation of the vocal tract resonances is handled by Kalman smoothing. For each resonance frequency, we use one Kalman filter. The critical point in this method is to choose correct measurement (resonance) candidate to update Kalman filter. For this purpose it is necessary to associate the resonance candidates with formant tracks. There are some methods in the literature to solve this problem [46, 55]. In this work, we use dynamic programming (DP) approach [36].

2.2.3.1 VTR Candidate Classification (Selection)

We use Viterbi-like dynamic programming algorithm to classify (select) resonance candidates for VTR estimation phase. The states of the Viterbi-like algorithm correspond to all possible formant track/candidate associations. As an example, if there are 4 tracks and 6 candidates, the number of states is $N_s = \frac{6!}{(6-4)!4!} = 15$. From the definition it is obvious that N_s may change for each frame since the candidates change. In applying DP algorithm, we define two types of costs: incremental local cost D_L and transitional cost D_T. Definition of them is similar to [36, 38, 37] and are given below. The local cost $D_L(\cdot)$ is related to our knowledge about VTR without using any temporal context and it is defined as

$$D_L(S_k = m) = \sum_{i=1}^{N} (\alpha\beta_{im} + \Gamma\eta_i \frac{|R_{im} - R_i^n|}{R_i^n}) \tag{2.9}$$

Here, S_k denotes the state at frame k. N is the number of VTR, β_{im} is the bandwidth of the ith resonance at mth state which is weighted by α that is

independent VTR index. R_{im} is the ith VTR candidate at the mth state. R_i^n is the ith nominal VTR. The normalized mean distance between the candidate and nominal VTR is weighted by η_i and Γ. The transitional cost $D_T(\cdot)$ which forces the resonance candidates to be continuous is defined as:

$$D_T(S_k = m|S_{k-1} = p) = \sum_{i=1}^{N} \varphi_i \left(\frac{R_{im}(k) - R_{ip}(k-1)}{R_{im}(k) + R_{ip}(k-1)} \right)^2 \tag{2.10}$$

where $R_{im}(k)$ is the ith resonance candidate at the kth frame for the mth state. Similarly, $R_{im}(k-1)$ is the ith resonance candidate at the $(k-1)$th frame for the mth state. φ_i is a weight, which is VTR dependent. Hence, the total cost at kth frame for $S_k = m$ is

$$D(S_k = m) = D_L(S_k = m) + \min_p \left[D_T(S_k = m|S_{k-1} = p) + D(S_{k-1} = p) \right] \tag{2.11}$$

The backtracking procedure of DP gives the best resonance frequency that means the VTR candidates are classified into VTR index and they are ready for final VTR estimation phase. On the contrary to the baseline method, the Γ parameter of DP is set for voiced and unvoiced parts differently.

2.2.3.2 VTR Estimation

Since the formant trajectories are physical quantities and that are slowly varying, they can be modeled as the output of a dynamic system model. The dynamics of the formants and their relation with observations can be approximated by the linear Gaussian model

$$x_{k+1} = Fx_k + Gw_k, \tag{2.12}$$

$$y_k = Hx_k + v_k(s), \tag{2.13}$$

where

- $x_k \in \mathbb{R}^2$ is the formant state vector defined as

$$x_k \triangleq \begin{bmatrix} f_k \\ \dot{f}_k \end{bmatrix}$$

f_k is corresponding to the resonance frequency that is estimated and $\dot{f}_k$ is it's time derivative. Its initial distribution is given as $x_1 \sim \mathcal{N}(x_1; \bar{x}, \Sigma)$. $F \in \mathbb{R}^{2\times 2}$ and $G \in \mathbb{R}^{2\times 1}$ is the transition and gain matrix respectively given as

$$F \triangleq \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, G \triangleq \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix},$$

where, T is the time difference of consecutive frames which is constant. w_k is the Gaussian process noise defined as $w_k \sim \mathcal{N}(w_k; 0, Q)$.

- $y_k \in \mathbb{R}^1$ is the observation vector. It is the best formant candidate obtained from output of the DP algorithm. It is considered to be a noisy observation of state vector x_k. $H \in \mathbb{R}^{2\times 1}$ observation matrix defined as;

$$H \triangleq \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

v_k is the Gaussian observation noise defined as $v_k \sim \mathcal{N}(v_k; 0, R_i)$.

Using above-defined model parameter, the smoothed state $\hat{x}_{k|N}$ can be estimated by a Kalman Smoother in an optimal manner [75].

2.3 Experimental Results

In this section we show experimental results and compare the performance of the proposed method with our baseline system, WaveSurfer [82] and Microsoft research (MSR) [1] methods. MSR center introduces the VTR database [1] that contains 516 sentences with hand labeled formant trajectories which is also used in our experiments. Together with this database MSR also developed some automatic tracking results. We compare our results with the one given in [1]. VTR estimation errors (in Hz) are measured by averaging absolute VTR differences between the estimated and hand labeled reference values over all frames, which is defined as

$$E_i \triangleq \frac{1}{N_c} \sum_{i=1}^{N_c} |\hat{f}_i - f_i^r|, i = 1, 2, 3 \tag{2.14}$$

where, E_i is the estimation error of ith VTR, $\hat{f}_i$ and f_i^r are the corresponding estimated and the hand labeled reference VTR's respectively and N_c is the total number of frames. The hand labeled database has 10 KHz sampling frequency. For error calculation, the hand labeled data is up-sampled so that it has the same sampling rate as the proposed system. For a more detailed examination, we measure VTR estimation error for broad phonetic classes as well. Fig. 2.4 is an example for format trajectories for voiced parts of the utterance given in Fig. 2.3. The estimated formant trajectories of this utterance can be seen in Fig. 2.5.

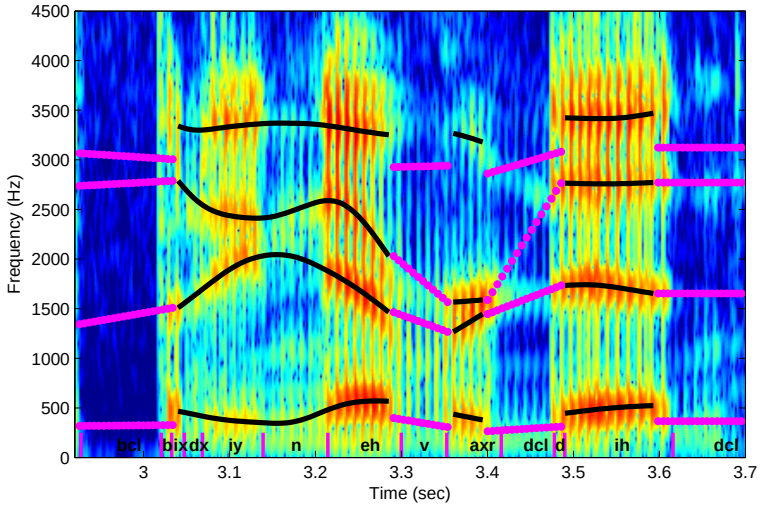


Figure 2.4: Formant (black line) and Nominal formant (magenta line) trajectories for voiced and unvoiced speech segments.

The comparison of the proposed and the baseline system is given Table 2.1.

The comparison of MSR and WaveSurfer is given in [1] (Although 538 sentences are used in [1], we use 516 of them since only 516 sentences are publicly available) and repeated here for over all evaluation of our method. The comparison of the proposed method with WaveSurfer and MSR's method [5] can be seen in Table 2.2, Table 2.3, Table 2.4, Table 2.5.

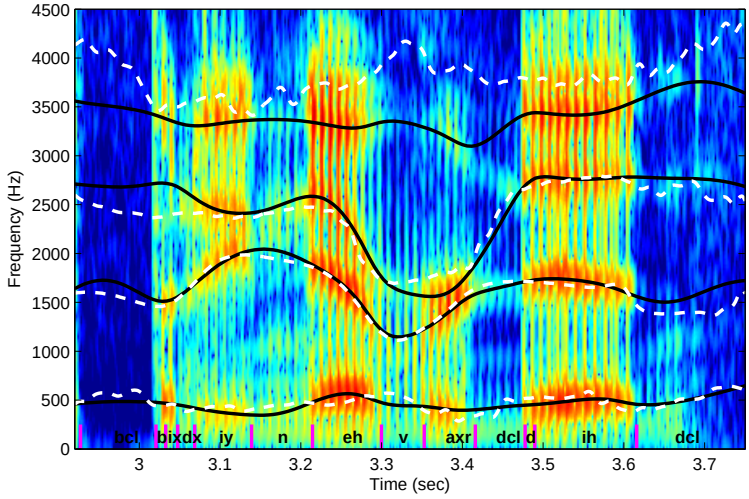


Figure 2.5: Estimated Formant trajectories for speech utterance given in Fig. 2.4. The white lines are the corresponding hand labeled formant trajectories.

Table 2.1: The error produced by the proposed and baseline methods for broad phonetic classes. (The unit of error is Hz)

Phonetic Class	Proposed			Baseline		
	E_1	E_2	E_3	E_1	E_2	E_3
Vowels	53	73	98	56	74	108
Semivowels	65	84	139	71	105	176
Nasals	93	194	156	108	236	178
Fricatives	119	126	156	224	185	227
Affricative	144	150	167	243	197	186
Stops	120	135	168	208	216	249
AVERAGE	83	105	131	122	137	169

2.4 Discussion And Conclusions

The experimental results show that the proposed method is significantly better than both the baseline system and WaveSufer. The method also has a significantly better performance compared to MSR's method [1] in vowel and semi-vowel phonetic classes where VTRs are well-defined. This result can be seen in Table 2.1, Table 2.2 and Table 2.4. On the other hand, it is comparable to the MSR's method for the remaining phonetic classes. The overall performance (for f_1, f_2 and f_3) of the proposed method is slightly better than MSR's

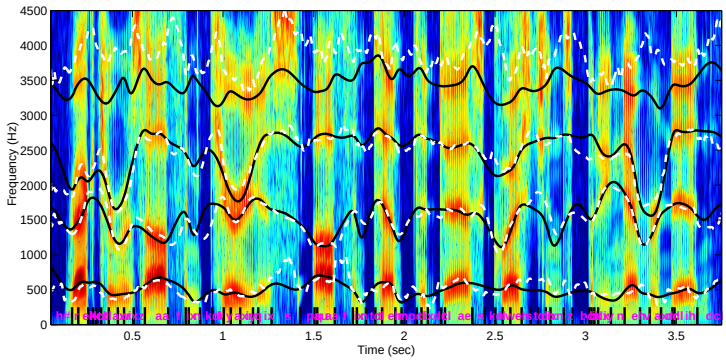


Figure 2.6: Estimated formant trajectories for a full sentence. The white lines are the corresponding to hand labeled formant trajectories.

Table 2.2: The error produced by the MSR and WaveSurfer methods for broad phonetic classes. (The unit of error is Hz)

Phonetic Class	MSR			WaveSurfer		
	E_1	E_2	E_3	E_1	E_2	E_3
Vowels	64	105	125	70	94	154
Semivowels	83	122	154	89	126	222
Nasals	67	120	112	96	229	239
Fricatives	129	108	131	209	263	439
Affricative	141	129	149	292	407	390
Stops	130	113	119	168	210	286

method, which can be seen in Table 2.5. Examination of Table-1 shows that the performance of the proposed method for nasal phonetic class is relatively low. The reason for this is that the resonance candidates of the proposed method are obtained using LPC analysis which chooses spectral peaks as VTRs. In hand labeled database, however, spectral valleys are chosen as VTRs for some nasal consonants as explained in [1].

Table 2.3: The error produced by the proposed method, MSR and WaveSurfer for all phonetic classes. (Note: MSR and WaveSurfer 's results are calculated using Table-2.2)

Method	f_1	f_2	f_3
Proposed	83	105	131
MSR	91	110	127
WaveSurfer	120	163	245

Table 2.4: The error produced by the proposed method and MSR for vowels and semivowels.

Method	Vowel	Semivowel
Proposed	74.6	96
MSR	98	119.6

Table 2.5: The error produced by the proposed method, baseline, MSR and WaveSurfer for overall average (f_1, f_2, f_3)

	Proposed	Baseline	MSR	WaveSurfer
Overall Average	106.3	164	109.3	176

CHAPTER 3

TRACKING VARIABLE NUMBER OF VOCAL TRACT RESONANCE FREQUENCIES

3.1 Introduction

The two terms, vocal tract resonances (VTRs) and formant frequencies are sometimes used interchangeably in the literature. VTR frequencies are related to physical system and defined as resonance frequencies of air path of articulatory system and they are independent of existence of air (excitation) in that articulatory system. Previously (VTR) frequencies are called formant frequencies and defined only for vowel-like sounds in the literature. However, scientists from different areas such as phonetics [48, 83] considered formant frequencies as spectral prominence and they used them as distinguishing features for some obstruent sounds such as plosives and fricatives. Therefore, nowadays, in general, formants are considered as frequencies that occur due to vocal tract resonances and are defined in acoustic domain with evidence of spectral prominence. Similarly, tracking resonance frequencies in speech utterance is handled in different perspective. Kopec [35] tracks formant frequencies only for vowel-like sounds and also labels some regions that formant frequencies do not exist. Lee et al. [38] tracks formant frequencies for speech utterances that contain unvoiced consonants but tracking performance is evaluated only for vowel-like regions. Their reason to track formant frequencies in unvoiced regions is to increase the tracking performance of vowel-like regions. Deng et al. [45, 84], extend formant frequency tracking into unvoiced regions including unvoiced closure and call it tracking VTRs. They consider formant and VTR as same for voiced regions and

assume that in non-sonorant regions VTRs are hidden (unobservable) but they should be tracked somehow to maintain the continuity of articulatory movement which is independent of the existence of air (excitation). Our perspectives are somehow different from previous studies:

- Since articulators change continuously, some resonance frequencies may have continuous trajectories (including sonorant and obstruent region) according to physical geometry of air path. However, some resonance frequencies can completely disappear or newly appear, especially when physical geometry of air path changes abruptly as in the case of nasal and plosive sounds.
- Number and frequency of a VVTR changes according to context and may even change for the same phone. As an example resonance frequencies of unvoiced closure (leakage formant) can occasionally be visible in the spectrogram. Also an unexpected resonance frequency may appear in a nasalized vowel.
- It is difficult to decide on the number of resonance frequencies that exists in a speech utterance. The generally excepted procedure of tracking 5 or 3 formant frequencies seems to give some wrong indications about VTRs.

In order to build a basis for our claim of varying number of formants, we investigate the effects of abrupt changes of the vocal tract in the state space framework in Sec. 3.2. The analysis shows that the number of formants appearing in the spectrum may vary in time. Considering all these observations we introduce the concept of tracking visible vocal tract resonance (VVTR) frequencies. VVTR includes well-known formant frequencies defined for vowel-like sounds, extra formants due to nasalization, leakage formant in obstruent regions etc, when they exist. Therefore, our tracking method is different from previous studies in the following ways:

- The number of visible resonance frequencies that we want to track depends on the utterance and is not known a priori. It may change in time for a

given speech utterance. Therefore, the tracking algorithm should have the capability of tracking different numbers of resonance frequencies along the speech utterance.

- The tracking algorithm should have the capability of initiating new trajectories and ending the already existing ones.

In this study we present a new strategy to track VVTR trajectories in a fully automatic manner without using any phonemic information. In Sec. 3.3, we will describe the tracking algorithm. Experimental results are given in Sec. 3.4 , and Sec. 3.5 is devoted to discussion and conclusions

3.2 Analysis of The Formants For Supra-Glottal Source Location

Formants, defined as the resonance frequencies of the vocal tract, are formed by the shape of the vocal tract, therefore are affected by any changes in the shape. In case of nasal sounds, almost an abrupt change occurs in the vocal tract that affects the second and the third formants observed in the spectrum. For some other phones, like fricatives and plosives, the location of the excitation changes due to heavy constriction of the vocal tract. The location of the constriction and the excitation are directly related. In this section, we analyze the effects of supra-glottal source location together with vocal tract constriction on formants. The analysis is based on concatenated tube model and the state space representation of the standard LPC filter. Our aim is to show that the disappearing formants correspond to *uncontrollable modes* of the given state space representation. The spectrogram in Fig. 3.1 is given as an example of varying number of formants during a speech utterance. The concatenated tube model, although oversimplified [15, 16] and ignores the effect of nasal cavity, is still widely used because of its tractability as well as fairly accurate modeling capability for at least some phones. A further advantage of the model is that it is based on vocal tract areas which are physical quantities. The well known concatenated tube model is shown in Fig. 3.2 In this figure ρ_k represents the kth reflection coefficient, and ρ_l and ρ_g represent the lips radiation and glottis parameters. The

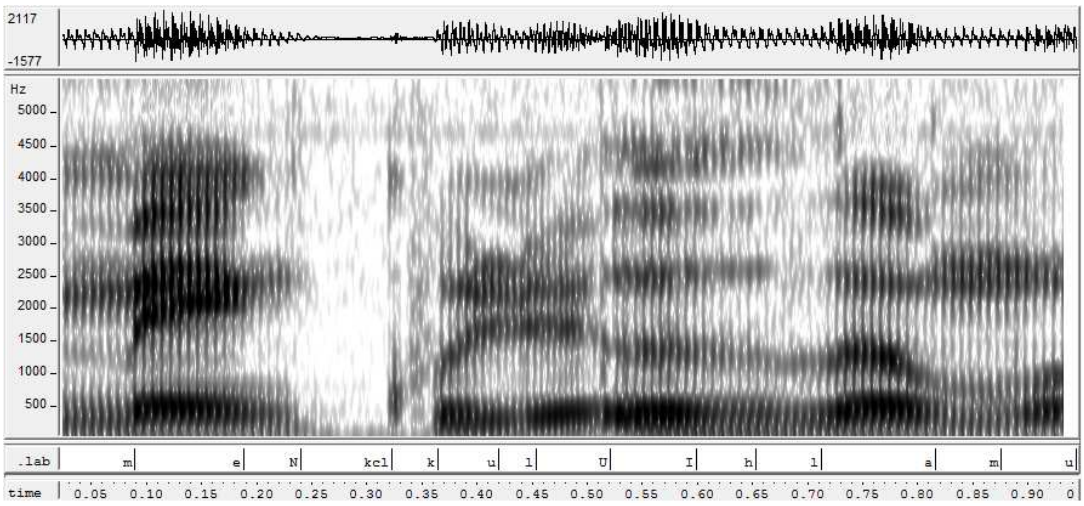


Figure 3.1: An example for the change in number of formants: The spectrogram of the utterance *menkulü ıhlamur*.

well known relationships between the reflection coefficients and the vocal tract areas are as follows.

$$\rho_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad \text{for} \quad k = 1, \ldots, n \tag{3.1}$$

An important property of this model is that it has no zeros. It is possible to write the transfer function of this system directly via the reflection coefficients. However, even for a second order system, the task is quite complicated [85]. Vocal tract resonances are usually defined as the resonance frequencies of the

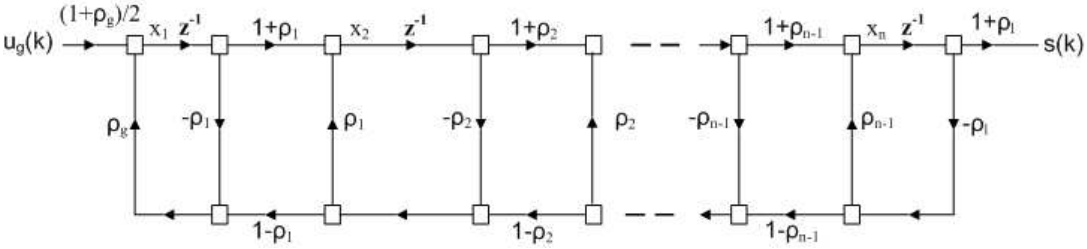


Figure 3.2: Concatenated tube model with the excitation at the glottis

system given by the above model. These frequencies are related with the poles of the transfer function of the system. In the literature, there exists some discussion regarding the *visible resonances* and/or cancelations in the transfer function by the introduction of zeros [86, 87, 46, 88]. Especially during the generation

of plosives or fricatives, vocal tract is excited from a constriction that is in the middle of the oral cavity and consequently, length of the vocal tract is reduced. Such a reduction in the length causes the generation of only high frequency resonances. In this part of the thesis, we will examine all these subjects by considering the state space representation of the system given in Fig. 3.2. The state space representation preserves the direct relationship between the parameters of the model and the reflection coefficients. In our representation, the state variables are chosen as the outputs of the adders on the forward path. Starting from the lips, i.e., from the output of the system we write the state equations as follows.

$$x_n(k+1) = -\rho_l\rho_{n-1}x_n(k) + (1+\rho_{n-1})x_{n-1}(k) \tag{3.2}$$

$$x_{n-1}(k+1) = -\rho_l\rho_{n-2}(1-\rho_{n-1})x_n(k) - \rho_{n-1}\rho_{n-2}x_{n-1}(k) + (1+\rho_{n-2})x_{n-2}(k) \tag{3.3}$$

$$\vdots$$

$$\begin{aligned} x_2(k+1) = &-\rho_l\rho_1\prod_{i=2}^{n-1}(1-\rho_i)x_n(k) - \rho_{n-1}\rho_1\prod_{i=2}^{n-2}(1-\rho_i)x_{n-1}(k) \\ &- \ldots - \rho_2\rho_1x_2(k) + (1+\rho_{n-1})x_1(k) \end{aligned} \tag{3.4}$$

$$\begin{aligned} x_1(k+1) = &-\rho_l\rho_g\prod_{i=1}^{n-1}(1-\rho_i)x_n(k) - \rho_{n-1}\rho_1\prod_{i=1}^{n-2}(1-\rho_i)x_{n-1}(k) \\ &- \ldots - \rho_{n-1}\rho_{n-2}x_2(k) + (1+\rho_{n-1})x_1(k) + \frac{1+\rho_g}{2}u(k) \end{aligned} \tag{3.5}$$

Using the above defined states and assuming that the vocal tract is excited from the glottis and the sound is produced at the lips, the state space model is written as follows.

$$x(k+1) = Ax(k) + Bu(k), \tag{3.6}$$

$$s(k) = Cx(k). \tag{3.7}$$

Where the matrices B and C are equal to

$$B = \begin{pmatrix} \frac{1+\rho_g}{2} & 0 & \ldots & 0 \end{pmatrix}^T \tag{3.8}$$

$$C = \begin{pmatrix} 0 & \ldots & 0 & (1+\rho_l) \end{pmatrix}. \tag{3.9}$$

The remaining part of the representation is given in Appendix-A. The eigenvalues of the matrix A that are close to the unit circle can be considered as

approximations of the formants. In this model, the place of the excitation is modeled by the location of the nonzero element of the matrix B. Fig. 3.3 shows the modification of the signal flow graph for supra-glotal source location. Such a modification in the signal flow graph can be reflected to the state space representation by means of the matrix B. If the constriction is at the mth tube, corresponding B matrix will be as follows.

$$B = \begin{pmatrix} 0 & \dots & \alpha_m & \dots & 0 \end{pmatrix}^T \tag{3.10}$$

An important observation here is that, a change in the position of the exci-

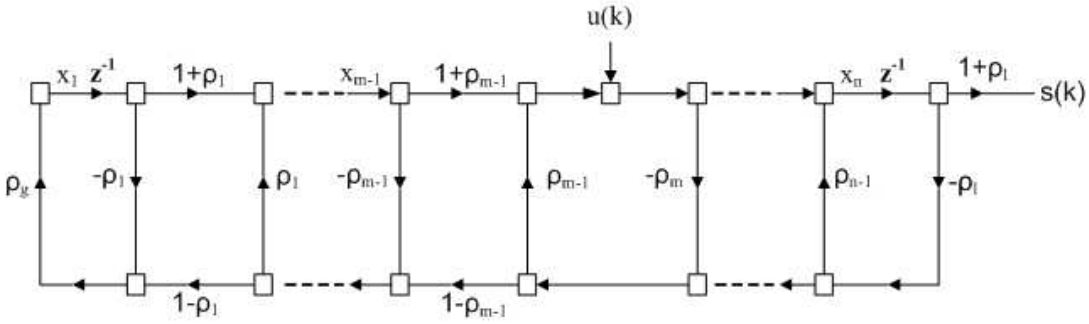


Figure 3.3: Concatenated tube model with the excitation at supra-glottis

tation introduces some zeros to the system. However, the poles are not affected. Constrictions during articulation basically changes vocal tract areas around the constriction region. Consider the case in which $A_m = 0$ for a certain m, and the corresponding reflection coefficients are $\rho_{m-1} = \frac{A_m - A_{m-1}}{A_m + A_{m-1}} = -1$ and $\rho_m = \frac{A_{m+1} - A_m}{A_{m+1} + A_m} = 1$. For this extreme case, the representation given above takes the following form.

$$x(k+1) = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ e_1 \end{pmatrix} \alpha_m u(k), \tag{3.11}$$

$$s(k) = \begin{pmatrix} 0 & e_n \end{pmatrix} x(k). \tag{3.12}$$

The sub-matrices are given in Appendix A. An interesting observation here is that A_{11} block of A is not controllable (first m states are not excited by the input) so the poles of this part can not affect the transfer function. Controllability is a concept that is borrowed from control theory that shows which modes

(eigenvalues of the A matrix) are not excited by the input of the system. These eigenvalues do not exist in the transfer function. The speech literature introduces this concept as unobservable formants that actually exist but canceled with zeros [1, 46, 45, 55]. Consider the case in which the constriction is quite close to the lips, so that $m = n - 2$. Under this condition the controllable part of the system becomes

$$A_{22} = \begin{pmatrix} -\rho_{n-2}\rho_{n-1} & -\rho_l\rho_{n-2}(1-\rho_{n-1}) \\ 1+\rho_{n-1} & -\rho_l\rho_{n-1} \end{pmatrix} = \begin{pmatrix} -\rho_{n-1} & -\rho_l(1-\rho_{n-1}) \\ 1+\rho_{n-1} & -\rho_l\rho_{n-1} \end{pmatrix}. \tag{3.13}$$

The eigenvalues of this matrix can be found as the roots of the equation $\lambda^2 + \rho_{n-1}(1+\rho_l)\lambda + \rho_l = 0$. These roots can be plotted as a function of ρ_{n-1} as $0 < \rho_{n-1} < 1$ as shown in Fig. 3.4. The figure is plotted for $\rho_l = 0.6$ and $\rho_{n-1} \geq 0$. The last reflection coefficient is clearly positive since it corresponds to the area after restriction. Fig. 3.4 indicates that there is one resonance frequency because of the second order model that we have used and its frequency is high if ρ_{n-1} is high. Furthermore, bandwidth is determined by the value of ρ_l which may not be very small. This model fits quite well with the observations related with the phone 's'. The analysis done within this section is an attempt to

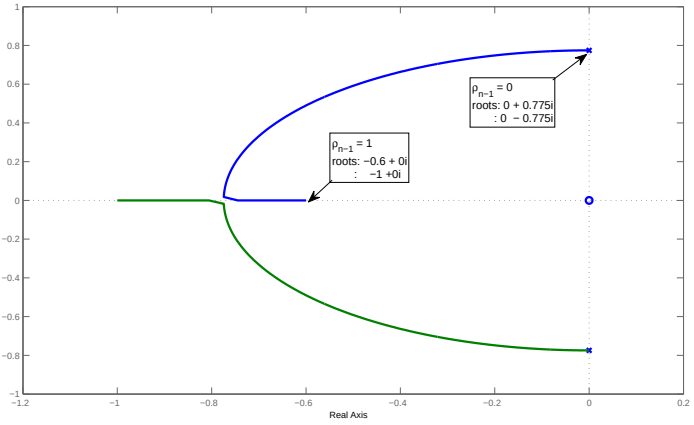


Figure 3.4: The root locus of the equation $\lambda^2 + \rho_{n-1}(1+\rho_l)\lambda + \rho_l = 0$, where $\rho_l = 0.6$ and ρ_{n-1} is varying between 0 and 1.

explain the phenomenon of appearance/disappearance of the formants in the spectrum. We conclude that the number of formants in the spectrum varies in

time. This variation can be important for fricatives and plosives.

3.3 VVTR Tracking Algorithm

Our proposed VVTR tracking algorithm is mainly based on multi-target tracking algorithms which are widely used in the tracking literature. Here we tailored the existing methods to suit VVTR tracking in speech utterance. Our VVTR tracking algorithm operates in four phases: speech analysis, track start/end decisions, gating/association and tracking, as shown in Fig.3.5. In the analysis phase, resonance candidates are obtained by finding the roots of the linear prediction polynomial obtained from LPC analysis. In the track start/end decision phase new trajectories start if there are consistent track candidates or old trajectories end if no more consistent candidates are available. In the gating and association phase, each trajectory is associated to a suitable resonance candidate if such a candidate exists in a sufficiently close neighborhood of it. Finally, in the tracking phase, the tracks are generated by Kalman filtering. Following sections explain each block in more detail.

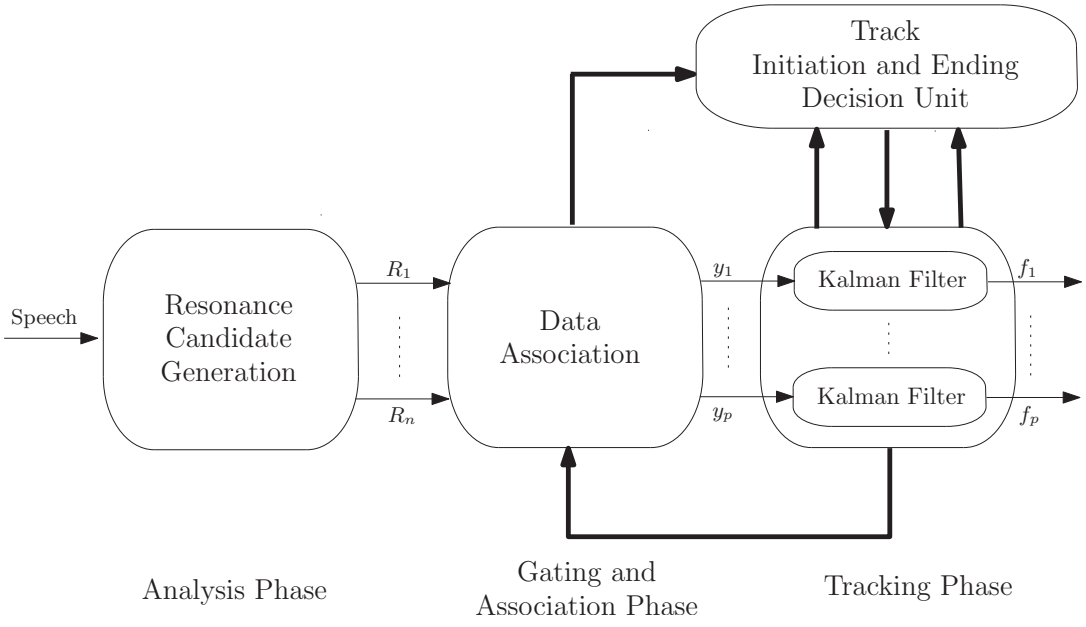


Figure 3.5: General VVTR tracking block diagram.

3.3.1 Speech Analysis Phase

Analysis phase generates candidates for resonance frequencies and their bandwidths by finding the roots of the linear prediction polynomial obtained from LPC analysis. The computaion of resonance candidates and their bandwidth is expalined in (2.7) and (2.8) in Ch.2. The sampling frequency can be chosen according to the interest of highest resonance frequency band. Bandwidth of the resonance frequencies are used as a threshold in the selection of resonance candidates. That is the roots of the LPC polynomial with large bandwidths may not be used as a resonance candidate.

3.3.2 Track Start/End Decision Phase

Track Start/End Decision phase provides the tracking algorithm with three decisions about trajectories (tracks); start a new VVTR trajectory, confirm VVTR trajectory and end an existing VVTR trajectory. The state diagram of this phase is given in Fig.3.6. State 1 is for track initiation. Any point obtained as a resonant frequency from the LPC analysis that is not in the gate of any existing track is considered as a possible future track in State 1. At the following frame the state of the candidate track will be changed either as 0 or 2. The change in the state is done according to the resonant candidates of the next frame. If there is a measurement, i.e. a candidate resonant frequency which is close to the track candidate generated at State 1 (in the gate of the track considered) than this track goes to state 2, otherwise it goes to state 0. States 0 and 2 are waiting states and waiting times are design parameters. In State 2 Kalman filters are initiated for the candidates. The trajectories which do not have measurements in their gates go to State 0 where they wait a certain time for a measurement. If they can not receive measurement in the gate of the trajectory, they go to State -1 and are deleted. However in the case of consistent measurements they go to State 2. The track candidates in State 2 go to State 100 and declared as tracks if they receive consistent measurements for a certain number of times. State 100 is the tracking state and all trajectories are tracked by the tracking algorithm if they take measurements. If they do not take any measurement in a certain

time interval track is ended.

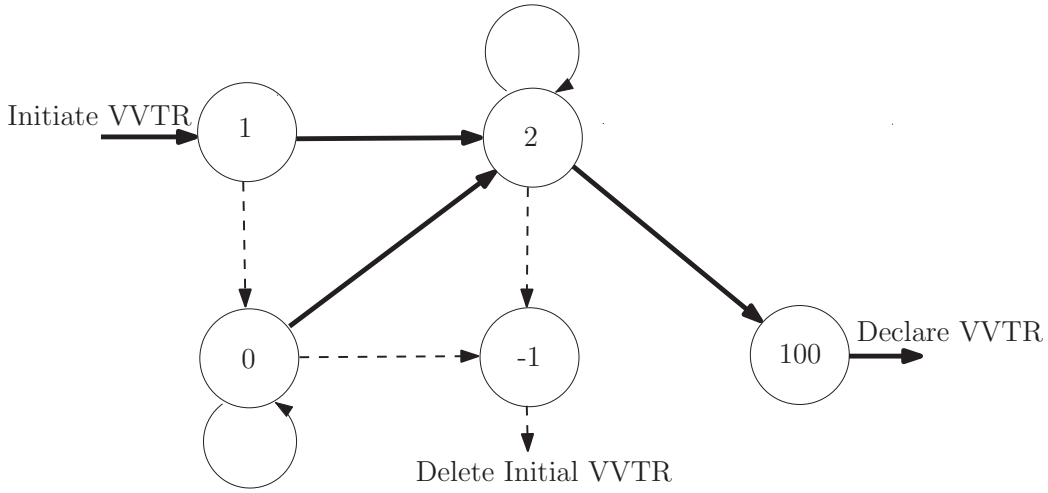


Figure 3.6: State flow diagram of the track decision phase (The solid lines denote trajectories that take consistent measurement and the dashed lines denote trajectories that do not take consistent measurement).

3.3.3 Gating and Association Phase

Gating and association phase is mainly related to Track Start/End Decision phase of the algorithm as well as the tracking phase. The gate of a trajectory determines upper and lower limits for the candidate resonance frequencies. Trajectories consider a measurement as consistent if it is in between these upper and lower limits. The VVTR tracking algorithm use two types of gates: constant and variable. The upper and lower frequency values of a constant gate are constant. As an example, if we define constant gate value as 150 Hz and current trajectory value is 500 Hz than the trajectory considers resonance candidates as consistent, if they are in between $500 - 150 = 350Hz$ and $500 + 150 = 650Hz$. The constant gate is used for those trajectories whose states are at State 1, State 0 or State 2. The variable gate is used for the trajectories in State 100, i.e. after VVTR is declared. The variable gate G changes by time and it is defined as:

$$G(k) = \sqrt{\gamma S_k} \tag{3.14}$$

where, S_k and γ are measurement variance and a predefined threshold respectively. The upper and lower limits L of gate are defined as follows.

$$L(k) = \hat{y}_k \pm \sqrt{\gamma S_k} \tag{3.15}$$

In this expression $\hat{y}_k$ is the predicted measurement at time instant k, which is obtained using prediction step of the Kalman filter. This expression indicates that the gate value is changed according to the measurement covariance. As the theory suggests, the gate value increases when gate contains no measuramet at a certain time. If more than one candidate falls in the gate of any trajectory, the nearest-neighbor procedure is applied to associate the candidate with the trajectory.

3.3.4 Tracking Phase

In the tracking phase Kalman filter is used to track resonance frequencies. The state-space representation of the dynamic system model is given as in Chapter 2 and is repeated here for convenience.

$$x_{k+1} = Fx_k + Gw_k, \tag{3.16}$$

$$y_k = Hx_k + v_k(s), \tag{3.17}$$

where

- $x_k \in \mathbb{R}^2$ is the formant state vector defined as

$$x_k \triangleq \begin{bmatrix} f_k \\ \dot{f}_k \end{bmatrix}$$

  f_k is corresponding to the resonance frequency that is estimated and $\dot{f}_k$ is it's time derivative. Its initial distribution is given as $x_1 \sim \mathcal{N}(x_1; \bar{x}, \Sigma)$. $F \in \mathbb{R}^{2\times 2}$ and $G \in \mathbb{R}^{2\times 1}$ are the transition and gain matrix respectively given as

$$F \triangleq \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, G \triangleq \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix},$$

  where, T is the time difference of consecutive frames which is constant. w_k is white Gaussian process noise $w_k \sim \mathcal{N}(w_k; 0, Q)$.

- $y_k \in \mathbb{R}^1$ is the observation vector. It is the best formant candidate obtained from the output of the DP algorithm. It is considered to be a noisy observation of state vector x_k. $H \in \mathbb{R}^{2\times 1}$ observation matrix defined as;

$$H \triangleq \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

  v_k is the Gaussian observation noise defined as $v_k \sim \mathcal{N}(v_k; 0, R)$.

Tracking algorithm uses standard Kalman filtering equations [51] where the measurements are the associated frequencies obtained at the association phase. First component of the filtered states give the related VVTR track.

3.4 Experimental Results

In this section, we will give spectrograms and VVTR frequency tracker outputs of some utterances that are taken from Continuous Speech Turkish Database [89]. Phonetic transcriptions (with SAMPA alphabet of Turkish [90]) of the fragments are given in the corresponding figures. We use WaveSurfer speech tool [82] to show results of VVTR tracker. We also use formant tracking results of WaveSurfer for comparison purposes. The first spectrogram is for the Turkish word *fakat*. The spectrogram of the *fakat* is given in Fig. 3.7. In this figure, the indicated regions denote leakages of resonance frequencies that appear in the closure before the plosives *k* and *t*. The solid lines are results of proposed VVTR tracker. It is clear that algorithm is successful in tracking and ending leakage resonance frequency tracks as well as the normal resonance frequencies of vowels. Fig. 3.8 shows spectrogram of a fragment *O hantaldı*. This speech utterance contains special phenomena about vowel nasalization which is on *a* that appears before the nasal sound *n*. Since the proposed VVTR tracker is capable of tracking different number of resonance frequencies it tracks extra resonance frequency due to nasalization as well the resonances of *n* as shown in Fig. 3.8. Also, the leakage resonance frequencies in the closure part of *d* are successfully tracked and these tracks are connected to resonance frequencies of the next vowel *1*(SAMPA representation of Turkish 'ı'). Note that tracking the leakages generates the required continuity of the vocal tract resonances when

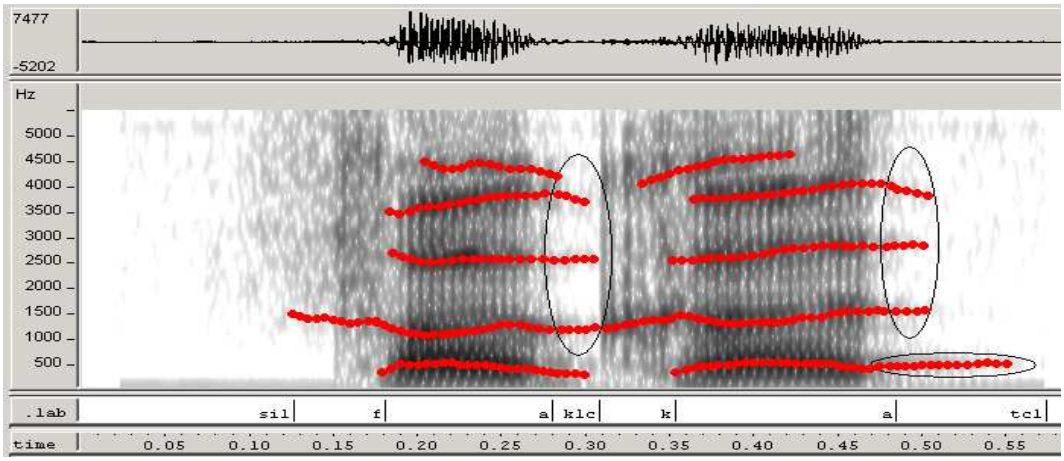


Figure 3.7: VVTR tracking result of the word *fakat* with transcription *(f-a-kcl-k-a-tcl-t)*.

such continuity exists. On the other hand it is clear that whenever the vocal tract shape change, this change causes an abrupt change in its resonance frequency, tracks are ended and new ones are started as explained above. Fig.

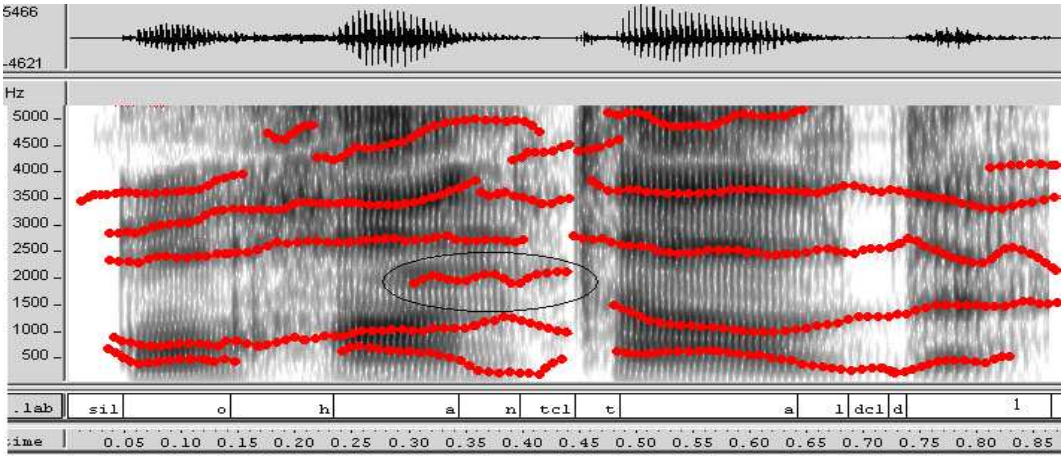


Figure 3.8: VVTR tracking result of the fragment *O hantaldı* with transcription *(o- 5-a-n-tcl-t-a-l-dcl-d-1)*.

3.9 shows VVTR tracker results of the statement *sürücügillerden*. This sound also contains other special phenomena which is called *velar pinch*, that occurs at velar consonant (k, g, N). At these phones two resonance frequencies merge together. The encircled region in the figure shows that the VVTR algorithm is also successful in detecting and tracking this phenomenon. In Fig. 3.10,

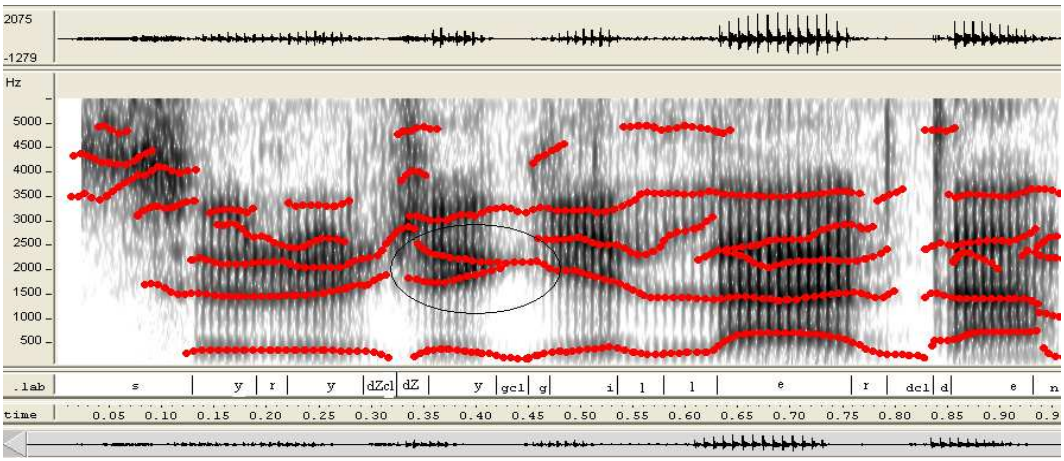
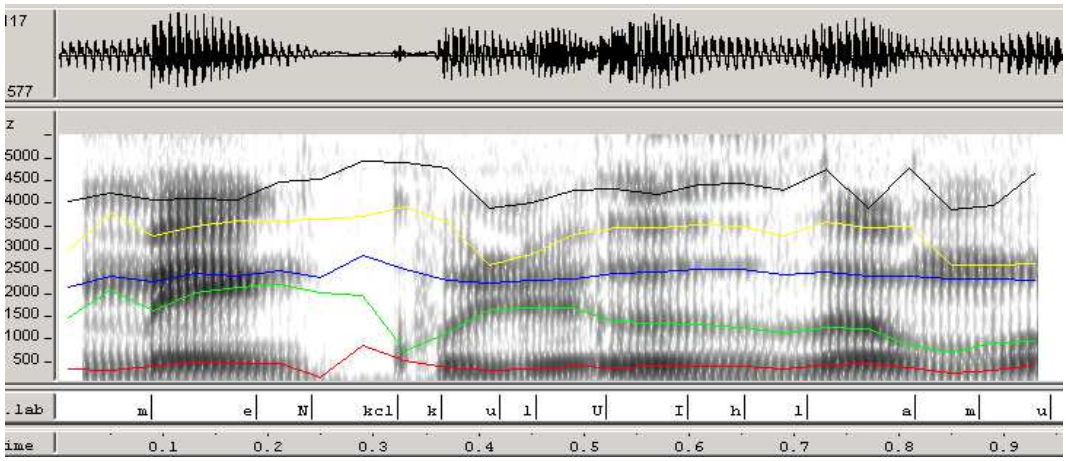


Figure 3.9: VVTR tracking result of the word *sürücügillerden* with transcription *(s-y-r-y-dZcl-dZ-y-gcl-g-i-l-l-e-r-dcl-d-e-n).*

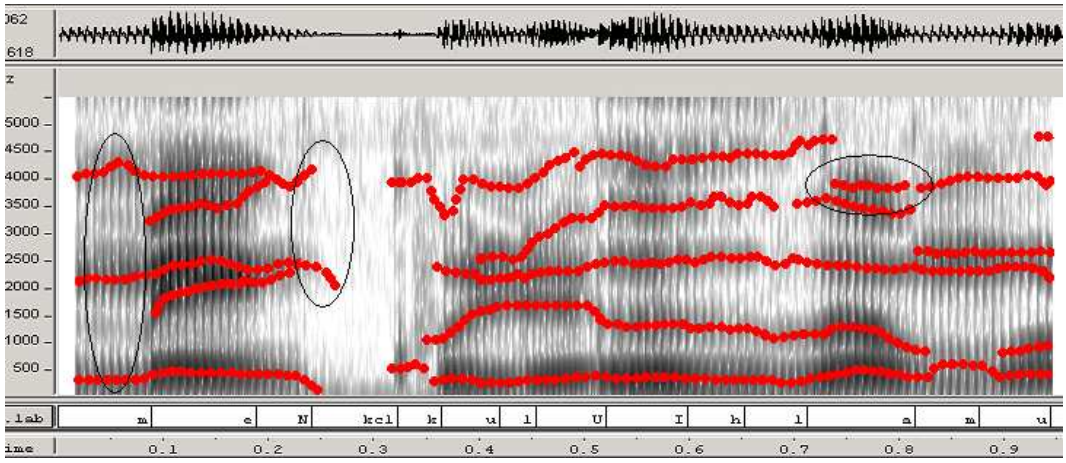
the tracking results of the proposed VVTR tracker and WaveSurfer speech tool are compared for the same fragment *menkulü ıhlamur*. WaveSurfer has fixed number of formants which is set to 5. A close examination of this utterance shows that the first nasal m, shown as first encircled region in the figure, has 3 VVTR tracks however the number increases to 5 at the vowel region of e. It can be seen that the VVTR tracks are again reduced to 3 corresponding to *velar pinch* before N which denotes phonetic velar n as the second encircled region indicates. The last marked region of in Fig. 3.10 shows the difference between the WaveSurfer tracks and our tracker for the last a. Our tracking system gives a much more satisfactory result for the 5th VVTR compared to WaveSurfer.

3.5 Discussion and Conclussions

In this study we have analyzed the vocal tract resonances and their visibility in the spectrogram. In the literature it is usually assumed that the vocal tract resonances are continuous due to the continuity of the movements of the articulators, however this continuity is not visible in the spectrogram of the speech. In our analysis we observed that vocal tract resonances may not be continuous especially at the nasals, but not restricted only to them, where vocal tract changes structurally. Based on our observations of spectrograms of a large num-



(a)



(b)

Figure 3.10: WaveSurfer and VVTR tracking result of fragment *menkulü ıhlamur (m-e-n-kcl-k-u-l-y- 1-5-l-a-m-u)* in part-a, part-b respectively

ber of sentences we proposed a tracking algorithm that can start, end and merge tracks of vocal tract resonances. The tracking algorithm is based on Kalman filtering and is inspired by multiple target tracking methods. Visibility of the vocal tract resonances in the spectrogram is another issue of this study. In the examples we showed that at some regions where vocal tract resonances exist due to *leakage*, continuity of the tracks is maintained by the algorithm. The experimental results are very promising and show that the approach given here can be used in many speech applications such as speech synthesis, recognition, segmentation etc. The VVTR frequencies can be considered as *anchor formant* frequencies hence the result also can be used for classical formant frequency tracking applications such as given by Xia et al.[37]

CHAPTER 4

AUDIOVISUAL-TO-ARTICULATORY INVERSION BASED ON GAUSSIAN MIXTURE MODEL REGRESSION

4.1 Introduction

In his book *The Acoustic Theory of Speech Production*, Gunnar Fant published a series of nomograms: plots of formant frequency as a function of articulatory constriction location [2]. His nomograms captured the imagination of the scientific world by depicting the relationship between articulation, $\mathcal{X}$, and acoustics, $\mathcal{Z}$ in the form of a mathematical function, $\mathcal{Z} = f(\mathcal{X})$. With Fant's nomograms in hand, the problems of speech technology seemed solvable: speech synthesis seemed to be no harder than implementation of the function $\mathcal{Z} = f(\mathcal{X})$, and speech recognition, no harder than implementation of the inverse function, $\mathcal{X} = g(\mathcal{Z})$. Although the nomogram is no longer considered to be a summary of all speech technologies, estimation of the function $\mathcal{X} = g(\mathcal{Z})$ (the so-called *articulatory inverse* function) is still a goal of speech technology research, partly because it is an interesting challenge, but also, in part, because recent results continue to demonstrate that the accuracy of automatic speech-to-text transcription using both acoustic and articulatory measurements is higher than the accuracy of transcription using only acoustics [91, 92, 93, 94].

The science of articulatory inversion has become far more precise than ever before, since the development of electromagnetic articulography (EMA) made it relatively easy to record simultaneous articulatory and acoustic measure-

ments [62], and especially since the publication of a relatively large single-talker EMA database as part of the MOCHA database [63]. Using EMA data, recent studies have demonstrated acoustic-to-articulatory inversion using hidden Markov models (HMMs) [68], neural networks [64, 65], and Gaussian mixture models (GMM) [67], using a wide variety of acoustic feature vectors [66]. The publication of visual features for the female speaker of the MOCHA database [95] has allowed experiments to test audiovisual-to-articulatory inversion [72]. This part of the thesis develops an articulatory inverse function based on GMM, using audio, video, and audiovisual input features. The first of the goals of this study is to simply extend previous research, by performing experimental tests of a wider variety of acoustic features and audiovisual fusion strategies than in previous work. In particular, we propose to step back to the days of Gunnar Fant, as it were, by including explicit formant frequency and formant energy estimates in the acoustic feature vector for purposes of articulatory inversion.

Although it demonstrated the potential utility of articulatory inversion, the nomogram also demonstrated that articulatory inversion is, strictly speaking, an ill-posed problem: the function $\mathcal{Z} = f(\mathcal{X})$ is non-monotonic, therefore it has no inverse. Schroeder [56] and Mermelstein [57] demonstrated that there is a one-to-one map between the shape of a tube, on one hand, and the set of pole and zero frequencies of its driving point impedance, on the other; since the formant frequencies carry information only about the poles of the driving-point impedance, they concluded, the problem of acoustic-to-articulatory inversion is underspecified by a factor of two. They proposed two potential solutions to the problem: (1) measure both the poles and zeros of the driving-point impedance, e.g., using external stimulation, or (2) halve the degrees of freedom of the problem, by assuming any particular distribution of losses within the vocal tract. Wakita [58] demonstrated a special case of the latter solution: he demonstrated that if all losses are at the lips, then the vocal tract shape is provided by the reflection-line implementation of linear predictive coding (LPC). Atal, Chang, Mathews and Tukey [59] proposed a third solution: dynamic programming. They proposed that the formant frequency vector at each time step should be inverted in one-to-many fashion, to find a list of matching articulatory vectors; dynamic

programming is then used to find the most likely temporal sequence of articulations matching the observed acoustics. Recent papers have replaced Atal's dynamic programming algorithm with an HMM [68] or low-pass filter [67, 64].

The second goal of this study is to demonstrate a better smoother for GMM-based articulatory inversion. In particular, we demonstrate that Kalman smoothing dramatically improves articulatory inversion. Furthermore, because the Kalman smoother is based on a Bayesian-normalized generative model for the observed articulatory and acoustic data, we demonstrate that the Kalman smoother can be easily adapted to make use of any available auxiliary information, e.g., of a phonetic transcription, in order to further reduce the error rate of the articulatory inverse.

Finally the third main goal of our work is to give the details of efficient and effective information fusion procedures that can be applied to combine acoustic and visual information. Although the description given above summarizes the highlights of our work, we would like to give below a list of major contributions of this chapter:

- Incorporation of various type of acoustic (and visual) features into the articulatory inversion process is shown to be effective.
- In contrast to the currently used (low-pass filter based post-processing) smoothers, we propose dynamic (model-based) smoothing in order to make use of as much information from the GMM inversion (both the means and covariances) as possible. We also present a way to include the extra information such as phonetic transcription (if any) into the smoothing framework effectively.
- In addition to the so-called early and late fusion types utilized in the literature, we propose a novel audio-visual fusion methodology (with we call "modified fusion algorithm") that is shown to give superior performance.

The rest of the chapter is structured as follows. We start in Sec. 4.2 by describing the set of audio and video features considered and emphasize the advantage of using various combinations of audio features in articulatory inversion. Sec. 4.3,

for the sake of completeness, describes GMM-based articulatory inversion, more or less exactly as it was described in [67]. The dynamic smoothing process based on global and multi Kalman smoothers that we propose is detailed in Sec. 4.4. Sec. 4.5 describes three different candidate methods for audiovisual fusion. The extensive experimental results that we performed are given in Sec. 4.6 while Sec. 4.6.3 lists and describes the results. Conclusions are drawn in Sec. 4.7.

4.2 Acoustic and Visual Features

Early studies of articulatory inversion (e.g. [57, 58]) often focused on the relative utility of different types of acoustic features; Qin and Carreira-Perpin, among others, have compared different acoustic feature sets in a modern probabilistic paradigm [66]. Video features have also been demonstrated to be useful for articulatory inversion [72]. This part of the thesis examines the extraction of different types of acoustic and visual features for GMM based articulatory inversion.

4.2.1 Audio Features

The papers of Fant [2], Mermelstein [57] and others suggest that good estimates of the formant frequencies may go a long way toward accurate articulatory inversion. Recent studies have approximated the formant frequencies using information from linear predictive coding (LPC), including the line spectral frequencies (LSFs). Qin and Carreira-Perpin [66], in particular, suggested that acoustic features related to the formants (LPC and LSF) are more useful for articulatory inversion than features not related to the formants. One of the goals of this work is to demonstrate that better formant estimates produce better articulatory estimates. This study will compare several of the same acoustic feature sets that were tested in [66], including mel-frequency cepstral coefficients MFCC, LPC, LSF. This study will also test two additional feature sets: a vector of formant related features (frequencies and energies) and log area ratios (LAR) of the vocal-tract tube model.

MFCCs have often been reported to give optimal results in acoustic-to-articulatory

inversion experiments [67]. In this study, 13 MFCCs are generated from 29 triangular band-pass filters, uniformly spaced on a Mel-frequency scale between 0 Hz and 8000 Hz. $M = [M_1, \ldots, M_{13}]$ denotes the 13-dimensional MFCC feature vector; $DM = [M, M_\Delta, M_{\Delta\Delta}]$ denotes a combination of the MFCCs and their velocity M_Δ, and acceleration $M_{\Delta\Delta}$ components.

The LPC coefficients are adequate for articulatory inversion during vowels and glides, if the vocal tract is assumed to be a lossless tube with a lossy termination [58]; even under more realistic assumptions, the LPCs are a useful description of the vowel spectrum [66]. LPCs are estimated by the autocorrelation method and the order of the LPC filter is chosen to be 18. $L = [L_1, \ldots, L_{18}]$ denotes an 18-dimensional LPC feature vector; $DL = [L, L_\Delta, L_{\Delta\Delta}]$ denotes the combination of LPCs and their velocity L_Δ and acceleration $L_{\Delta\Delta}$ components.

Log area ratio (LAR) coefficients are derived from LPC. In LAR analysis, the vocal tract is modeled by cascading uniform tubes (Wakita's model [58]). It is assumed that the first tube (the glottis) is closed (area = 0), and the last tube (just past the lips) has infinite area. The LAR coefficients are formed by the log area ratio of cross-section areas of consecutive tubes. Let g_i be the ith LAR coefficient; it is calculated as

$$A_i = \log \frac{G_i}{G_{i+1}} = \log \frac{1 - \rho_i}{1 + \rho_i}$$

where G_i and G_{i+1} are the ith and $(i+1)$th cross-sectional area respectively, and ρ_i is the corresponding partial correlation coefficient derived from the Levinson-Durbin recursion. LAR features are denoted $A = [A_1, \ldots, A_{18}]$; DA is the combined feature set, including velocity and acceleration.

The LSFs are the poles and zeros of the driving point impedance of Wakita's vocal tract model [96]. The LSFs are widely used in speech coding, because of their high intra-vector and inter-vector predictability [97]. LSF coefficients tend to cluster around spectral peaks, especially around formant frequencies, and are therefore closely related to articulation [98]. LSF features are denoted $S = [S_1, \ldots, S_{18}]$, and DS is the combined feature set, including velocity and acceleration.

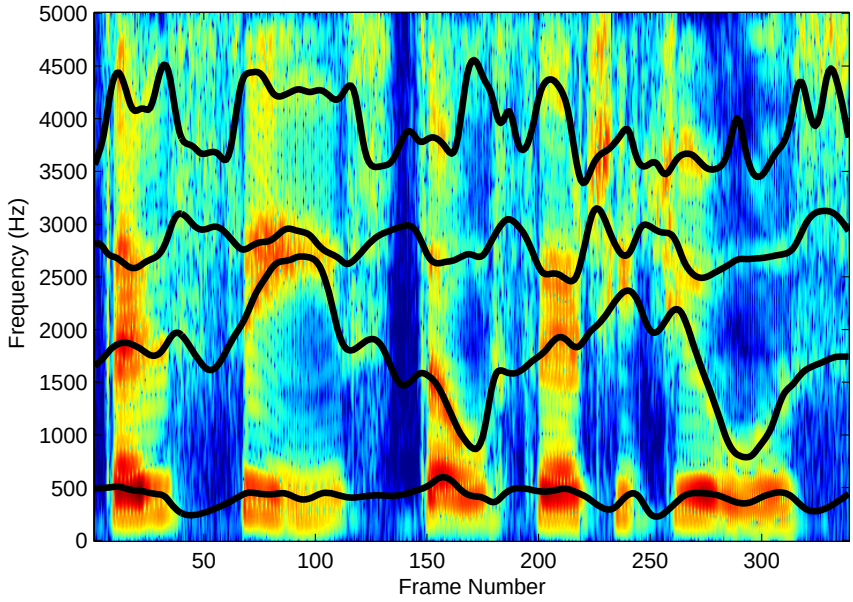


Figure 4.1: The spectrogram of the utterance 'Those thieves stole thirty jewels' from fsew0-Mocha-TIMIT database. Estimated formant trajectories are superimposed.

Formant frequencies are another type of acoustic features extracted from LPC features. Formants are the resonant frequencies of the vocal tract and they change according to movement of the articulators [99]. Therefore, this chapter considers the formant frequencies as effective acoustic features to be utilized for articulatory inversion. During vowels and glides, formant frequencies may be estimated by the poles of an autoregressive spectral estimator, though temporal smoothing improves the estimate. During obstruent consonants, formant frequencies must be interpolated using some type of dynamic programming model. In this study we use the formant tracker described in Chapter 2. The output of the formant tracking algorithm for one of the sentences from the MOCHA database is shown in Fig. 4.1.

The energy associated with each formant frequency is also extracted, using the algorithm described in [73]. First, a magnitude spectrum is computed for each frame. Second, for each formant, Gaussian windows are generated in the spectrum domain. The mean and variance of each Gaussian are related to the

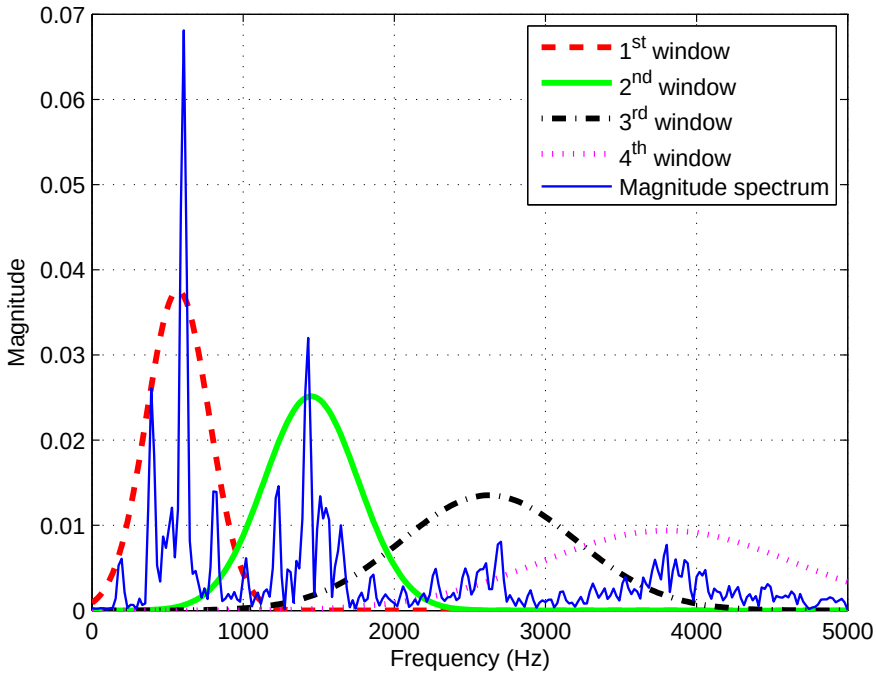


Figure 4.2: Magnitude spectrum and Gaussian windows for 155'th frame of Fig.4.1 Corresponding four formant frequencies are F=[586, 1457, 2628, 3803] Hz.

associated formant frequency and bandwidth respectively. The bandwidths of the first four formants are assumed to be fixed at the values of BW= [90, 110, 170, 220] Hz. The means of the Gaussian windows vary in time, tracking the estimated formant frequencies. Third, the energy level associated with the ith formant, E_i, is computed by multiplying the magnitude spectrum $|X(f)|$ by the ith Gaussian window, $\mathcal{W}_i(f)$, and summing over all frequencies:

$$E_i = \ln\left(\sum_{f=0}^{F_s/2} \mathcal{W}_i(f)|X(f)|\right) \tag{4.1}$$

Fig. 4.2 shows the magnitude spectrum and corresponding Gaussian windows for the 155'th frame of the spectrogram given in Fig. 4.1.

In this study, $FE = [F_1, \ldots, F_4, E_1, \ldots, E_4]$ denotes an 8-dimensional vector containing the frequencies and energies of the first four formants. $DFE = [FE, FE_\Delta, FE_{\Delta\Delta}]$ denotes the combination of FE and their velocity FE_Δ, and acceleration $FE_{\Delta\Delta}$ components.

4.2.2 Visual Features

In addition to acoustic features, visual features are also useful in articulatory inversion. Lips, jaw, teeth, and some part of tongue are visible articulators and visual information from these articulators can be extracted from a camera. Active Appearance Models (AAMs) are used to extract the visual features from the speaker face [72]. The visual features used in this chapter are AAM-based and consist of 12 shape and 27 texture features. We obtained these features from [95]. In this study, $V = [V_1, \ldots, V_{39}]$ denotes the 39-dimensional visual feature vector, and $DV = [V, V_\Delta, V_{\Delta\Delta}]$ denotes the combination of visual features and their velocity V_Δ, and acceleration $V_{\Delta\Delta}$ components.

4.3 Gaussian Mixture Model Based Articulatory Inversion

One of the well known acoustic to articulatory mapping methods is nonlinear regression based on GMMs [67]. The basic idea of the method is as follows. Let $\mathcal{Z}$, $\mathcal{X}$ denote vectors from acoustic (and/or visual) and articulatory spaces. Articulatory inversion methods look for an inverse mapping $\mathbf{g}(\cdot)$ defined as:

$$\mathcal{X} = \mathbf{g}(\mathcal{Z})$$

to estimate articulatory vectors from given acoustic and visual data. The mapping between the articulatory and the acoustic (and/or visual) spaces is quite nonlinear and analytically unknown. Therefore, it is not a trivial problem to estimate an inverse mapping directly. In a probabilistic framework, the inverse mapping function $\mathbf{g}(\cdot)$ can be approximated if enough data pairs ($\mathcal{Z}_i$, $\mathcal{X}_i$) are available. Let $\hat{\mathbf{g}}(\cdot)$ be an estimate of the true inverse mapping $\mathbf{g}(\cdot)$ defined as:

$$\hat{\mathcal{X}} \triangleq \hat{\mathbf{g}}(\mathcal{Z})$$

In the minimum mean square error (MMSE) sense, the optimal approximate mapping $\hat{\mathbf{g}}_{MMSE}(\cdot)$ can be estimated by minimizing the error variance $\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z}))$.

$$\hat{\mathbf{g}}_{MMSE}(\mathcal{Z}) = \arg\min_{\hat{g}(\cdot)} \mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z})) \tag{4.2}$$

where

$$\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z})) \triangleq \mathrm{E}\left[(\mathcal{X} - \hat{\mathbf{g}}(\mathcal{Z}))^T(\mathcal{X} - \hat{\mathbf{g}}(\mathcal{Z}))\right].$$

Taking the derivatives of $\mathcal{J}(\hat{\mathbf{g}}(\mathcal{Z}))$ with respect to $\hat{\mathbf{g}}(\mathcal{Z})$ and equating to zero, the approximate mapping can be found as;

$$\hat{x}_{MMSE} \triangleq \hat{\mathbf{g}}_{MMSE}(z) = \mathrm{E}(\mathcal{X}|\mathcal{Z} = z) \tag{4.3}$$

In other words, the best thing that we can do in the MMSE sense is to represent the inverse mapping $\mathbf{g}(\cdot)$ with the conditional expectation of articulatory data given audiovisual data. In order to find a mathematically tractable approximate mapping we need further assumptions. If it is assumed that $\mathcal{X}$, $\mathcal{Z}$ are jointly distributed according to a Gaussian mixture model (GMM), then the joint distribution can be written as

$$f_{\mathcal{X},\mathcal{Z}}(x,z) \triangleq \sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(x,z;\mu^i,\Sigma^i) \tag{4.4}$$

where $\mathcal{K}$ is the number of mixture components and π_i are the mixture weights, which satisfy $\sum_{i=1}^{\mathcal{K}} \pi_i = 1$. The μ^i and Σ^i denote the mean and the covariance of the GMM components and are defined as

$$\mu^i \triangleq \begin{bmatrix} \mu^i_{\mathcal{X}} \\ \mu^i_{\mathcal{Z}} \end{bmatrix}, \Sigma^i \triangleq \begin{bmatrix} \Sigma^i_{\mathcal{X}\mathcal{X}} & \Sigma^i_{\mathcal{X}\mathcal{Z}} \\ \Sigma^i_{\mathcal{Z}\mathcal{X}} & \Sigma^i_{\mathcal{Z}\mathcal{Z}} \end{bmatrix}.$$

The conditional distribution $f_{\mathcal{X}|\mathcal{Z}}(x|z)$ can be calculated from joint distribution $f_{\mathcal{X},\mathcal{Z}}(x,z)$ using the Bayesian rule

$$f_{\mathcal{X}|\mathcal{Z}}(x|z) \triangleq \frac{f_{\mathcal{X},\mathcal{Z}}(x,z)}{\int f_{\mathcal{X},\mathcal{Z}}(x,z)\,dx}$$

The resulting conditional distribution is again a GMM, given as

$$f_{\mathcal{X}|\mathcal{Z}}(x|z) \triangleq \sum_{i=1}^{\mathcal{K}} \beta^i(z) \mathcal{N}(x;\mu^i_{\mathcal{X}|\mathcal{Z}},\Sigma^i_{\mathcal{X}|\mathcal{Z}}) \tag{4.5}$$

where $\beta^i(z) \triangleq \mathcal{P}(c = i|\mathcal{Z})$ is the posterior probability of the ith mixture component, where $c \in (1,...,\mathcal{K})$ is the mixture indicator. The parameter $\mu^i_{\mathcal{X}|\mathcal{Z}}$ and $\Sigma^i_{\mathcal{X}|\mathcal{Z}}$ are the conditional mean of and covariance of $\mathcal{X}$ given mixture indicator $c = i$ and observation $\mathcal{Z}$, which are defined as

$$\begin{aligned} \mu^i_{\mathcal{X}|\mathcal{Z}} &= \mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}(z - \mu^i_{\mathcal{Z}}) \\ \Sigma^i_{\mathcal{X}|\mathcal{Z}} &= \Sigma^i_{\mathcal{X}\mathcal{X}} - \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}\Sigma^i_{\mathcal{Z}\mathcal{X}} \\ \beta^i(z) &= \frac{\pi_i \mathcal{N}(z;\mu^i_{\mathcal{Z}},\Sigma^i_{\mathcal{Z}\mathcal{Z}})}{\sum_{i=1}^{\mathcal{K}} \pi_i \mathcal{N}(z;\mu^i_{\mathcal{Z}},\Sigma^i_{\mathcal{Z}\mathcal{Z}})} \end{aligned} \tag{4.6}$$

Under these assumptions, the MMSE articulatory estimate $\hat{x}$ and corresponding error covariance Σ can be found as follows.

$$\begin{aligned}\hat{x} &\triangleq \mathrm{E}(\mathcal{X}|\mathcal{Z}=z) \\ &= \sum_{i=1}^{\mathcal{K}} \beta^i(z)\left[\mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}(z-\mu^i_{\mathcal{Z}})\right] \qquad (4.7)\\ \Sigma &\triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}=z) \\ &= \sum_{i=1}^{\mathcal{K}} \beta^i(z)\left[\Sigma^i_{\mathcal{X}|\mathcal{Z}} + (\mu^i_{\mathcal{X}|\mathcal{Z}}-\hat{x})(\mu^i_{\mathcal{X}|\mathcal{Z}}-\hat{x})^T\right] \qquad (4.8)\end{aligned}$$

The parameter set of the GMM, $\Theta = \{\pi_i, \mu^i, \Sigma^i\}_{i=1}^{\mathcal{K}}$, may be estimated using the expectation maximization (EM) algorithm, as described in [100].

4.4 Dynamic Smoothing of Articulatory Trajectories

Although the movements of the physical articulatory organs are slowly varying and quite smooth, the output of the various articulatory inversion methods given in the literature are not smooth enough to obtain good performance. Therefore, it is necessary to use further smoothing techniques to improve their estimates. For this purpose time derivative features like velocity and acceleration components are used in [68], and it is observed that these auxiliary features improve the smoothness of the articulatory trajectories. However, the trajectories obtained by this method are not sufficiently smooth. Zero-phase low-pass filters (FIR or IIR) can also be used to smooth output of the articulatory estimators [64, 67, 69]. Zero-phase filtering can be performed by first filtering forward in time, then filtering backward (in time) using the same filter. For each articulator, the cut-off frequencies of the low-pass filters are optimized by trial and error. Although being simple and fairly effective, this type of methodology cannot use the statistical output quantities like GMM calculated covariances. Furthermore, the inclusion of extra information such as phoneme boundaries into the smoothing process is quite difficult if not impossible. In order to address these drawbacks, this section of the thesis proposes a general statistical smoothing method which is based on Kalman smoothing, and which could, in principle, be applied to any articulatory inversion method given in the literature such as HMM ([72, 68]),

SVM ([101, 70]), or TMDN [65].

Since the articulatory trajectories are physical quantities and their motion is slowly varying, they can be modeled as the output of a dynamic system model. The dynamics of the articulators and their relation with the observed GMM articulatory estimates, $\hat{g}_{MMSE}(z)$, can be approximated by a linear Gauss-Markov model, for which the optimal Kalman smoother can be formulated. The Kalman smoother can be written as a Bayesian-normalized generative model, therefore it is relatively straightforward to integrate auxiliary information that will help to improve its inference power; as an example, this study demonstrates the use of a phonetic transcription as auxiliary information to improve the performance of the articulatory inversion.

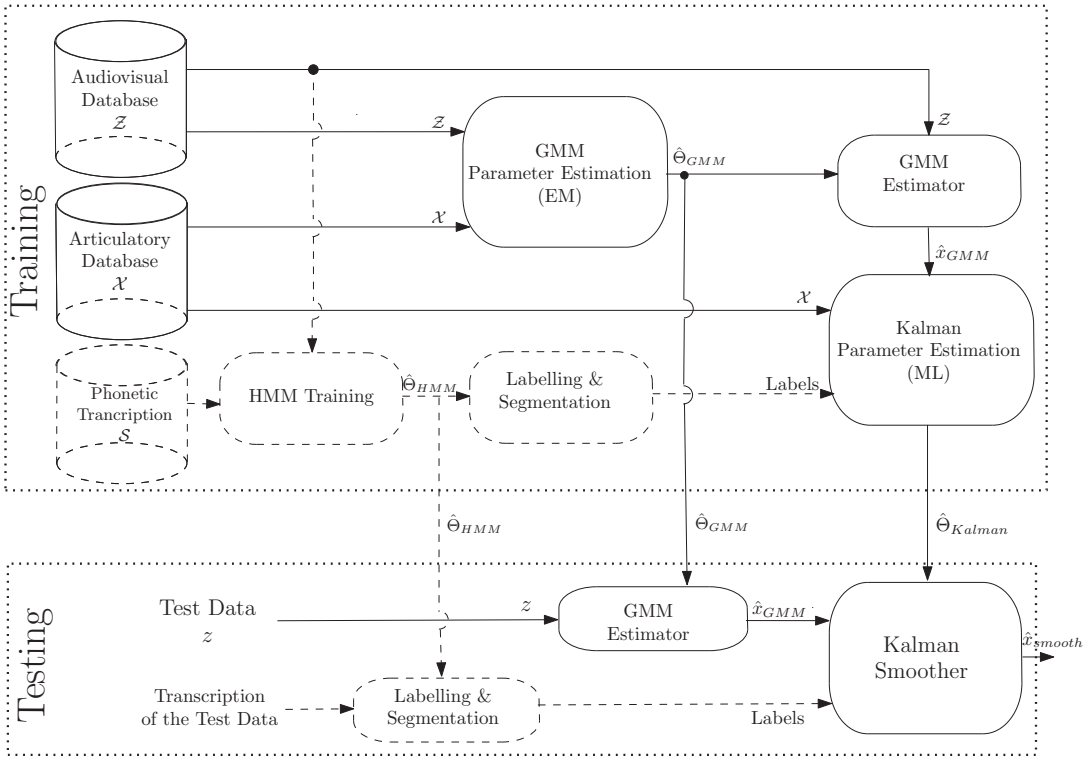


Figure 4.3: The general block diagram of the smoothed GMM inversion (both training and testing phases) proposed in this chapter.

4.4.1 Smoothing Problem Formulation

In this subsection, we will denote the output of the first stage articulatory inversion algorithm as y_n and the corresponding true articulatory state as x_n. The trajectories of the variables are shown by subscripts like $x_{1:N} \triangleq \{x_1, \ldots, x_N\}$. We assume that the dynamics of x_n, and the relationship between x_n and y_n are governed by nonlinear relations that can be approximated by the following piece-wise linear Gaussian dynamic system [102].

$$x_{k+1} = F_1(s_{k+1})x_k + F_2(s_{k+1})x_{k-1} + g(s_{k+1}) + \eta_k(s_{k+1}), \tag{4.9}$$

$$y_k = x_k + d(s_k) + v_k(s_k), \tag{4.10}$$

where $s_k \in \{1, \ldots, \mathcal{S}\}$ is the regime variable representing the model index and $\mathcal{S}$ is the total number of regimes. Each regime is a homogeneous unit (like, phoneme) and characterized by an s_k-dependent parameter set Θ_{s_k}. The second-order Markov dynamic system model of (4.9) and (4.10) can be converted into a first order model by defining the augmented state vector $\mathbf{x}_k$ which is composed of the current and the previous articulatory state vectors, i.e., $\mathbf{x}_k \triangleq \left[x_k^T, x_{k-1}^T\right]^T$. Then, the first order model is given by

$$\mathbf{x}_{k+1} = F(s_{k+1})\mathbf{x}_k + u(s_{k+1}) + w_k(s_{k+1}), \tag{4.11}$$

$$y_k = H\mathbf{x}_k + d(s_k) + v_k(s_k), \tag{4.12}$$

where

- $\mathbf{x}_k \in \mathbb{R}^{2n_x}$ denotes the augmented state vector related to articulatory data and n_x is the dimension of articulatory state vector x_k.
- $y_k \in \mathbb{R}^{n_x}$ denotes the observation vector related to the output of GMM estimator with dimension of n_x.
- the augmented initial state $\mathbf{x}_0$ has the regime-dependent distribution $\mathcal{N}(\mathbf{x}_0; \bar{\mathbf{x}}(s), \Sigma(s))$. $\bar{\mathbf{x}}(s)$ and $\Sigma(s)$ are the initial mean and covariance of the state respectively.
- $F(s) \in \mathbb{R}^{2n_x \times 2n_x}$ is the regime-dependent state transition matrix given as

$$F(s) \triangleq \begin{bmatrix} F_1(s) & F_2(s) \\ I_{n_x \times n_x} & 0_{n_x \times n_x} \end{bmatrix};$$

- $u(s) \in \mathbb{R}^{2n_x}$ is the regime-dependent bias vector given as $u(s) \triangleq [g^T(s), 0_{1\times n_x}]^T$;
- $w_k(s) \triangleq [\eta_k^T(s), 0_{1\times n_x}]^T$ where $\eta_k(s) \sim \mathcal{N}(\eta(s); 0_{n_x\times n_x}, Q(s))$ and $Q(s) \in \mathbb{R}^{n_x\times n_x}$ is the covariance matrix of process noise.
- $H \in \mathbb{R}^{2n_x\times 2n_x}$ is the observation matrix defined as $H \triangleq [I_{n_x\times n_x}, 0_{n_x\times n_x}]$.
- $d(s) \in \mathbb{R}^{n_x}$ is the regime-dependent measurement bias;
- $v_k(s) \sim \mathcal{N}(v(s); 0_{n_x\times n_x}, R(s))$ is the regime-dependent Gaussian observation noise. $R(s) \in \mathbb{R}^{n_x\times n_x}$ is the covariance matrix of the observation noise.

$I_{n_x\times n_x}$ and $0_{n_x\times n_x}$ are the identity and zero matrices with dimension $n_x \times n_x$. The vector $0_{1\times n_x}$ is the zero vector with dimension n_x. The regime dependent parameter set of the model can be defined as

$$\Theta_s = \{\bar{\mathbf{x}}_s, \Sigma_s, F(s), u(s), d(s), Q(s), R(s)\}.$$

In this work, it is assumed that the regime variables are known for each time instant n and they are going to specifically model the extra information such as phonetic transcription in the articulatory inversion. In such a scenario, the MMSE estimate $\hat{x}_{k|N}$ of the true articulatory state, which is defined as

$$\hat{x}_{k|N} \triangleq E[x_k|y_{1:N}, s_{1:N}],$$

is given by a Kalman smoother if the parameter set of the state space representation is known. Therefore, the process of smoothing articulatory trajectories involves two separate tasks: learning the parameter set, and inferring the state.

4.4.2 Learning the Parameter Set

The parameter set
$\Theta_s = \{\bar{\mathbf{x}}_s, \Sigma_s, F(s), u(s), d(s), Q(s), R(s)\}$ for regime s is to be estimated using a training data set. We consider here the following type of database: Our main database is composed of

- $\mathcal{M}$ acoustic trajectories shown as $z_{1:N_d}^{1:\mathcal{M}} \triangleq \{z_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$, where N_d is the length of the dth trajectory.
- $\mathcal{M}$ corresponding true articulatory trajectories shown as $x_{1:N_d}^{1:\mathcal{M}} \triangleq \{x_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$;
- $\mathcal{M}$ corresponding regime trajectories shown as $s_{1:N_d}^{1:\mathcal{M}} \triangleq \{s_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$.

Note that the availability of the regime variables does not cause a loss of generality because the case where the regime variables are not available is equivalent to the case that $\mathcal{S} = 1$ and $s_n^d = 1$ for all n and d.

The main database described above is processed to obtain the secondary database to train the model described in the previous subsection as follows.

- An acoustic to articulatory GMM is trained using the EM algorithm on the main database.
- All $\mathcal{M}$ acoustic trajectories $z_{1:N_d}^{1:\mathcal{M}}$ are input to the trained GMM model and the corresponding estimated articulatory trajectories shown as $y_{1:N_d}^{1:\mathcal{M}} \triangleq \{y_{1:N_d}^d\}_{d=1}^{\mathcal{M}}$ are obtained.
- All regime trajectories are partitioned into fragments of constant regime, i.e., each regime trajectory $s_{1:N_d}^d$ is divided into sub-regime trajectories $\{s_{n_s^j:n_e^j}^d\}_{j=1}^{m_d}$ such that when $s_{n_s^j:n_e^j}^d$ are concatenated, $s_{1:N_d}^d$ is obtained and all elements of $s_{n_s^j:n_e^j}^d$ are equal.
- All $\mathcal{M}$ estimated articulatory trajectories $y_{1:N_d}^{1:\mathcal{M}}$ and true articulatory trajectories $x_{1:N_d}^{1:\mathcal{M}}$ are partitioned according to their corresponding regime trajectories. Without loss of generality, all the partitioned sub-sequences are re-indexed and grouped according to the regime variables to obtain $\mathcal{M}_s$ estimated articulatory trajectories $y_{1:N_d}^{1:\mathcal{M}_s} \triangleq \{y_{1:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ and true articulatory trajectories $x_{1:N_d}^{1:\mathcal{M}_s} \triangleq \{x_{1:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ for each $s = 1, \ldots, \mathcal{S}$.
- The augmented states $\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s} \triangleq \{\mathbf{x}_{2:N_d}^d\}_{d=1}^{\mathcal{M}_s}$ are formed from $x_{1:N_d}^{1:\mathcal{M}_s}$. For the sake of simplicity, we call the combination of $\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s}$ and $y_{2:N_d}^{1:\mathcal{M}_s}$ as $\mathcal{D}_s$.

The database operations described above are illustrated schematically in the training part of Fig. 4.3 which shows both training and testing parts of the

GMM inversion and smoothing operations used in this work. Notice that the dashed blocks and lines in the figure represent the operations concerning the auxiliary information of phonetic transcriptions and in the case of unavailability of such information, they should be omitted.

Suppose now that the training data set $\mathcal{D}_s$ is given for the sth regime. The unknown parameter set Θ_s can be estimated by maximizing the logarithm of the likelihood $\mathcal{L}(\Theta_s)$ of the training data set $\mathcal{D}_s$, i.e.,

$$\hat{\Theta}_s = \arg\max_{\Theta_s} \mathcal{L}(\Theta_s) \tag{4.13}$$

where, $\mathcal{L}(\Theta_s) \triangleq \ln p(\mathbf{x}_{2:N_d}^{1:\mathcal{M}_s}, y_{2:N_d}^{1:\mathcal{M}_s}|\Theta_s)$. Under the assumption of Markov dynamics, the joint likelihood can be written as

$$\begin{aligned}\mathcal{L}(\Theta_s) &= \sum_{d=1}^{\mathcal{M}_s} \ln p(\mathbf{x}_{2:N_d}^d, y_{2:N_d}^d|\Theta_s) &(4.14)\\ &= \sum_{d=1}^{\mathcal{M}_s} \ln p(\mathbf{x}_2^d|\Theta_s) + \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d} \ln p(\mathbf{x}_k^d|\mathbf{x}_{k-1}^d, \Theta_s) \\ &\quad + \sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d} \ln p(y_k^d|\mathbf{x}_k^d, \Theta_s) &(4.15)\end{aligned}$$

where

$$\begin{aligned}\mathbf{x}_2|\Theta_s &\sim \mathcal{N}(\mathbf{x}_s; \bar{\mathbf{x}}(s), \Sigma(s)), &(4.16)\\ \mathbf{x}_k|\mathbf{x}_{k-1}, \Theta_s &\sim \mathcal{N}(\mathbf{x}_k; F(s)\mathbf{x}_{k-1} + u(s), Q(s)),\\ y_k|\mathbf{x}_k, \Theta_s &\sim \mathcal{N}(y_k; \mathbf{x}_k + d(s), R(s)).\end{aligned}$$

Taking derivatives of (4.14) for each unknown parameter, and setting the derivatives equal to zero, estimation formulas for parameter set Θ_s can be seen in Algorithm-4.4.1.

4.4.3 Inference (State Estimation)

After the parameter learning stage, the smoothed state $\hat{\mathbf{x}}_{k|N}$ can be estimated by a Kalman smoother. For this purpose, first a Kalman filter estimates the whole filtered state $\hat{\mathbf{x}}_{k|k}$ in the forward direction, and then in the backward direction, the Kalman smoother estimates the smoothed state $\hat{\mathbf{x}}_{k|N}$. The Kalman smoothing algorithm can be seen in Algorithm-4.4.2.

Algorithm 4.4.1 (Maximum Likelihood Parameter Estimation)

Define the following summations:

$$\bar{\mathbf{x}}_c \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}\mathbf{x}_k^d, \qquad \bar{\mathbf{x}}_p \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}\mathbf{x}_{k-1}^d, \qquad \bar{y}_c \triangleq \sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d} y_k.$$

The estimated parameters for regime s are given as

$$\begin{aligned}
\hat{F}(s) =& \left(\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_k^d(\mathbf{x}_{k-1}^d)^T] - \frac{\bar{\mathbf{x}}_c\bar{\mathbf{x}}_p^T}{\sum_{d=1}^{M_s}(N_d-2)}\right) \\
&\times \left(\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_{k-1}^d(\mathbf{x}_{k-1}^d)^T] - \frac{\bar{\mathbf{x}}_p\bar{\mathbf{x}}_p^T}{\sum_{d=1}^{\mathcal{M}_s}(N_d-2)}\right)^{-1} \\
\hat{u}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-2)}[\bar{\mathbf{x}}_c - \hat{F}(s)\bar{\mathbf{x}}_p] \qquad (4.17)\\
\hat{d}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}[\bar{y}_c - \bar{\mathbf{x}}_c] \qquad (4.18)\\
\hat{Q}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}\sum_{d=1}^{\mathcal{M}_s}\sum_{k=3}^{N_d}[\mathbf{x}_k^d - \hat{F}(s)\mathbf{x}_{k-1}^d - \hat{u}(s)] \\
&\times [\mathbf{x}_k^d - \hat{F}(s)\mathbf{x}_{k-1}^d - \hat{u}(s)]^T \qquad (4.19)\\
\hat{R}(s) =& \frac{1}{\sum_{d=1}^{\mathcal{M}_s}(N_d-1)}\sum_{d=1}^{\mathcal{M}_s}\sum_{k=2}^{N_d}[y_k^d - \mathbf{x}_k^d - \hat{d}(s)] \\
&\times [y_k^d - \mathbf{x}_k^d - \hat{d}(s)]^T \qquad (4.20)\\
\hat{\bar{\mathbf{x}}}(s) =& \frac{1}{\mathcal{M}_s}\sum_{d=1}^{M_s}\mathbf{x}_2^d \qquad (4.21)\\
\hat{\Sigma}(s) =& \frac{1}{\mathcal{M}_s}\sum_{d=1}^{\mathcal{M}_s}[\mathbf{x}_2^d - \hat{\bar{\mathbf{x}}}(s)][\mathbf{x}_2^d - \hat{\bar{\mathbf{x}}}(s)]^T \qquad (4.22)
\end{aligned}$$

Algorithm 4.4.2 (Kalman Smoother)

- *For* $k = 1, ..., N$ *(*$\hat{\mathbf{x}}_1 = \bar{\mathbf{x}}$ *and* $\Sigma_{1|1} = \bar{\Sigma}$*) (Forward Pass)*
  - *Prediction:*

$$\hat{\mathbf{x}}_{k+1|k} = F\hat{\mathbf{x}}_k + u \tag{4.23}$$

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + Q_{k+1} \tag{4.24}$$

  - *Filtering:*

$$\hat{y}_{k+1|k} = H\hat{\mathbf{x}}_{k+1|k} + d \tag{4.25}$$

$$S_{k+1} = H\Sigma_{k+1|k}H^T + R_k \tag{4.26}$$

$$K_{k+1} = \Sigma_{k+1|k}H^T S_{k+1}^{-1} \tag{4.27}$$

$$\hat{\mathbf{x}}_{\mathbf{k+1|k+1}} = \hat{\mathbf{x}}_{k+1|k} + K_{k+1}(y_{k+1} - \hat{y}_{k+1|k}) \tag{4.28}$$

$$\Sigma_{k+1|k+1} = \Sigma_{k+1|k} - \Sigma_{k+1|k}H^T S_{k+1}^{-1} H\Sigma_{k+1|k} \tag{4.29}$$

- *End For*
- *For* $k = N - 1, ..., 1$ *(Backward Pass)*
  - *Smoothing:*

$$L_k = \Sigma_{k|k}H^T\Sigma_{k+1|k}^{-1} \tag{4.30}$$

$$\hat{\mathbf{x}}_{k|N} = \hat{\mathbf{x}}_{k|k} + L_k(\hat{\mathbf{x}}_{k+1|N} - \hat{\mathbf{x}}_{k+1|k}) \tag{4.31}$$

$$\Sigma_{k|N} = \Sigma_{k|k} + L_k(\Sigma_{k+1|N} - \Sigma_{k+1|k})L_k^T \tag{4.32}$$

- *End For*

Remark 4.4.3 *When there is extra information like regime variables (phonetic transcriptions in our case) the measurement covariance estimate* R_s *calculated in* (4.20) *is an appropriate quantity to use for* R_k *above. However, when this extra information is absent,* R_s *becomes too coarse in general. In this case, the covariance* (4.8) *provided by the GMM based inversion method is more suitable to be used as* R_k*.*

4.5 Acoustic and Visual Information Fusion

The fusion of the audio and visual features improves the performance of the articulatory inversion. The fusion of audio and visual information is examined in [72] with an HMM based audiovisual inversion method and in [71] with a linear and nonlinear (Artificial Neural Network (ANN) based) audiovisual inversion. This part of the thesis examines the audiovisual information fusion for GMM based inversion. The fusion of audio and visual features is performed in three ways: Early, late, and modified-late fusion. The early-fusion and late-fusion strategies are modeled after the methods of [72]. The modified-late fusion strategy is a type of late fusion algorithm based on observability characteristics of articulatory state components.

4.5.1 Early Fusion

In early fusion (feature or centralized fusion), the audio features (MFCC, LPC,...) and visual features are augmented to form a large feature vector, and the GMM regression based inversion method is conducted using these combined features. Mathematically speaking, the acoustic measurement vector z in (4.7) is replaced by z_e defined as

$$z_e = \begin{bmatrix} z_a \\ z_v \end{bmatrix} \tag{4.33}$$

where z_a and z_v are vectors of audio and visual features respectively. The early-fusion MMSE articulatory estimate $\hat{x}_e$ is given as:

$$\begin{aligned} \hat{x}_e &\triangleq \mathrm{E}[\mathcal{X}|\mathcal{Z}_e = z_e] \\ &= \sum_{i=1}^{K} \beta^i(z_e) \left[\mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}_e}(\Sigma^i_{\mathcal{Z}_e\mathcal{Z}_e})^{-1}(z_e - \mu^i_{\mathcal{Z}_e})\right] \qquad (4.34) \\ \Sigma_e &\triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}_e = z_e) \\ &= \sum_{i=1}^{K} \beta^i(z_e) \left[\Sigma^i_{\mathcal{X}|\mathcal{Z}_e} + (\mu^i_{\mathcal{X}|\mathcal{Z}_e} - \hat{x}_e)(\mu^i_{\mathcal{X}|\mathcal{Z}_c} - \hat{x}_e)^T\right] \qquad (4.35) \end{aligned}$$

In this work, the early-fusion method is also used to examine various combinations of acoustic-only feature vectors (MFCC, LPC, LAR, ...).

4.5.2 Late Fusion

Late fusion (distributed fusion) combines separate estimation results which are based on audio and visual measurements. In this method, there are two different GMMs, one for audio and the other for visual data. For each time frame, their estimates are averaged with matrix weights to obtain the fused estimate. Matrix weights are generated from the local estimate covariances. A summary of the fusion algorithm is as follows.

Let $\hat{x}_a \triangleq \mathrm{E}[\mathcal{X}|\mathcal{Z}_a]$ and $\hat{x}_v \triangleq \mathrm{E}[\mathcal{X}|\mathcal{Z}_v]$ be the estimated articulatory data using audio and visual measurement respectively, and $\Sigma_a \triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}_a)$ and $\Sigma_v \triangleq \mathrm{Cov}(\mathcal{X}|\mathcal{Z}_v)$ be their corresponding covariance matrices. The audio based quantities are calculated as

$$\hat{x}_a = \sum_{i=1}^{K} \beta^i(z_a) \left[\mu^i_{\mathcal{X}} + \Sigma^i_{\mathcal{X}\mathcal{Z}_a}(\Sigma^i_{\mathcal{Z}_a\mathcal{Z}_a})^{-1}(z - \mu^i_{\mathcal{Z}_a})\right] \tag{4.36}$$

$$\Sigma_a = \sum_{i=1}^{K} \beta^i(z_a) \left[\Sigma^i_{\mathcal{X}|\mathcal{Z}_a} + (\mu^i_{\mathcal{X}|\mathcal{Z}_a} - \hat{x})(\mu^i_{\mathcal{X}|\mathcal{Z}_a} - \hat{x})^T\right] \tag{4.37}$$

The visual based estimate and its covariance are calculated similarly. The late fusion estimate $\hat{x}_l$ can be calculated for each time frame as

$$\hat{x}_l = W_a\hat{x}_a + W_v\hat{x}_v. \tag{4.38}$$

The weighting matrices W_a and W_v are calculated by minimizing the error covariance of $\hat{x}_l$. Considering the fact that each estimate $\hat{x}_a$ and $\hat{x}_v$ uses a different feature set, it can be assumed that estimation errors correponding to these estimates are independent. In that case, for the unbiased estimation, the weights W_a and W_v are found as follows

$$W_a = \Sigma_v(\Sigma_a + \Sigma_v)^{-1},$$
$$W_v = \Sigma_a(\Sigma_a + \Sigma_v)^{-1}.$$

Hence, the late fusion estimate $\hat{x}_l$ and its covariance Σ_l can be written as

$$\hat{x}_l = \Sigma_v(\Sigma_a + \Sigma_v)^{-1}\hat{x}_a + \Sigma_a(\Sigma_a + \Sigma_v)^{-1}\hat{x}_v, \tag{4.39}$$

$$\Sigma_l = \Sigma_a(\Sigma_a + \Sigma_v)^{-1}\Sigma_v. \tag{4.40}$$

The smoothing of the output of late fusion can be handled via Kalman smoother using $\hat{x}_l$ and Σ_l as an observation vector and measurement covariance as explained in Sec. 4.4. It is possible to combine the late fusion and the smoothing process into a single Kalman smoother that uses a state space model which has the same state equation as (4.11) and a modified observation equation given as

$$y_k^a = H_a \mathbf{x}_k + d_a(s) + v_k^a(s), \tag{4.41}$$

$$y_k^v = H_v \mathbf{x}_k + d_v(s) + v_k^v(s), \tag{4.42}$$

where y^a and y^v are the observation vectors which represent the output estimates $\hat{x}_a$ and $\hat{x}_v$ of the Audio-GMM and Visual-GMM respectively. Considering independent audio and visual measurements again, the combination of the observation equations (4.41) and (4.42) yields the following augmented measurement model.

$$y_k^e = H_l \mathbf{x}_k + q(s) + \nu_k(s) \tag{4.43}$$

where

$$y_k^e \triangleq \begin{bmatrix} y_k^a \\ y_k^v \end{bmatrix}, \qquad H_l \triangleq \begin{bmatrix} H_a \\ H_v \end{bmatrix}, \qquad q(s) \triangleq \begin{bmatrix} d_a(s) \\ d_v(s) \end{bmatrix}, \tag{4.44}$$

$$\nu_k(s) \triangleq \begin{bmatrix} v_k^a(s) \\ v_k^v(s) \end{bmatrix}, \qquad R(s) \triangleq \begin{bmatrix} R_a(s) & 0 \\ 0 & R_v(s) \end{bmatrix}. \tag{4.45}$$

The measurement matrices H_a and H_v are given as $H_a = H_v \triangleq [I_{n_x \times n_x}, 0_{n_x \times n_x}]$. The audio and visual GMM covariances Σ_a and Σ_v are used for the measurement covariances R_a and R_v respectively. The Kalman Smoother can be run for the augmented model above to find the smoothed late fusion results. This combined scenario is depicted in Fig. 4.4.

4.5.3 Modified Late Fusion

The modified version of the late fusion algorithm we propose here is based on the fact that the observability degree of the articulatory state components is changed according to their position. The movement of the apparent articulator

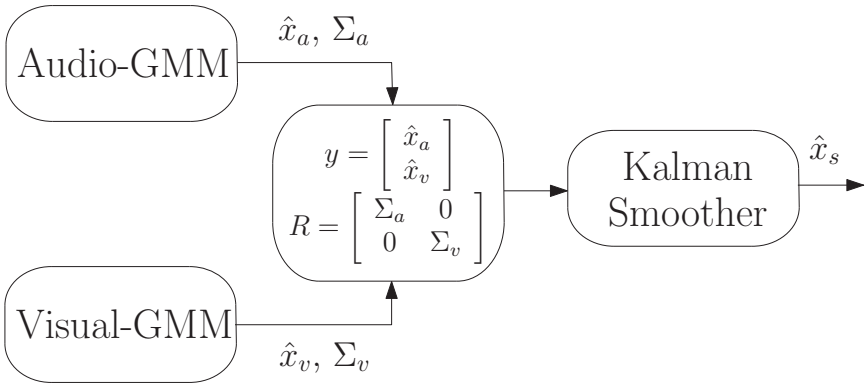


Figure 4.4: Combined late fusion and smoothing process as a single smoother.

such as, lips (upper, lower lip) and jaw (lower incisor) can be captured by the camera quite accurately. On the other hand, the movement of the tongue (tip, body, dorsum) are partly captured by camera and the movement of the velar articulator can not be captured by camera directly. Therefore it is possible to use this prior information in the fusion of the audio and the visual estimates. We here choose to modify the observation equations (4.41) and (4.42) of the late fusion process in order to incorporate these additional information into the combined fusion and smoothing process. The late fusion algorithm uses the observation matrices $H_a = H_v \triangleq [I_{n_x \times n_x}, 0_{n_x \times n_x}]$. In the modified late fusion method, these observation matrices are generalized as follows.

$$H_a^m \triangleq \begin{bmatrix} C_a & 0 \end{bmatrix} \qquad\qquad H_v^m \triangleq \begin{bmatrix} C_v & 0 \end{bmatrix}. \tag{4.46}$$

The acoustic C_a and visual C_v observation matrices are chosen diagonal by adjusting each diagonal element to represent the observability degree of the corresponding state component. This is done by selecting the diagonal elements in the interval [0,1] where 1 corresponds to a fully observable state and 0 a completely unobservable state. In the experimental studies we used the values 0.9, 0.6, and 0.1 for "highly observable", "moderately observable" and "poorly observable" components of the states.

4.6 Experimental Studies

4.6.1 Experiments

In this work, we use the MOCHA database [103]. The acoustic data and EMA trajectories of one female talker (fsew0) are used; these data include 460 sentences. Audio and visual features were computed using a 36 ms window with 18 ms shift. The summary of audiovisual feature types used in this chapter are given in Table 4.1. The articulatory data are EMA trajectories, which are the X and Y coordinates of the lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum. EMA trajectories are normalized by the methods suggested in [64] and down-sampled to match the 18 ms shift rate. All the model parameters of the GMM, HMM and Kalman smoother models are tested using 10-fold cross-validation. For each fold, nine tenths of the data (414 sentences) are used for training and one tenth (46 sentences) for testing. Cross-validation performance measures (RMS error and correlation coefficient) are computed as the average of all ten folds. The segmentation and labeling process is handled via HMM based on force alignment mode. In this work, three-sate, left-to-right HMM models are used for each phoneme. Total 46 models are trained for MOCHA database, 44 for the phonemes and 2 for breath and silence. The experiments can be classified into three classes according to the feature vectors that they use.

- Experiments that each use only one feature vector are given in Table 4.1
- Experiments that each use a combination of two or three different type of acoustic features are given in Table 4.1. Note that 2×2 or 3×3 combination of all features is a huge number. Hence, some combinations that are believed to be good are selected and only their results are given. The selected combination and their performance are shown in Table 4.2 as well as Fig. 4.7 and Fig. 4.8
- Experiments that fuse the acoustic and visual features. This set of experiments examines the performance of the various fusion types (e.g., early,

late and the modified late fusion) of acoustic and visual estimation results.

Another classification of the experiments is based on the post processing method that they use. The aim of the post processing is to smooth the estimated articulatory trajectories. In the literature, the smoothing is obtained either using time-derivatives features or passing the output of the estimation through a low-pass filter. In this work, we have applied two different Kalman smoothers, which is one of the novel parts of the work, and compare their performance with the performance of an optimized low-pass filter (LPF) as well as using time derivative features. LPF is Butterworth IIR filter of second order . Filter is first acted on the input, i.e., the estimated articulatory data forward and then the output of this process is applied to the same filter in reverse in order to satisfy the zero phase property. The cut-off frequency of each articulator is optimized over the training database as described in [64, 67]. The two Kalman smoothers either use a model generated by the global data (i.e., all phoneme are contributed to generate global model) or phoneme based data (i.e., only the data that belong to a certain phonetic class is used to obtain a model). In the first case only one model is generated and we call it *Global Kalman smoother* while in the second model there are 46 models and we call them *Phoneme (Phone) based Kalman smoother*. In phoneme based Kalman smoother, it is assumed that the phonetic transcription of test data is available and the test data can be segmented and labeled via HMM forced alignment.

4.6.2 Performance and Significance Test Measures

The performance of the algorithms is measured using three performance measures, namely, RMS error, normalized RMS error and correlation coefficient, all of which are described in [64, 72].

- RMS error:

$$E^i_{RMS} \triangleq \sqrt{\frac{1}{\mathcal{I}}\sum_{k=1}^{\mathcal{I}}(x^i_k - \hat{x}^i_k)^2}, \quad i = 1, \ldots, n_x \tag{4.47}$$

Table 4.1: Audio-visual feature types used in this study.

FE	Four formant frequencies and their energy levels $FE = [F_1, F_2, F_3, F_4, E_1, E_2, E_3, E_4]$
M	Mel-frequency cepstral coefficients ($MFCC$) $M = [M_1, \ldots, M_{13}]$
L	Linear Predictive Coding coefficients (LPC) $L = [L_1, \ldots, L_{18}]$
R	Log Area Ratio coefficients (LAR) $R = [R_1, \ldots, R_{18}]$
S	Line Spectral Frequencies coefficients (LSF) $S = [S_1, \ldots, S_{18}]$
V	Visual Active Appearance coefficients ($Visual$) $V = [V_1, \ldots, V_{39}]$
DX	Combination of X and its time derivatives; velocity and acceleration components $DX = [X, X_\Delta, X_{\Delta\Delta}]$ (X can be any feature type)

where x_k^i and $\hat{x}_k^i$ are true and estimated position, respectively, of the ith articulator in the kth frame. $\mathcal{I}$ and n_x are the total number of frames in the database and the total number of articulators respectively.

- Normalized RMS error:

$$E_{NRMS}^i \triangleq \frac{E_{RMS}^i}{\sigma_i}, \quad i = 1, \ldots, n_x \tag{4.48}$$

where σ_i is the standard deviation of ith articulator x^i.

- Correlation coefficient:

$$\rho_{x,\hat{x}}^i \triangleq \frac{\sum_{k=1}^{\mathcal{I}}(x_k^i - \bar{x}_k^i)(\hat{x}_k^i - \bar{\hat{x}}_k^i)}{\sqrt{\sum_{k=1}^{\mathcal{I}}(x_k^i - \bar{x}_k^i)^2}\sqrt{\sum_{k=1}^{\mathcal{I}}(\hat{x}_k^i - \bar{\hat{x}}_k^i)^2}} \tag{4.49}$$

for $i = 1, \ldots, n_x$ where $\bar{x}^i$ and $\bar{\hat{x}}^i$ are the average position of true and estimated ith articulator respectively.

- Significance Test: The following significance test is performed to compare two articulatory inversion methods (i.e., B_1 and B_2). According to comparison it is determined which one of the methods outperforms the other in terms of pre-defined performance measures (i.e., RMS error). For this

purpose two hypotheses are used: H_0 and H_1. The null hypothesis H_0 states that the performance of method B_1 is not better than that of B_2. The test hypothesis H_1 states that the performance of method B_1 is better than B_2.

$$H_0 : e = J^{B_2} - J^{B_1} \leq 0$$
$$H_1 : e = J^{B_2} - J^{B_1} > 0 \tag{4.50}$$

where, J^{B_1} is the performance measure using the method B_1 and J^{B_2} is performance measure obtained from the method B_2 we reject the null hypothesis if

$$Z = \frac{\bar{e}}{\frac{\sigma_{\bar{e}}}{\sqrt{K}}} > t_0(\alpha) \tag{4.51}$$

where, $\bar{e} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} e_i$, $\sigma_{\bar{e}} = \sqrt{\frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} (e_i - \bar{e})^2}$ and $t_0(\alpha)$ is the threshold based on the upper tail of the normal density with significance level α (for $\alpha = 0.01$, $t_0 = 2.33$). In order to validate the assumption of independent trials, each sentence is treated as a trial, rather than each frame; thus e_i is the average performance for the ith sentence and $\mathcal{M}$ is the total number of sentences. If the performance measure is chosen to be correlation coefficient, a similar significance test can be performed with some modification.

4.6.3 Experimental Results

4.6.3.1 Experimental Results for Single Feature Set

The experimental results given in this section are for the articulatory inversion using only the acoustic data and only visual data. The comparison of acoustic features (MFCC, LPC, LAR, Visual, etc.) and different smoothing methods can be seen in Fig. 4.5 and Fig. 4.6. The first observation from these figures is that the MFCC feature vector gives better results than any other feature vector. This result is similar to those reported in [72] and [104]. The LSF and FE are the second and third best features. The second observation from these

figures is that using dynamic features reduces RMS error about 2 - 6 % and improves the correlation coefficient about 3 - 8 % for different feature sets. The global Kalman and Low-pass smoothing method reduces RMS error about 5 - 10 % and improves the correlation about 8 - 21 %. It is also observed that global Kalman smoother gives slightly better performance compared to low-pass smoother. Moreover, phone based Kalman smoother (if an auxiliary phonetic transcription is available) reduces the RMS error about 16 - 19 % and improves correlation about 19 - 27 %. The best result for RMS error from *DM* is 1.428 mm and 1.39 mm and the best result for correlation coefficient is about 0.807 and 0.823 according the availability of the phonetic transcription.

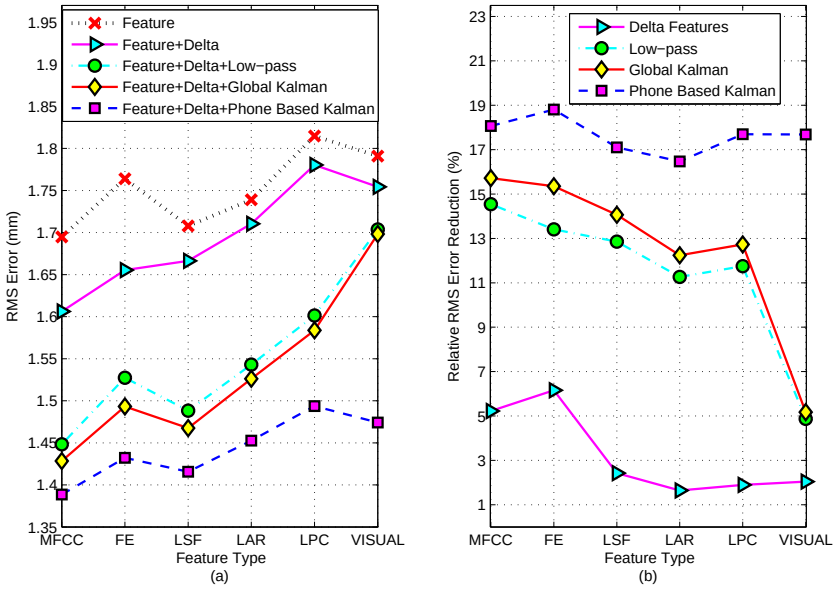


Figure 4.5: RMS errors of using the different audio-visual features and different smoothers (a), and the corresponding percentage RMS error reductions (b) compared to the standard case shown by *Feature* in the figure legends in (a).

It is interesting to observe that LSF and LAR perform better than LPC, although they are derived deterministically from LPC and therefore carry identical information. From the performance of the features we can say that LSF expresses the autoregressive spectral information in a way that is more useful for articulatory inference than LPC or LAR. FE contains four formants, derived from LPC, and formant energies derived from LPC and the power spectrum; although the for-

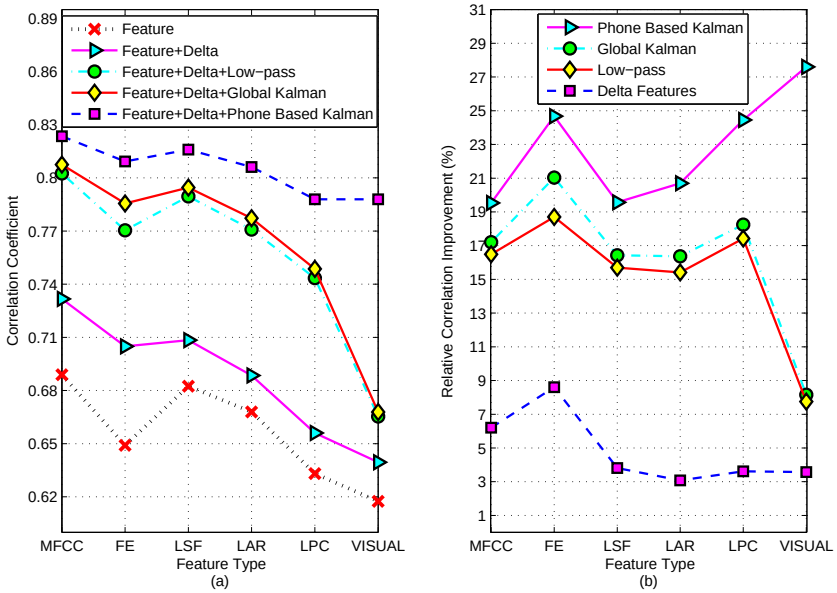


Figure 4.6: Correlation coefficients for use of different audio-visual features and smoothers (a), and the corresponding percentage correlation improvements compared to the standard case (b) (Standard case is shown by *Feature* in the figure legends in (a)).

mant energies are correlated with LPC, they may contain additional information not provided by the LPC vector. Therefore the better performance of FE may be caused by the inclusion of energy in the feature vector, the re-coding of LPC information into a more useful form, or some combinations of these two factors.

It is also interesting to observe from Fig. 4.5 that the performance is highly improved by applying smoothing. The improvement is especially significant when phoneme based Kalman smoother is used for visual data. It seems perhaps that the articulators not directly measured by the camera are estimated in an almost open-loop fashion, based almost entirely on their dynamic behavior. We can say that the resultant system is *state observable*: unmeasured articulator positions are also estimated with some accuracy, so much so that visual performance is better than LPC. When we analyze the correlation coefficient given in Fig. 4.6 we observe parallel results to RMS error, i.e., when correlation is high, RMS error is low.

4.6.3.2 Experimental Results for Combined Acoustic Features

The experimental results given in this section are for the articulatory inversion using some combinations of the various acoustic features. By a combination we mean the concatenation of the feature vectors, i.e., "early fusion". The combination pattern and performance results can be seen in Table 4.2, Fig. 4.7 and Fig. 4.8. In Table 4.2, RMS error given in each cell of the table corresponds to the performance obtained by combining the features in its corresponding row and column heading without using phonetic transcription. As an example the RMS error for combination of DM and DFE gives 1.395 mm error which can be seen on top of the last column. The best result for each row is highlighted. The first observation from this table is that the combination of formant-related features with almost any other feature vector gives better results than any other feature combination. The second observation is that the best combination is the combination of DM and DFE with the corresponding RMS error of 1.395 mm.

Table 4.2: RMS Errors for Combination of Various Acoustic Features.

Features	M	FE	S	A	L	DFE
DM	–	1.415	1.397	1.415	1.422	1.395
DFE	1.418	–	1.423	1.439	1.447	–
DS	1.425	1.447	–	1.454	1.471	1.419
DA	1.452	1.474	1.467	–	1.507	1.446
DL	1.478	1.51	1.498	1.514	–	1.471
DM+S	–	1.398	–	1.42	1.433	–
DS+M	–	1.416	–	1.433	1.441	–

The RMS error and correlation coefficient for highlighted future combinations given in Table 4.2 can be seen in Fig. 4.7 and Fig. 4.8 in detail. In these figures, performance of DM alone can also be seen for comparison purposes. DM is selected for comparison since it is the best single feature vector. The first observation from these figures is that using the low-pass filter for combined acoustic features reduces RMS error about 9.7 - 11.4 % and improves correlation coefficient about 9 - 11.3%. Similarly, Global Kalman smoother reduces the RMS

error about 11 - 13% and improves correlation coefficient about 9.7 - 12.3%.

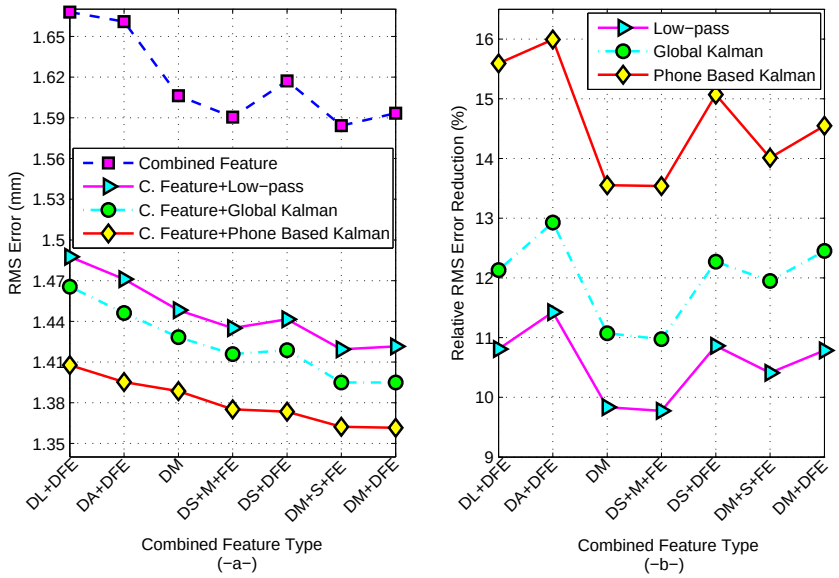


Figure 4.7: RMS errors of using combination of different acoustic features and different smoothers (a), and the corresponding percentage RMS error reductions compared to the standard case (b) (Standard case is shown by *Combined Features* in the figure legends in (a)).

Finally, if the phonetic transcription is available, the phone based Kalman smoother reduces RMS error about 13.5 - 16% and improve correlation coefficient about 12 - 15.5%. Moreover, it can also be seen from these figures that, the best combination of acoustic features is the combination of MFCC and formant-related features ($MD + DFE$). For this acoustic feature combination the RMS error and correlation coefficients for global Kalman smoother are 1.395 mm and 0.816 respectively. In case of phone based Kalman smoother, RMS error and correlation coefficients are 1.362 mm and 0.83 respectively. Observing the results of single features we have concluded that LPC contains irrelevant information and the features LSF and LAR are better than LPC although they are directly obtained from LPC. In Kalman filter terminology this can be explained as using measurement with less measurement error or higher SNR. On the other hand, MFCC had better performance than all the others the new experiments show that using MFCC and formant related features together improves the perfor-

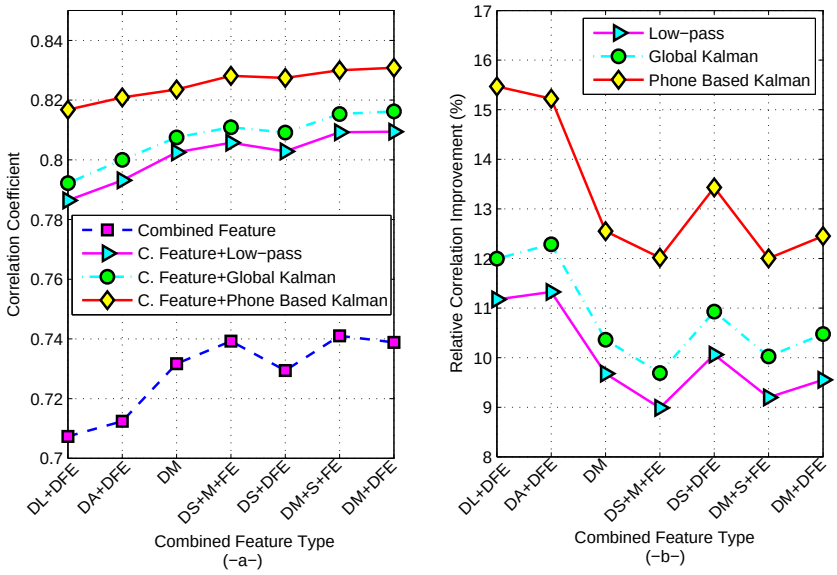


Figure 4.8: Correlation coefficients for use of different acoustic features and smoothers (a), and the corresponding percentage correlation improvements compared to the standard case (b) (Standard case is shown by *Combined Features* in the figure legends in (a)).

mance. That means MFCC and Formants carry complementary information about the position of the articulators.

In this set of experiment, we also conducted two types of statistical significance tests. The first type of significance test examines whether or not articulatory inversion using combined future set $DM + DFE$ (MFCC and formant related acoustic features) significantly outperforms articulatory inversion using only MFCC (DM) features.

Fig. 4.9 provides details regarding the utility of formant related acoustic features in inversion and shows the significance test of the results. From this figure, it is proven that RMS error reduction using formant related features in addition to MFCC features is significant at the $\alpha = 0.01$ level of significance for each articulator.

The second type of significance test is performed to show whether or not articulatory inversion using global Kalman smoother significantly outperforms articu-

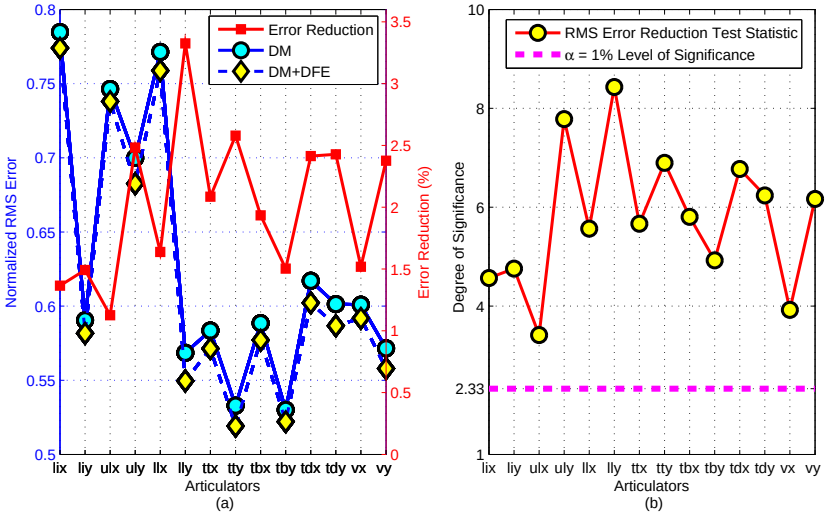


Figure 4.9: Normalized RMS errors (in blue lines and left axis) and corresponding percentage normalized RMS error reductions (in red lines and right axis) of DM+DFE with respect to DM (a). The corresponding significance test results for different articulators are shown in (b). The abbreviations li, ul,ll, tt,tb, td and v denote lower incisor, upper lip, lower lip, tongue tip, tongue body,tongue dorsum and velum, respectively. The suffixes x and y to the articulator abbreviations show the corresponding X and Y coordinates respectively.

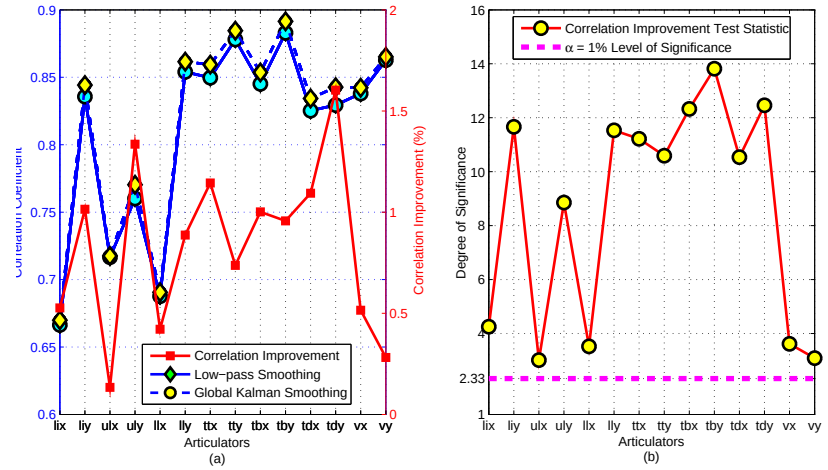


Figure 4.10: Correlation coefficients (in blue lines and left axis) and corresponding percentage correlation improvements (in red lines and right axis) of Low-pass smoother with respect to Kalman smoother (a). The corresponding significance test results for different articulators are shown in (b) Abbreviations related to the names of the articulators are explained in Fig.4.9.

latory inversion using only low-pass smoothing. It can be seen in Fig. 4.10 that using global Kalman smoother instead of low-pass smoother gives statistically significant correlation improvements for each articulator.

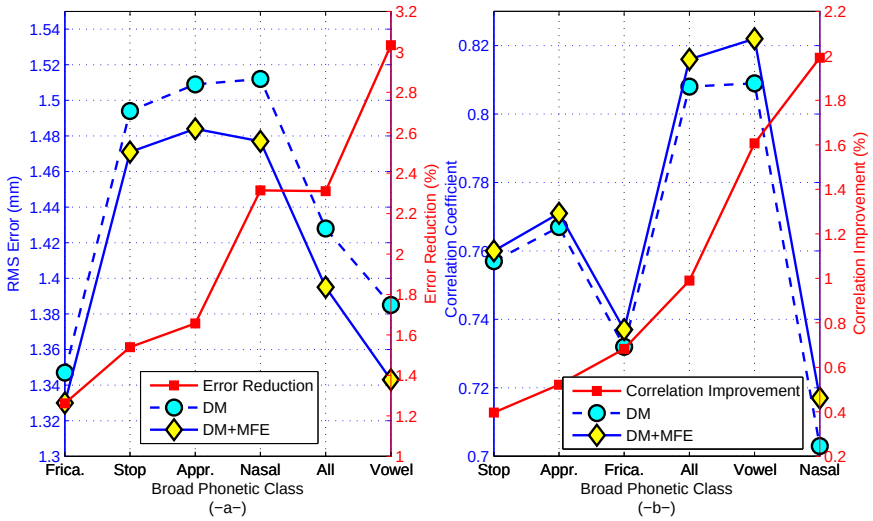


Figure 4.11: RMS errors (in blue lines and left axis) and corresponding percentage RMS error reductions (in red lines and right axis) of DM+DFE with respect to DM for each broad phonetic class (a). Correlation coefficients (in blue lines and left axis) and corresponding percentage correlation improvements (in red lines and right axis) of DM with respect to DM+DFE for each broad phonetic class (b).

Another interesting point is to analyze both types of improvement (augmenting feature vector DM by DFE and using Kalman smoother instead of LPF) for each broad phonetic class. Fig. 4.11 denotes the distribution of RMS error reduction and correlation improvement for each broad phonetic class using formant related acoustic features and MFCC together. From this figures, it can be seen that formant related acoustic features are especially useful when the sounds are vowel and nasal. As an example, RMS error reduction for vowel sound is about 3% and correlation improvement for nasal sound is about 2%. Similar experiment conducted to examine the distribution of RMS error reduction and correlation improvement for each broad phonetic class using global Kalman smoother instead of low-pass smoother. Fig. 4.12 shows that global Kalman smoother is especially useful for stop, fricative and approximant sounds. As an example, RMS error reduction and correlation improvement are about 2.8% and

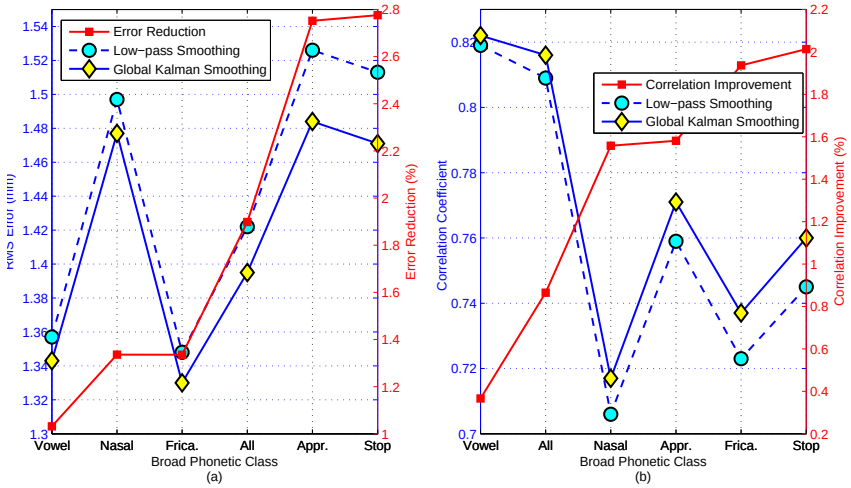


Figure 4.12: RMS errors (in blue lines and left axis) and corresponding percentage RMS error reductions (in red lines and right axis) of Kalman smoother with respect to low-pass filter for each broad phonetic class (a). Correlation coefficients (in blue lines and left axis) and corresponding percentage correlation improvements (in red lines and right axis) of low-pass smoother with respect to Kalman smoother for each broad phonetic class (b).

2% for stop sounds, respectively.

4.6.3.3 Experimental Results for Audiovisual Fusion

The results for the early fusion inversion can be seen in Fig. 4.13. This figure shows that the combination of the visual features with MFCC and formant related (FE) audio features gives better results than other combinations. The graphs also demonstrate the effectiveness of the filtering especially when a good model, i.e., phone based is used in smoothing. The results for different fusion types can be seen in Fig. 4.14 for three sets of articulators ; Lips & jaw, tongue and velar. A detailed examination of this figure illustrurates that inversion using only audio features, on average, gives better performance for tongue related and velar articulators, while inversion using only the visual gives better performance for lips and jaw (lower incisor). Therefore, the modified late fusion method uses the following C_a and C_v matrices (in equation (4.46)) which basically adjust the observability degrees of the state components and improve the inversion

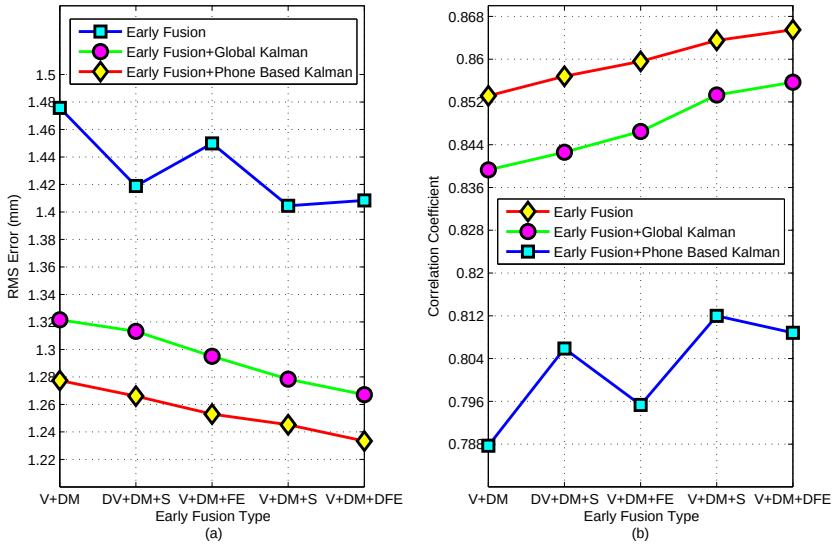


Figure 4.13: RMS errors for the different audio-visual features and smoothers (a), and the corresponding correlation coefficients (b) for the articulatory inversion with early fusion.

performance of the late fusion method.

$$C_a \triangleq [0.9I_{14x14}]$$

$$C_v \triangleq \text{blkdiag}\,[0.9I_{6x6}, 0.6I_{6x6}, 0.1I_{2x2}]$$

The overall results for different fusion types can be seen in Fig. 4.15 and Fig. 4.16. These figures demonstrate that the modified late fusion is the best fusion type for GMM based audiovisual-articulatory inversion. Furthermore, the error reduction curve shows that for the modified late fusion scenario the difference between the global Kalman and phone based Kalman is reduced. Fig.4.17 illustrates the advantages of using audio-visual information fusion compared to using only audio features. It is also shown in this figure that visual information especially improves the inversion performance of incisor and lip related articulators. Fig.4.18 illustrates an example for the estimated and true y-coordinates of articulatory trajectories for lower lip and velar. The utterance are taken form MOCHA database. The audiovisual information fusion with auxiliary information (phonetic transcription) gives better result from others. The summary of experimental results for proposed articulatory inversion method is can be seen in Table 4.3.

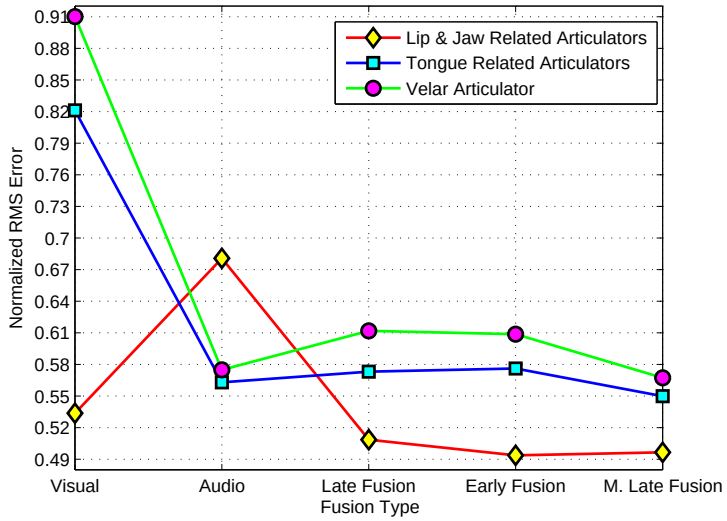


Figure 4.14: Normalized RMS error with different sets of articulators for the various fusion types.

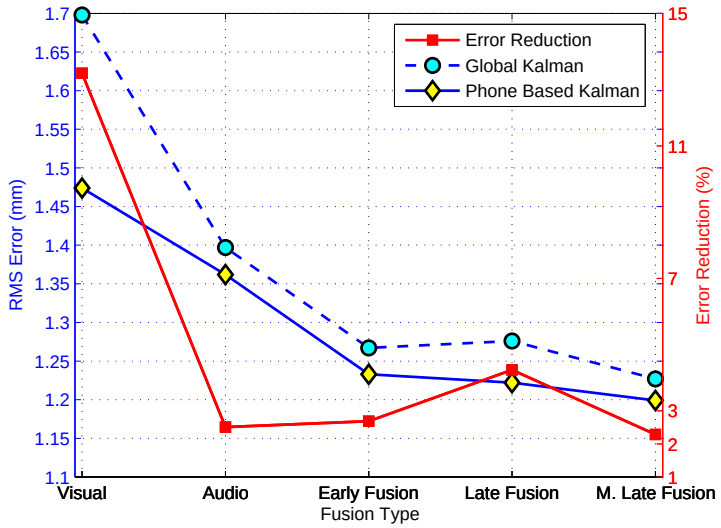


Figure 4.15: RMS errors for the various fusion types (blue lines and left axis). The corresponding percentage RMS error reductions of the phone based Kalman smoother compared to the global Kalman smoother are shown in red lines and right axis.

4.7 Conclusion

This chapter of the thesis demonstrates the utility of vocal-tract-related acoustic features, and of selective audiovisual fusion, for the purpose of audiovisual-to-

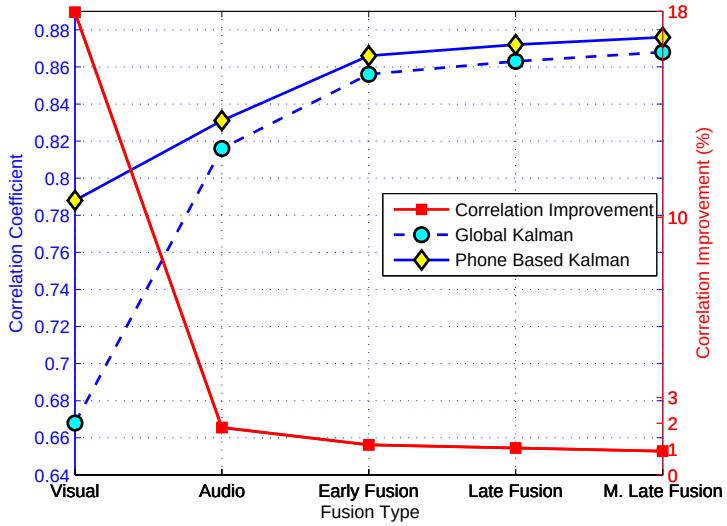


Figure 4.16: Correlation coefficients for the various fusion types (blue lines and left axis). The corresponding percentage correlation improvements of the global Kalman smoother compared to the phone based Kalman smoother are shown in red lines and right axis.

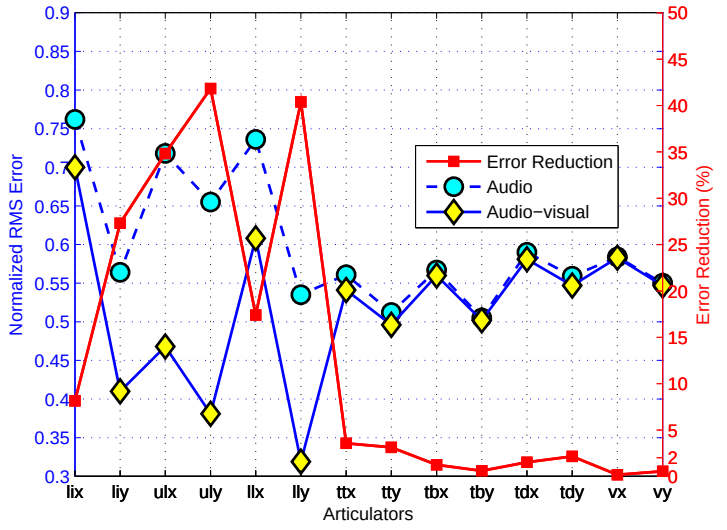
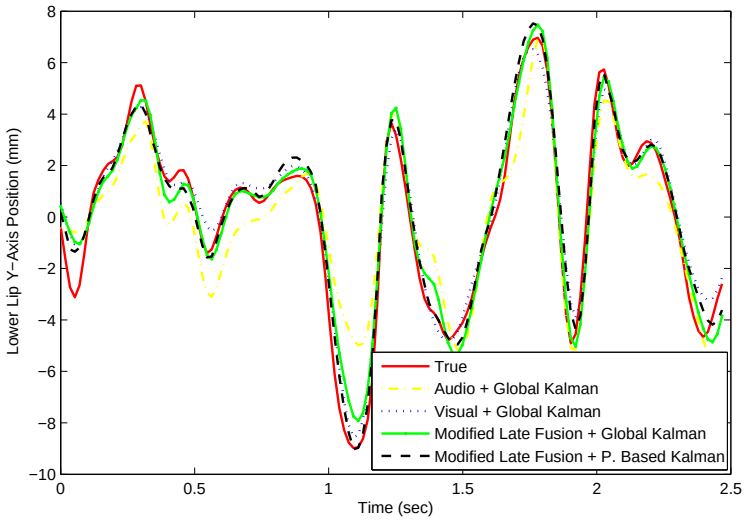
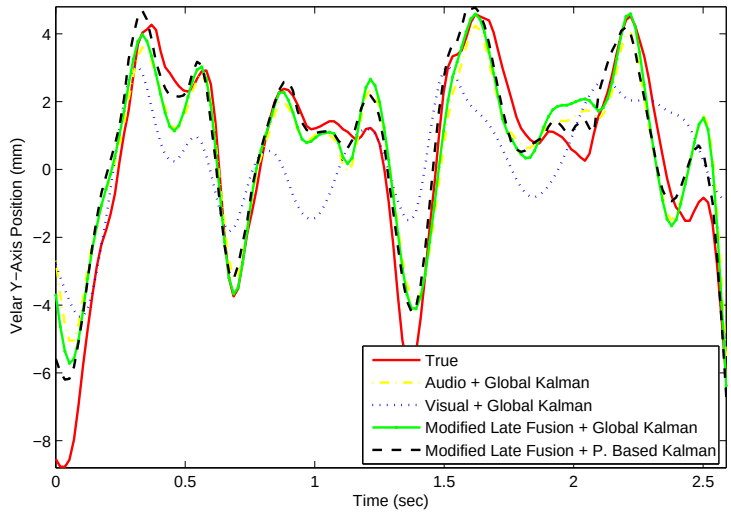


Figure 4.17: Normalized RMS error for each articulator in detail. Abbreviations related to the names of the articulators are explained in Fig. 4.9

articulatory inversion. As reported by other authors, MFCC are the best single feature set for articulatory inversion, but the performance of the MFCC-based inversion algorithm can be improved by appending a vector of line spectral



(a)



(b)

Figure 4.18: Estimated and true (measured) articulatory trajectories of y-coordinates for lower lip and velar for an example taken from MOCHA database. For this utterance, RMS errors are [1.54, 0.76, 0.77, 0.65] mm for lower lip and [0.18, 0.41, 0.17, 0.14] mm for velar ("RMS errors" in the brackets are ordered in the same order as the figure legends).

frequencies, or even better, a vector of formant frequencies and energies. As reported by other authors, early integration of audio and video features is better than audio-only inversion. We also find that audiovisual fusion is especially useful for articulators that are usually visible (the lips and jaw). Moreover it

Table 4.3: Best experimental results for articulatory inversion.

Transcription	Inversion	Feature	# Mix.	RMSE (mm)	Corr. Coef.
	Visual-only	DV	8	1.7	0.668
Unavailable	Audio-only	DM+DFE	154	1.395	0.816
	Audiovisual	DV & DM+DFE	8+154	1.227	0.868
	Visual-only	DV	8	1.474	0.788
Available	Audio-only	DM+DFE	154	1.362	0.831
	Audiovisual	DV & DM+DFE	8+154	1.199	0.876

is possible to obtain fairly good inversion results by using video features only for position estimation of the visible articulators (lips and jaw) and acoustic features for position estimation of the tongue and velum without applying any fusion techniques.

This chapter has also demonstrated a probabilistic smoothing method, based on the Kalman smoother, for the purpose of computing a smoothed articulatory inverse. The Kalman smoother outperforms a simple low-pass smoother by a relatively small margin even without auxiliary information, but the Kalman filter's key advantage is that it can incorporate auxiliary information in a natural way. Given information about the phoneme transcription, for example, the Kalman smoother is able to reduce RMS error of every inversion technique.

By using formant frequencies and energies in acoustic-to-articulatory inversion, Table 4.3 and Fig. 4.7 demonstrate a RMS error of 1.395 mm and correlation coefficient of 0.816 without using any phonemic information; to our knowledge, the best result reported in the literature for this task is 1.45mm and 0.79 [67]. By using audiovisual modified fusion, Table 4.3 and Fig. 4.15 reports an RMS error of 1.227 mm; to our knowledge, the best result reported in the literature for this task is 1.38 mm [72]. By providing a phonetic transcription as auxiliary information to the Kalman smoother, the RMS error is further reduced to only 1.199 mm, with a correlation coefficient of 0.876; to our knowledge, these are the best results reported in the literature for this task.

CHAPTER 5

ACOUSTIC-TO-ARTICULATORY INVERSION BASED ON JUMP MARKOV LINEAR SYSTEMS

5.1 Introduction

Articulation movement of the lips, tongue, jaw, soft palate, and larynx is initiated by the neural activation of muscle fibers. Muscle activation is well modeled as an active stress directed parallel to the muscle fibers [105, 106, 107]. Constant stress generates constant acceleration, limited only by the damping and stiffness of the muscles themselves, therefore movement of the articulators is well modeled, except during collision, as a second-order linear system [108]. Empirical measurements have demonstrated that planned movements are even smoother than would be predicted by the physical constraints of the muscle [109], increasing the utility of a linear dynamic model.

Unfortunately, the transform from articulatory state to acoustic spectrum is irreducibly nonlinear [110]. Given knowledge of both the poles and zeros of the vocal tract impedance, it is possible to specify the shape of the vocal tract [57, 56]; with knowledge of only the poles (the formant frequencies), the articulatory-to-acoustic transform is irreducibly many-to-one. Fortunately, the articulators move smoothly, with few sudden changes, therefore it is possible to use the smoothness of articulatory dynamics as a constraint to select among the many different articulatory configurations that, in theory, match any given acoustic spectrum [59, 111].

Several recent approaches have modeled the dynamics of articulation and acous-

tics using the formalism of a nonlinear dynamic system. Let $z_k \in \mathbb{R}^{n_z}$ be an acoustic observation at time k (e.g., a vector of Mel-frequency cepstral coefficients [104]), and let $x_k \in \mathbb{R}^{n_x}$ be a vector of articulatory state variables (at minimum, in order to model a second-order dynamic system, x_k must specify the positions and velocities of the articulators). The joint dynamics of these two variables may be described by

$$x_{k+1} = f(x_k, s_k, w_k), \tag{5.1}$$

$$z_k = h(x_k, s_k, v_k), \tag{5.2}$$

where $f(\cdot)$ is a deterministic nonlinear function representing the articulatory dynamics, $h(\cdot)$ is a deterministic nonlinear function representing the articulatory-to-acoustic transform, w_k and v_k are random disturbances, and many models represent conscious control of the articulators by the use of a discrete-valued modal state variable $s_k \in \{1, \ldots, r\}$. Thus, for example, Deng and Ma [84, 10] and Frankel et al. [112] both proposed to model $f(\cdot)$ by a linear second-order dynamic system, while $h(\cdot)$ is learned by a multilayer perceptron.

In this work, the nonlinear dynamic model in (5.1) and (5.2) is simplified by assuming that $f(\cdot)$ and $h(\cdot)$ are each piece-wise affine functions of their inputs: given the modal-state variable s_k, the dynamics and the articulatory-to-acoustic transform are both assumed to be linear. The resulting model has been called a Jump-Markov Linear System (JMLS), because it exhibits linear first-order Markov dynamics whose process and observation dynamics occasionally "jump" from one parameter set to another. Exact inference of x_k given observations of z_k is computationally intractable, but many papers have shown the utility of approximate inference algorithms including the interacting multiple models (IMM) algorithm [113, 114] and particle filters [115, 116, 117, 118, 119]. The goal of this chapter is to demonstrate acoustic-to-articulatory inversion using the interacting multiple models algorithm (IMM) applied to a piece-wise-affine (JMLS) representation of articulatory and acoustic dynamics. Sec. 5.2 describes the model, learning algorithm, and inference algorithm; Sec. 5.3 describes experimental methods and results and our conclusion can be found in Sec. 5.4

5.2 Description of Proposed Method

Assume that the nonlinear functions $f(\cdot)$ and $h(\cdot)$ given in (5.1) and (5.2) are approximated via piece-wise affine functions given the modal-state variable s_k. That means the dynamics and the articulatory-to-acoustic transform are both assumed to be linear. The resulting model, called a jump-Markov linear system (JMLS), can be formulated as follows.

$$x_{k+1} = F(s_k)x_k + u(s_k) + w_k(s_k), \tag{5.3}$$

$$z_k = H(s_k)x_k + d(s_k) + v_k(s_k), \tag{5.4}$$

The JMLS includes three essential random variables: a discrete modal-state variable s_k, a continuous articulatory-state vector x_k, and a continuous acoustic observation vector z_k. The stochastic dynamics of these variables are specified as

$$\begin{aligned} P(s_1 = i) &\equiv \pi_{0i} \\ P(s_{k+1} = j | s_k = i) &\equiv \pi_{ij} \\ p(x_1 | s_k = i) &\equiv \mathcal{N}(x_1; \bar{x}_i, \Sigma_i) \\ p(x_k | s_k = i, x_{k-1}) &\equiv \mathcal{N}(x_k; F_i x_{k-1} + u_i, Q_i) \\ p(z_k | s_k = i, x_k) &\equiv \mathcal{N}(z_k; H_i x_k + d_i, R_i) \end{aligned}$$

where

- $s_k \in \mathcal{S} = \{1, \ldots, r\}$ is the modal-state variable. Its dynamics are governed by a transition probability matrix $\Pi = [\pi_{ij}]$ and by an initial-value probability mass function $\pi_0 = [\pi_{0i}]$
- $x_k \in \mathbb{R}^{n_x}$ is the articulatory state vector. Its distribution at the start of each sequence, conditioned on mode variable $s_1 = i$, is Gaussian with mean vector $\bar{x}_i$ and covariance matrix Σ_i. It evolves dynamically, with mode-dependent transition matrix $F_i \in \mathbb{R}^{n_x \times n_x}$, mode-dependent bias vector $u_i \in \mathbb{R}^{n_x}$, and with mode-dependent Gaussian process noise $w_k \sim \mathcal{N}(w_k; 0, Q_i)$.
- $z_k \in \mathbb{R}^{n_z}$ is the acoustic observation vector. It is generated from the articulatory state by a mode-dependent measurement matrix $H_i \in \mathbb{R}^{n_z \times n_x}$,

with mode-dependent measurement bias $d_i \in \mathbb{R}^{n_x}$ and mode-dependent Gaussian observation noise $v_k \sim \mathcal{N}(v_k; 0, R_i)$.

- The parameter set of the model can be defined as

$$\Theta = \{\Pi, \pi_0, \bar{x}_{1,i}, \Sigma_{1,i}, F_i, u_i, Q_i, H_i, d_i, R_i\}_{i=1}^{r} .$$

Assume that we have a training database $D = \{X, Z\}$ that links acoustic observations Z and articulatory state, X. The problem of the acoustic-to-articulatory inversion involves two separate tasks, that we may call "learning" and "inference:"

- LEARNING: The estimation of the model parameters, Θ, given the training data set D and prior distribution $p(\Theta)$. We examine both maximum likelihood (ML) and a maximum *a posteriori* learning (MAP) criterion, therefore

$$\hat{\Theta}^{ML} = \arg\max_{\Theta} p(Z, X|\Theta)$$
$$\hat{\Theta}^{MAP} = \arg\max_{\Theta} p(Z, X|\Theta)p(\Theta)$$

- INFERENCE: The estimation of the articulatory state x_k given acoustic data $z_{1:\tau} = \{z_1, \ldots, z_\tau\}$ and estimated parameter set $\hat{\Theta}$ is found via MMSE estimate of the states as follows

$$x_{k|\tau} = \mathrm{E}[x_k|Z_{1:\tau}]$$

where $\mathrm{E}[\cdot]$ is the expectation operator. If $\tau = k$, the estimation is called filtering; if $\tau = N$ (where N is the length of the observation sequence), the estimation is called fixed-interval smoothing.

The following two sections explore, in detail, the problems of learning Θ from measured X, Z, and of inferring X from measured Z.

5.2.1 Learning of The Model Parameters

In order to learn the model parameter vector Θ, suppose that training database D contains L training sequences: acoustic observations $Z = \{z_{1:N_l}^l\}_{l=1}^{L}$ and artic-

ulatory observations, $X = \{x^l_{1:N_l}\}^L_{l=1}$, and suppose that the each of lth sequence contains N_l vector pairs, that is $x^l_{1:N_l} = \{x^l_1, \ldots, x^l_{N_l}\}$ and $z^l_{1:N_l} = \{z^l_1, \ldots, z^l_{N_l}\}$. That means, acoustic observation sequences Z and continuous-valued articulatory state trajectories X are known (observable), but the underlying modal-state sequences $S = \left\{s^l_{1:N_l}\right\}^L_{l=1}$ are unknown (unobservable).

5.2.1.1 Maximum Likelihood (ML) Based Learning

In the maximum likelihood learning criterion, the parameter set Θ is estimated based on training data set $D = \{X, Z\}$. That is, we do not use any informative prior distribution for the parameter set Θ .

$$\hat{\Theta}^{ML} = \arg\max_{\Theta} p(Z, X|\Theta) \tag{5.5}$$

Since the modal-state sequences S are unknown, there is no closed-form solution for this problem, therefore an iterative solution is required. The expectation maximization (EM) algorithm is known to converge more rapidly to a ML solution, in circumstances where it is applicable, than most other iterative learning methods [120]. Starting from an initial guess of the model parameters $\Theta^{(0)}$, the EM algorithm re-estimates model parameters $\Theta^{(t)}$ in such a way that the expected log-likelihood $\mathrm{E}\left[\ln p(X, Z, S|\Theta^{(t)})|\Theta^{(t-1)}\right]$, and therefore the likelihood $p(X, Z|\Theta^{(t)})$, are guaranteed to be non-decreasing [121]. In each iteration, the EM algorithm involves two steps. In the first step, the expected value of the complete-data log-likelihood (auxiliary function, $Q(\Theta^{(t+1)}, \Theta^{(t)})$ is calculated, with expectation computed over the distribution of the hidden variables S. In second step, the auxiliary function is maximized for each unknown parameter. Considering the conditional independence assumptions given in Fig. 5.1, the complete-data log-likelihood can be defined as

$$\begin{aligned}\ln p(X, Z, S|\Theta) = \textstyle\sum_{l=1}^{L} \ln \Big\{ \Big[P(s^l_1) \prod_{k=2}^{N_l} P(s^l_k|s^l_{k-1})\Big] \\ \Big[p(x^l_1|s^l_1) \prod_{k=2}^{N_l} p(x^l_k|x^l_{k-1}, s^l_k)\Big] \Big[\prod_{k=1}^{N_l} p(z^l_k|x^l_k, s^l_k)\Big] \Big\}\end{aligned} \tag{5.6}$$

In (5.6), the only random variable is the modal state sequence S. The expected log likelihood is therefore

$$Q(\Theta^{(t+1)}, \hat{\Theta}^{(t)}) = E\left\{\ln p(X, Z, S|\Theta^{(t+1)})|X, Z, \hat{\Theta}^{(t)}\right\}$$

where $\hat{\Theta}^{(t)}$ is the known parameter set at iterative step t and it will be used to evaluate the expectation operator. $\Theta^{(t+1)}$ is the new parameter set to be maximized in the maximization step to increase Q. For the remainder of this section, these parameter sets are abbreviated as $\hat{\Theta}$ and Θ, respectively.

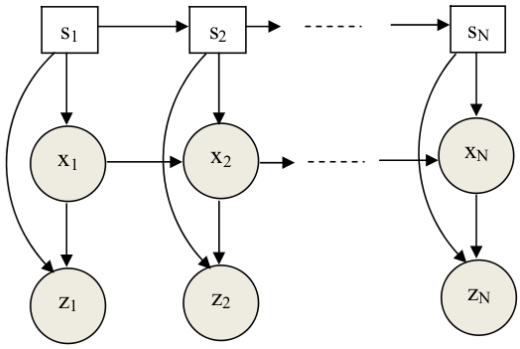


Figure 5.1: Dynamic Bayesian network representation of the jump Markov linear system in training mode.

Since random variable S is discrete-valued, the expectation can be written as a summation over all possible length-N_l modal-state sequences. That is,

$$Q(\Theta, \hat{\Theta}) = \sum_{S \in \mathcal{S}^{N_l}} \ln p(X, Z, S|\Theta) P(S|X, Z, \hat{\Theta})$$

and $\mathcal{S}^{N_l}$ is the set of all possible sets of length-N_l mode sequences. Taking the logarithm of each term in (5.6) reduces Q to a set of independent summations, each of which depends on the posterior probabilities of the mode variables in only one or two frames. The summation can be abbreviated using Levinson's notation [122]:

$$\gamma_k^l(i) \equiv P(s_k^l = i|X, Z, \hat{\Theta}), \tag{5.7}$$

$$\xi_{k-1}^l(i,j) \equiv P(s_{k-1}^l = i, s_k^l = j|X, Z, \hat{\Theta}), \tag{5.8}$$

yielding a relatively compact and easily differentiable form for Q:

$$
\begin{aligned}
Q(\Theta, \hat{\Theta}) = & \sum_{l=1}^{L} \sum_{i=1}^{r} \left[\gamma_1^l(i) \ln \pi_{0i} - \frac{\gamma_1^l(i)}{2} d_{\Sigma_i}(x_1^l, \bar{x}_i) \right. \\
& + \sum_{k=2}^{N} \left[\sum_{j=1}^{r} \xi_{k-1}^l(i,j) \ln \pi_{ij} \right] \\
& - \sum_{k=2}^{N} \left[\frac{\gamma_k^l(i)}{2} d_{Q_i}(x_k^l, F_i x_{k-1}^l + u_i) \right] \\
& \left. - \sum_{k=1}^{N} \left[\frac{\gamma_k^l(i)}{2} d_{R_i}(z_k^l, H_i x_k^l + d_i) \right] \right]
\end{aligned}
$$

where $|\cdot|$ is the matrix determinant, and $d_\Sigma(x, \bar{x}) \equiv (x - \bar{x})^T \Sigma^{-1} (x - \bar{x}) + \ln |2\pi\Sigma|$ is the normalized Mahalanobis distance.

The expectation maximization algorithm seeks, at each iteration, to maximize $Q(\Theta, \hat{\Theta})$ subject to stochastic normalization constraints for the vector π_0 and for each row of the matrix Π. Constrained optimization is performed by finding the saddle point of a Lagrangian function,

$$
Q_\lambda(\Theta, \hat{\Theta}) = Q(\Theta, \hat{\Theta}) + \lambda^T \left(e_{n_x+1} - \begin{bmatrix} \pi_0^T \\ \Pi \end{bmatrix} e_{n_x} \right) \tag{5.9}
$$

where e_{n_x} is a length-n_x vector of ones, and $\lambda \in \Re^{n_x+1}$ is a vector of Lagrange multipliers chosen to satisfy the constraints.

The local posteriors, $\gamma_k^l(i)$ and $\xi_{k-1}^l(i,j)$, can be computed using the Baum-Welch algorithm. In Levinson's notation,

$$
\gamma_k^l(i) = \frac{\alpha_k^l(i) \beta_k^l(i)}{\sum_i \alpha_k^l(i) \beta_k^l(i)} \tag{5.10}
$$

$$
\xi_{k-1}^l(i,j) = \frac{\alpha_{k-1}^l(i) \pi_{ij} b_j(z_k^l, x_k^l) \beta_k^l(j)}{\sum_i \sum_j \alpha_{k-1}^l(i) \pi_{ij} b_j(z_k^l, x_k^l) \beta_k^l(j)} \tag{5.11}
$$

where

$$
\alpha_k^l(i) \equiv p(X_{1:k}^l, Z_{1:k}^l, s_k^l = i | \hat{\Theta}) \tag{5.12}
$$

$$
\beta_k^l(i) \equiv p(X_{(k+1):N}^l, Z_{(k+1):N}^l | x_k^l, s_k^l = i, \hat{\Theta}) \tag{5.13}
$$

$$
b_j(x_k^l, z_k^l) \equiv \mathcal{N}(x_k^l; F_j x_{k-1}^l + u_j, Q_j) \mathcal{N}(z_k^l; H_j x_k^l + d_j, R_j) \tag{5.14}
$$

The "expectation" step of the EM algorithm computes the posterior probabilities $\xi_{k-1}^{l}(i,j)$ and $\gamma_k^l(i)$, using the formulae in (5.10) through (5.14). The "maximization" step then finds a parameter set Θ that maximizes $Q_\lambda(\Theta,\hat{\Theta})$ subject to constraints. Resulting re-estimation formulae are given in Table 5.1.

5.2.1.2 Maximum a Posteriori (MAP) Based Learning

Maximum likelihood estimation of a JMLS tends to over-fit the training data, leading to degraded test-set performance. In order to improve generalizability of the learned parameters, we propose a regularized learning algorithm based on MAP (maximum *a posteriori*) learning. Specifically, we propose to impose a prior distribution $p(u_i, F_i, Q_i)$ that encourages the regression matrix, F_i, to take values slightly smaller (therefore slightly more generalizable [123]) than its maximum-likelihood values. In the maximum a posteriori learning criterion, the parameter set Θ is estimated based on training data set $D = \{X, Z\}$ and prior distribution $p(\Theta)$, therefore

$$\hat{\Theta}^{MAP} = \arg\max_{\Theta} p(Z, X|\Theta)p(\Theta) \tag{5.15}$$

Since again the modal-state sequences S are unknown, there is no closed-form solution for this problem, therefore an iterative solution is required. The EM algorithm for ML solution can be easily adopted for MAP solution [124] . The E-step of the EM algorithm for MAP estimation is same as ML solution. In the M-Step a new auxiliary function $R(\Theta,\hat{\Theta})$ is maximized at each iteration instead of the maximization $Q_\lambda(\Theta,\hat{\Theta})$ in the ML solution procedures. The new auxiliary function $R(\Theta,\hat{\Theta})$ is defined as follows [124];

$$R(\Theta,\hat{\Theta}) = Q_\lambda(\Theta,\hat{\Theta}) + \ln p(\Theta) \tag{5.16}$$

The first function of the RHS of (5.16) is the conventional auxiliary function in the EM algorithm, as given in (5.9). The second part of the RHS of (5.16) is the prior distribution for the model parameter set Θ which can be defined as follows. In this work the model parameter set is divided into two subsets $\Theta = \{\Theta_1, \Theta_2\}$, where $\Theta_1 = \{\Pi, \pi_0, \bar{x}_{1,i}, \Sigma_{1,i}, H_i, d_i, R_i\}_{i=1}^{r}$ and $\Theta_2 = \{\mathbf{F}_i, Q_i\}_{i=1}^{r}$. In here, for simplicity, the augmented parameter $\mathbf{F}_i \triangleq [u_i, F_i]$ is used instead of separate

model parameters F_i and u_i. Under the prior independence assumption, the joint prior density can be written as follows:

$$p(\Theta) = p(\Theta_1)p(\Theta_2) \tag{5.17}$$

In this work prior density $p(\Theta_1)$ is assumed to be a noninformative uniform prior, i.e, $p(\Theta_1) = constant$. Under this assumption, (5.16) reduces to:

$$R(\Theta, \hat{\Theta}) = Q_\lambda(\Theta, \hat{\Theta}) + \ln \prod_{i=1}^{r} p(\mathbf{F}_i, Q_i) \tag{5.18}$$

The joint prior distribution for $p(\mathbf{F_i}, Q_i)$ can be written as

$$p(\mathbf{F_i}, Q_i) = p(\mathbf{F_i}|Q_i)p(Q_i) \tag{5.19}$$

The prior distribution $p(\mathbf{F}_i|Q_i)$ is a zero-mean matrix normal distribution [125, 126] defined as

$$\begin{aligned} p(\mathbf{F}_i|Q_i) &\sim \mathcal{N}(\mathbf{F}_i; 0, Q_i, \Omega_i) \\ &\propto |\Omega_i^{-1}|^{\frac{n_x}{2}} |Q_i^{-1}|^{\frac{n_x+1}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} \Omega_i^{-1} \mathbf{F}_i^T Q_i^{-1} \mathbf{F}_i\right) \end{aligned} \tag{5.20}$$

where, Ω_i and Q_i are two corresponding covariances.

The prior distribution $p(Q_i)$ is the ith inverse Wishart distribution [125, 126] defined as follows

$$\begin{aligned} p(Q_i) &\sim \mathcal{W}^{-1}(Q_i; \Psi_i, v_i) \\ &\propto |Q_i^{-1}|^{\frac{v_i+n_x+1}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} Q_i^{-1} \Psi_i\right) \end{aligned} \tag{5.21}$$

where v_i and Ψ_i are the degrees of the freedom and scale matrix for the inverse Wishart distribution. Combining (5.20) and (5.21), and defining $c_i \triangleq v_i + 2n_x + 2$, the joint prior density $p(\mathbf{F}_i, Q_i)$ in (5.18) can be written as

$$\begin{aligned} &p(\mathbf{F}_i, Q_i) \\ &\propto |\Omega_i^{-1}|^{\frac{n_x}{2}} |Q_i^{-1}|^{\frac{c_i}{2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Omega_i^{-1} \mathbf{F}_i^T Q_i^{-1} \mathbf{F}_i + Q_i^{-1} \Psi_i\right)\right) \end{aligned} \tag{5.22}$$

Combining, (5.9), (5.22) and (5.18) the new auxiliary function $R(\Theta, \hat{\Theta})$ can be

written as follows

$$R(\Theta,\hat{\Theta}) = Q(\Theta,\hat{\Theta}) + \lambda^T \left(e_{n_x+1} - \begin{bmatrix} \pi_0^T \\ \Pi \end{bmatrix} e_{n_x} \right) + \ln \prod_{i=1}^{r} |\Omega_i^{-1}|^{\frac{n_x}{2}} |Q_i^{-1}|^{\frac{c_i}{2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Omega_i^{-1} \mathbf{F}_i^T Q_i^{-1} \mathbf{F}_i + Q_i^{-1} \Psi_i \right) \right) \quad (5.23)$$

The parameter set $\theta_h = \{\Omega_i, \Psi_i, \alpha_i, v_i\}$ is the unknown hyperparameter set, whose assessment is explained in Sec.5.3.2.

To summarize, similar to ML case, the "expectation" step of the EM algorithm computes the posterior probabilities $\xi_{k-1}^l(i,j)$ and $\gamma_k^l(i)$, using the formulae in (5.10) through (5.14). The "maximization" step then finds a parameter set Θ that maximizes $R(\Theta,\hat{\Theta})$ given (5.23). Resulting re-estimation formulas are given in Table 5.2.

5.2.2 Estimation of The Articulatory Trajectories

After learning the parameter set of the model, we are given, one at a time, test sequences of the form $z_{1:N} = \{z_1, \ldots, z_N\}$. The goal of inference is to estimate the posterior probability distributions of the modal state sequence $s_{1:N} = \{s_1, \ldots, s_N\}$ and the articulatory state sequence $x_{1:N} = \{x_1, \ldots, x_N\}$.

5.2.2.1 Filtering

In filtering, the aim is to estimate the posterior distribution of the articulatory state given acoustic observations, $p(x_k|z_{1:k})$, and in particular, to compute the MMSE state estimate

$$\hat{x}_{k|k} = E[x_k|z_{1:k}]$$

The exact posterior distribution, given the model, is specified by

$$p(x_{1:k}|z_{1:k}) = \sum_{s_{1:k} \in \mathcal{S}_{1:k}} p(x_{1:k}|s_{1:k}, z_{1:k}) p(s_{1:k}|z_{1:k}) \quad (5.24)$$

where $\mathcal{S}_{1:k}$ is the set of all length-k sequences drawn from the alphabet $\mathcal{S}$, and the distribution $p(x_{1:k}|s_{1:k}, z_{1:k})$ is a Gaussian distribution of dimension kn_x, with

Table 5.1: EM Re-estimation formulae for the jump Markov linear system

Modal-State Parameters
$\hat{\pi}_{0i} = \frac{1}{L}\sum_{l=1}^{L}\gamma_1^l(i)$
$\hat{\pi}_{ij} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{N_l-1}\xi_k^l(i,j)}{\sum_{l=1}^{L}\sum_{k=1}^{N_l-1}\gamma_k^l(i)}$
Averages
$\bar{x}_i^p \equiv \frac{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)x_{k-1}^l}{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)}$
$\bar{x}_i^c \equiv \frac{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)x_k^l}{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)}$
$\bar{z}_i \equiv \frac{\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k^l(i)z_k^l}{\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k^l(i)}$
Articulatory State Process Parameters
$\hat{\bar{x}}_i = \frac{\sum_{l=1}^{L}\gamma_1^l(i)x_1^l}{\sum_{l=1}^{L}\gamma_1^l(i)}$
$\hat{\Sigma}_i = \frac{\sum_{l=1}^{L}\gamma_1^l(i)(x_1^l-\bar{x}_i)(x_1^l-\bar{x}_i)^T}{\sum_{l=1}^{L}\gamma_1^l(i)}$
$\hat{F}_i = \left(\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k(i)(x_k^l-\bar{x}_i^c)(x_{k-1}^l-\bar{x}^p)^T\right)\times \left(\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k(i)(x_{k-1}^l-\bar{x}_i^p)(x_{k-1}^l-\bar{x}^p)^T\right)^{-1}$
$\hat{u}_i = \bar{x}_i^c - \hat{F}_i\bar{x}_i^p$
$\hat{Q}_i = \frac{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)(x_k^l-\hat{F}_ix_{k-1}^l-\hat{u}_i)(x_k^l-\hat{F}_ix_{k-1}^l-\hat{u}_i)^T}{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)}$
Observation Parameters
$\hat{H}_i = \left(\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k(i)(z_k^l-\bar{z}_i)(x_k^l-\bar{x}^c)^T\right)\times \left(\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k(i)(x_k^l-\bar{x}_i^c)(x_k^l-\bar{x}^c)^T\right)^{-1}$
$\hat{d}_i = \bar{z}_i - \hat{H}_i\bar{x}_i^c$
$\hat{R}_i = \frac{\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k^l(i)(z_k^l-\hat{H}_ix_k^l-\hat{d}_i)(z_k^l-\hat{H}_iz_k-\hat{d}_i)^T}{\sum_{l=1}^{L}\sum_{k=1}^{N_l}\gamma_k^l(i)}$

Table 5.2: MAP based EM Re-estimation formulae for the jump Markov linear system.(Since the estimation formulae of the rest of the parameters are same as given in Table 5.1, we do not repeat them in here.)

$\hat{\mathbf{F}}_i \triangleq \left[\hat{u}_i, \hat{F}_i\right], {\mathbf{x_k}}^l \triangleq \left[1, (x_k^l)^T\right]^T$
$\hat{\mathbf{F}}_i = (\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k(i)x_k^l\mathbf{x}_{k-1}^l)\times(\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k(i)\mathbf{x}_{k-1}^l\mathbf{x}_{k-1}^l+\Omega^{-1})^{-1}$
$\hat{Q}_i = \frac{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)(x_k^l-\hat{\mathbf{F}}_i\mathbf{x}_{k-1}^l)(x_k^l-\hat{\mathbf{F}}_i\mathbf{x}_{k-1}^l)^T+\hat{\mathbf{F}}_i\Omega_i^{-1}\hat{\mathbf{F}}_i^T+\Psi_i}{\sum_{l=1}^{L}\sum_{k=2}^{N_l}\gamma_k^l(i)+v_i+2n_x+2}$

covariance dependent on the mode-dependent transition matrices $F(s_k)$. (5.24) is a mixture Gaussian PDF with r^k mixture components. Because the state vectors are correlated with one another, the marginal distribution at each instant in time, $p(x_k|z_{1:k})$, is also a mixture Gaussian distribution with r^k Gaussian components [127]. Exact inference is impossible for k larger than two or three, therefore approximate solutions are needed. In this work we use the interactive multiple model (IMM) algorithm, which approximates $p(x_k|z_{1:k})$ using a mixture Gaussian distribution with only r components [51]:

$$p(x_k|z_{1:k}) \approx \sum_{j=1}^{r} p(x_k|s_k=j,z_{1:k})p(s_k=j|z_{1:k}) \tag{5.25}$$

The posterior probability density consists of two parts. The first part, $p(x_k|s_k = j, z_{1:k})$, is modeled by a Gaussian PDF whose mean and variance depend on the mode and the observations. The second part, $p(s_k = j|z_{1:k})$, is the mode posterior probability mass function.

In the IMM algorithm, these two posterior functions are calculated recursively. The recursion is based on two phases: update and prediction. In the update phase, prior probability density functions $p(x_k|s_k = j, z_{1:k-1})$ and $p(s_k = j|z_{1:k-1})$ at time instant k are updated using the current observation z_k in order to produce $p(x_k|s_k = j, z_{1:k})$ and $p(s_k = j|z_{1:k})$. In the prediction phase, using these posterior density functions, prior density functions $p(x_{k+1}|s_{k+1} = j, z_{1:k})$

and $p(s_{k+1} = j|z_{1:k})$ are computed. Algorithms for the update and prediction steps are now described.

Suppose that the prior probability distributions are known for frame k. As specified previously, the mode-dependent distribution of x_k is assumed to be Gaussian, therefore

$$p(x_k|s_k = j, z_{1:k-1}) = \mathcal{N}(x_k; \hat{x}^j_{k|k-1}, \Sigma^j_{k|k-1}) \quad (5.26)$$

$$p(s_k = j|z_{1:k-1}) \equiv \mu^j_{k|k-1} \quad (5.27)$$

Indeed, (5.26) and (5.27) are true for $k = 1$: $\mu^i_{1|0} = \pi_{0i}$, $\hat{x}^j_{1|0} = \bar{x}_j$, and $\Sigma^j_{1|0} = \Sigma_j$. Induction continues as follows.

1. *Update:* Given (5.26) and (5.27), the update step computes $p(x_k|z_{1:k})$ and $\mu^j_{k|k} \equiv p(s_k = j|z_{1:k})$ exactly, without further approximation. Indeed, given knowledge of the mode variable $s_k = j$, the articulatory state distribution is updated using a standard Kalman update:

$$p(x_k|s_k = j, z_{1:k}) \propto p(z_k|s_k = j, x_k)p(x_k|s_k = j, z_{1:k-1})$$

where

$$p(z_k|s_k = j, x_k) = \mathcal{N}(z_k; H_j x_k + d_j, R_j)$$
$$p(x_k|s_k = j, z_{1:k-1}) = \mathcal{N}(x_k; \hat{x}^j_{k|k-1}, \Sigma^j_{k|k-1})$$

and therefore

$$p(x_k|s_k = j, z_{1:k}) = \mathcal{N}(x_k; \hat{x}^j_{k|k}, \Sigma^j_{k|k}) \quad (5.28)$$

Similarly, the update equations for the mode posterior are computed exactly as:

$$\begin{aligned} \mu^j_{k|k} &\equiv p(s_k = j|z_{1:k}) \\ &= \frac{p(s_k = j|z_{1:k-1})p(z_k|s_k = j, z_{1:k-1})}{p(z_k|z_{1:k-1})} \\ &= \frac{\mu^j_{k|k-1}\Lambda^j_k}{\sum_{j=1}^r \mu^j_{k|k-1}\Lambda^j_k} \end{aligned}$$

where

$$\Lambda_k^j = p(z_k|s_k = j, z_{1:k-1})$$
$$\mu_{k|k-1}^j = p(s_k = j|z_{1:k-1})$$

Combining these two update equations (5.28),(5.29), the overall state (5.25) is estimated given as

$$p(x_k|z_{1:k}) \approx \sum_{j=1}^{r} \mu_{k|k}^j \mathcal{N}(x_k; \hat{x}_{k|k}^j, \Sigma_{k|k}^j) \tag{5.29}$$
$$\approx \mathcal{N}(x_{k+1}; \hat{x}_{k|k}, \Sigma_{k|k}) \tag{5.30}$$

where

$$\hat{x}_{k|k} = \sum_{j=1}^{r} \mu_{k|k}^j \hat{x}_{k|k}^j$$
$$\Sigma_{k|k} = \sum_{j=1}^{r} \mu_{k|k}^j \left(\Sigma_{k|k}^j + (\hat{x}_{k|k}^j - \hat{x}_{k|k})(\hat{x}_{k|k}^j - \hat{x}_{k|k})^T \right)$$

2. *Prediction:* The prediction step computes posterior distributions of s_{k+1} and x_{k+1} by marginalizing over the possible values of s_k and x_k. Prediction of s_{k+1} may be exactly computed, but the exact posterior of x_{k+1} is a mixture Gaussian with r times as many mixture components as the posterior of x_k; if this iteration is repeated N times, the result is a posterior distribution with r^N mixture components [127]. Many approximate inference algorithms exist; this work uses the approximation called the interacting multiple model (IMM) hypothesis, according to which the posterior of x_{k+1} is approximated by a mixture Gaussian with only r mixture components. Thus the mode posterior is exactly computed as:

$$\begin{aligned} \mu_{k+1|k}^j &\equiv p(s_{k+1} = j|z_{1:k}) \\ &= \sum_{i=1}^{r} \pi_{ij} \mu_{k|k}^i \end{aligned}$$

The articulatory state posterior is exactly computed as:

$$
\begin{aligned}
& p(x_{k+1}|s_{k+1}=j, z_{1:k}) \\
=& \int p(x_{k+1}|x_k, s_{k+1}=j) \sum_{i=1}^{r} \mu_{k|k}^{ij} p(x_k|s_k=i, z_{1:k}) dx_k && (5.31) \\
=& \int \mathcal{N}(x_{k+1}; F_j x_k + u_j, Q_j) \sum_{i=1}^{r} \mu_{k|k}^{ij} \mathcal{N}(x_k; \hat{x}_{k|k}^{j}, \Sigma_{k|k}^{j}) dx_k && (5.32) \\
\approx& \int \mathcal{N}(x_{k+1}; F_j x_k + u_j, Q_j) \mathcal{N}(x_k; \hat{x}_{k|k}^{0j}, \Sigma_{k|k}^{0j}) dx_k && (5.33) \\
=& \mathcal{N}(x_{k+1}; \hat{x}_{k+1|k}^{j}, \Sigma_{k+1|k}^{j}) && (5.34)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_{k|k}^{ij} &\equiv p(s_k=i|s_{k+1}=j, z_{1:k}) \\
&= \frac{\pi_{ij}\mu_{k|k}^{i}}{\mu_{k+1|k}^{j}} \\
\hat{x}_{k|k}^{0j} &= \sum_{i=1}^{r} \mu_{k|k}^{i|j} \hat{x}_{k|k}^{i} \\
\Sigma_{k|k}^{0j} &= \sum_{i=1}^{r} \mu_{k|k}^{i|j} \left(\Sigma_{k|k}^{i} + (\hat{x}_{k|k}^{i} - \hat{x}_{k|k}^{0j})(\hat{x}_{k|k}^{i} - \hat{x}_{k|k}^{0j})^T \right)
\end{aligned}
$$

In the critical IMM approximation in (5.33), $\sum_{i=1}^{r} \mu_{k|k}^{ij} \mathcal{N}(x_k; \hat{x}_{k|k}^{j}, \Sigma_{k|k}^{j})$ is approximated by $\mathcal{N}(x_k; \hat{x}_{k|k}^{0j}, \Sigma_{k|k}^{0j})$. The summary of IMM filtering algorithm can be seen in Algorithm-5.2.1 and Algorithm 5.2.2

Algorithm 5.2.1 (Interacting Multiple Model (IMM) algorithm)

- *For* $j = \{1, \ldots, r\}$ *(Initialization)*
  - $\mu^j_{1|0} = \pi_j$, $\hat{x}^j_{1|0} = \bar{x}^j_1$, $\Sigma^j_{1|0} = \Sigma^j_1$
- *End For*
- *For* $k = \{1, \ldots, N\}$ $i, j = \{1, \ldots, r\}$
  - *Elementary Filter Update*
    $\left[\hat{x}^j_{k|k}, \Sigma^j_{k|k}, \Lambda^j_k\right] = KalmanUpdate\left(\hat{x}^j_{k|k-1}, \Sigma^j_{k|k-1}, z_k, H^j, d^j, R^j\right)$
  - *Mode Probability update*
    $$\mu^j_{k|k} = \frac{\Lambda^j_k \mu^j_{k|k-1}}{\sum_{j=1}^M \mu^j_{k|k-1}\Lambda^j_k}$$
  - *Main Estimation*
    $\left[\hat{x}_{k|k}, \Sigma_{k|k}\right] = Collapse\left(\hat{x}^j_{k|k}, \Sigma^j_{k|k}, \mu^j_{k|k}\right)$
  - *Mode probability prediction*
    $$\mu^j_{k+1|k} = \sum_{i=1}^{M} \pi_{ij}\mu^i_{k|k}$$
  - *Mixing Probability*
    $$\mu^{i|j}_{k|k} = \frac{\pi_{ij}\mu^i_{k|k}}{\mu^j_{k+1|k}}$$
  - *Mixing Estimate*
    $\left[\hat{x}^{0j}_{k|k}, \Sigma^{0j}_{k|k}\right] = Collapse\left(\hat{x}^i_{k|k}, \Sigma^i_{k|k}, \mu^{i|j}_{k|k}\right)$
  - *Elementary Filter Prediction*
    $\left[\hat{x}^j_{k+1|k}, \Sigma^j_{k+1|k}\right] = KalmanPrediction\left(\hat{x}^{0j}_{k|k}, \Sigma^{0j}_{k|k}, F^j, u^j, Q^j\right)$
- *End For*

Algorithm 5.2.2 (Kalman Filtering and Collapse)

- *Kalman Update*

  $\left[\hat{x}_{k|k}, \Sigma_{k|k}, \Lambda_k\right] = KalmanUpdate\left(\hat{x}_{k|k-1}, \Sigma_{k|k-1}, z_k, H, d, R\right)$

$$\begin{aligned}
\hat{z}_{k|k-1} &= H\hat{x}_{k|k-1} + d \\
S_k &= H\Sigma_{k|k-1}H^T + R \\
K_k &= \Sigma_{k|k-1}H^T S_k^{-1} \\
\tilde{z}_k &= z_k - \hat{z}_{k|k-1} \\
\hat{x}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + K_k\tilde{z}_k \\
\Sigma_{k|k} &= \Sigma_{k|k-1} - \Sigma_{k|k-1}H^T S_k^{-1} H\Sigma_{k|k-1} \\
\Lambda_k &= |2\pi\Sigma_{k|k}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\tilde{z}_k^T \Sigma_{k|k}^{-1}\tilde{z}_k\right)
\end{aligned}$$

- *Kalman Prediction*

  $\left[\hat{x}_{k+1|k}, \Sigma_{k+1|k}\right] = KalmanPrediction\left(\hat{x}_{k|k}, \Sigma_{k|k}, F, u, Q\right)$

$$\begin{aligned}
\hat{x}_{k+1|k} &= F\hat{x}_{k|k} + u \\
\Sigma_{k+1|k} &= F\Sigma_{k|k}F^T + Q
\end{aligned}$$

- *Collapse*

  $\left[\hat{x}_{k|k}, \Sigma_{k|k}\right] = Collapse\left(\hat{x}_{k|k}^j, \Sigma_{k|k}^j, \mu_{k|k}^j\right)$, j=$\{1,\ldots,r\}$

$$\begin{aligned}
\hat{x}_{k|k} &= \sum_{j=1}^{r} \mu_{k|k}^j \hat{x}_{k|k}^j \\
\Sigma_{k|k} &= \sum_{j=1}^{r} \mu_{k|k}^j \left(\Sigma_{k|k}^j + (\hat{x}_{k|k}^j - \hat{x}_{k|k})(\hat{x}_{k|k}^j - \hat{x}_{k|k})^T\right)
\end{aligned}$$

5.2.2.2 Smoothing

In the literature, [128, 129] use generalized pseudo-Bayesian of order two (GPB2) algorithm to calculate the smoothed state for JMLS. An IMM smoother is also described in [74]. It is based on two filter approach. In two filter approach, a smoothing algorithm is required to describe backward state dynamics which introduces computationall inefficiency in the estimation procedure. In this section of the thesis, we introduce a smoothing algorithm for JMLS which is based on the IMM approach. The proposed algorithm is based on RTS smoothing [130] method which does not need backward state dynamics. Therefore, the proposed algorithm is computationally much more efficient than the smoothing algorithm given in the literature [128, 129, 74]. In smoothing, the aim is to estimate posterior distribution of the state $p(x_k|z_{1:N})$ given observation sequence $z_{1:N} = \{z_1, \ldots, z_N\}$. Following Interactive Multiple-Models (IMM) estimation procedure, the posterior distribution $p(x_k|z_{1:N})$ can be written as;

$$p(x_k|z_{1:N}) = \sum_{i=1}^{r} p(x_k|s_k = i, z_{1:N}) P(s_k = i|z_{1:N}) \tag{5.35}$$

Now our aim is to find smoothed-elemental estimate $p(x_k|s_k = i, z_{1:N})$ and smoothed-mode probability $P(s_k = i|z_{1:N})$ in a recursive manner

- *Smoothed mode probability recursion*

$$\begin{aligned} \mu^i_{k|N} &\triangleq P(s_k = i|z_{1:N}) \qquad (5.36)\\ &= \sum_{j=1}^{M} P(s_k = i|s_{k+1} = j, z_{1:N}) P(s_{k+1} = j|z_{1:N}) \\ &\approx \sum_{j=1}^{M} P(s_k = i|s_{k+1} = j, z_{1:k}) P(s_{k+1} = j|z_{1:N}) \\ &= \sum_{j=1}^{M} \mu^{i|j}_{k|k} \mu^j_{k+1|N} \end{aligned}$$

- *Smoothed-elemental estimate recursion*

$$p(x_k|s_k=i,z_{1:N}) = \int dx_{k+1} \sum_{j=1}^{r} p(x_k, x_{k+1}, s_{k+1}=j|s_k=i,z_{1:N}) \quad (5.37)$$

$$= \int dx_{k+1} \sum_{j=1}^{r} I_1 \times I_2 \times I_3$$

$$= \sum_{j=1}^{r} I_3 \int dx_{k+1} I_2 \times I_1 \quad (5.38)$$

where,

$$I_1 \triangleq p(x_k|x_{k+1}, s_{k+1}=j, s_k=i, z_{1:N}) \quad (5.39)$$

$$I_2 \triangleq p(x_{k+1}|s_{k+1}=j, s_k=i, z_{1:N}) \quad (5.40)$$

$$I_3 \triangleq P(s_{k+1}=j|s_k=i, z_{1:N}) \quad (5.41)$$

if we examine each term separately

$$\begin{aligned} I_3 &\triangleq P(s_{k+1}=j|s_k=i,z_{1:N}) \quad (5.42)\\ &= \frac{P(s_k=i|s_{k+1}=j,z_{1:N})P(s_{k+1}=j|z_{1:N})}{P(s_k=i|z_{1:N})} \\ &= \frac{P(s_k=i|s_{k+1}=j,z_{1:k})P(s_{k+1}=j|z_{1:N})}{P(s_k=i|z_{1:N})} \\ &= \frac{\mu_{k|k}^{i|j}\mu_{k+1|N}^{j}}{\mu_{k|N}^{i}} \\ &\triangleq \mu_{k+1|N}^{j|i} \end{aligned}$$

The term $\int dx_{k+1} I_2 \times I_1$ can be approximated as

$$\int dx_{k+1} I_2 \times I_1 \approx \mathcal{N}(x_k; \hat{x}_{k|N}^{ij}, \Sigma_{k|N}^{ij}) \quad (5.43)$$

where

$$\hat{x}_{k|N}^{ij} = \hat{x}_{k|k}^{i} + L_k^{ij}(\hat{x}_{k+1|N}^{j} - \hat{x}_{k+1|k}^{ij}) \quad (5.44)$$

$$\Sigma_{k|N}^{ij} = \Sigma_{k|k}^{i} + L_k^{ij}(\Sigma_{k+1|N}^{j} - \Sigma_{k+1|k}^{ij})(L_k^{ij})^T \quad (5.45)$$

$$\begin{aligned} \hat{x}_{k+1|k}^{ij} &\triangleq F^j \hat{x}_{k|k}^{i} + u^j \\ \Sigma_{k+1|k}^{ij} &\triangleq F^j \Sigma_{k|k}^{i} (F^j)^T + Q^j \\ L_k^{ij} &\triangleq \Sigma_{k|k}^{i} (F^j)^T \Sigma_{k+1|k}^{ij} \end{aligned}$$

The proof of this approximation can be seen Appendix B.1. Therefore, the density $p(x_k|s_k = i, z_{1:N})$ can be written as follows.

$$p(x_k|s_k = i, z_{1:N}) = \sum_{j=1}^{r} \mu_{k+1|N}^{j|i} \mathcal{N}(x_k; \hat{x}_{k|N}^{ij}, \Sigma_{k|N}^{ij}) \tag{5.46}$$

The mixture of Gaussian given in the (5.46) is represented as a single Gaussian as

$$\begin{aligned} p(x_k|s_k = i, z_{1:N}) &\approx \sum_{j=1}^{r} \mu_{k+1|N}^{j|i} \mathcal{N}(x_k; \hat{x}_{k|N}^{ij}, \Sigma_{k|N}^{ij}) \\ &\approx \mathcal{N}(x_k; \hat{x}_{k|N}^{i}, \Sigma_{k|N}^{i}). \end{aligned}$$

where

$$\begin{aligned} \hat{x}_{k|N}^{i} &= \sum_{j=1}^{r} \mu_{k+1|N}^{j|i} \hat{x}_{k|N}^{ij} \\ \Sigma_{k|N}^{i} &= \sum_{j=1}^{r} \mu_{k+1|N}^{j|i} \left[\Sigma_{k|N}^{ij} + \left(\hat{x}_{k|N}^{ij} - \hat{x}_{k|N}^{i} \right) \left(\hat{x}_{k|N}^{ij} - \hat{x}_{k|N}^{i} \right)^T \right] \end{aligned}$$

Therefore, the over all estimate $p(x_k|z_{1:N})$ in (5.35) can be calculated as follows.

$$\begin{aligned} p(x_k|z_{1:N}) &= \sum_{i=1}^{r} p(x_k|s_k = i, z_{1:N}) P(s_k = i|z_{1:N}) \\ &\approx \sum_{i=1}^{r} \mu_{k|N}^{i} \mathcal{N}(x_k; \hat{x}_{k|N}^{i}, \Sigma_{k|N}^{i}) \\ &\approx \mathcal{N}(x_k; \hat{x}_{k|N}, \Sigma_{k|N}) \end{aligned}$$

where

$$\begin{aligned} \hat{x}_{k|N} &= \sum_{i=1}^{r} \mu_{k|N}^{i} \hat{x}_{k|N}^{i} \\ \Sigma_{k|N} &= \sum_{j=1}^{r} \mu_{k|N}^{i} \left[\Sigma_{k|N}^{i} + \left(\hat{x}_{k|N}^{i} - \hat{x}_{k|N} \right) \left(\hat{x}_{k|N}^{i} - \hat{x}_{k|N} \right)^T \right] \end{aligned}$$

The summary of the IMM smoothing algorithm can be seen in Algorithm-5.2.3.

Algorithm 5.2.3 (IMM Smoother algorithm)

- *For each* $k = \{N-1, \ldots, 1\}$, $i, j = \{1, \ldots, r\}$
  - *Smoothed Mode Probability recursion*

$$\mu^i_{k|N} = \sum_{j=1}^{r} \mu^{i|j}_{k|k} \mu^j_{k+1|N}$$

  - *Smoothed Mixing Probability recursion*

$$\mu^{j|i}_{k+1|N} = \frac{\mu^{i|j}_{k|k} \mu^j_{k+1|N}}{\mu^i_{k|N}}$$

  - *Elemental smoother*

$$\left[\hat{x}^{ij}_{k|N}, \Sigma^{ij}_{k|N}\right] = KalmanSmoother\left(\hat{x}^i_{k|k}, \Sigma^i_{k|k}, \hat{x}^j_{k+1|N}, \Sigma^j_{k+1|N}, F^j, u^j, Q^j\right)$$

  - *Combination of Elemental Smoothed Estimation*

$$\left[\hat{x}^i_{k|N}, \Sigma^i_{k|N}\right] = Collapse\left(\hat{x}^{ij}_{k|N}, \Sigma^{ij}_{k|N}, \mu^{j|i}_{k+1|N}\right)$$

  - *The overall Smoothed Estimate*

$$\left[\hat{x}_{k|N}, \Sigma_{k|N}\right] = Collapse\left(\hat{x}^i_{k|N}, \Sigma^i_{k|N}, \mu^j_{k|N}\right)$$

- *End for*

Kalman Smoother function is defined as

$$\left[\hat{x}_{k|N}, \Sigma_{k|N}\right] = KalmanSmoother\left(\hat{x}_{k|k}, \Sigma_{k|k}, \hat{x}_{k+1|N}, \Sigma_{k+1|N}, F, u, Q\right)$$

$$\begin{aligned}
\hat{x}_{k+1|k} &\triangleq F\hat{x}_{k|k} + u \\
\Sigma_{k+1|k} &\triangleq F\Sigma_{k|k}F^T + Q \\
L_k &= \Sigma_{k|k}F^T\Sigma^{-1}_{k+1|k} \\
\hat{x}_{k|N} &= \hat{x}_{k|k} + L_k(\hat{x}_{k+1|N} - \hat{x}_{k+1|k}) \\
\Sigma_{k|N} &= \Sigma_{k|k} + L_k(\Sigma_{k+1|N} - \Sigma_{k+1|k})L^T_k
\end{aligned}$$

5.3 Experimental Methods and Results

5.3.1 Experimental Methods

In this work, we use the MOCHA database [103]. The acoustic data and EMA trajectories of one female talker (fsew0) that include 460 sentences are used. Audio features (Mel-frequency cepstral coefficients (MFCC)) were computed using a 36 ms window with 18 ms shift. The articulatory data are EMA trajectories, which are the X and Y coordinates of the lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum. EMA trajectories are normalized by the methods suggested in [64] and down-sampled to match the 18 ms shift rate. All the model parameters of JMLS are tested using 10-fold cross-validation. For each fold, nine tenths of the data (414 sentences) are used for training and one tenth (46 sentences) are used for testing. Cross-validation performance measures (RMS error and correlation coefficient) are computed as the average of all ten folds.

5.3.2 Hyperparameter Assessment

The parameters of the prior distributions are called hyperparameters, and must be specified *a priori* in order to solve articulatory inversion problem. As explained in Sec.5.2.1.2 the hyperparameter set of the proposed model is $\theta_h = \{\Omega_i, \Psi_i, \alpha_i, v_i\}$. In this work, we choose the following parameters for the ith joint prior distribution

$$\Omega_i^{-1} = \alpha_i \sum_{l=1}^{L} \sum_{k=2}^{N_l} \gamma_k^l(i) \mathbf{x}_{k-1}^l \mathbf{x}_{k-1}^{lT} \tag{5.47}$$

$$\Psi_i = \frac{1}{v_i} I_{n_x \times n_x} \tag{5.48}$$

where, $\mathbf{x_k}^l$ is the augmented articulatory state trajectories defined as $\mathbf{x_k}^l \triangleq \left[1, (x_k^l)^T\right]^T$ In this way, prior distribution becomes an invariant prior distribution [125]. The number of degrees of freedom of the inverse Wishart distribution v_i

is also fixed to the expected number of transitions from state i, that is,

$$v_i = \sum_{l=1}^{L} \sum_{k=2}^{N_l} \gamma_k^l(i) \tag{5.49}$$

Therefore, we have only one unknown hyperparameter namely $\{\alpha_i\}$. α_i is fixed to the same value α for all models and it is estimated via trial and error. In this work, we test the values $\alpha = \{0.1, 0.3, 0.5\}$.

5.3.3 Performance Measures

The performance of the algorithms is measured using three performance measures, namely, RMS error, normalized RMS error and correlation coefficient, all of which are described in Chapter 4.6.2.

5.3.4 Experimental Results

Experimental results of the proposed methods are given in this sub-section. The comparison of the learning methods based on ML and MAP criteria for JMLS in terms of RMS error and correlation coefficient can be seen in Fig.5.2. Examination of the figure shows that the MAP based learning method significantly improves the performance of the articulatory inversion. The performance of the MAP based algorithm is tested for various α values and it is observed that $\alpha = 0.3$ gives the best performance. The RMS error and the correlation coefficient between the true (measured) and the estimated articulatory trajectories for filtering mode are about 2.22 mm and 0.59 respectively for ML learning method while the corresponding results for MAP ($\alpha = 0.3$) based learning are about 1.67 mm and 0.72 respectively. That means Filtering RMS error of MAP based learning algorithm is less than about 24.3% (from 2.22 mm to 1.67 mm) compared to ML and it improves the correlation coefficients about 22% (from 0.59 to 0.72). When we consider inferences based on smoothed estimates, the RMS error and correlation coefficient for ML based learning method are about 1.88 mm and 0.69, corresponding results for MAP ($\alpha = 0.3$) based learning method are 1.58 mm and 0.76. That means MAP based learning algorithm reduces

RMS error about 16% (from 1.88 mm to 1.58 mm) and improve the correlation coefficients about 10% (from 0.69 to 0.76) in smoothing.

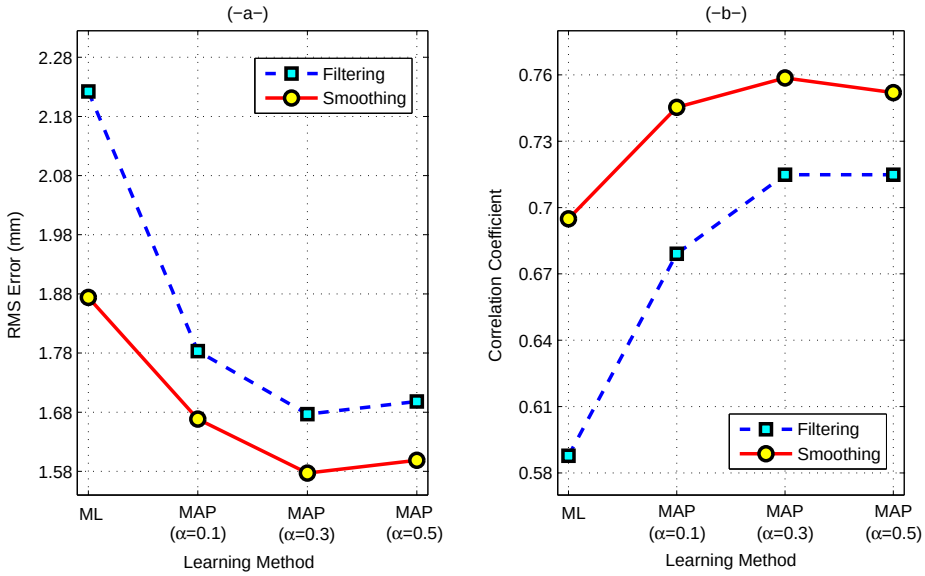


Figure 5.2: Two modal state case: RMS error (a) and correlation coefficient (b) between the true (measured) and the estimated articulatory trajectories for ML and MAP (with various α values) and the corresponding filtered and smoothed estimation results.

The second observation from Fig.5.2 is that, as expected, smoothing highly improves the performance compared to filtering. A similar result is reported in [67, 73] for articulatory inversion based on GMM and in [64] for articulatory inversion based on (TMDN). In our work smoothing reduces the RMSE from 2.22 mm to 1.88mm (a 15.7% relative improvement) and increases the correlation coefficient from 0.59 to 0.69 (a 17% relative improvement) when ML based learning is used. Similarly, smoothing inference for the MAP ($\alpha = 0.3$) based learning method reduces RMSE about 6% (from 1.68 mm to 1.58 mm) and improves correlation coefficient about 5.6% (from 0.72 to 0.76). The relatively small improvement in the MAP case shows that the prior densities used in MAP are sufficiently adequate to model the uncertainties of the system.

The RMS error and correlation coefficient for increasing number of JMLS modal-states can be seen in Fig.5.3. In this figure the global LDS performance is also given for comparison. Global LDS use only one model (that is, $s_k \in \mathcal{S} = \{1\}$) for

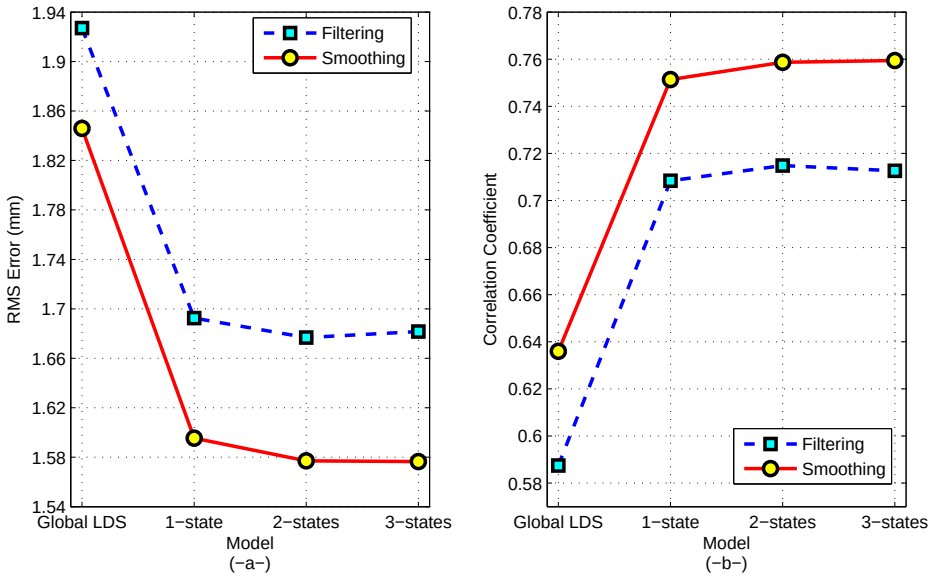


Figure 5.3: The MAP ($\alpha = 0.3$) based Learning: RMS error (a) and correlation coefficient (b) between the true (measured) and the estimated articulatory trajectories for increasing number of JMLS modal states and a global linear dynamic system. Both filtered and smoothed estimation results are given

all phonemes. The best result of RMSE and correlation coefficient for the global LDS are about 1.85 mm and 0.64 respectively. We compare our results with [131]. The approach and the experimental setup given in the referred work are similar to our work. RMSE and correlation coefficients reported in that study are about 2.15 mm and 0.59 respectively for global LDS. The outperformance of our study is due to the use of MAP based algorithm. This figure shows also the effect of increasing the number of modal states for each phoneme. '1-state' in this figure corresponds to one model for each phone, '2-state' corresponds to two models etc. Note that in Global LDS there is only one model for all phones. Increasing the number of modal states for each phoneme from one to two improves the performance of the JMLS based articulatory inversion, but the performance decreases slightly for three modal-states possibly due to insufficient training data.

Fig.5.4 provides more details regarding the utility of MAP based learning method in articulatory inversion. The abscissa distinguishes different articulators. As

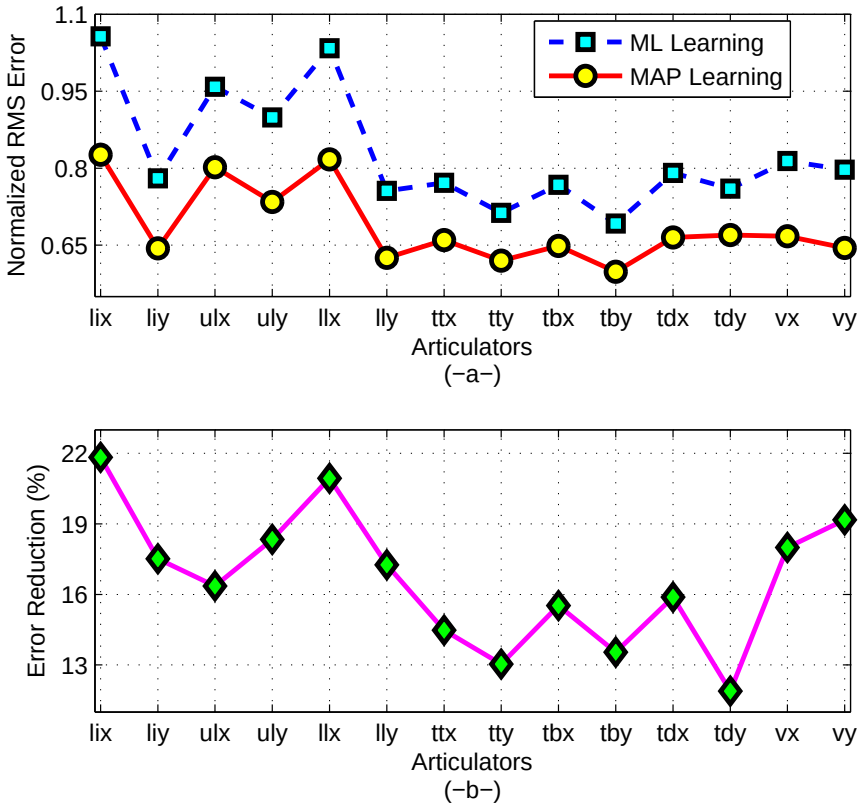


Figure 5.4: Normalized RMS errors for each articulator for ML and MAP ($\alpha = 0.3$) (a) and corresponding percentage normalized RMS error reductions of MAP ($\alpha = 0.3$) with respect to ML (b). The abbreviations li, ul,ll, tt,tb, td and v denote lower incisor, upper lip, lower lip, tongue tip, tongue body,tongue dorsum and velum, respectively. The suffixes x and y of the articulator abbreviations show the corresponding X and Y coordinates respectively.

an example, normalized RMS error for X axis of lower lip (llx) reduced from 1.04 to 0.82 (a 21% relative error reduction). Figure indicates that for all of the articulators there is a reduction in the estimation error. It can be said that the improvement is relatively small for the ones that are estimated by ML better. In general, this figure denotes that MAP based learning algorithm reduces normalized RMS about (11-22%). We presented an example of estimated and true trajectories to give an idea of how ML and MAP are performing. Fig.5.5 illustrates the estimated (based on ML and MAP ($\alpha = 0.3$)) and true x-coordinates of articulators for tongue tip and lower incisor. The utterances

are taken form MOCHA database.

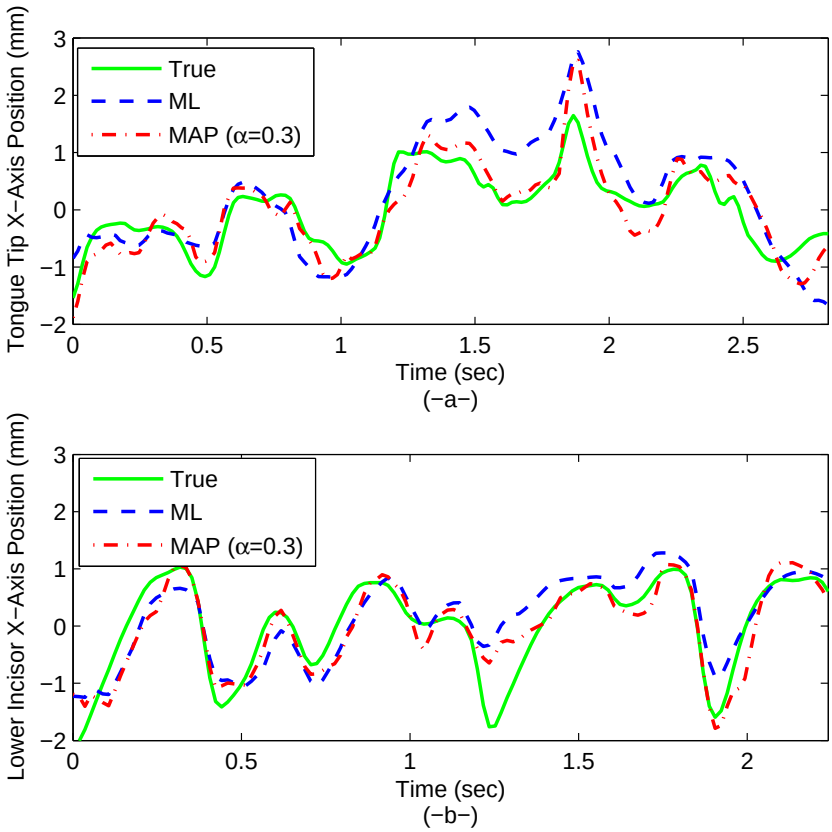


Figure 5.5: Estimated and true (measured) articulatory trajectories of x-coordinate for tongue tip (a) and lower incisor (b) for ML and MAP learning method. Data is taken from MOCHA database.

5.4 Conclusions

In this chapter of the thesis we have proposed a systematic framework for accurate estimation of articulatory trajectories from acoustic data based on jump Markov linear system (JMLS). JMLS can be considered as a generalized version of Hidden Markov model (HMM). Unlike the HMM, in which only discrete states are considered, JMLS models both discrete and continuous states simultaneously. That is, continuous articulatory trajectories are considered as the continuous hidden states of the JMLS, and discrete regimes (phonemes and subphoneme segments) are considered as the discrete modal states of the JMLS.

In that way, the acoustic to articulatory problem is converted into a state estimation problem for JMLS. In this study we first introduce a novel learning algorithm based on ML and MAP criteria that generates the model parameters of all linear systems used. This procedure can be considered as a generalization of Hidden Markov model (HMM) training. State estimation is done both online, by applying filtering methods, and off line, by applying smoothing methods. The IMM algorithm which is barrowed from tracking literature is used for filtering. This algorithm is adapted to articulatory inversion. Use of IMM in articulatory inversion problem is new. In addition to the filtering we present an efficient smoothing algorithm for the JMLS in Bayesian perspective. Smoothing algorithm that is used here is a novel one. In some applications (none of them are speech applications) GBP2 type of algorithms are modified to solve smoothing problem [128, 129]. The approach given here is different than GPB2. it is based on IMM inference method and is much more efficient from computational point of view.

Experiments have been conducted in the MOCHA database to test the performance of the proposed algorithms and the results are summarized in the previous section. Here, as a final conclusion we compare our work with two recent studies existing in the literature and the method that we have proposed in Chapter 4.

First we compare our results with [131] which uses HMM based switching linear dynamic system. The reason of selection of this paper is the similarity between two works. [131] also uses state space representation. There is one model for each phone and [131] assumes that phone boundaries are known. We do not use this restricting assumption; however we use a more sophisticated model as JMLS. Best RMSE and correlation coefficient results in [131] are 1.78 mm and 0.7. The inversion based on HMM is given in [72] and their RMSE and correlation coefficient results are about 1.64 mm and 0.725 (using only audio data) respectively. According to experimental results given in Sec.5.3, our best results is obtained using a MAP based learning algorithm with smoothing state estimate. The RMS error and correlation coefficient between the true (measured) and estimated articulatory trajectories are about 1.58 mm and 0.76 respectively.

In Chapter 4 of this theses a quite satisfactory inversion system, GMM regression, is proposed. The performance of GMM regression can be summarized as the RMS error 1.395 mm and correlation coefficient 0.81. As numbers indicate performance of GMM regression is better compared to JMLS formulation. In spite of these results we think that JMLS is much more flexible and open to improvements. The linear dynamic systems used in JMLS formulation can be improved in lots of different ways that correspond to better representation of the movements of the articulators. Even nonlinear state models are possible. Further improvements by model improvement are a future work.

CHAPTER 6

CONCLUSION

This thesis presents the research done on the estimation of the underlying slowly-varying continuous dynamic structure of speech signal. While speech signal is considered as a rapidly changing and quasi-periodic random signal, speech production mechanism is governed by vocal organs (such as, lip, jaw, tongue, etc.,) having highly constrained, quite continuous and relatively slow dynamic movements. During speech utterance, mainly two types of continuous structures change simultaneously in speech production mechanism. These are formant frequencies and positions of the articulators. Indeed, these quantities are quite related to each other. Movement of the articulators and formation of resonance frequencies according to the sound to be produced are main parts of the speech production mechanism and hence, they carry useful information about the speech signal, as well as the speaker. We propose novel methods for the estimation of these continuous quantities of the underlying speech signal. The main contributions of the thesis can be summarized as follows.

6.1 Contributions to Formant tracking

Formant tracking is studied in Chapter 2 and Chapter 3. Chapter 2 of the thesis is about tracking of fixed number of formants that are assumed to be continuous along an utterance. Since this is a well studied subject in the literature, our aim is to contribute to the performance. For this purpose one of the well-known methods of the literature, i.e., the one that uses dynamic programming (DP) [36, 37, 38] approach is selected. To achieve the desirable increase

in the performance formant tracking problem is divided into two sub-problems: the data association (which formant candidate belongs to which formant trajectory) and the state estimation. Data association problem is handled via DP method and the state estimation is done by Kalman filtering/smoothing. The experimental results show that the combination of DP and Kalman smoothing gives better performance compared to formant tracking using only DP algorithm. The proposed method is also compared to other methods given in the literature. It is shown that the proposed algorithm is significantly better than other algorithms in voiced like speech regions, where formant frequencies can be easily seen in speech spectrogram and its performance is comparable in unvoiced speech regions. During this study it is observed that continuity of the formants is not a reality but a commonly accepted idealization of the behavior of formants. This is also a reason for the relative degradation in the performance when formants of an unvoiced region are tracked. In Chapter 3 variable number of formant tracking idea is introduced. There are few studies in the literature that assume irregularities in the formant trajectories including changes in their number. We verified this observation partially by using a state space model for the acoustic tube. A modification is done on the model by changing the excitation point of the tube according to the constriction point for obstruent sounds like, *s,p,t,k*. Based on the observations and theoretical justification on changing number of formants we have developed an algorithm to track unknown and changing number of formants with a novel approach, which is inspired from multi-target tracking literature. In that way, formant frequencies are tracked in a flexible manner without pre-defined formant number or the assumption of continuity of formants along a given utterance. As a summary the following are the contributions of formant tracking part.

- The improvement of the performance of dynamic programming based tracking algorithm. This improvement is achieved by Kalman filtering/smoothing of the output of DP.
- The development of the state space representation of concatenated tube model and use it to explain the possibility of variable number of formants in speech signal. We believe that state space model can be used in vari-

ous speech applications including articulatory speech synthesis and voice conversion, etc.

- A novel algorithm for tracking variable number of formant frequencies.

6.2 Contributions in acoustic to articulatory inversion

Acoustic to articulatory inversion problem can be considered as one of the basic problems of speech so can be dated back to a several decades. However after the introduction of the UW-XRMB [132] and MOCHA [103] databases, which contain the positions of articulators as well as speech signal, there is a considerable increase in the related research. Our aim in this thesis is to improve the already existing methods as well as introducing some novel approaches to the subject. Several variations of 'GMM regression' method are developed to improve its performance. These studies are the subject of Chapter 4. Firstly the MFCC coefficients that are used as the input of the GMM regression method are enriched by augmenting them with formant frequencies as well as their energies and visual shape and texture features. These additional features are selected after an exhaustive search among other possibilities like LPC, LSF coefficients, etc. It is observed that the use of augmented features improves the performance of articulatory inversion. Besides enriching the input of the inversion block, its output is also changed by smoothing it by Kalman smoother. The proposed smoother has the capability of learning its own parameter set from training data and incorporating extra information such as phonetic transcription if it is available. The smoother algorithm developed in this study is one of the contributions of the thesis. The experimental results show that the proposed smoother significantly improves the performance of the articulatory inversion based on the GMM regression. Furthermore the proposed smoothing algorithm is a general smoothing algorithm can be used in a large group of inversion algorithms such as HMM [72, 68], SVM [101], or TMDN [65]. Feature vector augmentation can be considered as early fusion of acoustic and visual features. Late fusion, defined here as the fusion of inversion results, is also investigated in Chapter 4. The dynamic fusion algorithm (known as 'distributed fusion' in target tracking

in multi-sensors applications) is adapted to combine audio and visual estimates for articulatory inversion. Comparison of the early and late fusions gives the somewhat surprising result in the favor of late fusion. We believe that this is due to the insufficiency of the training data. To the best of our knowledge, the best performance that we have obtained about the articulatory inversion when compared with the existing work in the literature using any method and the same database is superior to all. A novel approach for inversion problem is introduced in Chapter 5. In this approach we model the articulatory data as the state of a Jump Markov Linear System (JMLS). The estimation of the states of JMLS gives us time variation of the positions of the articulators. State estimation requires measurements to estimate the state. The measurements used here are the acoustic features. This approach requires the estimation of the parameters of the JMLS. In the parameter estimation stage we use two different approaches, ML and MAP, which are generalization of Hidden Markov Model (HMM) training. The state estimation is performed via the interactive multiple model (IMM) algorithm. An IMM smoother is proposed to estimate the state. Since the approach used here is completely new to articulatory inversion literature, use of IMM is also new. Furthermore even in the tracking literature although IMM smoothers are used, the one that we developed here is somewhat different than the existing ones. In the experiments section, the two training algorithms ML and MAP are compared with each other and the outperformance of MAP is observed. This is not surprising since the training data is not sufficient for ML to converge. One other reason may be the incapability of the state and observation model used to track articulatory movements. The overall performance of this new approach is better than the performances of the many systems given in the literature; however it is slightly worse than our system that is using improved acoustic and visual features together described in Chapter 4. As a summary the following are the contributions of articulatory inversion part.

- The performance of articulatory inversion based on Gaussian Mixture Model (GMM) regression is improved by
  - improving the feature vectors,
  - applying smoothing to its output

  - dynamic fusing the outputs of a visual inverter and an acoustic one.

- A new approach of using Jump Markov linear system model (JMLS) for articulatory inversion is introduced.
  - The novel learning algorithms based on ML and MAP criterion for JMLS is presented.
  - An efficient smoothing algorithm for JMLS is introduced.

REFERENCES

[1] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP*, 2006, pp. 369–372.

[2] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton and Co., 1960.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 34–45, Mar. 1960.

[4] S. Gannot., D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 6, no. 4, pp. 373–385, Jul 1998.

[5] K. Y. Lee and S. Jung, "Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 282–291, May 2000.

[6] S. Subasingha, M. N. Murthi, and S. V. Andersen, "Gaussian mixture Kalman predictive coding of line spectral frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 379–391, Feb. 2009.

[7] T. Ramabadran and D. Sinha, "Speech data compression through sparse coding of innovations," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 274–284, Apr 1994.

[8] J. Z. Ma and L. Deng, "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 590–602, Nov. 2003.

[9] J. Deng, M. Bouchard, and T. H. Yeap, "Linear dynamic models with mixture of experts architecture for recognition of speech under additive noise conditions," *Signal Processing Letters, IEEE*, vol. 13, no. 9, pp. 573–576, Sept. 2006.

[10] J. Ma and L. Deng, "A mixed-level switching dynamic system for continuous speech recognition," *Computer Speech and Language*, vol. 18, pp. 49–65, 2004.

[11] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, and I. Andrianakis, "Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing," *Computer Speech and Language*, vol. 21, no. 3, pp. 543–561, 2007.

[12] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman-filter," in *Proc. ICASSP*, 1986, pp. 1229–1232.

[13] R. Togneri and L. Deng, "A state-space model with neural-network prediction for recovering vocal tract resonances in fluent speech from mel-cepstral coefficients," *Speech Communication*, vol. 48, pp. 971–988, 2006.

[14] L. D. S. Dusan, "Acoustic-to-articulatory inversion using dynamic and phonological constraints."

[15] G. Fant, Ed., *Acoustic Theory of Speech Production*. Mouton, 1960.

[16] J. L. Flanagan, Ed., *Speech analysis synthesis and perception*. Springer Verlang, 1965.

[17] L. R. Rabiner, "Digital-formant synthesizer for speech-synthesis studies," *J. Acoust. Soc. Am.*, vol. 43, no. 4, pp. 822–828, 1968.

[18] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[19] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 9, pp. 39–51, 2001.

[20] J. Bellegarda, "A global, boundary-centric framework for unit selectiontext-to-speech synthesis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 14, no. 4, pp. 990–997, 2006.

[21] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 6, pp. 36–48, 1998.

[22] P. Garner and W. Holmes, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition," in *Proc. ICASSP*, 1998.

[23] D. Rentzos, S. Vaseghi, Q. Yan, and C. Ho, "Parametric formant modeling and transformation in voice conversion," *International Journal of Speech Technology Springer*, vol. 8, pp. 227–245, 2005.

[24] T. Claes, I. Dologlou, L. ten Bosch, and D. van Compernolle, "A novel feature transformation for vocal tract length normalizationin automatic speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 6, pp. 549–557, 1998.

[25] R. S. S. Umesh, "A study of filter bank smoothing in mfcc features for recognition of children's speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 15, no. 8, pp. 2418–2430, 2007.

[26] A. Watanabe and T. Sakata, "Reliable methods for estimating relative vocal tract lengths from formant trajectories of common words," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 14, no. 4, pp. 1193–1204, 2006.

[27] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of british, australian, and american accents," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 15, pp. 676–689, 2007.

[28] A. Potamianosa and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3795 –3806, 1996.

[29] A. Watanabe, "Formant estimation method using inverse-filter control," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 9, no. 4, pp. 317–326, 2001.

[30] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 8, no. 1, pp. 240–254, 2000.

[31] M. Kamran and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 14, no. 2, pp. 435–444, 2006.

[32] J. Vargas and S. McLaughlin, "Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 16, no. 1, pp. 1–7, 2008.

[33] S. A. Fattah, W. Zhu, and M.0.Ahmad, "An approach to formant frequency estimation at low signal-to-noise ratio," in *Proc. ICASSP*, 2007.

[34] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[35] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 709–729, 1986.

[36] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Am.*, vol. 1, no. 6, p. 55, 1987.

[37] K. Xia and C. Espy-Wilson, "A new strategy of format tracking using dynamic programming," in *Proc. Internat. Conf. Spoken Language Processing*, 2000, pp. 1–4.

[38] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 13, no. 5, pp. 240–254, 2005.

[39] Y. Laprie and M. Berger, "A new paradigm for reliable automatic formant tracking," in *Proc. ICASSP*, 1994, pp. 201–205.

[40] P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri, "Bayesian modelling of the speech spectrum using mixture of Gaussians," in *Proc. ICASSP*, 2004, pp. 556–559.

[41] P. Zolfaghari, H. Kato, Y. Minami, A. Nakamura, and S. Katagiri, "Dynamic assignment of Gaussian components in modeling speech spectra," *The Journal of VLSI Signal Processing*, vol. 45, no. 1-2, pp. 7–19, 2006.

[42] D. T. Toledano, J. G. Villardebó, and L. H. Gómez, "Initialization, training, and context-dependency in HMM-based formant tracking," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 14, no. 2, pp. 511–523, 2006.

[43] G. Rigoll, "Formant tracking with quasilinearization," in *Proc. ICASSP*, 1988, pp. 307–310.

[44] M. Niranjan and I. Cox, "Recursive tracking of formants in speech signals," in *Proc. ICASSP*, 1994, pp. 205–208.

[45] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded inatemporal constraint," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 14, no. 2, pp. 425–434, 2006.

[46] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 15, no. 1, pp. 13–23, 2007.

[47] D. D. Rudoy and P. J. Wolfe, "Conditionally linear Gaussian models for tracking of vocal tract resonances," in *Proc. Interspeech*, 2007.

[48] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. ICASSP*, 2004, pp. 565–568.

[49] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, and I. Andrianakis, "Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing," *Computer Speech and Language*, vol. 21, no. 3, pp. 543–561, 2007.

[50] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA: Artech House, 1999.

[51] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques, and Software*. Norwell, MA: Artech House, 1993.

[52] İ. Y. Özbek and M. Demirekler, "Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval Kalman smoother," in *ICASSP*, 2008.

[53] ——, "Tracking of vocal tract resonances based on dynamic programming and Kalman filtering," in *SİU*, 2008.

[54] ——, "Tracking of speech formant frequencies," in *SİU*, 2006.

[55] ——, "Tracking of visible vocal tract resonances (VVTR) based on Kalman filtering," in *INTERSPEECH*, 2006.

[56] M. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, vol. 41, no. 4, pp. 1002–1010, 1967.

[57] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[58] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[59] 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," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[60] M. G. Rahim, W. B. Keijn, J. Schroeter, and C. C. Goodyear, "Acoustic to articulatory parameter mapping using an assembly of neural networks," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, Apr 1991, pp. 485–488 vol.1.

[61] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 133–150, Jan 1994.

[62] J. S. Perkell, M. H. Cohen, M. A. Svirsky, and M. L. Matthies, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am.*, vol. 92, no. 6, pp. 3078–3096, 1992.

[63] A. A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. Fifth Seminar on Speech Production: Models and Data*, Kloster Seeon, Germany, 2000, pp. 305–308.

[64] K. Richmond, "Estimating articulatory parameters from the speech signal," Ph.D. dissertation, The Center for Speech Technology Research, Edinburgh, KU, 2002.

[65] ——, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *INTERSPEECH*, 2006.

[66] C.Qin and M. Carreira-Perpin, "A comparison of acoustic features for articulatory inversion," in *INTERSPEECH*, 2007.

[67] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215–227, 2008.

[68] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 12, no. 2, pp. 175–185, March 2004.

[69] A. Toutios and K. Margaritis, "Contribution to statistical acoustic-to-EMA mapping," in *EUSIPCO*, 2008.

[70] V. Mitra, İ. Y. Özbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustic to vocal tract time functions," in *ICASSP*, 2009.

[71] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Communication*, vol. 51, no. 3, pp. 195–209, 2009.

[72] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 17, no. 3, pp. 411–422, 2009.

[73] İ. Y. Özbek, M. Hasegawa-Johnson, and M. Demirekler, "Formant trajectories for acoustic-to-articulatory inversion," in *INTERSPEECH*, 2009.

[74] R. E. Helmick, W. D. Blair, and S. A. Hoffman, "Fixed-interval smoothing for Markovian switching systems," *Information Theory, IEEE Transactions on*, vol. 41, no. 6, pp. 1845–1855, Nov 1995.

[75] Y. Bar-Shalom and X. R. Li, *Multitarget-multisensor tracking: principles and techniques*. YBS, 1995.

[76] J. Darch, B. Milnerand, and S. Vaseghi, "Map prediction of formant frequencies and voicing class from mfcc vectors in noise," *Speech Communication*, vol. 48, no. 11, pp. 1556–1572, 2006.

[77] C. Glaisert, M. Heckmann, E. Joublin, and C. Goerick, "Joint estimation of formant trajectories via specto-temporal smoothing and Bayesian techniques," in *Proc. ICASSP*, 2007, pp. 477–480.

[78] C. Myers and L. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 351–363, 1981.

[79] M. Sharma and R. Mammone, "Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge," in *Proc. Internat. Conf. Spoken Language Processing*, 1996, pp. 1237–1240.

[80] L. R. Rabiner and R. W. Schafer, *Digital Processing Of Speech Signals*. Prentice-Hall, 1978.

[81] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[82] K. Sjolander and J. Beskow. Tmh kth : Wavesurfer. [Online]. Available: http://www.speech.kth.se/wavesurfer/download.html, last visited on 29 April 2010.

[83] H. M. Sussman, "An investigation of locus equations as a source of relational invariance for stop place categorization," *The Journal of the Acoustical Society of America*, vol. 89, no. 4B, p. 1998, 1991.

[84] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[85] J. R. Deller, Jr., J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. IEEE Press, 2000.

[86] M. T.-T. Jackson, C. Espy-Wilson, and S. E. Boyce, "Verifying a vocal tract model with a closed side-branch," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2983–2987, 2001.

[87] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002.

[88] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Am.*, vol. 33, no. 5, pp. 589–596, 1961.

[89] O. Salor, P. L. Bryan, T. Ciloglu, and M. Demirekler, "Turkish speech corpora and recognition tools developed by porting sonic towards multilingual speech recognition." *Comput. Speech Lang.*, vol. 21, no. 4, pp. 580–593, 2007.

[90] J. Wells. Sampa for turkish. [Online]. Available: http://www.phon.ucl.ac.uk/home/sampa/turkish.htm, last visited on 29 April 2010.

[91] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. Internat. Conf. Spoken Language Processing*, 2000, pp. 145–8.

[92] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Communication*, vol. 48, pp. 161–175, 2006.

[93] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables," in *Proc. Internat. Conf. Spoken Language Processing*, 2000, pp. 951–954.

[94] T. A. Stephenson, M. Magimai-Doss, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 189–203, 2004.

[95] N. Katsamanis, T. Roussos, G. Papandreou, and P. Maragos. Computer vision, speech communication & signal processing group. Downloaded data: AAM-based visual features for the female speaker fsew0 of the MOCHA database. [Online]. Available: http://cvsp.cs.ntua.gr/research/inversion/, last visited on 29 April 2010.

[96] M. Hasegawa-Johnson, "Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 457–460, 2000.

[97] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*. New York, NY: John Wiley and Sons, 1995.

[98] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, p. 535, 1975.

[99] E. Özkan, İ. Y. Özbek, and M. Demirekler, "Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying dirichlet process mixture models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1518–1532, Nov. 2009.

[100] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc*, 1977.

[101] K. M. A. Toutios, "Contribution to statistical acoustic-to-EMA mapping," in *EUSIPCO*, 2008.

[102] L. Deng, *Dynamic Speech Models: Theory, Algorithms, and Applications*. Morgan & Claypool Publishers, 2006.

[103] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *In Proc. 5th Seminar on Speech Production*, 2000. [Online]. Available: http://www.cstr.ed.ac.uk/artic, last visited on 29 April 2010.

[104] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[105] F. Alipour-Haghigi and I. R. Titze, "Twitch response in the canine vocalis muscle," *J. Speech Hear. Res.*, vol. 30, pp. 290–294, 1987.

[106] ——, "Tetanic contraction in vocal fold muscle," *J. Speech Hear. Res.*, vol. 32, pp. 226–231, 1989.

[107] R. Wilhelms-Tricarico, "Physiological modeling of speech production: Methods for modeling soft-tissue articulators," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3085–3098, 1995.

[108] P. Perrier, D. J. Ostry, and R. Laboissière, "The equilibrium point hypothesis and its application to speech motor control," *J. Speech Hear. Res.*, vol. 39, pp. 365–378, 1996.

[109] H. Gomi and M. Kawato, "Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement," *Science*, vol. 272, pp. 117–120, 1996.

[110] T. Chiba and M. Kajiyama, *The vowel, its nature and structure*. Tokyo: TokyoKaiseikan Pub. Co., 1941.

[111] V. N. Sorokin, A. S. Leonov, and A. V. Trushkin, "Estimation of stability and accuracy of inverse problem solution for the vocal tract," *Speech Communication*, vol. 30, pp. 55–74, 2000.

[112] J. Frankel, K. Richmond, S. King, and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," in *Proc. Internat. Conf. Spoken Language Processing*, 2000, pp. 254–7.

[113] W. Wu, "A nonlinear IMM algorithm for maneuvering target tracking," *IEEE Trans. Aerospace and Electronic Systems*, vol. 30, no. 3, pp. 875–885, 1994.

[114] I. Potamitis and N. Fakotakis, "Tracking and voice separation of moving speakers based on IMM-PDA filters," in *Internat. Conf. Information Fusion*, 2004.

[115] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. ICASSP*, 2004, pp. 889–92.

[116] J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *Proc. EUROSPEECH*, 2003, pp. 1441–1444.

[117] M. Fujimoto and S. Nakamura, "Sequential non-stationary noise tracking using particle filtering with switching dynamical system," in *Proc. ICASSP*, 2006, pp. 769–772.

[118] M. Fujimoto, K. Ishizuka, and H. Kato, "Noise robust voice activity detection based on statistical model and parallel non-linear Kalman filtering," in *Proc. ICASSP*, 2007.

[119] M. A. Gandhi and M. A. Hasegawa-Johnson, "Source separation using particle filters," in *Proc. Interspeech*, Jeju Island, Korea, 2004, pp. 2673–6.

[120] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[121] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Am. Math. Soc.*, vol. 73, pp. 360–363, 1967.

[122] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[123] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[124] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[125] T. P. Minka, "Bayesian linear regression," 3594 Security Ticket Control, Tech. Rep., 1999.

[126] D. B. Rowe, *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. New York: CRC Press Company, 2003.

[127] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.

[128] C. Kim, "Dynamic linear models with Markov-switching," York (Canada) - Department of Economics, Working Papers, 1991.

[129] K. P. Murphy, "Switching Kalman filters," DEC/Compaq Cambridge Research Labs, Tech. Rep., 1998.

[130] H. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *American Institute of Aeronautics and Astronautics*, vol. 3, pp. 1445–1450, Aug 1965.

[131] A. Katsamanis, G. Ananthakrishnan, G. Papandreou, P. Maragos, and O. Engwall, "Audiovisual speech inversion by switching dynamical modeling governed by a hidden markov process," in *Proc. European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, Aug. 2008*, 2008.

[132] J. R. Westbury, "The signifcance and measurement of head position during speech production experiments using the x-ray microbeam system," *J. Acoust. Soc. Am.*, pp. 1782–1791, 1991.

APPENDIX A

ACOUSTIC TUBE MODEL

A.1 Acoustic Tube Model Based State Space Representation of LPC Filter

$$\begin{pmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_{n-1}(k+1) \\ x_n(k+1) \end{pmatrix} = \begin{pmatrix} -\rho_1\rho g & -\rho_1\rho g(1-\rho_1) & \dots & -\rho_{n-1}\rho_g \prod_{i=1}^{n-2}(1-\rho_i) & -\rho_l\rho_g \prod_{i=1}^{n-1}(1-\rho_i) \\ (1+\rho_1) & -\rho_2\rho_1 & \dots & -\rho_{n-1}\rho_1 \prod_{i=2}^{n-2}(1-\rho_i) & -\rho_l\rho_1 \prod_{i=2}^{n-1}(1-\rho_i) \\ 0 & (1+\rho_2) & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & -\rho_{n-1}\rho_{n-2} & -\rho_l\rho_{n-2}(1-\rho_{n-1}) \\ 0 & 0 & 0 & (1+\rho_{n-1}) & -\rho_l\rho_{n-1} \end{pmatrix} \times \begin{pmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_{n-1}(k) \\ x_n(k) \end{pmatrix} + \begin{pmatrix} \frac{1+\rho_g}{2} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} u(k) \tag{A.1}$$

$$s(k) = \begin{pmatrix} 0 & 0 & \dots & 0 & 1+\rho_l \end{pmatrix} \begin{pmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_{n-1}(k) \\ x_n(k) \end{pmatrix} \tag{A.2}$$

A.2 Sub-matrices in the Acoustic Tube Model Based State Space Representation of LPC Filter for supra-glotal excitation

$$A_{11} = \begin{pmatrix} -\rho_1\rho_g & -\rho_2\rho_g(1-\rho_1) & \dots & \rho_g\prod_{i=1}^{n-2}(1-\rho_i) & -2\rho_g\prod_{i=1}^{n-2}(1-\rho_i) \\ (1+\rho_1) & -\rho_1\rho_2 & \dots & \rho_1\prod_{i=2}^{n-2}(1-\rho_i) & -2\rho_1\prod_{i=2}^{n-2}(1-\rho_i) \\ 0 & \ddots & & \vdots & \vdots \\ 0 & 0 & & \rho_{m-2} & -2\rho_{m-2} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \tag{A.3}$$

$$A_{22} = \begin{pmatrix} -\rho_{m+1} & -\rho_{m+2}(1-\rho_{m+1}) & \dots & \rho_{n-1}\prod_{m+1}^{n-2}(1-\rho_i) & -\rho_l\prod_{m+1}^{n-1}(1-\rho_i) \\ (1+\rho_{m+1}) & -\rho_{m+1}\rho_{m+2} & \dots & -\rho_{n-1}\rho_{m+1}\prod_{m+2}^{n-2}(1-\rho_i) & -\rho_l\rho_{m+1}\prod_{m+2}^{n-1}(1-\rho_i) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -\rho_{n-2}\rho_{n-1} & -\rho_l\rho_{n-2}(1-\rho_{n-1}) \\ 0 & 0 & \dots & 1+\rho_{n-1} & -\rho_l\rho_{n-1} \end{pmatrix} \tag{A.4}$$

$$A_{21} = \begin{pmatrix} 0 & \dots & 0 & 2 \\ \vdots & \ddots & \vdots & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix}_{(n-m)\times m} \qquad e_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^T \quad e_2 = \begin{pmatrix} 0 & \dots & 0 & 1 \end{pmatrix} \tag{A.5}$$

APPENDIX B

IMM SMOOTHER

B.1 Proof of Equation-(5.43)

$$I_2 = \frac{p(x_{k+1}|x_k, s_{k+1}=j)p(x_k|s_k=i, z_{1:k})}{\int dx_k p(x_{k+1}|x_k, s_{k+1}=j)p(x_k|s_k=i, z_{1:k})} \tag{B.1}$$

The Gaussian densities $p(x_{k+1}|x_k, s_{k+1}=j)$ and $p(x_k|s_k=i, z_{1:k})$ can be written as

$$p(x_{k+1}|x_k, s_{k+1}=j) \sim \mathcal{N}(x_{k+1}, F_j x_k, Q_j), \quad p(x_k|s_k=i, z_{1:k}) \sim \mathcal{N}(x_k, \hat{x}^i_{k|k}, \hat{\Sigma}^i_{k|k})$$

After Computing (B.1), I_2 can be found as

$$I_2 = \mathcal{N}(x_k, L^{ij}_k x_{k+1} + c^{ij}_k, G^{ij}_k) \tag{B.2}$$

where,

$$L^{ij}_k \triangleq \Sigma^i_{k|k} F_j^T (\Sigma^{ij}_{k+1|k})^{-1}, \quad c^{ij}_k \triangleq \hat{x}^i_{k|k} - L^{ij}_k F_j \hat{x}^i_{k|k}, \quad G^{ij}_k \triangleq \Sigma^i_{k|k} - L^{ij}_k F_j \Sigma^i_{k|k}$$

I_3 is approximated as follows

$$I_3 \triangleq p(x_{k+1}|s_{k+1}=j, s_k=i, z_{1:N}) \approx p(x_{k+1}|s_{k+1}=j, z_{1:N}) \tag{B.3}$$

Suppose that $p(x_{k+1}|s_{k+1}=j, z_{1:N})$ is known and given as

$$p(x_{k+1}|s_{k+1}=j, z_{1:N}) \sim \mathcal{N}(x_{k+1}; \hat{x}^j_{k+1|N}, \hat{\Sigma}^j_{k+1|N}) \tag{B.4}$$

then, the integration given in the (5.38) can be approximated by using (B.2) and (B.4)

$$\begin{aligned}\int dx_{k+1} I_2 \times I_3 &\approx \int dx_{k+1} \mathcal{N}(x_k, L^{ij}_k x_{k+1} + c^{ij}_k, G^{ij}_k) \mathcal{N}(x_{k+1}; \hat{x}^j_{k+1|N}, \hat{\Sigma}^j_{k+1|N}) \\ &= \mathcal{N}(x_{k+1}; \hat{x}^{ij}_{k|N}, \hat{\Sigma}^{ij}_{k|N}) \end{aligned} \tag{B.5}$$

The definition of $\hat{x}^{ij}_{k|N}$ and $\hat{\Sigma}^{ij}_{k|N}$ are given in the (5.44) and (5.45) respectively.

VITA

PERSONAL INFORMATION

Surname, Name: Özbek, İ. Yücel

Nationality: Turkish (TC)

Date and Place of Birth: 04 Aug. 1975 , Oltu/Erzurum

Marital Status: Married

email: iozbek@metu.edu.tr and iyucelozbek@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
MS	METU Electrical and Electronics Engineering	2002
BS	Erciyes University Electrical and Electronics Engineering	1998
High School	Erzurum Lisesi, Erzurum	1994

WORK EXPERIENCE

Year	Institution
2000-2010	METU Electrical and Electronics Engineering, Research and Teaching Assistant
2006-2007	ASELSAN Military Electronics, Multi-Sensor Radar Data Fusion -I-, Researcher in the project
2007-2008	ASELSAN Military Electronics, Multi-Sensor Radar Data Fusion -II-, Researcher in the project
2008-2009	The Scientifc and Technological Research Council of Turkey (TUBITAK), Dynamic Speech Spectrum Representation and Tracking Variable Number of Vocal Tract Resonance Frequencies With Time-Varying Dirichlet Process Mixture Models, Researcher in the project. Project No: 108E097
2008-2009	Investigation of the Distinctive Properties of Turkish Fricatives and Plosives, Researcher in the project. Project No: 107E107

Publications:

- E. Özkan and **İ. Y. Özbek** and M. Demirekler, "Dynamic Speech Spectrum Representation and Tracking Variable Number of Vocal Tract Resonance Frequencies With Time-Varying Dirichlet Process Mixture Models" *IEEE Trans. Audio Speech Lang. Process.*, vol.17, no.8, pp. 1518-1532, Nov. 2009
- **İ. Y. Özbek** and M. Demirekler,"Audiovisual Articulatorty Inversion Based on Gaussian Mixture Model (GMM)" in *in Proc. SİU*, 2010
- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler,"Formant Trajectories for Acoustic-to-Articulatory Inversion" in *in Proc. Interspeech*, 2009
- V. Mitra and **İ. Y. Özbek** and H. Nam and X. Zhou and C. Espy-Wilson,"From Acoustic to Vocal Tract Time Functions" in *in Proc. ICASSP*, 2009

- **İ. Y. Özbek** and M. Demirekler,"Vocal Tract Resonances Tracking Based on Voiced and Unvoiced Speech Classification Using Dynamic Programming and Fixed Interval Kalman Smoother" in *in Proc. ICASSP*, 2008

- **İ. Y. Özbek** and M. Demirekler,"Tracking of vocal tract resonances based on dynamic programming and Kalman filtering" in *in Proc. SİU*, 2008

- M. Demirekler, E. Özkan, **İ. Y. Özbek** , M. Günay ve B. Özer, "Dağıtık mimarili savunma sistemlerinde füsyon katmanlarının birleştirilerek izleme başarımının iyileştirilmesi" in *in Proc. SAVTEK*, 2008

- **İ. Y. Özbek** and M. Demirekler,"Tracking of Visible Vocal Tract Resonances (VVTR) Based on Kalman Filtering" in *in Proc. Interspeech*, 2006

- **İ. Y. Özbek** and M. Demirekler,"Tracking of Speech Formant Frequencies" in *in Proc. SİU*, 2006

- **İ. Y. Özbek** , M. Demirekler and T. Çiloğlu, "LBDP Speech Segmentation based on Phoneme Duration Modelling." in *in Proc. SİU*, 2003

- **İ. Y. Özbek** , U. Orguner, T. Çiloğlu, K. Leblebicioğlu and M. Demirekler, "Genetic Algorithm for Speech Segmentation." in *in Proc. SİU*, 2003

- **İ. Y. Özbek** , U. Orguner,Ö. Salor, M. Demirekler and T. Çiloğlu, "Automatic Isolated Speech Segmentation." in *in Proc. SİU*, 2002

- **İ. Y. Özbek** "Speech Segmentation and Speech Database Annotation.", M.S. Thesis, Middle East Technical University, Ankara, Turkey, September 2002.

Submitted:

- **İ. Y. Özbek** and M. Demirekler,"ML vs. MAP Parameter Estimation of Linear Dynamic System for Acoustic-to-Articulatory Inversion: A Comparative Study" Submitted to *in Proc. EUSIPCO*, 2010

- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler, "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing" Submitted to *IEEE Trans. Audio Speech Lang. Process.,*

To be submitted:

- **İ. Y. Özbek** and M. Hasegawa-Johnson and M. Demirekler, "On Improving Dynamic State Space Approaches to Articulatory Inversion with MAP based Parameter Estimation" To be submitted to *IEEE Trans. Audio Speech Lang. Process.,*