BIMODAL AUTOMATIC SPEECH SEGMENTATION AND BOUNDARY REFINEMENT TECHNIQUES

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

BY

EREN AKDEMİR

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

MARCH 2010

Approval of the thesis:

BIMODAL AUTOMATIC SPEECH SEGMENTATION AND BOUNDARY REFINEMENT TECHNIQUES

submitted by **EREN AKDEMİR** in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical and Electronics Engineering 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 Engineering** _____

Assoc. Prof. Dr. Tolga Çiloğlu
Supervisor, **Electrical and Electronics Engineering Department** _____

Examining Committee Members:

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

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

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

Prof. Dr. Aydın Alatan
Electrical and Electronics Engineering Dept., METU _____

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

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: EREN AKDEMİR

Signature :

ABSTRACT

BIMODAL AUTOMATIC SPEECH SEGMENTATION AND BOUNDARY REFINEMENT TECHNIQUES

Akdemir, Eren
Ph. D., Department of Electrical and Electronics Engineering
Supervisor : Assoc. Prof. Dr. Tolga Çiloğlu

March 2010, 130 pages

Automatic segmentation of speech is compulsory for building large speech databases to be used in speech processing applications. This study proposes a bimodal automatic speech segmentation system that uses either articulator motion information (AMI) or visual information obtained by a camera in collaboration with auditory information. The presence of visual modality is shown to be very beneficial in speech recognition applications, improving the performance and noise robustness of those systems. In this dissertation a significant increase in the performance of the automatic speech segmentation system is achieved by using a bimodal approach.

Automatic speech segmentation systems have a tradeoff between precision and resulting number of gross errors. Boundary refinement techniques are used in order to increase precision of these systems without decreasing the system performance. Two novel boundary refinement techniques are proposed in this thesis; a hidden Markov model (HMM) based fine tuning system and an inverse filtering based fine tuning system. The segment boundaries obtained by the bimodal

speech segmentation system are improved further by using these techniques.

To fulfill these goals, a complete two-stage automatic speech segmentation system is produced and tested in two different databases. A phonetically rich Turkish audiovisual speech database, that contains acoustic data and camera recordings of 1600 Turkish sentences uttered by a male speaker, is build from scratch in order to be used in the experiments. The visual features of the recordings are extracted and manual phonetic alignment of the database is done to be used as a ground truth for the performance tests of the automatic speech segmentation systems.

Keywords: speech segmentation,bimodal, audiovisual, boundary refinement, fine tuning, visemes

ÖZ

ÇİFT DURUMLU OTOMATİK KONUŞMA BÖLÜTLEME VE SINIR İYİLEŞTİRME TEKNİKLERİ

Akdemir, Eren
Doktora, Elektrik Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Tolga Çiloğlu

Mart 2010, 130 sayfa

Otomatik konuşma bölütleme, konuşma işleme uygulamalarında kullanılacak geniş konuşma veritabanlarının hazırlanması için gereklidir. Bu çalışmada artikülatörlerin konum bilgileri ya da kamera kayıtlarından elde edilen görsel verileri, işitsel verilerle birlikte kullanan çift durumlu otomatik konuşma bölütleme sistemi önerilmiştir. Görsel verilerden, birçok konuşma tanıma uygulamasında faydalanılmıştır. Görsel bilgilerin varlığı bu sistemlerin performansını ve gürbüzlüklerini arttırmıştır. Bu çalışmada çift durumlu otomatik konuşma bölütleme sistemi kullanılarak, bölütleme başarımında kayda değer bir artış sağlanmıştır.

Otomatik konuşma bölütleme sistemlerinin çözünürlüğü arttırıldığında sistemin yaptığı büyük hatalar da artmaktadır. Bu durumun üstesinden gelebilmek için sınır iyileştirme teknikleri kullanılmaktadır. Bu çalışmada iki ayrı sınır iyileştirme yöntemi önerilmiştir; Saklı Markov Modeli (SMM) tabanlı ve ters süzgeçleme tabanlı sınır iyileştirme sistemleri. Çift durumlu otomatik konuşma bölütleme sistemiyle elde edilen ses sınırları önerilen iki yeni sınır iyileştirme sistemi kullanılarak elle işaretlenmiş sınırlarla aralarındaki ortalama mutlak fark

daha da azaltılmıştır.

Sonuç olarak iki basamaklı bir otomatik konuşma bölütleme sistemi oluşturulmuş ve bu sistemin başarımı iki ayrı veritabanı kullanılarak sınanmıştır. Ayrıca bu çalışmada kullanılmak üzere, erkek bir konuşmacının akustik kayıtlarını ve kamera kayıtlarını içeren 1600 cümlelik bir Türkçe görsel-işitsel konuşma veritabanı oluşturulmuştur. Bu veritabanına ait görüntü kayıtlarının görsel öznitelikleri çıkarılmış ve ayrıca veritabanının elle fonetik hizalaması yapılarak, veritabanı otomatik konuşma bölütleme sistemlerinin başarım ölçümünde kullanılmaya hazır hale getirilmiştir.

Anahtar Kelimeler: konuşma bölütleme, çift durumlu, görsel işitsel, sınır iyileştirme, vizem

To my family

ACKNOWLEDGMENTS

My time during the Ph. D. had been a very long and challenging period for me. I would like to express my gratitude to those who had been with me during this period and made it possible for me get over all the difficulties that I came across.

I owe so much to my supervisor Assoc. Prof. Dr. Tolga Çiloğlu, who had been a great mentor, lighted my way with his guidance, vision and motivation. I am very grateful to him for all the support that he provided to me in all these years.

My dear friend Emre Özkan had always been in my side since 1998, my first year in METU, walking with me through the ways from freshman year to finishing my Ph. D., I can never thank him enough for his with his support and friendship.

I also would like to thank to my fellows in METU Speech Laboratory; Turgay Koç, and İ. Yücel Özbek for the ideas they shared with me, their support and friendship through these years, and their help during the publication of this thesis.

I am also very grateful to Umut Orguner, Evren İmre, Özgül Salor and Evren Ekmekçi for their friendship and company, who made my times in the department enjoyable and worthy.

I owe much to my old friend Onur Yüce Gün who had been with me since high school and also to my friends Soner Yeşil, Sezen Yeşil and their lovely daughter Gökçe Yeşil, Ömer Köktürk, Zeynep Ekmekçi, Sinem Ulutürk, Beril Beşbınar, Dilruba Erkan and my friends at the METU Aviation Society who had been enriching my life during this long period.

Lastly, I would like to thank to my parents, Bilnur and Ensar Akdemir, who made me feel their complete trust and support at all moments in my life, and I also would like to thank to my brothers Kerem Akdemir and Emirhan Akdemir,

and to Ayşegül and Yücel Bağrıaçık who supported me in my education and became hosts for me in my first years in Ankara, and also to my grandmother Bakiye Güner who had been the gentle hand on my cheeks since my childhood.

This work is supported by Scientific and Technological Research Council of Turkey (TUBITAK) Project no: 107E101.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Speech has been the primary source of communication between humans since the dawn of the civilization. Researchers investigated the production of speech for centuries, the studies on producing mechanical synthetic speech goes back to 1773, Russian Professor Christian Kratzenstein studied physiological differences between five sustained vowels (/a/, /e/, /i/, /o/, and /u/) and made some apparatus to produce them artificially [1, 2], followed by Wolfgang and von Kempelen in Vienna who build an "Acoustic-Mechanical speech machine" [3]. The first full electrical synthesis device was introduced by Stewart in 1922 [4], since then the studies on mimicking the human speech production continues. The history of speech recognition is relatively recent. The subject became a topic of wide public interest after screening of several blockbuster movies starting from early 1960's [2], especially with Stanley Cubrick's legendary movie "2001: a Space Odyssey" and HAL (Heuristically programmed ALgorithmic Computer) the first leading character played by a computer and one of the greatest film villains of all time. HAL was an intelligent computer that understands the speech and responds by talking back to the users, that became a symbol of human and computer interaction for decades, followed by the droids, R2D2 and C3PO from the Star Wars Universe created by George Lucas. The earliest speech recognition systems stem back to digit recognition system by Davis et. al., in Bell laboratories in 1952 [5] and sped up in 1960's and 1970's by the Works of Itakura, Rabiner and Levinson and others [6, 7, 8]. The advances in both fields are boosted by the introduction of digital computers, and nowadays the human computer interaction is not a science fiction anymore, becoming a daily

concept of our lives, already starting to take its place in personal computers, home entertainment systems, game consoles, cellular phones, etc.

1.1 Automatic Speech Segmentation

State of the art speech processing tasks like speech synthesis and speech recognition strongly rely on corpus-based methodologies and because of this the need for well prepared speech databases arises. The databases are needed to be labeled and accurate and precise alignment between these labels and the speech waveform should be provided in order to be used in speech processing applications. The task of matching phonetic units to available acoustic waveform is called as the segmentation of the speech. The purpose of speech segmentation (also referred to as phonetic alignment, phonetic segmentation or text to speech alignment) is to time-align a sequence of phonetic labels and given acoustic data. The phonetic units to be aligned can be words, syllables, phones, etc. Ultimately, speech segmentation is the identification of phonetic boundaries in the speech waveform. In this dissertation phonemes are selected as the phonetic units. A phoneme is the smallest segment that is distinctive, in the sound system of a language [9, 10, 11]. Hence, by selecting the phonemes as segmentation units, the syllable case and word case are also covered. There are numerous applications in which large quantities of phonetically segmented speech is necessary. Nowadays, data driven, concatenation based Text to Speech (TTS) systems are mostly preferred in producing synthetic speech because of their naturalness, and fluency with respect to other speech synthesis systems. Huge amounts of phonetically labeled sentences are required to build a speech synthesizer based on waveform concatenation. As a result of the growing need for more versatile speech synthesis systems, which can adapt to new voices, and/or new languages quickly, with the maximum quality possible, the size of the needed phonetically aligned databases are multiplied. It is obvious that building large corpora with high quality as quickly as possible is invaluable for such tasks.

Despite the practical need of dividing the speech waveform into its segments, the notion that speech is a string of conjoined segments is not completely true. In

fact, apart from the cases where abrupt changes can be observed in the acoustical signal, the speech waveform is a continuously changing signal. Considering the speech waveform alone, isolating it from the meaning, perceptual and linguistic properties, it would be very difficult to split it to some segments. However, if the physical realization of the speech is considered including these factors, the psychological reality behind the notion of speech segments arises. Actually segmentation of the speech is so meaningful to humans, because both the listener's perceptual system and the speaker's utterance planning system operate on symbolic representation which is constrained to be in terms of sequenced discrete objects [12]. So the speech segmentation process should aim to find the degraded (in terms of discreteness) outcome of the human speech production system, by the help of the perceptual and linguistic properties of speech.

Given acoustic data for two consecutive phonemes, there is no unique definition for the location of boundary between the phonemes since phoneme to phoneme transitions evolve in a gradual and continuous manner. Manual phonetic alignment -where the boundaries between phonetic units are marked by expert human labelers - is still considered as the most reliable way of obtaining a labeled and a time-aligned corpus. It is shown that the perceptual quality of the concatenation based synthetic speech obtained by using a manually segmented database is better than the one obtained by using the same database that is automatically segmented, if the sizes of both databases are equal.

Manual segmentation of speech would probably be the most favorable method in corpus development had it not been an extremely time-consuming process and a considerably expensive one. Labelers have to spend 100-200 times of speech duration to undertake the process [13, 14]. This is a major concern as the developing speech systems require larger and larger databases. Furthermore, issues related to the training of labelers, differences among their levels of expertise, inter- and intra-labeler inconsistencies of labels, have to be resolved. It should also be noted that, manually segmented databases are not exactly reproducible even with the same labelers [14, 15]. These concerns, especially as the database sizes are multiplied, emphasize the creation of automatic methods of high accuracy, consistency and speaker/language independency.

1.2 Literature Review

Hidden Markov Models (HMM) are widely used in speech recognition. HMM speech recognizers are also used for automatic speech segmentation in forced alignment mode [13, 14, 16, 17]. In forced alignment, HMM speech recognizer is provided with the phonetic transcription of the speech to be recognized, this reduces the duty of the speech recognizer to determine the phonetic boundaries, from the optimal state sequence output of the HMM decoder. HMM based segmentation systems mostly require a small amount of manually segmented speech database to form initial phoneme models. Instead, linear initialization is also possible, where the speech utterance is uniformly segmented and each segment is used to initialize corresponding phoneme model in the utterance. However, linear initialization is rarely used due to its low performance, except the compulsory cases, where there is no manually segmented data available. Both context dependent and context independent HMMs are used for AS. In [13], the performances of HMM automatic segmentation systems using speaker independent monophone, speaker dependent monophone, and speaker dependent triphone models are compared against a hand labeled database by using perceptual tests. In [14], the performances of HMM automatic segmentation systems, using monophone, diphone, triphone, and tied state triphone models are compared against each other using hand labeled data as a baseline. It was shown that monophone HMM models with optimized number of states outperformed the others. Perception experiments were also made in order to evaluate the effect of segmentation errors on the naturalness of synthetic speech. It was observed that a manually segmented database produces superior results than the automatically segmented database when the amounts of segmented data are the same for both cases. However, doubling the automatically segmented database size (while keeping the manually segmented database size unchanged) yields better performance of automatic segmentation over the manual one in the listening tests.

Although HMM AS systems are basically designed to maximize the probability of the occurrence of the provided waveform for the given phonetic transcription but

not specifically to identify the segment boundaries, they are the most commonly used systems for AS. A different approach to automatic speech segmentation problem is based on the use of dynamic time warping (DTW) technique to align the given speech waveform with a corresponding synthetic speech waveform with known segment boundaries. The advantage of this approach is that the need for a training stage and a training database vanishes, and therefore the system can be adapted to different languages as long as a speech synthesizer in that language is available. In [16], a DTW AS system is compared to a conventional HMM AS system using Gaussian Mixture Models (GMM) to model the state emission probability density functions, and a hybrid HMM/Artificial Neural Network (ANN) system. The performances of such systems are usually worse than HMM AS systems [16].

Automatic speech segmentation systems, as they need more precise positioning of the phonetic boundaries, mostly operate at higher frame rates ($\sim$200-1000 frame/s) compared to speech recognition systems ($\sim$100 frame/s). As a result, the speech recognizers used in automatic speech segmentation should have increased frame rates, but this is not always possible. One drawback of increasing time resolution is the loss in the precision in frequency domain. The use of high frame rates in a HMM based segmentation system leads to increased number of gross errors due to decreased phone identification capability of the HMM speech recognizer, i.e, misrecognitions increase and this increases the number of gross errors [15]. To overcome this problem, two stage automatic speech segmentation systems are proposed. In the first stage, phonetic boundaries are estimated with a relatively lower frame rate system, and then these boundaries are refined through a second process using some spectral measures, or additional information from the database. This two-stage approach is similar to manual phonetic segmentation process, in which first boundaries are found roughly, and then refined by listening and inspecting the spectrogram, waveform, etc. Several methods for boundary refinement were proposed in the literature, using statistical models, average phone durations, acoustic discontinuities, average deviations from hand labeled boundaries, etc. [15, 18, 19, 21, 22, 23].

The performances of the AS systems are generally evaluated by comparing them

with the manually segmented boundaries. The mean of the absolute values of the differences between the boundaries found by the AS system and marked by the manual segmenters, called average absolute boundary error, is the mostly used measure [21, 24]. Another method is measuring the accuracy, by defining the accuracy as the percent of boundaries where the magnitude of the difference between the automatically found boundary and manually segmented boundary is smaller than a threshold (typically 10-20 ms.) [15, 25, 26]. Conducting perceptual listening tests over the synthetic speech obtained by using the automatically segmented database is another way of assessing the performances of AS systems. Although, this method depends on the opinions of the listeners, hence is not an objective method of assessment, it is used in some studies as a performance measure [13].

1.3 Outline of the Thesis

A bimodal AS proposed in this thesis in order to explore the possible benefits of the incorporation of visual data with the speech data. The suggested system is tested on two different databases. Several audiovisual feature vectors are proposed and tested on both databases. Afterwards, two new boundary refinement techniques are proposed in order to further decrease the absolute error of the boundaries estimated by the bimodal AS system.

The first stage is a HMM (Hidden Markov Model) automatic speech segmentation system, build by using HTK speech recognition toolkit [27]. The system is actually a speech recognizer that is used in forced alignment mode, by using the available phonetic transcriptions of the acoustic data. The HMM automatic segmentation system is used in the experiments for the assessment of various audiovisual feature vectors. The experiments run by using the MOCHA-TIMIT database, incorporating the articulator motion information (positions of the upper lip, the lower lip and the jaw) and the acoustic data in various ways, are presented in Chapter 2.

A phonetically rich Turkish audiovisual database is collected and prepared to

fulfill the need of a richer audiovisual speech database, which has more visual information available and has greater number of recordings that arises during the experiments on MOCHA-TIMIT database. The presence of such a database also enables the extension of the idea of audiovisual segmentation to Turkish as well. Several audiovisual feature vectors are proposed using the speech and the camera recordings of the database. Similar experiments to the ones in Chapter 2 are conducted to test the benefits of adding the visual information to acoustic information. The preparation and collection of the Turkish audiovisual database, the experiments on this database using various audiovisual feature vectors, and several methods for using the different boundary sets estimated by using different audiovisual features (decision fusion) are discussed in Chapter 3.

In Chapter 4 two algorithms are proposed for the refinement of the boundaries estimated by the automatic speech segmentation systems suggested in Chapter 2 and 3. One of the proposed algorithms uses a modified HMM topology in order to model different boundary types. The HMM boundary models are used for the precise detection of the boundary locations. The second one introduces a distance measure between consecutive speech segments, by using glottal inverse filtering of the speech. The refined boundaries are marked as the instants where the suggested distance measure between the consecutive speech segments is maximum. The overall results found by the refinement of the boundaries from the first stage systems are presented to conclude the chapter.

Chapter 5 concludes the thesis with a summary of the work done, and the discussions over the results achieved by the proposed automatic speech segmentation systems.

1.4 Contributions

The major contributions of this thesis can be summarized as follows:

- **Bimodal Automatic Speech Segmentation:** Although the use of visual modality had been benefited widely in speech recognition systems, its potential in automatic speech segmentation had not been investigated yet.

In this thesis, the electromagnetic articulograph data and camera data were used in collaboration with speech in order to inspect the possible improvements in the automatic segmentation results. The experiments had shown that the performance of the AS system increases significantly with the addition of information from visual modality.

- **Turkish Audiovisual Speech Database:** Databases are compulsory for developing speech processing applications. A rich database in terms of available acoustic and visual data and in terms of diversity of the visual data was needed for the studies in this thesis. There exists no publicly available, and rich enough audiovisual database, and also no publicly available Turkish audiovisual speech database in the literature. A phonetically rich audiovisual speech database is build in order to be able to test the audiovisual automatic speech segmentation system proposed.

- **Different Approach to Decrease Intralabeler inconsistency in manual segmentation of speech databases:** Nearly all speech processing systems requires labeled and time-aligned databases. The time-alignment process is usually handled by manual segmenters. Besides the time consumingness and expensiveness of the process, interlabeler and intralabeler inconsistency is a major problem. A boundary class wise manual segmentation approach is suggested and used in the manual segmentation of the Turkish audiovisual speech database in order to decrease the intralabeler inconsistency problem.

- **Two New methods for Boundary Refinement:** Boundary refinement is a process that tries to increase the precision and the accuracy of automatic speech segmentation systems. By using the segmentation results from first stage, acoustical properties of speech and some statistical information, the locations of the boundaries are fine tuned. Two different boundary refinement techniques that can also be used mutually are suggested in this thesis.

  - **HMM based boundary refinement:** HMM's are widely used in speech processing applications and they are known to have a very

high phone identification capacity with the sacrifice of low time resolution. In this thesis, a new HMM topology is suggested to be used for boundary modeling. The proposed HMM topology can work in high time resolution and also be used in a context dependent manner by using the advantage of having the phonetic transcriptions of the acoustic data.

– **Glottal Inverse filtering based boundary refinement:** Glottal inverse filtering is used in various speech processing applications with the aim of identification of glottal input waveform from the available speech waveform. In this thesis, an algorithm that uses a newly defined distance measure between successive speech segments is built using inverse filtering. Boundary refinement is done by using this algorithm by locating the boundary where the distance between successive speech segments is found to be maximum.

CHAPTER 2

BIMODAL AUTOMATIC SPEECH SEGMENTATION

2.1 Introduction

Speech production and perception are bimodal processes [30, 31] . Both acoustic and visual stimuli are produced during speech production, and at the other side, both visual and acoustic data are processed by the listeners to understand speech. The studies showing the benefit of visual data to speech intelligibility in noise was introduced by Sumby and Pollack in 1954 [32]. Since then, bimodal speech recognition systems aim to emulate multimodal structure of human perception in order to improve performances of audio-only systems.

The bimodality of the speech perception is also demonstrated by the famous experiment performed by McGurk and McDonald in 1976 [33]; when viewing the video of a person uttering /ga/, and listening the sound /ba/, most of the listeners perceives that the uttered sound is /da/. This phenomenon showing the bimodal nature of speech perception is called McGurk effect.

The presence of visual information helps the listeners by providing complementary information to the audio data, by helping the speaker localization and by providing speech segmental information, and by providing voice activity detection [34, 35]. Some phonemes that are similar in acoustic domain can be distinguished easily in visual domain. For example bilabial consonants /p/, /b/, and /m/ can be distinguished easily from their velar and alveolar counterparts /k/, /d/, and /n/ in visual domain, as distinctive place of articulation information can be obtained. The movement of the jaw also provides segmental information

about speech. Jaw syllabic oscillations and movement of the head, is highly correlated to acoustic data and proved to improve human speech perception [36, 37, 38].

The visual information had been used with speech in various human machine interaction systems. The help of visual modality is used in: person recognition/verification, laughter/smile detection, voice activity detection, emotion recognition, speech recognition etc. There are a number of studies in bimodal speech recognition, also referred as speechreading (uses both audio and visual data) and lipreading (uses visual data only) [39, 40, 41, 42, 44, 45, 46, 47, 48, 49]. It has been found that speech recognition benefits from the use of visual modality especially in noisy environments [39, 40]. However, to the best of the author's knowledge, there is no work published on bimodal speech segmentation, except the one published by Mak and Allen in 1994 [50]. In their study, the velocity of the lips is used with acoustical data in order to increase the performance of the segmentation system in noise. The visual segment boundaries are decided by peak picking and thresholding the velocity of the lips and they are fused with acoustic segmentation in order to improve segmentation accuracy in noise. The results are tested on a database that consists of only 5 sentences and segmentation errors are reduced by 10.4% when SNR is lower than 15 dB.

The studies presented in this chapter aim to investigate the possible improvements that can be achieved by the inclusion of visual data to automatic speech segmentation. Publicly available MOCHA-TIMIT database is used for this purpose. The visual information is fused to acoustic information in feature level in various ways, and the features obtained are used as input to a HMM automatic speech segmentation system build by using HTK speech recognition toolkit [27].

The inclusion of the information from the visual modality is investigated in this chapter. The organization of the chapter is as follows; some background information about speech production, phonemes, phones and phoneme types are given in Section 2.2. The MOCHA-TIMIT database used in the experiments is introduced in Section 2.3.1, the proposed audiovisual automatic speech segmentation system is described in Section 2.3.2, the feature vectors that are used

in the experiments are presented in Section 2.3.3. The experimental results are listed and investigated regarding different phoneme classes in Section 2.4. Finally, conclusion about the experiments is made in Section 2.5.

2.2 Speech Production

Speech signal is produced in the form of pressure waves emanating from the speaker's mouth to the ear of the listener. Despite the nonstationary characteristics, the speech is accepted to be composed of sound segments, which share some common acoustic and articulatory properties with one another for a short interval of time (Section 1.1) [58]. Each of these segments has a positioning/state of the vocal folds and vocal tract articulators; tongue, lips, teeth, velum and jaw.

The lungs are the source of energy for producing speech, generating the air flow through the vocal tract during exhalation of the air. The air flowing through the larynx and vocal tract creates the pressure necessary to produce speech. The almost constant pressure from the lungs is not able to produce sound itself, as a change in the pressure is needed for sound to be produced. The pressure change is provided by the vibration of the vocal cords or by the turbulence created by some constriction point on the vocal tract. The lung pressure vibrates the vocal cords to produce periodic excitation for voiced speech or creates a random noise source, by the compression of the air flow on some point/points in the vocal tract. The excitation from the vocal folds or some point in the vocal tract, is shaped by the vocal tract and radiated from the lips. This can be viewed as a filtering operation, where the vocal tract acts as filter for the sound source which can be periodic or noisy and aperiodic or both. The periodic excitation case caused by the vibrations of the vocal folds, results in voiced speech that is quasi-periodic (almost periodic). Aperiodic excitation case is caused by the constriction of the air flow at some point/points of the vocal tract and results in noisy, unvoiced speech.

2.2.1 Vocal Folds

Vocal folds are two masses of flesh ligament and muscle, which stretch between the front and the back of the larynx. They are free to move at the back and sides of the larynx, attached to two cartilages, that controls the position and the tension of the vocal folds in collaboration with the muscles within the folds. They can abduct (move apart) and adduct (move together) during phonation.

Vocal folds have three primary states: breathing, voiced and unvoiced. The glottis is wide open and the muscles within the vocal folds are relaxed in breathing state. The air from the lungs is free to flow with no significant obstruction by the vocal folds. However in production of the voiced speech, the air flow is obstructed by the folds. The vocal folds are tensed up and come close together, and the pressure from the lungs causes self-sustained oscillations of the vocal folds. The unvoiced state is similar to breathing state, but the folds are brought closer together and muscles are tenser than the breathing state. In this state vocal folds create turbulence which is called aspiration [59].

2.2.2 Vocal Tract

The vocal tract acts as a linear filter that filters the input from the vocal folds and/or other parts of the vocal cavity. The resonant frequencies of the vocal tract are called formants. The vocal tract amplifies the energy of the source around the resonant frequencies, while attenuating energy around antiresonant frequencies. The vocal tract can be modeled as a number of concatenated cylinders of varying cross-sectional area, but, in fact the actual shape is much more complex [60]. By moving the articulators the shape of the vocal tract is changed, which changes the frequency response of the system and the resonance frequencies. The shape of the vocal tract has the most decisive role on the produced sound.

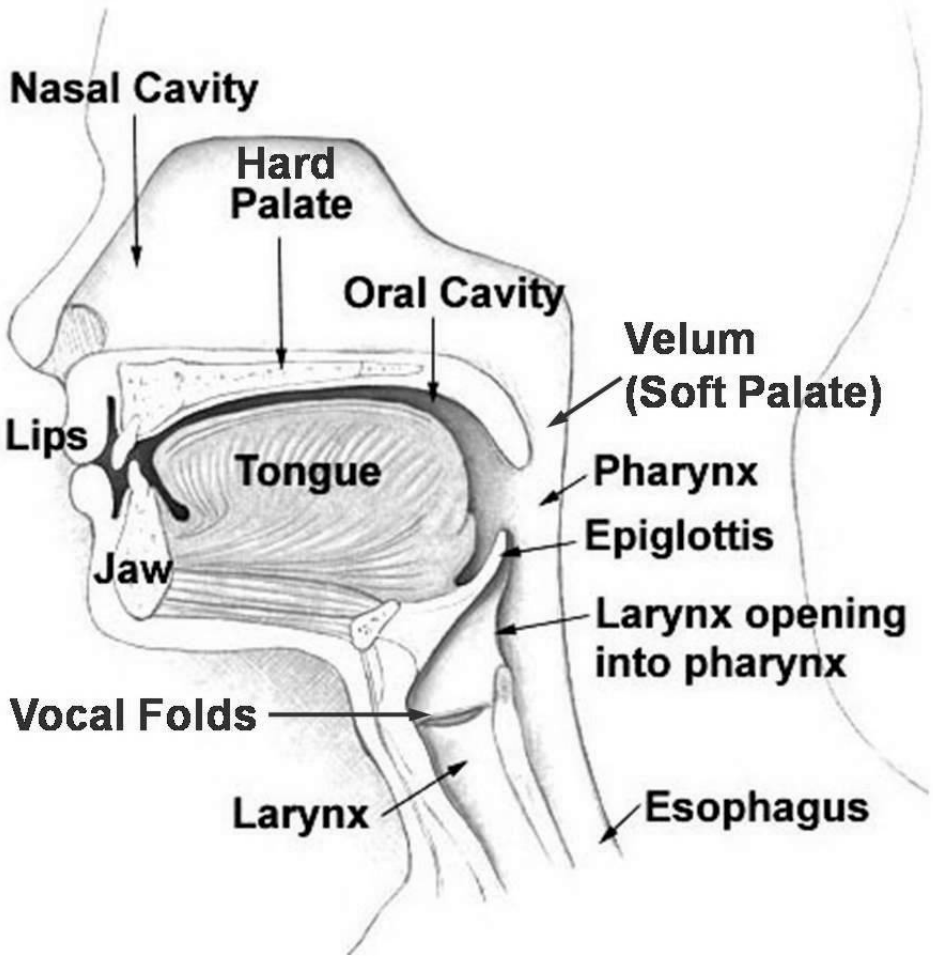

Figure 2.1: Vocal tract and vocal folds.

2.2.2.1 Articulators

Articulators are the structures in the vocal tract that can be moved to reshape the tubular form of the vocal tract. The tongue and lips are the most important articulators in terms of producing different number of different sounds produced, but velum and larynx also have important roles in speech production [58] (Figure 2.1).

- **Lips:** The most visible articulators are the lips. Lips can change the width (closure) and shape (rounding) of the end of the vocal tract. Closure is usually caused by the movement of the jaw and lower lip, while rounding is caused by the lip muscle which surrounds the lips.

- **Teeth:** The upper incisors are used to produce labio-dental (/f/,/v/) or dental (/t/,/d/ etc.) obstruents, by contacting to the lips, the tongue or lower incisors. They can be visible during the articulation of these sounds most of the time.

- **Tongue:** Tongue is composed of 12 muscle pairs and some passive tissues. It has ability to change shape of the vocal tract by changing its configuration, making contact with various parts of the oral tract, or creating narrow cross sections to cause constriction of the air flow. Most of the different vocal tract shapes are accomplished by different positioning of the tongue.

- **Velum (Soft Palate):** is a valve composed of muscles that separates oral cavity and nasal cavity. Velum is lowered and the air flow through nasal cavity is allowed during the articulation of nasal consonants. It is closed and blocks the air flow otherwise.

- **Larynx:** The primary function of the larynx is controlling the airflow through the vocal tract but it can also be lowered or raised to change the length of the vocal tract.

2.2.3 Phonemes and Phones

The notions of phoneme and phone are closely related and often used interchangeably. A phoneme is the smallest distinctive unit in the sound system of a language [9, 10, 11]. The word *distinctive* should not be confused with *distinct*. The distinctive refers the ability of a phoneme to make distinguish between words. The phonemes are called units because they are the smallest parts that must be substituted to make a different word. For example, the word kar in Turkish is represented by the phonemes /k/-/a/-/r/. One need to change one of the phonemes to change the meaning, the smaller changes can not change the meaning of the word.

On the other hand, a phone is the smallest identifiable unit found in a stream of speech that is able to be transcribed with a phonetic symbol. In other words

phones are the acoustic realizations of speech and phonemes. However the realization of a phoneme can be as different phones, the definition of allophones emerges here. Allophones are different phones that are the acoustic realizations of one phoneme. An example to this is the difference in the sounds corresponding to the letter t in the English words tea and trip [11]. The different phones in these words can be represented as $[t_v]$ and $[t_r]$. This set of phones corresponding to phoneme /t/ is called the allophones of phoneme /t/. Note that, substituting the allophones with each other does not change the meaning, even though the word can sound a bit weird. Phonemes are usually represented between slashes and phones are usually represented between brackets by convention. One example of allophones in Turkish can be given as; different realizations of the phoneme /e/ that are called as closed /e/ as pronounced in the word *gece* and open /e/ as pronounced in the word *eş*, these two phones have the representation of [é] and [e] respectively. These are the allophones of the phoneme /e/. However the similar sounds in the words *kar* and *kel* that are represented by k in Turkish alphabet are represented by different phonemes /k/ and /c/ in Turkish phonetic alphabet, respectively [61].

2.2.4 Phoneme Groups

The phonemes can be categorized in various ways. The realization of a phoneme arises from a combination of the states of the vocal folds and the vocal tract. The phonemes can fall into different categories by whether the vocal folds are vibrating or not; by different tongue positions; by different lip shapes; by whether the velum is open or not, etc. As the phonemes are language specific, the number of phonemes changes from language to language.

2.2.4.1 Vowels

Vowels are formed by the quasi-periodic input from the vocal folds, with an open vocal tract allowing the air flow. Vowels have three subgroups according to the position of the tongue; front, central or back. The sound is also quasi-periodic as the input. One "period" of the speech waveform is called the pitch period. In

Figure 2.4 time domain and frequency domain characteristics of the vowel /e/ are presented.

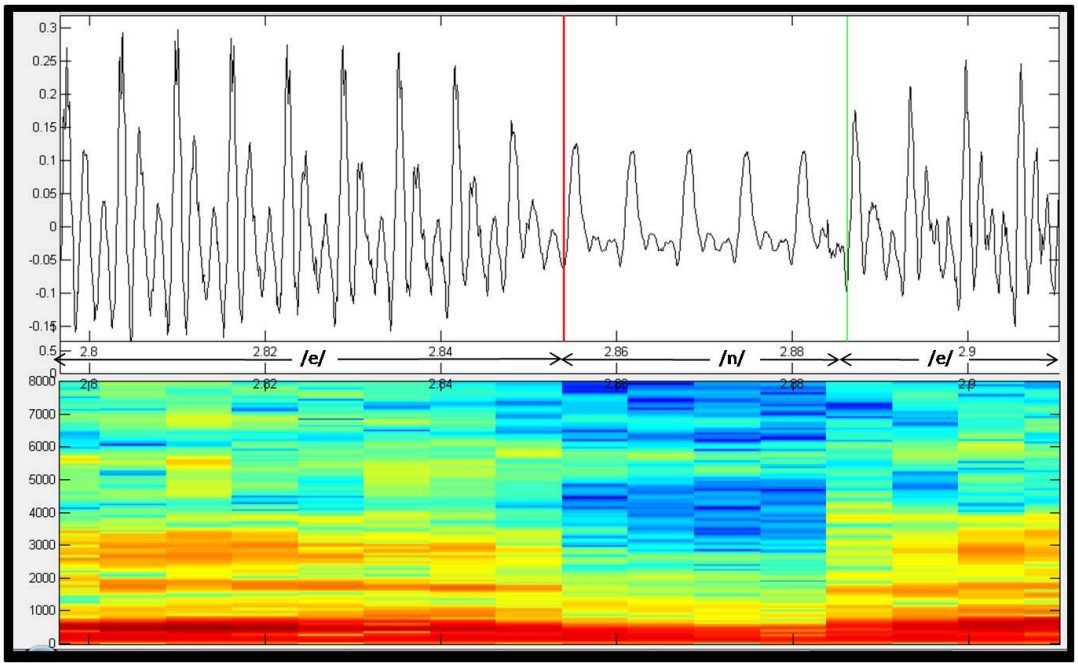

Figure 2.2: Time domain and frequency domain representations of /e/-/n/-/e/ triphone.

2.2.4.2 Nasals

The nasals are a subgroup of consonants. They are formed by the quasi-periodic input from the vocal folds, with an open vocal tract allowing the air flow, such as the vowels. The factor that causes the distinction from the vowels is the opening of the velum. This allows the flow of the air from the nasal cavity, and changes the vocal tract response immediately by introducing zeros to the transfer function. Because of this, the nasals have low resonances and these resonances have high bandwidths. Again the waveform is quasi-periodic, dominated by low first formant (Figure 2.2).

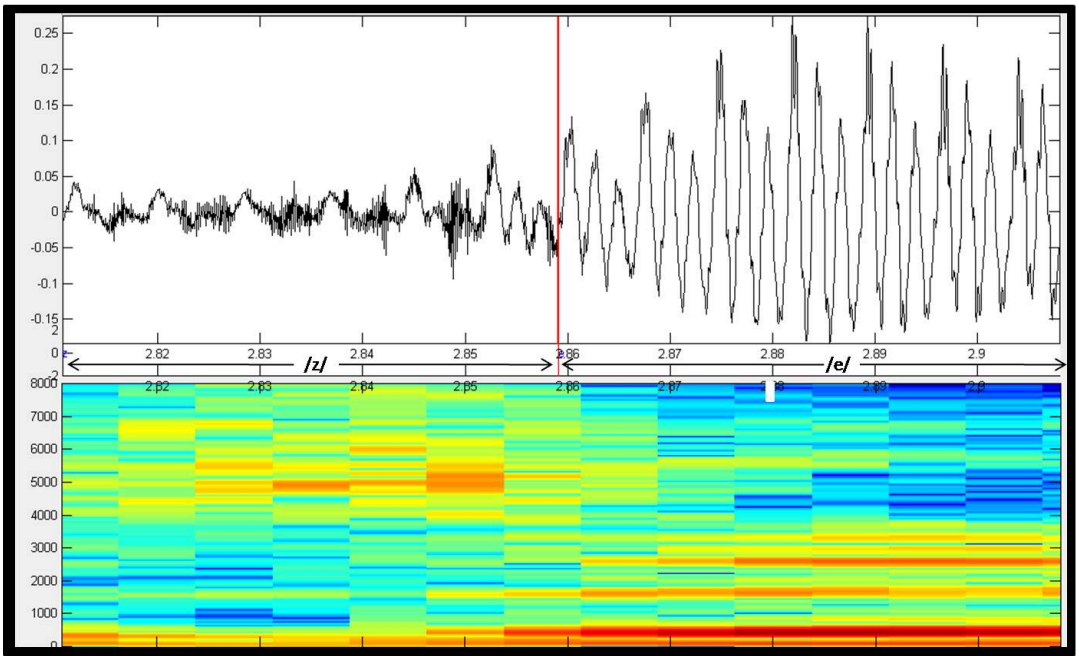

Figure 2.3: Time domain and frequency domain representations of /z/-/e/ diphone.

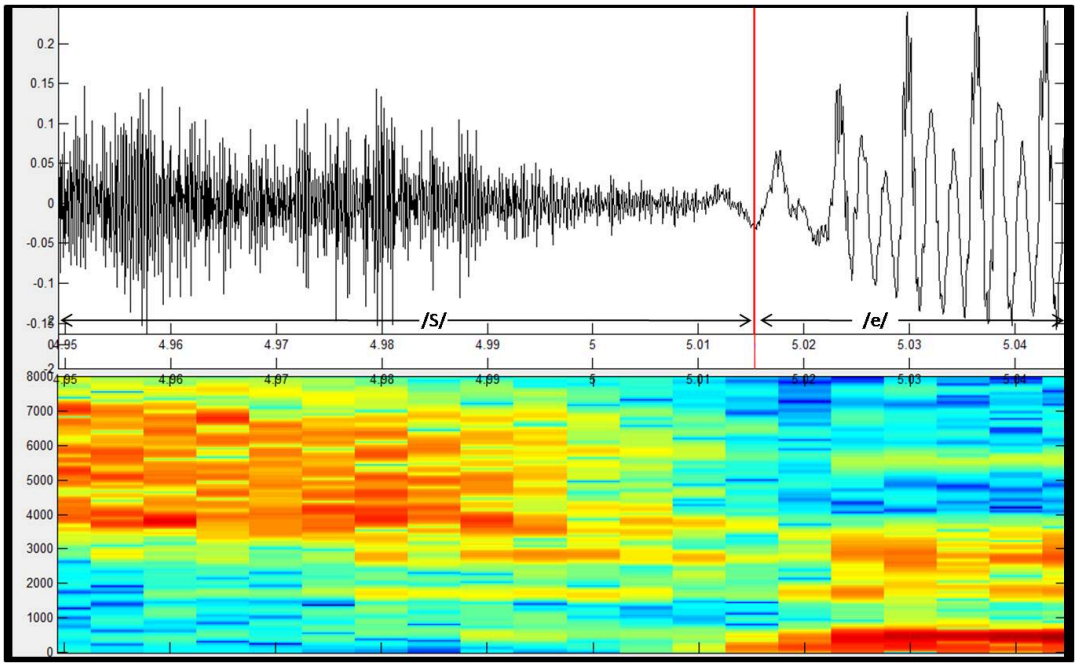

Figure 2.4: Time domain and frequency domain representations of /S/-/e/ diphone.

2.2.4.3 Fricatives

Fricatives are formed by the noisy input generated by the turbulent airflow caused by the constriction at some point on the vocal tract. They are divided to two subcategories according to the state of the vocal folds during articulation. In voiced fricatives vocal folds are tense and vibrate during the articulation and cause the airflow causing the noise to be periodic, resulting a periodic/noisy speech waveform (Figure 2.3). In unvoiced fricatives vocal folds are relaxed and the resultant speech waveform is completely noisy (Figure 2.4).

As the place of the input is closer to the lips, the length of vocal tract that shapes the input is shortened, and also the cavity behind the constriction point produces anti-resonances, the vocal tract transfer function consist of high frequency resonances. The resonance frequencies are determined by the place of constriction.

2.2.4.4 Plosives

Plosives are formed by an impulsive source caused by the burst generated by accumulated airflow behind a total constriction in the vocal tract. In voiced plosives vocal folds are tense and vibrate during the accumulation of the air pressure. During this step although the vocal tract is closed a low amplitude and low frequency speech waveform is observed at the output, due to the propagation of the input through the walls of the vocal tract (Figure 2.5). In unvoiced plosives the vocal folds are relaxed and no output is observed during the accumulation step (Figure 2.6). The time difference between the burst and the voicing before or after the burst is called the voice onset time. The value of voice onset time changes from plosive to plosive, it is negative for voiced plosives and positive for unvoiced plosives, by definition.

The plosives are composed of three parts; closure, burst and aspiration (Figure 2.6). Closure is the state of air flow accumulation behind the constriction. The burst stage follows the closure by the release of the accumulated airflow, and in some plosives the aspiration state fallows with continuing noisy airflow.

Aspiration state may not be observed in some plosives (Figure 2.5).

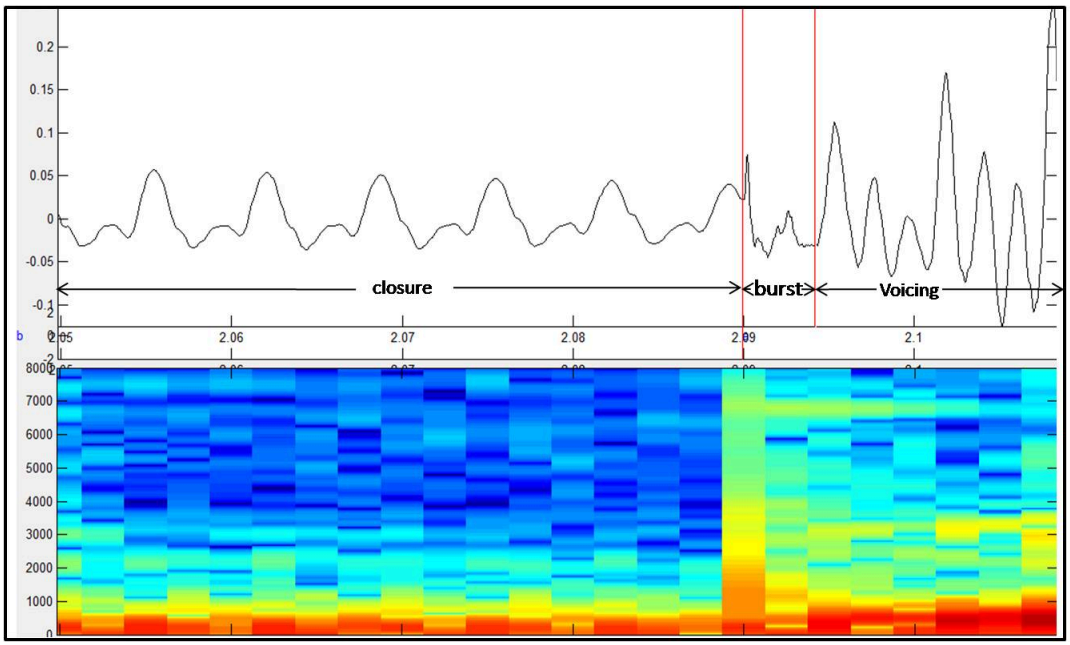

Figure 2.5: Time domain and frequency domain representations of closure and burst states of the voiced plosive /b/.

2.2.4.5 Liquids and Glides

Liquids and glides are semi-vowels that show the vowel-like characteristics with vibrating folds. These are dynamic and transitional sounds mostly occur before a vowel or between vowels, showing transition between preceding and the following vowel. Because of this transitional characteristic the boundaries of the semi-vowels are very hard to be located.

2.2.4.6 Affricates

Affricates are articulated like fricatives except the plosive like closure state at the beginning of the articulation. The examples are /tS/ as in the word *chew* in English and as in the word *çam* in Turkish.

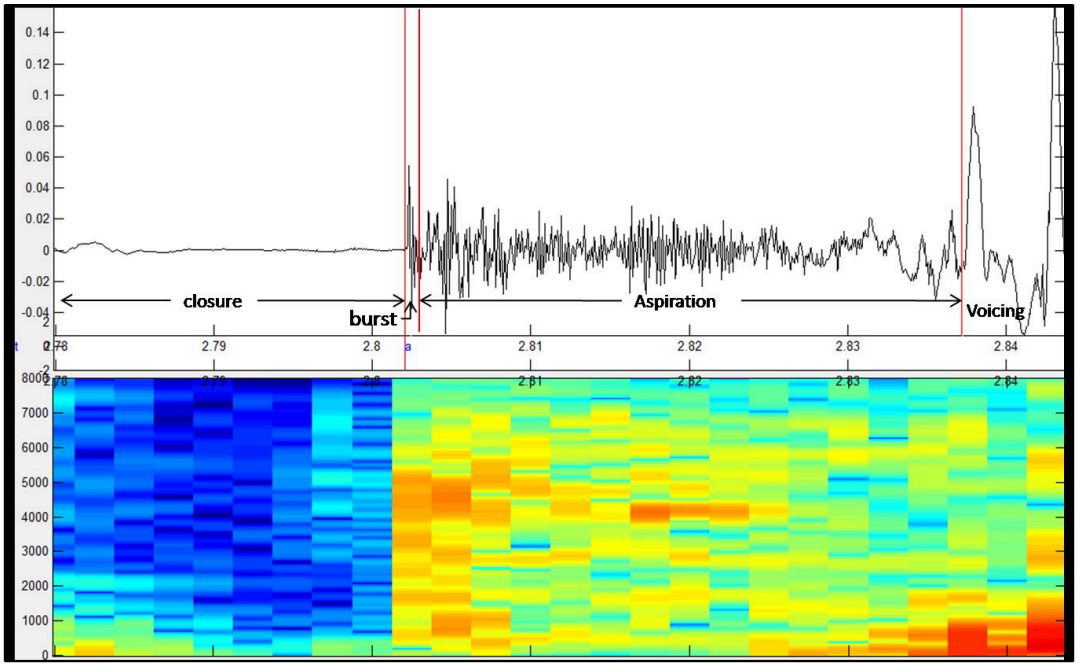

Figure 2.6: Time domain and frequency domain representations of closure, burst and aspiration states of the unvoiced plosive /t/.

2.3 The use of Articulator Motion Information in Automatic Speech Segmentation

Speech signal is formed by glottal excitation of the human vocal tract, composed of articulators, which shape and modify the sound passing from the articulator system, as mentioned in Chapter 2.2. The state of the articulator system (or the frequency response) is hidden in the speech waveform. Some phonemes can be represented exactly by one articulator configuration, some phonemes can be represented with more than one articulator configuration, and some phonemes can be represented by the transition of articulator configuration from one state to another. However acquisition of the state of the articulator system is not an easy task. The methods include mounting receiver coils inside the speaker's mouth as in Electromagnetic Articulograph method [28, 51], Magnetic resonance imaging (MRI) [47], or X-ray imaging [48]. The position and the formation of the lips are relatively easier to acquire by using a digital camera and they are more appropriate to be used in practical systems. Starting with the pioneering work

by Petajan [52], visual information has begun to be used in speech recognition systems, as it increases the robustness of the system to acoustic noise and cross talk among speakers [39, 40, 53, 54, 55, 56]. A major advantage of using visual information is a result of the complementariness among acoustic and visual data [34, 41, 57]. These systems use three types of features; shape based features, appearance based features or both. Shape based features include; horizontal and vertical apertures of the lips, the angle of the lips [39], parameters of ellipsoid models fitted to lips, positions of the sensors placed on the speakers face. Appearance based features include transform domain features found by transformation of the visual data with wavelets, DCT etc. [42]. As a first step to investigate using visual articulator positions in automatic segmentation a HMM AS system is build using the publicly available MOCHA-TIMIT database [28].

2.3.1 The MOCHA-TIMIT Database

The MOCHA-TIMIT database, [28], consists of recordings of 460 English sentences from TIMIT database, each uttered by a male and a female speaker. The database consists of ;

- Acoustic data at 16kHz
- Laryngograph data at 16kHz
- Electromagnetic Articulograph (EMA) data (Figure 2.7) 500Hz sample rate including vertical and horizontal positions of
  - upper incisor
  - lower incisor
  - upper lip
  - lower lip
  - tongue tip
  - tongue blade
  - tongue dorsum
  - velum

The plots of the acoustic signal, laryngograph data, vertical positions of upper lip and lower lip for the sentence "is this seesaw safe" from the database is shown in Figure 2.8. In this research, among the available articulator positions, the ones that can also be obtained visually are selected, as a result, only the vertical and horizontal positions (horizontal positions in EMA data is composed of advancement/retraction and not left/right.) of upper lip and lower lip and vertical position of jaw were used in order to aid an HMM based AS system.

The database also includes *.lab* files, that supplements the time alignment between the speech waveforms and the phonetic transcription of each utterance. The phonetic transcriptions contain 44 phonemes (/@/, / @@/, / a/, / aa/, / ai/, / b/, / ch/, / d/, / dh/, / e/, / ei/, / eir/, / f/, / g/, / h/, / i/, / i@/, / ii/, / iy/, / jh/, / k/, / l/, / m/, / n/, / ng/, / o/, / oi/, / oo/, / ou/, / ow/, / p/, / r/, / s/, / sh/, / t/, / th/, / u/, / uh/, / uu/, / v/, / w/, / y/, / z/, / zh/)[1], and silence and breath.

2.3.2 The Method

The speech segmentation system was build using HTK speech recognition toolkit [27]. 420 sentences of MOCHA-TIMIT database, uttered by the male speaker, were used to train the system, and 40 of them were used for testing. The feature vectors are computed at 100 Hz with an analysis window of length 25 ms. 3 state, left-to-right, continuous Gaussian density HMMs were used for modeling. Context independent HMMs are used as they are accepted to outperform context dependent HMMs in the literature [13, 14]. Monophone acoustic models for the 44 phonemes and silence and breath, are trained by initially using the segment boundaries given in the database, as monophone models are proven to produce better results for AS [14]. The HMM phoneme models are initialized by using the phoneme boundaries supplemented in *.lab* files, using the HINIT tool of HTK. HINIT accepts the prototype HMM as a generator of speech vectors. Supplied training examples are treated as the output of the HMM whose parameters are to be estimated. Thus, if the state that generated each vector

[1]Phonetic symbols are described in Appendix A

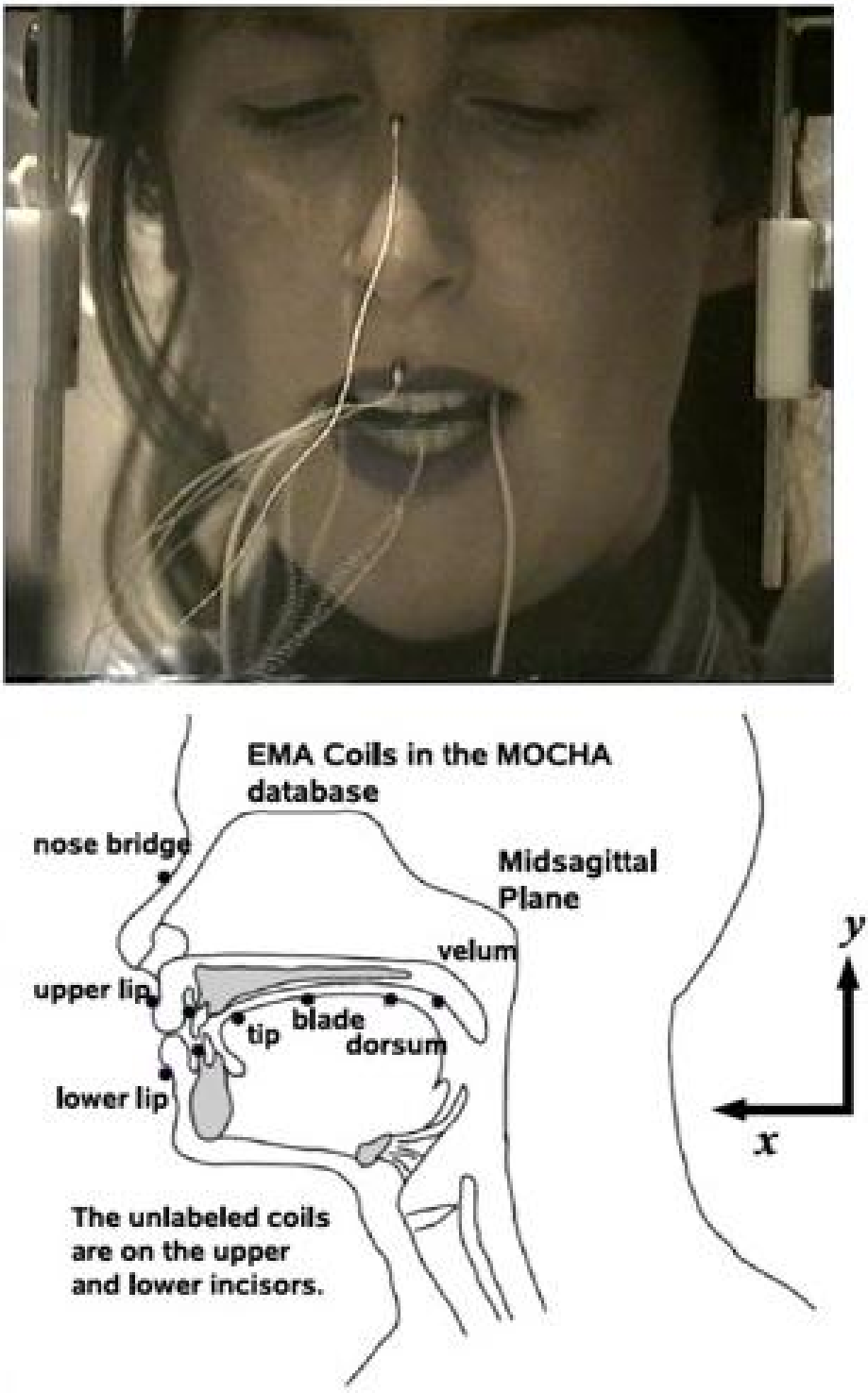

Figure 2.7: Placement of the Electromagnetic Articulograph coils in MOCHA Database [43]

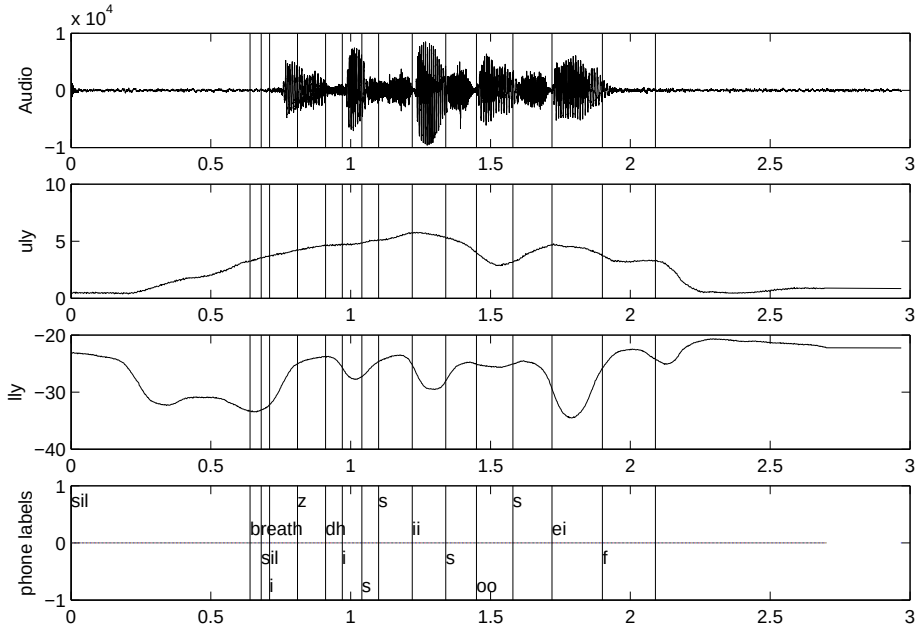

Figure 2.8: Acoustic signal, vertical positions of upper lip and lower lip for a speech file from database. The bottom panel shows the segment labels and segment boundaries.

in the training data was known, then the unknown means and variances could be estimated by averaging all the vectors associated with each state. Similarly, the transition matrix could be estimated by simply counting the number of time slots that each state was occupied. After initial models are set, the training was accomplished in an iterative manner by using Baum-Welch algorithm until the settlement of segment boundaries, HEREST tool of HTK. The test utterances were decoded using the Viterbi algorithm and forced alignment method. The segment boundaries so obtained are compared against the boundaries produced by manual segmentation.

Baseline automatic segmentation results were obtained by using audio feature vector only. Audio feature vector is formed by using Mel frequency cepstral coefficients (MFCCs) as they yield good representation and discrimination of speech, and widely used in speech processing applications. The feature vectors are computed at 100 Hz with an analysis window of length 25 ms. 13 MFCCs (including energy coefficient), and their first and second order derivatives are

extracted. The audio feature vectors were used as a benchmark for evaluating the performances of the systems that use articulator motion information (AMI).

The incorporation of Articulator motion information (AMI) was experimented by designing different feature vectors that include AMI in different forms in addition to MFCCs. EMA records the positions of the articulators at 500 Hz sampling rate. The noisy EMA data were low pass filtered, and then downsampled by 5 to synchronize with the 100 Hz extraction rate of MFCC vectors. Downsampling operation does not cause any problems as the EMA data is confined over very low frequencies (no significant frequency components beyond 10 Hz.).

In using AMI, the derivatives of lip and jaw positions were also utilized. Articulator position data from the MOCHA seems to be noisy, this makes the derivatives of the positions very noisy too. To prevent the noise the derivatives are found by fitting 10th degree polynomials to 250 point segments of the position data and then taking the derivative of the polynomials analytically.

2.3.3 Formation of Feature Vectors

The articulator position features are combined with MFCC vectors. In the experiments, they were either substituted for some of the acceleration coefficients or they were appended to the MFCC vector. The location for the best substitution was experimentally found as the 38th position in MFCC_0_D_A vector. Modified feature vectors of each speech file are stored as *.mfc* files in the HTK environment. The reason of experimenting with substitution of AMI data, in addition to appending them to the MFCC feature vectors, is the possibility of constraints on the size of the feature vectors. Some AS systems may require fixed sized feature vectors (39-element MFCC vectors are commonly used).

The effect of integration of the lip and jaw positions to the automatic segmentation was investigated with the following feature vector forms.

1. **MFCC_0_D_A:** This is the baseline system. Feature vector contains 12

MFCCs and 1 energy coefficient, their derivatives, and second derivatives. (39 elements)

2. **MFCC_0_D:** Feature vector contains 12 MFCCs and 1 energy coefficient, and their derivatives. (26 elements)

3. **MFCC_0_D_A-uly-lly:** Vertical positions of upper lip (uly) and lower lip (lly) are substituted at the 37th and the 38th positions of MFCC_0_D_A.

4. **MFCC_0_D_A-lly:** Vertical position of lower lip is substituted at the 38th position of MFCC_0_D_A.

5. **MFCC_0_D_A-uly:** Vertical position of upper lip is substituted at the 38th position of MFCC_0_D_A.

6. **MFCC_0_D_A-lly-dlly:** Vertical position of lower lip and its derivative are substituted at the 37th and the 38th positions of MFCC_0_D_A.

7. **MFCC_0_D_A-uly-lly-derivatives:** Vertical positions of upper lip and lower lip and their derivatives are substituted at the 35th-38th positions of MFCC_0_D_A.

8. **MFCC_0_D_A+lly:** Vertical position of lower lip is appended to the MFCC- _0_D_A as the 40th element.

9. **MFCC_0_D_A+uly:** Vertical position of upper lip is appended to the MFCC_0_D_A as the 40th element.

10. **MFCC_0_D_A+uly+lly:** Vertical positions of upper lip and lower lip are appended to the MFCC_0_D_A as the 40th and the 41st elements.

11. **MFCC_0_D_A+lly+dlly:** Vertical position of lower lip and its derivative are appended to the MFCC_0_D_A as the 40th and the 41st elements.

12. **MFCC_0_D_A+uly+duly:** Vertical position of upper lip and its derivative are appended to the MFCC_0_D_A as the 40th and the 41st elements.

13. **MFCC_0_D_A+(uly-lly):** Difference between vertical positions of upper lip and lower lip are appended to the MFCC_0_D_A as the 40th element.

14. **MFCC_0_D_A+lly+llx:** Vertical and horizontal positions of lower lip are appended to the MFCC_0_D_A as the 40th and the 41st elements.

15. **MFCC_0_D_A+distlips:** Euclidian distance between lips is appended to the MFCC_0_D_A as the 40th element.

16. **MFCC_0_D_A+lly+distlips:** Vertical position of lower lip and Euclidian distance between lips is appended to the MFCC_0_D_A as the 40th and 41th elements.

17. **MFCC_0_D_A+jy:** Vertical position of jaw (lower incisor) is appended to the MFCC- _0_D_A as the 40th element.

2.4 Experimental Results

2.4.1 Automatic Segmentation Results

Segmentation experiments were performed as described in Section 2.3.2 using the above 17 types of feature vectors. In each case, segmentation errors are computed by comparing the automatically found segment boundaries to those found manually. The AMI contributions using the feature vectors of items 3-17 above are evaluated by comparing their individual results to those obtained by using the MFCC_0_D_A vector (item 1), the results are given in Table 2.1.

The average absolute error for segmentation boundary is found to be 9.9 ms for the baseline system (MFCC_0_D_A). In our experiments, AMI features were substituted in place of some of the acceleration coefficients (items 3-7). To have an idea about the contribution of acceleration coefficients alone, an experiment with MFCC_0_D is performed. It is observed that the acceleration coefficients have vital importance as the performance of the system falls dramatically (average absolute error increases by 18%) when they are not used.

There are 5 cases (3-7) where AMI features are substituted in place of acceleration coefficients. First, the vertical positions of upper lip and lower lip (uly, lly) are substituted in place of two consecutive acceleration coefficients. All possi-

Table 2.1: Average Absolute Segment Boundary Errors for Different Feature Vectors

#	Feature Vector	Average Absolute Error (ms)	Variance of Error (X10 $^{-4}$)	Percent Decrease in Absolute Error Relative to the Segmentation by Using MFCC_0_D_A
1	MFCC_0_D_A	9.9	2.09	-
2	MFCC_0_D	11.5	3.36	-18.2
3	MFCC_0_D_A-uly-lly	9.1	1.59	8.1
4	MFCC_0_D_A-lly	9.0	1.93	6.1
5	MFCC_0_D_A-uly	9.7	1.67	2.0
6	MFCC_0_D_A-lly-dlly	9.2	1.63	7.1
7	MFCC_0_D_A-uly-lly-derivatives	9.7	2.27	2.0
8	MFCC_0_D_A+lly	**8.9**	1.51	**10.1**
9	MFCC_0_D_A+uly	9.6	1.60	3.0
10	MFCC_0_D_A+uly+lly	**8.9**	1.59	**10.1**
11	MFCC_0_D_A+lly+dlly	9.3	1.65	6.1
12	MFCC_0_D_A+uly+duly	10.7	1.89	-8.1
13	MFCC_0_D_A+(uly-lly)	9.2	1.64	7.1
14	MFCC_0_D_A+lly+llx	9.1	1.84	8.1
15	MFCC_0_D_A+distlips	9.1	1.60	8.1
16	MFCC_0_D_A+lly+distlips	9.1	1.84	8.1
17	MFCC_0_D_A+jy	9.4	1.67	5.1

bilities were experimented, and the system performed best when 37th and 38th locations are used for substitution. In this case average absolute error is 9.1 ms, corresponding to 8.1% error reduction. Then, the vertical position of either upper lip or lower lip is substituted into the feature vector individually. After a series of experiments 38th location produced the best results for both upper lip and lower lip position. Substitution of lly parameter or uly parameter results in an average absolute error of 9 ms or 9.7 ms, respectively. This shows that the information provided by the vertical motion of lower lip yields a contribution much more significant than that of the vertical motion of the upper lip.

The time derivatives of the vertical positions of upper lip and lower lip were also considered as components of the feature vector. Two substitution experiments were performed in this context. Substitutions of uly, lly and their derivatives in place of 35th to 38th locations yield an average absolute error of 9.7 ms. Without using derivative information, better results were obtained by using lly

alone compared to using lly and uly together. Along this track, experiments are done by substituting lly and its derivative to 37th and 38th locations and obtained an average absolute error of 9.2 ms.

Substitution experiments show that the largest reduction in average absolute error is achieved by using lly in place of the second derivative of 12th MFCC. It should be noted that in all substitution experiments usage of vertical lip positions improved the automatic segmentation performance. However, since lly alone yields the best result, it is not worth removing the acceleration coefficients to use the other AMI features

As a next step, the experiments are carried out by forming the new feature vectors by appending the data from lip positions to the MFCC feature vector of length 39. uly and lly, alone and together, one of them and its derivative, their difference, forward position of lower lip (llx), and Euclidian distance between lips alone and with lly and vertical position of jaw (jy) are appended to MFCC_0_D_A in different experiments (Items 8-17). The feature vectors so formed have 40 or 41 elements. As a result of these experiments, it is seen that the segmentation performance is improved in all cases except the addition of uly and its derivative (item 12). The best performance with an average absolute error of 8.9 ms (10.1% reduction compared to segmentation with MFCC_0_D_A) is obtained by using only lly. The minimum error variance is also observed in this case. Adding uly together with lly does not change the average absolute error compared to adding lly alone, however it increases the variance of error. The second best average absolute error performance (9.1 ms) is obtained in three cases: one of them is the addition of Euclidian distance between lips, another one is the addition of lly with the Euclidian distance between lips, and finally the addition of lly and llx, stated in order of increasing error variances. Appending the difference between vertical positions of upper lip and lower lip resulted an average absolute error of 9.2 ms which is the third best result. The fourth best result is obtained by the addition of lly and its derivative with an average absolute error to 9.3 ms. It is interesting that this error value is close to that observed when lly and its derivative are substituted in MFCC_0_D_A. Just as it is in the case of the substitution experiments, the least improvement is obtained by the uly

parameter also in the addition experiments with an average absolute error of 9.6 ms (which is also close to 9.7 ms average absolute error observed in the substitution experiments using uly parameter). The inclusion of vertical position of the jaw reduced the average absolute error to 9.4 ms. This is probably due to the syllabic oscillations of the jaw. It should be noted that this is a slightly inferior result compared to that obtained by using lower lip. The movement of the jaw is also included in the movement of lower lip. The results show that the movement of the lower lip provides more information about the boundary point.

Appending the lly parameter, forming a feature vector of length 40 and appending the lly and uly parameters together, forming a feature vector of length 41 reduced the average absolute error by 10% to 8.9 ms, the minimum average absolute error achieved in the experiments. So if the automatic segmentation system to be used allows feature vectors of any size one of these feature vectors can be used. But, if the system is restricted to use a fixed size of 39, MFCC_0_D_A-lly can be considered as the candidate feature vector to get a performance close to the best one (the performance differs approximately by 1% in the experiments.).

2.4.2 Analysis of Segmentation Errors According to Phoneme Classes

2.4.2.1 Acoustic Phoneme Classes

The segmentation results using feature vectors 8-17 were investigated in a phoneme class based manner. For each class, the feature vector yielding the minimum average absolute error is found and compared to the baseline system. The phonemes are clustered into 5 classes as described in Section 2.2.4 regarding their acoustical properties, as;

The phonemes are grouped acoustically as follows;

- **Vowels:** /aa/, /iy/, /i/, /@/, /ii/, /oo/, /ei/, /ou/, /@@/, /uu/, /ai/, /o/, /e/, /eir/, /a/, /i@/, /u/, /oi/, /w/, /ow/, /uh/, /v/, /y/;
- **Plosives:** /b/, /p/, /d/, /t/, /g/, /dh/, /th/, /ng/, /k/;

- **Fricatives:** /f/, /s/, /sh/, /z/, /zh/, /h/, /ch/, /jh/;
- **Nasals:**/m/, /n/;
- **Liquids:** /r/, /l/;

The affricates are included in the fricatives class and the glides are included in the vowels class. Silence is considered as the sixth class. The average absolute errors of class to class boundaries are calculated and the percent deviations of the average absolute errors from those of the baseline system are given in Table 2.2. For example, the feature vector resulting in the minimum average absolute error and the change in the average absolute error for vowel-silence boundary is shown in the cell of the vowel-row and silence-column; 8th feature vector resulted the minimum average absolute error, the baseline average absolute error for vowel-silence boundary is 9.6 ms and it becomes 7.6 ms when MFCC_0_D_A+lly is used, which corresponds to approximately 21% decrease in the average absolute error. The values in parentheses are the number of occurrences of that boundary class in the training set and in the test set, respectively. The feature vectors including horizontal positions of upper lip and lower lip are used in the experiments for the sake of completeness. However it is hard to find horizontal positions using visual information, because of this the next best result without using horizontal positions are also given in Table 2.2.

It is observed that the segmentation performances on all boundary classes *starting* with a vowel are more or less increased. It is seen that relatively larger number of training and testing samples are available for those cases. In particular, vowel-vowel and vowel-liquid transitions of the articulator system are dominated by the movement of inner articulators but not by the movement of the lips. These boundaries are very hard to label for even manual labelers. For those transitions at which the lip motion is not so salient, a significant contribution of lip motion features is not expected. This is validated by the small gain in performance (3.6% and 4.9%, respectively) for vowel-vowel and vowel-liquid boundaries.

Table 2.2: In each cell, the first line is the number of the feature vector yielding minimum absolute average error, and the second line is the percent decrease with respect to baseline for each acoustic class to class boundary and the 3rd line is the numbers of training and test data. For each cell, row id is the left phonetic class type and the column id is the right phonetic class type.

	Silence	Vowel	Plosive	Liquid	Fricative	Nasal
Silence	1 - (0/0)	11 12.5% (116/15)	8 33.5% (204/15)	1 - (18/0)	1 - (66/8)	11 23.0% (16/2)
Vowel	8 20.7% (50/10)	17 3.6% (1140/117)	13 7.5% (1715/107)	13 4.9% (615/66)	10 9.8% (1111/78)	14-8 45.4- 40.6% (936/74)
Plosive	1 - (114/7)	10 15.5% (1922/118)	1 - (363/16)	9 4.1% (430/18)	15-10 38.7- 28.1% (320/16)	8 75% (60/4)
Liquid	1 - (58/4)	16-10 18.2- 17.5% (952/75)	1 - (92/8)	1 - (9/2)	1 - (61/6)	1 - (14/0)
Fricative	14-10 63.7- 57.5% (148/14)	15-10 16.7-9.7% (948/77)	17 45.8% (448/28)	17 24.2% (89/6)	1 - (131/9)	11 14.3% (115/9)
Nasal	10 25.9% (50/5)	14-10 39.8- 36.2% (489/50)	13 33.0% (387/15)	1 - (25/3)	15-10 20.0- 16.7% (190/16)	9 19.3% (27/3)

The phoneme class couples that have sufficient data and accompanied by a significant change in the performance (more than 10% increase or decrease) are selected and related comments are given below.

The 45.4% increase in the performance for vowel-nasal boundary can be seemed to be surprising, as it is known that for this boundary class the most dominant change in the articulator system is the opening of the nasal cavity. However, it is observed that this change in the articulator system is accompanied by lip movements when ending or starting a vowel (Figure 2.9). Similarly, 39.8% increase in segmentation accuracy for nasal-vowel boundaries can be justified accordingly.

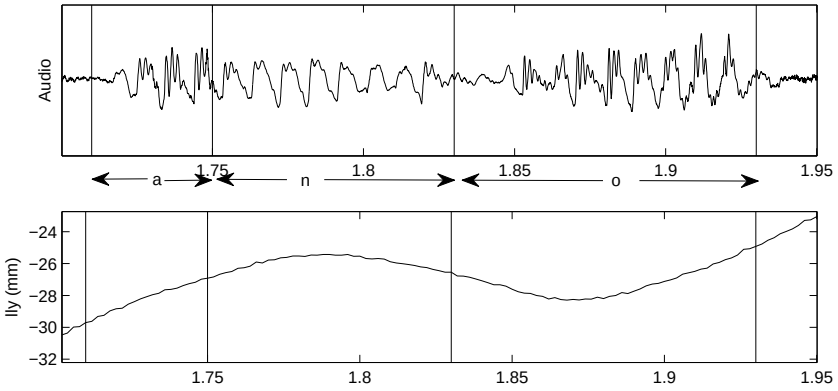

Figure 2.9: Acoustic waveform (upper panel) and the vertical position of lower lip (lower panel) for a-n and n-o boundary.

The biggest increase in the segmentation accuracy is observed in plosive-nasal boundary (75% decrease in average absolute error.) where there are only 4 test samples. It is seen that the segmentation accuracy for the nasal-plosive boundary is also increased by 33.0% as evaluated over 15 test samples. An example of /p/-/n/ boundary is shown in Figure 2.10.

In fricative-silence boundaries average absolute error is decreased by 63.7%, that is the second best degradation in average absolute boundary achieved by adding lly feature. The fricative-silence boundaries in the database are mostly consisting of /z/-/sil/ and /s/-/sil/ boundaries. Examining these boundaries, it is observed that the position of the boundaries mostly coincide with local minima of the vertical position of lower lip. An example of this is given in Figure 2.11. One can expect a similar performance for silence-fricative boundary as well, but baseline feature vector results in the best average absolute error for this case. This result may not be so well-built as there are only 8 test samples for this boundary class. Examining these samples it is observed that the degradation in the performance is because of the 29% increase in the average absolute error for /sil/-/h/ samples (3 of 7 samples) where the lip motion is weak.

For those boundary types that have richer training and testing datasets (so that related results are more reliable), such as vowel- fricative, fricative-vowel, plosive-vowel, liquid -vowel, fricative-plosive, nasal-plosive, nasal-vowel, vowel-

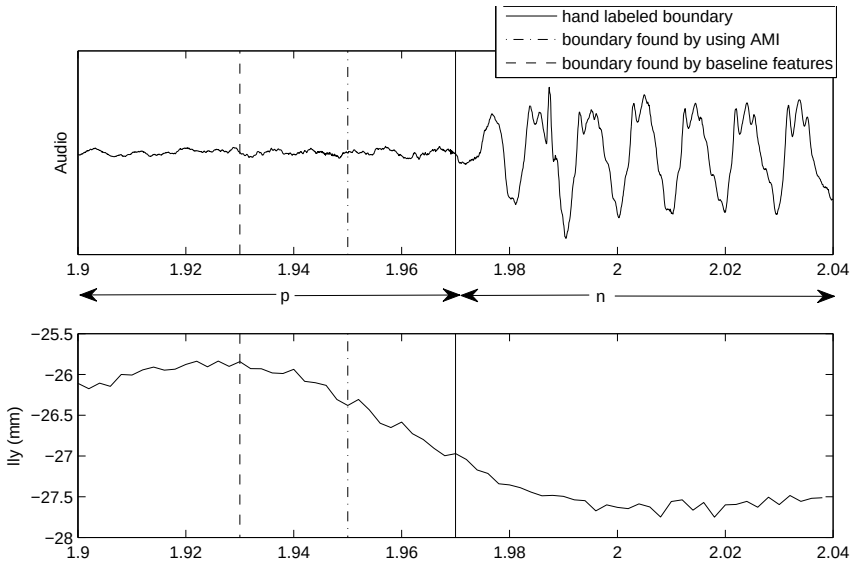

Figure 2.10: Acoustic waveform (upper panel) and the vertical position of lower lip (lower panel) for p-n boundary.

nasal boundaries, average absolute errors are decreased by 9.8% to 40.6% (45.4% with llx). On the other hand, examining Table 2.2 one can see that the class to class boundaries, at which AMI features do not reduce the average absolute errors, have insufficient test and training data ((114,7), (66,8), (115,9), (61,6) (92,8), (89,6)). So it can be stated that for the boundaries that have enough training and test data there is not a significant decrease in segmentation accuracy due to the use of lip motion information.

2.4.2.2 Visual Phoneme Classes

The phoneme classes used in Section 2.4.2.1 were formed using the acoustical properties of phonemes. In an attempt to use visual information for segmentation, it is reasonable to investigate the performance of AMI with respect to visual phoneme classes. The vowels are divided into two classes as: rounded vowels and unrounded vowels. The consonants are divided into 3 classes as: (1) bilabial and labio-dental, (2) dental and alveolar, (3) palatal, velar and glottal consonants. The resulting visual phoneme classes are;

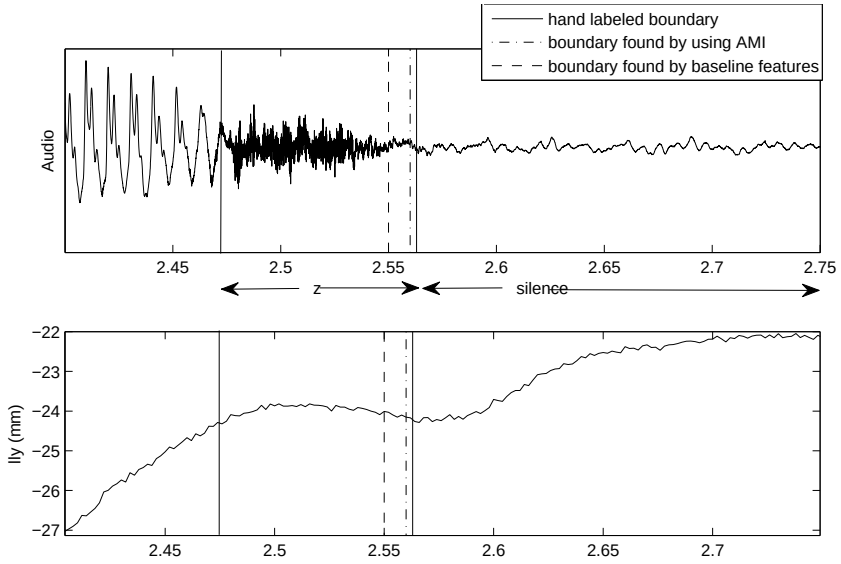

Figure 2.11: Acoustic waveform (upper panel) and the vertical position of lower lip (lower panel) for z-silence boundary.

- **Rounded Vowels:** /oo/, /ou/, /uu/, /o/, /oi/, /ow/, /uh/, /u/, /w/, /v/;
- **Unrounded Vowels:** /aa/, /iy/, /i/, /@/, /ii/, /ei/, /@@/, /ai/, /e/, /eir/, /a/, /i@/, /y/;
- **Bilabial and Labio-dental consonants:** /b/, /p/, /m/, /f/;
- **Dental and Alveolar Consonants:** /t/, /dh/, /th/, /n/, /s/, /sh/, /z/, /zh/, /r/, /l/, /d/;
- **Palatal, Velar and Glottal Consonants:** /h/, /ch/, /jh/, /g/, /ng/, /k/;

The method used in Section 2.4.2.1 is repeated for these new classes. The results are given in Table 2.3.

It is observed that, similar to the results in Section 2.4.2.1, at the class to class boundaries with sufficient training and test data the average absolute boundary error is decreased. There are only three exceptions to this; silence-dental

Table 2.3: In each cell, the first line is the number of the feature vector yielding minimum absolute average error, and the second line is the percent decrease with respect to baseline for each visual class to class boundary and the 3rd line is the numbers of training and test data. For each cell, row id is the left phonetic class type and the column id is the right phonetic class type.

	Silence	Rounded vowels	Unrounded vowels	Bilabial and labio-dental consonants	Dental and alveolar consonants	Palatal glottal and velar consonants
Silence	1 - (0/0)	12 19.2% (37/7)	11 12.3% (79/8)	8 17.4% (48/5)	1 - (190/12)	1 - (66/8)
Rounded vowels	11 86.9% (8/2)	8 18.6% (135/15)	12 13.3% (374/34)	14-8 52.9-45.3% (196/15)	1 - (727/58)	8 13.1% (171/7)
Unrounded vowels	10 25.4% (42/8)	17 17.0% (418/48)	8 40.5% (213/20)	17 12.2% (567/40)	15-10 22.9-21.6% (209/172)	10 18.3% (624/33)
Bilabial and labio-dental consonants	10 66.6% (31/3)	11 26.2% (206/10)	10 15.9% (583/45)	1 - (86/4)	10 19.0% (345/21)	9 31.6% (30/3)
Dental and alveolar consonants	14-10 28.4-25.3% (302/24)	10 38.2% (573/40)	15-10 21.2-17.7% (2235/183)	11 4.7% (334/20)	10 10.4% (1269/74)	8 16.4% (243/10)
Palatal glottal and velar consonants	1 - (37/3)	8 33.2% (242/11)	8 11.6% (472/31)	17 9.4 (50/2)	1 - (333/14)	1 - (7/1)

and alveolar consonants boundary, rounded vowels-dental and alveolar consonants boundary and palatal glottal and velar consonants- dental and alveolar consonants boundary. In all three cases, the boundaries end with a dental or alveolar consonant. This does not sound strange since the lip motion is not so dominant at these boundary types. The biggest decrease in the average absolute error is achieved in rounded vowels-bilabial and labio-dental consonants boundary (52.9%). This is reasonable as the lips motion is very definite for

this boundary type. There are also considerable decreases in unrounded vowels-unrounded vowels (40.5%) and dental and alveolar consonants-rounded vowels (38.2%) boundaries.

Table 2.4: Average Absolute Segment Boundary Errors

System	Average Error(ms)
Baseline system	9.90
AMI included using acoustical classes (without horizontal positions)	8.35 (15.6%)
AMI included using acoustical classes	8.24 (16.8%)
AMI included using visual classes (without horizontal positions)	8.26 (16.6%)
AMI included using visual classes	8.12 (18.0%)

2.4.3 The use of AMI in Segmentation in a Phoneme Class Based Manner

Lastly, after analyzing the boundary errors in a phoneme class based manner, for a particular boundary, it is considered to use the feature vector that performs best for that type of boundary. In most of the practical cases the text transcription of the acoustic data to be segmented is available. Hence knowing the boundary types for which AMI improves segmentation accuracy, it may be possible to use relevant AMI features for these boundaries. This can be considered as a ROVER like approach [62]. However since the boundary type information is available decision mechanism for the type of feature to be used is simpler. The results for both acoustic and visual phoneme classes are given in Table 2.4. In the table results are given for the feature vectors both including and excluding vertical positions of the lips.

2.5 Discussion

The concept of bimodal automatic speech segmentation is introduced in this chapter. The introduction of the visual modality to speech segmentation was achieved by using the positions of visible coils, which are the three coils on the upper lip, lower lip and jaw, from the MOCHA-TIMIT database. The EMA recordings include vertical and horizontal-forward positions of these coils alongside the acoustic recordings of the 460 English utterances.

The articulator positions and their derivatives were embedded into the MFCC feature vector in different ways. Experiments have shown that using only the vertical position of the lower lip together with 13 MFCCs and their first and second derivatives results in the best performance. The contribution of using AMI was examined with respect to the phoneme classes around the phoneme boundaries. The examination is done in two different ways; first using acoustic properties of the phonemes and second using the visual properties. The phonemes are divided into 5 phoneme classes, regarding their acoustic properties; vowels, plosives, liquids and glides, fricatives, nasals. It is observed that including AMI decreases the AABE in almost all of the boundary classes in which sufficient training and test data exist. Degraded segmentation accuracy was observed in the cases of insufficient training and/or test data. Furthermore, in some of these boundary classes where there is no apparent lip motion, this does not imply that one has to expect degradation of segmentation accuracy because of insignificant lip motion, since nondegraded segmentation results are observed for the boundaries of weak lip motion when richer training and test data are available (i.e., vowel-vowel, liquid/glide- vowel, vowel- liquid/glide boundaries.). Then, the phonemes are divided into 5 phoneme classes regarding their visual properties; (1) rounded vowels, (2)unrounded vowels, (3) bilabial and labio-dental consonants, (4) dental and alveolar consonants, (5) palatal, velar and glottal consonants. The observations were similar to the previous case. The largest decrease in average absolute boundary error is achieved in rounded vowels-bilabial and labio-dental consonants boundary.

Considering the availability of boundary type information in the segmentation

process, the results of AMI based segmentation are used selectively depending on the type of boundary class. For a particular boundary, it is considered using the feature vector that performs best for that type of boundary. Average absolute error decreases by 15.6% when acoustic classification of the phonemes is used, and average absolute error decreases by 16.6% when visual classification of the phonemes is used. The average absolute error can be decreased by 16.8% and 18.0% respectively, if horizontal positions of the lips are available.

The studies in this chapter constituted a primary step proving the benefits of integrating the visual modality to automatic speech segmentation process. The promising results achieved by using MOCHA-TIMIT database had directed the course of this thesis to seek for improvements in automatic speech segmentation using more visual information and a richer database, which leads to the studies that will be presented in Chapter 3

CHAPTER 3

AUDIOVISUAL AUTOMATIC SPEECH SEGMENTATION

3.1 Introduction

Bimodal automatic speech segmentation is investigated in Chapter 2. By integrating the positions of only the upper lip and the lower lip, a decrease of 18% in average absolute boundary error (AABE) is achieved. Although that study had shown that the visual modality could provide very beneficial information for the segmentation process, this idea should be extended by integrating different visual features that can be extracted from the visual images of the speaker during the utterance of the speech. In order to be able to test the idea of audiovisual AS, a phonetically rich audiovisual database is needed. Although there are some public databases available, they are more or less have the same size with the MOCHA-TIMIT database, also there is no Turkish audiovisual speech database that is publicly available. The lack of a public Turkish audiovisual database limits the research in audiovisual speech processing applications for Turkish. Because of this, the collection and preparation of a Turkish audiovisual speech database had been the first step of the studies in audiovisual automatic speech segmentation. After the recording of audiovisual data, manual segmentation had to be done in order to have a ground truth data for the AS experiments. Visual data is also prepared to be used in AS system by extracting several shape based and appearance based visual feature vectors. Following the preparation of the database procedures similar to the ones used in Chapter 2 are applied in order to assess the benefits of the integration of the visual modality to automatic

speech segmentation.

This chapter is organized as follows. In Section 3.2 the collection and preparation of the Turkish audiovisual speech database is represented. The preparation of the text corpus, the properties of the acoustic and visual data is presented in Section 3.2.1 and Section 3.2.2. Manual segmentation of the database and the proposed method to decrease the intralabeler inconsistency in manual segmentation are discussed in Section 3.2.3. After that, the detection and tracking of the markers and normalization process are explained in Section 3.2.4 and Section 3.2.5. In Section 3.3 the visual features to be used in the experiments are introduced. The audiovisual AS system is proposed and several audiovisual feature vectors that are used in the experiments are introduced in Section 3.4.1 and in Section 3.4.2. The improvements achieved by using this system are represented in Section 3.4.3, Section 3.4.4 and Section 3.4.5. The discussion about the Chapter is made in Section 3.5.

3.2 Database Preparation

3.2.1 Preparation of the Text Corpus

The first step of building a speech database is the preparation of the text corpus that will be uttered by the speakers. A phonetically rich and balanced text corpus is compulsory for a useful speech database. A text corpus that is composed of 1600 Turkish sentences is prepared.

The sentences in the text corpus are annotated using the Speech Assessment Methods Phonetic Alphabet (SAMPA) for Turkish [64]. SAMPA notation has 42 phonemes (including eight vowels with length mark) for representing Turkish sounds [2]. Text corpus contains 85266 bigrams. The statistics and the sufficiency of the database is observed with respect to acoustic and visual phoneme classes, concerning the work done in Chapter 2. The numbers of the bigrams in means of acoustic phoneme classes in the database are shown in Table 3.1 and the

[2]Phonetic symbols are described in Appendix B

numbers of the bigrams in means of visual phoneme classes in the database are shown in Table 3.2.

The phonemes are grouped acoustically as follows;

- **Vowels:** /a/,/ax/, /e/, /ex/, /I/, /Ix/, /i/, /ix/, /o/, /ox/, /O/, /Ox/, /u/, /ux/, /y/, /yx/ ;
- **Plosives:** /b/, /p/, /d/, /t/, /g/, /gj/, /k/, /c/;
- **Fricatives:** /f/,/h/,/v/,/s/,/S/,/z/,/Z/,/tS/,/dZ/;
- **Nasals:** /m/, /n/, /N/;
- **Liquids and glides:** /G/,/r/,/L/,/j/,/l/,/w/;

The database seems to have 'sufficient' number of occurrences for most of the bigram classes for either acoustic or visual classification cases. Only bigram classes starting with a rounded vowel and ending with an unrounded vowel or a rounded vowel have lower number of occurrences with respect to other bigram classes for the visual classification case. For example, most of the visual bigram classes have thousands of occurrences in the text corpus.

Table 3.1: The number of occurrences of acoustic bigram classes. Row id denotes the phonetic class type of the starting phoneme and the column id denotes the phonetic class type of second phoneme.

	Vowels	**Plosives**	**Fricatives**	**Nasals**	**Liquids-glides**
Vowels	10088	8284	5919	7765	9223
Plosives	11905	923	335	287	795
Fricatives	6227	952	189	204	439
Nasals	5604	1779	821	278	569
Liquids-glides	7587	1754	594	575	508

The phonemes are grouped according to visual properties as follows;

- **Rounded vowels:** /o/, /ox/, /O/, /Ox/, /u/, /ux/, /y/ ;

Table 3.2: The number of occurrences of visual bigram classes. Row id denotes the visual class type of the starting phoneme and the column id denotes the visual class type of second phoneme.

	Rounded vowels	Unrounded vowels	Bilabial and labio-dental consonants	Dental and alveolar consonants	Palatal glottal and velar consonants
Rounded vowels	113	132	759	2482	4279
Unrounded vowels	501	758	3683	9660	13695
Bilabial and labio-dental consonants	1173	5012	201	348	561
Dental and alveolar consonants	2580	9531	929	1978	2088
Palatal, glottal and velar consonants	3352	13123	1319	2497	2850

- **Unrounded vowels:** /a/, /ax/, /e/, /ex/, /I/, /Ix/, /i/, /ix/;
- **Bilabial and labio-dental consonants:** /p/ , /b/, /m/, /f/, /v/;
- **Dental and alveolar consonants:** /t/, /d/, /s/, /z/, /n/, /tS/, /Z/;
- **Palatal glottal and velar consonants:** /dZ/, /g/, /gj/, /h/, /k/, /c/, /l/, /r/, /S/ /j/;

3.2.2 Acoustic and Visual Data

A Matlab program with a graphical user interface is developed for collecting the database. The user articulates the sentence displayed the screen after pressing the "record/stop" button, and stops the recording upon finishing the sentence. The recording can be repeated, or can be continued to the next sentence by pressing the "next" button. A screenshot of the program is given in Figure 3.1.

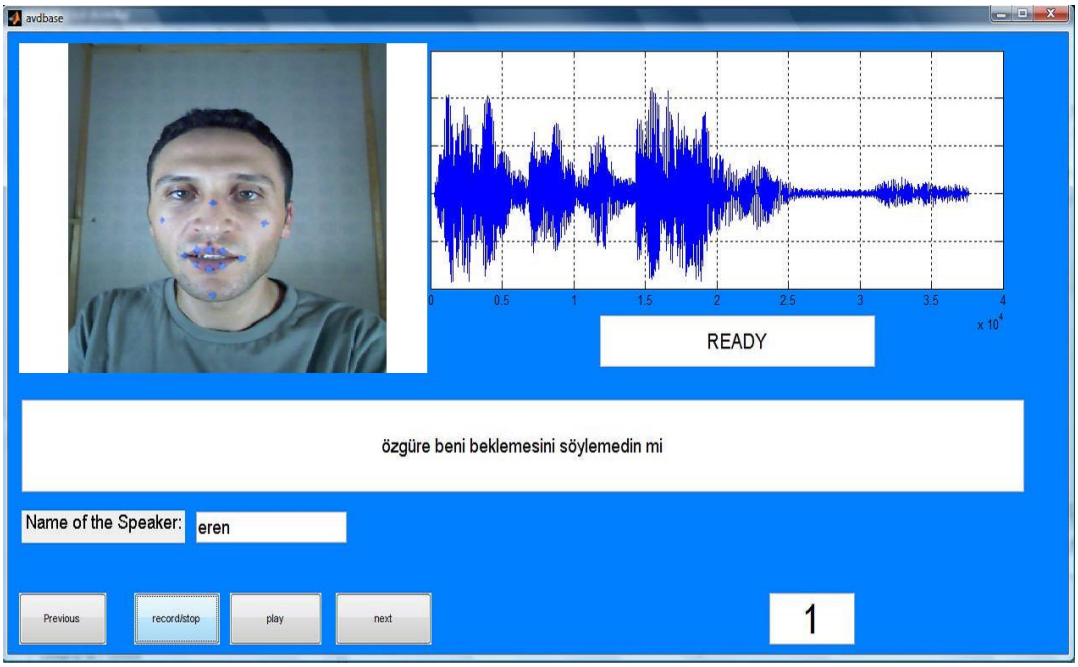

Figure 3.1: Graphical User Interface for the Recording of the Database

The utterances are recorded in a sound proof recording cabin (Figure 3.2), using Edirol UA-1000 audio capture device. The audio files are recorded at 16 kHz with 16-bit resolution. The cabin is illuminated with two white light bulbs on two sides to decrease the shadows on the speaker's face and have clear visual records. The visual data is recorded at 30 fps at 800×600 resolution using a commercially available Philips SPC1300 webcam.

12 blue markers are placed on the speaker's face (Fig 3.3, Fig 3.4). The placement of the markers enables easier and more accurate detection of the position and the shape of the lips, chin and the speaker's head. Three markers (on the nose and cheeks) are used to calibrate the data, i.e., compensate the head movements, change in the position of the speaker etc. There are 8 markers located on the lips and one marker at the chin to capture the movement of the chin. The visual data are recorded at 30 frames/s, with a resolution of 800×600 pixels, and stored as image files (*.jpg*). Our previous research using the MOCHA database had shown that, the highest frequency component of the visual articulators is about 10 Hz, so data acquisition rate of 30 fps is adequate.

Figure 3.2: The soundproof recording cabin and the recording environment.

To sum up, the database consists of 1600 Turkish sentences uttered by a male speaker. The duration of the acoustic data is approximately 160 minutes, and visual data is composed of approximately 23700 frames (800×600 pixels). For each sentence, one *.wav* file and the corresponding image files (*.jpg*) are stored in a directory named as the speaker's name concatenated with utterance number.

3.2.3 Manual Segmentation of the Database

The audiovisual speech database should be manually segmented because of the need for a ground truth for the automatic speech segmentation system to be developed. Besides the inevitable disadvantage of being time consuming and expensive, manual segmentation also suffers from the inconsistency problem. Interlabeler inconsistency is the difference between the boundary marks of different segmenters for the same boundaries and intralabeler inconsistency is the

Figure 3.3: A Series of Frames Captured

variation of the boundary points of similar boundary types that are marked by the same segmenter at different times. It is not possible to get rid of the inconsistency problem in manual segmentation, but it is tried to be minimized by setting some ground rules describing where the boundary point should be marked between two phonetic units. A widely used approach in the literature is mentioned in [63]. Two principal rules used in this approach are:

Rule 1) The boundaries that can be found unambiguously by visual (speech waveform/spectrogram) and audio inspection (listening) are marked directly.

Rule 2) For the ambiguous cases, speech is listened at a window placed on the left and on the right side of the hypothesized boundary position and the window is iteratively extended until each phonetic unit is perceived, respectively.

In our work, we extended these rules to decrease the interlabeler and intralabeler inconsistency further. The additional rules for marking the boundary points manually are listed below;

- **Change in the source type:** The type of the input/excitation differs for different phoneme types. The excitation is periodic for voiced phonemes, impulsive for plosives and noisy for fricatives. If a change in the type of the excitation is detected at the speech waveform, the point where the change occurred is marked as the boundary point (Figure 3.5).
- **Instant change at the vocal tract shape:** The vocal tract has the function of shaping the input from the source to form the speech waveform. The change in the vocal tract is not discrete; it changes from phoneme to phoneme in a gradual manner. However, at some phoneme to phoneme

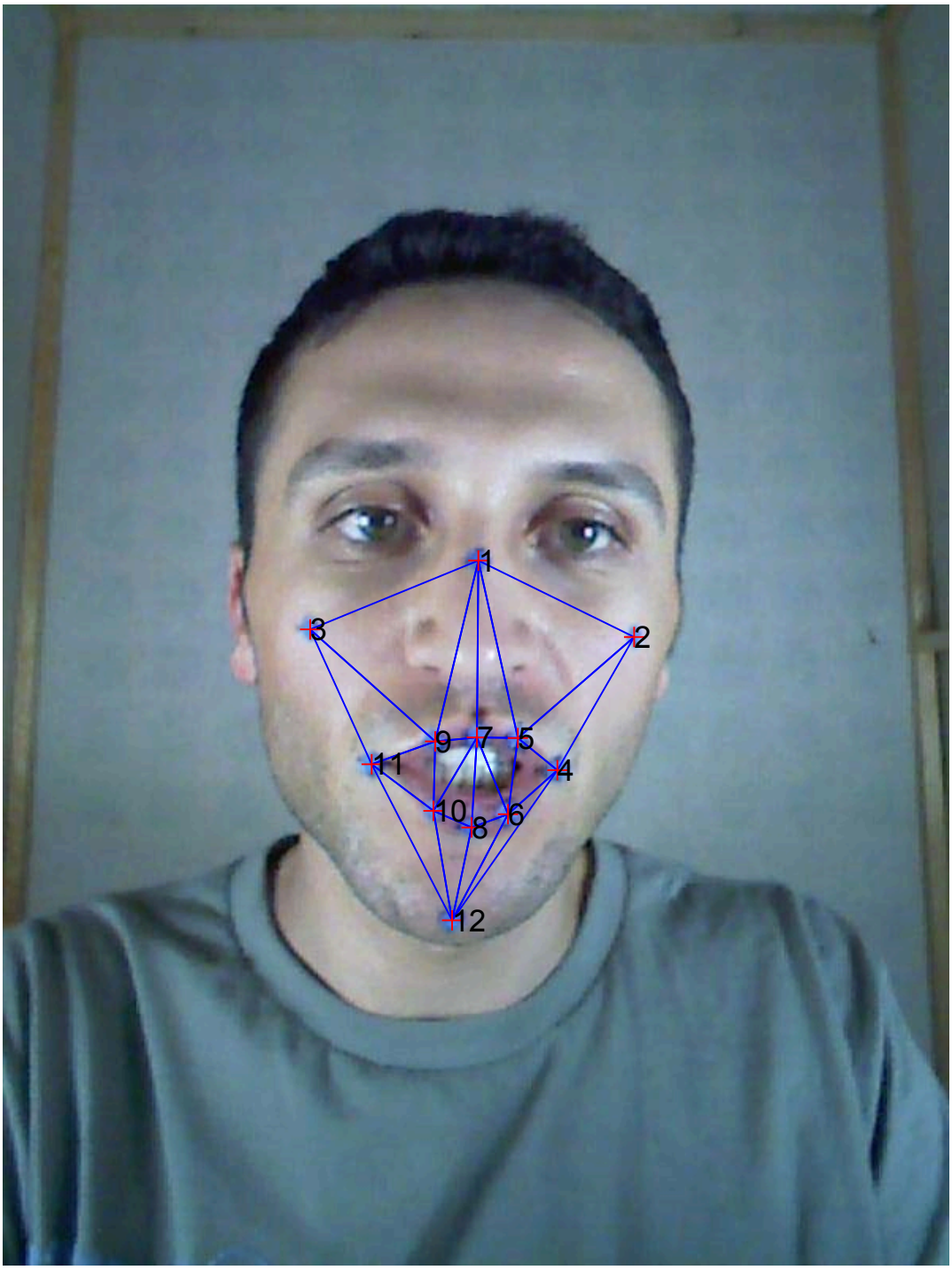

Figure 3.4: Markers on the Speaker's Face

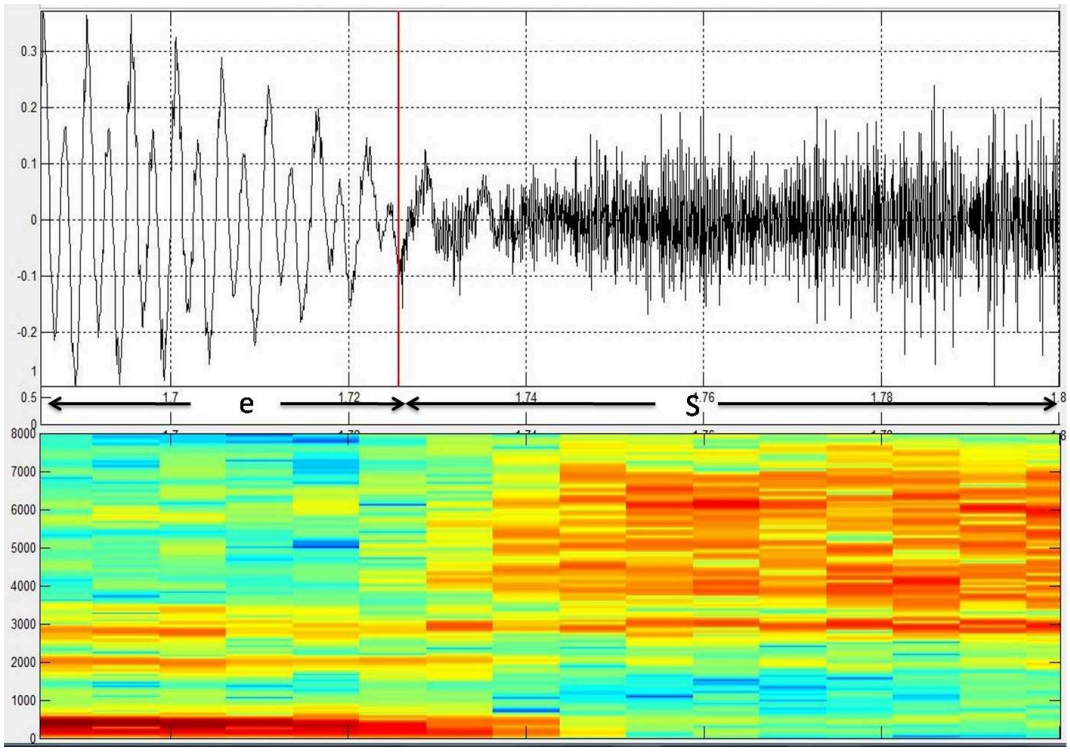

Figure 3.5: A vowel-fricative boundary (/e/-/S/). The red line shows the manually marked boundary location.

boundaries an instant change at frequency domain and/or time domain can be observed. Vowel-Nasal boundaries are good examples for this. At these boundaries, the opening of the vellum is instantaneous and this introduces new zeros to the vocal tract response function immediately, which changes the output waveform significantly. At these type of phoneme couples, the boundary point should be marked as the instant where the significant change in the time and/or frequency domain is observed (Figure 3.6).

- **Ambiguous cases:** At some of the boundary types, especially at the ones between similar phonemes, the change in the characteristics of the speech waveform is gradual. This makes the detection of the boundary point very difficult. At these type of boundaries rule 2 stated above is used to delimit the location of the boundary point. After the margins for the possible location of the boundary are found, each quasi-periodic speech segment between glottal closure instants (GCIs) are considered. The point where the 'difference' between two consecutive quasi-periodic segments is maxi-

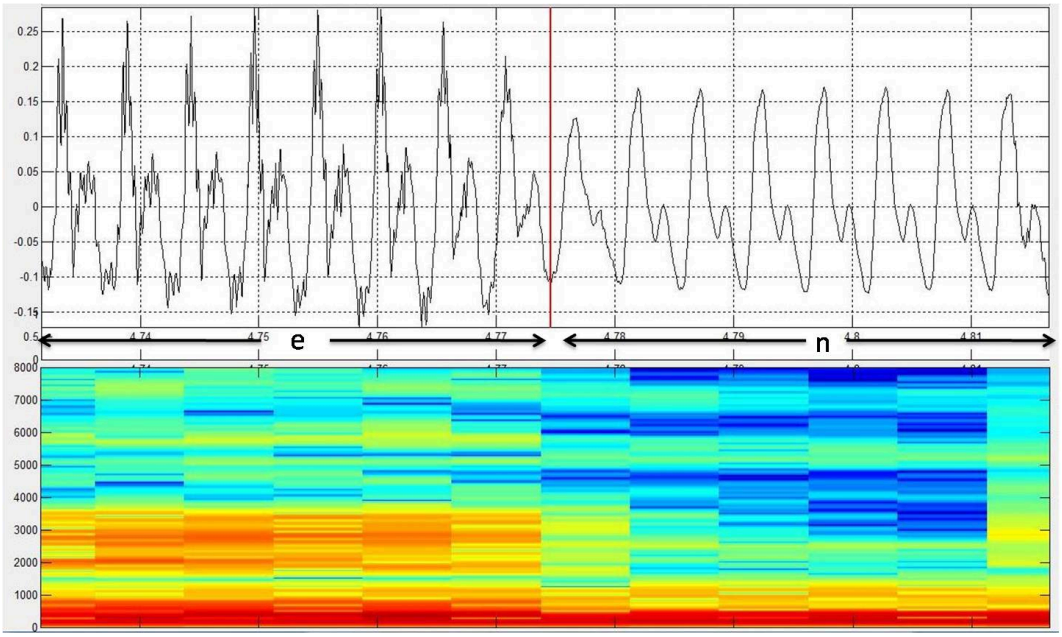

Figure 3.6: A vowel-nasal boundary (/e/-/n/). The red line shows the manually marked boundary location.

mum is marked as the boundary point (Figure 3.7). Perhaps these types of boundaries are the ones, where interlabeler, and intralabeler differences are maximum, and also where the automatic segmentation systems produce greatest errors.

In addition to applying the rules stated above, the manual segmenters should have extensive knowledge about the phonemes, phoneme types, manner of articulation and place of articulation of different phonemes and speech production process in order not to get confused at some ambiguous boundaries. Two senior electrical and electronics engineering students are lectured about basic speech concepts such as speech production, phoneme types, place of articulation and manner of articulation of different phonemes and speech segmentation, for a one month period for this purpose. Afterwards, the trainees manually segmented a 75 sentence training set together, under supervision of the writer of this thesis and then, they are allowed to handle the marking of the boundaries at the database. The supervision is continuously provided during the segmentation of the database. Each manual segmenter marked 43000 of the 86000 boundaries

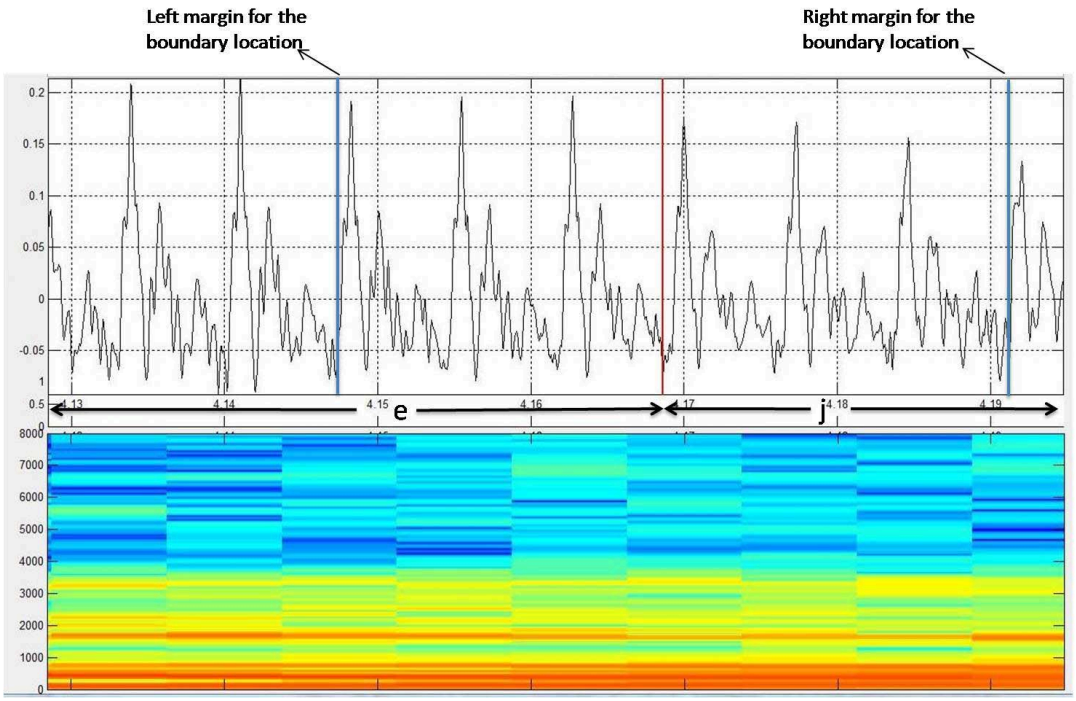

Figure 3.7: A vowel-liquid boundary (/e/-/j/). The red line shows the manually marked boundary location.

in the database. The initial phonetic boundaries are found using HTK in order to speed up the segmentation process and increase the accuracy. Some examples about locating the boundary at different boundary types will be presented below;

- **(Vowel/Nasal/Liquid)-Plosive boundaries:** The plosives are composed of three parts; closure, burst, and aspiration. The lips are closed when articulating a plosive sound in order to increase pressure in the mouth to be able to produce a burst. The existence of closure state makes it easy to detect the starting of plosives. For the unvoiced plosives (/p/, /t/, /k/) at the closure part the amplitude of the speech waveform decreases almost rapidly, that point should be marked as the point. In some cases some sinusoid-like oscillations which seem to be very different from the previous voiced phoneme can be observed after the closure of the lips, in that case the boundary should be marked as the starting of these oscillations (Figure 3.8.a).

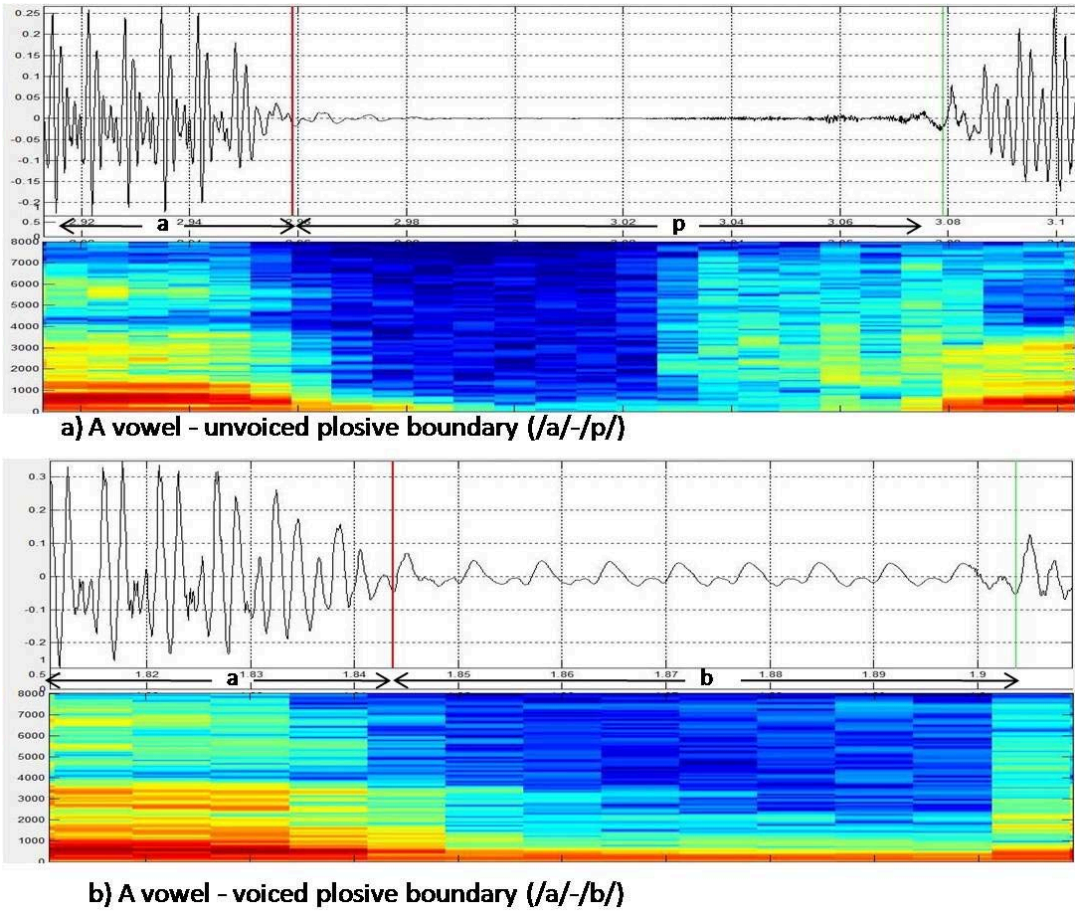

Figure 3.8: Vowel-plosive boundaries. The red lines show the manually marked boundary locations.

The oscillations after the closure of the lips are observed from beginning to end of the closure part of the phoneme for the voiced plosives (/b/, /d/, /g/). For most of the cases the transition is significant, the high frequency component vanishes almost immediately. In that case the transition point should be marked as the boundary point (Figure 3.8.b). Unlike the other voiced phonemes to plosive boundaries, at nasal-(voiced plosive) boundaries the transition may not be observed clearly in some cases as the shape of the waveform is simple and lack the high frequency components for both nasals and burst phase of the voiced plosives (Figure 3.9). In that case the procedure for the ambiguous boundaries which was described above should be used to mark the boundary point.

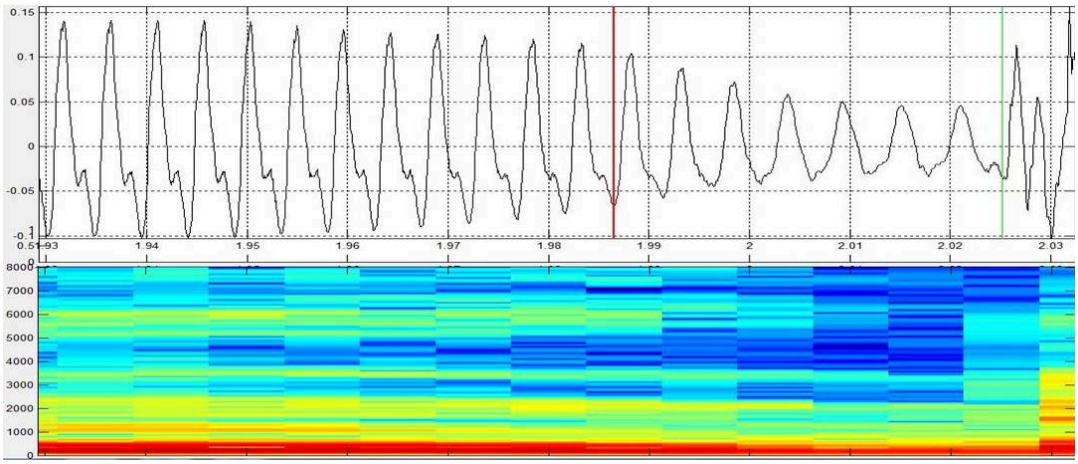

Figure 3.9: A nasal - plosive boundary

3.2.3.1 Boundary Class-wise Segmentation

Two major problems of manual speech segmentation are; it is a very time consuming process (It is known that it takes 100-200 times of real time to manually segment and align a database [13, 14].), and interlabeler and intralabeler inconsistency. To overcome interlabeler inconsistency problem the number of manual segmenters was kept as minimum (2 manual segmenters.) and the segmenters are trained together in a one month long training. The author of this thesis had been present as a supervisor during the segmentation process and the interaction between the segmenters about the 'challenging boundary types' had been encouraged during the manual segmentation process.

To increase the intralabeler consistency of the manual segmenters and to speed up the manual segmentation process, a new concept of boundary class wise manual segmentation is proposed. The phonemes are divided into 5 acoustic classes as described in Section 3.2.1. After marking a boundary, the user interface makes the manual segmenter mark similar boundaries until marking the boundaries belonging to that type of bigram class is finished. For example, when the manual segmenter starts to mark the boundary between phonemes /a/ and /b/, he or she continues to mark the boundaries between these phonemes until segmentation of all /a/-/b/ boundaries in the database are done. When manual segmentation

of /a/-/b/ boundaries is finished, the program continues with another phoneme couple belonging to /class(a)/-/class(b)/ (i.e. /vowel/-/plosive/). Dealing with the same type of boundaries repeatedly, will help the user on making similar and more consistent decisions about the locations of the boundary points, and also focusing in similar boundaries will speed up the segmentation process. Although there exists no accepted method to measure intralabeler inconsistency, depending on our experience on manual segmentation, we can strongly claim that intralabeler consistency was increased significantly, using this approach. This approach also decreases the time needed for manual segmentation of the database. To the best of author's knowledge, an approach like this is not used before in manual segmentation process.

A Matlab program with a graphical user interface is prepared for boundary class based manual segmentation process. A screen capture of the manual segmentation user interface is shown in Figure 3.10.

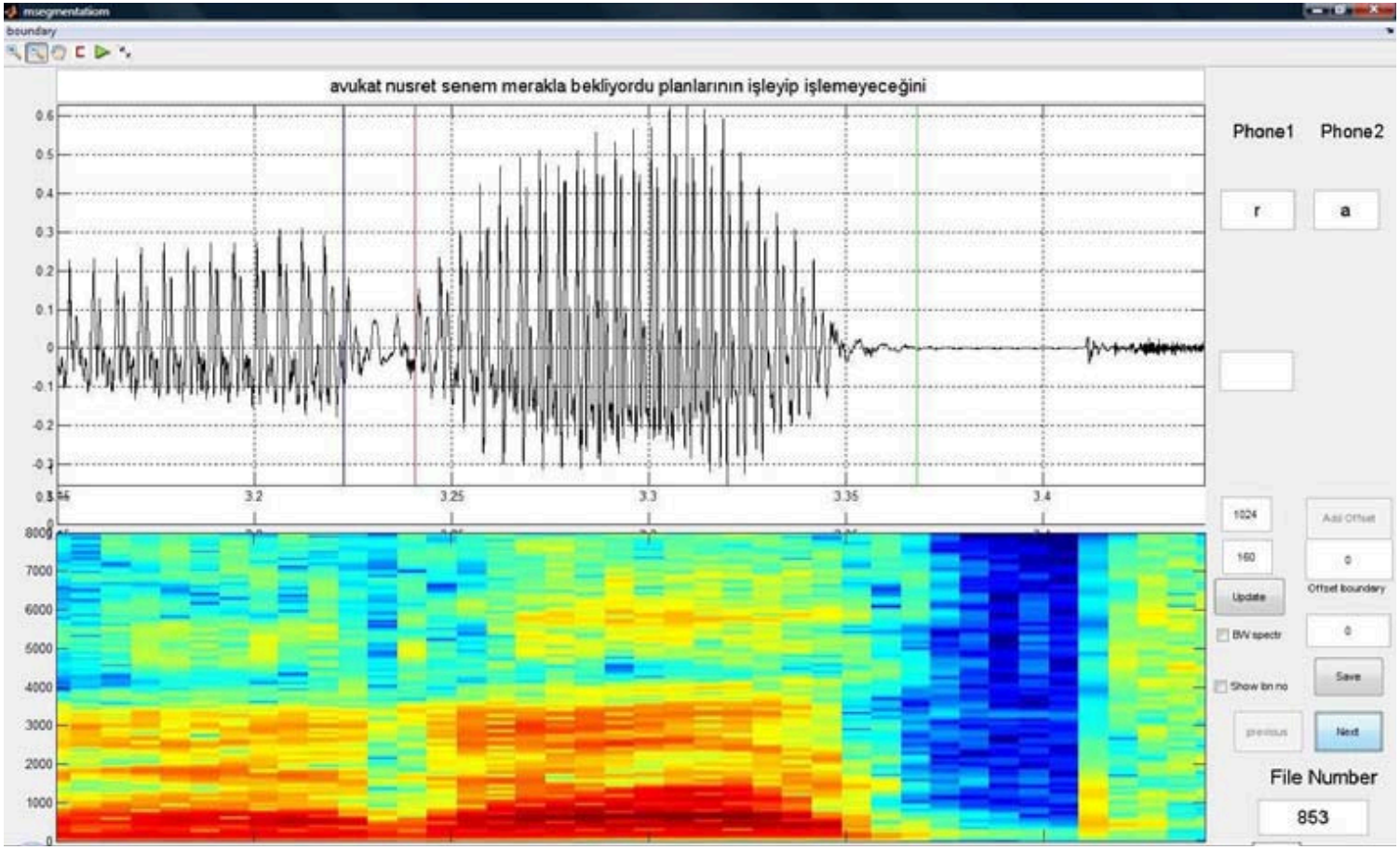

Figure 3.10: User interface for the manual segmentation of the database

3.2.3.2 Interlabeler Inconsistency

Each of the two manual segmenters aligned one half of the database, their work included 75 utterances in common. The inconsistency between these segmenters is obtained using the manual segmentation results of these utterances. The mean error, average absolute boundary error (AABE) and variance of the error between the manually segmented boundaries belonging to different segmenters are shown in Table 3.3. AABE is used as a metric to compare different segmentation results in this study. The comparison can be made between different manual segmentation results to find the interlabeler inconsistency or between manual segmentation results and automatic segmentation results to assess the performance of corresponding AS system. The calculation of the AABE is shown in equation 3.1 for interlabeler inconsistency case and in equation 3.2 for performance measurement of an AS system. The AABE between the manual segmeters, and the mean and the variance of the boundary differences between them are presented in Table 3.3.

$$InterlabelerAABE = \frac{1}{N_{bint}} \sum_{j=1}^{N_{uttint}} \sum_{k=1}^{N_j} |t_{s1}(j,k) - t_{s2}(j,k)| \tag{3.1}$$

$$AABE = \frac{1}{N_b} \sum_{j=1}^{N_{utt}} \sum_{k=1}^{N_j} |t_m(j,k) - t_a(j,k)| \tag{3.2}$$

Where,

N_j=Number of phonetic boundaries in j^{th} utterance

N_{utt}=Number of all utterances

N_{uttint}=Number of utterances segmented by both segmenters

N_b=Number of phonetic boundaries in the database

N_{bint}=Number of phonetic boundaries segmented by both segmenters

$t_{s1}(j,k)$= k^{th} manual segmentation boundary at the j^{th} utterance marked by the first segmenter

$t_{s2}(j,k)$= k^{th} manual segmentation boundary at the j^{th} utterance marked by the second segmenter

$t_m(j,k)$= k^{th} manual segmentation boundary at the j^{th} utterance

$t_a(j,k)$= k^{th} boundary found by AS system at the j^{th} utterance

$$N_b = \sum_{j=1}^{N_{utt}} N_j \tag{3.3}$$

Table 3.3: The mean error, average absolute error and variance of the error between the manual segmentation results of the two segmenters.

Mean	Average absolute error	Variance
2.43ms	9.09ms	4.56e-004

3.2.4 Detection and Tracking of the Markers on the Speaker's Face

12 blue markers were located on the speaker's face during the recording of the visual data in the database. The position information of these markers enables more accurate and more precise extraction of the visual features. These positions will be used in the extraction of both shape based visual features and appearance based visual features which will be discussed in Section 3.3.

The color of the markers is selected as blue in order to allow a chroma key approach to detect the positions of the markers, as blue is the most unlikely color to be found on a human face unlike red and yellow. The detection of the blue pixels is needed in order to find the positions of these markers. The identification of the colors can be tricky in RGB format as different illumination conditions result in completely different RGB values for the same color. The images at RGB format are converted to YC_bC_r formatted images in order to get rid of the illumination problem [65, 66]. In YC_bC_r format, Y is the luminance value, C_b is the difference from blue, C_r is the difference from red (Equations 3.4,3.5,3.6).

$$Y = K_r * R' + (1 - K_r - K_b) * G' + K_b * B' \tag{3.4}$$

$$C_b = 1/2 * (B' - Y')/(1 - K_b) \tag{3.5}$$

$$C_r = 1/2 * (R' - Y')/(1 - K_r) \tag{3.6}$$

Where R', G', B' are the normalized values of R,G,B, and K_b and K_r are normalization parameters usually taken as 0.114, 0.219.

The resulting luma (Y) value will then have a nominal range from 0 to 1, and the chroma (C_b and C_r) values will have a nominal range from -0.5 to +0.5. The reverse conversion process can be readily derived by inverting the above equations.

The Y values are omitted in order to get rid of the problems can be caused by different lighting conditions. The target candidates for the positions of the blue markers are found as the pixels, where the difference between C_b and C_r is maximum. Thresholding must be applied to the difference values in order to find the connected regions that have the highest C_b-C_r difference. Thresholds are decided iteratively for each image until the number of target candidates found is between 15 and 20. Note that, these target candidates include the positions of the markers and a few false alarms. A tracking algorithm should be used in order to eliminate the false alarms and track the true targets.

After the detection process, the markers should be tracked within following images. Auction algorithm [67] is used for tracking the blue markers. The algorithm uses previous tracks $(T_1[n-1], T_2[n-1], ..., T_2[n-1])$, and the track candidates for current frame $(C_1[n], C_2[n], ..., C_N[n], N > 12)$, and finds the mapping $\mathbf{M}$ (Equation 3.7) that chooses the current tracks $(T_1[n], T_2[n], ..., T_{12}[n])$ by minimizing the total distance between previous tracks and possible current tracks. A screen shot of the tracking user interface is shown in Figure 3.11. The positions of the markers are found for all of the images in the database. The position tracks for each uttered sentence are saved to a .mat file in the corresponding directory for later use.

$$\left[\hat{T}_1[n], \hat{T}_2[n], ..., \hat{T}_{12}[n]\right] = M\left(\left[C_1[n], C_2[n], ..., C_N[n]\right]\right) \tag{3.7}$$

$$C = \sum_{k=1}^{12} |\hat{T}_k[n] - T_k[n-1]|^2 \tag{3.8}$$

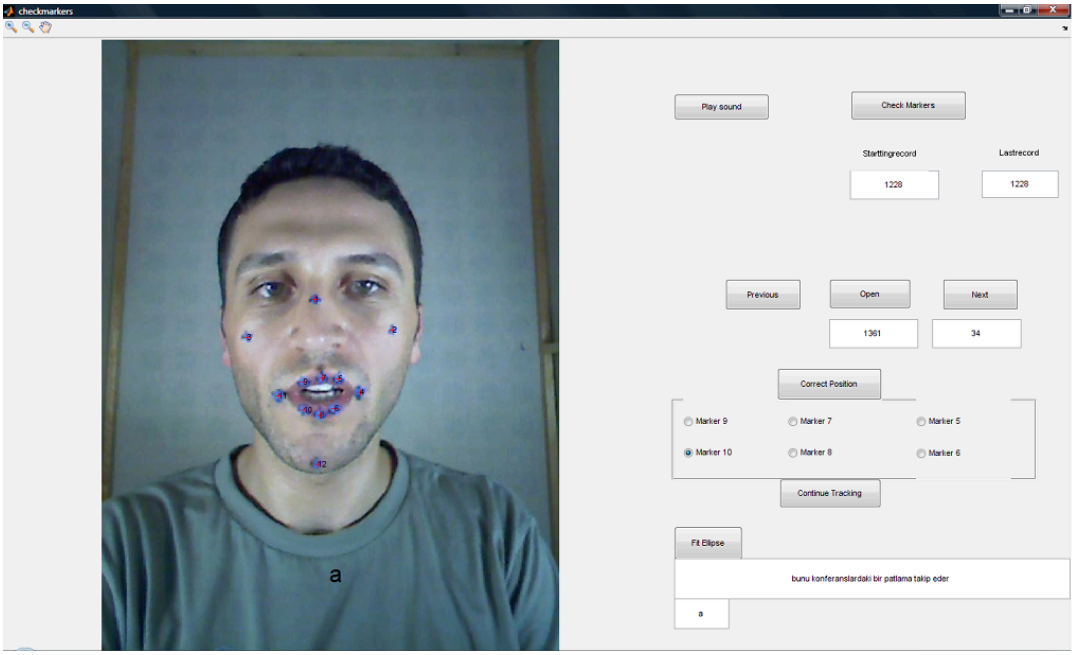

Figure 3.11: User interface for the tracking of the markers on the user's face

3.2.5 Normalization of Tracking Results

The visual data are recorded by a still camera mounted to the wall of the sound proof cabin (Figure 3.2). Although there is no camera movement, the position of the speaker and the distance between the speaker and the camera vary between different recordings and also during the same recording as well. The position data should be normalized in order to compensate the rotation and translation of the markers and to be able to capture the lip and chin motion accurately. The static markers on the face (The first three markers; one on the nose and two on the cheeks) are used for this purpose. The distance between the 2^{nd} and the 3^{rd} markers(d_{2-3}) is used to normalize the x components of the position data,

and the perpendicular distance between the 1^{st} marker and the line between the 2^{nd} and the 3^{rd} markers is used to normalize the y components of the position data. In order to be able to perform this scaling operation, the positions of the markers should be projected on a plane, where the line between the 2^{nd} and the 3^{rd} markers is taken as X axis and the perpendicular line from the 1^{st} marker to the X-axis is taken as Y-axis (Fig 3.12).

The normalization process then reduces to finding the rotation (R) and translation (T) matrices that satisfy equations 3.9, 3.10, 3.11 for each frame.

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & T_x \\ R_{21} & R_{22} & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ d_{1-o} \\ 1 \end{bmatrix} \tag{3.9}$$

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & T_x \\ R_{21} & R_{22} & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{2-o} \\ 0 \\ 1 \end{bmatrix} \tag{3.10}$$

$$\begin{bmatrix} x_3 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & T_x \\ R_{21} & R_{22} & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{3-o} \\ 0 \\ 1 \end{bmatrix} \tag{3.11}$$

Where x_n, y_n are the x and y components of the n^{th} point, and d_{n-o} is the distance between Nth point and the origin of the axes.

There are six equations for six unknowns ($R_{11}, R_{12}, R_{21}, R_{22}, T_x$ and T_y). By solving these equations the translation and rotation matrices, T and R are found. Then the 12 points are back projected using these matrices. Then all x components are multiplied by $(9.83/d_{2-3})$, as the actual distance between the 2^{nd} and the 3^{rd} markers is 9.83 cm and all y components are multiplied by $(3.2/d_{1-o})$, as the actual perpendicular distance between the 1^{st} marker and the line between the 2^{nd} and the 3^{rd} markers is 3.2 cm. It is worth to be noted that the selection of the actual distances for normalization is not important for practical issues, these values are selected in order to be able to compare the position data to actual physical values.

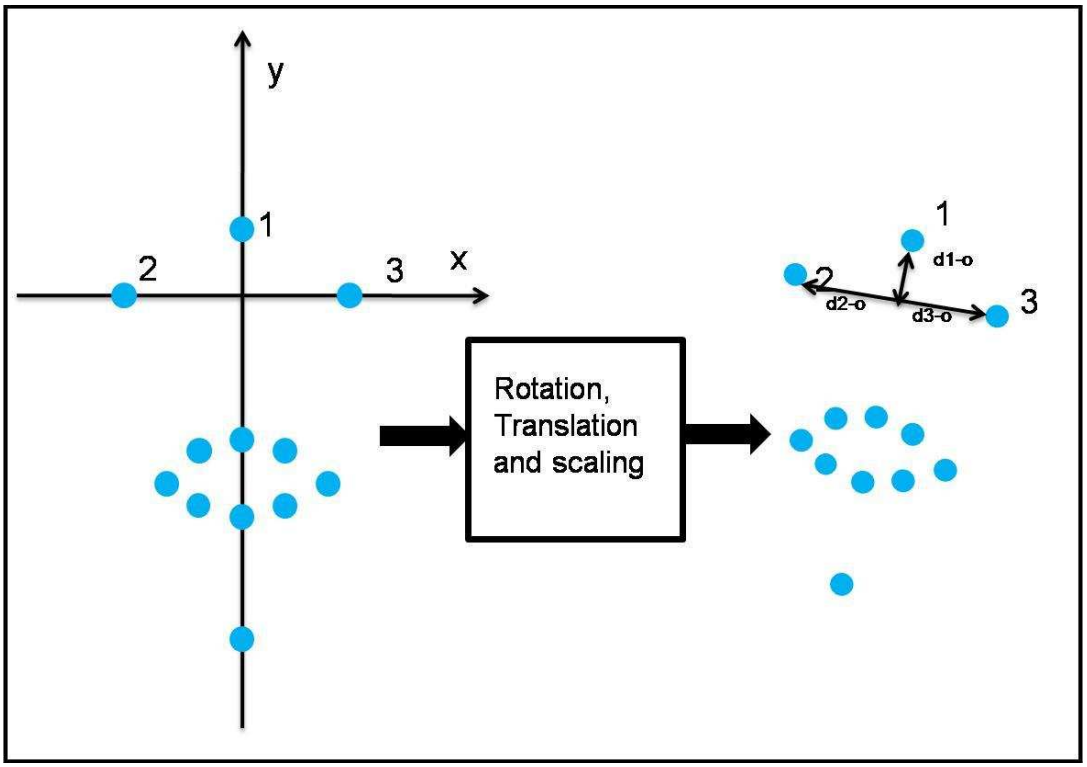

Figure 3.12: Normalization of the positions of the markers

3.3 Extraction of Visual Features

The visual features used in literature can be separated into two major classes, namely; shape based visual features and appearance based visual features, where the former includes the information extracted about the shape and contour of the lips, the latter extracts the useful information from the pixel values of the face or the region of interest.

3.3.1 Shape Based features

Shape based feature extraction assumes that most speechreading information is contained in the shape (contours) of the speaker's lips, or more generally, in the face contours (e.g., jaw and cheek shape, in addition to the lips). Two types of features fall within this category: Geometric type features, and shape model based features. In both cases, an algorithm that extracts the inner and/or outer

lip contours, or in general, the face shape, is required. The detection of the markers located on the speaker's face allows easier detection of the lip contours, and extraction of the shape based features.

3.3.1.1 Lip geometric Features

Given the lip contour, a number of high level features, meaningful to humans, can be extracted, such as the contour height, width, perimeter, as well as the area contained within the contour. As demonstrated in Figure 3.13, such features do contain significant speech information.

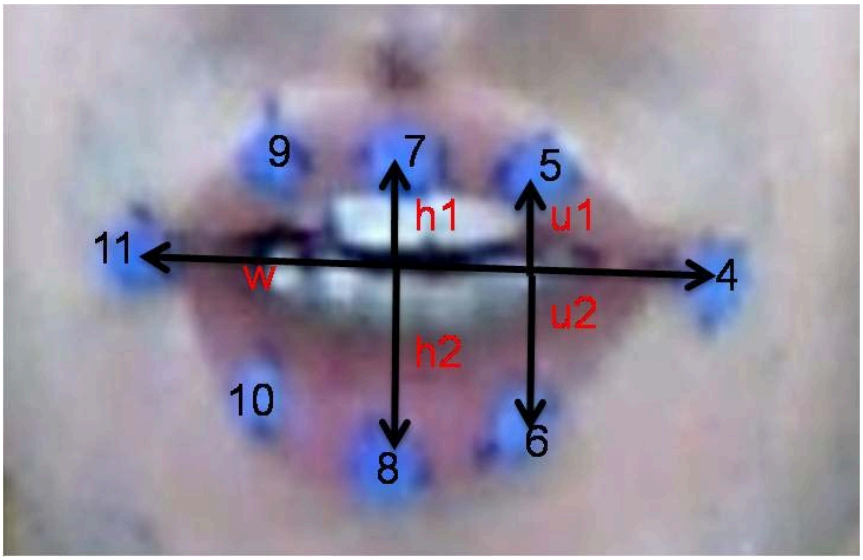

Figure 3.13: Lip geometric features

In this work, the lip geometric features are found using the normalized tracking data (normalization process was discussed in Section 3.2.5), that makes these features invariant to affine image transformations (i.e., translation, rotation, scaling.). Width and height of the lips and the perpendicular distances of the markers to the line between two markers on two sides of the lips are also found (Figure 3.13). The distance between 1^{st} and the 12^{th} marker indicates the movement of the chin. The chin movements are accepted to have high correlation with the syllabic structure of the speech [68]. The area between 8 markers is

also calculated and recorded to the database as the lip area for each image.

3.3.1.2 Lip model features

A number of parametric models have been used for lip or face-shape tracking in the literature, are mentioned in a previous section. The parameters of these models can be used as visual features. In this study the positions of the markers located on the speaker's lips can be used as lip model features.

Another popular lip model is the active shape model (ASM). ASMs are flexible statistical models that represent an object by a set of labeled points. Such object can be the inner and/or outer lip, or the union of various face shape contours. To derive an ASM, a number of K contour points are first labeled on available training set images, and their coordinates are placed on 2K-dimensional "shape" vectors.

In our study, because of the presence of the positions of the markers on the lips, the extraction of lip model features is not needed. The positions of the markers can replace the lip model parameters, or model parameters can be easily obtained by fitting polynomials or splines to the position data.

3.3.2 Appearance Based features

In this approach to visual feature extraction the image typically containing the speaker's mouth is considered as informative for lipreading. The region of interest (ROI) may have a rectangular shape and it may also embrace a larger portion around the mouth including jaw and cheeks. ROI's from consecutive images can be bundled to form a three dimensional (dynamic) information content. A feature vector is obtained by concatenating the ROI pixel grayscale, or color values. This vector is expected to contain most visual speech information. The dimensions of these vectors are too large to allow successful statistical modeling of speech classes, or to be used in classifiers. Therefore, appropriate transformations of the ROI pixel values are used as visual features. The most popular appearance based feature representations achieve such reduction by us-

ing traditional image transforms. The image compression transforms are used for feature reduction with the hope that they preserve information most relevant to speechreading.

In this work tracked markers are used to extract the ROI, the mean of the markers located on the lips is used to find the center of the ROI. 128x128 pixels frame around this center is extracted as ROI. Then RGB data is converted to grayscale, and discrete cosine transform of the image is calculated. 128x128 DCT matrix is converted to a vector of length 16384, this vector can be used as a feature vector, but as described above the size of this vector is too long to be used in speech processing applications. In order to reduce the size of the feature vector, the method of principal component analysis (PCA), borrowed from eigenfaces approach to face recognition, is used [69].

Principal Component Analysis (PCA) (also named as discrete Karhunen-Loève transform (DKLT)) involves a mathematical procedure that is used to represent a function/signal/random variable as a linear combination of orthogonal basis functions. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The DKLT is the most efficient representation of a random process if the expansion is truncated to use **m** number components, where **m** is less than the length of the random process.

The method is based on representing the feature vector $\mathbf{x}$, as a combination of the eigenvectors of its covariance matrix. Where $\mathbf{u}_i$ are the eigenvectors of the covariance matrix of $\mathbf{x}$, with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$

$$\mathbf{x} = \sum_{i=0}^{P} x_i \mathbf{u}_i \tag{3.12}$$

The minimum mean square approximation of the feature vector $\mathbf{x}$, with m coefficients can be achieved by

$$\hat{\mathbf{x}} = \sum_{i=0}^{m} x_i \mathbf{u}_i \tag{3.13}$$

To implement PCA, eigenvectors of the covariance matrix of $\mathbf{x}$ should be calculated. Before that, the mean of the data should be set to zero. The mean vector of all the data is calculated and subtracted. Then the covariance matrix $\mathbf{C}$ can be estimated by

$$\mathbf{C} = E\left[\mathbf{x}\mathbf{x}^T\right] \approx \frac{1}{N}\sum_{i=0}^{P}\mathbf{x}_k\mathbf{x}_k^T \tag{3.14}$$

Where, $\mathbf{x}_k$ are N training vectors from the database. If we pack these vectors into a matrix $\mathbf{X}$.

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_N\right]^T \tag{3.15}$$

Then C can be written as

$$\mathbf{C} \approx \frac{1}{N}\mathbf{X}\mathbf{X}^T \tag{3.16}$$

The dimensions of covariance matrix C are 16384 by 16384. It is very difficult to find the eigenvectors of such a large matrix. The singular value decomposition of X becomes useful at this point [70].

$\mathbf{X}$ has an SVD

$$\mathbf{X} = \sum_{k=1}^{r}\sqrt{\lambda_k}\mathbf{u}_k\mathbf{v}_k^T \tag{3.17}$$

Where $\sqrt{\lambda_k}$, $\mathbf{u}_k$, and $\mathbf{v}_k$ are singular values and left and right singular vectors of $\mathbf{X}$, and r is the rank of $\mathbf{X}$, and r $\leq$ N. Then

$$\mathbf{X}^T\mathbf{X} = \sum_{k=1}^{r}\sum_{k=1}^{r}\sqrt{\lambda_k}\sqrt{\lambda_l}\mathbf{v}_l\mathbf{u}_l^T\mathbf{u}_k\mathbf{v}_k^T \tag{3.18}$$

$$\mathbf{X}^T\mathbf{X} = \sum_{k=1}^{r}\lambda_k\mathbf{v}_k\mathbf{v}_k^T \tag{3.19}$$

Since

$$\mathbf{u}_l^T \mathbf{u}_k = 1, if k = l \tag{3.20}$$

$$= 0, if k \neq l \tag{3.21}$$

and similarly

$$\mathbf{X}\mathbf{X}^T = \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{u}_k^T \tag{3.22}$$

Thus, λ_k are nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$, and $\mathbf{u}_k$ (P×1 vector)and $\mathbf{v}_k$ (NX1 vector) are eigenvectors of , $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ respectively. By multiplying both sides of equation 3.17 with $\mathbf{v}_k$ from right,

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{X}\mathbf{v}_k \tag{3.23}$$

So by finding the eigenvectors of $\mathbf{X}^T\mathbf{X}$ ($\mathbf{v}_k$), which is a much easier task (N<<P, Typical value for N changes from 100 to 1000 where P=16384 in this case.), the eigenvectors of $\mathbf{X}\mathbf{X}^T$ can be found. After finding the eigenvectors of $\mathbf{C}$, $\mathbf{L}$ eigenvectors with the highest eigenvalues are selected to represent $\mathbf{x}$. $\hat{x}_i$ in eqn 3.13 can be found as the inner product of $\mathbf{x}$ and $\mathbf{u}_k$ as the eigenvectors are orthogonal. This reduces the size of the feature vector from 16384 to $\mathbf{L}$ (fig 3.14). Some examples of the original images and image found by back transformation of the reduced length feature vectors are shown in Figure 3.15. Examining the images from reduced sized feature vectors, it can be stated that nearly all information that is meaningful to human perception seems to be preserved by using 10-20 parameters, i.e., the features that are visually meaningful to humans, such as lip shape, lip opening, lip width, teeth visibility are detectable. For each feature vector size, average absolute boundary errors per pixel are presented in Table 3.4.

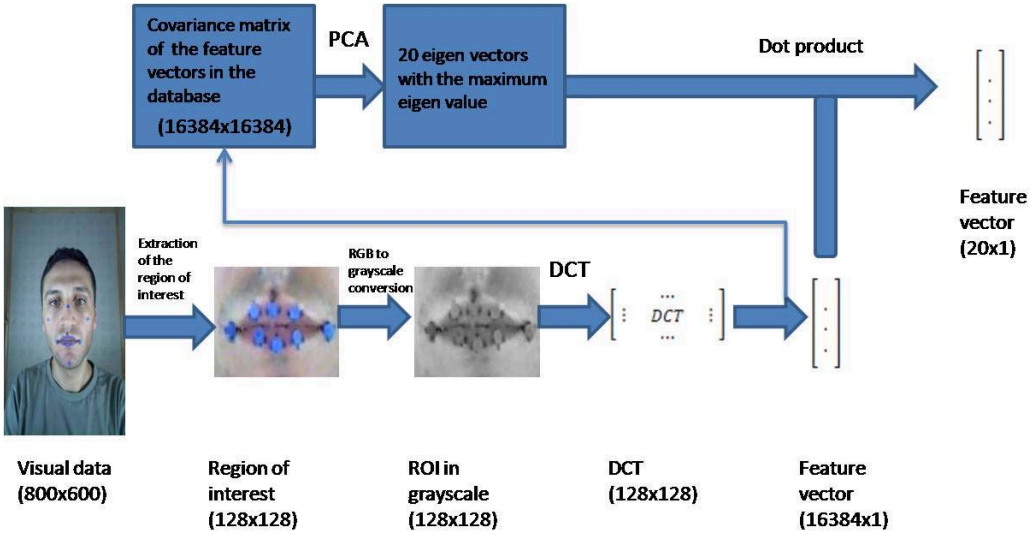

Figure 3.14: Extraction of appearance based features

Table 3.4: Average absolute error per pixel using different length appearance based feature vectors

Phoneme couple	20 coefficients	10 coefficients	5 coefficients
/r/-/i/	7.57	8.18	8.54
/v/-/a/	9.41	11.23	13.01
/i/-/l/	9.22	11.96	12.57
silence	6.65	7.14	8.97
/t/-/ü/	10.99	11.90	14.53

The inverse DCT of the eigenvectors with the largest five eigenvalues are presented in Figure 3.16. It can be observed that the eigenvector with the highest eigenvalue seems like the average of the visual features and the remaining eigenvectors are the deviations from this average.

Teeth visibility: The visibility of the teeth is also an important cue for humans in recognizing the uttered speech. Teeth visibility is a very distinctive property for the recognition of labio-dental, dental and alveolar phonemes. A measure of visible teeth area can be also useful for our purposes. The teeth area is found by locating the brightest pixels (white area) in the ROI by thresholding the image (Fig 3.17). The area of the white pixels are normalized by using the track

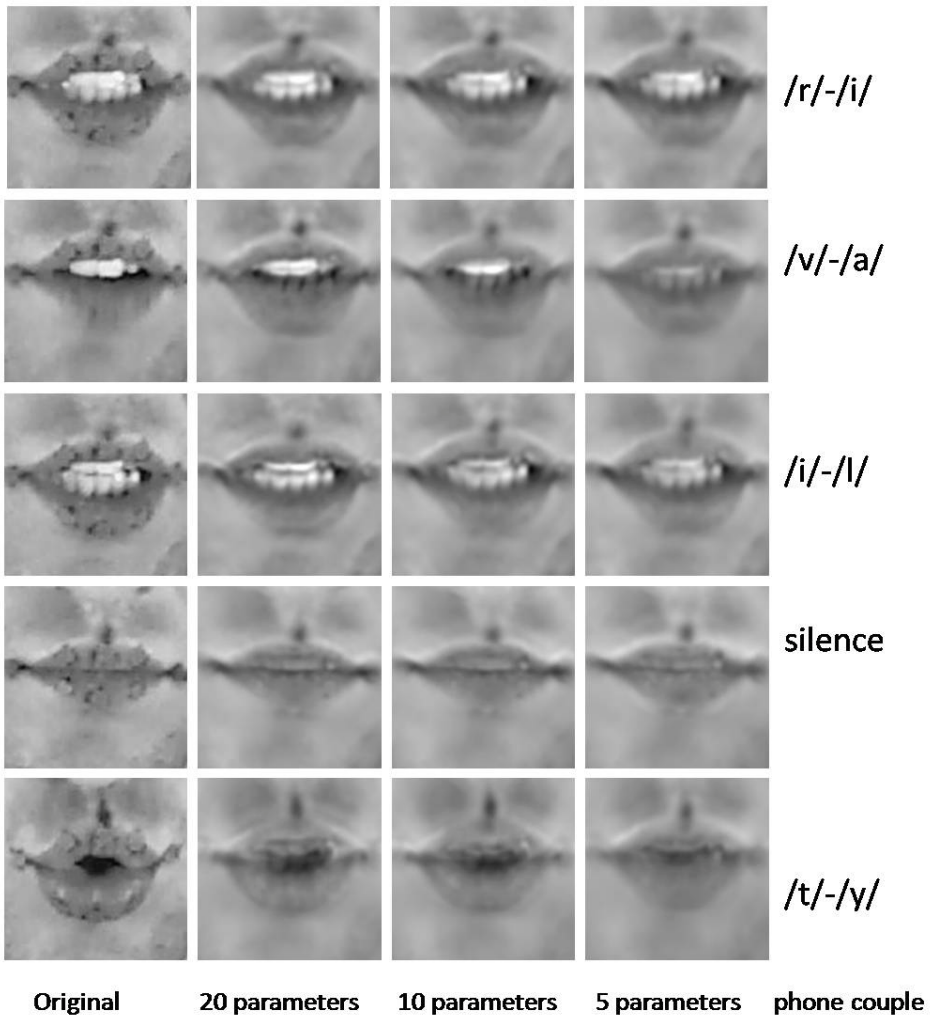

Figure 3.15: Some examples of the images represented by reduced sized feature vectors of length 20, 10 and 5

markers, the number of teeth pixels is divided by the area between the static markers, and then saved to the database, for each utterance.

3.3.3 Visual Parameters Stored in the Database

The database contains different folders for each utterance. Each folder contains a *.wav* file containing the audio data, a *.lab* file in HTK format, containing the time labels of manual segmentation, visual data composed of *.jpg* files, and a *.mat* file containing the visual features discussed in Chapter 3.3.

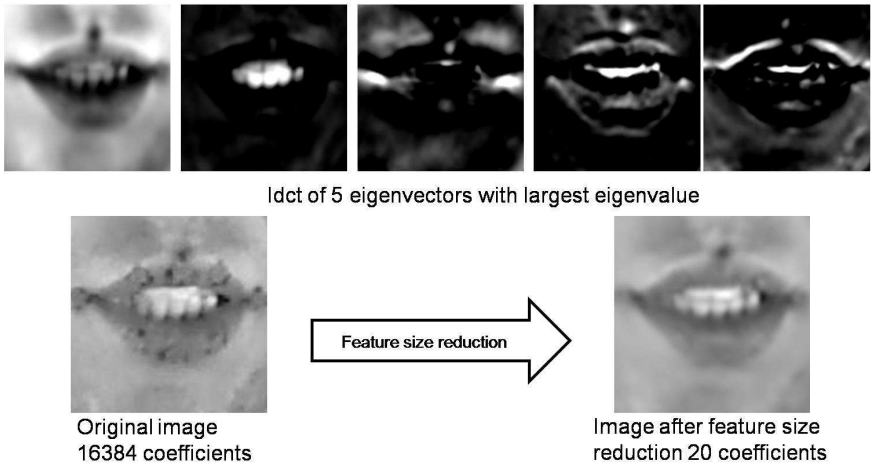

Figure 3.16: The inverse DCT of five eigenvectors with the largest eigen values

The visual data recorded at 30 fps are interpolated to 200 fps in order to be used with the audio data. The visual features saved to the database are;

1. **tracked_markers:** vector sequence of the positions of the markers (length 24)

2. **Normalized tracked_markers:** vector sequence of the normalized positions of the markers (length 24)

3. **Wh:** vector sequence of width and height of the lips (length 2).

4. **Wh1h2u1u2:** vector sequence of width and 4 type of height parameters (Figure 3.13) (length 5).

5. **Pcavect:** vector sequence of PCA visual features (length20).

6. **Teethscore:** vector sequence containing teeth score.

7. **Liparea:** vector sequence containing normalized lip area.

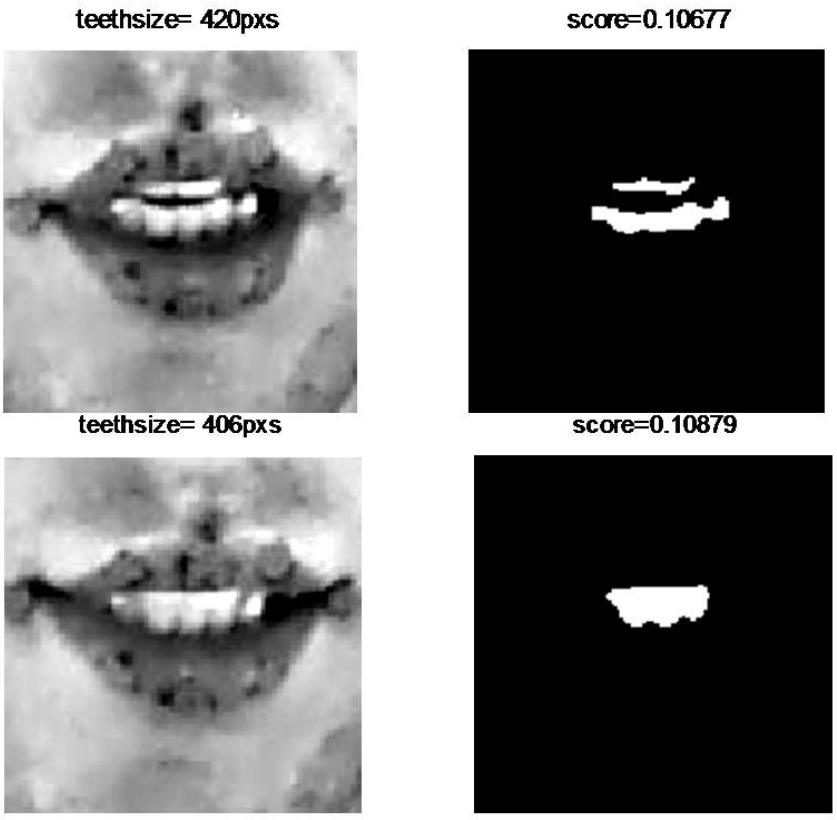

Figure 3.17: Finding the teeth area

3.4 Audiovisual Automatic Speech Segmentation

The studies in Chapter 2 had shown the potential of the improvement that can be achieved by the integration of the articulator positions to an automatic speech segmentation system. The experiments had proven that even with limited information about the articulator positions (only the positions of upper lip and lower lip had been used) the average absolute boundary error can be reduced by 18%. A Turkish audiovisual speech database is prepared in order to extend the idea of automatic audiovisual speech segmentation. The studies on this subject will be described in following sections.

3.4.1 The Method

The AS system build for audiovisual automatic speech segmentation is similar to the one described in Section 2.3.2 (Figure 3.18). The speech segmentation system was build using HTK speech recognition toolkit [27].

Among 1600 utterances 1000 of utterances are used for training, 500 of utterances are used for test, and 100 of the utterances are used for decision fusion between the outputs of the system with different feature vectors (The decision fusion method will be explained in Section 3.4.5). Phonemes are selected as phonetic units for AS in the experiments.

3 state, left-to-right, continuous Gaussian density HMMs with 3 mixtures, were used for modeling 42 phonemes (/o/, /ox/, /O/, /Ox/, /u/, /ux/, /y/, /a/, /ax/, /e/, /ex/, /I/, /Ix/, /i/, /ix/, /p/ , /b/, /m/, /f/, /v/, /t/, /d/, /s/, /z/, /n/, /tS/, /Z/), and silence and breath.

The system uses monophone acoustic models, as it was shown that monophone HMM models with optimized number of states outperformed diphone and triphone models [14]. The monophone models are trained initially by using the segment boundaries given in the database, and then the model parameters are re-estimated in an iterative manner by using Baum-Welch algorithm until the settlement of segment boundaries. The test is done by finding the automatic seg-

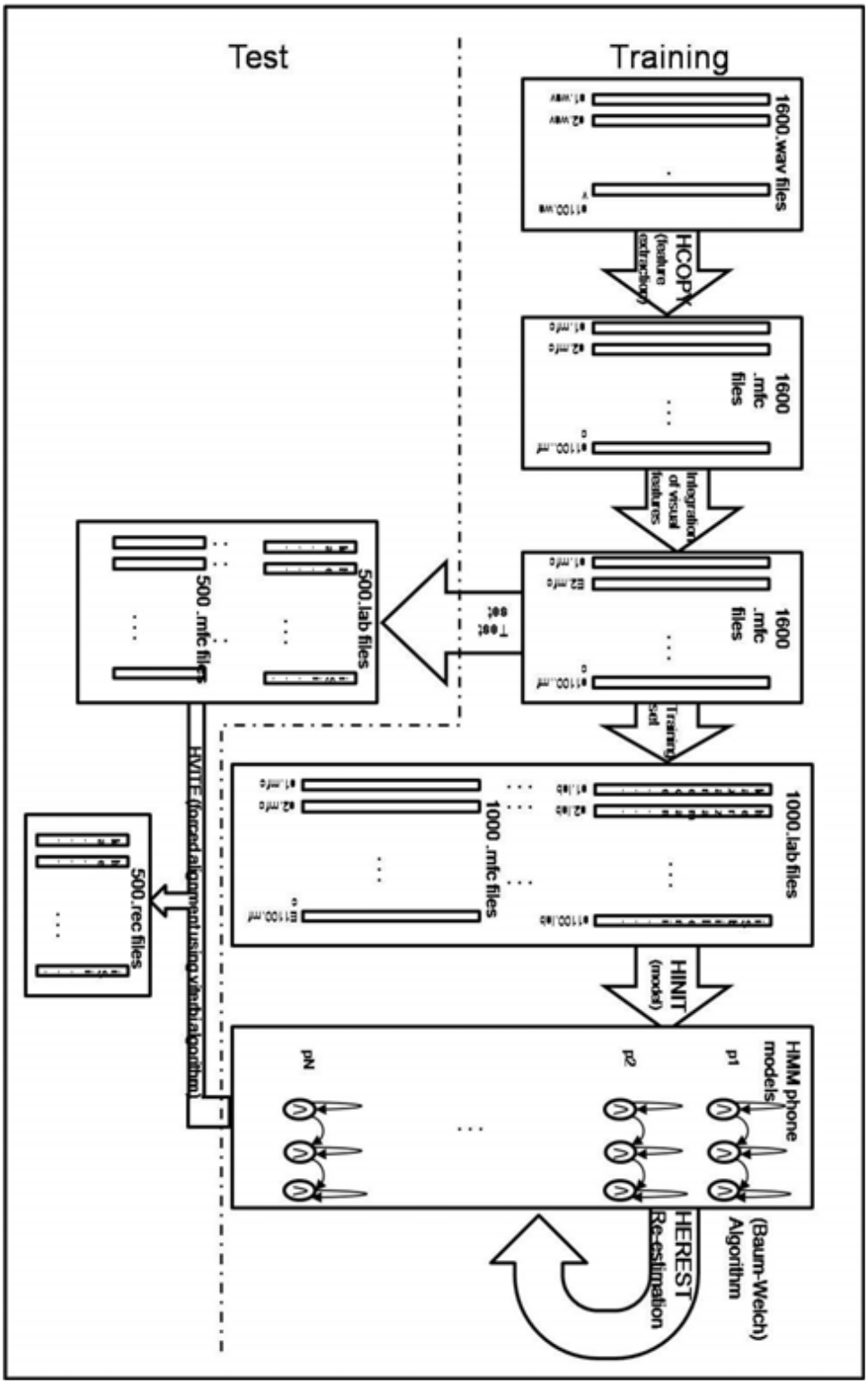

Figure 3.18: The Overview of the Automatic Speech Segmentation System

mentation results using Viterbi algorithm by supplying the phonetic transcription of the test utterances (forced alignment method). The segment boundaries so obtained are compared against the boundaries produced by manual segmentation. Average absolute boundary error (AABE) is used as a performance measure for different systems/feature sets.

The acoustical feature vectors are computed at 100 Hz with an analysis window of length 25 ms. 3 state, left-to-right, continuous Gaussian density HMMs with three mixtures were used for modeling. Context independent HMMs are used as in the previous case. Monophone acoustic models for the 42 phonemes and silence and breath, are trained initially by using the manually segmented boundaries. The training and test are done using HTK as described in Section 2.3.2.

3.4.2 Feature Vectors

3.4.2.1 Acoustical Feature Vectors

Mel-Frequency Cepstral Coefficients (MFCC) that are frequently used in speech processing applications are used as acoustical features in this study. The vector includes 12 Mel-Frequency Cepstral Coefficients, an energy coefficient, and their first and second order derivatives, resulting in an acoustical feature vector of length 39. The feature vectors are computed at 100 Hz with an analysis window of length 25 ms.

3.4.2.2 Visual Feature Vectors

The visual features stored in the database are used in the experiments (Section 3.3.3). The visual features on the database are at a rate of 200 fps. They are downsampled to 100 fps in order to be compatible with the frame rate of the acoustic features.

3.4.2.3 Early Fusion of the Audio and Visual Features

The information from two different modalities; namely, the acoustic and visual modalities have to be fused at some level. The acoustic and visual data may be combined at a variety of levels from raw data level, feature level, state vector level, to decision level, etc. The fusion at raw data level is preferred when the data is available from different sensors measuring the same physical phenomena (sensors are commensurate.). For the noncommensurate case, data are fused at feature/state vector or decision level. In feature level fusion, different features are extracted from the data of different types of sensors, and combined into a single concatenated feature vector that will be used as input to a pattern recognition process [71].

In this study available information are fused at feature level (early fusion), by concatenating the acoustical and visual feature vectors. Multiple combinations of the visual feature vectors are appended to acoustical feature vector (noted as MFCC_0_D_A), resulting features are saved to *.mfc* files for each utterance in the database.

Different bimodal feature vectors obtained by appending the acoustical and visual feature vectors are as follows:

1. **MFCC_0_D_A:** This is the baseline feature vector that contains 12 MFCCs and 1 energy coefficient, their derivatives, and second derivatives. (39 elements)

2. **+h:** The height of the lips is appended to the MFCC_0_D_A as the 40th element.

3. **+h_D:** The height of the lips and its derivative are appended to the MFCC_0_D_A as the 40th and the 41st elements.

4. **+wh:** The width and height of the lips are appended to the MFCC_0_D_A as the 40th and the 41st elements.

5. **+la:** The lip area is appended to the MFCC_0_D_A as the 40th element.

6. **+wh1h2u1u2:** The width of the lips and h1h2u1u2 parameters (Section 3.3) are appended to the MFCC_0_D_A resulting in a 44 element feature vector.

7. **+wh1h2u1u2_D:** The wh1h2u1u2 vector described above and its derivative are appended to the MFCC_0_D_A resulting in a 49 element feature vector.

8. **+h1h2u1u2:** The h1h2u1u2 vector described above is appended to the MFCC_0_D_A resulting in a 43 element feature vector.

9. **+h1h2u1u2_D:** The h1h2u1u2 vector described above and its derivative are appended to the MFCC_0_D_A resulting in a 47 element feature vector.

10. **+h1h2u1u2+la:** The h1h2u1u2 vector described above and lip area are appended to the MFCC_0_D_A resulting in a 44 element feature vector.

11. **+h1h2u1u2+teeth** The h1h2u1u2 vector described above and and visible teeth area are appended to the MFCC_0_D_A resulting in a 44 element feature vector.

12. **+normalized markers:** Normalized positions of the 9 markers on the speakers face are appended to the MFCC_0_D_A resulting in a 57 element feature vector.

13. **+normalized uly-lly:** The vertical positions of the upper lip and the lower lip (normalized) markers (7 and 8) are appended to the MFCC_0_D_A resulting in a 41 element feature vector.

14. **+pca12:** Appearance based feature of length 12 that is explained in Section 3.3.3 is appended to the MFCC_0_D_A resulting in a 51 element feature vector.

Determination of the Size of the PCA Vector to be used: Visual inspection of the different length appearance based feature vectors obtained by using PCA shows that most of the relevant features are preserved for the feature sizes above 10. Obviously, increasing the feature size of the PCA feature vector would

result in a better representation of the ROI, but because of the curse of dimensionality, increasing the feature size would degrade the accuracy of the models developed, as the training data is limited. Over the different sized PCA vectors, the one resulting in the minimum AABE is used in the experiments. The average pixel errors and the AABE for different length PCA vectors are presented in Figure 3.19, it can be observed that the PCA vector having 12 elements results in the minimum AABE although the larger PCA vectors have smaller average pixel errors.

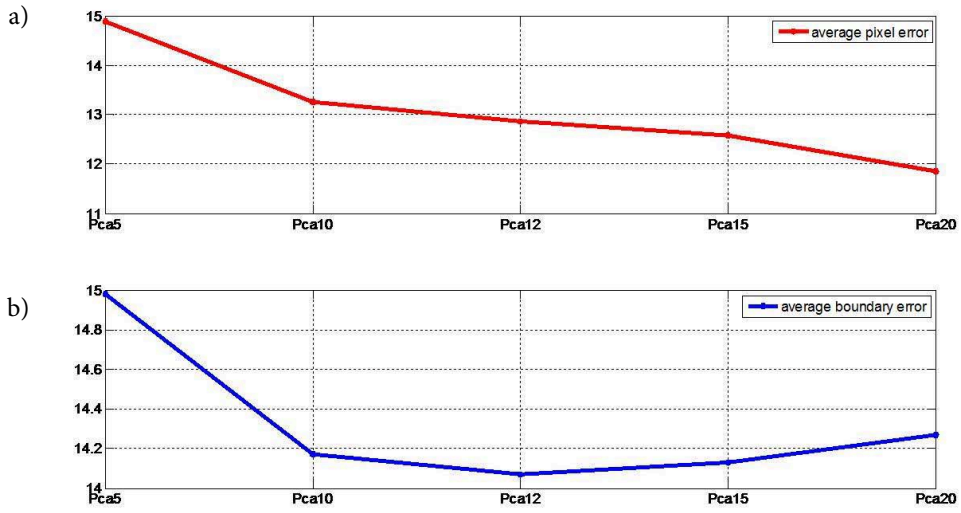

Figure 3.19: a) the average pixel error and b) the average absolute boundary error for different PCA vector sizes.

3.4.3 Audiovisual Automatic Segmentation Results

The system described in Section 3.4.1 is used to test the effect of using different visual feature vectors on AS. The performance of the proposed automatic speech segmentation system is observed for different bimodal feature vectors and these performances are compared to each other in following work. AABE is used as a performance measure of different feature vectors.

Table 3.5: Average Absolute Boundary Errors Variances of Errors and Decrease in the Average Absolute Boundary Errors for Different Feature Vectors

	Feature	AABE (ms)	Variance (X10-4)	AABE with baseline (ms)	Percent Decrease in AABE Relative to Baseline System
1	MFCC_0_D_A	17.77	7.34	17.77	-
2	+h	15.39	5.26	15.33	13.73
3	+h_D	15.5	4.93	15.38	13.45
4	+wh	15.16	5.25	14.94	15.93
5	+liparea	15.34	5.24	15.29	13.96
6	+wh1h2u1u2	14.99	6.03	14.49	18.46
7	+wh1h2u1u2_D	20.72	28.21	16.57	6.75
8	+h1h2u1u2	**14.67**	5.64	14.4	18.96
9	+h1h2u1u2_D	20.20	24.13	16.8	5.46
10	+ h1h2u1u2 +liparea	15.29	5.61	14.54	18.18
11	+ h1h2u1u2 +teeth	15.04	7.02	14.75	16.99
12	+normalized markers	19.21	5.41	17.05	4.05
13	+normalizedulylly	15.36	6.15	15.6	12.21
14	+Pca12	16.18	8.39	**14.07**	20.82

The performance of the AS system using different feature vectors are presented in Table 3.5. The value at the third column is the average absolute boundary error (AABE) obtained by using the corresponding feature vector and the value at the fourth column is the variance of that error. The fifth column shows the AABE achieved when the boundaries obtained by the corresponding feature vector are used selectively with the boundaries obtained by using the baseline feature vector. The selection is done by using visual phoneme classes, each boundary is selected from the outputs of the systems using two different feature vectors, and the output of the system that gives the best overall performance by that boundary class is selected.

The 8^{th} feature vector, +h1h2u1u2, yields the smallest AABE. The AS system using the 14^{th} feature vector which includes appearance based features, leads to the smallest AABE when used in collaboration with the baseline system, resulting 20.8% decrease in the AABE. This means that the 14^{th} feature vector causes high improvements in some boundary classes, and high degradation in

the others. This is predictable as the pca visual feature vector is the longest vector and thus appending it to the acoustical features affects the performance of that feature most.

The 13^{th} vector includes the vertical positions of upper and lower lip (the 7^{th} and 8^{th} markers). In our previous work [72] (Chapter 2), using the vertical positions of upper lip and lower lip had decreased the AABE by 8.1%. Although the acquisition of the positions using a camera is expected to be less precise with respect to EMA, it is seen that using the positions of the same articulators 12.21% decrease is achieved in AABE. This may partially be accounted for the increase in the size of the database. The database in this work is also in a different language and recorded by a different speaker. However, it is difficult to speculate clearly on the possible effects of these differences.

3.4.4 Visual Phoneme Classes

The segmentation results using different feature vectors are investigated in a phoneme class based manner. Firstly, AABE of the baseline feature vector belonging to each phoneme class to phoneme class boundary is calculated and listed in Table 3.6.

After that, for each class, the feature vector yielding the minimum average absolute error is found and compared to the baseline system. The phonemes are clustered into 5 classes regarding their visual properties; the vowels are divided into two classes as: rounded vowels and unrounded vowels. The consonants are divided into 3 classes as: (1) bilabial and labio-dental, (2) dental and alveolar, (3) palatal, velar and glottal consonants. The phonemes are grouped according to visual properties as follows;

- **Rounded vowels:** /o/, /ox/, /O/, /Ox/, /u/, /ux/, /y/ ;
- **Unrounded vowels:** /a/, /ax/, /e/, /ex/, /I/, /Ix/, /i/, /ix/;
- **Bilabial and labio-dental consonants:** /p/ , /b/, /m/, /f/, /v/;
- **Dental and alveolar consonants:** /t/, /d/, /s/, /z/, /n/, /tS/, /Z/;

- **Palatal glottal and velar consonants:** /dZ/, /g/, /gj/, /h/, /k/, /c/, /l/, /r/, /S/ /j/;

Table 3.6: AABE between phoneme classes. In each cell, the first line is the AABE in milliseconds, and the second line presents the number of occurrences of the boundary class in the database. For each cell, row id is the left phonetic class type and the column id is the right phonetic class type.

	Silence	Rounded vowels	Unrounded vowels	Bilabial and labio-dental consonants	Dental and alveolar consonants	Palatal glottal and velar consonants
Silence	37.20 4	10.42 509	9.70 279	17.17 675	21.42 300	31.65 1019
Rounded vowels	63.88 1106	24.94 113	27.12 132	19.30 759	25.58 2482	24.02 4279
Unrounded vowels	41.56 188	29.46 501	18.04 758	16.51 3683	24.78 9660	23.80 13695
Bilabial and labio-dental	26.29 358	9.51 1173	9.70 5012	18.35 201	13.90 348	14.75 561
Dental and alveolar consonants	16.72 300	10.18 2580	9.95 9531	13.52 929	22.59 1978	14.87 2088
Palatal glottal and velar consonants	30.47 823	12.18 3352	11.24 13123	19.38 1319	20.30 2497	21.19 2850

The experimental results regarding the visual phoneme classes are presented in Table 3.7 and Table 3.8. The results are listed in a phoneme class based manner

Table 3.7: The best feature vector for each class to class boundary. AABE between phoneme classes. In each cell, the first line is the number of the feature vector yielding minimum average absolute error, and the second line is the percent decrease with respect to baseline for each visual class to class boundary. The third cell represents the number of occurrences of the boundary class in the database. For each cell, row id is the left phonetic class type and the column id is the right phonetic class type.

	Silence	Rounded vowels	Unrounded vowels	Bilabial and labio-dental consonants	Dental and alveolar consonants	Palatal glottal and velar consonants
Silence	14 53.75% 4	12 -15.02% 509	12 -5.94% 279	2 3.98% 675	3 24.11% 300	8 14.44% 1019
Rounded vowels	14 32.15% 1106	14 14.14% 113	14 9.08% 132	14 45.78% 759	14 23.82% 2482	14 35.75% 4279
Unrounded vowels	4 52.97% 188	14 26.15% 501	12 5.60% 758	11 23.29% 3683	10 26.22% 9660	6 25.63% 13695
Bilabial and labio-dental	4 37.71% 358	5 15.54% 1173	2 13.38% 5012	12 12.23% 201	14 25.44% 348	4 12.61% 561
Dental and alveolar consonants	8 9.97% 300	14 25.58% 2580	11 7.21% 9531	12 18.07% 929	8 3.91% 1978	14 22.38 % 2088
Palatal glottal and velar consonants	3 23.10% 823	11 12.16% 3352	13 3.92% 13123	4 20.71% 1319	10 26.16% 2497	14 24.00 % 2850

in Table 3.7, each row shows the first phoneme class and each column shows the second phoneme class at a boundary. The average absolute boundary errors of class to class boundaries are calculated and the percent deviations of the average absolute errors from those of the baseline system are given in the table (negative values indicate that the best audiovisual feture vectors performs worse than the baseline feature vector.). For example, the feature vector resulting in the minimum average absolute error and the change in the average absolute error for (rounded vowels)-(silence) boundary is shown in the cell of the (rounded vowels) row and (silence) column; 14^{th} feature vector resulted the minimum AABE, approximately 32.2% decrease in the average absolute error is achieved. The third value listed in that cell is the number of occurrences of that boundary

Table 3.8: AABE for different boundary classes, sorted in descending order of percentage decrease achieved.

Feature Type	Best 3 Features	% decrease	Number of samples
(Silence)-(Silence)	14, 2, 3	53.75	4
(Unrounded vowels)-(Silence)	4, 12, 6	52.97	188
(Rounded vowels)-(Bilabial and labio-dental)	14, 6, 10	45.78	759
(Bilabial and labio-dental)-(Silence)	4, 11, 13	37.71	358
(Rounded vowels)-(Palatal glottal and velar)	14, 8, 10	35.75	4279
(Rounded vowels)-(Silence)	14, 4, 13	32.15	1106
(Unrounded vowels)-(Dental and alveolar)	10, 8, 11	26.22	9660
(Palatal glottal and velar)-(Dental and alveolar)	10, 6, 8	26.16	2497
(Unrounded vowels)-(Rounded vowels)	14, 2, 5	26.15	501
(Unrounded vowels)-(Palatal glottal and velar)	6, 11, 10	25.63	13695
(Dental and alveolar)-(Rounded vowels)	14, 11, 10	25.58	2580
(Bilabial and labio-dental)-(Dental and alveolar)	14, 7, 6	25.44	348
(Silence)-(Dental and alveolar)	3, 7, 9	24.11	300
(Palatal glottal and velar)-(Palatal glottal and velar)	14, 8, 10	24.00	2850
(Rounded vowels)-(Dental and alveolar)	14, 6, 10	23.82	2482
(Unrounded vowels)-(Bilabial and labio-dental)	11, 6, 8	23.29	3683
(Palatal glottal and velar)-(Silence)	3, 2, 13	23.10	823
(Dental and alveolar) -(Palatal glottal and velar)	14, 11, 10	22.38	2088
(Palatal glottal and velar)-(Bilabial and labio-dental)	4, 2, 5	20.71	1319
(Dental and alveolar) -(Bilabial and labio-dental)	12, 13, 4	18.07	929
(Bilabial and labio-dental)-(Rounded vowels)	5, 2, 3	15.54	1173
(Silence)-(Palatal glottal and velar)	8, 6, 2	14.44	1019
(Rounded vowels)-(Rounded vowels)	14, 6, 11	14.14	113
(Bilabial and labio-dental)-(Unrounded vowels)	2, 5, 13	13.38	5012
(Bilabial and labio-dental)-(Palatal glottal and velar)	4, 6, 8	12.61	561
(Bilabial and labio-dental)-(Bilabial and labio-dental)	12, 13, 5	12.23	201
(Palatal glottal and velar)-(Rounded vowels)	11, 8, 6	12.16	3352
(Dental and alveolar) -(Silence)	8, 10, 5	9.97	300
(Rounded vowels)-(Unrounded vowels)	14, 13, 8	9.08	132
(Dental and alveolar) -(Unrounded vowels)	11, 1, 13	7.21	9531
(Unrounded vowels)-(Unrounded vowels)	12, 13, 11	5.60	758
(Silence)-(Bilabial and labio-dental)	2, 5, 13	3.98	675
(Palatal glottal and velar)-(Unrounded vowels)	13, 12, 5	3.92	13123
(Dental and alveolar) -(Dental and alveolar)	8, 9 ,16	3.91	1978
(Silence)-(Rounded vowels)	1, 12, 13	0.00	509
(Silence)-(Unrounded vowels)	1, 12, 8	0.00	279

class in the database. The results are listed by sorting the boundaries according to the decrease achieved in AABE, by using the best feature vector for that class to class boundary, from greatest to lowest in Table 3.8. The three feature types yielding the lowest AABE also given at the second row of the table in descending AABE order.

AABE is decreased at all boundary types except (silence)-(rounded vowels) and (silence)-(unrounded vowels) boundaries. This can be explained by the low correlation between lip movement and the boundary location in these types of

boundary classes, i.e., the lips are opened much more before the production of the vowel and there is no change in the positions of the articulators after the silence. Note that, this is not the case in (silence)-(consonant) type boundaries, there is a strong correlation between lip movement and the boundary location in these boundaries, i.e., production of the consonants mostly starts with the movement of the lips.

At most of the boundary types ending with silence more than 30% of decrease in the AABE is observed. Also, within all the boundary types, the greatest reduction in the AABE is observed in (unrounded vowels)-(silence) boundaries (52.97%). These results seem reasonable, as the lip movement vanishes and all face remains still almost simultaneously with the start of the silence. Also, the third highest improvement is achieved at (bilabial and labio-dental consonants)-(silence) boundaries may be explained by a similar reasoning.

Second highest reduction in AABE is achieved at boundaries between (rounded vowels)-(bilabial and labio-dental consonants). Using PCA visual vector in concatenation with MFCC_0_D_A vector caused a decrease of 45.78% in AABE in these boundaries. Similar decrease had also been observed at this boundary type in our previous study [72]. This seems relevant as the lip movement is most clearly observed in these types of boundaries.

Furthermore, it is worth noting that, two of the three cases where the 11^{th} feature vector (including the visible teeth area) performs the best are the (unrounded vowels)-(bilabial and labio-dental consonants) and (bilabial and labio-dental consonants)- (unrounded vowels) boundaries as the teeth are mostly visible at (vowel)-(bilabial consonants) and (bilabial consonants)-(vowel) boundaries. The improvement by 11^{th} feature vector can be observed at boundaries including unrounded vowels but not in rounded vowels. At most of the boundaries including a rounded vowel 14^{th} feature vector performs best, when the rounded vowels are articulated, beside the rounding, the lips are retracted to forward too, it seems that this combined movement can be more successfully detected by the 14^{th} feature.

Investigating the results on Table 3.8 it can be observed that the boundaries with

the least improvement are the boundaries ending with an (unrounded vowel), five of the eight cases where the improvements are minimum are belonging to this case.

3.4.5 Fusion of the Automatic Segmentation Results Obtained by Using Different Audiovisual Feature Vectors

The decision fusion can be defined as the selection of a decision from a set of available decisions from different systems. The experimental assessment of different pattern recognition systems may put forward one of the system as the best system, but in most of the problems, different systems may provide complementary information in 'some parts of the problem', and combining these systems appropriately would lead better performance than the system performing best when used alone. This led the research on decision fusion techniques. The idea is not to rely on a single decision making scheme. Instead, all available decisions from different classifiers, or a subset, are used for finding a combined decision [76].

The combination of the decisions of different classifiers would be particularly useful when these classifiers are really 'different'. This can be achieved by the classifiers using different methods, different feature sets, or different training sets.

The decision fusion techniques had also been used in speech segmentation systems to combine the boundary locations estimated by different AS systems. Suppose that there are N_{AS} boundary decisions from different AS systems. The boundary locations found by these systems T_i=$\{\{t_i(j,k)\}_{k=1}^{N_j}\}_{j=1}^{N_{utt}}$. Where,

N_{utt}=Number of all utterances

N_j=Number of boundaries at the j^{th} utterance

$t_i(j,k)$= k^{th} boundary mark at the j^{th} utterance found by the i^{th} AS system.

The goal of decision fusion is to find final set of boundaries T_F=$\{\{t_F(j,k)\}_{k=1}^{N_j}\}_{j=1}^{N_{utt}}$, by using the set of boundaries $T_1, T_2, ..., T_{N_{AS}}$.

Several ways to combine the boundaries found by different AS systems had been proposed. Using the average of the boundary locations estimated by different AS systems is proposed as a combination method in [77] (Eqn 3.24).

$$t_F^{avg}(j,k) = \frac{1}{N_{AS}} \sum_{i=1}^{N_{AS}} t_i(j,k) \tag{3.24}$$

The final boundaries can also be found as a weighted sum of the boundaries found by different AS systems (Eqn 3.25).

$$t_F^{w}(j,k) = \sum_{i=1}^{N_{AS}} w_i t_i(j,k) \tag{3.25}$$

where,

$$\sum_{i=1}^{N_{AS}} w_i = 1 \tag{3.26}$$

then the estimation of $\{w_i\}_{i=1}^{N_{AS}}$ is the problem. Note that if w_i's are set to be $\frac{1}{N_{AS}}$ then $t_F^w(j,k) = t_F^{avg}(j,k)$. Different methods for calculating the weights exist. The training set can be used for finding the weight values minimizing the final error. A common approach to assign values to the weights is setting the weight value corresponding to a system, inversely proportional to the variance of the boundary errors achieved by that system.

$$w_i = \frac{1}{V_T \sigma_i^2} \tag{3.27}$$

Where,

σ_i^2 is the variance of boundary errors found by the i^{th} AS system and V_T is the sum of the inverses of the variances, used to make the sum of the weights equal to 1.

$$V_T = \sum_{i=1}^{N_{AS}} \frac{1}{\sigma_i^2} \tag{3.28}$$

The combination of the decisions of different AS systems makes more sense when used in a context dependent manner. In that case the weights are calculated for each phoneme couple or phoneme class couple. Then the final boundary, $t_F^c(j,k)$, will be calculated as;

$$t_F^c(j,k) = \sum_{i=1}^{N_{AS}} w_i(p(j,k), p(j,k+1)) t_i(j,k) \tag{3.29}$$

where, $p(j,k)$ is the label of k^{th} phoneme at the j^{th} utterance. $w_i(p(j,k), p(j,k+1))$ is the weight for i^{th} AS system for the boundary between the phonemes or phonemecouples denoted by $p(j,k)$ and $p(j,k+1)$, respectively. The weights should sum up to 1 for each class to class boundary.

$$\sum_{i=1}^{N_{AS}} w_i(C_j, C_k) = 1, ... \forall j,k \tag{3.30}$$

where, C_i is the i^{th} phoneme class.

The weights for each phoneme class couple can be calculated by using the variances of errors corresponding to the boundary classes found in Section 3.4.4 as explained before. The 100 utterances that were not included to training and test sets are used for this purpose. By using this method AABE of the boundaries is found to be 14.87 ms (Table 3.9).

This strategy can be modified to use **N** best boundary types for each boundary type again by weighting them inversely proportional to their variances. The fusion results for **N**= 3 and **N**= 5 are presented in Table 3.9.

Another strategy of fusing the decisions is selecting the feature vector yielding to minimum AABE for each boundary class, by assessing the AABE found by using the 100 utterances reserved for this pupose. This boundary class wise selection based approach is also called hard decision fusion and can also be achieved by setting the weight of the system resulting the minimum AABE to 1 and all the others to 0. Using this approach led to the minimum error among all fusion techniques, resulting an AABE of 13.49 ms meaning a 24.09% decrease with

respect to the original system (Table 3.9).

Also, to be able to see the best performance that can be achieved by the system, the best feature vectors for each boundary class is found by using the whole test data. Using this information in the selection of the audiovisual feature vectors for each boundary class resulted in a 25.27% decrease in the AABE (13.28 ms AABE).

Table 3.9: AABE for different audiovisual feature vectors and for selective features

Feature	AABE (ms)	% decrease
MFCC_0_D_A	17.77	
+h1h2u1u2	14.41	18.91%
+Pca12	14.07	20.82%
Weighted Decision Fusion	14.87	16.32%
Weighted Decision Fusion 5 Best	13.91	21.72%
Weighted Decision Fusion 3 Best	13.62	23.35%
Hard Fusion	13.49	24.09%
Hard Fusion Using All Test Set	13.28	25.27%
Hard Fusion Using All Test Set and 10 Phoneme Classes	12.95	27.12%

Table 3.10: Accuracy of different segmentation systems for different thresholds

Segmentation System	AABE	<5ms	<10ms	<20ms	<50ms
Baseline	17.77 ms	18.16%	38.41%	70.86%	96.12%
+wh	15.16 ms	21.28%	44.03%	76.92%	97.65%
Hard Fusion	13.28 ms	28.45%	52.33%	81.38%	97.92%
Manual discrepancies	9.09 ms	50.67%	73.91%	86.52%	98.07%

The phonemes in the Turkish audiovisual speech database are clustered into 6 phoneme classes for the results obtained to be comparable with the results obtained in the previous section. Also it is difficult to visualize phoneme class to phoneme class results when the class size increases. But, on the other hand availability of the large database enables the clustering of the phones into more classes. To be able to see the effect of this possibility, the experiments are repeated by using 10 phoneme clusters. The experiments resulted in a lower

AABE as expected. The AABE is decreased to 12.95 ms when hard decision fusion is applied.

Note that, the phoneme class based approach in decision fusion is applicable to speech segmentation, as at most of the cases, the phonetic transcriptions of the speech data to be segmented are available, which is also true for the forced alignment method used in this study. The results achieved by different fusion techniques are presented in Table 3.9, with the results using baseline feature vector, +h1h2u1u2 vector, +PCA12 vector.

The accuracies of the system using different fusion methods and different feature vectors are represented in Table 4.8. The accuracy is calculated as the percentage of the boundaries, which has an absolute value smaller than the corresponding threshold value. It is observed that accuracies always increase as the AABE decreases. The accuracies of the system with minimum AABE is close to the manual segmentation for the thresholds of 20 ms and 50 ms. However, the accuracies for the lower thresholds is very low for the AS systems and should be increased further.

3.5 Discussion

The incorporation of the visual modality to automatic speech segmentation problem is investigated in this chapter. In Chapter 2 AABE had been decreased by up to 18% by using the horizontal and the vertical positions of the upper lip and the lower lip. The visual data collected by a camera contains much more information about the state of the articulators, than the data from electromagnetic articulograph that is used in previous chapter. Also the collection of the camera recordings is much more feasible than the other techniques such as; electromagnetic articulograph, X-Ray, etc. The aim of the studies in this chapter is extending the improvements achieved in the previous chapter by using the camera recordings of the speaker. A Turkish audiovisual speech database is collected and prepared in our laboratory for this purpose. The database includes audio recordings of 1600 Turkish sentences and camera recordings of the speaker's

head during the articulation. There are 12 markers located on the speaker's face to make the visual feature extraction easier. Several shape based visual features are extracted using positions of the markers, and an appearance based feature vector is extracted by using eigenface approach. These visual features are appended to the MFCC acoustic feature vector and the performance of the AS system using different concatenated feature vectors is examined. AABE is decreased by 18% by using +h1h2u1u2 visual features in concatenation with the acoustic features.

The performance of the audiovisual feature vectors are examined in a boundary class based manner, where the phones are clustered into visual boundary classes. The experiments have shown that using visual information caused decrease in AABE at almost all types of boundaries, however the type of the visual feature vector causing the least AABE varies between boundary classes. Multiple decision fusion algorithms are used in order to use the complementary outputs of the different systems using different feature vectors. As the phonetic transcriptions are available in AS problem, the fusion algorithm is used in a boundary class based manner. By using the visual feature vectors selectively for different boundary types (hard fusion), approximately 27% decrease in AABE is achieved.

In this study, it is shown that integration of visual information increase the performance of an AS system even recorded with a cheap, widely available webcam. It is known that as the automatic segmentation results get closer to the manual segmentation results, the quality of the TTS systems using the database increase. Bimodal AS approach enables the time alignment of wide databases with boundary locations closer to manual segmentation.

CHAPTER 4

BOUNDARY REFINEMENT TECHNIQUES

4.1 Introduction

Boundary refinement (also referred to as boundary correction or fine tuning) is the process of locating the actual boundaries of the utterances of a speech database in a more precise way by using the boundary locations previously estimated by an AS system, acoustic properties of speech, statistical information about the speech segments and other available information. In other words, it is a second stage that takes the boundaries estimated by an automatic segmentation system as input and searches for a better boundary location around these initial estimates. The refinement process is generally applied to HMM AS systems, as these systems are very good in phone identification, but they work in lower frame rates (∼100 frame/s), where the frame rate of AS systems are typically needed to be about 200 to 1000 frame/s. The time precision of a HMM AS system can be increased by using a smaller window size and a smaller slide rate, but increasing the time resolution decreases the frequency resolution of the system and decreases phone identification capacity, which is also the case for AS systems other than HMM AS systems. However, decreased phone identification capacity causes a drastic increase in the number large errors that ruins the overall system performance. For example, the HMM AS systems introduced in previous chapters have frame rates of 100 frame/s. To increase the frame rate, the experiments on the Turkish audiovisual speech database are carried out by adapting the proposed system to a frame rate of 200 frame/s. The expected results are observed in these experiments. The AABE error increased drastically

from 17.77 ms. to 88.25 ms., while the percentage of absolute boundary errors smaller than 50 ms. decreases to 37.16% from 96.12%. The AABE errors and the accuracies for different thresholds are presented in Table 4.1.

Table 4.1: AABE and Accuracy of different frame rate segmentation systems for different thresholds

Segmentation System	AABE	<5ms	<10ms	<20ms	<50ms
Baseline (100fps)	17.77	18.16%	38.41%	70.86%	96.12%
Increased Frame Rate (200fps)	61.39	2.26%	4.50%	9.82%	37.59%

Because of the need of increasing the precision of the AS systems, two stage approaches are widely used in the literature. Usually the boundaries obtained by the first stage AS system have very few large errors and many small errors due to the poor time resolution. Refinement process should decrease the magnitudes of the small errors without adding new large errors to the ones obtained in the first stage. In the first stage, phonetic boundaries are found with a relatively lower frame rate system, and then these boundaries are refined through a second process using some spectral measures, or additional information from the database.

Two stage approaches for AS are also similar to the strategy followed by the segmenters during manual segmentation process. In manual segmentation, the segmenters find the location of the boundary roughly by using the spectrogram of the speech utterance and then try to mark the boundary precisely by listening the utterance, by inspecting the speech waveform in time domain, at most of the phonetic boundaries.

Several approaches exist in the literature; in [18], average deviations from the hand labeled boundaries are calculated for different boundary classes and the boundaries from the first stage are shifted by boundary specific average deviation. A context dependent approach uses boundary models composed of a fixed length sequence of GMMs for every phoneme pair. Feature vectors contain elements such as mean energy, voicing, MFCCs and their deltas. Ultimately, the

boundary is found around the boundary point estimated in the first stage so as to maximize its likelihood given the model [19, 20]. GMM based refinement is also applied in [21], a homogeneity criterion is defined and speech is divided into segments using this criterion. Another similar method aims to minimize audible signal discontinuities caused by spectral mismatches when concatenating these units [22]. Voicing information is used to refine the boundaries in [23]. Weighted spectral slope metric, [29], is adapted to find the boundary as the point at which the spectral discontinuity is maximum. The search interval for the maximization is determined according to the boundary class. A more comprehensive work [15] involves building an ANN boundary model for the second stage, which uses statistical information such as average durations of the phonemes in the database and probability distribution function of the boundary around the boundary found by first stage and also acoustic features such as energy, correlation and log energy spectrum of the signal.

The organization of this chapter is as follows; The proposed HMM topology for boundary modeling is presented in Section 4.2, the use of the topology in boundary refinement is explained and refinement results are presented. In Section 4.3 the boundary refinement algorithm based on a glottal inverse filtering based distance measure between consecutive speech segments is explained, and refinement results belonging to two databases are presented. In Section 4.4 two boundary refinement systems are used in combination, and the overall improvements in the AABE are listed. Discussions about the Chapter are stated in Section 4.5.

4.2 HMM Based Boundary Refinement

HMMs are widely used in speech processing applications because of their excellent time warping abilities. It is also explained in Section 4.1 that increasing the frame rate of HMM systems decreases the phone identification capacity of the system. Moreover, the goal of speech segmentation is to determine a point (a frame) on the speech waveform, namely the boundary point, but traditional HMM systems divide speech waveform into group of frames that are assigned to

corresponding states of the models. In other words, HMM AS systems do not try to find the boundary locations, but try to find the state sequence that maximizes the probability of the observation of the given feature sequence. A new HMM topology that eliminates these problems, while keeping the time warping ability, is designed for the second stage of segmentation.

A context dependent approach is used in building the boundary refinement stage. For each boundary type between phoneme classes, a 3 stage left to right HMM topology is proposed to model the boundaries [78] (Figure 4.3). 1^{st} state of the HMM is associated with the first phoneme class, the 2^{nd} state is associated with the boundary frame and the model jumps to 3^{rd} state immediately in order to restrict the boundary state to last only one frame, this is achieved by setting the transition probability from 2^{nd} state 3^{rd} state (a_{23}) to 1. The 3^{rd} state is associated with the second phoneme class (Fig. 4.2). Boundary models for each phoneme to phoneme boundary could have been developed if the size of the training and the test data were adequate. Instead, phonemes are divided into 10 classes and the boundary models are developed for each phoneme class to phoneme class boundary, resulting in 121 boundary models (Breath and silence is the 11^{th} class for both cases).

The phonemes in MOCHA-TIMIT database are assigned to 10 classes with respect to their acoustical properties as follows;

- **Close/near-close front vowels :** /iy/, /i/, /ii/, /i@/;
- **Closed middle vowels and semi vowels :** /ei/, /e/, /eir/, /y/, /w/, /v/;
- **Open vowels :** /aa/, /ai/, /a/, /@/, /@@/;
- **Back vowels :** /oo/, /ou/, /ow/, /o/, /oi/, /u/, /uu/, /uh/;
- **Voiced plosives :** /b/, /d/, /dh/, /g/, /ng/;
- **Unvoiced plosives :** /p/, /t/, /th/, /k/;
- **Liquids :** /r/, /l/;

- **Voiced fricatives :** /s/, /z/, /zh/, /jh/;
- **Unvoiced fricatives :** /sh/, /f/, /h/, /ch/;
- **Nasals :** /m/, /n/;

The phonemes in Turkish audiovisual database are assigned to 10 classes with respect to their acoustical properties as follows;

- **Close/near-close front vowels :** /ix/, /i/, /I/, /Ix/;
- **Close/front vowels :** /e/, /ex/, /y/, /yx/;
- **Open vowels :** /a/, /ax/, /o/, /ox/;
- **Close/ back vowels :** /O/, /Ox/, /u/, /ux/;
- **Voiced plosives :** /b/, /d/, /g/, /gj/;
- **Unvoiced plosives :** /p/, /t/, /c/, /k/;
- **Liquids and glides :** /G/, /r/, /L/, /j/, /l/, /w/;
- **Voiced fricatives :** /v/, /s/, /z/, /Z/, /dZ/;
- **Unvoiced fricatives :** /f/, /h/, /S/, /tS/;
- **Nasals :** /m/, /n/, /N/;

/breath/ and /silence/ are also added as 11^{th} phoneme group, for each case.

4.2.1 Training HMM Boundary Models

A HMM boundary model is trained for each phoneme class to phoneme class type. Assuming that the phonemes are clustered to **N** phoneme classes, there will be **NXN** HMM boundary models (121 boundary model is trained in this case). An overview of the boundary model training stage is shown in Figure 4.1.

The phonetic transcriptions of the utterances and the locations of the boundaries marked by the manual segmenters are stored in *.lab* files in both databases. The

transcriptions are processed according to the phoneme classes defined above and corresponding phoneme class names are substituted with the phoneme names.

The training data for each boundary class is extracted by using the phoneme class based transcriptions. The training data for each boundary model is composed of the acoustic data belonging to phoneme couples belonging to that boundary class and the location of the boundary for corresponding bigram. Hence, 121 data sets are generated for the training of the HMM boundary models.

The proposed HMM topology has the advantage of avoiding complex, iterative and time consuming HMM training methods such as Baum-Welch algorithm [79], or some gradient based techniques [80] that are generally used in HMM training. Instead, the training is very fast and practical as the feature vectors belonging to the states of the HMMs are strictly determined by the definition of the topology. For each training set the starting 30% of the acoustic feature vectors belonging to the first phoneme class are omitted in order to get rid of the portions of the starting phoneme that also carries the properties of the preceding phoneme (the phoneme before the diphone.). The remaining feature vectors belonging to first phoneme are used to estimate the probability distribution function belonging to 1^{st} state of the boundary model. The feature vectors corresponding to the boundary location are used to estimate the probability distribution belonging to 2^{nd} state and the feature vectors belonging to second phoneme are used to estimate the probability distribution of the 3^{rd} state after the last 30% of the features of each sample is omitted as described before.

The probability distributions of the states are modeled with Gaussian mixture models (GMMs). GMM parameters are extracted using the frames that are associated to the phoneme classes and the boundary points in the training data. The parameters of the GMMs with one mixture are calculated by finding the means and the variances of the feature vectors belonging to the corresponding state.

After the probability distributions of each state of the HMM are estimated, the transition probabilities of the HMM topology are calculated as follows:

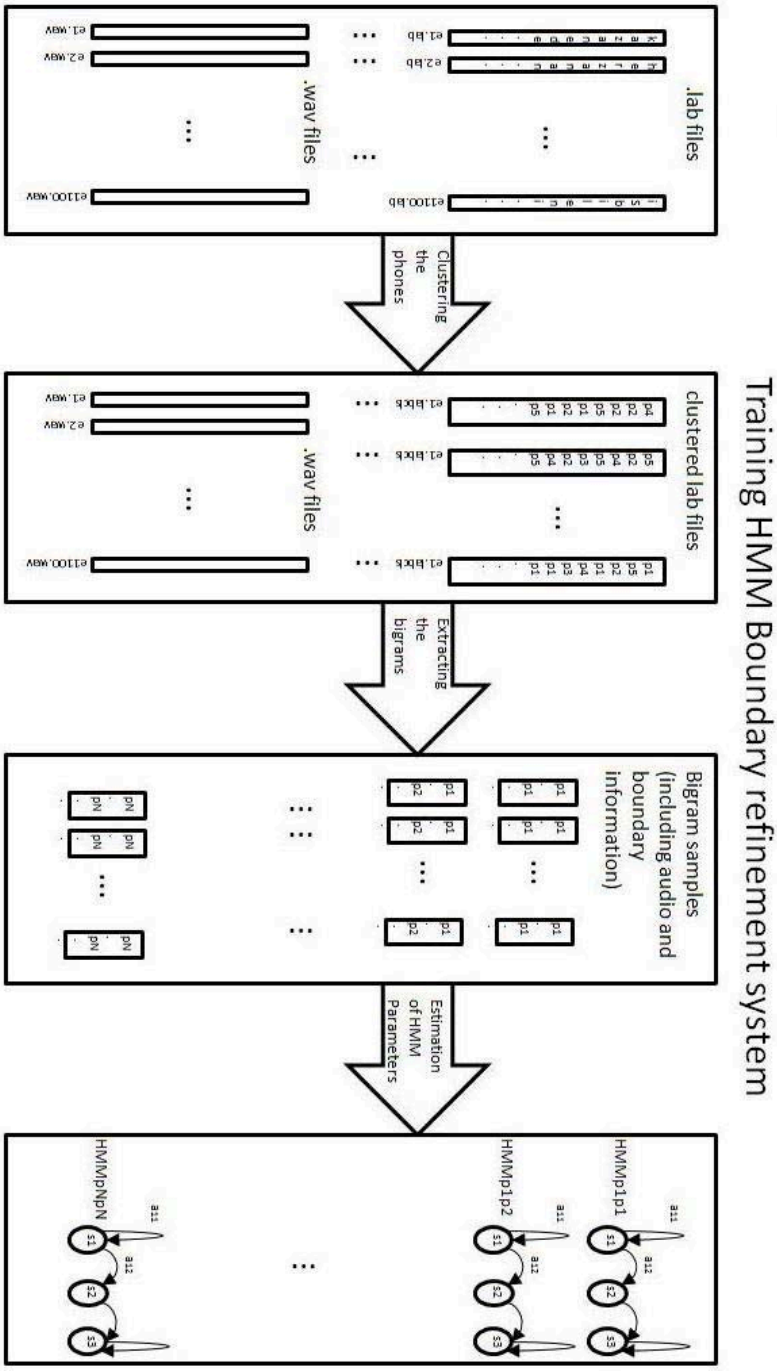

Figure 4.1: Training the HMMs

$$\mathbf{a}_{11} = 1 - \frac{1}{N_1} \tag{4.1}$$

$$\mathbf{a}_{12} = \frac{1}{N_1} \tag{4.2}$$

$$\mathbf{a}_{22} = 0 \tag{4.3}$$

$$\mathbf{a}_{23} = 1 \tag{4.4}$$

$$\mathbf{a}_{33} = 1 \tag{4.5}$$

$$\tag{4.6}$$

where, a_{ij} is the transition probability from i^{th} state to j^{th} state and N_1 is the average number of the frames observed in the 1^{st} state.

The transition probability from first state to second state (a_{12}) is inverse of number of average feature vectors observed before boundary point. The probability to stay in 1^{st} state (a_{11}) is 1-a_{12}. The transition probability from 2^{nd} state to 3^{rd} state and from 3^{rd} state to itself is 1, as the 2^{nd} state should last only one frame according to the proposed HMM topology (Figure 4.3).

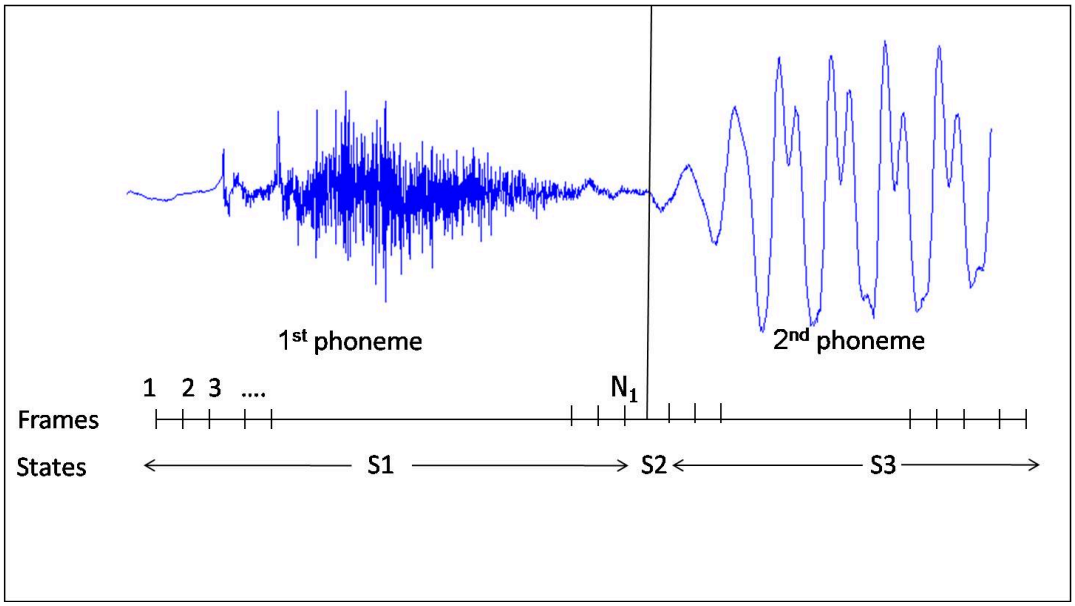

Figure 4.2: Assigning frames to HMM states

The acoustic feature vector that is used in Chapter 2, containing 13 MFCCs (including energy coefficient), and their first and second order derivatives and

they were used in the proposed system. The feature vectors are obtained by using 10 ms windows. The sliding rate of the windows are set to be 5 ms for voiced-voiced phoneme couples, and 1 ms for the phoneme couples including an unvoiced phoneme group (voiced-unvoiced, unvoiced-voiced and unvoiced-unvoiced), as boundary location should/can be detected more precisely in the boundaries including an unvoiced phoneme. The precision for the voiced-voiced boundaries are less, as the change in the speech waveform is examined in a pitch synchronous manner during manual segmentation and automatic segmentation most of the time.

The training of the HMM boundary models are finished upon finding the pdf for each state and state transition probability for the first state of each HMM.

4.2.2 Refining Boundaries Using HMM Boundary Models

Viterbi algorithm is generally used for associating the observed frames to HMMs. The algorithm determines the most likely sequence of hidden states using the observed frames and HMM parameters [81]. In this work, Viterbi algorithm is used for the decoding of the HMMs and detection of boundary frame.

The locations of the boundaries estimated by the AS systems discussed in the previous chapters were saved in *.rec* files by HTK. These boundary locations will be refined in this stage. The boundary refinement stage is similar to the training stage described in Section 4.2.1. The phonetic transcriptions of the automatically segmented utterances in the *.rec* files are arranged according to phoneme classes as in the training stage, i.e., the names of the phonemes are substituted with the corresponding phoneme classes. For each utterance, the boundary refinement process is handled diphone by diphone. The starting 30% of the first phoneme and the last 30% of the second phoneme are omitted for each diphone, like at the training stage. Then, boundary model corresponding to that phoneme class to phoneme class boundary is used to decode the observed features using Viterbi algorithm. The location of the feature vector that is decoded as belonging to second state of the HMM boundary model is marked as the refined boundary.

The proposed HMM based boundary refinement system is firstly tested on two boundary classes from the MOCHA-TIMIT database, a glide-vowel boundary (/y/-/uu/) and a plosive-vowel boundary (/t/-(/uu//o/)). There are 125 utterances of /y/-/uu/ boundary, The experiments are carried out by using 5-fold validation set i.e., at each time 100 utterances are used to train the boundary model and 25 utterances are used for test, this is repeated 5 times to use all data. Same procedure is repeated for 50 utterances of /t/-(/uu/-/o/) boundary, 40 utterances are used for training and 10 are used for test each time. The test results showing the AABE achieved by using each boundary model are listed in Table 4.2. The AABE is decreased by 44% for /y/-/uu/ boundary and 63% for /t/-(/uu//o/) boundary.

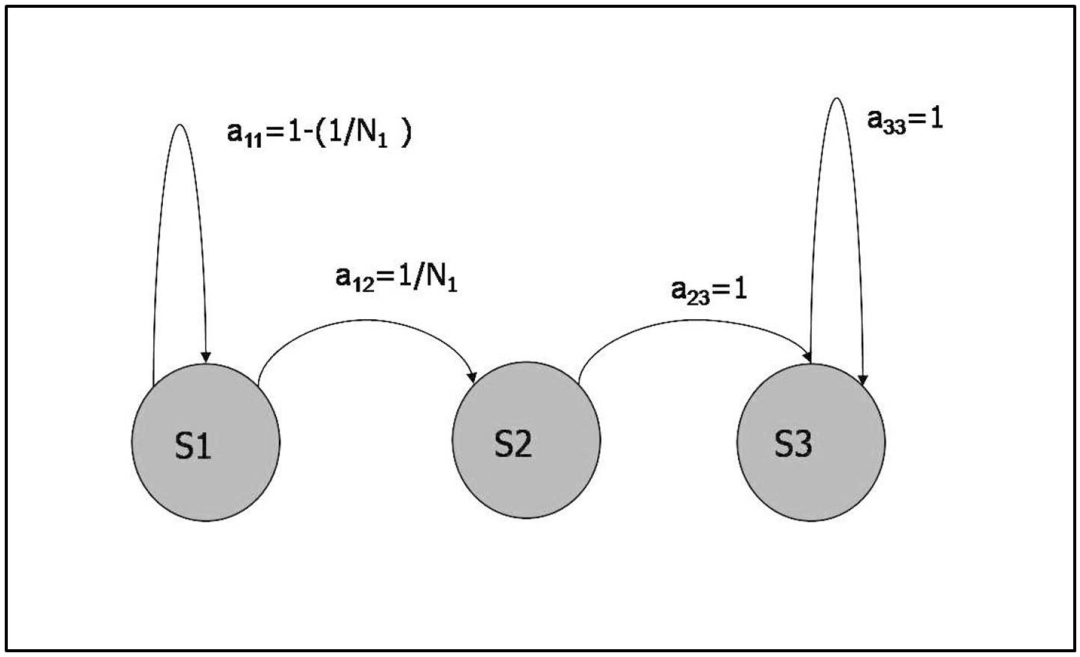

Figure 4.3: Three state HMM topology: States and transition probabilities

The proposed HMM boundary refinement algorithm is tested on both of the databases using the AS results obtained by the AS systems introduced in Chapter 2 and Chapter 3. The HMM boundary refinement experiments over MOCHA-TIMIT database are done by using the same training and the test data in chapter 2 as the boundaries found by the 1^{st} stage will be used in the 2^{nd} stage. Hence

Table 4.2: Average absolute boundary errors (ms) for /t/-(/uu//o/) boundaries

Boundary	Average absolute boundary error for New HMM (ms)	Average absolute boundary error for 1st stage (ms)
/y/-/uu/	8.9 (44% decrease)	15.8
/t/-(/uu//o/)	3.1 (63%)	8.3

420 of the utterances are used for recording and 40 of them are used for test.

The HMM boundary refinement algorithm decreased the average absolute boundary error approximately by 18-20% (Table 4.3).

Table 4.3: Average absolute boundary errors (AABE) (ms) for 1^{st} stage, and after HMM and inverse filtering based boundary refinement (MOCHA-TIMIT Database)

1^{st} stage AS system	AABE at 1^{st} stage (ms)	AABE after HMM boundary refinement (ms)
Baseline	9.94	7.96 (19.9%)
Lly-uly	8.91	7.32 (17.75%)
Clusterbased-visual	8.26	6.77 (18.04%)

The training set of 1000 utterances are used for training the HMM boundary models, 500 of them are used for test. Using the same algorithm on Turkish audiovisual speech database slightly better results are achieved, probably due to the better estimation of parameters of the HMM boundary models and better evaluation of the test results, as a result of the availability of more training and test data (Table 4.4).

The experiments had shown that using the proposed boundary models based on the modified HMM topology the AABE decreased about 20% for both databases.

Table 4.4: Average absolute boundary errors (AABE) (ms) for 1^{st} stage, and after HMM and inverse filtering based boundary refinement (Turkish audiovisual speech database)

1^{st} stage AS system	AABE at 1^{st} stage (ms)	AABE after HMM boundary refinement (ms)
Baseline	17.77	13.75 (22.6%)
wh	15.16	12.14 (19.92%)
Clusterbased-visual	13.28	10.92 (17.77%)

4.3 Glottal Inverse Filtering Based Boundary Refinement

4.3.1 Glottal Inverse Filtering

Glottal inverse filtering (GIF) is used in a wide range of speech processing applications. The production of speech is approximated as a source-filter process and obtaining these source and filter components provides a flexible representation of speech signal. Inverse filtering is primarily used in assessment of laryngeal aspects of voice quality and for correlations between acoustic and vocal fold dynamics etc [73]. Extraction of glottal signal is not primarily used but accurate inverse filtering remains important in modeling the speech signal flexibly, in some applications such as; harmonic pulse noise modeling [74], sinusoidal modeling [75], voice conversion, or speech synthesis.

In this study we used GIF to develop a distance measure between speech segments for the detection of the boundary point between two voiced phonemes. The boundary point is described as the sample where the dissimilarity between successive frames of the estimated glottal waveform is maximum. Iterative adaptive inverse filtering is used in order to estimate GIF.

For the voiced speech case, speech waveform can be represented as the convolution of glottal flow, vocal tract response and radiation effect of the lips [73]. In Z domain this equation can be written as;

$$S(z) = A_v P(z)G(z)V(z)R(z) \tag{4.7}$$

Where A_v is gain of the system and S(z), P(z), G(z), V(z), R(z) are z transforms of speech waveform, glottal excitation, glottal filter, vocal tract impulse response and radiation effect of the lips, respectively.

The glottal excitation p[n] is not actually a physical signal but a mathematical input to a filter which will generate the glottal flow waveform. Lip radiation is usually represented by a simple differencing filter:

$$R(z) = 1 - z^{-1} \tag{4.8}$$

Glottal input can be represented as:

$$P(z)G(z) = \frac{S(z)}{AV(z)R(Z)} \tag{4.9}$$

Radiation term can be included in the glottal waveform by defining the signal q(n) with z transform:

$$Q(z) = P(z)G(z)R(z) \tag{4.10}$$

by using eqution 4.9:

$$Q(z) = \frac{S(z)}{AV(z)} \tag{4.11}$$

In equation 4.11 only Z-transform of the speech waveform, S(z), is known. Solving for both Q(z) and V(z) is a blind deconvolution problem. One method is using the time instants when the glottis is closed (g[n]=0). While the glottis is closed, the speech waveform must be simply a decaying oscillation which is only a function of the vocal tract and its resonances or formants i.e., it represents the impulse response of the vocal tract. Solving for the system during this closed phase should exactly capture the linear vocal tract filter, V (z), which may then be used to inverse filter and recover Q(z). G(z) may then be reconstructed by integration (equivalently by inverse filtering by 1 /R(z)). This approach is known

as closed phase inverse filtering and is the basis of most approaches to recovering the glottal flow waveform. However, especially for high pitch speakers, closed phase is very small and the data is not enough to estimate the vocal tract filter.

There are a number of source filter decomposition algorithms, such as Zeros of Z-Transform (ZZT), Cepstrum based minimum-maximum phase decomposition, analysis by synthesis or iterative methods such as iterative adaptive inverse filtering (IAIF).

4.3.2 Iterative Adaptive Inverse Filtering Algorithm

It is mentioned in previous section that estimation of the vocal tract response is needed to find the glottal waveform. Iterative Adaptive Inverse Filtering (IAIF) Algorithm proposed by Alku [82], operates in two repetitions. The overall structure of the algorithm is shown in Figure 4.4 [83]. The first phase, blocks 2 to 6, generates an initial estimate of the glottal excitation, which is subsequently used as input of the second phase, blocks 7 to 12, to achieve a more accurate estimate. The steps of the method are described in detail below.

1. The input signal is first high-pass filtered to remove disturbing low-frequency fluctuations captured by the microphone. The cut-off frequency should be lower than the fundamental frequency of the speech signal in order to avoid filtering out relevant information. The found signal s[n] will be used in following steps.

2. A first-order LPC analysis is calculated in order to compensate for the -12dB/octave effect caused by voice source and +6dB/octave high pass effect from the lip radiation. Thus, a first order, discrete all-pole model (DAP) $H_{gl}(z)$ for the combined effect of -6dB/octave effect is calculated.

3. s[n] is inverse filtered using $H_{gl}(z)$ to effectively remove the spectral tilt caused by the spectrum of the excitation signal and the lip radiation effect and to obtain pressure signal $s_{gl}[n]$ that contains only the effects of the vocal tract response and the impulse-train excitation.

4. The output of the previous step is analyzed by a p^{th} order LPC to obtain a model of the vocal tract transfer function ($H_{vt1}(z)$). The order p of the LPC analysis is usually two times the sampling frequency in kHZ, i.e., p is chosen as 32 for 16 kHz sampling rate.

5. The effect of the vocal tract is canceled from s[n], by inverse filtering s[n] with the signal found in step 4.

6. The output of the previous step, $\dot{g}_1[n]$, is integrated in order to cancel the lip radiation effect. This yields the first estimate of the glottal flow, $g_1[n]$, and completes the first repetition.

7. The second repetition starts by calculating a g^{th} order analysis of the obtained glottal flow estimate. This gives a spectral model of the effect of glottal excitation on the speech spectrum. The value of g is usually between 2 and 4.

8. The input signal is inverse filtered using the model of the excitation signal to eliminate the glottal contribution.

9. Lip radiation is canceled by integrating the output of the previous step.

10. A new model of the vocal tract filter is formed by an r^{th} order LPC analysis.

11. The effect of the vocal tract is removed from the input signal by inverse filtering it with the vocal tract model obtained in the previous step.

12. Finally, the lip radiation effect is canceled by integrating the signal.

This yields the final estimate of the glottal flow, which is the output of the IAIF method.

4.3.3 Boundary Estimation Using a Dissimilarity Metric Based on IAIF

The speech waveform changes slowly in a gradual manner from one voiced phoneme to another voiced phoneme, especially at the boundaries between vowels, liquids and glides. When the spectrum of the speech waveform is observed

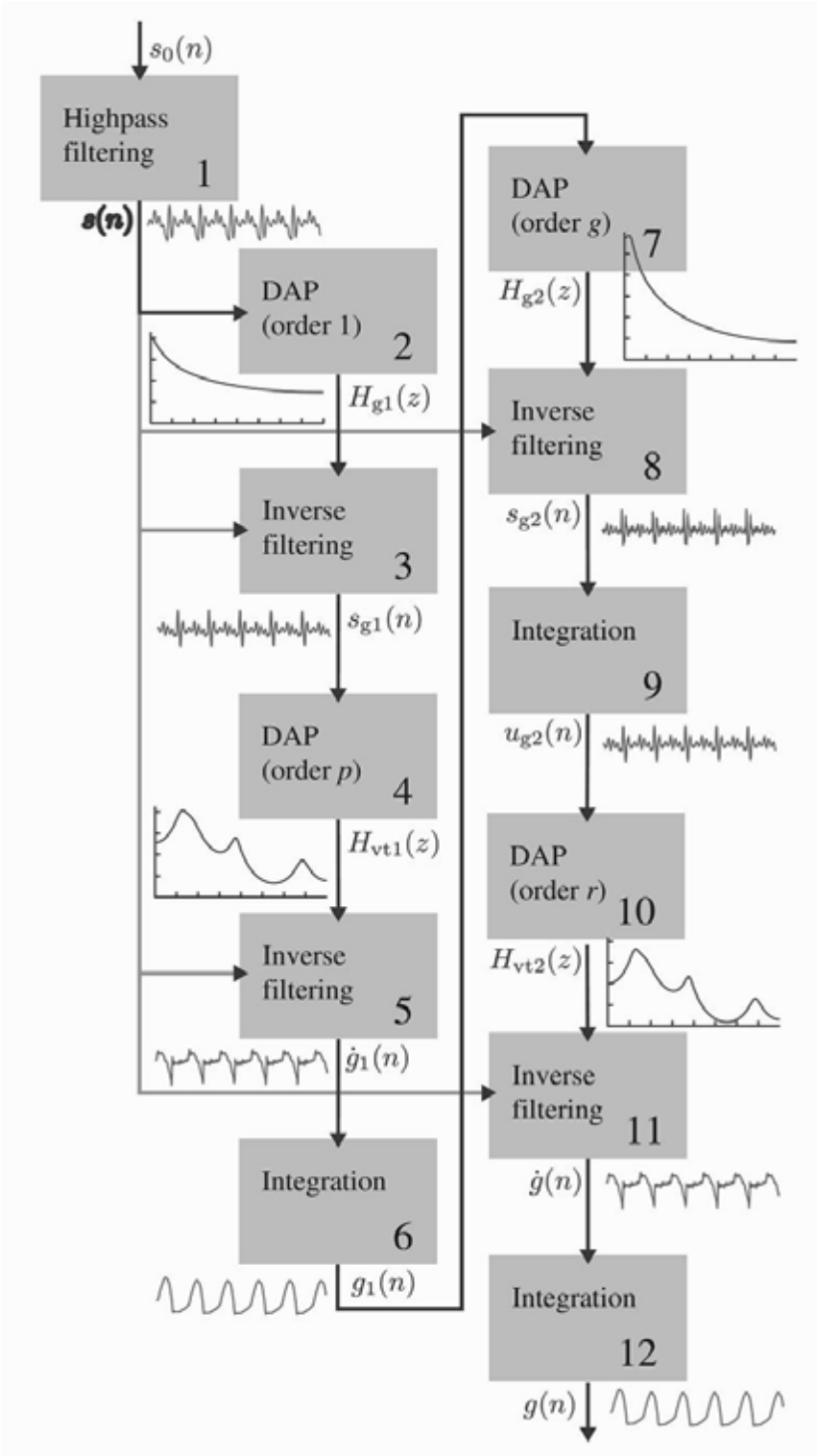

Figure 4.4: The Structure of IAIF Algorithm [83]

it can be noticed that, the formants of the speech waveform slowly evolves from the formants of the first phoneme to the formants of the following phoneme as stated in Section 3.2.3. The change in the speech waveform is similar, the waveform changes from initial state to target state slowly in each period of speech, the gradual change in the speech waveform can be seen in Figure 4.5, the waveform belonging to /e/-/a/ bigram. In such ambiguous cases, where it is hard to locate the boundary point, manual segmenters are asked to find upper and lower margins for the possible location of the boundary, and mark the point where the difference between two consecutive quasi-periodic segments is maximum as the boundary point (Section 3.2.3). The approach suggested to mark the ambiguous boundaries can be adapted to boundary refinement problem.

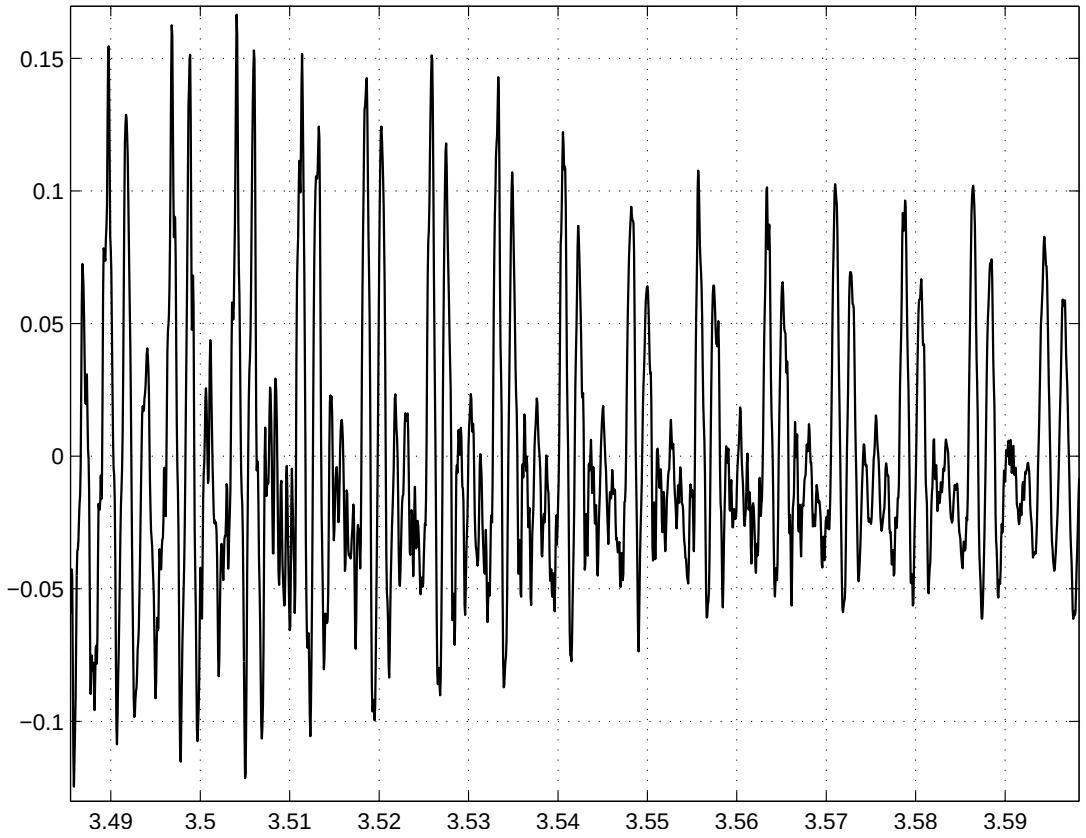

Figure 4.5: The acoustic waveform belonging to /e/-/a/ diphone

The boundary information from the 1^{st} stage provides the upper and lower margins for the location of the boundary to be found. As seen in Figure 4.5

the change from one 'period' of speech to the next one is hard to notice or compare with the following periods most of the time. To develop a metric of change or dissimilarity is hard for both manual segmenters and for automatic segmentation system. In this study IAIF algorithm is used to obtain a 'simpler' waveform that allows easier comparison of the consecutive periods and also much more susceptible to the change in the vocal tract transfer function.

When using IAIF algorithm it is generally assumed that the speech signal to be investigated is stationary, as a result the whole signal available is used for the blind deconvolution of the glottal input and vocal tract response functions. After the vocal tract response function is estimated, by inverse filtering the speech waveform with this function glottal input is found. However, the stationarity assumption is not true around the boundaries. Actually, the algorithm will be used to detect the point where the signal (vocal tract) changes the most, by operating on the speech segments around the boundaries.

A similar approach to the one in Section 4.2.1 is followed in this section. Using the boundary locations estimated in the 1^{st} stage, the bigram in which the boundary will be refined is extracted. The starting 30% of the first phoneme and the last 30% of the second phoneme are omitted. Then the samples from the starting 50% of the remaining of the first phoneme is extracted to be used in IAIF algorithm (Figure 4.6). The vocal tract transfer function belonging to the extracted segment from the 1^{st} phoneme is estimated by IAIF algorithm and after that, the whole waveform (bigram) is inverse filtered with the estimated vocal tract transfer function. The output of the filtering operation is the glottal waveform of the bigram, estimated by using the properties of the first phoneme only. As the estimation process is dependent on the vocal tract transfer function solely, the estimated waveform is very susceptible to the change in the vocal tract that is hidden in the speech waveform. At most of the cases, if the estimation of the vocal tract transfer function of the first phoneme is successful, the resultant waveform shows a drastic change near the boundary point (Figure 4.7, 4.8).

The consecutive segments of the estimated waveform should be compared with each other in order to find the instant where maximum change between these

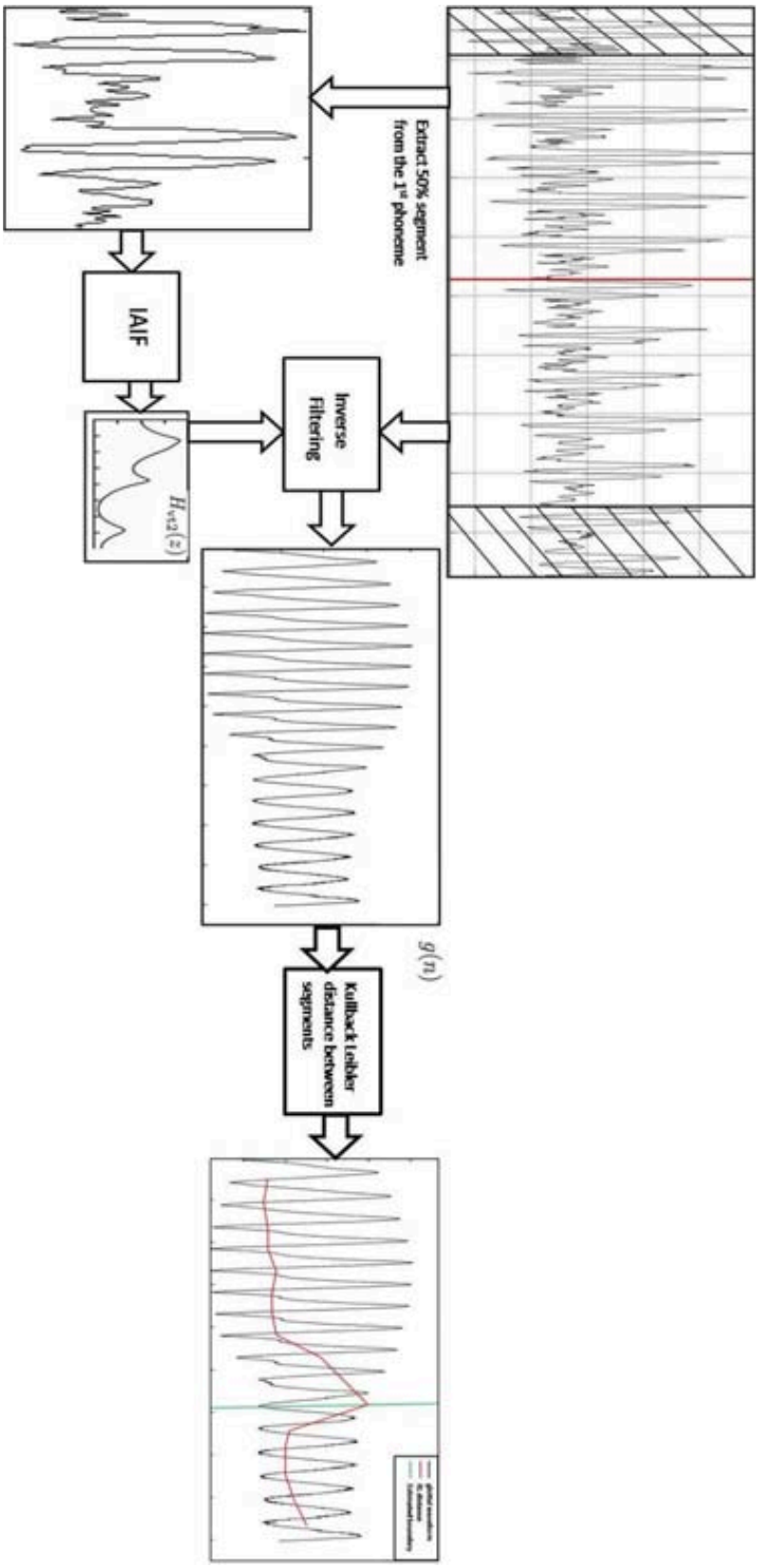

Figure 4.6: Fine tuning using IAIF

segments occurs. The change in the glottal waveform function is detected by calculating the symmetrised Kullback-Leibler distance (divergence) [84] between consecutive pitch periods of the waveform. This distance measure is selected as it omits the change in the amplitude and uses the shape of the waveform. The Kullback-Leibler divergence is actually measures distance between two probability distribution functions. The Kullback-Leibler divergence between two continuous random variables **P** and **Q** is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx \tag{4.12}$$

where p(x) and q(x) denote the densities of P and Q. The symmetrised distance is

$$D_{KLs}(P||Q) = D_{KLs}(Q||P) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{4.13}$$

The comparisons of the dissimilarity between consecutive speech segments are made pitch synchronously. First the pitchmarks of the estimated glottal waveform belonging to the bigram are found and then, at each pitchmark, the distance between the speech segments among that pitcmark and the previous and next pitchmarks is calculated by using the symmetrised Kullback-Leibler distance. The instant where the symmetrised Kullback-Leibler distance between consecutive segments is found to be maximum is marked as the boundary location.

Proposed method is tested for some boundaries between vowels and liquids, the results of the experiments are presented in table 4.5. Proposed algorithm outperforms HMM fine tuning algorithm in some diphone classes, in other classes HMM fine tuning is better. The fine tuning using inverse filtering does not need any training stage; this is an important advantage of the system compared to HMM fine tuning. These systems can be used selectively or in cascade in order to decrease the overall AABE further.

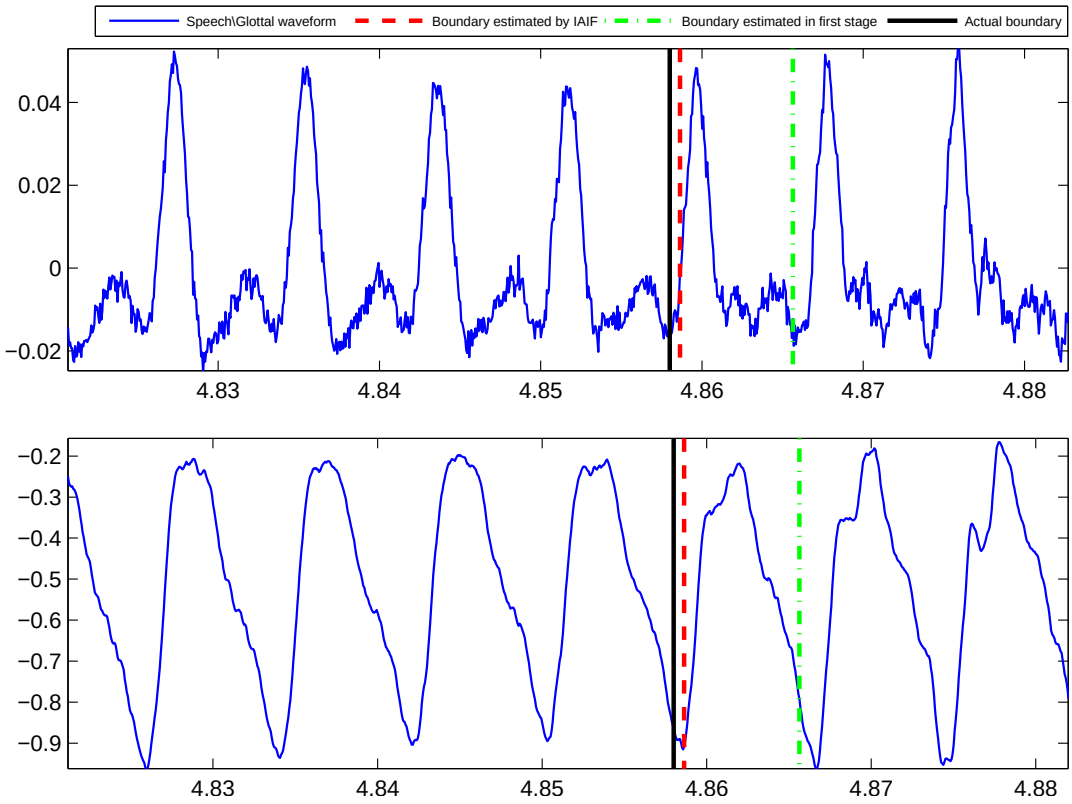

Figure 4.7: The acoustic waveform, glottal waveform and the boundary locations of /I/-/L/ diphone

Boundary refinement stage using IAIF is tested on the MOCHA database. As training stage is not needed for glottal inverse filtering based boundary refinement, the training set is omitted and the algorithm is used on 40 utterances that were previously used as the test set for the 1^{st} stage. Three different systems defined in Section 2.4 are used as the first stage AS systems, namely; the systems using the baseline feature vector, MFCC_0_D_A+(uly-lly) feature vector and the one using all the feature vectors selectively. The results are presented in Table 4.6. The proposed boundary refinement algorithm managed to decrease the AABE by approximately 8.5% for all three 1^{st} stage systems. Note that, the algorithm can only be used in voiced-voiced boundaries as the glottal inverse filtering is defined only for voiced phonemes. The SAMPA notation for English has 27 voiced phonemes within 46 phonemes. This means that the proposed algorithm can be used 729 of 2116 phoneme boundary types that means 34% of the boundary types.

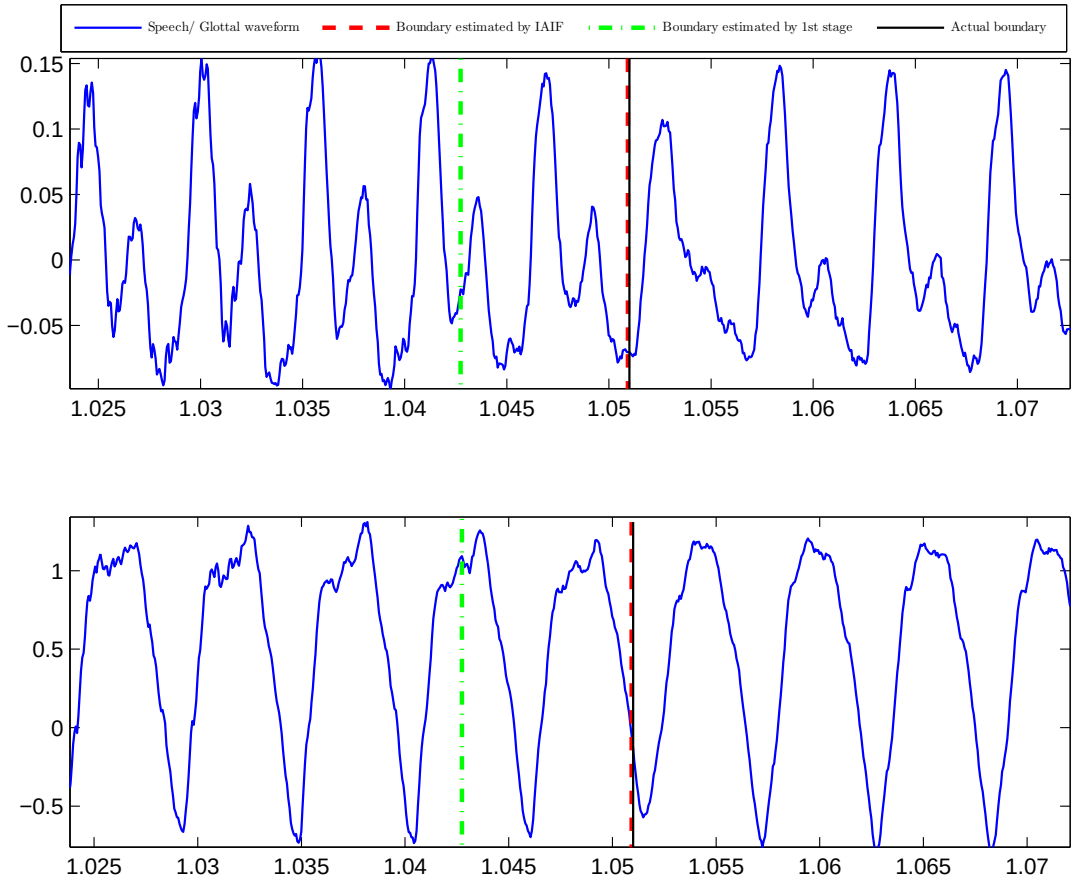

Figure 4.8: The acoustic waveform, glottal waveform and the boundary locations of /y/-/m/ diphone

Glottal inverse filtering based boundary refinement algorithm is also tested on Turkish audiovisual speech database by using 500 test utterances that are used in Chapter 3. The boundaries found by using the baseline feature vector, +wh feature vector, and the boundaries found by using selective decision fusion are used as input to the boundary refinement algorithms. The results are presented in Table 4.7. The proposed boundary refinement algorithm managed to decrease the AABE by 5% to 8% for different 1^{st} stage systems. In SAMPA notation for Turkish, 24 of 44 phonemes are voiced. This means that the proposed algorithm can be used 672 of 1936 phoneme boundary types that means 35% of the boundary types. Also it is worth noting that, there are actually 42197 voiced-voiced boundaries in the Turkish audiovisual database, which is approximately 50% of all boundaries (Table 3.1). This means that, by using glottal inverse filtering based boundary refinement, a decrease of 10-16% is achieved in the AABE be-

Table 4.5: Average absolute boundary errors (AABE) (ms) for some phoneme couples after 1^{st} stage, and after HMM and inverse filtering based boundary refinement

Phoneme couple	1st stage	HMM fine tuning	Inverse filtering
/a/-/r/	23.5ms	12.9ms (45.11%)	8.2ms (65.11%)
/a/-/l/	18.9ms	6.3ms (66.67%)	6ms (68.25%)
/l/-/a/	25.5ms	8ms (68.69%)	11.6ms (54.51%)
/j/-/uu/	15.8ms	8.9ms (43.67%)	13.0ms (17.72%)

longing to voiced-voiced boundaries while the AABE belonging to the remaining boundary types stays the same.

Table 4.6: Average absolute boundary errors (AABE) (ms) for 1^{st} stage, and after HMM and inverse filtering based boundary refinement

	AABE at 1^{st} stage (ms)	AABE after HMM boundary refinement (ms)	AABE after inverse filtering based boundary refinement (ms)	AABE after combined boundary refinement (ms)
Baseline	9.94	7,96 (19.9%)	9.08 (8.65%)	7.64 (23.1%)
+Lly-uly	8.91	7.32 (17,75%)	8.15 (8.53%)	6.92 (22.3%)
Clusterbased-visual	8.26	6.77 (18.04%)	7,56 (8.47%)	6,43 (22.15%) (35.3% overall)

4.4 Combining HMM Based Boundary Refinement and Inverse Filtering Based Boundary Refinement

Two novel boundary refinement techniques are proposed in previous sections. The former builds HMM boundary models for each phoneme class to phoneme class boundary, and tries to locate the refined boundary around the boundaries supplied by the 1^{st} stage using these boundary models, where the latter operates

Table 4.7: Average absolute boundary errors (AABE) (ms) for 1^{st} stage, and after HMM and inverse filtering based boundary refinement

	AABE at 1^{st} stage (ms)	AABE after HMM boundary refinement (ms)	AABE after inverse filtering based boundary refinement (ms)	AABE after combined boundary refinement (ms)
Baseline	17.77	13.91 (21.72%)	16.82 (5.35%)	13.20 (25.72%)
+wh	15.16	12.14 (19.92%)	14.01 (7.55%)	11.85(21.83%)
Clusterbased-visual	13.28	10.92 (17.77%)	12.36 (6.92%)	10.42 (21.54%) (41.36% overall)

on only the voiced-voiced boundaries and without the knowledge of context, it tries to locate the instant where the 'change' in the speech waveform is maximum. Two systems operate in a very different manner, thus using these systems together has the high potential of decreasing the AABE further. The combination of two boundary refinement systems is achieved by using these systems in cascade, i.e., the boundaries found by HMM boundary refinement system are used as input to the boundary refinement system based on IAIF and a second boundary refinement is applied to those boundaries, resulting in a 3 stage system.

The experiments are carried out on MOCHA-TIMIT database using three differents systems as 1^{st} stage and then using the proposed boundary refinement algorithms separately and in cascade. The segmentation results are presented in Table 4.6. The second column shows the segmentation results of the 1^{st} stage system that uses the feature vector in the 1^{st} column. The 2^{nd} and the 3^{rd} columns show the segmentation results of HMM based and inverse filtering based boundary refinement, respectively, and the segmentation results obtained by using two of these systems in cascade are given in the 4^{th} column.

The combination of the HMM based and inverse filtering based boundary refinement decreases average absolute boundary error by 22-23% for each case. The

Table 4.8: Accuracy of different segmentation systems for different thresholds

Segmentation System	AABE	<5ms	<10ms	<20ms	<50ms
Baseline	17.77 ms	18.16%	38.41%	70.86%	96.12%
Baseline+HMM FT	13.91 ms	29.04%	53.53%	81.53%	96.53%
Baseline+Combined FT	13.20 ms	32.21%	59.22%	83.15%	96.53%
+wh	15.16 ms	21.28%	44.03%	76.92%	97.65%
+wh+HMM FT	12.14 ms	31.53%	56.75%	85.06%	97.99%
+wh+Combined FT	11.85 ms	33.30%	58.91%	86.29%	98.11%
Hard Fusion	13.28 ms	28.45%	52.33%	81.38%	97.92%
Hard Fusion+HMM FT	10.92 ms	34.03%	59.65%	87.23%	98.15%
Hard Fusion+Combined FT	10.42 ms	38.25%	63.01%	88.35%	98.91%
Manual discrepancies	9.09 ms	50.67%	73.91%	86.52%	98.07%

feature vectors used for 1^{st} stage systems are MFCC_0_D_A, MFCC_0_D_A+uly+lly, and class wise selected feature vectors as explained in Section 2.4.2.2, respectively. By using the last one the average absolute boundary error is decreased to 6.43 ms achieving a decrease of 35.3% with respect to the baseline system.

Same procedure is repeated for the Turkish audiovisual speech database using boundaries from the AS system with the baseline feature vector, MFCC_O_D_A+wh feature and the results obtained by selectively using all the features. The segmentation results are presented in Table 4.7. The second column shows the segmentation results of the 1^{st} stage system that uses the feature vector in the 1^{st} column. The 2^{nd} and the 3^{rd} columns show the segmentation results of HMM based and inverse filtering based boundary refinement, respectively, and the segmentation results obtained by using two of these systems in cascade are given in the 4^{th} column.

Using both boundary refinement algorithms decreases AABE by 21-25% for each case. When the boundaries obtained by the selectively usage of the visual features are employed as input to 2^{nd} stage, boundary refinement by using the proposed algorithms in cascade results in an average absolute boundary error of 10.42 ms which means a decrease of 41.36% with respect to the baseline system.

The accuracies of different systems with respect to different thresholds are also presented in Table 4.8. It is observed that the goal of decreasing the numbers

of small errors without increasing the number of large errors is achieved on all three AS systems. Actually the number of the large errors is slightly decreased as well. For example, the percentage of the errors smaller than 5 ms is increased to 38.25% from 28.45% for the boundaries obtained by the hard fusion of the boundaries estimated by different systems, a small decrease in the percentage of the errors greater than 50 ms is achieved as well. It is also observed that the accuracy of the combined boundary refinement systems for the thresholds of 20 ms and 50 ms is better than the accuracy of the manual discrepancies, but for smaller thresholds the accuracies of the manual discrepancies are higher.

4.5 Discussion

Refinement of the boundaries is compulsory for all AS systems in order to increase the precision of these systems. In this chapter, two new methods for the refinement of the boundaries estimated by an AS system are proposed. The first method uses a new HMM topology for the modeling of the boundaries between two phoneme classes and finds the refined boundary by using these boundary models, while the second method locates the refined boundary where the dissimilarity between consecutive speech segments is maximum, by using a distance metric over the glottal waveform estimated around the previously located boundary point. Both of these two methods are new approaches to the boundary refinement problem.

The experiments on the MOCHA-TIMIT database and the Turkish audiovisual speech database had shown that the HMM based boundary refinement algorithm decreased the AABE about 20% and the glottal inverse filtering based boundary refinement decreased the AABE about 6-9% with respect to the AABE achieved by using the AS systems in the first stage. The proposed boundary refinement algorithms were also used in combination, by supplying the output of the HMM based method to the input of the glottal inverse filtering method. Actually, using these methods in cascade, results in a 3 stage AS system that includes two boundary refinement stages (Figure 4.9). Note that, the same training and

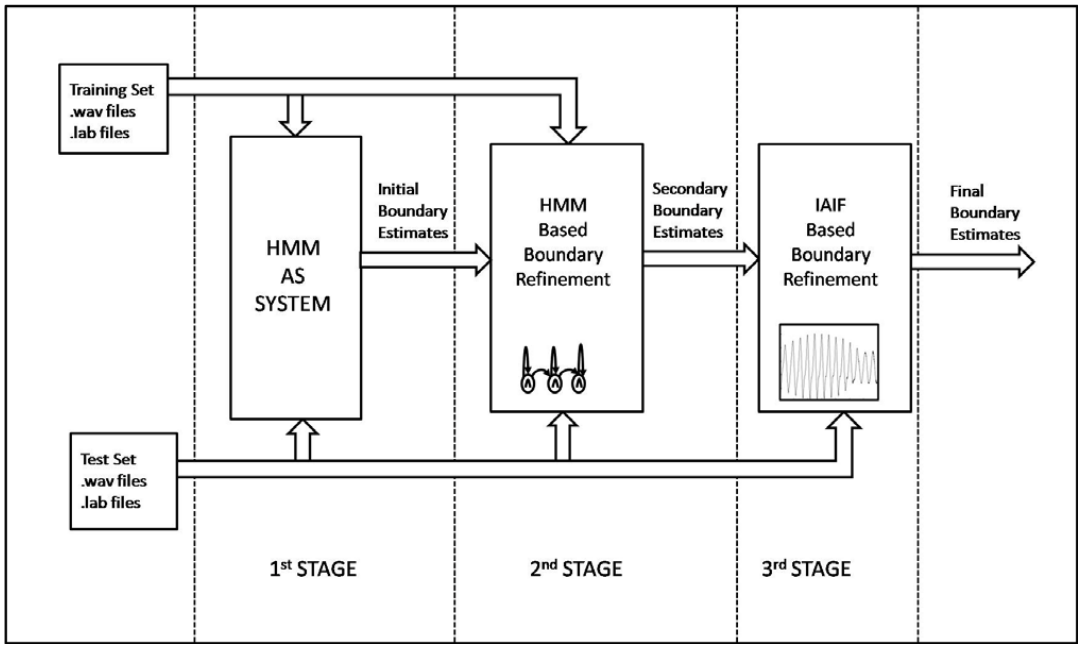

Figure 4.9: The overview of the three stage AS system

test data are used for the first and the second stage, the third stage does not need a training set, the same test data is used in the third stage too. The AABE from the first stage is decreased by 22-23% for the MOCHA database, and 21-25% for the Turkish audiovisual speech database, by the cooperation of newly proposed boundary refinement algorithms. If the decrease in the AABE achieved by including visual information in AS is also taken into account the AABE is decreased by 35.3% for the MOCHA database and 41.36% for the Turkish audiovisual speech database, by using the complete AS system that is proposed in this thesis.

CHAPTER 5

CONCLUSION

In this thesis, a complete automatic speech segmentation system is presented. Each stage of the proposed system includes novel approaches to automatic speech segmentation problem; firstly an unexplored concept of bimodal automatic speech segmentation is investigated and successfully implemented to two speech databases, and then two novel methods for the refinement of the boundaries estimated by an automatic speech segmentation system are introduced and the estimated boundaries are improved further using these methods.

- **Bimodal Automatic Speech Segmentation:** The speech researches make use of the visual modality to improve the performances of speech processing systems in several applications. However, the incorporation of the visual modality with the speech in automatic speech segmentation systems is an uninvestigated concept yet. In this thesis, the visual modality is successfully integrated to acoustic modality, resulting in improved automatic segmentation results. The information from different modalities are fused in feature level. The experiments are held on the MOCHA-TIMIT database and the audiovisual Turkish speech database developed for the studies in this thesis. Several audiovisual feature vectors are investigated, and the performances are inspected in different boundary classes, it is observed that different visual features are more useful at different boundary types. In order to benefit all the audiovisual feature types, the decisions (boundary locations) of the systems using different feature vectors are also fused. As a result, the AABE is decreased by 18% for MOCHA-TIMIT

database and 27% for the audiovisual speech database by the integration of the visual features to automatic speech segmentation process.

- **Turkish Audiovisual Speech Database:** A new database is collected and prepared to be used in the experiments in this thesis. Although some public audiovisual speech databases exist, they have either limited size or limited visual information. The prepared database covers the need for a phonetically rich data set as well as having appropriate visual data. The phonetic transcriptions and manual segmentation of the database is done and the visual features of the recordings are extracted and saved making the database ready to be used by the researchers studying on audiovisual speech processing.

  A novel approach to manual segmentation process is proposed and used in the manual segmentation of the prepared database. The boundaries are located in a boundary class-wise manner in order to decrease the intralabeler inconsistency.

- **Novel Approaches to Boundary Refinement:** Two new approaches to boundary refinement problem are introduced in this thesis. The proposed boundary refinement stages are used to refine the boundaries estimated by the audiovisual speech segmentation systems, constituting a complete automatic speech segmentation system. The proposed boundary refinement methods are;

  - **Boundary Refinement based on a New HMM Topology:** A new left to right HMM topology is proposed for developing boundary models for boundary refinement. The models developed using the proposed HMM topology are able to detect the boundary point and also can be used at high frame rates unlike the HMM systems used in the first stage. The AABE of the boundaries obtained in the first stage is decreased by 18-20% in MOCHA-TIMIT database and 18-22% in Turkish audiovisual speech database by using the proposed HMM base boundary refinement method.
  - **Glottal Inverse Filtering Based Boundary Refinement:** A

new boundary refinement algorithm using a distance metric between the glottal waveform estimate of the speech segment belonging to a phoneme couple. The glottal waveform is estimated around boundaries found at the first stage of AS. The estimated glottal waveform is very susceptible to the change in the vocal tract as only the speech segment from the first phoneme is used in the estimation process. Then the refined boundary is located where the Kullback-Leibler distance between consecutive phonemes is maximum. The drawback of the algorithm is that, it only operates in voiced-voiced boundaries. However, the ability of operating without the need of training stage is a major advantage of this method. The AABE of the boundaries obtained in the first stage is decreased by 8-9% in MOCHA-TIMIT database and 5-8% in Turkish audiovisual speech database by using the proposed HMM base boundary refinement method.

The proposed boundary refinement algorithms are integrated in cascade. The boundaries estimated by HMM based boundary refinement are refined one more time using the glottal inverse filtering based method. The AABE from the first stage is decreased by 22-23% in MOCHA-TIMIT database and 21-26% in Turkish audiovisual speech database by using the proposed boundary refinement methods together in this way.

A three stage automatic speech segmentation system is build by integrating the proposed automatic speech segmentation systems in this thesis (Figure 4.9). First stage is a HMM speech recognizer based bimodal automatic speech segmentation system, that uses several audiovisual feature vectors, the second stage is a HMM based boundary refinement system and the third stage is the glottal inverse filtering based boundary refinement system. By using the 3 stage AS system AABE of a standard HMM based AS system is reduced by 35.3% for MOCHA-TIMIT database and 41.36% for Turkish audiovisual speech database.

REFERENCES

[1] C.G. Kratzenstein, Sur la raissance de la formation des voyelles, J. Phys., Vol 21, pp. 358-380, 1782

[2] B. H. Juang and L. R. Rabiner, Technometrics, Vol. 33, No. 3 (Aug., 1991), pp. 251-272

[3] H. Dudley and T. H. Tarnoczy, The Speaking Machine of Wolfgang von Kempelen, J. Acoust. Soc. Am., Vol. 22, pp. 151-166, 1950.

[4] Klatt D. (1987) Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America, JASA vol. 82 (3), pp.737-793.

[5] K. H. Davis, R. Biddulph, and S. Balashek, Automatic Recognition of Spoken Digits, J. Acoust. Soc. Am., Vol 24, No. 6, pp. 627-642, 1952.

[6] F. Itakura and S. Saito, A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies, Electronics and Communications in Japan, Vol. 53A, pp. 36-43, 1970.

[7] F. Itakura, Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-23, pp. 57-72, Feb. 1975.

[8] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon, Speaker Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. Assp-27, pp. 336-349, Aug. 1979.

[9] L. M. Hyman, Phonology: theory and analysis. New York: Holt, Rinehart Winston, 1975. Pp. xiii+268.

[10] D. A. Burquest, and D. L. Payne, Phonological analysis: A functional approach. Dallas, TX: Summer Institute of Linguistics, 1993.

[11] P. Coxhead, Natural Language processing & Applications, Phones and Phonemes, retrieved from the Internet: http://www.cs.bham.ac.uk/ pxc/nlpa/2002/NLPA-HOPhon.pdf, last visited on 21 February 2010.

[12] M. Tatham, K. Morton, Speech Production and Perception. Basingstoke: Palgrave Macmillan, 2006.

[13] M. J. Makashay, C. W. Wightman , A. K. Syrdal, A. Conkie, Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis. In: Proc. Int. Conf. on Spoken Language Processing, Beijing, pp. 431-434, 2000.

[14] H. Kawai, T. Toda, An evaluation of automatic phone segmentation for concatenative speech synthesis. In: Acoustics, Proceedings. (ICASSP '04). IEEE International Conference on Speech, and Signal Processing, 2004.

[15] D. T. Toledano, L. A. H. Gómez, L.V. Grande, Automatic Phonetic Segmentation. In: IEEE Transactions on Speech and Audio Processing vol. 11, no. 6., 2003.

[16] F. Malfrere, O. Deroo, T. Dutoit, C. Ris, Phonetic alignment: speech synthesis-based vs. Viterbi-based. In:Speech Communication, Volume 40, Issue 4, pp. 503-515, 2002.

[17] F. Brugnara, D. Falavigna, M. Omologo, Automatic segmentation and labeling of speech based on Hidden Markov Models. In: Speech Communication, Volume 12, Issue 4, pp. 357-370, 1993.

[18] J. Matousek, D. Tihelka, J. Psutka, Automatic Segmentation for Czech Concatenative Speech Synthesis with Boundary-Specific Correction. In: Eurospeech 2003.

[19] A. Sethy, N. Shrikanth, Refined speech segmentation for concatenative speech synthesis. In: Proc. ICSLP-2002, pp.145-148, Denver, CO, Sept. 2002.

[20] L. Wang, Y. Zhao, M. Chu, F. Soong, J. Zhou, Z. Cao, contextdependent boundary model for refining boundaries segmentation of TTS units. IEICE Trans. Inform. Syst. E89-D (3), 1082-1091, 2006.

[21] A. Bonafonte, A. Nogueiras, A. Rodriguez-Garrido, Explicit segmentation of speech using Gaussian models, In ICSLP-1996, 1269-1272, 1996.

[22] K. Yeon-Jun, C. Alistair, Automatic segmentation combining an HMM-based approach and spectral boundary correction. In: ICSLP-2002, 145-148, 2002.

[23] D. Meen,T. Svendsen, J. E. Natvig, Improving Phone Label Alignment Accuracy by Utilizing Voicing Information In SPECOM 2005 Proceedings. Moscow State Linguistic University, ISBN 5-7452-0110-X, p. 683-686, 2005.

[24] S. S. Park; N. S. Kim, On Using Multiple Models for Automatic Speech Segmentation, Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.8, pp.2202-2212, Nov. 2007.

[25] S. Jarifi, D. Pastor, O. Rosec, A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis, Speech Communication, Volume 50, Issue 1, January 2008, Pages 67-80, ISSN 0167-6393, DOI: 10.1016/j.specom.2007.07.001.

[26] V. Kamakshi Prasad, T. Nagarajan, Hema A. Murthy, Automatic segmentation of continuous speech using minimum phase group delay functions, Speech Communication, Volume 42, Issues 3-4, April 2004, Pages 429-446, ISSN 0167-6393, 2004.

[27] S. Young , G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland The HTK Book (for HTK Version 3.2). Cambrige University Engineering Departmen, 2002.

[28] A. Wrench, MOCHA-TIMIT Database, 1999, http://www.cstr.ed.ac.uk/research/projects/ artic/mocha.html, last visited on 21 February 2010.

[29] Lawrence Rabiner and Bing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[30] D.G. Stork, M.E. Hennecke, Speechreading: an overview of image processing, feature extraction, sensory integration and pattern recognition techniques, Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on , vol., no., pp.XVI-XXVI, 14-16 Oct 1996.

[31] R. Campbell, B. Dodd, and D. Burnham, Eds., Hearing by Eye II. Hove, U.K.: Psychology, 1998.

[32] W. H. Sumby and Irwin Pollack, Visual Contribution to Speech Intelligibility in Noise J. Acoust. Soc. Am. 26, 212 (1954), DOI:10.1121/1.1907309

[33] H. MacGurk and J. MacDonald, Hearing lips and seeing voices, Nature, vol. 264, pp. 746-748, 1976.

[34] Q. Summerfield, Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Associates, London, pp. 3-51, 1987.

[35] D. Massaro, D. Stork, Speech recognition and sensory integration: a 240-year old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. American Scientist, 86(3):236-245, 1998.

[36] A. Q. Summerfield, A. MacLoed, M. McGrath, N.M. Brooke, Lips, teeth, and the benefits of lipreading. In A.W. Young H.D. Ellis (Eds.) Handbook of Research in Face Processing, Amsterdam: North Holland, 1989.

[37] P.M.T. Smeele, et al. Laterality in visual speech perception J. Exp. Psychol. Hum. Percept. Perform. 24, 1232-1242, 1998.

[38] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, E.Vatikiotis-Bateson, Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological Science., 15, 133-137, 2004.

[39] M.N. Kaynak, Qi Zhi, A.D. Cheok, K. Sengupta, Zhang Jian, Ko Chi Chung, Analysis of lip geometric features for audio-visual speech recognition. In: Systems, Man and Cybernetics, Part A, IEEE Transactions on, Vol.34, Iss.4, pp.: 564- 570, 2004.

[40] B.P. Yuhas, M.H. Goldstein, Jr., T.J. Sejnowski, R.E. Jenkins, Neural network models of sensory integration for improved vowel recognition. In: Proceedings of the IEEE, Vol.78, Iss.10, pp.:1658-1668, 1990.

[41] T. Chen, R.R. Rao, 1998. Audio-visual integration in multimodal communication. In: Proceedings of the IEEE, Vol.86, Iss.5, pp.:837-852, 1998.

[42] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, J. Zhou, Audio-visual speech recognition, Final Workshop 2000 Report , Center for Language and SpeechProcessing, The Johns Hopkins University, Baltimore, 2000 (http://www.clsp.jhu.edu/ws2000/finalreports/avsr/).

[43] A. Katsamanis, G. Papandreou and P. Maragos, Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation, IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 3, pp. 411-422, Mar. 2009.

[44] J. Wouters and M. Macon, Perceptual evaluation of distance measures for concatenative speech synthesis, in Proc. ICSLP, vol. 6, Sydney, Australia, pp. 2747-2750, 1998.

[45] F. Itakura, Line spectrum representation of linear predictor coefficientsof speech signals, J. Acoust. Soc. Amer., vol. 57, p. S35(A), 1975.

[46] J. Vepa, S. King, Kalman-filter based join cost for unit-selection speech synthesis, In EUROSPEECH-2003, 293-296, 2003.

[47] O. Engwall, A revisit to the application of MRI to the analysis of speech production. Testing our assumptions. In 6th Intl Seminar on Speech Production (pp. 43-48). Sydney, 2003.

[48] K.G. Munhall, E. Vatikiotis-Bateson, Y. Tohkura, X-ray film database for speech research. J. Acoust. Soc. Am. 98, 1222-1224, 1995.

[49] G. Potamianos, C. Neti, G. Gravier, A. Garg, Senior, A.W., Recent advances in the automatic recognition of audiovisual speech, Proceedings of the IEEE , vol.91, no.9, pp. 1306-1326, Sept. 2003.

[50] M.W. Mak, W.G. Allen, Lip-motion analysis for speech segmentation in noise. Speech Comm. 14 (3), 279-296, 1994.

[51] S. Fagel, C. Clemens, An articulation model for audiovisual speech synthesis-Determination, adjustment, evaluation. Speech Communication 44, 141- 154, 2004.

[52] E.D. Petajan, Automatic lipreading to enhance speech recognition. Doct. Thesis, Univ. of Illinois, 1984.

[53] A. Adjoudani, C. Benont, On the integration of auditory and visual parameters in an HMM-based ASR. In D.G. Stork & M.E. Hennecke (Eds.) Speechreading by Man and Machine: Models, Systems and Applications (pp. 461-472). NATO ASI Series, Springer, 1996.

[54] P. Duchnowski, U. Meier, A. Waibel, See me, hear me: integrating automatic speech recognition and lip-reading. Proc. Int. Conf. on Spoken Language Processing, pp. 547-550. Yokohama, Japan, 1994.

[55] P. Teissier, J. Robert-Ribes, J. L. Schwartz, A. Guirin-Dugui, Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans. Speech and Audio Processing, 7, 629-642, 1999.

[56] D.G. Stork, G. Wolff, E. Levine, Neural network lipreading system for improved speech recognition. Proc. IJCNN-92, Baltimore MD, vol. 2, pp. 285-295, 1992.

[57] J.L. Schwartz, J. Robert-Ribes, P. Escudier, Ten years after Summerfield ... a taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing (pp. 85-108). Hove (UK) : Psychology Press, 1998.

[58] D. O'Shaughnessy, Speech communication: Human and ma-chine (USA: Wiley-IEEE Press, 2001).

[59] Quatieri, Discrete-time Speech Signal Processing, Prentice-Hall, 2002

[60] S. L. Hamlet, H. T. Bunnell, B. Struntz, Articulatory asymmetries. Journal of Acoustic Society of America, 79,(4), 1164-1169, 1986.

[61] A. Karadüz, Yazı ve Yazım kavramlarının dilin anlam ve ses öğeleriyle ilişkisi - The relation between the concepts of literary and writing with meaning and phonic components of language, TURKISH STUDIES - International Periodical For The Languages, Literature and History of Turkish or Turkic, Volume 3/6 Fall 2008, www.turkishstudies.net, p. 422-448, 2008.

[62] J.G. Fiscus, A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), Automatic Speech Recognition and Understanding. Proceedings., 1997 IEEE Workshop on , vol., no., pp.347-354, 14-17 Dec 1997.

[63] P. Cosi, D. Falavigna and M. Omologo, A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies, Proceedings of European Conference on Speech Communication and Technology, pp. 693-696, 1991.

[64] SAMPA - computer readable phonetic alphabet, www.phon.ucl.ac.uk/home/sampa, last visited on 21 February 2010.

[65] MPEG-2 specification, ITU H.262 2000 E pg. 44.

[66] A. V. Barbosa, H. C. Yehia, E. Vatikiotis-Bateson, MATLAB toolbox for audiovisual speech processing, In AVSP-2007, paper P38, 2007.

[67] M. Bayati, D. Shah, M. Sharma, A Simpler Max-Product Maximum Weight Matching Algorithm and the Auction Algorithm, Information Theory, 2006 IEEE International Symposium on , vol., no., pp.557-561, 9-14 July 2006

[68] Jintao Jiang, Abeer Alwan, Patricia A. Keating, Edward T. Auer Jr., and Lynne E. Bernstein, On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics, EURASIP Journal on Applied Signal Processing, vol. 2002, no. 11, pp. 1174-1188, 2002.

[69] J. Zhang, Y. Yan and M. Lades, Face Recognition: Eigenface, Elastic Matching, and Neural Nets, Proc. IEEE, vol. 85, no. 9, pp. 1,423-1,435, 1997.

[70] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, Classifying facial actions, IEEE Trans. Pattern Anal. Machine Intell., vol. 21, pp. 974-989, Oct. 1999.

[71] D.L. Hall, J. Llinas, An introduction to multisensor data fusion, Proceedings of the IEEE , vol.85, no.1, pp.6-23, Jan 1997

[72] E. Akdemir, T. Çiloğlu, The use of articulator motion information in automatic speech segmentation Speech Communication, 50 (7), pp. 594-604, 2008.

[73] Walker J., Murphy P. 2007 A review of glottal waveform analysis, Springer Berlin/Heidelberg

[74] Stylianou, Y.: Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. Speech Audio Process., 9(2001) 21-29

[75] T. F. Quatieri, R. J. McAulay, Shape invariant time-scale and pitch modification of speech. IEEE Trans. Signal Process., 40 497-510, 1992.

[76] J. Kittler, M. Hatef, R.P.W. Duin, Matas, J., On combining classifiers, Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.20, no.3, pp.226-239, Mar 1998

[77] J. Kominek and A. W. Black, A family-of-models approach to HMM-based segmentation for unit selection speech synthesis, in Proc. ICSLP, Jeju, Korea, 2004.

[78] E. Akdemir, T. Çiloglu, A HMM based system for fine tuning in automatic speech segmentation, T.,Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th 9-11 April 2009 Page(s): 381-384

[79] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm 1. Roy. Stat. Soc., vol. 39, no. 1, pp. 1-38, 1977.

[80] 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, SeIISyst. Tech., vol. 62, no. 4, pp. 1035-1074, Apr. 1983.

[81] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, IEEE Trans. Informat. Theory, vol. IT-13, pp. 260-269, Apr. 1967. G. D. Forney, The Viterbi algorithm, Proc. IEEE vol.61, pp. 268-278, Mar. 1973.

[82] P. Alku, Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering, Speech Communication, Volume 11, Issues 2-3, Eurospeech '91, June 1992, Pages 109-118, ISSN 0167-6393, 1992.

[83] M. Airas, TKK Aparat: An environment for voice inverse filtering and parameterization, Logoped Phoniatr Vocol. 2008;33(1):49-64.

[84] S. Kullback, R. A. Leibler, On Information and Sufficiency. Annals of Mathematical Statistics 22 (1) 79-86, 1951.

APPENDIX A

PHONEMES in MOCHA-TIMIT DATABASE

Table A.1: Phones in MOCHA-TIMIT Database

Phone label	Word	Phonetic transcription	Phone label	Word	Phonetic transcription
@	was	w @ z	m	man	m a n
@@	thirty	th @@ t iy	n	new	n y uu
a	exam	i g z a m	ng	swing	s w i ng
aa	hard	h aa d	o	on	o n
ai	bright	b r ai t	oi	voyage	v oi I jh
b	bright	b r ai t	oo	wall	w oo l
ch	Charlie	ch aa l i	ou	yellow	y e l ou
d	good	g u d	ow	allow	a l ow
dh	those	dh ou z	p	petrol	p e t r @ l
e	every	e v r iy	r	result	r i z uh l t
ei	fails	f ei l z	s	safe	s ei f
eir	rare	r eir	sh	musician	m y uu z i sh @ n
f	fails	f ei l z	t	votes	v ou t s
g	gallon	g a l @ n	th	walth	w e l th
h	healty	h e l th iy	u	good	g u d
i	this	dh i s	uh	money	m uh niy
i@	near	n i@	uu	formula	f oo m y uu l @
ii	museum	m y uu z ii @ m	v	vodka	v o d k @
iy	lily	l i l iy	w	will	w I l
jh	orange	o r i n jh	y	you	y uu
k	scholar	s k o l @	z	is	i z
l	lay	l ei	zh	pleasure	p l e zh @ r

APPENDIX B

PHONEMES in SAMPA for TURKISH

Table B.1: Phones in Audiovisual speech database in SAMPA notation.

Phone label	Word	Phonetic transcription	Phone label	Word	Phonetic transcription
a	ara	a r a	m	masal	m a s a L
ax	arıza	ax r I z a	n	yarın	j a r I n
b	bar	b a r	N	ankara	a N k a r a
c	kitap	c i t a p	o	olay	o L a j
d	deniz	d e n i z	ox	koordinasyon	k ox r d i n a s j o n
dZ	cam	dZ a m	O	önem	O n e m
e	emek	e m e c	Ox	-	-
ex	tesis	t ex s i s	p	pas	p a s
f	faz	f a z	r	ray	r a j
g	gam	g a m	s	ses	s e s
gj	genç	gj e n tS	S	şan	S a n
G	sağır	s a G I r	t	tam	t a m
h	ham	h a m	tS	çiçek	tS i tS e c
I	sır	s I r	u	futbol	f u t b o L
Ix	-	-	ux	cumhuriyet	dZ u m h ux r i j e t
i	ilik	i l i k	v	vergi	v e r g i
ix	dakika	d a c ix k a	w	tavuk	t a w u k
j	ayar	a j a r	y	ürün	y r y n
k	akıl	a k I L	yx	güya	gj ux j a
l	lale	l ax l e	z	zar	z a r
L	olay	o L a j	Z	masaj	m a s a Z

VITA

PERSONAL INFORMATION

Surname, Name: Akdemir, Eren

Nationality: Turkish (TC)

Date and Place of Birth: 14 Oct 1980 , Zonguldak

Marital Status: Single

Phone: +90 532 396 50 30

email: eakdemir@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
BS	METU Electrical and Electronics Engineering	2002
High School	Özel Arı Fen Lisesi, Ankara	1998

WORK EXPERIENCE

Year	Institution
2002-2008	METU Electrical and Electronics Engineering, Research and Teaching Assistant
2007-2010	The Scientific and Technological Research Council of Turkey, Researcher in project 107E101

Publications:

- Akdemir E., Çiloğlu T., The use of articulator motion information in automatic speech segmentation, Speech Communication, 50 (7), pp. 594-604, 2008.

- Akdemir, E.; Çiloğlu, T., "A HMM based system for fine tuning in automatic speech segmentation," Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th , vol., no., pp.381-384, 9-11 April 2009.

- Akdemir, E.; Çiloğlu, T., "Using visual information in automatic speech segmentation," Signal Processing, Communication and Applications Conference, 2008. SIU 2008. IEEE 16th , vol., no., pp.1-4, 20-22 April 2008.

- Akdemir, E.; Çiloğlu, T., "The Use of Lip Motion and a New Spectral Criterion for Speech Segmentation ", ECESS Advances in Speech Technology 2007, 14th International workshop, 2007.

- Akdemir, E.; Çiloğlu, T., "Incorporation of Visual Cues In Speech Segmentation", ECESS Advances in Speech Technology 2006, 13th International workshop, 2006.

Submitted:

- Akdemir E., Çiloğlu T., Bimodal Automatic Speech Segmentation Based on Audio and Visual Information Fusion, Submitted to: IEEE Transactions on Audio, Speech, and Language Processing

To be submitted:

- Akdemir E., Çiloğlu T., A new HMM topology for boundary refinement in automatic speech segmentation, Electronics Letters, IEEE

- Akdemir E., Çiloğlu T., A new measure based on inverse filtering of speech for boundary refinement in automatic speech segmentation, Electronics Letters, IEEE

- Akdemir E., Özbek Y., Çiloğlu T., Time Domain and Frequency Characteristics of Turkish Plosives, The Journal of the Acoustical Society of America
- Özbek Y., Akdemir E., Çiloğlu T., Acoustic characteristics of Turkish fricatives, The Journal of the Acoustical Society of America