MODELING PHONEME DURATIONS AND FUNDAMENTAL FREQUENCY CONTOURS IN TURKISH SPEECH

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

BY

ÖZLEM ÖZTÜRK

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

OCTOBER 2005

Approval of the Graduate School of Natural and Applied Sciences

Prof. Dr. Canan ÖZGEN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. İsmet ERKMEN
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Tolga ÇİLOĞLU
Supervisor

Examining Committee Members

Prof. Dr. Mübeccel DEMİREKLER (METU, EE)

Assoc. Prof. Dr. Tolga ÇİLOĞLU (METU, EE)

Prof. Dr. Uğur HALICI (METU, EE)

Dr. Meltem TURHAN YÖNDEM (METU, CENG)

Asst. Prof. Dr. H. Gökhan İLK (Ankara University, EE)

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.

Özlem ÖZTÜRK

ABSTRACT

MODELING PHONEME DURATIONS AND FUNDAMENTAL FREQUENCY CONTOURS IN TURKISH SPEECH

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

October 2005, 202 pages

The term prosody refers to characteristics of speech such as intonation, timing, loudness, and other acoustical properties imposed by physical, intentional and emotional state of the speaker. Phone durations and fundamental frequency contours are considered as two of the most prominent aspects of prosody. Modeling phone durations and fundamental frequency contours in Turkish speech are studied in this thesis.

Various methods exist for building prosody models. State-of-the-art is dominated by *corpus-based* methods. This study introduces corpus-based approaches using classification and regression trees to discover the relationships between prosodic attributes and phone durations or fundamental frequency contours. In this context, a speech corpus, designed to have specific phonetic and prosodic content has been recorded and annotated.

A set of prosodic attributes are compiled. The elements of the set are determined based on linguistic studies and literature surveys. The relevances of prosodic attributes are investigated by statistical measures such as mutual information and information gain.

Fundamental frequency contour and phone duration modeling are handled as independent problems. Phone durations are predicted by using regression trees where the

set of prosodic attributes is formed by forward selection. Quantization of phone durations is studied to improve prediction quality. A two-stage duration prediction process is proposed for handling specific ranges of phone duration values. Scaling and shifting of predicted durations are proposed to minimize mean squared error.

Fundamental frequency contour modeling is studied under two different frameworks. One of them generates a codebook of syllable-fundamental-frequency-contours by vector quantization. The codewords are used to predict sentence fundamental frequency contours. Pitch accent prediction by two different clustering of codewords into accented and not-accented subsets is also considered in this framework. Based on the experience, the other approach is initiated. An algorithm has been developed to identify syllables having perceptual prominence or pitch accents. The slope of fundamental frequency contours are then predicted for the syllables identified as accented. Pitch contours of sentences are predicted using the duration information and estimated slope values.

Performance of the phone duration and fundamental frequency contour models are evaluated quantitatively using statistical measures such as mean absolute error, root mean squared error, correlation and by kappa coefficients, and by correct classification rate in case of discrete symbol prediction.

Keywords: Duration modeling, fundamental frequency contour modeling, speech database, prosody, intonation, classification and regression trees.

ÖZ

TÜRKÇE KONUŞMADA SESBİRİM SÜRELERİNİN VE TEMEL FREKANS EĞRİLERİNİN MODELLENMESİ

Öztürk, Özlem
Doktora, Elektrik Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. Tolga Çiloğlu

Ekim 2005, 202 sayfa

Ezgi, konuşmanın süre, vurgu, genlik ve diğer akustik özelliklerinden oluşan, konuşmacının fiziksel ve duygusal durumuna bağlı olarak değişiklikler gösteren nitelikleridir. Sesbirim süreleri ve perde eğrileri, ezginin en önemli bileşenlerinden ikisi olarak kabul edilmektedir. Bu tezde, sesbirim süreleri ve perde eğrileri Türkçe konuşma için modellenmiştir.

Birçok ezgi modelleme yöntemi bulunmaktadır. Yapılan son çalışmalarda çoğunlukla derlem-tabanlı yöntemler kullanılmaktadır. Bu çalışma, ezgi öznitelikleri ile perde eğrisi ve sesbirim süreleri arasındaki ilişkiyi meydana çıkarmak için *sınıflandırma ve bağlanım* (classification and regression) ağaçları kullanarak derlem-tabanlı çalışmaları içermektedir. Bu çerçevede, istenilen ezgisel ve sesbirimsel içerikte bir derlem kaydedilmiş ve işaretlenmiştir.

Dilbilimsel çalışmalar ve yazın araştırmaları doğrultusunda ezgi öznitelikleri derlenmiştir. Karşılıklı bilgi (mutual information) ve bilgi kazancı (information gain) gibi istatistiksel ölçütler kullanılarak, ezgi özniteliklerinin ezgi ile olan ilgileri belirlenmiştir.

Perde eğrisi ve sesbirim süresi modelleme çalışmaları bağımsız problemler olarak ele alınmıştır. Sesbirim süreleri bağlanım ağaçları kullanılarak ileri seçme (forward selection)

yöntemi ile oluşturulmuş ezgi özniteliklerinden öngörülmüştür. Sesbirim süreleri başarımı artırmak için nicemlenmiştir. Süre aralıklarının ayrı ayrı ele alınabilmesi için iki aşamalı süre modelleme yöntemi öne sürülmüştür. Ortalama karesel hatanın düşürülmesi için öngörülen süreler üzerinde ölçekleme ve öteleme yapılmıştır.

Perde eğrisi modelleme iki ayrı çatı altında incelenmiştir. Birinde hece perde eğrileri için vektör nicemleme kullanılarak kod defteri oluşturulmuştur. Hece kodları tümce perdesi öngörümünde kullanılmıştır. Ayrıca, iki farklı sınıflandırma yöntemi kullanılarak hece kodlarının vurgulu ve vurgusuz alt kümeleri belirlenmiş; bu bilgi kullanılarak perde vurgusu öngörülmüştür. Bu bölümde elde edilen deneyimler diğer yaklaşım için başlangıç olmuştur. Algısal önemi olan ya da perde vurgusu alan hecelerin belirlenmesi için bir algoritma geliştirilmiştir. Perde vurgusu alan hecelerin perde eğimleri öngörülmüştür. Süre ve eğim öngörüleri kullanılarak tümce perdeleri elde edilmiştir.

Sesbirim süreleri ve perde eğrisi modelleme başarımları nicel olarak değerlendirilmiştir. Sayısal değerlendirmeler mutlak hata, etkin hata ve ilinti gibi istatistiksel ölçütlerle gerçekleştirilmiştir. Ayrık işaret kestirimlerinde ise kappa katsayıları ve doğru kestirim oranları kullanılmıştır.

Anahtar Kelimeler: Ezgi, entonasyon, süre modelleme, perde eğrisi modelleme, derlem, sınıflandırma ve bağlanım ağaçları.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitudes to my thesis supervisor Assoc. Prof. Dr. Tolga Çiloğlu. He spent endless hours discussing and reviewing various aspects of my work. He even spent more time with me than his family. I owe much to his family for their patience. Special thanks go to my thesis comittee members, Prof. Dr. Mübeccel Demirekler, Assoc. Prof. Dr. Cem Bozşahin, and Dr. Meltem Turhan Yöndem for their advises on the subject. I am very grateful to Asst. Prof. Dr. Bilge Say who has attended once my thesis comitte and introduced me METU-Sabancı Turkish TREEBANK project.

During my course of study, I was employed at TCTS Lab at Belgium for two months. I am grateful to Prof. Dr. Thierry Dutoit and Dr. Barış Bozkurt for giving me the opportunity to work there.

I would also like to express my sincere gratitudes to Prof. Dr. Hiroya Fujisaki. We have never met before but he has not hesitated a moment to send me his copies of former papers.

Prof. Dr. İclal Ergenç, Prof. Dr. Güneş Müftüoğlu, and Assoc. Prof. Dr. Engin Sezer deserve my gratitudes for their precious help in Turkish prosody.

Thanks to the faculty members and administrative staff in Electrical and Electronics Department of Dokuz Eylül University for their endless support and patience. Special thanks go to my office-mate Mehmet Ali Yarım for his continuous support.

I am also grateful to Yücel, Eren and Turgay for their assistance in manual correction of the speech corpus and to Dinç for his helps in using HTK toolkit. Banu was the first who introduced me the incredible computer program PRAAT. Özgül, you were the one I owe much for your intellectual and mental support. And, Erdinç, you were an excellent friend, I will never forget you.

I owe a great deal to my parents, Menekşe and Muzaffer Özkan, to my sister Özge and brother Mert, and lastly to my soul mate, my husband Uğur, for all they have done for me over the years. It is to them that this thesis is dedicated.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

The term prosody refers to characteristics of speech such as intonation, timing, stress, loudness, and other acoustical properties imposed by articulatory, emotional, mental, and intentional states of the speaker. The most prominent components of prosody are considered as phoneme duration and pitch contour. This study started with the aim of predicting phoneme durations and pitch contour of a Turkish sentence given its written form. The resultant phoneme durations and pitch contour are expected to resemble natural speech. From a practical point of view, such information are needed in Text-to-Speech (TTS) synthesis systems. Text-to-Speech synthesis is used in many areas such as information retrieval systems; language education, and reading machines for visually impaired [Fortinea 1999; Kenney 1998; Lemmetty 1999].Without appropriate prosody models, synthetic speech is perceived as monotonous, boring and less intelligible [Ross 1995].“It has been shown that poor prosody is worse than no prosody (Benoit 1990)” [Monaghan 1997].

TTS systems can be divided into two major subsystems (**Figure 1-1**) Natural language processing (NLP) subsystem and 2) Signal processing subsystem The NLP module performs the task of converting input text into a linguistic representation including phonetic and prosodic information. The DSP module generates output speech waveform using information provided by NLP subsystem [Dutoit 1997; Lemmetty 1999].

The NLP subsystem can be further divided into two parts (**Figure 1-2**): *Text-to-phonetic conversion* module is responsible for the transformation of text into corresponding phonetic units that specify sounds to be produced. *Text-to-prosodic parameter conversion* module performs the generation of prosodic parameters, fundamental frequency, duration and intensity in general, which specify how these sounds are to be produced [Dutoit 1997; Huang *et. al.* 1997; Lemmetty 1999].

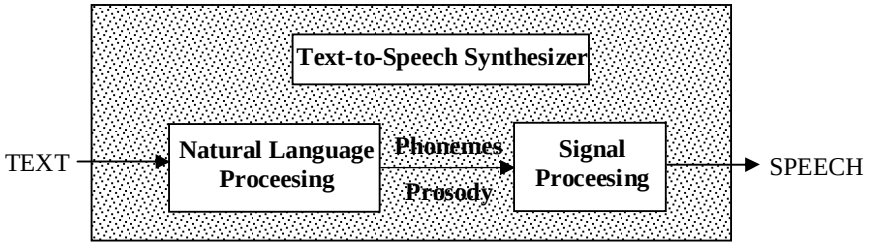

Figure 1-1: Functional block diagram of a general TTS synthesizer

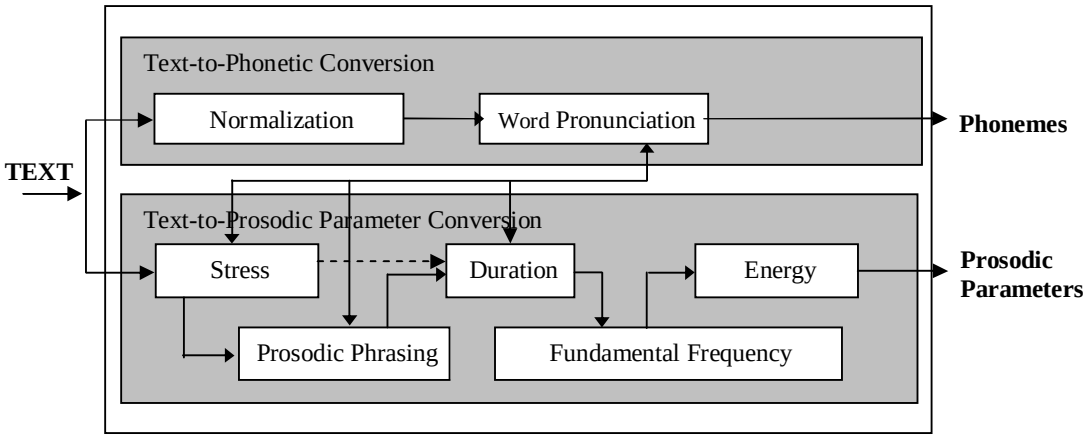

Figure 1-2: Block Diagram of NLP subsystem.

The fundamental problem encountered in speech synthesis systems is the poor generation of prosody for unrestricted text [Sun 2002; Huang *et. al*. 1997]. Every spoken language has its own prosody. However, the prosody of an utterance is not unique. A sentence may have a number of perceptually acceptable but significantly different (in a mathematical sense) prosody patterns while carrying the same semantic information. Any spoken utterance is produced with a particular sound pattern shaped by its prosody. Prosody is a means of conveying information. It plays an important role in human speech communication. In natural speech, prosody of an utterance may depend on semantics, context, syntax of the statement, intended audience, and emotional or physical state of the speaker. Today, prosody modeling constitutes one of the main arenas of speech research [Agüero et. al 2004; Fujisaki and Nagashima 1969; Sun 2002; Taylor 1992; Taylor and Isard 1992].

The elements of prosody commonly derived from speech are:

- Intonation; variations in pitch which gives speech its melody.
- Timing; variations in phoneme duration, speaking rate (tempo) and pausing.
- Loudness pattern.
- Stress, perceived in terms of pitch, loudness and duration

All these ingredients are present in varying amounts in every spoken utterance. Specific mixtures of these elements orient the listener in interpreting the utterance. When we speak we do not only produce a sequence of speech sounds but also impose stress and intonation patterns to convey a meaning. For example, in Turkish, words having identical orthographies can bear different meanings that can only be differentiated by their semantic context and prosody: the only difference between the noun *yazma'* (a kind of scarf) and the verb *yaz'ma* (do not write) is that of prosodic variation due to lexical stress placement.

Recently, a number of speech studies on Turkish have come out. Automatic Speech Recognition [Bayer 2005; Büyük *et. al.* 2005; Çarkı *et. al.* 2000; Çilingir 2003; Çömez 2003; Orkan 2005; Salor *et. al.* 2002a, 2002b; Yapanel 2000; Yılmaz 1999], Language Modeling [Bayer 2005; Çiloğlu 2004, Çiloğlu *et. al.* 2004; Şahin, 2003], Voice Transformation [Salor 2004; Arslan 1997, 1999; Türk and Arslan 2002, 2004], and Text-to-Speech [Abdullahmeşe 1998; Fidan 2002; Oskay 2000, 2001; Özge, 2003; Vural and Oflazer 2004] are some of them. There is no study which covers a comprehensive prosody modeling in Turkish. Existing studies either handle a part of the modeling process or they do not rely on detailed linguistic analysis [Abdullahmeşe 1998; Fidan 2002; Oskay 2000, 2001; Özge, 2003].

Various methods exist for building prosody models [Agüero 2004; Batusek 2002; Black and Hunt 1996; Chen *et. al.* 1996; Dusterhoff 2003; Pierrehumbert 2000; Lee and Oh 1999a, 1999b, 2001; Mixdorff 2000, 2001; Riedi 1998; Sakurai *et. al.* 2003; Shih and Kochanski 2002; Sun 2002; Taylor 1992, 1995, 2002; Vegnaduzzo 2003]. Those used at the initial stages of prosody modeling are generally known as *rule-based approaches*. Rule-based heuristic systems such as Klatt's duration modeling system [Klatt, 1987] combine linguistic expert knowledge and manual analysis of quite limited amount of text and their recordings. They are often unsatisfactory and case-dependent. Hence, they exhibit less flexibility against, for example, personality and speaking style. State-of-the-

art is dominated by *corpus-based approaches*. They have appeared due to the increasing computational power and availability of large corpora. Corpus-based (data-driven) modeling utilizes large text and speech corpora to discover rules as a function of prosodic attributes. Prosodic attributes, defined on text, are linguistic features (phonetic context, number of words in sentence, number of syllables in word, etc.) that are considered to be affecting prosody. Corpus-based modeling involves machine learning techniques such as Artificial Neural Networks (ANN), and Classification and Regression Trees (CART) to reveal the relation between prosody and prosodic attributes. They can be adapted to new speaking styles by providing new data. This study concentrates on corpus-based modeling and uses machine learning techniques to develop models of phoneme duration and pitch contour for Turkish.

Each modeling method mentioned above has its advantages and disadvantages. Neural networks are very popular machine learning algorithms. They are known for their ability to generalize according to the similarity of their inputs. With sufficient data, neural networks can approximate any nonlinear function. However, the trained model is not human readable which is a disadvantage if one needs to understand the conceptual relationship between inputs and outputs [Campbell 2000; Chen *et. al.* 1996; Taylor 1995; Witten and Frank 1999].

A decision tree is a predictive model that can be viewed as a tree. It is a popular nonparametric supervised learning method. In decision trees, each branch of the tree represents a choice and the leaves of the tree represent decisions. Decision trees provide interpretability. They can also be applied to any data and requires less parameter tuning [Agüero *et. al.* 2004; Black and Taylor 1997; Breiman *et. al.* 1984; Batusek 2002; Witten and Frank 1999]. Within the framework of this dissertation, decision tree learning is incorporated for phoneme duration and pitch contour modeling.

State-of-the-art TTS systems use pre-recorded acoustic units, such as phones, diphones, or polyphones, to perform synthesis [Bulyko and Ostendorf 2002; Chen *et. al.* 1996; Violaro and Böeffard 1998]. To improve the naturalness of synthetic speech, continuous speech databases composed of multiple representations of these units are developed. In general, language can be considered as the set of all possible combinations of these units. However, it is not practical to record all combinations.

Success of prosody modeling is mainly related to chosen corpus used for training. If the speech corpus is rich enough to represent the prosodic and contextual variety of the language, higher performance can be achieved in modeling [Iida and Campbell 2001; Campione and Veronis 1998a]. Hence, speech corpora design is one of the key issues to improve the naturalness and intelligibility of synthetic speech.

A speech database can be built randomly or by means of optimizing the units acoustically or with respect to their textual properties. Random selections may not be adequate to provide sufficient variability for prosody research. To develop appropriate prosody models, we also need a speech database of sufficient phonetic and prosodic coverage. Phonetic coverage can be obtained by supplying sufficient representatives of each unit [Iida and Campbell 2001]. Prosodic coverage is achieved by considering various types of syntactic constituents with sufficient representatives. Within the scope of this thesis, a phonetically and prosodically rich speech database is developed.

General assumption for intonation modeling is that it can be successfully handled only by fundamental frequency, thus, the ultimate goal is to develop a model to generate fundamental frequency contours. Various intonation models have been proposed in the past. They are contrasted by different viewpoints [Monaghan 1992; Veronis *et. al.* 1998; Taylor and Isard 1992]: the systems may be phonological or phonetic; pitch contours can be produced by parametric or nonparametric methods; or the systems may use level tones or pitch movements. These viewpoints can be summarized in a more compact form as shown in **Figure 1-3**.

Phonological models employ a set of discrete symbols to represent the pitch contour [Dusterhoff 2000; Frid 2001; Jilka *et. al.* 1999; Taylor 1992; Veronis *et. al. 1998*]. The most influential one is Pierrehumbert's model later evolved into a standard (Tones and Breaks Indices, ToBI) for transcribing American English. As stated in Silverman *et. al.* (1992), ToBI is the most widely used system for the symbolic transcription of intonation at present. It provides a four level transcription system to the researchers, which obeys the general outline proposed by Beckman and Pierrehumbert [Pierrehumbert 2000]. In Beckman and Pierrehumbert, six different pitch accents (H*, L*, L+H*, L*+H, H+L*, H*+L) and two levels of intonational phrasing (intermediate and full intonational phrase) were proposed [Pierrehumbert 2000]. Pitch accents are mainly aligned with accented *syllables*. A boundary tone is associated to each intonational phrase boundary. The symbol L- (H-) describes a low (high) tone at an intermediate phrase boundary. The

symbols L-L%, L-H%, H-L% and H-H% are used to represent full intonational phrase boundaries.

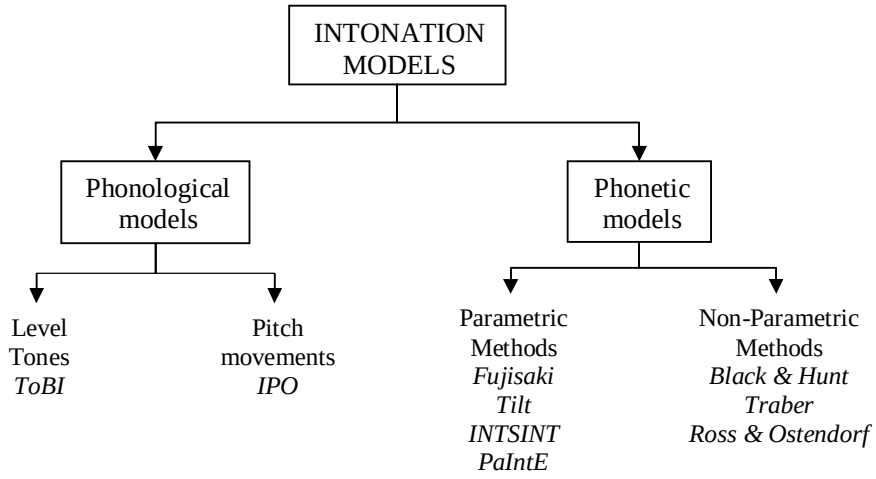

Figure 1-3: Classification of intonation models.

Another example of phonological models is the Instituut voor Perceptie Onderzoek's (IPO) perceptual model which relies on identifying perceptually relevant pitch movements and approximating them with straight lines. The main point of the approach is to simplify the F0 curve and preserve the same melodical impression to the listener [Monaghan 1992; *Veronis et. al. 1998*].

Parametric models that belong to the broader class of phonetic models use a set of continuous parameters to describe intonation patterns [Campione and Veronis 1998a, 1998b, 1988c; Hirst *et. al.* 1994; Syrdal *et. al.* 1998; Möhler 1998, 1999; Conkie and Möhler 1998]. A well known parametric model is the Fujisaki's superpositional model. The actual F0 contour is obtained by the superposition of baseline F0, phrase and accent components on a logarithmic scale. A second-order, critically damped linear filter in response to an impulse function called *phrase command* generates the phrase component. Accent component is generated by another second-order, critically damped linear filter in response to a step function called *accent command*. Basically, filters used in accent and phrase component generation differ in the effective length of their impulse responses [Fujisaki and Hirose 1984; Fujisaki and Nagashima 1969; Fujisaki 2003; Mixdorff 2000, 2001; Mixdorff and Jokish 2001; Sakurai *et. al*. 2003].

The Tilt intonation modeling proposed by Taylor can be considered both as phonological and phonetic because continuous tilt parameters are computed only at event locations and non-event parts of the pitch contour are generated by linear interpolation. Pitch accents and boundary tones are defined as events. Events have rise-fall patterns. Each event is represented by three *tilt* parameters: duration, amplitude and tilt [Taylor 1992; Taylor and Isard 1992; Taylor 2000]. Duration is the sum of the rise and fall durations. Amplitude is the sum of the magnitudes of the rise and fall amplitudes. The tilt parameter is a dimensionless number which expresses the overall shape of the event [Taylor 2000, Taylor 1998].

Nonparametric approaches use F0 values themselves. Samples from the pitch contour are taken to develop intonation models. Examples of nonparametric methods are rare. Black and Hunt used a linear regression based method to predict F0 target values for the start, mid-vowel, and end of every syllable [Black and Hunt, 1996]. In his approach, Traber (1991, 1992) utilized neural networks to identify the regular relations among German sentences. Traber predicted eight F0 values per syllable by recurrent neural networks [Keller and Werner, 1997].

Main trend in intonation modeling studies is towards the utilization of intermediate representation such as ToBI [The Ohio State University Department of Linguistics 1999], tilt, etc. described above [Campione and Veronis 1998; Conkie and Möhler 1998; Möhler 1998; 1999; Pierrehumbert 1983, 2000; Ross 1995; Taylor 1992, 1995, 1998, 2000; Sakurai *et. al.* 2003]. A great deal of the studies involve labeling of pitch accents and intonational phrases introduced by Pierrehumbert [Black and Hunt 1996; Jilka *et. al.*, 1999; Pierrehumbert 2000; Taylor 2000; Sun 2002a, 2002b]. Pitch contours are annotated with respect to those pitch accents and boundary tones by expert labelers considering language specific constraints [Bulyko and Ostendorf 2002]. Phonetic transcription of the speech signals is also provided. Phonetic transcriptions together with abstract labels for the pitch contours constitute prosodically labeled speech databases.

Boston University Radio Speech Corpus, speaker F2B is widely used among researchers studying English intonation [Clark 2003; Dusterhoff *et. al.* 1999; Jilka *et. al.* 1999; Ross 1995; Sun 2002a; Taylor 1998]. The database consists of about 40 minutes of speech read aloud by a female professional announcer. It is also labeled using ToBI transcription [The Ohio State University Department of Linguistics 1999] system. The total number of syllables in the database is 14377. The database is also labeled with

phoneme, syllable, and word boundaries, part-of-speech tags and includes lexical stress markings. It is also labeled with intonation labels based on the Tilt intonation model. [Dusterhoff *et. al.* 1999]

For pitch contour modeling in Turkish, we do not have a prosodically labeled speech database. Consequently, we do not have labels that identify accent status of the syllables in the database. Besides, there is no concrete definition of pitch accent for Turkish as for other languages such as English, German, or etc. The only source is the pitch contours of the sentences. Hence, for pitch contour modeling in Turkish, labels are derived from the pitch contours themselves.

Two methods are proposed for pitch contour modeling in Turkish. One method involves a nonparametric approach whereas the other can be considered as a phonological approach; both incorporate syllable units. Yet, both of the proposals yield a mathematical accent definition.Using proposed methods, syllables are associated to “pitch accents”. Pitch accents are associated to syllables having sudden pitch excursions. This choice of pitch accent assignment is motivated by perceptual listening tests as a result of which prominence is decided to be perceived on sharp rises.

Decision trees are used to map the relation between intonational (prosodic) attributes and accent status of the syllables. Performance of accent classification is evaluated by correct/incorrect classification rates, kappa coefficient and confusion matrix. For accented syllables, regression trees are incorporated to predict the gradient of the syllable pitch contours. The performance of gradient estimation is evaluated by objective measures such as correlation coefficient, mean absolute error, and root mean squared error. Syllable pitch contours are reconstructed using syllable duration information and gradient estimates. Overall pitch contour is reconstructed by concatenating individual syllable pitch contours.

Timing or duration plays as much important role as intonation in the encoding/decoding of speech by the speaker/listener. Duration can be defined as the time taken to utter an acoustic unit such as phoneme, syllable, etc. Duration modeling studies mainly concentrate on phoneme duration [Batusek 2002; Cordoba *et. al.* 1999; Cordoba *et. al.* 2002; Febrer *et. al.* 1998; Klatt 1987; Krishna *et. al.* 2004; Krishna and Murthy 2004; Lee and Oh 1999a, 1999b, 2001; Möbius and van Santen 1996; Riedi 1998; Venditti and van Santen 1998] however there are studies also on syllable duration

[Campbell 2000; Chen *et. al.* 2003; Lee *et. al.* 1989; Sreenivasa and Yegnanarayana 2004]. Durational patterns are part of the prosody and contain important cues for understanding the spoken text [Riedi 1998]. As stated by Campbell, variations in duration provide assistance for the listener to extract the meaning [Campbell 2000].

As a rule-based approach, Klatt used the notion of intrinsic duration introduced by Peterson and Lehiste (1960) [Campbell 2000; Klatt 1987]. Intrinsic duration is the average duration of the syllable nucleus. His model assumes that each phonetic segment type has an inherent duration that can be modified by a set of rules, but phonemes cannot be compressed shorter than a certain minimum duration [Klatt 1987]. Riley (1990, 1992) used a 1500 hand-labeled speech database from a single male speaker for segmental duration prediction using CART trees [Campbell 2000]. van Santen states that classification trees require huge amount of training data to cover all possible feature space and proposed the sum-of-products models reference. Sum-of-products model find phoneme durations by a summation of functions of attributes (van Santen 1992, 1993, 1994) [van Santen 1997, Venditti and van Santen 1998, Möbius and van Santen 1996]. Campbell (1992) utilized neural networks for predicting syllable timing [Cambpell 2000]. He used a categorical factor analysis to find out the factors that influence the syllable duration.

For phoneme duration modeling, a collection of attributes are derived from the database such as phoneme identity, left/right context, lexical stress, Part-of-Speech (POS), and etc. The selection of these features is guided by those for other languages in literature and the suggestions of Turkish linguists Prof. Dr. İclal Ergenç, Prof. Dr. Engin Sezer, and Assoc. Dr. Engin Uzun. Relevance of attributes affecting phoneme duration in Turkish are determined by means of statistical analyses. Using regression trees durational attributes are mapped to phoneme durations. The performance of the mapping is evaluated by objective measures such as correlation coefficient (CC), mean absolute error (MAE), and root mean squared error (RMSE). In order to increase phoneme duration prediction performance, modifications on attribute values are proposed. Error minimization in the least squares sense is applied to the resulting predictions in order to further improve RMSE between predicted and actual phoneme durations. Performance of decision trees on predicting discretized segmental duration is evaluated.

1.1 Goals and Outline of the Thesis

In the work presented here, our primary goal is to build pitch contour and phoneme duration models for Turkish using classification and regression trees. A prosodically and phonetically rich speech corpus has been built as part of this study. Relevant prosodic attributes appropriate for pitch contour and phoneme duration modeling are identified

The outline of the thesis is as follows: First chapter introduces a brief definition of prosody and its components. Objectives and motivations are discussed in this chapter.

Focusing on the most influencing research, an overview of different approaches to intonation and duration modeling is given in Chapter 2. Intonation modeling studies are discussed under two broad categories: Phonological and phonetic modeling approaches. Well-known intonation and duration models are introduced.

Chapter 3 introduces the produced text and speech databases. The text database is designed to provide phonetic and prosodic balance. A set-covering algorithm is used to select sentences from a larger set to guarantee phonetic coverage. The phonetically balanced set is modified syntactically to attain prosodic coverage. Designed text is recorded by a native female speaker in a soundproof booth. SAMPA transcriptions of speech files are provided.

Chapter 4 introduces durational attributes used for phoneme duration modeling. Attributes used for phoneme duration modeling involves phoneme (segment) identity, preceding/following phoneme identities, lexical stress, and positional features for segments, syllables, and words.

Phoneme duration modeling studies are presented and their results are discussed in Chapter 5. Forward selection method is used to optimize durational attributes. Performances of the models are quantitatively analyzed. To improve performance several modifications; duration quantization, modification of attribute values, outlier analyses, and mean square error correction, are proposed.

Chapter 6 presents the attributes used in pitch contour modeling. The attributes are associated to syllable units while the ones discussed in Chapter 5 belong to phonemes.

Pitch contour modeling studies are presented in Chapter 7. Two different methods are proposed for pitch contour modeling. One method can be associated to phonetic modeling described in Chapter 1. The other method can be viewed as a phonological model since it

captures the prominence of syllables. Both aim at identifying syllable pitch contours with a limited set of symbols. The symbols used in the former method involved a large codebook that resulted in lower prediction performance. However, the latter uses only binary levels for syllable prominence. Slope values used to reconstruct overall pitch contour of a given sentence are predicted for prominent syllables.

Chapter 8 comprises final conclusions and future directions.

CHAPTER 2

PROSODY MODELING

This chapter reviews intonation and duration modeling studies in the literature.

2.1 Intonation Modeling

General assumption for intonation modeling is that it can be successfully generated with fundamental frequency only, thus, the ultimate goal is to develop a model that generates the fundamental frequency contour of the original utterance. Various intonation models have been have been proposed since 1960's. They are contrasted by different point of views: Phonological versus phonetic models; linear or superpositional models; or models involving level tones or pitch movements. Linear models interpret F0 contour as a linear sequence of phonologically distinctive units (tones or pitch accents), which are local in nature. Superpositional models interpret F0 contour as a complicated pattern of components that are superimposed on each other.

2.1.1 Phonological Models

The goal of a phonological model is to study the organization and underlying structure of intonation [Sun 2002a; Taylor 1992; Dusterhoff 2000; Clark 2003; Monaghan 1992b; Ross 1995]. Intonation patterns are described by a set of abstract descriptions which are regarded as the primitive entities in representing intonation. Generally, the symbol inventory is developed by means of phonetic analysis of F0 curves either from a production perspective or from a perception perspective [Clark 2003; Dusterhoff 2000; Monaghan 1992b; Ross 1995; Sun 2002a; Taylor 1992; Vegnaduzzo 2003].

2.1.1.1 Autosegmental-Metrical (AM) Approach

The most influential work on intonational phonology is the Autosegmental-Metrical (AM) approach which constitutes the basics of American School. Pike (1945) and

Bolinger (1965) were the first who paid more attention to intonation in the American School which was dominated by the notion of phoneme. Pike used four phonemically distinct levels (L, LM, HM, and H) whereas Bolinger used F0 changes to decompose melodies [Goldsmith 1999; Grice *et. al.* 2000; Pierrehumbert, 2000; Pirker *et. al*. 1997]. With the development of Autosegmental analysis (Goldsmith, 1976) and metrical phonology (Liberman, 1975), American phonological community shifted considerably to intonation studies. Autosegmental analysis involves breaking down phonological systems into parallel interacting systems of tones and syllables. In 1975, Liberman proposed metrical phonology as a complementary system to autosegmental analysis. He argued that there were not *two* (accented and unaccented) but *three* functionally distinct roles in which a High/Low contrast arises in English intonation. Liberman called the tone playing the third role a "boundary tone, and indicated by a % adjacent to the tone [Goldsmith 1999; Pierrehumbert 2000].

Pierrehumbert in her dissertation (1980) studied English intonation using the autosegmental-metrical framework. She used the term "pitch accent" developed by Bolinger (Bolinger 1958, 1965) for the tone associated with the accented syllable. Pierrehumbert's intonation model used two basic tone levels (H and L). She proposed bitonal pitch targets which are phonologically located at metrically prominent syllables. She also distinguished pitch accents from boundary tones. Pierrehumbert (1980) proposed seven pitch accents which were then reduced to six by Beckman and Pierrehumbert (1986). The six pitch accents include H*, L*, L+H*, L*+H, H+L*, H*+L. In Beckman and Pierrehumbert (1986), there were two levels of intonational phrasing: intermediate phrase and intonational phrase which were also associated to boundary tones (L or H) [Pierrehumber 2000; Pirker *et. al.* 1997]. A full grammar of possible patterns is given in **Figure 2-1**.

During reconstruction of pitch contours, each tonal element is mapped onto F0 targets which were then interpolated to produce intonation contour referans var mı?. The F0 targets depend on the speaker's choice of pitch range.

Pierrehumbert also employed the concept of downstep for successive high tones occurring in alternating (H L H L H…) patterns [Pierrehumbert 2000]. Downstep involves the lowering of succeeding Hs. Besides, she used an upstep rule which applies only to intonation phrase boundary tones following an H tone.

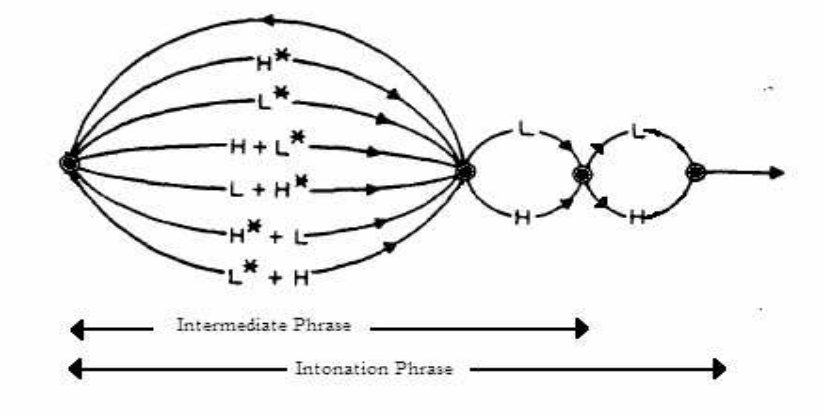

Figure 2-1: The grammar of English intonation patterns according to Beckman and Pierrehumbert (1986) [Pierrehumbert, 2000].

The studies of Beckman and Pierrehumbert (1986) and Pierrehumbert (1980) then evolved into a standard for transcribing American English. Tone and Breaks Indices (ToBI) is the most widely used intonation transcription system at present [Pierrehumbert 2000; Pirker *et. al.* 1997; Silverman *et al*. 1992].

"*ToBI is a framework for developing community-wide conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety. A ToBI framework system for a language variety is grounded in careful research on the intonation system and the relationship between intonation and the prosodic structures of the language.*" [The Ohio State University Department of Linguistics 1999].

ToBI provides a four level transcription system to the researchers:

1) Orthographic/phonetic transcription of the words,

2) *Tone tier* that follows the general outline of Beckman and Pierrehumbert (1986) [Pierrehumbert 2000]

3) *Break indices tier* for indicating the strength of connection between words ranging from 0 (no boundary) to 4 (a maximal, fully-marked intonation boundary) and

4) Miscellaneous tier for any comments.

The prosodic features of ToBI include four intonation features: pitch accent, phrase accent, boundary tone and break index. Pitch and phrase accents and boundary tones are depicted in the tone tier while break indices tier show corresponding break indices.

Pitch accent is the intonational prominence that makes a particular word or syllable implicit in a stream of speech. It corresponds to the local maximum or minimum of the fundamental frequency taking the values H*, L*, L*+H, L+H*, H+!H.

- H* is an apparent tone target on the accented syllable in the upper part of the speaker's pitch range for the phrase.
- L* is an apparent tone target on the accented syllable in the lowest part of the speaker's pitch range.
- L*+H is a low tone target on the accented syllable followed by a sharp rise to a peak in the upper part of the speaker's pitch range.
- L+H* is a high peak target on the accented syllable preceded by a relatively sharp rise from a valley in the lowest part of the speaker's pitch range.
- H+!H* is a clear step down onto the accented syllable from a high pitch but can not be accounted as a high pitch itself.

Phrase accent is the pitch level, which extends the last accent in an intermediate phrase, namely nuclear accent, to the end of the intermediate phrase. It can be either L- or H-.

Boundary tone is the tone type at the end of each intonational phrase. It can be either L% or H%.

Break index indicates the degree of the perceived juncture between adjacent words. It can take values ranging from 0 to 4 [The Ohio State University Department of Linguistics 1999]. An example annotation is given in **Figure 2-2**.

American-English-ToBI is adapted to many other languages such as German, Korean, Japanese, Mandarin, Greek, and etc. Oskay studied ToBI labeling scheme on Turkish [Oskay 2002]. In her thesis, she associates pitch accents and phrasal tones to words assuming that pitch accents always occur on the lexically stressed syllable of words. Her tone inventory is not as rich as Pierrehumbert's though she obtained quite satisfactory results.

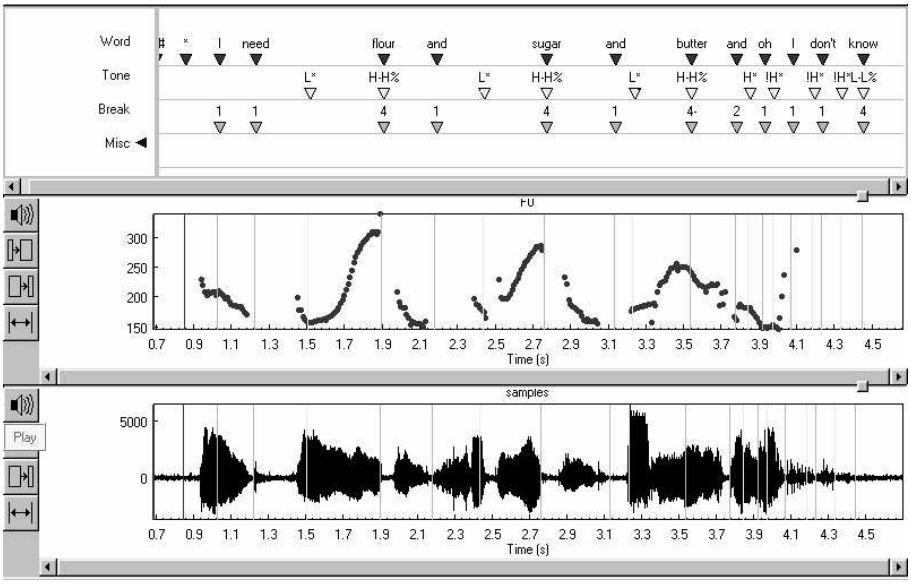

Figure 2-2: TOBI annotation of the Sentence “I need flour and sugar and butter and oh I don’t know”.

2.1.1.2 The IPO (Instituut voor Perceptie Onderzoek) Approach

The IPO approach, developed at Institute of Perception Research at Eindhoven, is probably the best-known perceptual model of intonation ('t Hart et al., 1990) [Koutny *et. al*. 2000; Sun 2002a; Campione *et. al*. 1997; Vegnaduzo 2003]. Although it is counted as a phonological model, it is phonetic in nature. It was originally developed for Dutch and later for English intonation (de Pijper 1983) [Clark, 2003]. The main point of the IPO approach is that only perceptually relevant pitch movements are important to intonation and natural F0 contours can be simplified by means of stylization [Sun 2002a]. Stylization is the process of reducing the amount of information that the fundamental frequency possess while keeping perceptually equivalent [Campione *et. al.*, 1997; Hirst *et. al*, 2000].

In late 70s, De Pijper (1979) introduced the concept of close-copy stylization. The stylized contours are generated by means of straight line segments in log-domain. There is no limit in the number of straight line segments used; however, in order to maintain simplification, their quantity is restricted to the smallest possible value with which the perceptual equivalence can be obtained. A close copy is obtained when subjects are unable to distinguish the synthesized version from the original [Campione *et. al.* 1997; Hirst *et. al.* 2000; Vegnaduzzo 2003].

The IPO model inventory is composed of pitch movements rather than pitch levels. These movements occur between three levels of pitch making *eight* distinct movements which occur as steep and or? Shallow shallow, steep'in karşıtı değil rises and falls between any two levels of the pitch range. Each movement can be aligned with a syllable in *three* ways according to its location: early, middle or late. So, the total number of elements in the inventory of IPO model is 24 movements in total [Campione *et. al*. 1997; Clark 2003; d'Imperio 2000; Vegnaduzzo 2003].

Figure 2-3 contains a representation of the IPO method applied to the sentence 'malaria will worry anyone'. The figure shows original, close copy, and stylized F0 contours of the utterance. Close copy and stylized F0 contours are approximations to original F0 contour. According to the IPO model, the three F0 contours should be perceptually equivalent and acceptable by listeners. The figure is presented to illustrate the data reduction process used in the IPO model [Campione *et. al.* 1997].

The stylization approach proposed for Dutch intonation by Cohen and 't Hart is applied to other languages such as English, German, Russian, French and Indonesian [Campione *et. al.* 1997; Hirst *et. al.* 2000].

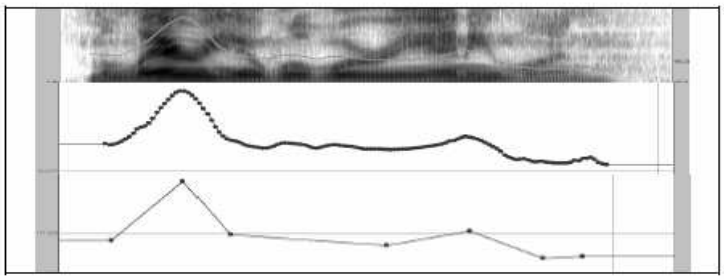

Figure 2-3: IPO data reduction method as applied to the sonorous utterance, "Malaria will worry anyone." Original (top), close copy (middle), and stylized F0 contour (bottom) of the utterance.

2.1.1.3 INTSINT (INternational Transcription System for INTonation)

INTSINT proposed by Hirst and Di Cristo. (1998) is used for coding the intonation pattern of an utterance [Campione *et. al*. 1997; Campione and Veronis 1998a, 1998b, 1998c; Hirst *et. al.* 2000; Hirst *et. al.* 1994; Veronis *et. al.* 1998]. They use the proposed

system for transcribing the intonation patterns of several languages. Pitch patterns are represented as a sequence of discrete tonal symbols (**T**op, **M**id, **B**ottom, **H**igher, **S**ame, **L**ower, **U**pstepped, **D**ownstepped). The pitch patterns are coded either as *absolute tones* (T, M, and B) or *relative tones* (H, S, L, U, and D). Absolute (global) tones are assumed to refer to the speaker's overall pitch range whereas relative (local) tones refer only to the value of the preceding tone. Relative tones can be further split into two categories: *non-iterative* (H, S, and L) and *iterative* (U and D) tones. Iterative raising or lowering uses a smaller F0 interval than non-iterative raising or lowering. There is no corresponding iterative tone for S tones. **Figure 2-4** shows the abstract symbols used in the INTSINT labeling system [Campione *et. al.* 1997; Campione and Veronis 1998a, 1998b, 1998c; Hirst *et. al.* 2000; Hirst *et. al.* 1994; Louw and Barnard 2004; Veronis *et. al.* 1998].

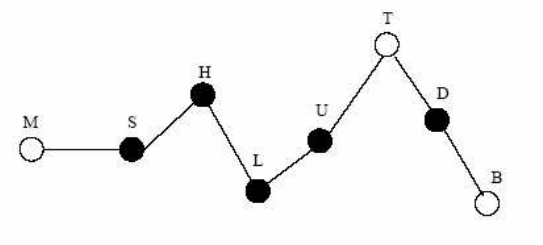

Figure 2-4: INTSINT labelling scheme.

The INTSINT codes for a given speech signal is computed as follows:

1) Code highest and lowest target F0 values as T and B, respectively.
2) Code first target point or any target point that follows a silent pause as M (unless already coded as T or B).
3) Code all other target points with relative tones. Assign an S tone to the targets which are below a predetermined threshold value, otherwise code as H, L, U, or D depending on the targets configuration with respect to its preceding and following target points.
4) Compute the statistical value of each category for each target point. Assign mean values for absolute tones and handle relative tones by a linear regression on the preceding target.

5) To improve statistical model, recode target points H and L as T, U, B, or.

6) Repeat 4 and 5 until no more recoding.

Figure 2-5 contains INTSINT representation of the sentence 'özgüre beni beklemesini söylemedin mi' (didn't you tell özgür to wait for me). The figure also shows the sound waveform, original pitch contour, original and predicted pitch values of the corresponding tones.

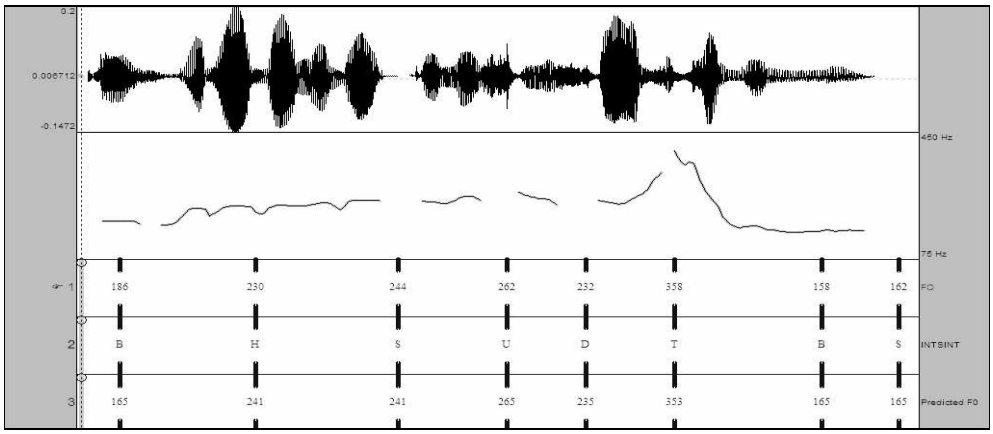

Figure 2-5: Sound waveform, (upper panel), original pitch contour (mid panel), and INTSINT codes of the sentence 'özgüre beni beklemesini söylemedin mi' [Auran 2005].

2.1.2 Phonetic Models

Phonetic models use a set of waveform elements and related parameters to describe intonation patterns of an F0 contour [Dusterhoff *et. al.* 1999; Dusterhoff 2000; Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; Fujisaki 2003; Lee and Oh 2001; Mixdorf 2000, 2001; Mixdorf and Jokish 2001; Möhler and Conkie 1998; Möhler 1999; Ross 1995; Sakurai et. al 2003; Sun 2002a, 2002b; Taylor 1992, 1995, 1998, 2000; Taylor and Isard 1992; Vegnaduzzo 2003; Wright and Taylor 1997]. The ultimate goal of phonetic models is to reconstruct the F0 contour given model parameters. However, for functionality, a phonetic model has to be convenient for linking model parameters and linguistic entities. In fact, the challenge of phonetic models lies in the mapping of linguistic cues to model parameters [Sun 2002a, 2002b;; Ross 1995;].

Depending on their representation of F0 contours, phonetic models are mainly examined under two categories: parametric versus nonparametric. The former approach

transforms the original F0 values into some predictable parametric forms while the latter uses the F0 values themselves [Black and Hunt 1996; Dusterhoff 2000; Sun 2002a; Vegnaduzzo 2003].

2.1.2.1 Parametric Methods

2.1.2.1.1 Fujisaki's Superpositional Model

Fujisaki and Nagashima [Fujisaki and Nagashima 1969] presented a model that generates pitch contours from a set of binary steps corresponding to phrase and accent commands. The model was further improved by Fujisaki and Hirose (1984) to the well-known superpositional model. The later assumes that the actual F0 curve can be expressed by superimposing phrase and accent components in log-domain as given by the equations **(2-1)** – **(2-3)**. A second-order, critically damped linear filter, *phrase command filter*, generates the phrase component in response to an impulse, and the accent component is generated by another second-order, critically damped linear filter, *accent command filter*, in response to a step function [Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; Fujisaki 2003; Mixdorff 2000, 2001; Mixdorf and Jokish 2001].

$$\ln F0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} Aa_j \left[Ga(t - T_{1j}) - Ga(t - T_{2j}) \right] \tag{2-1}$$

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ 0, & \text{for } t < 0, \end{cases} \tag{2-2}$$

$$Ga(t) = \begin{cases} \min\left[1 - (1 + \beta t)\exp(-\beta t), \gamma\right], & \text{for } t \geq 0, \\ 0, & \text{for } t < 0. \end{cases} \tag{2-3}$$

where $Gp(t)$ represents the impulse response function of the phrase control mechanism and $Ga(t)$ represents the step response function of the accent control mechanism. The symbols in the above equations indicate

Fb : baseline value of fundamental frequency,

I : number of phrase commands,

J : number of accent commands,

Ap_i : magnitude of the ith phrase command,

Aa_j : amplitude of the jth accent command,

T_{0i} : timing of the ith phrase command,

T_{1j} : onset of the jth accent command,

T_{2j} : end of the jth accent command,

α : natural frequency of the phrase control filter,

β : natural frequency of the accent control filter,

γ : relative ceiling level of accent components.

Parameters α and β are assumed to be constant at least within an utterance, while the parameter γ is typically set to 0:9.

A block diagram of Fujisaki's superpositional model is given in **Figure 2-6**.

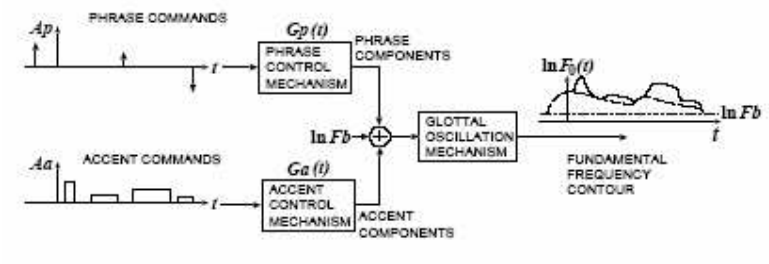

Figure 2-6: A command-response model for F0 contour generation of Japanese utterances [Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; Fujisaki 2003].

Two examples of Analysis-by-Synthesis of F0 contours using Fujisaki's superpositional model is given in **Figure 2-7**.

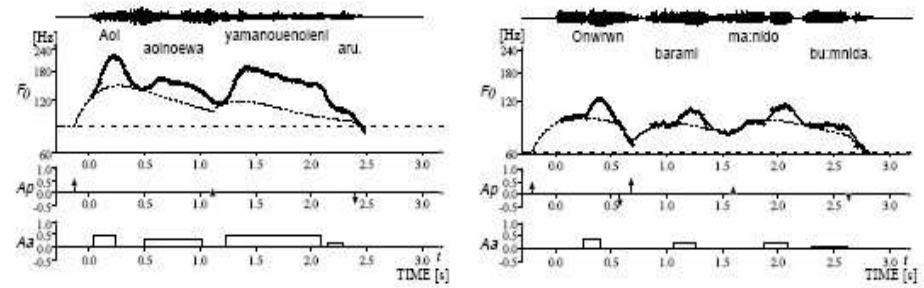

Figure 2-7: Examples of Analysis-by-Synthesis of F0 contours utterances [Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; Fujisaki 2003].

2.1.2.1.2 Tilt Model

In his studies, Taylor [Dusterhoff *et. al.* 1999; Dusterhoff 2000; Taylor 1992, 1995, 1998, 2000; Taylor and Isard 1992; Wright and Taylor 1997] presentes Rise/Fall/Connection (RFC) model, which analyzes an *F0* curve as a sequence of three elements: *rise*, *fall* and *connection* (RFC) [Taylor 1992]. The rise and fall are *parabolic* while the connection element is linear. The basic unit of investigation is the intonational event, which is either a pitch accent or a phrase boundary. Each event is characterized by the amplitudes and durations of the rises and falls. Hence, four parameters, rise amplitude, rise duration, fall amplitude and fall duration, are used to represent events. Taylor, subsequently, introduces the *Tilt* intonational model where the three Tilt parameters, namely *duration*, *amplitude* and *tilt* are obtained by transforming the four RFC parameters. Duration is the sum of the rise and fall durations. Amplitude is the sum of the magnitudes of rise and fall amplitudes. The tilt parameter is a *dimensionless* number taking values between [-1, 1]. Tilt parameter expresses the overall *shape* of the event [Taylor 2000]. A pure rise (fall) takes a value of 1 (-1) while a rise-fall (fall-rise) pattern whose rise and fall magnitudes are equal values takes a value of 0. Tilt parameters are computed as follows:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{2-4}$$

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{2-5}$$

$$tilt = \frac{1}{2} tilt_{amp} + \frac{1}{2} tilt_{dur} \tag{2-6}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \tag{2-7}$$

$$D_{event} = |D_{rise}| + |D_{fall}| \tag{2-8}$$

Graphical representations of Tilt parameters are given in **Figure 2-8**. **Figure 2-9** depicts a schematic representation of F0 contour and corresponding Tilt parameters associated to the syllable nuclei.

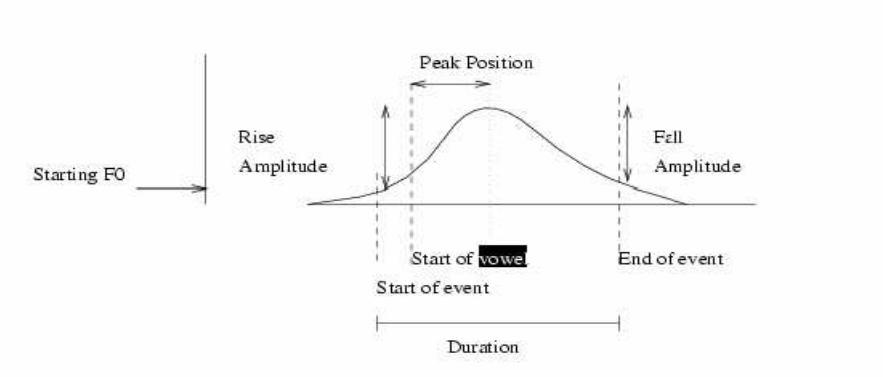

Figure 2-8: Tilt parameters [Dusterhoff 2000]

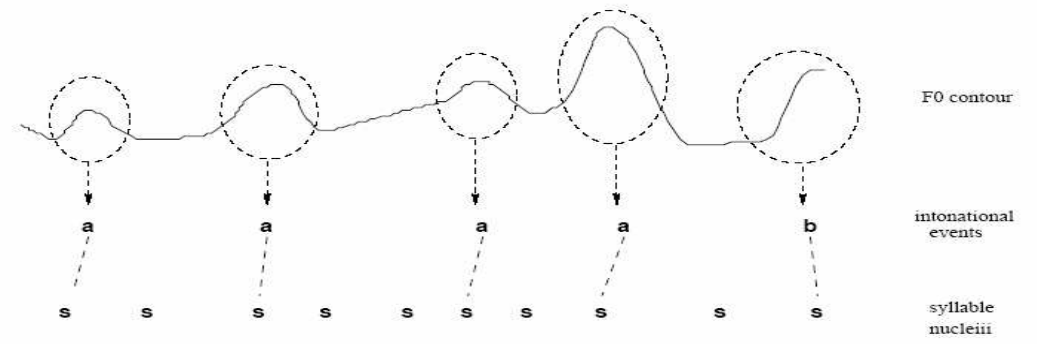

Figure 2-9: Schematic representation of F0, intonational event stream (circled events) and segment stream in the Tilt model. Events, labelled **a** for pitch accent and **b** for boundary are associated to syllable nuclei of syllable stream [Taylor, 2000].

2.1.2.1.3 MOMEL (MOdélisation de MELodie)

MOMEL was originally proposed by Hirst (1980, 1983, 1987, and 1992) and automated by Hirst and Espessser (1993). MOMEL represents the fundamental frequency as a sequence of target points (relevant local variations of F0 curve) in frequency and time pairs, <*F*0, *t*>. For interpolation, MOMEL uses a quadratic spline function resulting in a continuous, smooth curve. Quadratic splines provide a simpler codification. In order to maintain the continuity of the resulting curve, interpolation is performed over the

unvoiced segments [Campione *et. al.* 1997; Campione and Veronis 1998a, 1998b, 1998c; Hirst *et. al.* 1994; Hirst *et. al.* 2000].

The model has been used for the analysis of F0 contours of other languages including English, French, Spanish, Italian, and Arabic [Campione *et. al.* 1997; Campione and Veronis 1998a, 1998b; Hirst *et. al.* 2000].

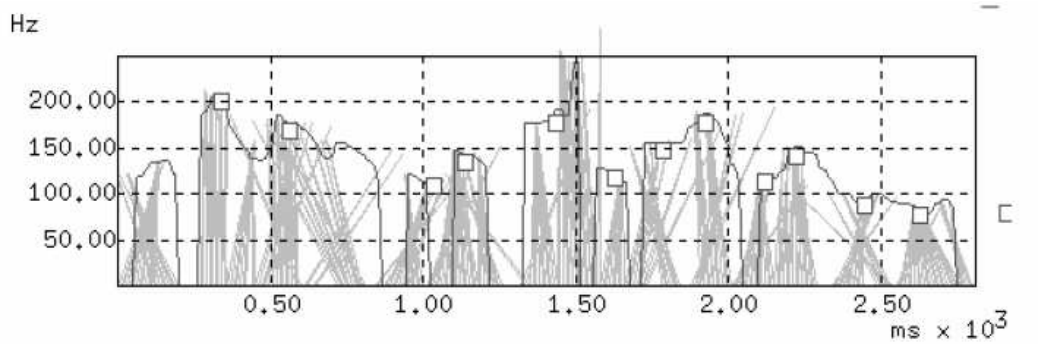

Figure 2-10: Estimation of candidate target point (grey lines) and final targets (white squares). The grey lines connect the centre of the moving window to the extremum of the parabola estimated for that window [Campione *et. al*, 2000].

2.1.2.1.4 Parametric representation of Intonation Events (PAintE)

Mohler and Conkie (1998) describe an intonation event using two *sigmoids* with a fixed time delay given by the following equation:

$$f(x) = d - \frac{c_1}{1+\exp(-a_1(b-x)+\gamma)} - \frac{c_2}{1+\exp(-a_2(x-b)+\gamma)} \qquad (2\text{-}9)$$

where a_1 and a_2 represent the steepness of sigmoids, c_1 and c_2 model the amplitude of sigmoids, b stands for the alignment of the function and d corresponds to the function's peak. The syllable length is defined as unity. PAintE model function and parameters are given in Figure 2-11.

This model emphasizes intonation events like Tilt model does. F0 contours' parameterization is applied to the accented syllable as an anchor point. The

approximation is performed within a three syllable window around the syllable carrying accent.

Möhler and Conkie (1998) introduced Vector Quantization (VQ) of PaIntE parameters. They use codebooks of different size. They argue that:

1) intonation can be described by a number of distinct shapes,

2) reducing data can improve machine learning performance;

3) VQ allows predicting all six parameters together rather than individually [Möhler and Conkie 1998; Möhler 1998, 1999; Sun 2002a].

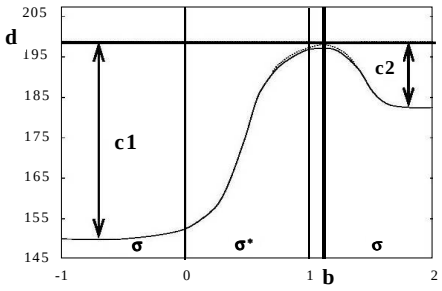

Figure 2-11: The PaIntE model function is the sum of a rising and a falling sigmoid with a fixed time delay. Time axis is in syllable units [Möhler and Conkie 1998].

2.1.2.2 Non-Parametric Models

Non-parametric methods use F0 values themselves to generate intonation based on available linguistic information [Sun 2002a]. Non-parametric methods use samples of smoothed and interpolated F0 contours usually associated to syllable units. Traber (1992) uses Recurrent Neural Networks (RNN) to predict a number of pitch values per syllable [Buhmann *et. al.* 2000]. Based on Traber (1992), Buhmann *et. al.* incorporates RNNs to predict 5 F0 values per syllable for 6 languages [Buhmann *et. al.* 2000]. Lee and Oh predictes 10 F0 values values per syllable using a CART tree to generate an F0 contour for a given sentence [Lee and Oh 2001]. Ross and Ostendorf (1999) develop a dynamical system model to predict normalized F0 values in a syllable. They also use regression trees to predict dynamic range of the F0 contours [Ross 1995; Ross and Ostendorf 1999].

2.2 Duration Modeling

Generally three prosody components are modeled: Intonation, duration and intensity [Batusek 2002]. Duration plays as much important role as intonation in the encoding/decoding of speech by the speaker/listener. The durational patterns are part of the prosody and contain important cues for understanding the spoken text [Riedi 1998]. Variations in duration provide assistance for the listener to understand the meaning [Campbell 2000]. Different representational factors specify and modify several aspects of speech during speech production [Klatt 1987]. Historically, duration prediction models can be split into two as *rule-based* and *corpus-based*-duration models.

One of the most salient rule-based duration models has been proposed by Klatt [Möbius and Santen 1996]. Klatt's work on duration modeling has pioneered the development of several duration models. Klatt's immediate antecedents are Peterson and Lehiste (1960) and Barnwell (1971). Klatt use the notion of intrinsic duration introduced by Peterson and Lehiste (1960) [Campbell 2000]. They define intrinsic duration as "the average duration of syllable nucleus measured from minimal pairs differing in the voicing of the final consonant". Peterson and Lehiste carry on a comparative study on the durations of read 1263 single words in a sentence and report specifically that the syllable nuclei tend to be shortened when followed by a voiceless consonant. Barnwell (1971) presents an algorithm to model vowel duration as a function of [Campbell 2000]:

- the word-level stress of the parent syllable,
- the structural location of the parent word,
- the number of syllables in the parent word,
- the proximity of any word or syllable juncture.

Dennis Klatt summarizes Barnwell's work (1971) and proposes a set of rules to model duration. Both Klatt and Barnwell (1971) use a context-related percentage change following Peterson and Lehiste's (1960) findings. Klatt's model evolves into its final form in 1987 (Klatt 1987). The model assumes that each phonetic segment type has an inherent duration that is specified as one of its distinctive properties, each rule assigns a percent increase or decrease in the duration of the segment but segments cannot be shortened less than a certain minimum duration (Klatt 1987). The model is summarized as:

$$DUR = MINDUR + (INHDUR-MINDUR) \times PRCNT / 100 \quad (2\text{-}10)$$

where INHDUR is the inherent duration of a segment in ms, MINDUR is the minimum duration of a segment if stressed and twice that if unstressed, and PRCNT is the percent shortening determined by applying Klatt's eleven rules.

Many other rule systems have been developed for different languages. However, these systems were developed when sufficient speech data and computational power to analyze the data did not exist. Recently, with the availability of large speech corpora and advances in computational power, a general interest in corpus-based methods has arised [Möbius and Santen 1996]. Corpus-based statistical models employ natural speech data. Generally, model parameters are trained over the data to optimize some criteria [Kenney 1998; Lemmetty1999].

Application of Classification and Regression Trees (CART) [Breiman *et. al.* 1984] to segmental duration prediction appears in the context of Corpus-Based statistical modeling. The input is formed by attribute-value pairs in CART modeling. Successive splitting of data into two sub-trees, in which the variance of newly formed subsets is minimal with respect to dependent variable, forms a regression tree. For each node of the tree, observed average duration of the associated subset of the corpus is listed. Riley (1990, 1992) uses a 1500-sentence hand-labeled speech database from a single male speaker for segmental duration prediction using CART [Campbell 2000]. Lee and Oh proposed 10 features to predict segmental duration from a set of 400 sentences using CART trees [Lee and Oh 1999].

One of the great advantages of CART is that the algorithm has the validation of the model. CART builds a very complex tree and then pruns it back to an optimal tree based on the results of cross validation or test set validation. The tree is pruned back based on the performance of the various pruned versions of the tree on the test set data. The most complex tree rarely fares the best on the held aside data as it has been over fitted to the training data. By using cross validation the tree that is most likely to do well on new, unseen data can be chosen.

CART algorithm is relatively robust with respect to missing data [Breiman *et. al.* 1984]. If the value is missing for a particular predictor in a particular record, that record

will not be used in making the determination of the optimal split when the tree is being built. In effect CART utilizes as much information as it has on hand in order to make the decision for picking the best possible split.

Van Santen (1992) summarizes different approaches employed in duration modeling for synthesis applications. He identifies four principle classes:

- Sequential rule systems such as Klatt's model (1987);
- equation systems,
- look-up tables,
- binary classification trees referring to Riley' studies (1992) [Campbell 2000; van Santen *et. al.* 1997].

He states that lookup tables and classification trees require huge amount of training data to cover all possible feature space and proposes the sum-of-products models (1992, 1993,1994) [Santen 1997, Venditti and Santen 1998, Möbius and Santen 1996]. Sum-of-products model combines scales of attribute values by forming sums and products. According to this model, segment duration is given by

$$DUR(f) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_i) \tag{2-11}$$

where $S_{i,j}$ is the function representing the influence of factors *i, j* and f_i is the i^{th} element of descriptor vector $\boldsymbol{f}$ [Batusek 2002; Möbius and van Santen 1996; van Santen 1997; Venditti and van Santen 1998].

Neural networks constitute another method for prosody modeling. Campbell (1992) utilizes neural networks for predicting syllable timing "to account for the interaction between higher and lower level of timing control" [Cambpell 2000]. He employs a categorical factor analysis to find out the factors that influence the syllable duration. A three-layer back-propagation neural network is used to predict syllable durations as a first approximation. In the second stage, a top-down accommodation process determines the durations of each segment in syllable where syllable duration is partitioned among phonemes of the syllable according to their intrinsic duration.

In their work, Rao and Yegnanarayana use a four layer feedforward neural network trained with standard backpropagation algorithm for predicting syllable durations of Indian languages. Riedi applied a neural network to duration modeling in German and obtained very good results [Riedi, 1995]. Cordoba *et. al.* uses phoneme as the base unit in their model involving neural networks [Cordoba *et. al.* 2002].

Ostendorf and Roukos propose the stochastic segment model, the recognition algorithm, and an iterative training algorithm for estimating segment models from continuous speech [Ostendorf and Roukos, 1986].

2.3 Research on Turkish Prosody

Prosodic analysis experienced a considerable boost about 20 years ago, and there have been an increasing number of publications on prosodic research for the last two decades [van Santen *et. al.* 1997]. However, Turkish has been left almost untouched. The only attempt to model Turkish prosody is Oskay's in her masters' thesis [Oskay 2002]. In her studies, she adapts ToBI [Black and Hunt 1996; Pierrehumbert 2000] for Turkish and models word–level labels via machine learning. She uses a database of 400 sentences recorded by herself. The sentences are selected from the Turkish Treebank Corpus [Metu-Sabancı Turkish Treebank Corpus 2005; Nart, Oflazer and Say 2003]. Treebank provides the morphosyntactic information of the sentences. In her thesis, Oskay performs word, syllable and phoneme level transcriptions of speech waveforms automatically and incorporates modified ToBI labels (H and L tones) to the transcription. She uses an inductive learning scheme, RIPPER [Cohen 1996] developed at AT&T to predict word-level ToBI labels.

Abdullahmeşe devised a fundamental frequency contour synthesis system relying on a sentence database and utilizing the syntactic structure of sentences based on word categories and stress information [Abdullahmeşe 2001].

In his thesis, Özge argues that prosody is the sole structural determinant of information structure and proposes a tune-based account for the structural realization of information structure in Turkish [Özge 2003].

CHAPTER 3

TEXT AND SPEECH CORPORA DEVELOPMENT

Speech corpus design is one of the key issues to improve the naturalness of synthetic speech. A language can be considered as the set of all possible combinations of speech units. However, it is not possible to have all combinations in a database. A speech database can be built randomly or by means of optimizing the units acoustically or with respect to their textual properties. For our purposes, random selections may not be adequate to provide sufficient variability. Thus, it is aimed to construct an optimal continuous speech database consisting most frequent units with more than one representation.

Recent studies about prosody modeling use speech corpora of limited size. In their research about Korean prosody modeling, Lee and Oh used 400- and 500-sentence databases, 60% of which are used for modeling and remaining for testing [Lee and Oh 1999, 2001]. For modeling Spanish duration, Cordoba *et. al.* employe 732 phrases [Cordoba 2002]. In their studies on automatic classification of intonational phrase boundaries, Wang and Hirschberg uses 298 utterances from the 774 sentences in the DARPA collected Air Travel Information Service (ATIS) database [Wang and Hirschberg 1992]. Black and Hunt test their regression-based F0 contour modeling on Boston University FM Radio Corpus (Speaker f2b) (include 14778 syllables) [Black and Hunt 1996]. Dusterhoff, Black and Taylor uses 3 different databases: Boston University Radio News Corpus (Speaker f2b), 450 TIMIT sentences (10% of which are questions), and an instructional text database consisting 43 excerpts of text describing a museum exhibition [Dusterhoff *et. al.* 1999]. For duration modeling in German, Möbius and van Santen employ Kiel Corpus of Read Speech which includes 23490 phonemes [Möbius and van Santen 1996]. Venditti and van Santen perform an optimization over 34000-sentence database and obtaines 197 sentences covering their feature space. They use this

197 sentence to model Japanese duration [Venditti and van Santen 1998]. In Sakurai, Hirose and Minematsu, 486 sentences are used to generate F0 contours [Sakurai *et. al*. 2003]. For experimental purposes, Agüero, Wimmer and Bonafonte use a Spanish corpus of 500 sentences for the joint extraction and prediction of Fujisaki's parameters [Agüero *et. al*.2004]. In a very recent work on Hindi duration modeling a corpus of 250 sentences is utilized [Krishna *et. al.*2004]. The list may not be complete, however, it points out the importance of speech corpus in prosody modeling.

Construction of a speech database requires three stages: creation of a text corpus, recording and annotation. In the following sections each step involved in text and speech corpora development is introduced.

3.1 Text Corpus

Text design by random selection of sentences from various topics is one of the most frequently used techniques for speech corpora design. However, corpus formation is a long and difficult task and therefore some means of optimization are necessary. Especially for building open domain applications, optimization becomes a must since recording every possible speech event is practically impossible.

The coverage concept is very appropriate in formulating the problem and searching for solutions. The aim can be stated as optimal design of a text corpus, which has highest coverage for a target synthesis domain. Coverage of a domain can be defined via the concept of unit. Units in this research are determined to comprise *phonemes* and *sentence types* to account for phonetic and prosodic variety.

As a prosodic corpus, it has to be representative of the prosodic variations of the language. The corpus may also be available for synthesis research thus another point is to provide phonetic balance. Along with the specifications, a two-step approach is taken. In the first stage, phonetic coverage is provided and in the second one, the resulting database is forced to present prosodic variations by means of adding new sentences or changing the types of the sentences obtained after the first stage.

Our initial text database is a collection of sentences selected from various resources such as grammar books [Adalı 1979; Aksan 1995; Atabay *et. al.* 1981; Hatiboğlu 1972; Kornfilt 1997], newspapers and TREEBANK project [Metu-Sabancı Turkish Treebank Corpus 2005; Nart, Oflazer and Say 2003]. A first selection has been performed on

newspaper and TREEBANK text to remove very long and short sentences. The resulting text together with the sentences taken from the grammar books[1] constitutes the main source text (5802 sentences) used in this study. There are 43867 words 16708 of which are distinct. Phonetic transcription of the text database has been performed using Turkish SAMPA conventions [Well 2003]. The total number of occurrences for 42 SAMPA characters is 305341. The sentence lengths in terms of word counts of the source text are given in **Figure 3-1**.

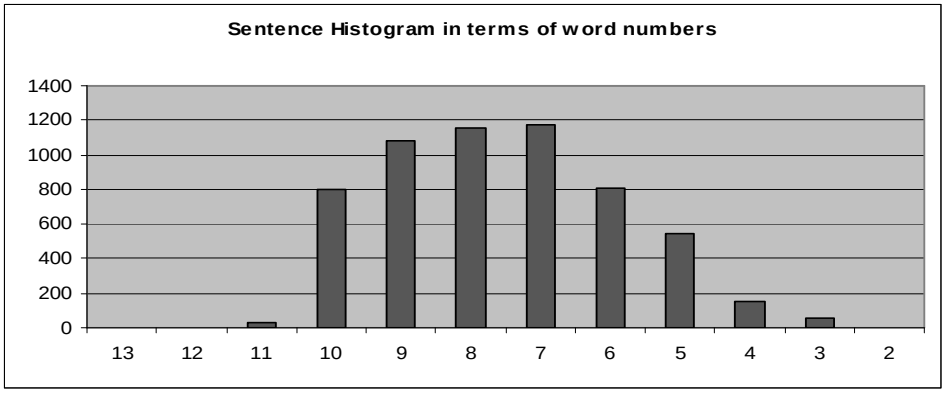

Figure 3-1: Sentence histogram of the original database in terms of word numbers.

Since large databases require much time and effort for building speech corpora, it is needed to condense the text database to a convenient size. Database reduction is presented in the following sections.

3.1.1 First Step: Phonetic Coverage

One of the constraints on the resulting database is to have phonetic variety and phonetic balance. There exist rare phonemes in Turkish such as Z in aZan (agent) and w in tawuk (chicken). When random selection is considered to reduce database size, the probability of the coverage of the sentences containing rare phonemes becomes very low. Since phonetic balance is desired, a greedy-like approach to cover the rarest phonemes in the source text is followed.

[1] Sentences from grammar books do not need a filtering in length since they are already in desired form.

A set of sentences (approximately 650) including rarest phonemes is selected from the source text database (5802 sentences). Since there is no other constraint on the selection process it is expected to maintain the original distribution in some manner. Resulting phoneme distributions reveal that selecting the sentences with rarest phonemes result in a text having nonrare phoneme distribution almost the same as that of the original one, yielding phonetic balance.

Diphone coverage is another important issue. The original source text has a total number of 1130 distinct diphones consisting approximately 64.1% of all-possible diphone combinations (41 SAMPA + pause = 42; 42*42-1 = 1763 diphones). The diphone coverage of the original database seems to be very low; however some combinations are not encountered. The total number of diphone occurrences is 299539. Resulting database has 949 distinct diphones, which is approximately 84% of the original diphone set, with 36280 occurrences.

With the help of a simple greedy approach, the dimension of our original database is reduced from 5802 sentences to 675 sentences (approximately 88.4% reduction ratio) and achieved full phonetic coverage and partial diphone coverage.

3.1.2 Second Step: Prosodic Coverage

Sentence types and phonemes are taken as units to be covered in the database. To provide phonetic coverage, 675 sentences are chosen from the text database consisting 5802 sentences using a greedy-like approach. For prosodic coverage, it is aimed to cover sufficient representations of each sentence type in Turkish. Following section introduces sentence types encountered in Turkish.

3.1.2.1 Turkish Sentence Types

Broadly, Turkish sentences can be investigated under three categories that can be further split into subcategories [Adalı 1979; Aksan 1995; Atabay *et. al*. 1981; Demircan 2001; Hatiboğlu 1972; Kornfilt 1997]:

Depending on Syntactic Constituents

Simple Sentence

Simple sentences are composed of only one judgment with one verb. Since they are simple in structure, their prosodic variations are also simple. In most of the sentences, preverbal word carries sentence focus. However this is not so when different types are

considered: Question enclitics affect preceding word; question words constitute focus of the sentences, and etc. Some examples of simple clauses are given below (words in italic indicate verbs of the sentences):

Ahmet annesini ziyaret etti (Ahmet visited his mother) (**Figure 3-2**)

Kim televizyon seyretti (Who watched television?)

Hasan kitabı okumadı (Hasan didn't read the book)

Compound Sentence

Sentences consisting more then one verb are known to be compound sentences. Each sub-sentence can be viewed as an intonation group. They have a complex prosodic structure. Compound sentence examples are given below (intonation groups are enclosed within /'s):

Hasan nereye gitmişse/ orada kaldı (Hasan stayed wherever he went) (**Figure 3-3**)

Bir adam /ki çocuklarını sevmez/ yalnız yaşamalıdır (A man who does not love his children must live alone)

Complex Sentence (Clauses)

Complex sentences are composed of a main clause and one or more nominal, adjectival or adverbial clauses. In most of the cases, prosodically, complex sentences can be handled as simple sentences. However, as given in the examples below, they may show complex prosodic structures.

Yarın benimle sinemaya gelmeni istiyorum (I want you to come to the movie with me tomorrow) (**Figure 3-4**)

Ahmet /çok çalışarak/ hedefine ulaştı (Ahmet attained his goal by working a lot)

/Müdürün tatile çıkmasından sonra/ ofis kapandı (After the director went on vacation, the office is closed)

Coordinate Sentence

Coordinate sentences are composed of more than one simple, compound, or complex sentences that are related to each other in terms of meaning. Although there are complete simple sentences, the meaning of coordination requires different intonational patterns (continuation rise). Examples are given as follows:

Hasan işe gitti, Ali evine döndü, ben de parkta kaldım (Hasan went to work, Ali returned home, and I stayed in the park)

Hasan arabayı yıkadı ve evi süpürdü (Hasan washed the car and swept the house) (**Figure 3-5)**

Hasan istakozu pişirdi, Ali de balığı (Hasan cooked the lobster and Ali cooked the fish)

Reported Speech

Although reported speech sentences can be thought as compound sentences, we investigate them as another type to emphasize the difference in their pronunciation. Below are the examples:

Komşular /yarın seyahate çıkacağız/ dediler (The neighbors said: We will go on a trip) (**Figure 3-6**)

Ahmet /sinemaya gideceğim/ diye mırıldandı (Ahmet muttered 'I will go the movie)

Depending on Verbal Composition

Verb-Final Sentence

Turkish is a standard Subject-Object-Verb (SOV) order language. However, variations to this structure exist, i.e. OSV, SVO, OVS, VSO, and VOS depending on the focus of the sentence. The word to be focused comes to preverbal location in general. Among these various compositions, SOV and OSV structures are known to be verb-final sentences Sentences given in previous subsection are examples of verb-final sentences.

Non-Verb Final Sentence

Sentences in SVO, OVS, VSO, and VOS are named to be non verb-final sentences. In general, non verb-final sentences are used in conversations. They are mostly encountered in poems and daily communication like e-mails and messages. They are used to express emphasis in formal writing. Example of a non verb-final sentence is given below:

Hasan bugün yedi istakozu (Hasan ate the lobster today) (**Figure 3-7**)

Depending on Semantics

Affirmative Sentence

They are the sentences carrying positive sense. All of the previous examples are in affirmative form. Some examples are given below:

Ahmet annesini ziyaret etti (Ahmet visited his mother) (**Figure 3-2**)

Hasan işe gitti, Ali evine döndü, ben de parkta kaldım (Hasan went to work, Ali returned home, and I stayed in the park)

Negative Sentence

Negation in sentences is either marked by negation enclitics (-me, -ma) or by the word ‘değil (not)’. When used, the meaning of the sentence should be sensed in the opposite manner. Examples are as follows:

Hasan istakozu bugün yemedi (Hasan did not eat the lobster today) (**Figure 3-8**)

Hasan kitabı okumadı (Hasan didn’t read the book)

Interrogative Sentence

These are question forms. The questions can be formed using question enclitics (-mi, -mı, and variants) or question words like kim (who), ne (what), nerede (where), and etc. Verb-final sentences are turned into question sentences by placing the question enclitic before the subject under suspect or using the question words instead of the subject itself. Examples are:

Hasan bugün istakoz mu yedi (did Hasan eat lobster today) (**Figure 3-8**)

Ahmet neyi öğrencilere sattı (Ahmet sold what to the students)

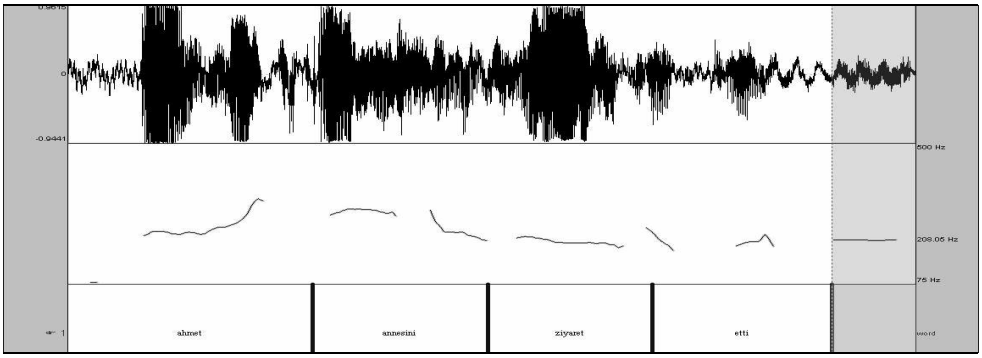

Figure 3-2: Example of an affirmative, simple, and verb-final sentence: “Ahmet annesini ziyaret etti”. Speech waveform (upper) corresponding F0 contour (middle) and word segmentation (bottom). The pitch contour declines throughout the utterance.

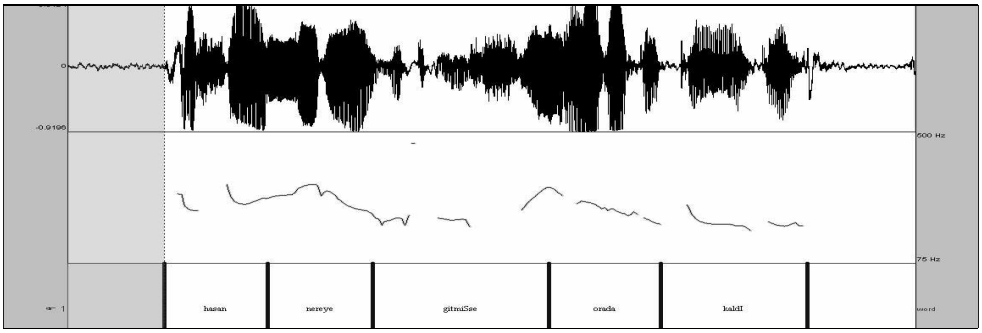

Figure 3-3: Example of an affirmative, compound, and verb-final sentence: “Hasan nereye gitmişse orada kaldı”. There are two intonational phrases: Second intonational phrase starts at the word ‘orada’.

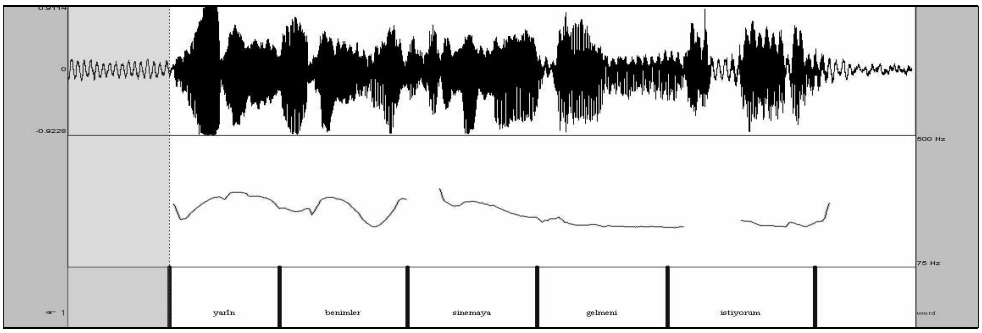

Figure 3-4: Example of an affirmative, complex, verb-final sentence: “Yarın benimle sinemaya gelmeni istiyorum”.

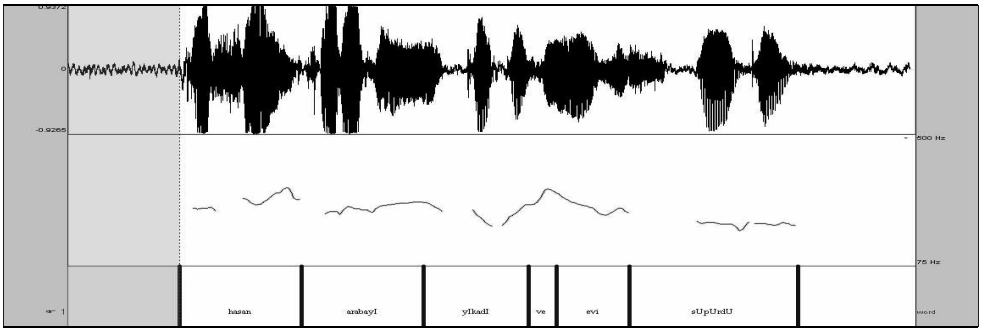

Figure 3-5: Example for an affirmative, coordination, and verb-final sentence: “Hasan arabayı yıkadı ve evi süpürdü”.

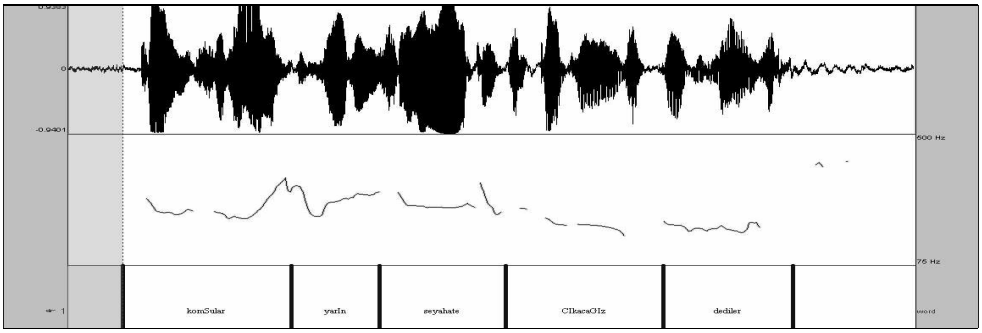

Figure 3-6: Example for an affirmative, reported, verb-final sentence: “Komşular yarın seyahate çıkacağız dediler”.

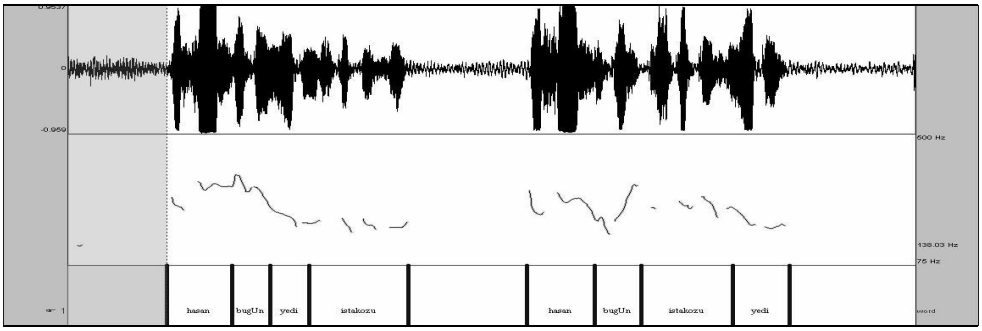

Figure 3-7: Examples for affirmative, simple, non verb-final and affirmative, simple, verb-final sentences: “Hasan bugün yedi istakozu” and “Hasan bugün istakozu yedi”.

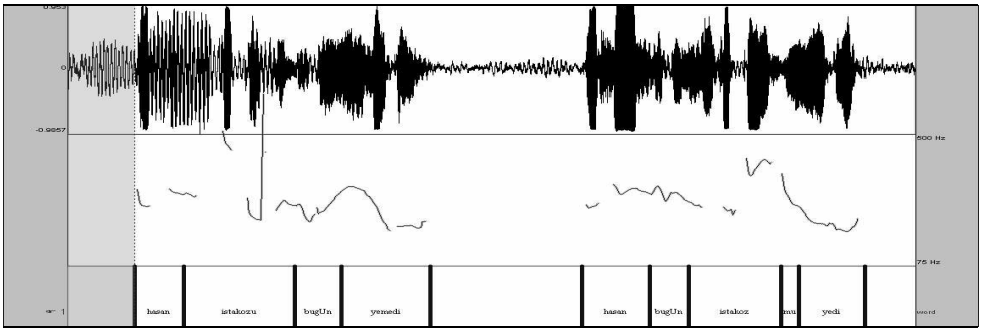

Figure 3-8: Examples for negative, simple, verb-final and affirmative, simple, question forms: “Hasan istakozu bugün yemedi” and “Hasan bugün istakoz mu yedi”.

Each sentence in the 675-sentence reduced database is annotated with respect to their sentence structures. Although this reduced subset contains almost all sentence structures,

there are not sufficient representatives for non verb-final, compound and coordination, reported speech sentences as well as interrogative sentences. However, non verb-final constructions are frequently used in daily life, informal written materials as well as literature especially in poetry and theatrical texts [Adalı 1979; Aksan 1995; Atabay *et. al*. 1981; Demircan 2001; Hatiboğlu 1972; Kornfilt 1997]. Hence, it is necessary to increase the size of non verb-final sentences in our database. They have been increased to acccount for the 29.4% of the overall database).

Reported speech sentences are another frequently used pattern existing in daily life, informal and news texts, and literature [Adalı 1979; Aksan 1995; Atabay *et. al.* 1981; Demircan 2001; Hatiboğlu 1972; Kornfilt 1997]. Their original size should also be increased, however not to repeat words indicating reported speech such as 'dedi (said)', we avoid increasing the number of such sentences to a comparable level as that of other types (10% of the overall database).

Mostly by adding new sentences, the number of compound and coordination sentences is increased to a comparable level (14% and 20% respectively). Sentences are made interrogative by appropriate question words and morphemes.

While selecting sentences and/or converting them into an appropriate form, special care is taken to ensure easily pronounceable sentences. Sentences seemed to be nonsense or hard to pronounce are deleted during prosodic coverage.

Some of the previous 675 sentences have undergone small perturbations or deletions to provide balance between sentence structures. The remaining 335 sentences are selected either from the original database or from other resources to fulfill the sentence type balance. The total number of changed sentences in the resulting database to provide prosodic balance is 555 (55.5%).

The original database has been enlarged to 5903 sentences with 311542 segments and the final reduced database is composed of 1000 sentences (16.9% of the original database) with a total of 54892 segments (17.6% of the original segment size). The total number of distinct diphones in the expanded and the resulting text are 1116 and 991, respectively. The resulting database covers approximately 89% of the diphones of the expanded database. The total number of diphone instances for expanded and the resulting database are 305639, and 53892 (approximately 18%), respectively.

Sentence distribution of the resultant database is given in **Figure 3-9**. Approximately 30% of the database is composed of non verb-final sentences, while 70% is of verb-final sentences as desired. The *Reported Speech* sentences are allowed to reach a threshold value of 10% which seemed to be a good resolution. Among the sentence types, the complex sentence number is above the average, around 34%. This is an expected result since our original database sentences are composed mostly of complex sentences. It should be mentioned that in written Turkish most of the grammatical forms observed are verb-final, simple/complex affirmative sentences.

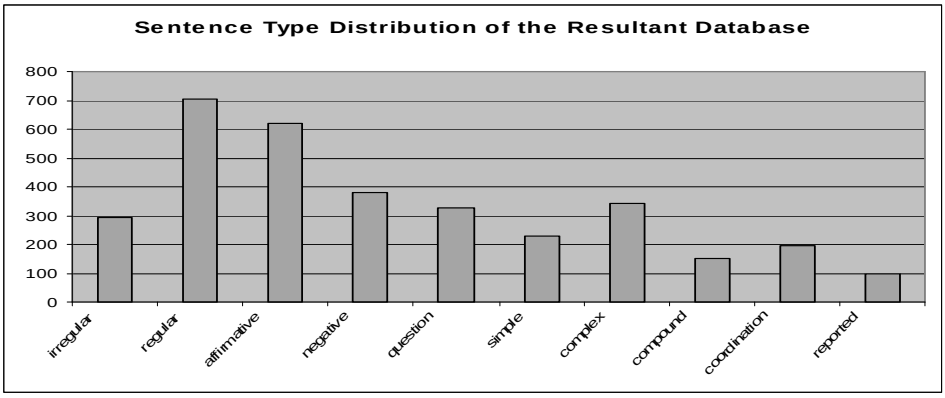

Figure 3-9: Sentence type distribution of the resultant database.

3.2 Speech Corpus

Last step in building METU Prosodic Speech Corpus is the recording process. Selected prompts are recorded in a soundproof booth located at METU's speech lab (**Figure 3-10**). The speaker uses a Sennheiser microphone with a ME102 modular mini-microphone capsule.

Selected sentences have been recorded using EMU Speech Tools [Cassidy and Harrington 1996]. EMU is a collection of software tools for the creation, manipulation and analysis of speech databases.

After recording, speech waveforms are examined perceptually. It has been noted that the speaker made natural deviations: i.e. pronounce words such as yapmayacağım as yapmıca:m or değil as di:l, etc. These are carefully examined and the speech database is then rebuilt considering these modifications.

3.2.1 Labeling

Resulting speech database is designed to serve as a research material for prosodic modeling; hence it is necessary to provide a basic annotation scheme along with the speech corpus. Many researchers employ phonemes as segments in duration modeling whereas more complex units such as syllables or words are used in F0 contour modeling. Phoneme boundaries can be used to obtain syllable or word boundaries.

Automatic phonetic labeling of speech corpus is performed using HTK Speech Recognition Tool [University of Cambridge 2005] developed for building and manipulating Hidden Markov models. 70% of the labels are then manually corrected and used in phoneme duration and pitch contour modeling studies.

Figure 3-10: Soundproof booth.

CHAPTER 4

IDENTIFICATION OF DURATIONAL ATTRIBUTES

In natural speech, two similar speech sounds rarely have exactly the same durations due to many factors [Campbell 2000]. The influences of these factors on durational characteristics of speech have been investigated from the very beginning of prosody research. Selecting incomplete or inappropriate set of attributes results in an erroneous prediction of duration. Therefore, determination of attributes that have greater influence on speech timing is a crucial step in duration modeling process.

Various durational attributes have been used in the literature for modeling purposes. Campbell (1992) used *number-of-phones-in-the-syllable*, *nature-of-syllabic-peak*, *position-in-tone-group*, *type-of-foot*, *stress* and *word-class* to predict syllable timing [Campbell 2000]. Shih and Ao utilize *segment-identity*, *tone-identity*, *previous/next-segment-identity*, *previous/next-tone-identity*, *degree-of-discourse-prominence*, *number-of-preceding-syllables-in-the-word/phrase/utterance*, *number-of-following-syllables-in-the-word/phrase/utterance*, *syllable-type* for modeling Mandarin duration [van Santen *et. al.* 1997]. van Santen (1994) use *phone-identity*, *surrounding-phones-identity*, *pitch-accent*, *syllabic-stress*, *within-syllable/word/utterance-position* [Cordoba *et. al.* 1999, 2002]. Venditti and van Santen employ *current/preceding/following-phone-identities*, *left/right-prosodic-context*, *accent-status*, *syllable-structure* and *special-morpheme-status* for Japanese duration modeling [Venditti and van Santen 1998]. *Segment-identity*, *segment-type*, *word-class*, *position-of-phrase-in-utterance*, *phrase-length-in-number-of-words*, *position-of-word-in-phrase*, *word-length-in-number-of-syllables*, *position-of-syllable-in-word*, *stress*, *segment-position-in-syllable*, *segmental-context*, *segmental-context-type* are used by Möbius and van Santen for modeling German duration [Möbius and van Santen 1996]. For modeling Catalan duration, Febrer *et. al.* utilize *vowel-identity*, *stress*, *sentence-position*, *post-vocalic-phone-class* and *manner-of-articulation* [Febrer *et.*

al. 1998]. For duration modeling in Spanish, *phone-identity*, *contextual-phones*, *stress*, *stress-in-the-syllable*, *syllable-beginning-with-vocal*, *diphthong*, *phone-in-a-function-word*, *phrase-type*, *positioning-phrase* and *number-of-units-in-the phrase* are employed [Cordoba *et. al.* 1999]. Attributes used to predict Hindi duration are as follows: *segment-identity*, *segment-features*, *previous/next-segment-features*, *parent-syllable-structure*, *position-in-parent-syllable*, *parent-syllable-initial/final*, *parent-syllable-position-type*, *number-of-syllables-in-parent-word*, *position-of-parent-syllable*, *parent-syllable-break-information*, *phrase-length-in-number-of-words*, *position-of-phrase-in-utterance*, and *number-of-phrases-in-utterance* [Krishna *et. al.* 2004]. For the prediction of Czech duration, *current/previous/next-phone-identities*, *syllable/word/phrase-lengths-in-phones*, *phone-position-in-syllables-from-beginning/end*, *phone-position-in-word-from-beginning/end* and *word-position-in-phrase* are utilized [Batusek 2002]. Lee and Oh use morphological and syntactic features as well as positional attributes for predicting Korean duration [Lee and Oh 1999].

We have been in contact with linguists Prof. Dr. İclal ERGENÇ and Assoc. Prof. Dr. Engin UZUN from Ankara University, Prof. Dr. Güneş MÜFTÜOĞLU from Middle East Technical University and Assoc. Prof. Dr. Engin SEZER from Bilkent University. They state that the most influencing attribute for segmental duration in Turkish is the phonetic context, i.e. the phonetic identity of preceding and following segments, and especially that the next segment has a higher impact on segmental duration. Another influencing attribute for consonant duration is the position of consonant in parent syllable. All three agreed on the fact that contrary to other languages like English or Spanish, Turkish segment durations are not significantly affected by stress. Sezer mentions that long vowels do not appear in open syllables (syllables ending with short vowels). He also states that segments occurring in the last syllable of a word are longer in duration if there is a clear word boundary.

Regarding the attributes selected in previous research on other languages and remarks by Turkish researchers, we select a set of attributes for modeling Turkish duration. Each phone in the database is assigned a feature vector describing the phone and the values of its attributes. The attributes and their values used in this study are presented next.

4.1 Performance Measures

We employ several statistical measures for the quantitative analysis of the durational attributes used in this study. Let us assume that durations in an *n*-dimensional database is represented by the vector $\mathbf{x}=[x_1, x_2, \ldots, x_n]$. Then, the expressions of *Mean*, *Standard Deviation* (SD) and *Coefficient of Variation* (CV) for the corresponding **x** vector are given by **(4-1)** through **(4-3).** In order to reveal duration-segment relations, we mainly rely on CV which is a dimensionless measure. CV is a suitable measure that describes the degree of spread around the mean of the data.

$$Mean(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4-1}$$

$$SD(\mathbf{x}) = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - mean(\mathbf{x}) \right)^2 \right)^{1/2} \tag{4-2}$$

$$CV(\mathbf{x}) = \frac{SD(\mathbf{x})}{Mean(\mathbf{x})} \tag{4-3}$$

4.2 Durational Attributes

Attributes used in phoneme duration modeling are presented in the followinf sections.

4.2.1 Phone Identity

Phone identity (*Phn*) is known to be the most influencing attribute on speech timing as duration is directly related to the characteristics of the phone and its closest neighbors. We use SAMPA convention [Wells 2003] to identify phones in our database. No allophonic variations are used for the vowels and the consonant ‘r’ but allophones of ‘g’, ‘k’, ‘n’, and ‘l’ are used. Long vowels are separated from their short counterparts as well. The total number of symbols used is 43 including *silence*. The lists of symbols and their frequency in the speech corpus are given in **Table 4-1**.

4.2.2 Manner of Articulation

Although manner of articulation for consonants and backness/frontedness for vowels are added as influencing attributes, the feature is not used for modeling purposes. Instead, we use this feature to reveal the relationship between segments and their durations. The values of the feature are: {Affricate, Fricative, Nasal, Liquid, Semivowel, Plosive, Back,

Front, and Silence}. **Table 4-2** shows the phone coverage of each feature value. **Table 4-3** demonstrates the duration distribution of phones with respect to their manner of articulations. As the table reveals, maximum deviation is seen on the duration of the consonant ‘r’ with a CV ratio of 0.797 while the least is observed on the vowel ‘e:’.

Table 4-1: Frequencies of the phones in the speech corpus

Phone	**Frequency**	**Phone**	**Frequency**
a	5790	m	2228
a:	268	n	3627
b	1292	N	156
c	1007	o	1521
d	2142	o:	31
dZ	731	2	493
e	4451	2:	1
e:	94	p	436
f	235	r	3570
g	163	s	1503
G	685	S	747
gj	546	silence	2000
h	459	t	1761
1	2415	tS	547
1:	42	u	1980
i	4378	u:	84
i:	141	v	391
j	1931	w	178
k	1389	y	972
l	1656	y:	14
5	1705	z	757
		Z	133

Table 4-2: Phone clusters with respect to their manner of articulation property

Manner of Articulations	**Phones**
Affricate	tS, dZ
Fricative	f, v, w, s, z, S, Z, h, G
Nasal	m, n, N
Liquid	l, 5, r
Semivowel	j
Plosive	p, b, t, d, c, gj, k, g
Back	a, I, o, u, a:, I:, o:, u:
Front	e, i, 2, y, e:, i:, 2:, y:

Table 4-3: Mean, standard deviation (SD) and standard deviation over mean (CV) for the segments in the database with respect to their manner of articulations (MOA) in decreasing CV ratio.

Segment	MOA of Segment	Mean	SD	CV	Frequency
r	Liquid	43.253	34.493	0.797	2450
1	Back	51.890	29.328	0.565	1730
z	Fricative	70.385	38.231	0.543	512
i	Front	57.414	30.672	0.534	3031
y:	Front	84.300	44.659	0.530	10
u	Back	55.442	25.11	0.453	1356
G	Fricative	34.698	15.311	0.441	470
g	Plosive	55.598	23.558	0.424	112
h	Fricative	53.073	22.475	0.423	330
y	Front	58.492	24.046	0.411	693
l	Liquid	41.635	16.148	0.388	1152
n	Nasal	52.968	20.524	0.387	2522
b	Plosive	48.244	17.809	0.369	902
j	Semivowel	40.582	14.83	0.365	1349
k	Plosive	79.642	28.736	0.361	948
5	Liquid	38.535	13.861	0.360	1154
c	Plosive	80.040	28.399	0.355	696
a	Back	82.300	29.027	0.353	4028
e	Front	79.768	27.59	0.346	3114
v	Fricative	45.822	15.464	0.337	264
gj	Plosive	52.419	17.586	0.335	372
o:	Back	113.250	37.629	0.332	16
d	Plosive	47.706	15.848	0.332	1647
u:	Back	89.574	29.498	0.329	61
t	Plosive	72.472	23.382	0.323	1214
m	Nasal	54.174	17.372	0.321	1594
w	Fricative	40.133	12.874	0.321	120
S	Fricative	101.423	32.363	0.319	523
o	Back	82.013	25.869	0.315	1090
i:	Front	87.451	26.913	0.308	102
2	Front	87.266	25.99	0.298	353
dZ	Affricate	50.225	14.986	0.298	520
f	Fricative	81.830	23.937	0.293	171
N	Nasal	54.874	14.931	0.272	111
p	Plosive	76.919	20.27	0.264	307
1:	Back	90.407	23.352	0.258	27
tS	Affricate	86.013	22.175	0.258	387
Z	Fricative	67.022	15.32	0.229	91
s	Fricative	99.217	22.349	0.225	1069
a:	Back	133.888	26.924	0.201	187
e:	Front	119.191	23.802	0.200	68
2:	Front	63.000	0	0.000	1

4.2.3 Voicing

Since the phone and its characteristics play an important role on the duration mechanism of speech, we consider the effects of voicing of the phones on duration by considering voicing property of the segments. **Table 4-4** reveals the mean durations of the segments' manner of articulation with respect to their voicing property. According to the table, the differences between voiced and voiceless consonants are very significant, in the order of 30-40 ms; voiceless segments are longer in duration then their voiced counterparts.

Table 4-4: Mean, SD and CV for the MOA of the segments with respect to their voicing in decreasing CV ratio.

MOA of Segment	Voicing of Segment	Mean	SD	CV	Frequency
liquid	voiced	41.716	26.947	0.646	4756
fricative	voiced	51.721	30.129	0.583	1457
front	vowel	69.41	31.159	0.449	7372
back	vowel	73.055	32.652	0.447	8495
semivowel	voiced	40.582	14.83	0.365	1349
nasal	voiced	53.473	19.267	0.36	4227
plosive	voiced	48.736	17.112	0.351	3033
plosive	voiceless	76.715	26.183	0.341	3165
fricative	voiceless	91.046	30.64	0.337	2094
affricate	voiced	50.225	14.986	0.298	520
affricate	voiceless	86.013	22.175	0.258	387

4.2.4 Previous/Next Phone Identities

The preceding (*Left*) and following (*Right*) phone identities are used to model segment duration. As the work of Klatt [Campbell 2000; Klatt 1987] revealed that the segmental context highly influences the segment's duration, it is beneficial to use a larger window however when the number of segments are considered it is difficult to find a database covering sufficient representatives for modeling each combination.

4.2.5 Manner of Articulation of Previous/Next Phones

Using the identity of previous/next phones increases the dimension of the search space to the order of 43x43. Since we use a database of limited dimension, the frequency of every possible combination is not at a considerable level. In order to reduce the dimension, we included manner of articulation property for consonants and

backness/frontedness property for vowels of preceding and following phonemes. The attributes are named as *Leftc1* and *Rightc1* for previous and following phones, respectively. Using manner of articulations for the neighboring phones, the dimension of the search space is reduced to 9x9. **Table 4-5** reveals segmental durations with respect to manner of articulations of following phones. The most striking result is segments followed by a silence, i.e. a possible phrase break, have the largest durations. This result agrees with that reported by Klatt about segments just before a pause [Campbell 2000].

Table 4-5: Mean, SD and CV for the voicing of the segments with respect to their right neighbour's manner of articulation in decreasing CV ratio.

Voicing of Segment	**MOA of Right Segment**	**Mean**	**SD**	**CV**	**Frequency**
voiceless	silence	152.174	64.981	0.427	121
voiceless	semivowel	94.529	14.877	0.157	51
voiceless	Nasal	89.077	23.778	0.267	235
voiceless	liquid	86.576	21.32	0.246	406
voiceless	Front	82.631	26.323	0.319	1734
voiceless	Back	81.415	24.834	0.305	2248
voiceless	plosive	74.815	22.466	0.3	655
voiceless	affricate	64.977	15.741	0.242	44
voiceless	fricative	60.908	24.224	0.398	152
voiced	silence	122.484	53.797	0.439	376
voiced	semivowel	57.524	20.691	0.36	170
voiced	affricate	55.587	17.604	0.317	240
voiced	plosive	54.814	20.035	0.366	1697
voiced	fricative	48.745	19.57	0.401	514
voiced	liquid	47.057	19.157	0.407	768
voiced	Nasal	45.776	16.228	0.355	522
voiced	Front	44.496	17.287	0.389	5309
voiced	Back	42.631	16.238	0.381	5746
vowel	silence	134.441	30.447	0.226	513
vowel	Front	80.710	41.187	0.51	214
vowel	affricate	73.062	26.649	0.365	581
vowel	liquid	72.228	29.689	0.411	3571
vowel	fricative	69.732	31.16	0.447	2732
vowel	Back	69.318	36.011	0.519	333
vowel	plosive	69.012	25.152	0.364	3417
vowel	Nasal	66.829	31.607	0.473	3413
vowel	semivowel	62.385	29.781	0.477	1093

4.2.6 Voicing of Previous/Next Phones

Klatt also reported that vowel duration is shortened if it is followed by a voiceless consonant in the same word [Campbell 2000; Klatt 1987]. In order to observe the effect of voicing property of neighboring phones to segmental duration, we used neighboring phones voicing property as well. The name for the attribute holding the voicing property of previous phone is *Leftc2* and for the following phone, it is *Rightc2*. **Table 4-6** shows the statistical results related to the segment duration and voicing property of neighboring phones. According to the data in the table, when followed by a voiced segment, phone duration increases. It should also be mentioned voiced fricative followers influence voiceless phone durations more than voiced plosive and affricate followers. However, the maximum average segmental duration is attained by voiceless segments followed by semivowels. Besides, the effect of a semivowel follower on voiceless segments is highly influential since the CV ratio attains very small values. One important attribute to be mentioned about the data in the table is the insufficient representatives for some of the combinations leading to less information about such data.

4.2.7 Lexical Stress

It is reported that the stress attribute is a relevant feature in duration control [Batusek 2002; Campbell 2000; Cordoba *et. al*. 1999; Cordoba *et. al*. 2002]. Therefore, the effects of stress to segmental duration in Turkish are also considered. The attribute is named as *Accent* while using for model development. There exist two levels for lexical stress: Accented (A) or Not-Accented (NA). A segment is associated with an A if the vowel of the parent syllable is stressed and an NA otherwise.

It is reported in Cordoba *et. al.* (2002) that stressed vowels are 20% longer on average than unstressed vowels. For our database, mean durations for stressed and unstressed segments turned out to be 62.43 ms and 63.47 ms, respectively. **Table 4-7** through **Table 4-9** reveals the statistics related to stressed and unstressed segment durations with respect to segments' features. Last column in **Table 4-7** gives the percentage change of the corresponding vowel when it occurred in a stressed syllable. The former summarizes the mean durations for vowels occurring in stressed/unstressed syllables. There is a tendency of a slight increase in vowel durations occurring in stressed syllables but this is not the case for all vowels. So, a generalization about vowel length with respect to stress attribute cannot be done. Second table reveals the average durations of segments occurring in stressed/unstressed syllables with their voicing property. According to the voicing of the

segments, stressing of a segment does not play an important role in phones' durations. The latter table shows the segments' manner of articulation properties with respect to stressing property. As the table indicates a direct relation with stressing property of the segments and their duration.

Table 4-6: Mean, SD and CV for the voicing of the segments with respect to their right neighbour's manner of articulation in decreasing CV ratio.

Voicing of Segment	**MOA of Right Segment**	**Voicing of Right Segment**	**Mean**	**SD**	**CV**	**Frequency**
voiceless	semivowel	voiced	94.529	14.877	0.157	51
voiceless	nasal	voiced	89.077	23.778	0.267	235
voiceless	liquid	voiced	86.576	21.32	0.246	406
voiceless	fricative	voiced	84.531	24.09	0.285	32
voiceless	plosive	voiced	84.481	26.071	0.309	156
voiceless	front	vowel	82.631	26.323	0.319	1734
voiceless	back	vowel	81.415	24.834	0.305	2248
voiceless	affricate	voiced	76.667	8.733	0.114	6
voiceless	plosive	voiceless	71.794	20.319	0.283	499
voiceless	affricate	voiceless	63.132	15.875	0.251	38
voiceless	fricative	voiceless	54.608	20.083	0.368	120
voiced	semivowel	voiced	57.524	20.691	0.36	170
voiced	affricate	voiced	57.157	16.313	0.285	159
voiced	plosive	voiced	55.610	20.473	0.368	1187
voiced	plosive	voiceless	52.961	18.866	0.356	510
voiced	affricate	voiceless	52.506	19.64	0.374	81
voiced	fricative	voiced	52.298	20.208	0.386	121
voiced	fricative	voiceless	47.651	19.264	0.404	393
voiced	liquid	voiced	47.057	19.157	0.407	768
voiced	nasal	voiced	45.776	16.228	0.355	522
voiced	front	vowel	44.496	17.287	0.389	5309
voiced	back	vowel	42.631	16.238	0.381	5746
vowel	front	vowel	80.710	41.187	0.51	214
vowel	affricate	voiced	74.209	26.743	0.36	339
vowel	liquid	voiced	72.228	29.689	0.411	3571
vowel	fricative	voiced	71.944	34.302	0.477	1272
vowel	affricate	voiceless	71.455	26.488	0.371	242
vowel	plosive	voiceless	69.893	24.709	0.354	2045
vowel	back	vowel	69.318	36.011	0.519	333
vowel	fricative	voiceless	67.805	28.007	0.413	1460
vowel	plosive	voiced	67.698	25.752	0.38	1372
vowel	nasal	voiced	66.829	31.607	0.473	3413
vowel	semivowel	voiced	62.385	29.781	0.477	1093

Table 4-7: Mean and percentage values for Stressed and Unstressed vowels.

Vowel	Lexical Stress	Mean	Frequency	Percentage
a	N	78.754	2861	13.45
a	A	90.994	1167	
a:	N	132.073	151	6.66
a:	A	141.5	36	
e	N	76.618	2090	11.11
e	A	86.196	1024	
e:	N	119.639	61	-3.78
e:	A	115.286	7	
1	N	50.112	1037	8.14
1	A	54.551	693	
1:	N	88.05	20	9.36
1:	A	97.143	7	
i	N	57.421	2022	-0.04
i	A	57.4	1009	
i:	N	87.468	62	-0.05
i:	A	87.425	40	
o	N	81.29	980	8.12
o	A	88.455	110	
2	N	86.915	342	11.48
2	A	98.182	11	
o:	N	113.25	16	-
2:	N	63	1	-
u	N	54.269	1009	7.79
u	A	58.856	347	
u:	N	89.574	61	-
y	N	57.382	516	7.04
y	A	61.729	177	
y:	N	84.3	10	-

Table 4-8: Mean, SD, and CV values for Stressed and Unstressed segments with respect to their voicing. There is no abrupt change in stressed and unstressed segments.

Voicing of Segment	Lexical Stress	Mean	SD	CV	Frequency
voiced	N	47.955	23.704	0.494	10102
voiced	A	46.57	21.274	0.457	5240
voiceless	N	82.722	28.491	0.344	4170
voiceless	A	82.513	28.561	0.346	1476
vowel	N	70.28	31.684	0.451	11239
vowel	A	73.987	32.669	0.442	4628

Table 4-9: Mean, SD, and CV values for Stressed and Unstressed segments with respect to their manner of articulations.

MOA of Segment	Lexical Stress	Mean	SD	CV	Frequency
affricate	N	66.157	25.083	0.379	618
affricate	A	64.08	26.451	0.413	289
fricative	N	76.023	36.058	0.474	2546
fricative	A	72.093	35.918	0.498	1005
nasal	N	54.027	18.47	0.342	2484
nasal	A	52.683	20.329	0.386	1743
plosive	N	63.427	26.349	0.415	4573
plosive	A	61.888	25.936	0.419	1625
liquid	N	42.43	29.238	0.689	3069
liquid	A	40.417	22.129	0.548	1687
semivowel	N	41.221	15.11	0.367	982
semivowel	A	38.872	13.928	0.358	367
back	N	71.831	32.166	0.448	6135
back	A	76.238	33.682	0.442	2360
front	N	68.417	30.996	0.453	5104
front	A	71.646	31.417	0.439	2268

Although the tables and the overall statistics do not reveal an influence on segmental duration with respect to the occurrence in stresses/unstressed syllables as mentioned by Cordoba *et. al.* (2002), we would like to observe this phenomenon in more detail therefore we included this attribute to our feature set.

Lexical stresses [Barker 2002; Demircan 2001; Lees 1961; Sezer 1981; Underhill 1976] of the words in the database are obtained through an automatic stress assignment algorithm developed within the course of this study.

4.2.8 Position in Syllable

We consider phone position in syllable (*PosInSyllable*) as another feature influencing duration in Turkish. A syllable is composed of an *onset* + *rhyme*. Onset is described to be the consonants before the syllable vowel that forms the syllable *nucleus*. Rhyme is the remaining part of the syllable, i.e. it is composed of the nucleus + *coda* where coda is the consonants following the syllable nucleus. In our feature coding we use a three level representation for the segment position in syllable: Nucleus (N), Onset (O) and Coda (C). Decomposition of the syllable '*semt*' in the word '*semtten*' is described in **Figure 4-1**.

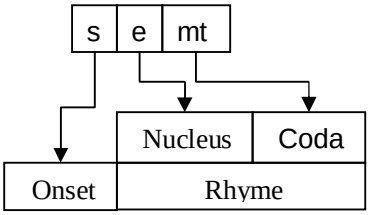

Figure 4-1: Decomposition of the syllable '*semt*' into its *PosInSyllable* tags

By this coding scheme, every vowel is set to N and consonants are set to either O or C. In our database, the average durations are 43.33 ms and 82.285 ms for voiced and voiceless consonants occurring at onset position while the average durations are 57.102 ms and 83.504 ms for voiced and voiceless consonants occurring at coda position. **Table 4-10** and **Table 4-11** show the statistical figures obtained through our database for the phones occurring in various positions in the parent syllable. It can be concluded that segment durations for voiceless consonants are independent of the position in syllable since they are almost the same in either position. Thus, it can be deduced that there is a slight increase on the average duration for the consonants occurring at onset. Besides, voiced consonants are slightly longer when occur in coda position. Close examination of **Table 4-11** reveals that there is a significant difference between the durations of affricates, nasals, plosives and liquids occurring at the onset and coda positions.

Table 4-10: Mean, SD, and CV values of segment duration with respect to Position in Syllable feature. Segments are clustered according to their voicing property.

Voicing of Segment	**Position In Syllable**	**Mean**	**SD**	**CV**	**Frequency**
voiced	O	43.33	16.629	0.384	10717
voiced	C	57.102	31.117	0.545	4625
voiceless	O	82.285	25.726	0.313	3875
voiceless	C	83.504	33.796	0.405	1771
vowel	N	71.362	32.018	0.449	15867

Under the light of the discussions with experts on Turkish and our statistical findings, it is concluded that *PosInSyllable* is an important parameter for consonant duration in Turkish. Statistical information gathered from the database supports this conclusion.

Table 4-11: Mean, SD, and CV values of segment duration with respect to Position in Syllable feature. Segments are clustered according to their manner of articulation.

MOA of Segment	Position In Syllable	Mean	SD	CV	Frequency
affricate	O	62.716	23.605	0.376	818
affricate	C	91.034	28.393	0.312	89
fricative	O	72.785	33.682	0.463	2534
fricative	C	80.206	40.925	0.51	1017
nasal	O	47.548	15.022	0.316	2265
nasal	C	60.313	21.263	0.353	1962
plosive	O	59.465	23.262	0.391	5043
plosive	C	78.56	32.226	0.41	1155
liquid	O	33.981	12.687	0.373	2898
liquid	C	53.781	37.003	0.688	1858
semivowel	O	40.067	15.539	0.388	1034
semivowel	C	42.273	12.085	0.286	315
back	N	73.055	32.652	0.447	8495
front	N	69.41	31.159	0.449	7372

4.2.9 Syllable Type

We also include the type of the parent syllable (*SylType*) in our annotations. Two levels are used to denote syllable types: Heavy (H) and Light (L). Heavy and light syllables are sometimes called *open* and *closed* syllables, respectively. Average segment durations are 67.87 ms for heavy segments and 62.46 ms for light segments. **Table 4-12** and **Table 4-13** reveal the overall segment durations with respect to their parent syllable type. In general, all segments have shorter durations in open syllables then in closed syllables. From the overall view, it can be deduced that syllable type is an influencing attribute in segment duration in Turkish.

Table 4-12: Mean, SD, and CV values of segments in Heavy (H) and Light (L) syllables. Segments are clustered with respect to their voicing property.

Voicing of Segment	Syllable Type	Mean	SD	CV	Frequency
voiced	L	43.073	16.812	0.39	6441
voiced	H	50.673	26.001	0.513	8901
voiceless	L	80.691	26.03	0.323	2099
voiceless	H	83.837	29.818	0.356	3547
vowel	L	65.061	31.297	0.481	9135
vowel	H	79.91	30.988	0.388	6732

Table 4-13: Mean, SD, and CV values for segments in Heavy (H) and Light (L) syllables. Segments are clustered with respect to their manner of articulations.

MOA of Segment	Syllable Type	Mean	SD	CV	Frequency
affricate	L	59.649	22.837	0.383	405
affricate	H	70.211	26.613	0.379	502
fricative	L	70.595	34.103	0.483	1379
fricative	H	77.65	36.991	0.476	2172
nasal	L	47.882	15.065	0.315	1428
nasal	H	56.325	20.514	0.364	2799
plosive	L	58.391	22.861	0.392	3036
plosive	H	67.471	28.434	0.421	3162
liquid	L	33.351	12.49	0.374	1737
liquid	H	46.529	31.479	0.677	3019
semivowel	L	39.115	16.111	0.412	555
semivowel	H	41.607	13.782	0.331	794
back	L	65.012	30.713	0.472	4851
back	H	83.762	32.08	0.383	3644
front	L	65.118	31.949	0.491	4284
front	H	75.365	29.002	0.385	3088

4.2.10 Syllable-Position-in-Word

Like *PosInSyllable* feature, the location of parent syllable in the parent word (*SyllablePosInWord1*) is used by many researchers as an influencing attribute on segmental duration. Klatt reported [Klatt 1987; Campbell 2000] especially that boundary syllable segments are longer in duration. In order to examine the affects of the parent Syllable-Position-in-Word, we tried different coding schemes.

In the first type of coding (*SyllablePosInWord1*), the syllables of the same word are counted from the left starting from 1. The database contains words of at most 10 syllables thus the feature can take at most 10 as value. However, the database lacks words containing 9 syllables. **Table 4-14** reveals the database statistics related to segment durations and parent syllable locations. The table does not provide direct information related to the segment durations of the last syllables but gives an intuition about the shortening in segment durations with increasing number of syllables in parent word.

In the second type of coding (*SyllablePosInWord2*), a discrete set of symbols is used to represent the location of parent syllable in the parent word. The segments of the parent syllable take the value *Initial* if they constitute the first syllable of the parent word, or *Final* if they form the last syllable, or *Middle* otherwise. The segments of the words

containing single syllables are represented by the value Single. With this coding scheme, we have the advantage of differentiating initial and final syllables clearly. However, with this coding, we loose the information relating segment duration with the actual location of parent syllable. To overcome this issue, a third coding scheme is proposed. **Table 4-15** demonstrates the quantitative results obtained from the database with second coding. According to the data in the table, initial and final segments are longer in duration and segment durations in words with single syllables attain the maximum average value.

Table 4-14: Mean, SD, and CV values for segments in syllables with respect to different *Syllable Positions.*

Syllable Position	**Mean**	**SD**	**CV**	**Frequency**
1	69.344	29.833	0.43	12371
2	60.067	28.989	0.483	11462
3	59.405	31.671	0.533	7334
4	59.535	34.235	0.575	3807
5	62.658	37.577	0.6	1423
6	64.306	39.707	0.617	359
7	67.888	35.01	0.516	80
8	66.867	41.834	0.626	15
10	47.5	3.536	0.074	2

According to the third coding scheme (*SyllablePosInWord3)*, we scale the raw syllable positions (*SyllablePosInWord1*) with the total number of syllables in the parent word. The coding can be formulated as follows:

$$SyllablePo\ sInWord\ 3 = \frac{SyllablePo\ sInWord\ 1 - 1}{WordLength\ - 1} \tag{4-4}$$

With this coding, the initial and final syllables are differentiated while preserving syllable position information in polysyllabic words. The initial syllables as well as single syllables take the value 0 and the final syllables take the value 1.

Table 4-15: Mean, SD, and CV values for segments in Initial (I), Middle (M), Final (F) and Single (S) syllables.

Syllable Position	Mean	SD	CV	Frequency
I	68.485	28.369	0.414	10283
M	54.154	25.232	0.466	12800
F	66.456	35.946	0.541	11684
Single	73.578	35.896	0.488	2088

4.2.11 Word Position in Sentence

In our experimental studies, we consider position of parent word in the sentence (*WordNo*) to have an impact on segment duration since it has been reported that increasing the number of words in a sentence results in shorter segments [Klatt 1987; Campbell 2000] The feature values are set to be numeric and ranges from 1 to 19. All segments take the same value in a parent word. **Figure 4-2** demonstrates the distribution of feature values in the database. **Table 4-16** reveals the statistics of the database related to the current coding scheme for *WordNo* feature. It can be observed that there is an increase in average segment duration around the 9^{th} word in the sentences. However, we think that this increase is not related to the number of words but a possible sentence final.

Table 4-16: Mean values for segments of words in different locations in the utterance.

WordNo1	Mean	Frequency	WordNo1	Mean	Frequency
1	61.607	4187	10	69.771	984
2	62.771	4592	11	68.689	283
3	63.649	4482	12	69.704	142
4	62.205	4545	13	60.171	35
5	62.296	4582	14	69.902	41
6	61.644	4247	15	86.929	14
7	63.707	3939	16	61.471	17
8	63.627	2872	17	71.385	13
9	67.267	1863	18	56.25	8
			19	52.222	9

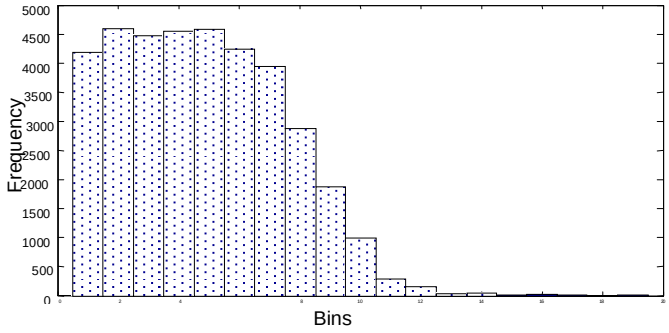

Figure 4-2: Histogram plot of Word Position in Utterance feature

We used the same coding scheme for *WordNo* feature as in *SyllablePosInWord* feature to observe the effects of initial and final words in segment duration. The attribute values for the second representation are as follows: The segments of the parent syllable attain a value I for sentence initial words, a value F for the words at the end of sentences and a value M, elsewhere. The tag *Single* is discarded from the set since there is no single word sentence in our database. The statistics with the new coding scheme is given in **Table 4-17**. The table reveals the lengthening effect of sentence finality on the segment duration.

4.2.12 Word Part of Speech

In numerous studies, Part-of-Speech (POS) tags are used to observe effects on duration. The attribute's name is *WordPOS*. We also employ POS tags for parent word in our attribute space. The segments are annotated with their major POS tags as being NOUN, PRON, VERB, QUES, INF, POSTP, CONJ, ADV, ADJ, CNOUN, or EXC. These tags are obtained through a morphological analysis procedure [Oflazer 1994]. **Table 4-18** shows the average segment durations with respect to different POS tags in the database. According to the table, question words reveal one of the largest average duration. Our hypothesis is that question words mainly locates at phrase ends and the maxima occurred in average segmental duration in question words is nothing but a clause-final lengthening as reported by Klatt [Klatt 1987; Campbell 2000].

Table 4-17: Mean, SD, and CV values in Initial (I), Middle (M) and Final (F) Words.

WordNo2	Mean	SD	CV	Frequency
I	61.607	28.38	0.461	4187
M	61.452	29.083	0.473	27567
F	73.612	40.962	0.556	5101

Table 4-18: Mean, SD, and CV values for segment durations according to POS values of tags of the parent word.

Part of Speech	Mean	SD	CV	Frequency
CONJ	79.992	38.295	0.479	622
QUES	74.417	43.025	0.578	410
VERB	65.982	35.04	0.531	8105
ADV	65.96	30.643	0.465	2075
POSTP	65.771	30.765	0.468	682
PRON	63.253	30.542	0.483	716
INF	62.039	23.314	0.376	256
NOUN	61.872	29.428	0.476	17411
ADJ	59.678	28.688	0.481	6406
CNOUN	58.138	24.672	0.424	167
EXC	56.6	22.865	0.404	5

4.2.13 Word Length

Previous studies indicates that segment duration is directly related to the number of syllables in a word; increase in the number of syllables results in a squeezing in segment duration. The number of syllables in the parent word is represented by the name *NumOfSyl*. Therefore, we also include the number of syllables in the parent word as an influencing attribute. The attribute values are numeric and ranges from 1 to 10. **Figure 4-3** reveals the histogram of *NumOfSyl* of the database. **Table 4-19** gives the database statistics related to segment durations and *NumOfSyl* attribute. From the table, it can be inferred that as the number of syllables of a word increase, the average segment duration is shortened. Thus, there is an inverse proportion between the number of syllables in the parent word and average segment duration.

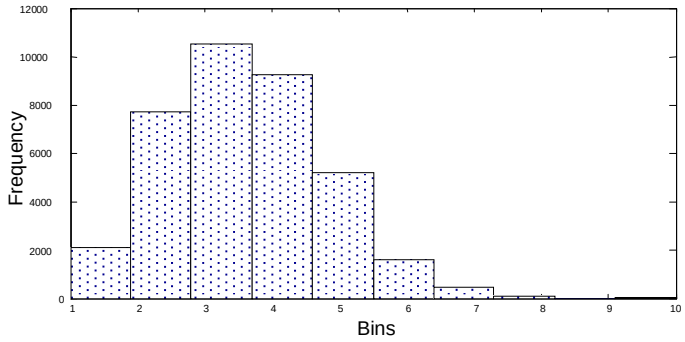

Figure 4-3: Histogram plot of Word Length.

Table 4-19: Mean, SD, and CV values for segment durations according to *Word Length*.

Word Length	**Mean**	**SD**	**CV**	**Frequency**
1	73.578	35.896	0.488	2088
2	67.187	31.443	0.468	7704
3	64.259	30.968	0.482	10496
4	60.521	30.234	0.5	9237
5	58.436	29.819	0.51	5182
6	56.518	29.579	0.523	1601
7	53.929	27.912	0.518	434
8	53.967	26.733	0.495	91
10	52.182	21.456	0.411	22

4.2.14 Total Number of Words in Utterance

We also employ the total number of words in the utterance (*NumOfWord*) as a separate attribute. The attribute values are in the range [3-19]. *WordNo* attribute is related to *NumOfWord* such that the last word of each utterance attains the same value assigned to the current attribute. **Figure 4-4** and **Table 4-20** show the histogram and mean segment duration for the attribute values. As the table reveals, the segment durations do not show a characteristic change with respect to the number of words in the sentence. The figure and the table also show the utterance - number of word situation of our database.

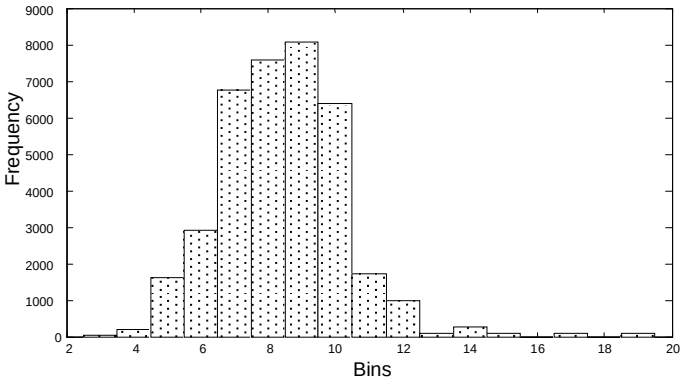

Figure 4-4: Histogram plot of Total number of Words in Utterance.

Table 4-20: Mean, SD, and CV according to Total Number of Words in Utterance.

NumOfWord	Mean	SD	CV	Frequency
3	62.119	29.937	0.482	42
4	64.065	30.343	0.474	185
5	64.418	33.482	0.52	1606
6	63.285	32.221	0.509	2907
7	63.789	32.17	0.504	6753
8	62.535	30.521	0.488	7576
9	63	30.713	0.488	8058
10	62.798	31.11	0.495	6373
11	63.686	30.432	0.478	1715
12	62.681	28.878	0.461	994
13	64.037	28.211	0.441	80
14	65.419	33.008	0.505	270
15	62.283	33.791	0.543	92
17	65.143	30.896	0.474	105
19	63.939	27.035	0.423	99

4.2.15 Syllable Position in Utterance

The position of the parent syllable in the utterance (*SylNo*) has been considered to investigate the effects on segment duration. It has been reported that increasing number of syllables affect English timing to be shorter. The feature levels have numerical values ranging from 1 to 45. The statistics of the database used for modeling are demonstrated in **Figure 4-5** and **Table 4-21**. From the table, it can also be concluded that the segments having larger *SylNo* values are longer in duration then the rest of the segments. The main

reason for this is the higher the value of *SylNo* value, the more probable that the syllable is an utterance final syllable in confirmation with Klatt's rule about phrase-final lengthening in segment duration.

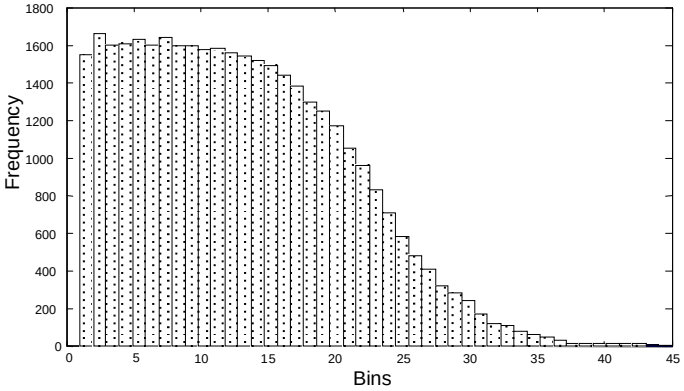

Figure 4-5: Histogram plot of Syllable Position in Utterance.

Attributes below are obtained through the speech corpus identified perceptually or by direct extraction.

4.2.16 Phrase Break Information

Speech corpus has been evaluated perceptually several times and the major perceptual breaks in the utterances were marked manually. The marks mainly correspond to the speaker's breathings however some correspond to lengthening in the segment durations causing perceptual differences in the utterance. Therefore, we doubt that the current speaker is not efficient to make predictions on the pause durations. For phrase break information (*PhrInfo*), we have only used three categorical levels. Segment takes a Phrase Initial (PI) value if it immediately follows a phrase break, a Phrase medial (PM) value if there is no phrase break engagement and a Phrase Final (PF) if a phrase break immediately follows the segment. **Table 4-22** reveals the average segment durations and their standard deviations for the three categorical values. According to the table, it can be deduced that our speaker has a tendency to lengthen segment durations at phrase boundaries. The extension is more in phrase final segments.

Table 4-21: Mean values according to *Syllable Position in Utterance*.

SylNo	Mean	Frequency	SylNo	Mean	Frequency	SylNo	Mean	Frequency
1	64.18	1551	16	61.625	1441	31	70.418	170
2	62.59	1660	17	63.3	1383	32	63.084	119
3	63.991	1601	18	61.932	1298	33	65.121	107
4	63.404	1610	19	62.726	1248	34	66.908	76
5	62.671	1630	20	63.477	1170	35	70.237	59
6	62.887	1601	21	61.93	1050	36	73.5	44
7	62.521	1640	22	64.469	961	37	75.071	28
8	63.328	1599	23	66.438	831	38	76.133	15
9	64.084	1598	24	65.018	708	39	59.6	15
10	62.86	1575	25	65.653	580	40	59.933	15
11	60.023	1579	26	63.761	481	41	62.091	11
12	62.984	1561	27	64.73	408	42	53.5	12
13	61.534	1541	28	63.322	320	43	81.231	13
14	61.636	1521	29	68.158	284	44	97.714	7
15	62.484	1490	30	71.627	241	45	52.333	3

Table 4-22: Mean, SD, and CV values according to *Phrase Break*.

Phrase Break	Mean	SD	CV	Frequency
I	69.447	30.606	0.441	4218
M	58.806	26.498	0.451	28011
F	83.734	45.604	0.545	4626

4.2.17 Number of Words from (to) the Preceding (Following) Phrase Break

When we examine the *WordNo* feature, we observed that there is an increase in the average segment duration around the 9th word, a possible utterance final word. Therefore, it can be deduced that the word position in the utterance has an impact on segment duration. Since the segmental duration is affected by word location, we would like to examine the influence of word position in the perceived phrases. Current attributes identify the number of words between the parent word and the preceding (following) phrase break counting from 0. The attributes are named as *NumOfWordFromPrevBr* and *NumOfWordToFolBr* for the corresponding attributes, respectively. **Table 4-23** and **Figure 4-6** show the mean segment durations and histogram of the current attributes. As seen from the left table, as the number of words from the preceding phrase break increases, the average segment duration also increases since the probability of encountering a new phrase break increases. For the right panel of the table, the maximum

average segment duration has occurred at 0^{th} level meaning that the word is immediately followed by a phrase break.

Table 4-23: Mean values for segment durations according to Number of Words from the Preceding Phrase Break (Left) and Number of Words to the Following Phrase Break (Right).

NumOfWord FromPrevBr	Mean	Frequency	NumOfWord ToFolBr	Mean	Frequency
0	62.921	11753	0	68.83	13740
1	62.032	11052	1	58.908	10759
2	63.741	7694	2	59.808	6852
3	64.008	3929	3	61.708	3298
4	65.769	1502	4	60.31	1400
5	65.769	577	5	61.72	511
6	66.671	255	6	63.173	208
7	72.4	75	7	63.333	75
8	74.667	18	8	67.083	12

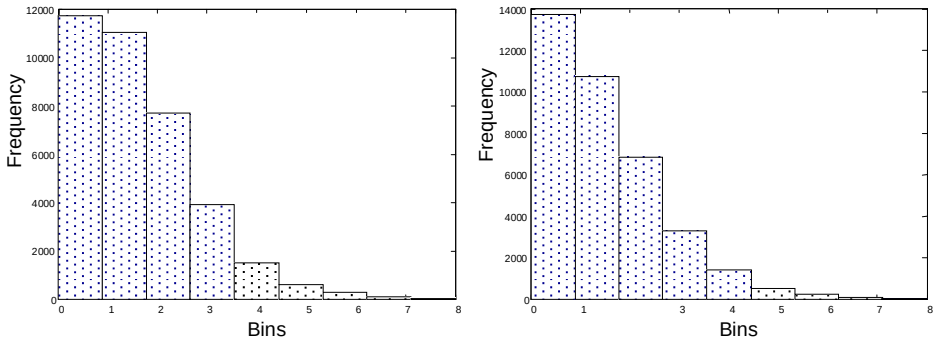

Figure 4-6: Histogram plots of *Number of Words from the Preceding Phrase* (Left) and *Number of Words to the Following Phrase Break* (Right).

4.2.18 Number of Syllables from (to) the Preceding (Following) Phrase Break

This attribute is almost the same as the number of words from the preceding phrase break attribute. Instead of using word numbers, this attribute uses syllable counts. The names for the attributes are *NumOfSylFromPrevBr* and *NumOfSylToFolBr*, respectively. The values range from 0to 27. **Table 4-24** and **Table 4-25**, and **Figure 4-7** show the

mean segment durations and histogram of the current attributes. From **Table 4-24**, it can be observed that there is a maximum at the first entry, then a sharp decrease of approximately 9 ms and a regular increase starting around 15th syllable in segmental durations. The first maxima is related to lengthening phenomenon in phrase initial syllables and as the syllables tend to approach phrase finality the average segment durations increase. For the **Table 4-25**, we have a maximum at the 0th level. This lengthening in segmental duration is due to the fact that the syllable that the segment belongs to is located at phrase final.

Table 4-24: Mean values for segment durations according to Number of Syllables from the Preceding Phrase Break (Left) and Number of Syllables to the Following Phrase Break (Right).

NumOfSyl FromPrevBr	**Mean**	**Frequency**	**NumOfSyl FromPrevBr**	**Mean**	**Frequency**
0	69.447	4218	14	60.373	365
1	61.592	4510	15	66.786	295
2	62.162	4182	16	65.224	196
3	61.777	3981	17	68.232	151
4	61.053	3628	18	65.636	99
5	61.051	3202	19	70.463	82
6	62.577	2841	20	66.111	45
7	62.949	2356	21	69.842	38
8	63.912	1880	22	73	32
9	62.959	1471	23	58.08	25
10	62.838	1148	24	78.214	14
11	64.481	896	25	59.857	7
12	64.813	673	26	80.714	7
13	63.878	509	27	101.75	4

4.2.19 Duration

Phoneme durations are measured in miliseconds. Raw durations are extracted from the text files containing segmentation information. Alignment of speech files with corresponding orthography has been achieved via embedded training using HTK toolkit [University of Cambridge 2005].

The raw duration distribution is given in **Figure 4-8**. In this figure, several fits to raw duration distribution such as *Normal* distribution, *Gamma* distribution and *Inverse*

Gaussian distribution are also given. Phoneme duration distribution is more likely to have a Gamma distribution.

Table 4-25: Mean values for segment durations according to Number of Syllables from Preceding Phrase Break (Left) and Number of Syllables to Following Phrase Break (Right).

NumOfSyl ToFolBr	**Mean**	**Frequency**	**NumOfSyl ToFolBr**	**Mean**	**Frequency**
0	83.755	4632	14	60.094	372
1	59.95	4451	15	60.8	285
2	60.532	4149	16	62.075	200
3	60.133	3909	17	66.597	144
4	59.223	3594	18	61.816	98
5	59.076	3149	19	62.946	74
6	59.908	2795	20	59.882	51
7	60.385	2283	21	65.108	37
8	60.238	1862	22	57.75	28
9	60.428	1469	23	72.783	23
10	61.429	1166	24	77	14
11	62.239	887	25	60.857	7
12	60.55	664	26	44.167	6
13	62.133	503	27	58.333	3

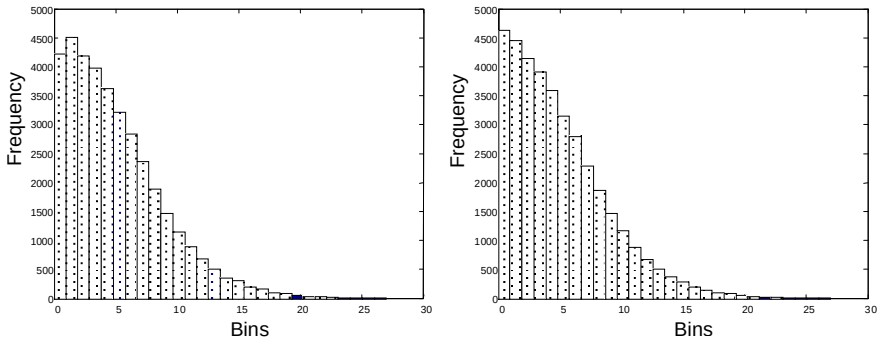

Figure 4-7: Histogram plot of Number of Syllables from the Preceding Phrase Break (Left) and Number of Syllables to the Following Phrase Break (Right).

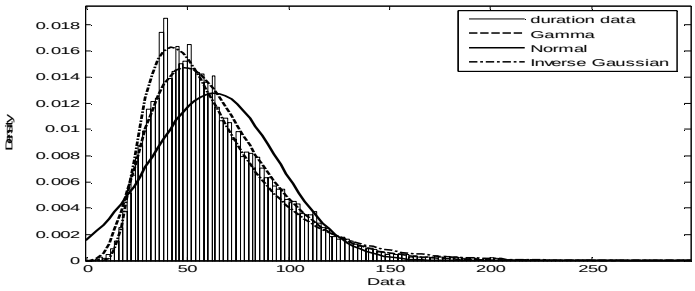

Figure 4-8: Gamma, Normal and Inverse Gaussian and phoneme duration distributions.

CHAPTER 5

DEVELOPING PHONEME DURATION MODELS

Generally three prosody components are modeled: Intonation, duration and intensity [Batusek 2002]. Phoneme durations are part of the prosody and contain important cues for understanding the spoken text [Riedi 1998]. Variations in duration provide assistance for the listener to understand the meaning [Campbell 2000]. Different representational factors specify and modify several aspects of speech during speech production (Klatt 1987).

In our studies, a corpus-based approach is considered to model phoneme duration in Turkish. To this aim, as presented in Chapter 3, a phonetically and prosodically balanced text corpus is designed and corresponding speech corpus is generated through a careful recording procedure. In Chapter 4, durational attributes used in phoneme duration modeling process are introduced. This chapter addresses phoneme duration modeling studies.

5.1 Duration Modeling Using Decision Trees

In duration modeling studies, a hierarchical framework is followed, i.e. attribute combinations are successively analyzed, and best attribute set is obtained in a greedy manner. Experiments are performed mostly with the REPTree algorithm of WEKA. For some of the experiments, the results obtained with the M5P algorithm of WEKA are presented to have a better understanding of the models developed. Both REPTree and M5P algorithms are used for building decision trees. The reason why we preferred decision tree based algorithms is explained in Chapter 4. As mentioned in Chapter 4, both decision tree based algorithms yield better performance than other machine learning algorithms for phoneme duration prediction.

All phoneme types except *silence* are used in training. Most of the phoneme duration prediction studies develop models for vowels and consonants separately. They split database into two subsets, vowel subset and consonant subset, then training is performed on each subset to predict vowel and consonant durations. We study predicting vowel and consonant duration one at a time.

Prediction performance of each experiment is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Correlation Coefficient (CC). It is better to predict the performance of a model on new data (test data) rather than old data (training data). In our studies, the whole database, which consists of 36855 phonemes, is split into two subsets: training set and test set. The test set consists approximately 19.9% of the database and the remaining phonemes constitute the training set (80.1%). The total number of instances in the training and test sets is 29527 and 7328, respectively.

5.2 Experimental Work

In the original database, there are 17 attributes, 16 predictors and 1 dependent attribute, *duration*. **Table 5-1** demonstrates the attribute and value pairs. The first column of the table shows the indices of the attributes used in the experiments. *Phoneme Identity* is considered as the discriminating attribute; hence it is included in all of the experiments. *Left/Right* (23) attribute is considered to form a single pair in order to take into account context. In order to evaluate relative importance of each attribute, models using all possible attribute combinations are to be developed. For a set of *N* attributes the total number of combinations is 2^N -1. Discarding *Phoneme Identity* and considering Left/Right attributes as a single attribute, the dimension of duration attribute set is reduced to 15. Thus $2^{15} - 1 = 32767$ experiments have to be performed to uncover the relation between the phoneme durations and the chosen attributes. However, this is not a feasible value. Therefore, experiments are performed considering limited number of attribute combinations and making generalizations through the results. To this aim, each of 15 attributes are paired with *Phoneme Identity* and used for dphoneme duration modeling at first stage. At the second stage, two of the 15 attributes are combined with *Phoneme Identity* and used in training. This procedure is repeated untill fifth stage. At each stage, the number of attributes combined with *Phone Identity* is increased by 1. When all five experiments are reviewed, it is observed that best performances of different attribute combinations progress one step further. For example, among the experiments

performed with two attribute combinations, *1, 2-3, 6* resulted in the minimum RMSE. When the experiments involving three factor combinations are examined, the minimum RMSE is obtained with *1, 2-3, 6, 12*. Reviewing the two results, it is observed that two of the three factors that resulted in the best performance using three factors are already present in the two-attribute combination that yielded the best RMSE. Considering this performance shift, it is concluded that the best six can be found using the best five and so on. Last stage involves all 17 attributes in training. This procedure is known as *Forward Selection*. Forward selection is used in many machine learning applications for selecting the best attribute set. Following section presents best prediction error performances and related discussions.

Table 5-1: Attribute-Value pairs in the original database.

Index	Attribute	Value
1	Phoneme Identity	42 phones (42 levels)
2-3	Left/Right	43 phonemes (43 x 43 levels)
4	Accent	N, A (2 levels)
5	PosInSyllable	N, C, O (3 levels)
6	SylType	H, L (2 levels)
7	SylNoInWord	Numeric
8	WordNo	Numeric
9	WordPOS	N, P, V, Q, I, T, C, A, J, B, E (11 levels)
10	NumOfSyl	Numeric
11	NumOfWord	Numeric
12	PhrInfo	I, M, F
13	NumOfWordFromPreBr	Numeric
14	NumOfWordToFolBr	Numeric
15	SylNo	Numeric
16	NumOfSylFromPreBr	Numeric
17	NumOfSylToFolBr	Numeric
18	Duration(ms)	Numeric

5.2.1 Forward Selection of Duration Attributes

Each attribute described in Chapter 5 is evaluated on its own to observe the individual affects on phoneme duration. *Phoneme Identity* is considered to be the discriminating attribute; hence corresponding results are used as a reference (baseline) for the rest of the experiments. Individual performances of the attributes in terms of CC, MAE and RMSE are given in **Table 5-2**. As illustrated in the table, *Phoneme Identity* (1) is the best predictor of all attributes. Contextual attributes (2-3) turn out to be the second best

predictors. Third best predictor is *PosInSyllable*. Worst predictor performance given at the bottom of the table corresponds to *NumOfWord* attribute (11). Best CC, MAE and RMSE obtained are 0.5958, 18.2003, and 25.7872, respectively.

Table 5-2: Individual performances of attributes for predicting phoneme durations. Results are given in increasing RMSE order.

Index	CC	MAE (ms)	RMSE (ms)
1	0.5958	18.2003	25.7872
2-3	0.53	20.8325	27.1914
5	0.3106	23.3704	30.5724
12	0.2641	24.3414	30.9329
17	0.2443	24.5178	31.0977
6	0.1473	24.4769	31.7265
14	0.1381	24.8184	31.7601
10	0.1212	24.5606	31.8327
7	0.1218	24.4285	31.8344
9	0.0873	24.7954	31.9577
16	0.0713	24.6631	31.9872
8	0.0539	24.7744	32.0196
15	0.0386	24.7759	32.0445
13	0.0234	24.7784	32.0597
4	0.0193	24.7751	32.0604
11	0	24.7806	32.0658

Table 5-2 illustrates the individual impacts of attributes on phoneme duration however it does not present their combinatorial affect. Attributes that seem to predict phoneme durations individually may fail to perform well when used in combination with other attributes. Combinatorial affects of durational attributes are presented in the subsequent paragraphs.

Table 5-3 shows the regression tree obtained incorporating only best predictor, *Phoneme Identity*. As shown in the table, the tree is split into the values of *Phoneme Identity*. Estimated phoneme durations are the average durations for each split. The numbers in brackets are "(coverage in the training set/errors in the training set)" and "[coverage in the pruning set/errors in the pruning set]". Because there may be fractional instances (i.e. instances with weight < 1) the numbers are not necessarily integers.

Table 5-3: Resulting regression tree using *Phoneme Identit*y attribute only.

```
Phoneme = 2 : 87.55 (190/691.95) [95/654.65]
Phoneme = z : 70.83 (288/1578.03) [110/1401.13]
Phoneme = gj : 52.96 (219/275.32) [77/381.84]
Phoneme = y : 58.29 (368/604.41) [203/548.51]
Phoneme = r : 43.31 (1345/1126.61) [642/1217.83]
Phoneme = e : 79.19 (1647/755.94) [816/785.64]
Phoneme = b : 49.19 (484/322.79) [238/331.99]
Phoneme = n : 52.15 (1358/409.46) [637/371.6]
Phoneme = i : 57.36 (1612/928.38) [802/950.19]
Phoneme = c : 81.02 (377/714.41) [179/959.8]
Phoneme = l : 41.58 (584/215.13) [334/308.08]
Phoneme = m : 53.99 (847/269.03) [440/305.98]
Phoneme = s : 98.36 (546/529.65) [313/524.36]
Phoneme = j : 41.63 (735/210.33) [354/255.33]
Phoneme = d : 48.91 (848/237.58) [486/226.71]
Phoneme = o: : 116.31 (9/311.8) [4/4258.92]
Phoneme = g : 57.95 (50/361.39) [34/926.46]
Phoneme = a : 81.57 (2122/847.94) [1101/843.88]
Phoneme = 5 : 39.53 (608/172.32) [331/219.96]
Phoneme = 1 : 51.87 (941/827.81) [465/804.55]
Phoneme = o : 81.59 (566/660.83) [312/724]
Phoneme = k : 81.09 (505/877.47) [248/741.05]
Phoneme = t : 73.12 (660/540.29) [307/641.85]
Phoneme = tS : 86.88 (204/526.35) [97/430.16]
Phoneme = S : 100.4 (292/1053.54) [143/658.95]
Phoneme = v : 45.89 (142/262.75) [66/201.85]
Phoneme = G : 35.93 (260/249.37) [119/275.03]
Phoneme = u : 55.53 (741/616.94) [360/682]
Phoneme = a: : 133.89 (114/690.86) [41/1025.36]
Phoneme = f : 81.21 (86/449.4) [39/498.95]
Phoneme = w : 40.56 (62/148.42) [34/169.39]
Phoneme = dZ : 52.01 (273/250.07) [134/157.02]
Phoneme = p : 77.59 (174/392) [77/401.49]
Phoneme = i: : 87.61 (48/706.27) [26/775.95]
Phoneme = h : 51.93 (183/438.83) [74/566.99]
Phoneme = N : 55.17 (53/238.68) [42/245.01]
Phoneme = e: : 118.1 (35/421.53) [19/728.05]
Phoneme = u: : 88.82 (34/967.03) [17/737.49]
Phoneme = Z : 68.58 (55/153.56) [17/457.41]
Phoneme = 1: : 89.79 (12/534.17) [7/335.14]
Phoneme = y: : 82.12 (6/2281) [3/2006.67]
Phoneme = 2: : 63 (1/0) [0/0]
```

To obtain best error performances, every possible attribute combination together with *Phoneme Identity* is used to model phoneme durations. The number of attributes is increased by 1 at each stage. At the k^{th} stage, k+1 attributes are used and $\binom{N}{k}$ experiments are performed to obtain the best error performance. Best error performances

obtained at each stage is given in **Table 5-4**. Columns of the table correspond to CC, MAE and RMSE, respectively.

Table 5-4: Best prediction error performances obtained with forward selection.

Best Results			
Attributes	**CC**	**MAE (ms)**	**RMSE (ms)**
1, 2-3	0.7576	15.1605	20.9321
1, 2-3, 6	0.7706	14.7089	20.44
1, 2-3, 6, 12	0.7744	14.6039	20.2937
1, 2-3, 6, 9, 12	0.7772	14.5887	20.184
1, 2-3, 4, 6, 9, 12	0.7798	14.5613	20.0792
1, 2-3, 4, 6, 9, 12, 14	0.7806	14.5574	20.0456
1, 2-3, 4, 6, 9, 12, 14, 7	0.7807	14.5607	20.0478
All	0.7718	14.6678	20.4236

After the fifth stage, the number of experiments, so the time to conduct the experiments is increased. Therefore, at the sixth stage, the best set obtained in the fifth stage is used as the base for forthcoming experiments. First five experiments showed that attribute combination that resulted in the best error performances is encountered in the larger dimensional attribute set that result in the best error performance. So, in order to find the six attributes that gives the maximum CC and the minimum RMSE, five attributes, 1, 2-3, 4, 6, 9, and 12 that yield best performance are used. Every other attribute is combined with the best-five to obtain best-six set. Stages for $k \geq 6$ are performed using the same framework.

At the seventh stage, it is observed that RMSE obtained with six attribute analysis is 0.01% better than best RMSE obtained at seventh stage. So, including further attributes to the best six attribute set does not improve the error performances further.

In order to observe the total effects of all the attributes on phoneme duration, a last experiment involving all attributes is conducted. Error performances of all attributes are worse than the best-six attribute. Considering all the results, it is concluded that *Phoneme Identity*, *Left/Right*, *Accent*, *SylType*, *WordPOS*, *PhrInfo*, and *NumOfWordToFolBr* (1, 2-3, 4, 6, 9, 12, and 14) constitutes optimum attribute set for phoneme duration modeling.

5.3 Performance Improvements

In the previous section, the experiments performed using various numbers of attributes to predict phoneme durations are presented. With the set of attributes (1, 2-3, 4, 6, 9, 12, and 14), an RMSE of 20.0456 ms at best is obtained. When all attributes are used, RMSE and CC is becomes 20.4236 ms and 0.77, respectively.

5.3.1 Attribute Modification

This section describes the modifications on the original attribute set for possible improvements on the model performance. The modifications given in the subsequent sections include utilization of phonetic class (Manner of Articulation) instead of SAMPA transcriptions [Wells 2003] for neighboring phonemes (*Left/Right*) and utilization of SylPosInWord2 and SylPosInWord3 attributes instead of SylPosInWord1 attribute.

5.3.1.1 Phonetic Class Instead of SAMPA Transcriptions

In Chapter 4, the effects of *Left/Right* context on phoneme duration is discussed. It is verified that phoneme duration is highly correlated with *Manner of Articulation* and *Voicing* property of *Left/Right* phonemes as well as the phoneme itself. According to the results of previous sections, *Left/Right* attribute turn out to be the most effective attribute on phoneme duration. However, there is a drawback of using Left/Right attributes as they are. As discussed in Chapter 5, each Left/Right attribute consists of 42 SAMPA characters [Wells 2003] plus the *silence*, so a total of 43 values. When the two attributes are considered together, their span is 43x43 = 1849 different *Left-Right* pairs. When the phoneme itself is also considered, the search space is of size 43x42x43 = 77658. Let us consider the database used in our experiments. Our database contains a total of 36855 instances/phonemes, 29527 of which are used for training and the remaining is used for testing. The total number of instances is far below the total number of possible triphone combinations. So, even if we assume that our database contains distinct entries for *Left-Phoneme-Right* trio, the database is still beyond the limits of being sufficient. Besides, the database contains multiple entries. So, available data is insufficient to represent all possibilities.

A reasonable choice for dimension reduction can be the utilization of *manner–of-articulations* of the phonemes instead of their phonetic identities. As the identity of a phoneme is the discriminating factor for its duration, we use manner-of-articulation for left and right phonemes only. Values for the *manner–of-articulation* attribute are set as

follows: {Affricate, Fricative, Nasal, Liquid, Semivowel, Plosive, Back, Front, and Silence}. With this modification, the number of possible triphones is reduced by approximately 95.6% (9x42x9 = 3402).

Table 5-5 shows the attribute and value pairs in the revised database. The first column of the table shows indices of the attributes used in the experiments. The index for *LeftC/RightC* is named as 1920 since they come as a new pair of attributes. The experimental results using manner-of-articulations instead of phonetic identity and all other attributes are given in **Table 5-6**. When the prediction performances of the two experiments (last row of **Table 5-4** and **Table 5-6**) are compared, it is observed that although the information content of the Left/Right attributes is reduced, a slight improvement (approximately 3%) is achieved in the RMSE value.

Table 5-5: Attribute-Value pairs in the modified database.

Index	**Attribute**	**Value**
1	Phoneme Identity	42 phones (42 levels)
1920	LeftC/RightC	Affricate, Fricative, Nasal, Plosive, Back, Front, Semivowel, Liquid, Silence (9 x 9 levels)
4	Accent	N, A (2 levels)
5	PosInSyllable	N, C, O (3 levels)
6	SylType	H, L (2 levels)
7	SylNoInWord	Numeric
8	WordNo	Numeric
9	WordPOS	N, P, V, Q, I, T, C, A, J, B, E (11 levels)
10	NumOfSyl	Numeric
11	NumOfWord	Numeric
12	PhrInfo	I, M, F
13	NumOfWordFromPreBr	Numeric
14	NumOfWordToFolBr	Numeric
15	SylNo	Numeric
16	NumOfSylFromPreBr	Numeric
17	NumOfSylToFolBr	Numeric
18	Duration(ms)	Numeric

Table 5-6: Prediction performance obtained using all attributes with MOAs.

CC	**MAE (ms)**	**RMSE (ms)**
0.79	14.47	19.81

5.3.1.1.1 Modification of Phonetic Class (1)

Several modifications on the revised attribute set are considered. Obtaining better results led us make new arrangements on the attributes *leftC/rightC*. As discussed in Chapter 4, right context plays a very crucial role on phoneme duration. However, the effects of right context within the same syllable are not considered up to now. So, in order to reveal the situation about the effects of right context on phoneme duration within the same syllable, statistical analyses are carried out on the dataset. The mean, SD, CV and frequency of every occurrence in the dataset are given in **Table 5-7**, **Table 5-8**, and **Table 5-9**. The first two columns of the tables give the right neighbor's characteristics while the top most entry is the characteristics of the phoneme. For example, in the first table, the first entry states that Voiceless phonemes followed by Voiceless Fricatives have an average duration of 90.667ms and a standard deviation of 30.271ms. Some combinations in the tables may seem unrealistic but close examination of the dataset reveals that they exist. When the three tables are examined, it can be concluded that the phoneme durations vary abruptly according to the following phonemes' manner–of-articulation within the same syllable. Besides, it can be deduced that with the same right context, different phone classes have different durations.

Table 5-7: Mean, SD, CV and frequencies of the voiceless phones according to the *Manner of Articulation* and *Voicing* property of their *Right* neighbour in the same syllable.

		Voiceless			
RightC	**RightV**	**Mean**	**SD**	**CV**	**Frequency**
Fricative	Voiceless	90.667	30.271	0.334	3
Plosive	Voiceless	78.444	24.656	0.314	18
Liquid	Voiced	79.757	30.386	0.381	37
Back	Vowel	81.579	25.035	0.307	2178
Front	Vowel	83.18	26.478	0.318	1653

Manner of Articulation of the following phoneme is enriched by adding some new values. To this aim, statistically significant phonemes of each class are found and added to the value inventory of *RightC*. Statistical significance of a phoneme is decided upon its CV value. The smaller the CV value, the more significant the phoneme is. So, for each *Manner of Articulation* class, most significant phoneme is selected as a new candidate for

additional attribute value. The total number of additional attribute values is selected to be 7 since *Semivowel* class contains only one type of phoneme ('j').

Table 5-8: Mean, SD, CV and frequencies of the voiced phones according to the *Manner of Articulation* and *Voicing* property of their *Right* neighbour in the same syllable.

		Voiced			
RightC	**RightV**	**Mean**	**SD**	**CV**	**Frequency**
Affricate	Voiceless	50.1	22.679	0.453	10
Fricative	Voiceless	39.368	5.036	0.128	19
Fricative	Voiced	53	0	0	1
Plosive	Voiceless	36.07	11.583	0.321	57
Plosive	Voiced	57.667	8.386	0.145	3
Liquid	Voiced	62.429	27.724	0.444	7
Nasal	Voiced	33.667	11.846	0.352	3
Back	Vowel	42.435	16.131	0.38	5562
Front	Vowel	44.271	17.07	0.386	5148

Table 5-9: Mean, SD, CV and frequencies of the vowels according to the *Manner of Articulation* and *Voicing* property of their *Right* neighbour in the same syllable.

		Vowel			
RightC	**RightV**	**Mean**	**SD**	**CV**	**Frequency**
Affricate	Voiceless	75.132	23.416	0.312	68
Affricate	Voiced	81.364	15.062	0.185	11
Fricative	Voiceless	68.908	28.622	0.415	1563
Fricative	Voiced	92.97	31.605	0.34	432
Plosive	Voiceless	72.207	23.707	0.328	1040
Plosive	Voiced	79.093	23.637	0.299	43
Liquid	Voiced	82.492	28.548	0.346	1858
Nasal	Voiced	74.896	32.361	0.432	1959
Semivowel	Voiced	81.81	25.378	0.31	315

Table 5-10 presents the attributes used for phoneme duration modeling with the recent modifications on the *RightC* attribute. The corresponding prediction performances are given in **Table 5-11**. When the two tables are compared (**Table 5-6** and **Table 5-11**), it is seen that an improvement is not achieved in the RMSE as a result of increased dimension of *Right* neighborhood.

Table 5-10: Attribute-Value pairs in the modified database.

Index	Attribute	Value
1	Phoneme Identity	42 phones (42 levels)
19	LeftC	A, F, N, P, B, F, S, L, Silence (9 levels)
21	RightC1	A + tS, F + S, N + n, P + t, B + a, F + 2, S, L + l, Silence (9 + 7 =16 levels)
4	Accent	N, A (2 levels)
5	PosInSyllable	N, C, O (3 levels)
6	SylType	H, L (2 levels)
7	SylNoInWord	Numeric
8	WordNo	Numeric
9	WordPOS	N, P, V, Q, I, T, C, A, J, B, E (11 levels)
10	NumOfSyl	Numeric
11	NumOfWord	Numeric
12	PhrInfo	I, M, F
13	NumOfWordFromPreBr	Numeric
14	NumOfWordToFolBr	Numeric
15	SylNo	Numeric
16	NumOfSylFromPreBr	Numeric
17	NumOfSylToFolBr	Numeric
18	Duration(ms)	Numeric

Table 5-11: Prediction performance obtained using all attributes with modified MOAs.

CC	MAE (ms)	RMSE (ms)
0.78	14.51	19.91

5.3.1.1.2 Modification of Phonetic Class (2)

Among the SD and CV ratios of the newly added values, we observe that the significance of ‘tS’ is much below than those of the others. So, we carried out an experiment to observe the effect of discarding ‘tS’ from the value set of *RightC* attribute. The new set of attributes is given in **Table 5-12**.

The prediction performance of the developed model is given in **Table 5-13**. When the last two prediction performances are compared, it is observed that by eliminating ‘tS’ from the value set a minor improvement is obtained but the performance is still worse than the performance of the original modification (**Table 5-6**).

Table 5-12: Attribute-Value pairs in the modified database.

Index	Attribute	Value
1	Phoneme Identity	42 phones (42 levels)
19	LeftC	A, F, N, P, B, F, S, L, Silence (9 levels)
22	RightC2	A, F + S, N + n, P + t, B + a, F + 2, S, L + l, Silence (9 + 6 =15 levels)
4	Accent	N, A (2 levels)
5	PosInSyllable	N, C, O (3 levels)
6	SylType	H, L (2 levels)
7	SylNoInWord	Numeric
8	WordNo	Numeric
9	WordPOS	N, P, V, Q, I, T, C, A, J, B, E (11 levels)
10	NumOfSyl	Numeric
11	NumOfWord	Numeric
12	PhrInfo	I, M, F
13	NumOfWordFromPreBr	Numeric
14	NumOfWordToFolBr	Numeric
15	SylNo	Numeric
16	NumOfSylFromPreBr	Numeric
17	NumOfSylToFolBr	Numeric
18	Duration(ms)	Numeric

Table 5-13: Prediction performances obtained using all attributes with modified MOAs.

CC	MAE (ms)	RMSE (ms)
0.78	14.5	19.89

5.3.1.2 Transformation of Numeric Attribute Values

Performance of decision tree algorithms is highly correlated to the selected attribute-value pairs [Ross 1995]. In Chapter 3, we discuss the selected attribute-value pairs and their effect in the chosen database. Among the selected attributes, some of them have categorical values while the rest of them have numerical values. Most of the numerical attributes are related to positional attributes, like *position-of-word-in-utterance*. They take integer values. In seeking for improved prediciton performance, we use two types of transformations on attribute values: 1) Discretizing the attribute values and 2) Mapping the attribute values to [0, 1] range. The effects of numeric transformations are tested via the *Syllable-Position-in-Word* attribute. The transformations are described previously in Chapter 3, Section 3.11. The simulations are held removing the original *Syllable-Position-in-Word* attribute and replacing the transformed attributes one by one. **Table 5-14** shows the results of the experiments performed with original attribute set. The first

row corresponds to the original attribute set, in the second row, *SylPosInWord1* attribute is replaced by *SylPosInWord2* and *SylPosInWord2* is replaced by *SylPosInWord3* in the third row. Transformation in the values of the numerical attribute *Syllable Position in Word* results in slight improvements in error performances. Best performance is obtained with the discretized version of the chosen attribute.

Table 5-14: Prediction performances of the original and transformed attributes with original attribute set.

Attribute	**CC**	**MAE (ms)**	**RMSE (ms)**
All with SylPosinWord1	0.7718	14.6678	20.4236
All with SylPosinWord2	0.7741	14.5973	20.331
All with SylPosinWord3	0.7729	14.6359	20.3754

5.3.2 Duration Quantization

As mentioned in Chapter 4, the duration range in our database varies with a Gamma distribution. The statistics of the duration data is given in the **Table 5-15**. The table indicates that the duration data is widely spread in the range 2 ms-295 ms with a mean and standard deviation of 63.15 ms and 31.21 ms, respectively. There are 242 distinct duration values.

Table 5-15: Duration statistics of the database

Min	Max	Mean	Median	SD	Rang
2	295	63.1529	57	31.2074	293

Quantization of phone durations is considered for possible improvement in prediction performance. A non-uniform quantization is applied to the original phone durations and duration attributes are used to model quantized phone durations (**Figure 5-1**). The quantization step size is set to 1.1 and each quantization level is mapped to the mean of the corresponding duration interval. As seen in **Figure 5-1**, the total number of levels used is 54. So the number of distinct duration values is reduced from 242 to 54 (approximately 77.7% reduction in variation). The histogram plot of the quantized durations is given in **Figure 5-2**.

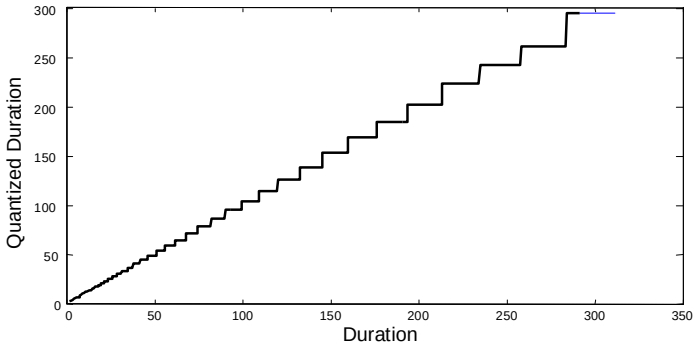

Figure 5-1: Mapping function.

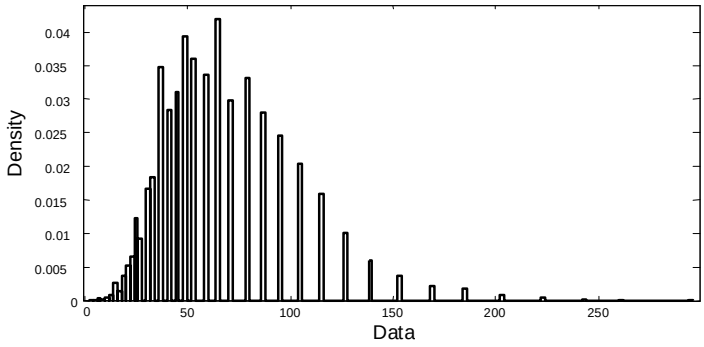

Figure 5-2: Histogram plot of the quantized duration

Table 5-16 shows the quantitative results obtained using all attributes. Comparison with the model developed using original durations (last column of **Table 5-4** and **Table 5-16**) shows that resulting error performances are slightly worse than obtained on training original durations.

Table 5-16: Quantitative results obtained for modeling quantized durations.

CC	MAE	RMSE
0.7702	14.6901	20.4755

5.3.3 Removing the Outliers

Largest deviations in prediction errors on test data generally occur around the boundary values of duration range, i.e., around 2ms and 295 ms. **Figure 5-3** demonstrates

MAE performance on test data sorted in decreasing order. As shown in the figure, MAE drops to 20 ms, which corresponds to approximately 75% of the test data, around 1800th instance.

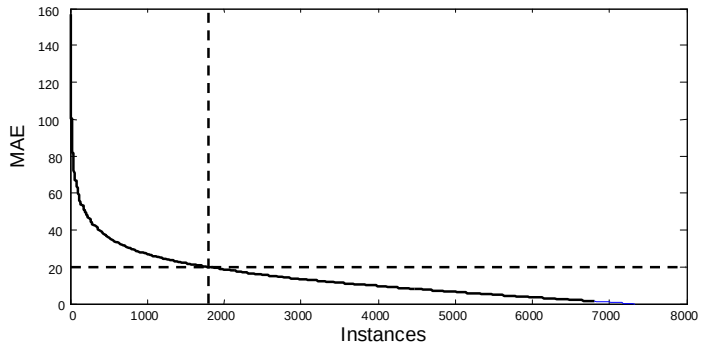

Figure 5-3: Mean Absolute Error (MAE) performance on test data using original 17 attributes.

Table 5-17 shows the MAE performances on test data. As observed in the table, approximately 90% of the data have a MAE less than or equal to 30 ms and 75% have a MAE less than or equal to 20 ms. Considering these information, leaving part of the phonemes that have extreme duration values out of the modeling process is experimented.

Table 5-17: Prediction performances of test data portions.

MAE <= 20 ms	75.4503% (5529)
MAE <= 25 ms	83.7063% (6134)
MAE <= 30 ms	89.2058% (6537)
MAE <= 35 ms	92.6310% (6788)
MAE <= 40 ms	94.7871% (6946)
MAE <= 45 ms	96.3155% (7058)
MAE <= 50 ms	97.3253% (7132)
MAE <= 55 ms	98.1441% (7192)
MAE <= 60 ms	98.5671% (7223)

Figure 5-4 shows the cumulative frequencies of the instances both in the test and train data with respect to their duration values. The text boxes indicate approximately 20 ms and 150 ms duration values, respectively. The corresponding cumulative instance

frequencies are 256 (3.5%) and 7205 (98.3) for the test data and 303 (~1%) and 29087 (98.5%) for the train data, respectively.

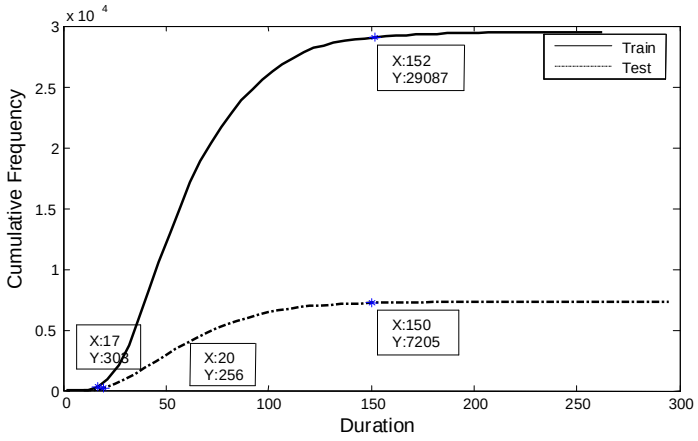

Figure 5-4: Cumulative frequency of instances with respect to their duration values evaluated on test data.

According to the given statistical figures, removing some of the data both from the training and test set do not cause significant exception since most of the data still lies in the selected range. Considering all of the data (both train and test data), 91.3% of the data lies in the 22 ms – 117 ms range. Considering the distribution of durations and the number of extreme instances both in the test and training data, 10 ms - 150 ms range is selected as a reliable duration data range. This new range of duration data contains 98.32% of the test data and 98.28% of the train data. **Table 5-18** demonstrates the quantitative results obtained after removing the outliers of the test and train data. Removing the outliers of the data resulted in an 8.8% RMSE improvement.

Table 5-18: Prediction performances obtained using all 17-attributes to model newly constructed data.

CC	**MAE (ms)**	**RMSE (ms)**
0.7541	13.9756	18.6214

5.3.4 Attribute Selection Using Mutual Information

Performance of decision tree learning is highly related to the quality of attributes selected for modeling. Using attributes that are highly correlated or bear high mutual information would yield degraded performance. Therefore, it is important to carefully determine the attributes to be used in model development. We have previously studied the selection of attributes according to their performance in model development. First, single attributes are used for model development and their prediction performances are given. This result is informative about the contribution of each attribute to duration modeling. Then, optimal subsets of attributes with increasing sizes of one element are developed and used for model development. Trying different attribute combinations, we observe that the optimal attribute sets that differ by one element in size are also different by one element in type. The optimal subset that revealed the best result is composed of the attributes 1, 2-3, 4, 6, 9, 12, and 14.

Determining the relevance of attributes can be performed in various ways. Another criterion of attribute selection can be mutual information among attributes. Mutual information of two random variables is a quantity that measures the independence of two variables. The unit of measurement of mutual information is *bits*.

Formally, in discrete case, if the joint probability mass function of **X** and **Y** is $p(x, y)$ = Prob(**X**=x, **Y**=y), the marginal probability mass function of **X** is $f(x)$ = Prob(**X**=x), and the marginal probability mass function of **Y** is $g(y)$ = Prob(**Y**=y), then the mutual information of **X** and **Y**, $I(\boldsymbol{X}, \boldsymbol{Y})$, is defined as:

$$I(X,Y) = \sum_{x,y} p(x,y) \times \log_2 \left(\frac{p(x,y)}{f(x)g(y)} \right) \tag{5-1}$$

and for the continuous case, probability mass functions are replaced by the corresponding probability density functions and the summation is replaced by the integral:

$$I(X,Y) = \int_{(-\infty,\infty)\times(-\infty,\infty)} p(x,y)\ \log_2\left(\frac{p(x,y)}{f(x)g(y)}\right) d(x,y) \qquad (5\text{-}2)$$

Mutual information is a measure of independence in the following sense: $I(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if **X** and **Y** are independent random variables. This is easy to see in one direction: if **X** and **Y** are independent, then $p(x,y) = f(x)\ g(y)$, and therefore:

$$\log\frac{p(x,y)}{f(x)g(y)} = \log 1 = 0 \qquad (5\text{-}3)$$

Moreover, mutual information is nonnegative (i.e. $I(\boldsymbol{X},\boldsymbol{Y}) \geq 0$) and symmetric (i.e. $I(\boldsymbol{X},\boldsymbol{Y}) = I(\boldsymbol{Y},\boldsymbol{X})$).

Table 5-19 demonstrates the mutual information between the attributes (1-17, 19-22) used in developing phoneme duration models and durations of the training data. Duration is considered to be a discrete random variable and the probability of each duration value is calculated accordingly. The left part of the table shows the mutual information for original durations (262 levels) and the right part of the table reveals the mutual information values for quantized duration values (53 levels). As observed in the table, top seven entries are 1, 2-3, 15, 17, 16, 8, and 5 which are different from the elements of the attribute set that yields the best prediction performance previously (1, 2-3, 4, 6, 9, 12, and 14).

Table 5-20 and **Table 5-21** show the mutual information values of the attributes with respect to each other. The diagonal entries in the table attain the largest values of the corresponding column and row and the mutual information matrix is symmetric. According to the tables, the attributes are not independent but have some dependencies. The mutual information of the Phoneme-Identity and the Left/Right, PosInSyllable, LeftC/RightC, RightC1 and RightC2 attributes are slightly larger compared to the mutual information values with other attributes revealing a stronger relation between them. The relation between the phoneme and its context is reasonable but the relation between the PosInSyllable and the Phoneme Identity can be elaborated. The training database consists of 5126 codas, 11684 onsets and 12717 nuclei. As mentioned in the third chapter, all the

vowels of the database is labeled as nucleus while the consonants are eithered labeled as onset or coda. Because of the labeling style, there is a strong relation between the PosInSyllable attribute and the Phoneme Identity. Other entries do not indicate strong relation among any attribute pair.

Table 5-19: Mutual information of attributes with respect to original durations (Left) and with respect to quantized durations (Right) in decreasing bits.

Using original duration values		**Using quantized duration levels**	
Attribute	**I**	**Attribute**	**I**
1	0.5852	1	0.4688
3	0.2952	3	0.1721
2	0.2195	21	0.1345
21	0.1904	22	0.134
22	0.1874	20	0.1216
15	0.1809	2	0.0898
20	0.154	17	0.0677
17	0.1376	5	0.0638
16	0.1186	12	0.057
8	0.0763	15	0.0419
5	0.0732	7	0.0388
7	0.066	16	0.0346
12	0.0657	19	0.0313
19	0.0647	14	0.0268
11	0.0644	10	0.0244
9	0.0579	9	0.022
10	0.0549	6	0.0199
14	0.0518	8	0.0192
13	0.0425	11	0.0133
6	0.0252	13	0.0109
4	0.0072	4	0.0015

Table 5-20: Mutual information matrix.

	1	2	3	4	5	6	7	8	9	10	11
1	4.635	0.9181	0.9117	0.0331	1.0776	0.0687	0.1333	0.0273	0.1154	0.0481	0.0146
2	0.9181	4.6985	0.5788	0.0273	0.5825	0.0571	0.16	0.0659	0.0926	0.0518	0.0145
3	0.9117	0.5788	4.6743	0.032	0.8982	0.0627	0.0704	0.0388	0.1019	0.0354	0.0142
4	0.0331	0.0273	0.032	0.8913	0.0011	0.0047	0.1156	0.001	0.017	0.0704	0.0002
5	1.0776	0.5825	0.8982	0.0011	1.4912	0.1877	0.0041	0.0004	0.0025	0.0031	0.0002
6	0.0687	0.0571	0.0627	0.0047	0.1877	0.9988	0.0067	0.0021	0.0118	0.0147	0.001
7	0.1333	0.16	0.0704	0.1156	0.0041	0.0067	2.1287	0.0083	0.0436	0.4859	0.0033
8	0.0273	0.0659	0.0388	0.001	0.0004	0.0021	0.0083	3.3243	0.1417	0.0387	0.2836
9	0.1154	0.0926	0.1019	0.017	0.0025	0.0118	0.0436	0.1417	2.1441	0.1543	0.0254
10	0.0481	0.0518	0.0354	0.0704	0.0031	0.0147	0.4859	0.0387	0.1543	2.421	0.0202
11	0.0146	0.0145	0.0142	0.0002	0.0002	0.001	0.0033	0.2836	0.0254	0.0202	2.8556
12	0.0453	0.1265	0.1122	0.0462	0.0012	0.0017	0.234	0.0637	0.0569	0.0507	0.0008
13	0.0128	0.0564	0.0189	0.0007	0.0002	0.0012	0.0046	0.657	0.061	0.0257	0.0231
14	0.0172	0.0256	0.054	0.001	0.0001	0.0005	0.0081	0.1209	0.134	0.0328	0.0217
15	0.0587	0.1312	0.0654	0.0209	0.0013	0.0045	0.1812	1.5787	0.1295	0.0697	0.1191
16	0.0603	0.1467	0.0531	0.0398	0.0013	0.0031	0.4466	0.2843	0.0692	0.1101	0.0148
17	0.0548	0.0607	0.1444	0.0199	0.0007	0.0029	0.0747	0.1002	0.1544	0.0409	0.014
19	0.6233	2.8638	0.321	0.0144	0.5158	0.0497	0.0748	0.0443	0.0292	0.0204	0.0027
20	0.6461	0.298	2.8585	0.0146	0.8303	0.0324	0.0247	0.0187	0.0366	0.0119	0.0028
21	0.7373	0.3774	3.544	0.0169	0.8548	0.0366	0.0378	0.0245	0.0612	0.0184	0.0053
22	0.7329	0.3745	3.5214	0.0159	0.8548	0.0365	0.0366	0.0239	0.0595	0.0161	0.0049

Table 5-21: Mutual information of matrix (continued).

	12	13	14	15	16	17	19	20	21	22
1	0.0453	0.0128	0.0172	0.0587	0.0603	0.0548	0.6233	0.6461	0.7373	0.7329
2	0.1265	0.0564	0.0256	0.1312	0.1467	0.0607	2.8638	0.298	0.3774	0.3745
3	0.1122	0.0189	0.054	0.0654	0.0531	0.1444	0.321	2.8585	3.544	3.5214
4	0.0462	0.0007	0.001	0.0209	0.0398	0.0199	0.0144	0.0146	0.0169	0.0159
5	0.0012	0.0002	0.0001	0.0013	0.0013	0.0007	0.5158	0.8303	0.8548	0.8548
6	0.0017	0.0012	0.0005	0.0045	0.0031	0.0029	0.0497	0.0324	0.0366	0.0365
7	0.234	0.0046	0.0081	0.1812	0.4466	0.0747	0.0748	0.0247	0.0378	0.0366
8	0.0637	0.657	0.1209	1.5787	0.2843	0.1002	0.0443	0.0187	0.0245	0.0239
9	0.0569	0.061	0.134	0.1295	0.0692	0.1544	0.0292	0.0366	0.0612	0.0595
10	0.0507	0.0257	0.0328	0.0697	0.1101	0.0409	0.0204	0.0119	0.0184	0.0161
11	0.0008	0.0231	0.0217	0.1191	0.0148	0.014	0.0027	0.0028	0.0053	0.0049
12	1.0156	0.2259	0.2235	0.175	0.5359	0.5663	0.0963	0.0906	0.095	0.0945
13	0.2259	2.2495	0.1312	0.2755	1.0574	0.1093	0.0449	0.009	0.0107	0.0105
14	0.2235	0.1312	2.1778	0.1012	0.129	1.0353	0.0122	0.04	0.0436	0.0431
15	0.175	0.2755	0.1012	4.8819	1.2866	0.1319	0.0833	0.0244	0.035	0.0338
16	0.5359	1.0574	0.129	1.2866	3.809	0.1624	0.1012	0.0208	0.026	0.0255
17	0.5663	0.1093	1.0353	0.1319	0.1624	3.7997	0.0202	0.0951	0.1063	0.1048
19	0.0963	0.0449	0.0122	0.0833	0.1012	0.0202	2.8638	0.1915	0.2387	0.2379
20	0.0906	0.009	0.04	0.0244	0.0208	0.0951	0.1915	2.8585	2.8585	2.8585
21	0.095	0.0107	0.0436	0.035	0.026	0.1063	0.2387	2.8585	3.544	3.5214
22	0.0945	0.0105	0.0431	0.0338	0.0255	0.1048	0.2379	2.8585	3.5214	3.5214

5.3.5 Shift and/or Scale Modification

In an attempt to decrease MSE further, we suggest making shift and/or scale modifications on the predicted duration values. The modifications are described in the following subsections.

5.3.5.1 Shift Modification

Suppose that we have developed a model using all attributes and therefore we have the phoneme duration predictions, $\hat{d}$ for each train and test instance. We define the new predictions such that $\hat{d}_{shitf} = \hat{d} + a$ where a is a constant shift value. Here, the aim is to find *a*, such that the MSE is minimized. To this aim, we first define the modified MSE:

$$\begin{aligned} MSE_{shift} &= \frac{1}{N}\sum_{k=1}^{N}\left(d_k - \hat{d}_{shift,k}\right)^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(d_k - \hat{d}_k - a\right)^2 \end{aligned} \tag{5-4}$$

To find *a* that minimizes MSE_{shift}, we take the derivative of MSE_{shift} with respect to *a* and equate to 0:

$$\frac{\partial MSE_{shift}}{\partial a} = 0 \tag{5-5}$$

Solving above equation,

$$\hat{a} = \frac{1}{N}\sum_{k=1}^{N}\left(d_k - \hat{d}_k\right) \tag{5-6}$$

We use (**5-6**) to calculate $\hat{a}$'s for each phoneme class in the database. Then, the corresponding MSE_{shift} is calculated using (**5-4**).

5.3.5.2 Scale Modification

The same assumptions hold for the scale modification. The new predictions are defined as $\hat{d}_{scale} = b.\hat{d}$ where *b* is a constant. The modified MSE is defined as follows:

$$\begin{aligned} MSE_{scale} &= \frac{1}{N}\sum_{k=1}^{N}\left(d_k - \hat{d}_{scale,k}\right)^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(d_k - b.\hat{d}_k\right)^2 \end{aligned} \tag{5-7}$$

In order to find *b* that minimizes MSE_{scale}, we take the derivative of MSE_{scale} with respect to *b* and equate to 0:

$$\frac{\partial MSE_{scale}}{\partial b} = 0 \tag{5-8}$$

Solving above equation for *b*, we obtained

$$\hat{b} = \frac{\sum d_k \hat{d}_k}{\sum \hat{d}_k{}^2} \tag{5-9}$$

For every phoneme type in the database, we calculate the predictions for *b* and MSE_{scale} is calculated the accordingly.

5.3.5.3 Shift and Scale Modification

Another possibility is to apply shift and scale modification simultaneously. The new predictions are defined as $\hat{d}_{shift\_scale} = c\hat{d} + e$ where *c* and *e* are constants. Here, the aim is to find *c* and *e* such that *MSE* is minimized. The new MSE is defined as:

$$MSE_{shift\_scale} = \frac{1}{N}\sum_{k=1}^{N}\left(d_k - \hat{d}_{shift\_scale,k}\right)^2 = \frac{1}{N}\sum_{k=1}^{N}\left(d_k - c\hat{d}_k - e\right)^2 \quad (5\text{-}10)$$

In order to find c and e that minimizes MSE_{shift_scale}, the partial derivatives of MSE_{shift_scale} with respect to c and e are equated to 0 and solved simultaneously:

$$\frac{\partial MSE_{shift\_scale}}{\partial c} = 0 \qquad \frac{\partial MSE_{shift\_scale}}{\partial e} = 0 \quad (5\text{-}11)$$

Solving above equation for c and e, we obtained

$$e = \frac{1}{N}\sum d_k - c * \hat{d}_k$$

$$c = \frac{\sum_k d_k\hat{d}_k - \frac{1}{N}\sum_j\sum_k d_k\hat{d}_j}{\sum_k \hat{d}_k{}^2 - \frac{1}{N}\sum_j\sum_k \hat{d}_k\hat{d}_j} \quad (5\text{-}12)$$

For every phoneme type in the database, we calculate the predictions for c and e, then MSE_{shift_scale} is calculated accordingly.

5.3.5.4 Application of Shift and/or Scale Modification

Predictions obtained using original 17-attribute dataset is used (last column of **Table 5-4**) for the modifications described in the previous sections. Modification parameters are calculated both from the test set and the training set. **Table 5-22** shows the MSE values after applying the corresponding modifications on each database. The last row of the table shows the results of using *a*, *b*, *c* and *e* on the test data when they are calculated from the training data.

Table 5-23 and **Table 5-24** demonstrate corresponding RMSE and CC values for the original and modified predictions, respectively. As observed in the tables, both RMSE and CC are improved for the case where shift and scale modification are applied simultaneously. It can be noticed that the improvements are slightly better when the modification parameters are trained on the test data.

Table 5-22: Original and modified MSE values.

Database	MSE	MSE_{shift}	MSE_{scale}	MSE_{shift_scale}
Test	417.1231	400.9892	399.455	381.0388
Train	280.6778	280.6568	280.6429	280.5861
Train_Test	417.1231	417.2679	417.1450	415.8239

Table 5-23: Original and modified RMSE values ($RMSE = \sqrt{MSE}$).

Database	$RMSE$	$RMSE_{shift}$	$RMSE_{scale}$	$RMSE_{shift\_scale}$
Test	20.42	20.02	19.99	19.52
Train	16.75	16.75	16.75	16.75
Train_Test	20.42	20.43	20.42	20.39

Table 5-24: Original and modified CC values

Database	CC	CC_{shift}	CC_{scale}	CC_{shift_scale}
Test	0.77175	0.78305	0.78375	0.79335
Train	0.84129	0.84129	0.84130	0.84133
Train_Test	0.77175	0.77166	0.77170	0.77240

CHAPTER 6

SYLLABLE PITCH CONTOUR PREDICTION DATABASE AND PROSODIC ATTRIBUTES

The database for predicting syllable pitch contours contains 15867 syllables of the 692 sentences in the database. There are 1254 distinct syllables in the database. For prediction purposes, the database is split into two subsets: training and test datasets. Training set contains approximately 80% (12483 instances) of the whole database and the remaining 20% (3384 instances) of the data is used for testing.

6.1 Features Used in Syllable Pitch Contour Prediction

Every syllable in the database is coded with a feature vector. Feature vector contains information related to syllable, word and sentence levels.

The features used in syllable pitch contour prediction experiments are given the succeeding sections.

6.1.1 Lexical Stress

This feature represents the lexical stress of the syllable. Analysis of the pitch contours in our database reveals that pitch accents are mainly aligned with the lexically stressed syllables of the words. Such an alignment can be observed in **Figure 6-1.** The figure illustrates sound waveform, pitch contour and the syllable labels of the sentence ‘doğduğum büyüdüğüm memleketime biraz faydam olsun istedim dedi’ (I want to be helpful to the country I was born and held he said). Pitch accents of the sentence are aligned with the words ‘doğduğum’ (I was born), ‘büyüdüğüm’ (I was held), ‘memleketime’ (to my country) and ‘faydam’ (helpful). The lexical stresses of the words are aligned with the syllables ‘ğum’, ‘ğüm’, ‘me’, and ‘dam’, respectively. As shown in the figure, pitch accents are also aligned with the same syllables of the words. It should also be noted that identical syllables have different lexical stresses and therefore different

accent types[2]. Let us consider the syllable ‘me’ of the sentence ‘özgüre beni beklemesini söylemedin mi’ (didn’t you tell özgür to wait for me) given in **Figure 6-2**. Although both of the ‘me’s are orthographically identical because of their lexical stress property, their pitch contours differ. While the lexically stressed ‘me’ shows an increasing pattern, the unstressed ‘me’ exhibits a decreasing pattern. Therefore, in order to develop an accurate prosodic model in Turkish, lexical stress should be employed in the prediction procedure.

Lexically stressed syllables are obtained through a morphological analysis procedure. Then, stress assignment rules for Turkish are applied to obtain the lexical stress of each syllable. During coding, a syllable is represented with an ‘A’ if it is stressed and with an ‘N’, otherwise.

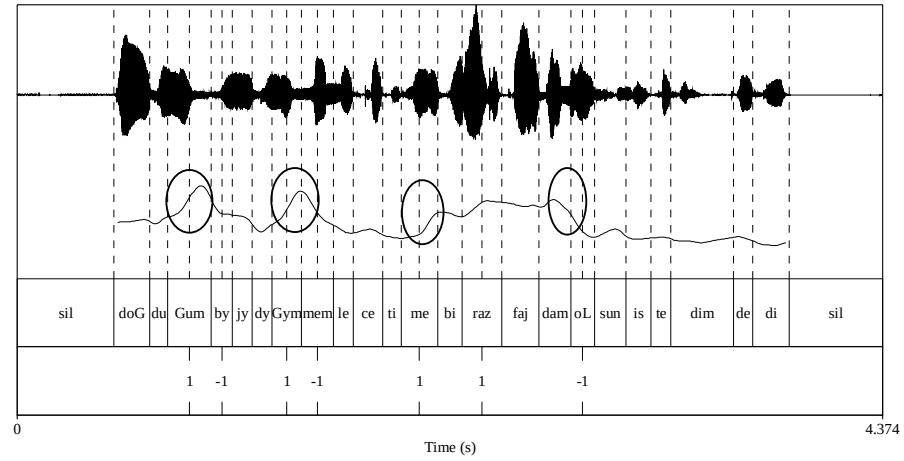

Figure 6-1: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels (lower panel) of the sentence ‘doğduğum büyüdüğüm memleketime biraz faydam olsun istedim dedi’.

[2] *No Accent* should also be considered as a type of pitch accent.

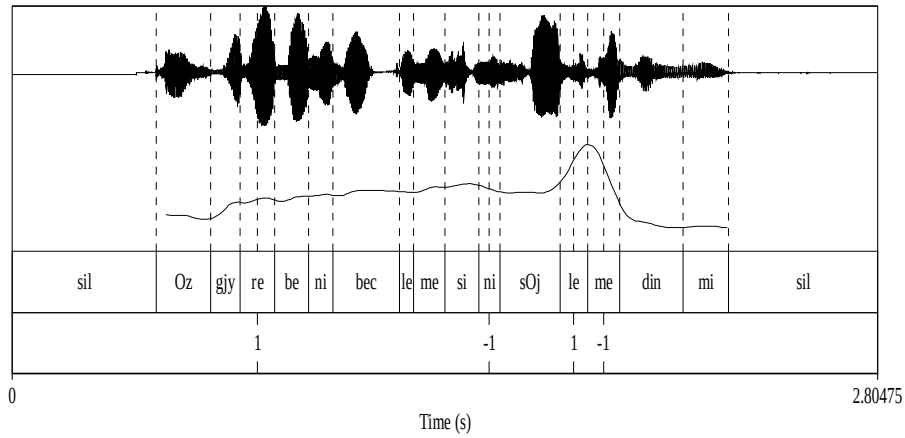

Figure 6-2: Sound waveform (upper), pitch contour (middle), and syllabic segmentation (lower) of the sentence ‘özgüre beni beklemesini söylemedin mi’.

6.1.2 Negation Flag (NegFlag)

NegFlag is a binary feature representing the syllables’ morphemic constitute. If a negation suffix is enclosed within syllable or the syllable is composed of a negation suffix, then the NegFlag is set to 1 otherwise it is set to 0.

As discussed in the stress subsection, the type of the lexical stress of a syllable plays an important role in determining the pitch accents. In the previous section, we examine the syllable ‘me’ with two different lexical stress constituents. In the first case, where the syllable was enclosed within the sentence ‘doğduğum büyüdüğüm memleketime biraz faydam olsun istedim dedi’, ‘me’ was lexically stressed and has an increasing pitch pattern. In the second case, where the syllable was enclosed within the sentence ‘özgüre beni beklemesini söylemedin mi’, the syllable ‘me’ is lexically unstressed and shows a decreasing pitch pattern. However, not all the unstressed ‘me’ syllables show a decreasing pitch pattern. **Figure 6-3** demonstrates a counter example for the lexically unstressed ‘me’ with almost flat pitch contour. The figure shows the sound waveform, pitch contour, pitch accents and syllable segmentation of the sentence ‘avukat nusret senem merakla bekliyordu planlarının işleyip işlemeyeceğini’ (lawyer nusret senem was waiting curiously whether his plans was working or not). The difference in the pitch contours of both ‘me’ principally is due to the morphemic structure. The ‘me’

shown in **Figure 6-2** is an extrametrical[3] negation morpheme while the 'me' shown in **Figure 6-3** belongs to root of the word and is not related to an extrametrical suffix or enclitic. Therefore, we use a binary flag to discriminate the negation suffix due to its stress blocking affect.

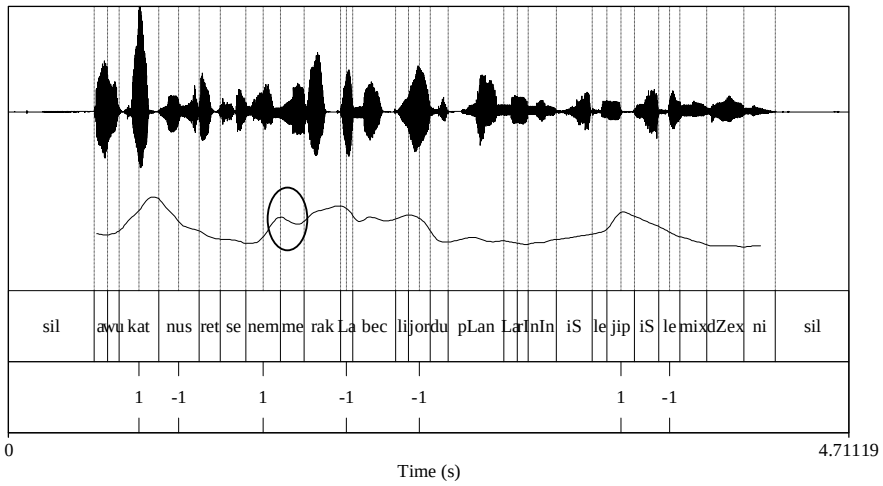

Figure 6-3: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels (lower panel) of the sentence 'avukat nusret senem merakla bekliyordu planlarının işleyip işlemeyeceğini'.

6.1.3 Syllable Type (SylType)

Type of parent syllable (*SylType*) is included in the annotations. Two levels are used to denote syllable types: Heavy (H) and Light (L). Heavy and light syllables are sometimes called *open* and *closed* syllables, respectively.

6.1.4 Syllable Structure (SylStruct)

The structure of the syllable in terms of its constituents is represented by the *SylStruct* feature. As mentioned before, the consonants occurring before syllable nucleus are called onsets while the consonants occurring after the nucleus are called codas. A syllable should possess a single nucleus and our syllabification algorithm mainly relies on this principle. Current feature codes the syllable according to the order of the onsets, nucleus and codas, i.e., N, ON, ONC, OON, ONCC, and OONCC.

[3] Extrametricality is equivalent to stress blocking and can be used interchangeably.

It may be argued that syllable structure highly influences the location of pitch accent peak within the syllable. An early or late peak can be observed depending on the succeeding syllable or more specifically succeeding phoneme.

6.1.5 Syllable-Position-in-Word (SylNoinWord)

This attribute codes the position of the syllable in the word. Counting is performed in syllable units. The feature attains numerical values ranging from 1 up to 10. Our database contains words of at most 10 syllables thus the feature can take at most 10 as value.

The default stress in Turkish is generally assigned to the last syllable of the word. Therefore, the position of the syllable in the parent word plays a crucial role in pitch accent and prominence level prediction. Accented syllables at the beginning of the phrases have higher pitch values from the rest of the syllables except for the lexically stressed syllable of the focus word. Syllables at the end of the phrases show gradual decrease in the prominence level if the sentences are not in the question forms.

In **Figure 6-4 - Figure 6-6** , three aspects of the syllable ‘ba’ from different words are given. All the three pitch contours reveal almost similar shapes but with different pitch levels.

6.1.6 Syllable-Position-in-Word 1 (SylPosinWord1)

Categorical feature representing the position of syllable in parent word. The feature takes a value I (F) when the syllable is a word initial (final) syllable, a value Single when the word consists of only one syllable word, or a value M for other cases. With this coding scheme, we have the advantage of differentiating initial and final syllables as well as words with single syllables. Using discrete symbols instead of integer values also reduces the dimension of the attribute from 10 to 4. However, we loose the information relating prominence with the actual location of parent syllable.

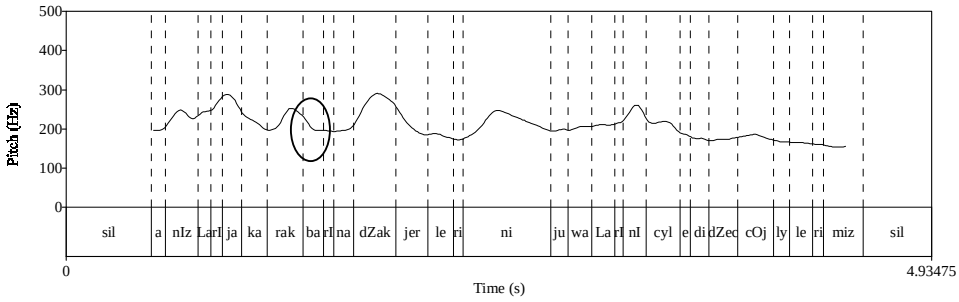

Figure 6-4: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels (lower panel) of the words ‘barınacak yerlerini’ (the places they will live). Minimum pitch observed on the syllable ‘ba’ is around 192 Hz.

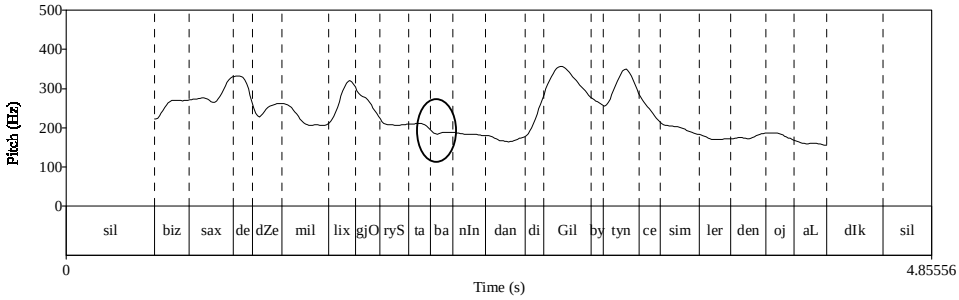

Figure 6-5: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels (lower panel) of the words ‘görüş tabanından’ (base sight). Minimum pitch observed on the syllable ‘ba’ is around 176 Hz.

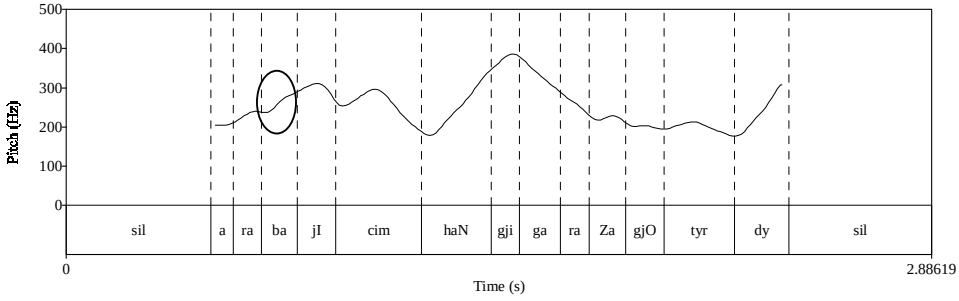

Figure 6-6: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels (lower panel) of the words ‘arabayı kim’ (the car who). Minimum pitch observed on the syllable ‘ba’ is around 225 Hz.

6.1.7 Word Position in Sentence (WordNoinSent)

WordNoinSent feature is a numeric feature that represents the position of word in the sentence in terms of word units. The values change in the range [1, 14].

The position of the word within the sentence affects the word pitch contours such that the words to the end of the sentences have lower peak values due to declination effect. *Downstep* and *declination* are important aspects of intonation. *Downstep* refers to the lowering effect observed in successive high (H) pitch targets in recursive H L patterns; while *declination* refers to the tendency for F0 to gradually decline over the course of an utterance [Pierrehumbert, 2000; Xu and Wang, 2001]. But this is not the case for the word that is intended. The word that is focused attains the maximal peak value independent of its location in the sentence. We can not give examples from our database to reveal the locative effects of the words on the pitch contours since the database sentences are not designed to have such variability. Words mostly appear in their inflected and derived surface forms. Hence, it is almost impossible to find a word at various locations of sentences with identical surface forms. But all other effects can be observed throughout the sentences given as examples previously with different words.

6.1.8 Word Position in Sentence 1 (WordPosinSent1)

Symbolic attribute representing the position of parent word in the sentence. The feature takes a value I (F) when the word is located at sentence initial (final), or a value M for other cases. The main reason employing *WordPosinSent1* attribute is to discriminate the sentence initial and final words. As mentioned in the previous section, declination is a global aspect of intonation. Hence, words at phrase initials have higher peaks than phrase final words in declarative sentences. Therefore, using symbolic attribute for identifying the word location in sentence helps to track this phenomenon in a better way. At the same time, using symbolic representation reduces the attribute space considerably, i.e. the former representation has 14 values while the latter has only 3 values.

6.1.9 Number of Phones in Syllable (NumofPhninSyl)

Numerical attribute indicating the number of phones in the parent syllable. The values change in the range [1, 5]. This attribute is used to observe the effect of syllable length to pitch contour.

6.1.10 Number of Syllables in Word (NumofSylinWord)

Numerical attribute indicating the number of syllables in the word. The value is same for all syllables of the same word. The values change in the range [1, 10]. This attribute is used to observe the effect of word length to pitch contour.

6.1.11 Number of Words in Sentence (NumofWordinSent)

Numerical attribute indicating the number of words in the sentence. The value is same for all syllables of the same sentence. Since, the database consists of sentences having 3 to 19 words; the attribute attains values in this range.

6.1.12 Part-of-Speech of Current Word (POSw)

Part-of-Speech (POS) of parent word is also used in syllable pitch contour prediction. Generally, in Turkish, contrastive stress is realized by locating the most prominent word (focus) to preverbal position. Verb focusing is performed by carrying the verb to the sentence initial but such kind of a contrastive stress does not exist in our database. Commonly, the words placed after the verb, have no prominence. Thus, the contours observed after the verb are very smooth with no abrupt pitch changes. **Figure 6-7** manifests a non-verb final sentence, 'böyle bir dönemde oynatılması tesadüf olamaz bu filmin' (at such a time it is impossible to play this film on purpose). The verb of the sentence is 'olamaz' (it is impossible) and it is located before the sentence final. The words located after the verbs contribute to the pitch contour only with their microprosody.

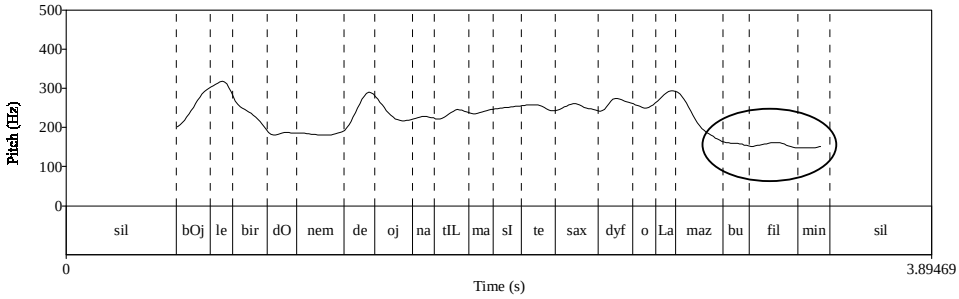

Figure 6-7: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels of the sentence.

In Yes/No questions, the contrastive stress is realized by locating the prominent word in front of the question enclitic 'mi'. **Figure 6-8** demonstrates focusing in Yes/No

questions by means of the enclitic ‘mi’. The waveform given in the figure belongs to the sentence ‘başçavuş tüm takıma koşu cezası mı verdi’ (did the sergeant major give the whole team running punishment). The word placed before the enclitic ‘mi’ is naturally focused and thus bears the highest pitch of the sentence.

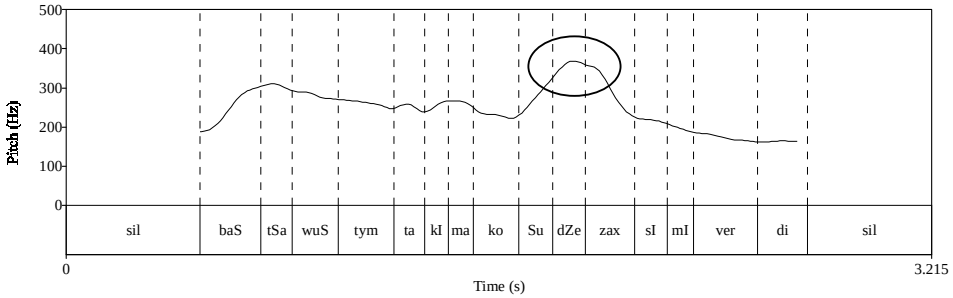

Figure 6-8: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels of the sentence ‘başçavuş tüm takıma **koşu cezası** mı verdi’.

When the question enclitic ‘mi’ is placed at the end of the sentence, it either makes the preceding syllable accented or it obeys to the previous rule and focuses the preceding word, so the word is accented at its lexically stressed syllable. **Figure 6-9** illustrates focusing using the enclitic ‘mi’. The waveform given in the figure belongs to the sentence ‘özgüre beni beklemesini söylemedin mi’. The word placed before the enclitic ‘mi’ is focused and bears the highest pitch of the sentence at the lexically stressed syllable ‘le’. In **Figure 6-10**, an example for pre-accenting is illustrated. The sentence studied in the example is ‘çıplak doğrudan doğruya tadını duyuran içkiler var biliyor musun’ (do you know that there are naked beverages that flavor directly). Here, the syllable ‘yor’ bears the accent of the syllable although it is a part of a stress-blocking morpheme that never bears lexical stress. A counter example for the unaccented version of the syllable ‘yor’ is given in **Figure 6-11**. The sentence inspected in the figure is ‘zirveden once bu hususta anlaşılması gerekmiyor mu’ (isn’t it necessary to make an agreement on the subject before the summit). In this sentence, the word placed before the enclitic ‘mi’ is focused and sharpest slope is observed at its lexical stress.

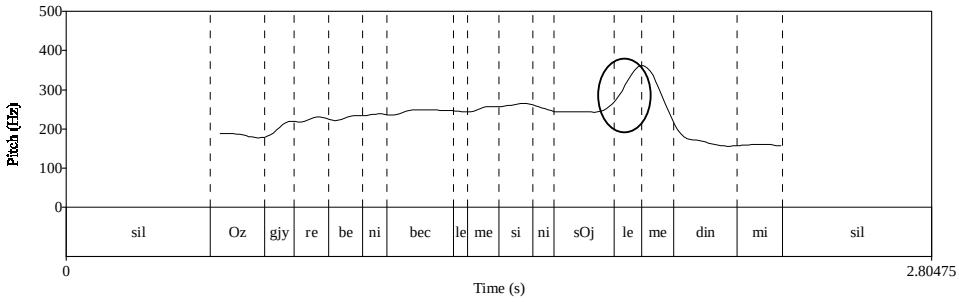

Figure 6-9: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels of the sentence ‘özgüre beni beklemesini söylemedin mi’.

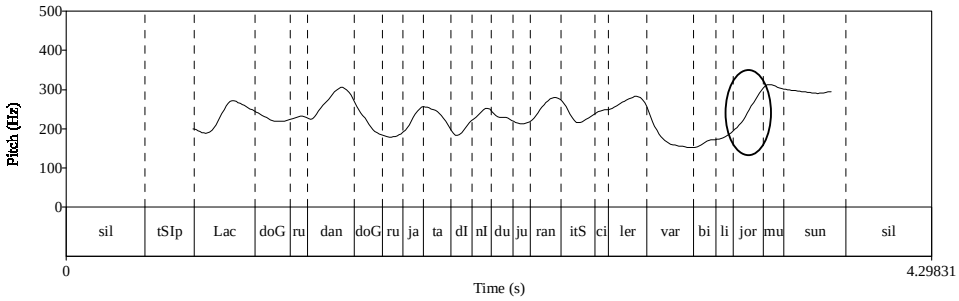

Figure 6-10: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels of the sentence ‘çıplak doğrudan doğruya tadını duyuran içkiler var biliyor musun’.

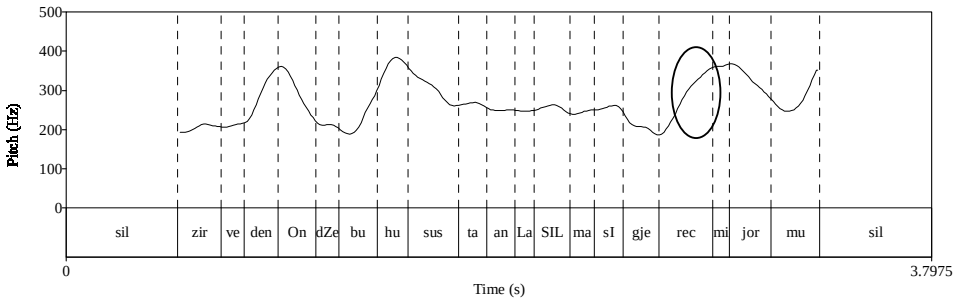

Figure 6-11: Sound waveform (upper panel), pitch contour (middle panel), and syllable labels of the sentence ‘zirveden once bu hususta anlaşılması gerekmiyor mu’.

Therefore, it is necessary to determine the part-of-speech of words to develop an accurate model of Turkish prosody. The categories and their occurrence frequency are given in **Table 6-1**.

Table 6-1: POS categories and their occurrence frequency.

NOUN	VERB	ADJ	ADV	PROPN	PROPP	POSTP	CONJ	MODAL
6547	2765	2701	878	401	348	314	305	300
TELL	**PRON**	**CNOUN**	**QUES**	**WH**	**NOT**	**INF**	**EXC**	
280	252	199	188	184	114	89	2	

6.1.13 Part-of-Speech of Succeeding Word (POSw+1)

This feature represents the POS of the word that immediately follows the parent word. The feature has the same values as the POSw feature plus a symbol *none* for the sentence finals. As discussed in the previous section, verbs and enclitics play an important role on the pitch contours of their predecessors. Generally in Turkish, preverbal word is focused. In the previous section, we also discuss about the effects of the enclitic 'mi'. Since enclitics are stress blockers words preceding the enclitics are generally accented. Hence, using a three-word POS window (POSes of preceding, current, and succeeding words) as attribute positively affects prediction performance.

6.1.14 Part-of-Speech of Preceding Word (POSw-1)

This feature represents POS of the word that immediately precedes parent word. The feature has the same values as the POSw feature plus a symbol *none* for the sentence initials.

6.1.15 Part-of-Speech of Word Root (POSRoot)

Part-of-Speech (POS) of the root of word is also used in syllable pitch contour prediction in order to capture the nature of words. The attribute attains values as in *POSw* attribute.

6.1.16 Break Index (Break)

This attribute encodes the perceptual break category of the syllable. The attribute takes categorical values such as I, F, and M as well as SI and SF. An I (F) value denotes that the syllable is located immediately after (before) a break, while the M value is assigned to the syllables that do not occur at boundary locations. A SI (SF) value denotes that syllable is at the beginning (end) of the sentence. Labels I and F refer to minor breaks while SI and SF refer to major breaks. Major breaks are also considered as minor breaks but the reverse is not true.

The database was previously manually labeled with respect to perceptual breaks. Close examination of the perceptual breaks reveal that, final lengthening occurs at break positions. Besides, Turkish shows continuation rises at phrase finals like English. Also, declarative Verb-Final sentences have a falling pitch pattern while question forms have rising patterns. Therefore, using break indices provide assistance in discriminating such F0 patterns.

In **Figure 6-12**, the pitch contour and syllable segmentation of the sentence sentence 'mikroorganizmaları yok etmek için şok ısıtma ve soğutma yöntemi kullanılır' (in order to exterminate the microorganisms shock heating and cooling method is used) is given. The figure illustrates continuation rise and phrase-final. Phrase final rising in interrogative sentences is depicted in **Figure 6-13** on the sentence 'neden nurinin sinemaya gitmesini istemiyorsun' (why don't you want Nuri to go to the cinema).

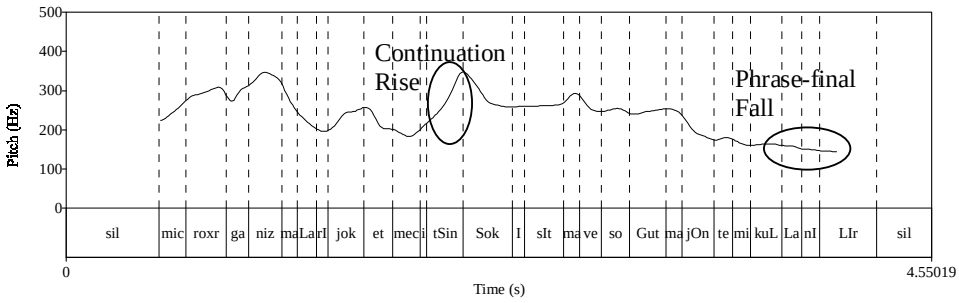

Figure 6-12: Pitch contour, and syllable labels of the sentence 'mikroorganizmaları yok etmek için şok ısıtma ve soğutma yöntemi kullanılır'.

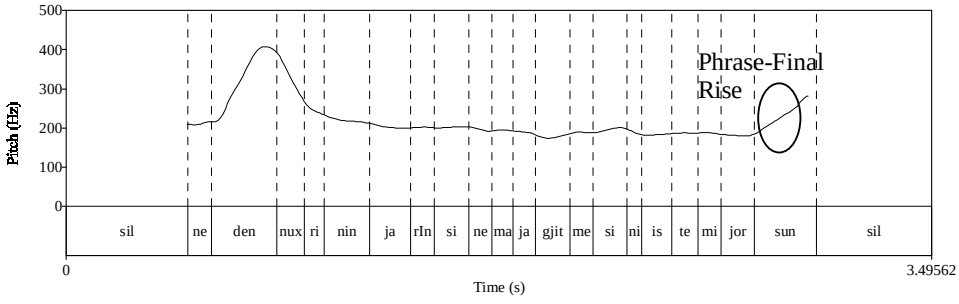

Figure 6-13: Sound waveform, pitch contour, syllable labels and break indices of the sentence 'neden nurinin sinemaya gitmesini istemiyorsun'.

6.1.17 Sentence Type Index (STIndex)

Each syllable is coded with 4 attributes denoting sentence type categories. Sentences are divided into four categories depending on their semantics (Affirmative, Negative, and Interrogative), on their syntactic constituents (Simple, Compound, and Complex), on their verbal compositions (Verb-final and non Verb-final) and for interrogatives, depending on the question word or enclitic (Wh-ques and Yes/No_ques). Each category is represented by an attribute in the feature set. Sentence type combinations observed in the database are given in **Table 8-1**.

6.1.18 Number of Words (Syllables) to the Following Major (Minor) Break (NumofWordToFolMajorBreak, NumofSylToFolMajorBreak, NumofWordToFolMinorBreak, and NumofSylToFolMinorBreak)

Attributes are used to denote the positions of words (syllables) to the next major (minor) phrase break. The attributes are all numeric. Characteristics of attribute values are given in **Table 6-2**. The columns of the table correspond to minimum, maximum, mean, and standard deviations, respectively.

Table 6-2: Characteristics of NumofWordToFolMajorBreak, NumofSylToFolMajorBreak, NumofWordToFolMinorBreak, and NumofSylToFolMinorBreak.

	Minimum	Maximum	Mean	STD
NumofWordToFolMajorBreak	0	18	3.782	2.649
NumofSylToFolMajorBreak	0	44	11.67	7.899
NumofWordToFolMinorBreak	0	9	1.406	1.406
NumofSylToFolMinorBreak	0	27	4.918	4.176

6.1.19 Number of Words (Syllables) from the Previous Major (Minor) Break (NumofWordFromPrevMajorBreak, NumofSylFromPrevMajorBreak, NumofWordFromPrevMinorBreak, and NumofSylFromPrevMinorBreak)

Attributes are used to denote the positions of words (syllables) from the previous major (minor) phrase break. The attributes are all numeric. Characteristics of attribute values are given in **Table 6-3**. The columns of the table correspond to minimum, maximum, mean, and standard deviations, respectively. The attributes are used in the second approach (ref Section 7.2).

Table 6-3: Characteristics of NumofWordToFolMajorBreak, NumofSylToFolMajorBreak, NumofWordToFolMinorBreak, and NumofSylToFolMinorBreak.

	Minimum	Maximum	Mean	StdDev
NumofWordFromPrevMajorBreak	0	18	3.591	2.661
NumofSylFromPrevMajorBreak	0	44	11.67	7.899
NumofWordFromPrevMinorBreak	0	9	1.284	1.378
NumofSylFromPrevMinorBreak	0	27	4.918	4.176

6.1.20 Position of Words (Syllables) in Major (Minor) Phrases (PosofWordMajor, PosofSylMajor, PosofWordMinor, and PosofSylMinor)

Attributes are used to denote the positions of words (syllables) in major (minor) phrases. All are represented by three categorical values: Initial, Middle and Final as in Break feature. The attributes are used in the second approach (ref Section 7.2).

6.1.21 Duration

Syllable durations are also used for syllable pitch contour prediction purposes since they are effective in slope prediction.

6.1.22 Cluster Index of Previous Syllable (Cluster-s1)

For each syllable in the database, the cluster values of the syllables that immediately precede the current syllable are also used in syllable pitch contour prediction. The main reason for this choice is to embed an initial value constraint on the pitch prediction process. Depending on the model used, the attribute attains different values. However, its value is always categorical. Whatever approach is used, the attribute attains values of the dependent variable (ref Section 6.1.23) plus a symbol *none* for sentence initials.

Pitch prediction employing secondly proposed method uses two different versions of this attribute: *Cluster-s1-major* and *Cluster-s1-minor*. The former attains *none* value for sentence initials only. The latter attains *none* value for intermediate phrase initials as well as sentence initials.

6.1.23 Dependent Variable

This feature contains the values to be predicted by decision tree learning. For pitch contour prediction, two different approaches are used and each approach underwent some modifications. Therefore, the feature attains different values to perform different decision tasks.

For parametric modeling, cluster indices obtained by k-means clustering algorithm are used as dependent variable (ref Section 7.1.2). Each syllable is represented by one of 24 clusters given in Appendix A. Hence, the dependent variable attains categorical values changing from 1 to 24. Then, accented and unaccented syllables are determined from the cluster centroids (ref Section 7.1.3). So, the attribute attains only two values (accented versus not-accented) for this task.

Second approach associates accent categories to syllables (ref Section 7.2). Therefore, dependent variable is configured with respect to the analysis results. The attribute either takes three (positive, negative, and no-accent) or two (accented vs no-accent) categorical values depending on the task performed.

6.2 Attribute Evaluation

Three statistical measures are used to reveal prosodic attribute (PA) - dependent variable (DV) relation: Information Gain, Gain Ratio, and Symmetrical Uncertainty.

$$InfoGain(DV, PA) = H(DV) - H(DV|PA) \tag{6-1}$$

$$GainRatio(DV, PA) = \frac{H(DV) - H(DV|PA)}{H(PA)} \tag{6-2}$$

$$SymmetricalUncertainty(DV, PA) = 2 * \frac{H(DV) - H(DV|PA)}{H(DV)} + H(PA) \tag{6-3}$$

where *H* denotes entropy.

Table 6-4 illustrates *Information Gain*, *Gain Ratio* and *Symmetrical Uncertainty* of the attributes with respect to the 24 cluster centroids proposed in Chapter 8 section X. According to the table, the *Cluster-s1* attribute attains the best *Information Gain* and *Gain Ratio* value while the *Duration* attribute attains the best *Symmetrical Uncertainty* measure. Attribute relevance according to *Information Gain* can be listed as follows: *Cluster-s1*, *Normalized_SylNoInSent*, *WordPosinSent1*, *WordNoinSent*, *SylPosinWord1*,

Break, STIndex, *Duration*, *POSw*, *SylNoinWord*, *SylStruct*, *Stress*, *LeftStress*, *Break*-s1, *POS-w1*, *POS+w1*, *RightSylStruct*, *LeftSylStruct*, and *NegFlag*.

Table 6-4: Information gain, Gain ratio and symmetrical uncertainty measures of the attributes with dependent variable in 24-cluster centroid prediction. Shaded values correspond to the maxima of each measure.

Attribute Index	Attribute Name	Information Gain	Gain Ratio	Symmetrical Uncertainty
1	NegFlag	0.0137	0.1025	0.019
2	Stress	0.053	0.0608	0.008
3	LeftStress	0.0503	0.0577	0.006
4	SylStruct	0.0539	0.0369	0.02
5	LeftSylStruct	0.0161	0.0109	0.056
6	RightSylStruct	0.0234	0.0164	0.094
7	SylNoinWord	0.0701	0.0444	0.048
8	WordNoinSent	0.1727	0.0988	0.011
9	POSw	0.0868	0.0403	0.031
10	POS-w1	0.0367	0.0165	0.018
11	POS+w1	0.0295	0.0122	0.023
12	Break	0.1108	0.1076	0.026
13	Break-s1	0.0471	0.0509	0.005
14	Normalized_SylNoInSent	0.2996	0.1562	0.017
15	SylPosinWord1	0.1493	0.0831	0.008
16	WordPosinSent1	0.2051	0.1937	0.02
17	STIndex	0.0964	0.0196	0.04
18	Duration	0.0957	0.0555	0.282
19	Cluster-s1	1.2511	0.2824	0.074

Table 6-5 illustrates *Information Gain, Gain Ratio* and *Symmetrical Uncertainty* of the attributes with respect to the three accent categories proposed in Section 7.2. According to the table, *SylPosinWord1* is the best predictor when *Information Gain* measure is considered. For the other measures *Cluster-s1_major* turns out to be the best predictor.

Attribute relevance with respect to *Information Gain* measure is as follows:

1. *Syl-Pos-in-Word1*
2. *Cluster-s1_major*
3. *Cluster-s1_minor*
4. *Stress Syl-No-in-Word*
5. *Break*
6. *Num-of-Syl-To-Prev-Minor-Break*
7. *Pos-of-Syl-Minor*
8. *Num-of-Syl-To-Fol-Major-Break*

9. *Num-of-Syl-in-Word*
10. *Word-Pos-in-Sent1*
11. *Pos-of-Word-Major*
12. *Word-No-in-Sent*
13. *Num-of-Word-To-Fol-Major-Break*
14. *Num-of-Syl-To-Fol-Minor-Break*
15. *Syl-Struct*
16. *POSw-1*
17. *Num-of-Word-To-Prev-Major-Break*
18. *Num-of-Syl-To-Prev-Major-Break*
19. *Num-of-Word-To-Fol-Minor-Break*
20. *POS*
21. *POSRoot*
22. *POSw+1*
23. *Num-of-Phn-in-Syl*
24. *Pos-of-Word-Minor*
25. *Pos-of-Syl-Major*
26. *Duration*
27. *Num-of-Word-To-Prev-MinorBreak*
28. *Syl-Type*
29. *Neg-Flag*
30. *ST4*
31. *ST2*
32. *ST3*
33. *ST1*
34. *Num-of-Word-in-Sent.*

Table 6-5: Information gain, Gain ratio and symmetrical uncertainty measures of the attributes with dependent variable in three accent prediction.

Attribute Index	Attribute Name	Information Gain	Gain Ratio	Symmetrical Uncertainty
10	NumofWordinSent	0	0	0
14	ST1	0.000445	0.000464	0.000396
16	ST3	0.001035	0.000477	0.000598
15	ST2	0.001315	0.001487	0.001209
17	ST4	0.001362	0.001194	0.00112
4	NegFlag	0.001546	0.011555	0.00217
2	SylType	0.001574	0.0016	0.001384
27	NumofWordToPrevMinorBreak	0.001947	0.002968	0.002
33	Duration	0.013851	0.008467	0.009464
30	PosofSylMajor	0.015827	0.030363	0.017466
31	PosofWordMinor	0.016846	0.01068	0.011746
11	NumofPhninSyl	0.019095	0.016422	0.015563
19	POSw+1	0.021362	0.006822	0.009661
8	POSRoot	0.022873	0.009615	0.012465
7	POS	0.022943	0.008377	0.011386
23	NumofWordToFolMinorBreak	0.023119	0.014697	0.016144
26	NumofSylToPrevMajorBreak	0.023726	0.011414	0.014082
25	NumofWordToPrevMajorBreak	0.024803	0.012845	0.015396
20	POSw-1	0.025837	0.008752	0.012178
3	SylStruct	0.026689	0.018277	0.019401
24	NumofSylToFolMinorBreak	0.028706	0.011767	0.015389
21	NumofWordToFolMajorBreak	0.032663	0.015309	0.019075
6	WordNoinSent	0.032811	0.015392	0.019172
29	PosofWordMajor	0.032992	0.030858	0.027956
13	WordPosinSent1	0.033041	0.031202	0.028119
9	NumofSylinWord	0.033867	0.015538	0.019516
22	NumofSylToFolMajorBreak	0.036041	0.015648	0.020055
32	PosofSylMinor	0.038753	0.038078	0.03357
28	NumofSylToPrevMinorBreak	0.044632	0.030221	0.032249
18	Break	0.07128	0.057245	0.056209
5	SylNoinWord	0.0922	0.06375	0.067364
1	Stress	0.0963	0.110415	0.089033
35	cluster-s1_minor	0.145895	0.089457	0.09986
34	cluster-s1_major	0.170166	0.112902	0.121621
12	SylPosinWord1	0.184099	0.102418	0.119211

CHAPTER 7

PITCH CONTOUR MODELING

General assumption for intonation modeling is that it can be successfully generated with fundamental frequency only, thus, the ultimate goal is to develop a model that generates the fundamental frequency contour of the original utterance.

A great majority of intonation modeling studies consider syllables as the smallest unit that bears prosody and uses syllable based intonational attributes in predicting pitch contours. Main trend is towards associating syllables with accents and boundary tones as described in Chapter1 and Chapter2. Pitch contour estimate is reconstructed using pre-determined values for the target labels. These values are generally computed using machine learning approaches. Syllables are taken as basic units in this study.

This chapter introduces stylization, prediction and reconstruction of pitch contours using syllable units. Two approaches are proposed for modeling syllable pitch contours. One of the approaches is phonetic in nature. The other is a kind of phonological method.

Prediction performance of each experiment involving regression tree is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Correlation Coefficient (CC) while quantitative analysis of decision trees is performed using True Positives rate (TP rate), False Positives rate (FP rate), Precision, Recall, and F-measures. A hypothetical confusion matrix (**Table 7-1**) for binary classification can be used to define these measures.

Table 7-1: Hypothetical confusion matrix of binary classification.

		predicted class	
		yes	no
actual class	yes	true positive (TP)	false negative (FN)
	no	false positive (FP)	true negative (TN)

$$\text{True Positives (TP) Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% \quad (7\text{-}1)$$

$$\text{False Positives (FP) Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} * 100\% \quad (7\text{-}2)$$

$$\text{Recall} = \text{True Positive (TP) Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% \quad (7\text{-}3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100\% \quad (7\text{-}4)$$

$$\text{F} - \textit{measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} * 100\% \quad (7\text{-}5)$$

7.1 Pitch Contour Modeling – A Phonetic Approach

Pitch contour prediction using phonetic approach involves stylization of pitch contours, sampling, clustering, prediction, and reconstruction. Following sections describe the steps involved in phonetic modeling of pitch contours.

Within the framework of this study, original pitch contours are stylized by means of non-parametric methods to make analysis and synthesis possible. Then, a codebook that represents different linguistic aspects of the speech waveform is developed from the stylized pitch contours.

7.1.1 Pre-Processing of Pitch Contours

Pitch contours reveal discontinuities mainly caused by unvoiced regions in speech signal. Some of these may correspond to abrupt changes in the pitch contours. These local changes have significant effects in model development hence removal of microprosody is an important issue to be handled in pitch contour modeling studies.

In this study, pitch contours of various types of sentences observed in Turkish are aimed to be modeled. Dynamic range of sentences changes depending on sentence type and speaker's emotional and physical state. Therefore, the dynamic ranges of the pitch contours are normalized for further studies.

Following sections present studies to remove microprosodic effects and normalization of syllable pitch contours.

7.1.1.1 Removal of Microprosodic Effects

In order to develop a model, microprosodic effects need to be eliminated from the original pitch contour. Microprosodic effects are mainly observed in the unvoiced regions of the speech signal as abrupt changes in the contour. Since most of the pitch contour analysis tools rely on the periodicity of the speech waveforms, unvoiced regions are discarded or erroneously calculated. In our pitch contour analysis and synthesis studies, PRAAT [Boersma and Weenink 2005; Wood 2005], a free speech processing tool that provides various analysis and synthesis functions, is employed. Succeeding paragraphs describe the methods used to handle microprosody removal.

Pitch contours of individual sound files are compouted using PRAAT. Minimum and maximum pitch values are set to be 75 Hz and 450 Hz, respectively. Speech waveform and pitch contour of the sentence 'özgüre beni beklemesini söylemedin mi' are given in **Figure 7-1**. The total time of the speech is 1.984 seconds. There are 44876 samples and the speech is sampled at 16 kHz. As shown in the lower part of the figure, the pitch contour is not continuous around unvoiced regions of the speech signal.

Discontinuities caused by the unvoiced regions of the speech signal are eliminated by *Interpolation* and *Smoothing*. By means of interpolation, the discontinuities are interpolated linearly and possible candidates are eliminated to produce the best path. Upper part of **Figure 7-2** shows interpolated pitch contour of the example sentence. As shown in the figure, discontinuities are mostly eliminated but there are still abrupt changes due to microprosody.

Remaining microprosody can be removed by means of a smoothing filter. A 10 Hz smoothing filter is used to remove microprosody from the original contour. Resultant interpolated and smoothed pitch contour is given in the lower part of **Figure 7-2**.

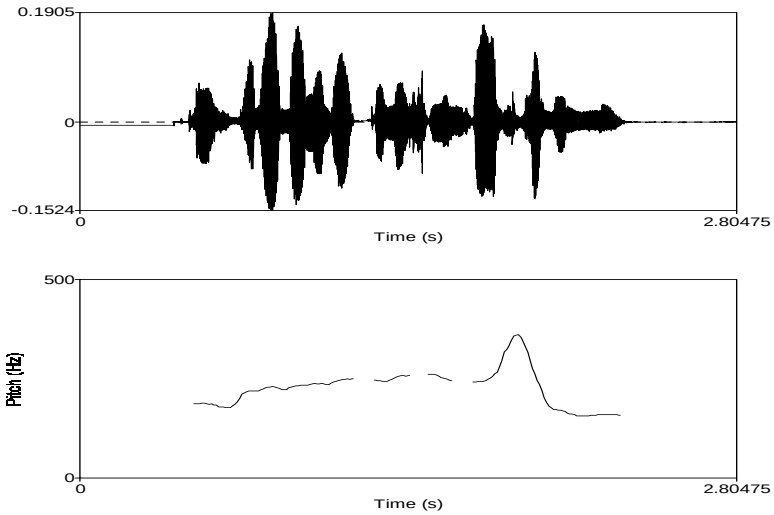

Figure 7-1: Sound waveform (upper window) and pitch contour (lower window) of the sentence ‘özgüre beni beklemesini söylemedin mi’.

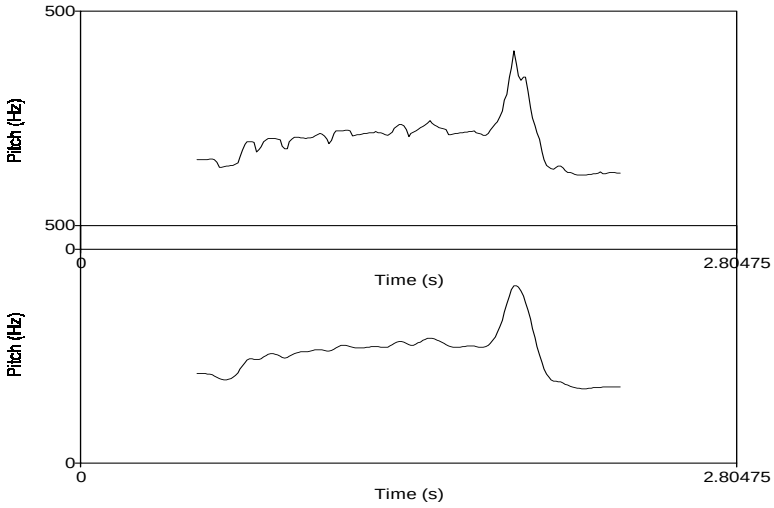

Figure 7-2: Interpolated (upper window) and smoothed (lower window) pitch contour of the example sentence.

For each syllable, 10 equidistant pitch values are selected from the pitch contour and a syllable pitch contour inventory is developed.

7.1.1.2 Normalization

Syllable pitch contours exhibit similar patterns such as falls, rises or combination of them as well as level curves. Therefore, contours of similar patterns are grouped to develop a syllable pitch contour codebook. However, sentences have different dynamic ranges and therefore pitch contours cannot be grouped to yield reliable information as they are. To eliminate level differences, normalization is performed.

For each sentence, minimum and maximum pitch values are determined from interpolated and smoothed F0 contours. These values are used for normalization (F0mins and F0maxs). Syllable pitch contour normalization is performed as follows:

$$F0_{i,norm1} = \frac{F0_i - F0_{i,min}}{F0_{i,max} - F0_{i,min}} \tag{7-6}$$

Here $F0_i$, $F0_{i,min}$, and $F0_{i,max}$ represent sample, minimum and maximum pitch values drawn from interpolated and smoothed pitch contour of the i^{th} sentence and, $F0_{i,norm1}$ is the normalized F0 value.

Normalized pitch contour of the example sentence is given in **Figure 7-3**. This scheme produces best match to the original pitch contour, however, for unobserved data, corresponding sentence minimum and maximum is not known; hence it requires prediction of sentence minimum and maximum as well.

7.1.2 Non-Parametric Representation of Pitch Contours

Normalized syllable pitch contours show similar pitch patterns. They can be identified by a set of predefined contours. Those pre-defined contours, templates, can be obtained by means of a clustering algorithm. K-means algorithm is used for clustering. Resulting templates mostly bear *level*, *rising* and *falling* patterns. Therefore, there should be at least four clusters in the templates to represent high, low, rising and falling patterns. An upper bound should also be set since increasing the number of clusters may result in perceptually equivalent pitch contour clusters.

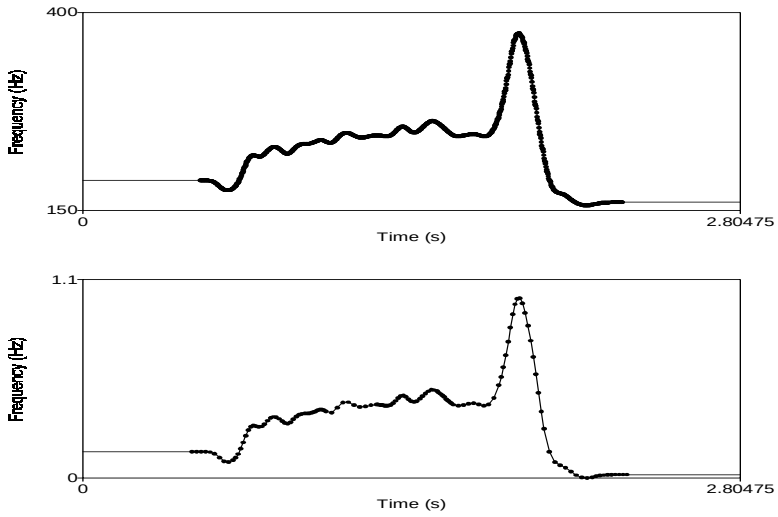

Figure 7-3: Original (upper) and normalized (lower) pitch contours of the example sentence ‘özgüre beni beklemesini söylemedin mi’.

For experimentation, 4, 5, 8, 16, and 24 clusters are generated from the normalized syllable pitch contours. Cluster centroids and corresponding elements for k-means clustering with 24 clusters are given in Appendix A.

For 24-clusters, the normalized pitch contour of the sentence ‘özgüre beni beklemesini söylemedin mi’ is reconstructed considering the cluster centroids only. **Figure 7-4** shows the sound waveform, corresponding glottal pulses, original and reconstructed pitch contours of the example sentence. Reconstructed pitch contour is obtained by using 24 cluster centroids given in Appendix A. The reconstructed pitch contour is perceptually equivalent to the original pitch contour although it exhibits deviations from the original contour.

Some of the 24-cluster centroids show almost similar contour shapes. For example, cluster centroid-1 and -2 as well as the centroid-6 and -7 are quite similar. For the rest of the cluster centroids such similarities can also be found. This is a direct result of the number of clusters used to partition the data space. Increasing the number of clusters results in centroids with almost similar shapes but with different levels. Since similar pitch contour shapes are represented by different cluster centroids, the predictive capability of the algorithms in determining unseen pitch contours is expected to get

lowered. Therefore, the number of clusters has a direct impact on contour prediction. So, determining the appropriate number is important for the success of the predictions.

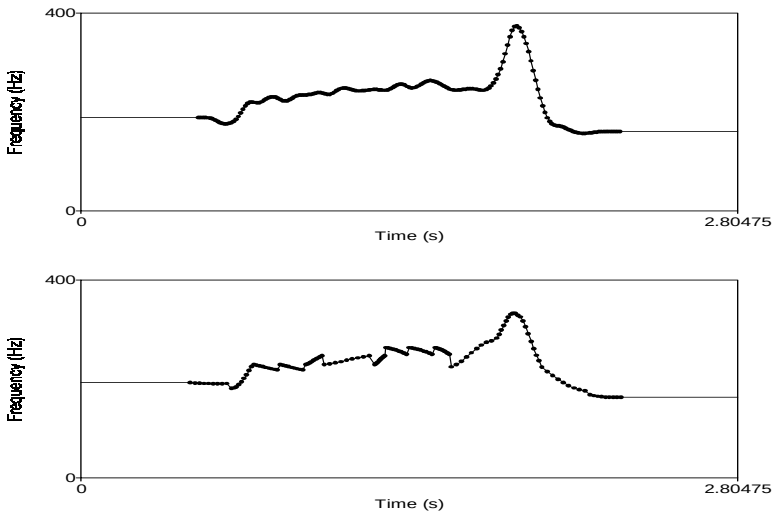

Figure 7-4: Original (upper) and reconstructed (lower) pitch contours of the example sentence ‘özgüre beni beklemesini söylemedin mi’.

Another critical point about the clusters is that although there are pitch patterns having multimodal characteristics, none of the cluster centroids exhibit such kind of patterns. This phenomenon does not cause any alterations in the reconstruction process; however it may cause performance reduction in the contour prediction process. This problem can be tackled by increasing the number of clusters. However, increasing clusters results in redundancy in partitioning of contours.

7.1.2.1 Decision Tree Learning Using Non-Parametric Representation

This section addresses syllable pitch contour prediction studies. Linguistic and acoustical attributes presented in Chapter 6 are mapped to 24-cluster centroids associated to the syllables in the database using decision trees. The data set is split into training (79%) and test (21%) sets to observe the prediction performance of the decision trees on unobserved data. Training and test datasets contain 12483 and 3384 syllables, respectively. Decision tree is developed using training set and prediction performance is evaluated using test set.

Results of the prediction of syllable pitch contours obtained are given in **Table 7-2** - **Table 7-4**. **Table 7-2** shows the size of the tree generated by decision tree learning, the kappa statistics, and the total number of correctly and incorrectly classified instances.

Table 7-2: Total number of leaves, size of the resultant tree and total number of correctly classified and misclassified syllables.

Number of Leaves	9538	
Size of the Tree	11046	
Kappa statistics	0.22	
Correctly Classified Instances	904	26.71 %
Incorrectly Classified Instances	2480	73.29 %

Cohen's **kappa coefficient** is a statistical measure of inter-annotator agreement. It is generally thought to be a more robust judge than simple percent agreement calculation. Kappa coefficient is defined as

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{7-7}$$

where $\Pr(a)$ is the relative observed annotator agreement, and $\Pr(e)$ is the probability that agreement is due to chance. $\Pr(a)$ is computed from the predictions of decision tree, i.e., proportion of total number of correct predictions over all predictions. $\Pr(e)$ is computed in the same manner but with different estimates: for calculating $\Pr(e)$, syllables are associated random cluster indices. A kappa coefficient of 1 means a statistically perfect modeling whereas a 0 means every model value was different from the actual value. A kappa statistic of 0.7 or higher is generally regarded as good statistic correlation, but the higher the value, the better the correlation.

The percentage of correct classifications is about %27. This is a rather low rate; however, as mentioned previously, most of the clusters centroids resemble each other and causes low true classification rate.

Table 7-3 reveals the TP rate, FP rate, Precision, Recall, and F-measure values for each cluster (last column indicates Class). As demonstrated by the table, the best TP rate is obtained for the 2^{nd} cluster and the worst TP rate is obtained for the 12^{th} cluster.

Table 7-3: TP rate, FP rate, Precision, Recall, and F-measure.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.79	0.04	0.66	0.79	0.72	2
0.42	0.04	0.27	0.42	0.32	11
0.36	0.08	0.28	0.36	0.32	20
0.35	0.03	0.29	0.35	0.31	23
0.33	0.09	0.18	0.33	0.23	7
0.31	0.06	0.33	0.31	0.32	1
0.26	0.03	0.26	0.26	0.26	13
0.25	0.02	0.26	0.25	0.25	14
0.21	0.04	0.26	0.21	0.24	6
0.21	0.03	0.17	0.21	0.19	9
0.20	0.03	0.20	0.20	0.20	8
0.20	0.05	0.23	0.20	0.22	3
0.20	0.03	0.25	0.20	0.22	19
0.18	0.03	0.13	0.18	0.15	21
0.17	0.04	0.27	0.17	0.21	17
0.11	0.02	0.12	0.11	0.11	10
0.10	0.01	0.15	0.10	0.12	24
0.09	0.02	0.13	0.09	0.11	22
0.08	0.03	0.11	0.08	0.10	18
0.08	0.02	0.12	0.08	0.09	4
0.07	0.01	0.14	0.07	0.10	5
0.07	0.03	0.08	0.07	0.08	16
0.06	0.02	0.10	0.06	0.08	15
0.04	0.01	0.07	0.04	0.05	12

Table 7-4 demonstrates the confusion matrix for resulting predictions. The confusion matrix shows the predicted clusters for each cluster in the test database. First column shows the frequency of each cluster in the test database. The last column holds the original cluster values while the rows correspond to the predicted clusters. The diagonal elements of the matrix indicate true predictions while the off-diagonals indicate false predictions.

Predicted pitch contours are compared with original pitch contours. Original and predicted pitch contours of the example sentence are given in **Figure 7-5**. Bold discontinuous curve belongs to the predicted contour while gray continuous curve corresponds to the original pitch contour. Predicted pitch contour is reconstructed using original $F0_{min}$ and $F0_{max}$.

Table 7-4: Confusion Matrix.

	Predicted Clusters																								
#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	
282	88	61	3	1	4	0	7	0	0	12	0	0	0	0	0	4	22	0	2	63	9	3	0	3	**1**
309	34	244	0	0	4	0	2	0	0	7	0	0	0	0	0	1	3	0	0	11	2	0	0	1	**2**
222	6	1	45	6	0	11	43	0	3	7	1	0	0	1	2	11	27	1	2	26	20	5	0	4	**3**
90	0	1	5	7	0	10	9	1	2	0	22	0	1	2	0	5	0	11	6	2	0	3	0	3	**4**
41	5	8	1	0	3	0	6	0	0	6	1	0	0	0	0	3	1	0	0	3	0	2	0	2	**5**
192	1	0	14	11	0	41	45	4	2	0	10	0	0	1	6	12	5	16	10	2	5	2	1	4	**6**
196	6	2	28	4	0	13	64	2	0	3	7	0	0	0	6	10	21	2	4	7	12	3	0	2	**7**
113	0	0	2	0	0	4	0	23	15	0	7	3	12	5	5	1	1	8	13	0	0	0	14	0	**8**
97	0	0	0	2	0	1	1	7	20	0	12	3	16	7	1	0	0	4	13	0	0	0	10	0	**9**
104	16	17	5	0	3	1	12	0	0	11	0	0	0	1	0	1	5	0	0	23	4	5	0	0	**10**
98	0	0	3	5	0	8	10	4	0	1	41	1	4	1	0	4	2	4	2	0	3	0	1	4	**11**
46	0	0	0	0	0	0	0	10	9	0	0	2	3	2	1	0	0	1	1	0	1	0	16	0	**12**
110	0	0	1	0	0	1	2	9	20	0	5	5	28	11	0	0	0	5	2	0	1	0	20	0	**13**
69	0	0	0	0	0	0	0	1	9	0	2	3	17	17	0	0	0	1	1	0	0	0	18	0	**14**
98	0	0	2	1	0	13	11	10	6	1	3	2	1	2	6	2	0	13	12	0	9	0	4	0	**15**
109	7	4	7	3	2	10	24	1	0	3	3	0	0	0	2	8	6	4	1	9	2	10	0	3	**16**
294	31	1	35	0	0	7	47	0	0	13	0	0	0	0	2	9	51	0	0	68	17	12	0	1	**17**
143	0	0	1	14	0	22	8	14	9	0	18	1	9	0	10	4	0	12	12	1	6	0	1	1	**18**
156	0	0	5	1	0	10	20	20	8	0	9	0	4	0	14	3	1	19	31	0	5	0	6	0	**19**
263	64	17	13	0	0	1	12	0	0	17	1	0	0	0	0	7	20	0	2	95	5	9	0	0	**20**
85	3	4	17	1	0	0	9	0	0	0	1	0	0	0	6	0	9	3	6	10	15	0	1	0	**21**
100	5	7	5	1	1	1	19	0	0	9	3	0	0	0	1	8	11	0	1	13	0	9	0	6	**22**
107	0	0	0	0	0	0	0	10	13	0	1	9	15	16	0	0	0	2	3	0	0	1	37	0	**23**
60	2	4	1	2	4	2	11	0	0	0	8	0	0	0	0	5	1	0	1	6	4	3	0	6	**24**

Original Clusters

Two striking points come out of the observations:

1) Most of the predictions reported to be FALSE classifications are caused by level mismatches: The predicted clusters more or less resemble the original cluster shapes but have level shifts.

2) The contours reconstructed from the predicted clusters mainly missed multimodalities observed in the original contour: Although the predicted cluster seems to match the starting point of the syllable contour, it cannot impose an inclination/declination or both on the syllable.

Level mismatches and lack of inclination can be observed from **Figure 7-5**..

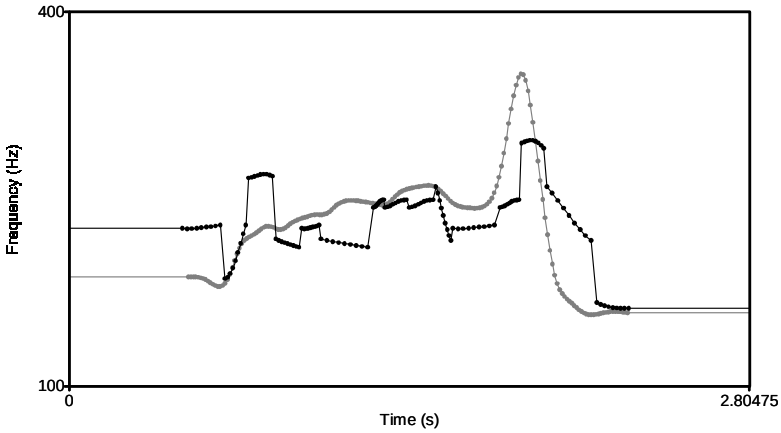

Figure 7-5: Original (gray) and predicted (black) pitch contours of the example sentence. 'özgüre beni beklemesini söylemedin mi'.

From preliminary prediction results given above, it is concluded that the predictions by the decision tree learning method used in the experiments are not satisfactory in the 24 cluster case. The main reason for this is that several clusters exhibit almost similar contour shapes. Therefore, although the decision tree predicts similar shapes, they are counted to be misclassified.

When reconstructed pitch contours using predicted clusters are considered, it is observed that the predictions cannot seize the multimodalities that exist in the original contour. This results from the lack of cluster centroids that represent multimodalities in the cluster set. Multimodalities seen in pitch contours correspond to the perceptual differences in the speech signal. Therefore, it is necessary to capture multimodalities of the pitch contour. So, a multi-level clustering algorithm is proposed and presented in the following sections.

7.1.3 Parametric Representation to Phonological Representation

According to our observations on pitch contours, relevant/perceptual pitch changes are demonstrated by multimodalities on pitch contour, i.e. peaks and valleys of pitch contours. Levels and pure rises/falls are used to link events (**Figure 7-6**). Predicted contours exhibit levels patterns. However, levels are not perceptually relevant. Therefore, a two-level clustering approach is used to capture dynamic cluster centroids.

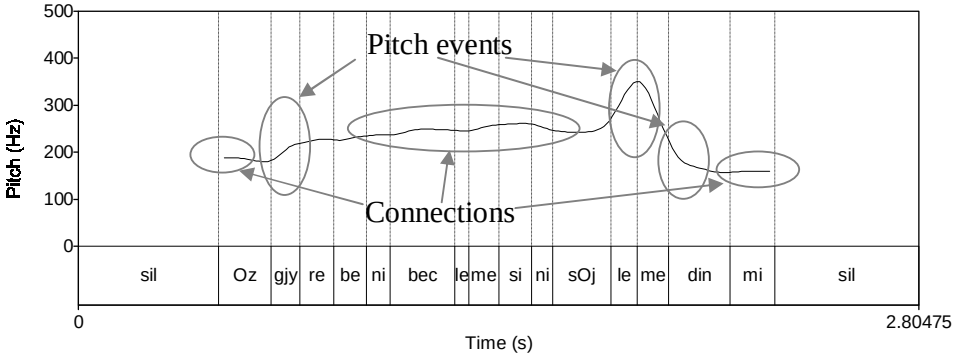

Figure 7-6: Pitch contour of the example sentence.

For two-level clustering, first, level differences among pitch contours are eliminated by a least-squares optimization algorithm. Cost function for the optimization problem is as follows:

$$[\hat{a}_i] = \arg\min_{a} \sum_{x=1}^{M} (a_i + f_i(x) - c)^2 \tag{7-8}$$

where c is a constant, and $f_i(x)$ represents the pitch points for the i^{th} syllable in the database. As mentioned before, every syllable pitch contour is sampled at every tenth of the overall syllable duration so the value for M is 10. Least-squares optimization is performed on $\{a_i\}$ values. For consistency, c is set to 0.5. Solution to the problem is given as

$$a_i = c - \frac{1}{10} \sum_{x=1}^{M} f_i(x) \tag{7-9}$$

K-means clustering algorithm is performed on level-removed syllable contours to obtain 100 clusters. These 100 centroids are fed to a second k-means algorithm to obtain 25 clusters. **Figure B-1 - Figure B-11** demonstrate 100 centroids grouped into 25 cluster centroids. As shown in the figures, resultant 25 clusters have almost similar patterns. Number of clusters is reduced 1) by eliminating clusters with centroids representing levels or pure rises and falls and 2) by merging clusters of the same shape (determined by

25 cluster centroids) into single clusters. Merged clusters mainly coincide to multimodal forms of pitch contours and various levels of them.

7.1.3.1 Decision Tree Learning Using Phonological Representation

According to our previous observations, selected clusters correspond to pitch events whereas eliminated contours correspond to connections. Therefore, a binary decision is performed on the resultant database to predict locations of pitch events. For binary prediction, syllables represented by codewords that are manually selected are assigned TRUE values while remaining syllables are assigned FALSE values. Prosodic attributes described in Chapter 6 are used to train decision tree. The statistical results for the binary prediction of pitch events are given in **Table 7-5** through **Table 7-7**.

However, this binary decision did not result in a better statistical performance than before. This is mainly due to the fact that there is a major difference between the number of TRUEs and FALSEs in the database. The number of TRUEs is 3039 while the number of FALSEs is 9444 in the training data and 763 versus 2621 in the test data set. Therefore, the decision tree algorithm cannot cope with the less observable cases which correspond to pitch events.

Table 7-5: Total number of leaves, size of the decision tree and total number of correctly classified and misclassified syllables.

Number of Leaves	860	
Size of the tree	1164	
Kappa statistics	0.38	
Correctly Classified Instances	2729	80.64%
Incorrectly Classified Instances	655	19.36%

Table 7-6: TP rate, FP rate, Precision, Recall, and F-measure for each class.

TP Rate	FP Rate	Precision	Recall Class	F-Measure	Class
0.92	0.58	0.85	0.92	0.88	False
0.43	0.08	0.6	0.43	0.5	True

Table 7-7: Confusion matrix of binary prediction.

Predictions			
False	True		
2405	216	False	Original
439	324	True	

TRUE and FALSE labels associated to syllables in the database are examined. According to our observations, some of the selected clusters do not really correspond to pitch events (i.e. syllables with very smooth rises and falls are also considered as pitch events) and some of the clusters that are previously discarded and assigned FALSE should be considered as pitch events (i.e. there are sharp rises and falls that are not included in selection but cause audible/ perceptual pitch changes). Therefore, cluster centroid selection is performed one more time taking into account the dynamic ranges of the clusters: Cluster centroids having dynamic ranges greater than or equal to 40 Hz are selected as prominent (TRUE) syllables and leave the others as FALSE. Threshold value corresponds approximately 10% of the difference of minimum of $F0_{i,min}$ and maximum of $F0_{i,max}$ values.

Correspondign results are given in **Table 7-8** and **Table 7-10**.

Table 7-8: Total number of leaves, size of the binary classification tree and total number of correctly classified and misclassified syllables.

Number of Leaves	1555	
Size of the tree	2010	
Kappa statistic	0.41	
Correctly Classified Instances	2426	80.8806 %
Incorrectly Classified Instances	958	28.3097 %

Table 7-9: TP rate, FP rate, Precision, Recall, and F-measure values for each class.

TP Rate	FP Rate	Precision	Recall Class	F-Measure	Class
0.60	0.19	0.71	0.60	0.65	False
0.81	0.40	0.72	0.81	0.76	True

Table 7-10: Confusion matrix for the binary prediction of pitch events.

Predictions			
False	True		
886	598	False	Original
360	1540	True	

According to the statistical results, prediction using codebook of varying size does not result in reasonable performances. However, such kind of an implementation would benefit by increasing the dimension of the database and making the cluster distributions even. Another important result rising from the current study is that microprosody still plays an important role in the clustering algorithm since microprosody cannot be removed completely from the pitch contours and causes spurious pitch contours for syllables. **Figure 7-7** demonstrates significant microprosodies. The speech signal and the pitch contour given in the figure belong to the sentence 'mikroorganizmaları yok etmek için şok ısıtma ve soğutma yöntemi kullanılır' (shock heating and cooling method is used to exterminate microorganisms). Though, successive application of hierarchical clustering may provide better performances.

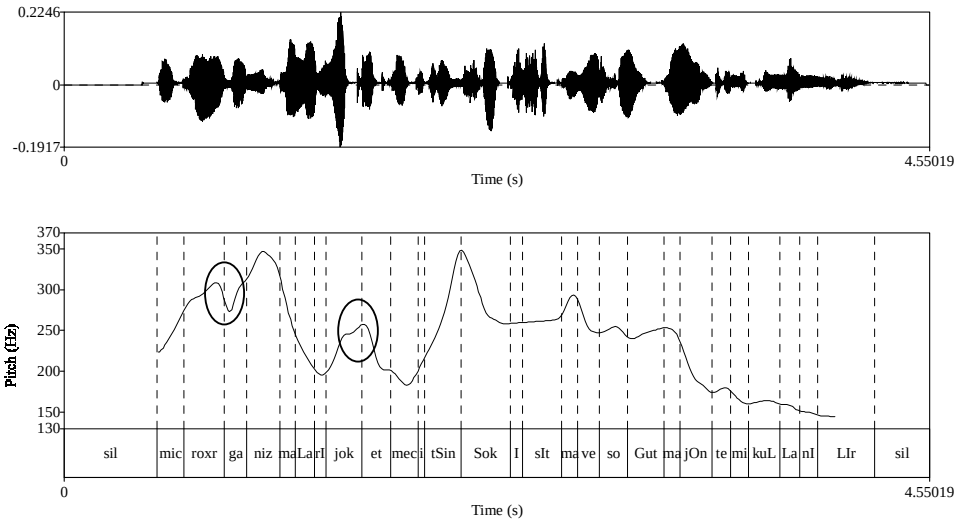

Figure 7-7: Circles mark significant microprosody that can not be removed completely.

7.2 Pitch Contour Modeling - A Phonological Approach

Results of early attempts on pitch contour modeling using pre-defined clusters for each syllable and predicting them through machine learning led us develop a new approach that involves less contour dependency by decreasing the number of pre-defined clusters. Proposed approach mainly relies on representing pitch contours as a sequence of discrete events.

In this approach, syllables are assigned to pitch accents depending on the decisions made over syllable pitch contours. Our primary aim is to distinguish accented syllables from not-accented ones by means of a binary decision. For each accented syllable, a prominence level is predicted by means of regression trees.

Resulting work resembles to the intonation modeling studies that involve ToBI labeling where each syllable of the database is labeled according to their accent status and then a binary decision is performed whether the current syllable is accented or not depending on acoustic and linguistic features derived from the speech and text corpora. Then, each accent is further discriminated by another decision tree or some other learning method such as a neural network. After identifying the labels of each syllable, prominence levels are found. This is performed by means of numeric prediction methods. Phrase and boundary tones are assigned and corresponding prominence levels are predicted with almost same procedures but using different feature sets.

Studies incorporating ToBI transcription system rely on label-rich corpora. Generally, pitch contours are manually transcribed using ToBI system and attributes are obtained by deep linguistic analyses. Within the course of this study, such a corpus is not available. Hence, intermediate representations, such as ToBI labels, that represent pitch contours do not exist as well. Lexical stress scheme is available however it is still insufficient to resolve stress assignments for compound words or even more complex forms such as noun phrases.

Previous studies on other languages reported that pitch accents are observed as local minima and maxima on pitch contours and should be aligned with the lexically stressed syllable of the words that are accented [Pierrehumbert 2000]. Acoustic behavior and their perceptual equivalences are captured by means of perceptual listening tests. According to our observations, sharper rises turn out to be perceptually more prominent from the rest of

the pitch contours. It is concluded that selecting rises instead of maxima would be better for predicting prominent, accented, syllables.

Developed pitch accent assignment algorithm locates accents according to syllable pitch contours. For the example sentence given in **Figure 7-6**, the data used to assign pitch accents are given in **Table 7-11**. First column of the table is related to the index of the sentences within the database. In the second column, the syllable identities in their SAMPA [Wells 2003] format are given. The third, fourth and fifth columns correspond to the position of syllable in word (*SylPosinWord*), position of word in sentence (*WordPosinSent*), number of syllables in word (*NumofSylinWord*) obtained from the text. The sixth column of the table holds the slope amplitudes calculated from the original syllable pitch contours. The seventh column manifests the sign of the slopes given in the sixth column. Mean F0 values of corresponding syllables are given as the last column of the table. Slope of syllable *k* is calculated as follows:

$$Slope_k = \frac{\left(\frac{F0_k(8)+F0_k(9)}{2}\right)-\left(\frac{F0_k(0)+F0_k(1)}{2}\right)}{D_k} \qquad (7\text{-}10)$$

where $F0_k(x)$ and D_k correspospond to 10 equidistant F0 values picked from the syllable pitch contour and syllable duration.

Table 7-11: Data used by pitch accent assignment algorithm.

ID	Syllable Label	SylPosinWord	WordPosinSent	NumofSylinWord	Slope Amp	Sign	Mean
0	Oz	1	1	3	-56.94	-1	184
0	gjy	2	1	3	351	1	202
0	re	3	1	3	96.18	1	224
0	be	1	2	2	101.36	1	229
0	ni	2	2	2	45.13	1	237
0	bec	1	3	5	22.83	1	242
0	le	2	3	5	-23.9	-1	244
0	me	3	3	5	118.7	1	252
0	si	4	3	5	56.09	1	262
0	ni	5	3	5	-164.21	-1	253
0	sOj	1	4	4	23.1	1	248
0	le	2	4	4	902.56	1	320
0	me	3	4	4	-995.55	-1	303
0	din	4	4	4	-127.9	-1	172
0	mi	1	5	1	11.1	1	160

Pitch accent assignment algorithm searches for relevant positive and negative slopes for every syllable of candidate words which are selected with respect to positive and negative thresholds: Words having syllables with slopes greater (less) than positive (negative) threshold are selected as candidates for positive (negative) accents.Threshold value plays an important role in pitch accent assignment process. If the value is chosen too small, every tiny slope is considered as a potential candidate. If the threshold is chosen too large, then most of the potential pitch accents are discarded.

Among the syllables of the candidate words, the ones with considerable positive and negative slopes are associated with positive and negative pitch accents. Relevancy is determined by means of slope amplitudes calculated from the interpolated and smoothed contours.

For positive accent assigment, two syllables with the highest mean F0 among all positive sloped syllables in the candidate word are selected. This choice is based on the observations that

- the sharpest slope may not be aligned with the syllable with the highest mean F0,
- highest mean F0 on a syllable may be an indicator of pitch accent on the syllable,
- both syllables have higher mean F0 than the rest of the syllables in the candidate word.

Positive accent is associated to the syllable with the lower mean F0 if its slope amplitude exceeds a scaled version of the highest mean F0 valued syllable's slope magnitude or to the syllable with the highest mean F0 otherwise.

Negative accents are assigned to the first syllable with the largest negative slope in the candidate word.

Figure 7-8 shows positive and negative pitch accents accents associated to the syllables of the sentence 'mikroorganizmları yok etmek için şok ısıtma ve soğutma yöntemi kullanılır' by accent assignment algorithm.

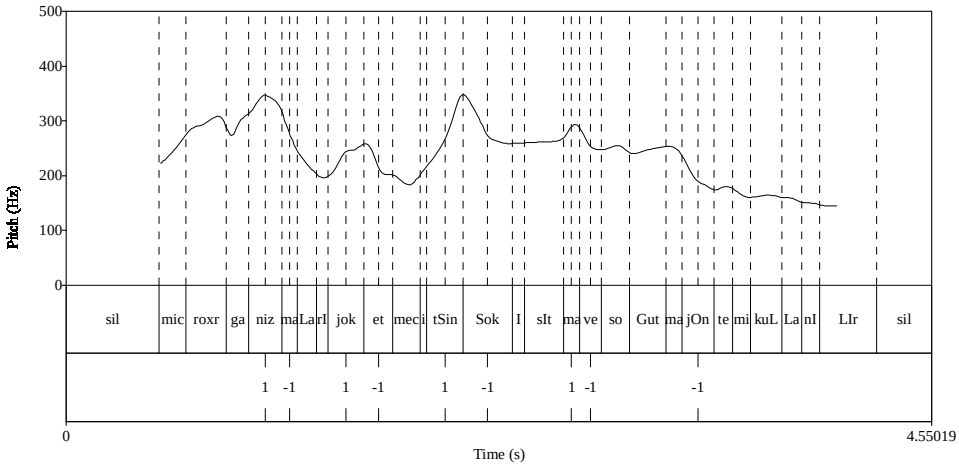

Figure 7-8: Positive (1's) and negative (-1's) pitch accents of the sentence 'mikroorganizmları yok etmek için şok ısıtma ve soğutma yöntemi kullanılır'.

The process of deciding whether a syllable is accented or not, can be viewed as a binary classification problem.

7.2.1 Prediction of Pitch Contour Parameters

A three-step procedure is followed for modeling pitch contours. First step involves pitch accent placement, second step involves numeric prediction of accent slopes. In the last step, syllable pitch contours are re-generated using slopes estimates and sentence pitch contours are reconstructed by concantenating syllable pitch contours..

In the first part which involves a classification task, the decision tree algorithm (J48) of WEKA package [Witten and Frank 1999] is used. In the second part, the problem is to predict slope amplitudes, hence, requires numeric prediction methods. Therefore, at this step, regression tree algorithm (REPTree) of the WEKA package [Witten and Frank 1999] is used.

For accent location prediction, two experiment sets are utilized. Both experiment sets utilize prosodic attributes presented in Chapter 6. In the first experiment set (Experiment Set 1), the decision tree is used to predict accent states of the syllables which are determined by the accent assignment algorithm. The syllables are either positive/negative accented or not-accented (cf. **Figure 7-8**). Hence, three accent states are predicted in Experiment Set 1.

In the second experiment set (Experiment Set 2), positive and negative accent types are merged to construct a single *accented* class.

For accent slope prediction, the slope values computed from the normalized syllable pitch contours are predicted by the regression tree. Accent states of syllables are also used in the learning. Slopes are predicted considering the experimental setups employed in accent location prediction.

For each prediction task, the performance of the learning algorithm is evaluated using three methods given below.

Evaluation on training data (Training): Training data can be used to observe the performance of the decision tree however performance on the training data is not a good indicator of future performance, i.e. performance on new data: During the learning process, the classifier tries to make the best prediction for every sample in the training database, so the resulting error rate would be an optimistic one and very likely to overestimate the true predictive performance of the learning method. However, it is still useful to look at these results, for they generally represent an upper bound on the model's performance. Therefore, for each of the prediction tasks, an evaluation on training dataset is performed to have an idea about decision tree's prediction capacity.

Evaluation on test data (Test): To predict the performance of decision tree on new data, the performance should be calculated over a *test set*. The data in the test set is not used in the learning phase of the decision tree. The test data may be completely distinct from the training data or may be part of it. The only constraint on the test data is: The test data should not be employed in the development of the classifier. Hence, for each prediction task, evaluations on test data are performed.

Evaluation using 10-fold cross validation (CrossVal): In *10-fold cross validation*, the data is divided into *10* subsets of equal size. The decision tree is trained *10* times, each time leaving out one of the subsets from training. The remaining subset is used to compute the error rate. The overall error for the classifier is the average of errors computed in *10* training. In most practical applications, 10-fold cross validation is used for predicting error. Since 10-fold cross-validation method is a good indicator of decision tree's future performance, 10-fold cross validation is performed.

In order to perform **Training** and **Test**, the database is split into two subsets: Training and Test datasets. Training dataset is used to develop appropriate classification/regression

tree while test dataset is used to evaluate the performance of resultant tree. The training dataset consists of 12483 samples of the syllable database and the test dataset consists of the remaining 3384 samples of the syllable database. These values correspond to 78.67% and 21.33% of the syllable database, respectively.

Following sections summarizes the results obtained from accent location and slope prediction.

7.2.1.1 Accent Prediction

Statistical observations given in tables **Table 7-12** through **Table 7-18** belong to the accent prediction experiments. For both of the experiment sets, **Test** and **CrossVal** performances are lower than **Training** performances however they are more reliable.

Table 7-12: Correct and incorrect classification rates for Experiment Set 1 & 2.

			Number of Syllables	**Percentage**
Experiment Set 1	Training	Correctly Classifed Syllables	10515	84.23%
		Incorrecly Classified Syllables	1968	15.77%
	Test	Correctly Classifed Syllables	2523	74.56%
		Incorrecly Classified Syllables	861	25.44%
	CrossVal	Correctly Classifed Syllables	11988	75.55%
		Incorrecly Classified Syllables	3879	24.45%
Experiment Set 2	Training	Correctly Classifed Syllables	10558	84.58%
		Incorrecly Classified Syllables	1925	15.42%
	Test	Correctly Classifed Syllables	2568	75.89%
		Incorrecly Classified Syllables	816	24.11%
	CrossVal	Correctly Classifed Syllables	12025	75.79%
		Incorrecly Classified Syllables	3842	24.21%

According to **Table 7-12** , best correct classification rates are observed in the cases where evaluations are performed on training set in both experiment sets. These results illustrate the upper limits of the decision trees. In Experiment Set 1, classification rates are worser than that of the cross validation performance which denotes that selected test to evaluate decision tree performance is not optimal. It is also observed that classification rates obtained in both experiment sets are almost same although Experiment Set 1 is more complicated since it relies on the prediction of three accent states: positive, negative and no-accent.

Kappa coefficients for both experiment sets are given in **Table 7-13**. Kappa statistics of Experiment Set 1 are better than those of Experiment Set 2. Evaluations using training sets approach 0.7 which is regarded as good statistic correlation.

Table 7-13: Kappa statistics of Experiment Set 1 & 2.

		Kappa Statistics
Experiment Set 1	Training	0.68
	Test	0.47
	CrossVal	0.50
Experiment Set 2	Training	0.65
	Test	0.45
	CrossVal	0.46

The confusion matrices of Experiment Set 1 & 2 are given in **Table 7-14** and **Table 7-16**, respectively. Diagonal entries of the tables correspond to correct predictions while off-diagonals correspond to false predictions. Confusion matrix of Experiment Set 1 (**Table 7-14**) shows that decision trees cannot discriminate accented syllables from the not-accented syllables. However, they perform a better discrimination in between positive and negative accented syllables.

Table 7-14: Confusion matrices observed in Experiment Set 1 (*positive*, *negative* and *no-accent*).

		Classified as:	no-accent	positive	negative
Experiment Set 1	Training	no-accent	7411	260	335
		positive	577	1685	59
		negative	677	60	1419
	Test	no-accent	1906	196	144
		positive	219	344	16
		negative	256	30	273
	CrossVal	no-accent	8889	627	736
		positive	1185	1600	115
		negative	1116	100	1499

In order to compare performances of correct class predictions of both experiment sets, confusion matrix of Experiment Set 1 (**Table 7-14**) is converted to two-class confusion

matrix given in **Table 7-15**. The converson is performed by merging the statistics of positive and negative classes. Comparison of the cofusion matrices of both experiment sets (**Table 7-15** and **Table 7-16**), it is observed that prediction of *accented* class is better in Experiment Set 2 but Experiment Set 1 predicts *no-accent*s better.

Table 7-15: Confusion matrices observed in Experiment Set 1 (*accented* vs *no-accent*).

		Classified as:	no-accent	accented
Experiment Set 1	Training	no-accent	7411	595
		accented	1254	3223
	Test	no-accent	1906	340
		accented	475	663
	CrossVal	no-accent	8889	1363
		accented	2301	3314

Table 7-16: Confusion matrices observed in Experiment Set 2 (*accented* vs *no-accent*).

		Classified as:	no-accent	accented
Experiment Set 2	Training	no-accent	7335	671
		accented	1254	3223
	Test	no-accent	1881	365
		accented	451	687
	CrossVal	no-accent	8671	1581
		accented	2261	3354

As shown in **Table 7-14** - **Table 7-16**, the total number of *no-accent*s are greater than the total number of positive and negative (*accented*) classes. The number of *no-accent*ed syllables in Training, Test and CrossVal are 8006, 2246 and 10252 accented syllables are 4477, 1138, and 5615, respectively. Hence, resulting decision trees predict *no-accent*s more accurately than *accented* classes.

TP rates, FP rates, Precisions, Recalls and F-Measures for Experiment set 1 & 2 are given in **Table 7-17** and **Table 7-18**, respectively. It is observed that best TP rates are observed for *no-accent*s.

Table 7-17: TP rate, FP rate, Precision, Recall and F-measures of Experiment Set 1.

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Experiment Set 1	Training	0.93	0.28	0.86	0.93	0.89	no-accent
		0.73	0.03	0.84	0.73	0.78	positive
		0.66	0.04	0.78	0.66	0.72	negative
	Test	0.85	0.42	0.80	0.85	0.82	no-accent
		0.59	0.08	0.60	0.59	0.60	positive
		0.49	0.06	0.63	0.49	0.55	negative
	CrossVal	0.87	0.41	0.79	0.87	0.839	no-accent
		0.55	0.06	0.69	0.55	0.61	positive
		0.55	0.07	0.64	0.55	0.59	negative

Table 7-18: TP rate, FP rate, Precision, Recall and F-measures of Experiment Set 2.

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Experiment Set 2	Training	0.92	0.28	0.85	0.92	0.88	no-accent
		0.72	0.08	0.83	0.72	0.77	accented
	Test	0.84	0.40	0.81	0.84	0.82	no-accent
		0.60	0.16	0.65	0.60	0.63	accented
	CrossVal	0.85	0.40	0.79	0.85	0.82	no-accent
		0.60	0.15	0.68	0.60	0.64	accented

Figure 7-9 shows the decision tree generated in Experiment Set 1 for **Training** and **Test** cases. The tree has six splitting levels. As shown in the table, the first split occurs at *cluster-s1_major* attribute which corresponds to the most significant attribute of the attribute set. The significance order of the attributes can be considered by means of the ordering in the trees, that is, the higher the attribute is observed on the branching, the more significant the attribute is. Second level splits occur at *SylPosinWord1*, *Stress*, and *SylnoinWord*. These two splits correspond to the most relevant prosodic attributes.

Third level splits occur at *NumofSylToPrevMinorBreak*, *PosofWordMajor*, *Duration*, and *NumofSylinWord*. Fourth levels splits occur at *Break*, *NumofSylToPrevMinorBreak*, and *NumofSylTo*-PrevMajorBreak attributes. The fifth level splits are *NumofWordToFolMinorBreak*. At the last level, the only splitting is observed at the attribute *NumofWordToPrevMajorBreak*. The rest of attributes do not play any role in the resultant decision therefore the rest of the attributes are irrelevant to accent prediction using decision trees.

In the tree structure, a colon (:) introduces the class label that is assigned to a particular leaf, followed by the number of instances that reach that leaf, expressed as a decimal number. The decimals are used because of the way the algorithm uses fractional instances to handle missing values [Witten and Frank 1999].

```
cluster-s1_major = NONE: cl0 (547.0/150.0)
cluster-s1_major = cl0
|  Stress = N
|  |  NumofSylToPrevMinorBreak <= 0
|  |  |  Break = SI: cl0 (0.0)
|  |  |  Break = M: cl0 (0.0)
|  |  |  Break = SF: cl0 (203.0/40.0)
|  |  |  Break = F: cl (148.0/73.0)
|  |  |  Break = I: cl0 (0.0)
|  |  |  Break = I/F: cl0 (4.0)
|  |  NumofSylToPrevMinorBreak > 0: cl0 (4511.0/942.0)
|  Stress = A
|  |  PosofWordMajor = I: cl (320.0/34.0)
|  |  PosofWordMajor = M
|  |  |  NumofSylToPrevMinorBreak <= 0: cl (433.0/142.0)
|  |  |  NumofSylToPrevMinorBreak > 0
|  |  |  |  NumofWordToFolMinorBreak <= 0
|  |  |  |  |  NumofWordToPrevMajorBreak <= 2: cl0 (161.0/77.0)
|  |  |  |  |  NumofWordToPrevMajorBreak > 2: cl (312.0/110.0)
|  |  |  |  NumofWordToFolMinorBreak > 0: cl0 (1129.0/431.0)
|  |  PosofWordMajor = F: cl0 (365.0/72.0)
cluster-s1_major = cl
|  SylPosinWord1 = I: cl-1 (1392.0/454.0)
|  SylPosinWord1 = M: cl-1 (228.0/89.0)
|  SylPosinWord1 = F: cl0 (353.0/139.0)
|  SylPosinWord1 = Single
|  |  Duration <= 0.168: cl-1 (109.0/50.0)
|  |  Duration > 0.168: cl0 (132.0/67.0)
cluster-s1_major = cl-1
|  SylNoinWord <= 1: cl0 (426.0/164.0)
|  SylNoinWord > 1
|  |  NumofSylinWord <= 2
|  |  |  NumofSylToPrevMajorBreak <= 8: cl0 (142.0/46.0)
|  |  |  NumofSylToPrevMajorBreak > 8
|  |  |  |  NumofSylToFolMinorBreak <= 3: cl (128.0/45.0)
|  |  |  |  NumofSylToFolMinorBreak > 3: cl0 (108.0/52.0)
|  |  NumofSylinWord > 2: cl0 (1332.0/90.0)
```

Figure 7-9: Decision tree obtained in Experiment Set 1 using training set.

Figure 7-10 demonstrates the resultant decision tree generated in Experiment Set 2 for **Training** and **Test** cases.The tree has five splitting levels. As the table reveals, the first split occurs at *SylPosinWord1* attribute. Second level splits occur at *cluster-s1_major*, *NegFlag*, *PosofSylMajor*, and *NumofWordToFolMajorBreak*. Third level splits occur at *NumofSylToFolMajorBreak*, *NumofWordToPrevMinorBreak*, *PosofWordMinor*, and *SylType*. Fourth levels splits occur at *Duration*, *cluster*-s1_minor, *POSw+1*, *ST4*,

NumofSylToPrevMajor-Break, and POSRoot attributes. Last level splits occur at *ST4*, and *SylNoinWord*. The rest of attributes do not play any role in the resultant decision therefore they are irrelevant to accent prediction using decision trees. When current decision tree is compared with the former (**Figure 7-9**), it can be seen that they are different from each other although they share common attributes but at different splits.

```
SylPosinWord1 = I
|  cluster-s1_major = NONE: cl0 (455.0/73.0)
|  cluster-s1_major = cl0
|  |  NumofSylToFolMajorBreak <= 4: cl_cl-1 (150.0/73.0)
|  |  NumofSylToFolMajorBreak > 4: cl0 (1383.0/341.0)
|  cluster-s1_major = cl_cl-1
|  |  NumofWordToPrevMinorBreak <= 1: cl_cl-1 (1067.0/320.0)
|  |  NumofWordToPrevMinorBreak > 1
|  |  |  Duration <= 0.201
|  |  |  |  ST4 = MI-Ques: cl0 (113.0/47.0)
|  |  |  |  ST4 = NONE: cl_cl-1 (402.0/144.0)
|  |  |  |  ST4 = WH-Ques: cl_cl-1 (40.0/13.0)
|  |  |  Duration > 0.201: cl0 (116.0/41.0)
SylPosinWord1 = M
|  NegFlag = FALSE: cl0 (4170.0/597.0)
|  NegFlag = TRUE
|  |  NumofSylToFolMajorBreak <= 17: cl_cl-1 (103.0/48.0)
|  |  NumofSylToFolMajorBreak > 17: cl0 (102.0/31.0)
SylPosinWord1 = F
|  PosofSylMajor = I: cl0 (0.0)
|  PosofSylMajor = M
|  |  PosofWordMinor = I
|  |  |  cluster-s1_minor = NONE: cl0 (1.0)
|  |  |  cluster-s1_minor = cl0: cl_cl-1 (841.0/212.0)
|  |  |  cluster-s1_minor = cl_cl-1
|  |  |  |  SylNoinWord <= 2: cl_cl-1 (171.0/72.0)
|  |  |  |  SylNoinWord > 2: cl0 (194.0/56.0)
|  |  PosofWordMinor = M
|  |  |  POSw+1 = PRON: cl0 (12.0/5.0)
|  |  |  POSw+1 = NOUN: cl0 (392.0/124.0)
|  |  |  …
|  |  PosofWordMinor = F
|  |  |  ST4 = MI-Ques: cl0 (129.0/28.0)
|  |  |  ST4 = NONE: cl_cl-1 (462.0/93.0)
|  |  |  ST4 = WH-Ques: cl0 (36.0/4.0)
|  PosofSylMajor = F: cl0 (481.0/86.0)
SylPosinWord1 = Single
|  NumofWordToFolMajorBreak <= 1: cl_cl-1 (165.0/40.0)
|  NumofWordToFolMajorBreak > 1
|  |  SylType = H
|  |  |  NumofSylToPrevMajorBreak <= 8: cl0 (107.0/38.0)
|  |  |  NumofSylToPrevMajorBreak > 8: cl_cl-1 (114.0/55.0)
|  |  SylType = L
|  |  |  POSRoot = NOUN: cl0 (5.0/2.0)
|  |  |  POSRoot = PRON: cl_cl-1 (5.0/2.0)
|  |  |  POSRoot = VERB: cl_cl-1 (0.0)
|  |  |  POSRoot = TELL: cl_cl-1 (0.0)
|  |  |  POSRoot = QUES: cl_cl-1 (107.0/28.0)
|  |  |  POSRoot = CNOUN: cl_cl-1 (0.0)
|  |  |  POSRoot = MODAL: cl_cl-1 (0.0)
|  |  |  POSRoot = POSTP: cl_cl-1 (0.0)
|  |  |  POSRoot = CONJ: cl_cl-1 (120.0/42.0)
|  |  |  POSRoot = ADV: cl_cl-1 (1.0)
|  |  |  …
```

Figure 7-10: Decision tree obtained for Experiment Set 2 using training set.

The best performance in accent prediction experiments is obtained by means of binary prediction, that is, the case where two accent types are merged into *accented* class and not-accented ones are used for the other class. The evaluations on the test set shows 74.56% correct prediction for triple accent prediction and 75.89% correct prediction for binary prediction of accents. When the two experiments are examined in detail, it is seen that the former predicts the not-accented syllables better while the latter does it for accented syllables.

Among the three evaluation methods utilized in binary prediction, the best performance is obtained via EvalA, as expected. The second best statistical performance is obtained using EvalB. The worst performance is obtained by means of EvalC; however, it is the most reliable evaluator.

In both of the experiments, the first two of the evaluations are recovered from the split train and test datasets while the third one belongs to the whole syllable database.

Although the results are much more promising then our previous attempts on classifying pitch accents by means of syllable pitch contour clusters, they still need improvement. When the results are examined in detail, it is observed that prediction accuracy for not-accented classes (class0) is better than the other classes in both of the schemes. This is mainly due to the uneven distribution of accented and not-accented syllables in the database. In the first set of experiments, the percentages of class1, class-1, and class0 are 18.59%, 17.27%, and 64.14% in the training database and 17.11%, 16.52%, and 66.37% in the test database, respectively. In the second set of experiments where two accent types are merged into one accent category (class1_class-1), the percentages of the accented and not-accented syllables are 35.86% and 64.14% in the training dataset and 33.63% and 66.37% in the test database, respectively. Not-accented syllables may cover feature combinations that can also be observed for accented syllables. Significant difference in the amounts of not-accented and accented syllables in the database can explain the tendency of decision trees to label a majority of syllables as not-accented.

7.2.1.2 Slope Prediction for Accented Syllables

In the second stage of pitch contour modeling, syllables are associated with the corresponding slopes by means of numeric prediction methods. With slope and duration information, abstract accent labels can be transformed into continuous pitch contours.

Slope prediction is performed considering the two experimental setup employed in accent prediction for three evaluation methods (**Training**, **Test**, and **CrossVal**). In the first experiment set (Experiment Set 1), slopes are predicted for three accent states (positive, negative, and no-accent). In Experiemtn Set 2, slopes are predicted for two accet states (accented vs no-accent).

To decrease variation of data, slope values are computed from the normalized pitch contours instead of original pitch values. The histogram plot of syllable slopes computed from the normalized pitch contours are illustrated in **Figure 7-11** - **Figure 7-13**. **Figure 7-11** demonstrates the slope histogram of all syllables. In **Figure 7-12**, only the slope histogram of syllables associated to *no-accent*s is demonstrated. As revealed by the figure, the slopes of no-accents vary from -5 to 4. This range partially overlaps the slope ranges of negative accent (**Figure 7-13** – left window) and positive accent (**Figure 7-13** – right window).

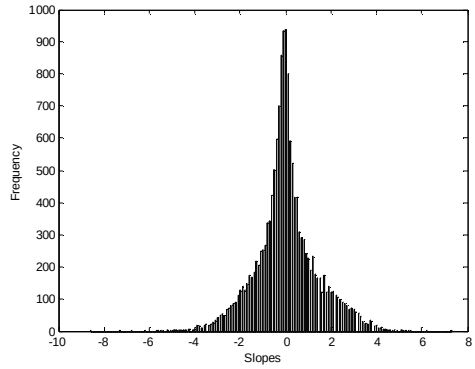

Figure 7-11: Histogram plot of the slopes of all syllables in the database.

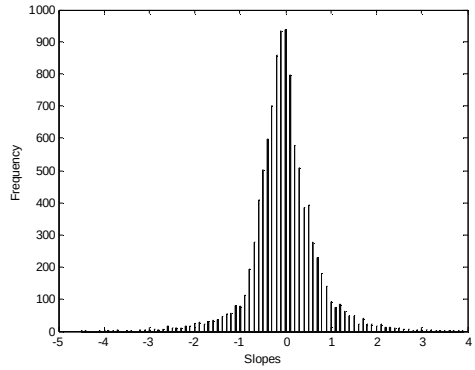

Figure 7-12: Histogram plot of the slopes of syllables associated to no-accent.

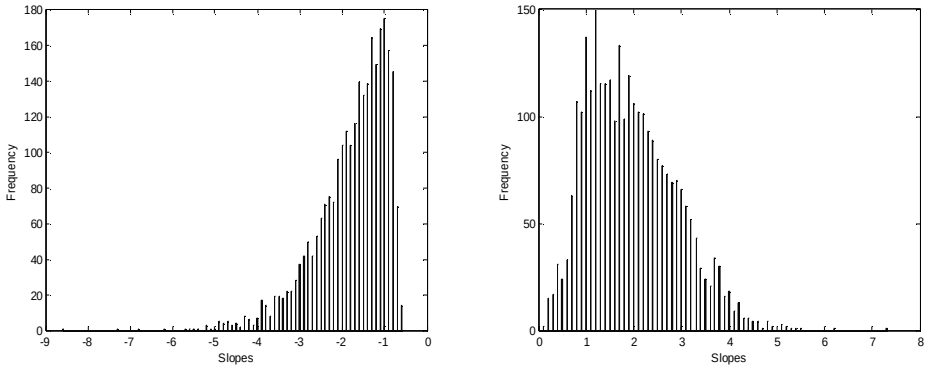

Figure 7-13: Histogram plots of the slopes of syllables associated to negative (Left window) and positive (Right window).

Table 7-19 demonstrates slope statistics of the training, test and overall datasets, respectively. According to the table, the minimum and maximum slope values observed in training database is -8.6 and 6.2, respectively. Corresponding values for the test database are -5.6 and 7.3, respectively. These values correspond to rather steep slopes. Generally, these values are observed on syllables having shorter durations. According to the table, the mean value for the slope is around 0 which denotes that all databases are dominated by *no-accent*s. However, when the histogram plots in **Figure 7-11** - **Figure 7-13** are considered, it is observed that the slopes of accented syllables are either at the positive or negative half of the slope line.

When slope statistics of the train and test dataset are compared, it is observed that the upper slope limit of the test data set is beyond the scope of the train dataset. This may result in a performance reduction.

When slope statistics of complete dataset is considered, it is observed that it covers the statistics of the train and test datasets, but resembles the statistics of the train dataset.

Table 7-19: Slope statistics of the training and test data.

	Minimum	**Maximum**	**Mean**	**SD**
Training	-8.6	6.2	0.03	1.378
Test	-5.6	7.3	0.015	1.288
AllSyllables	-8.6	7.3	0.027	1.359

For comparison purposes, mean slope of corresponding dataset is used as a baseline model for slope prediction experiments.

Table 7-20 illustrates the performance statistics of the decision trees and the baseline model for both experiment sets. As shown in the table, the slope prediction statistics are fairly good. Utilization of decision trees for slope prediction outperforms baseline model.

Table 7-20: Performance statistics of the the baseline, Experiment Set 1 & 2.

		CC	**MAE**	**RMSE**
Baseline	Train	~0	0.99	1.90
	Test	~0	0.92	1.66
	CrossVal	~0	0.97	1.85
Experiment Set 1	Training	0.86	0.49	0.69
	Test	0.83	0.53	0.73
	CrossVal	0.84	0.54	0.75
Experiment Set 2	Training	0.74	0.62	0.93
	Test	0.67	0.66	0.97
	CrossVal	0.66	0.69	1.03

For both experiment sets, best performances are obtained using the training set and worst performance are obtained with cross validation evaluation. However, the performances are more or less similar to each other numerically.

Considering all results obtained in Experiment Set 1 & 2, it can be said that decision trees outperform baseline model. The percentage improvement in MAE is around 46% for Experiemtn Set 1 and 31% for Experiemtns Set 2.

When the results of Experiment Set 1 and 2 are compared, it is observed that Set 1 exhibits better performances. This is due to the fact that Experiment Set 1 uses three accent states (positive, negative, and no-accent) while Experiment Set 2 uses two accent states (accented vs no-accent). Merging positive and negative accents reduces the prediction capacity. Experiment Set 1 performs approximately 20% and 26% better than Experiment Set 2 in MAE and RMSE, respectively.

Figure 7-14 demonstrates the resultant decision tree generated in Experiment Set 1 for **Training** and **Test** cases. The tree has five splitting levels. Attributes at each split are as follows:

1. *Accent* (most significant attribute)
2. *cluster-s1_major*, *POSw+1*, and *SylStruct*
3. *PosofWordM*ajor, *NumofSylToFolMinorBreak*, and *cluster-s1_minor*
4. *POSw+1*
5. *SylPosinWord1*

Rest of the attributes do not play any role in the resultant regression tree therefore they are irrelevant for accent prediction using regression trees.

```
Accent = cl0
|  cluster-s1_major = NONE : 0.53 (258/0.44) [139/0.45]
|  cluster-s1_major = cl0
|  |  PosofWordMajor = I : 0.62 (188/0.66) [94/0.45]
|  |  PosofWordMajor = M
|  |  |  POSw+1 = PRON : 0.06 (32/0.49) [15/0.87]
|  |  |  POSw+1 = NOUN : 0.05 (847/0.32) [418/0.31]
|  |  |  …
|  |  |  |  SylPosinWord1 = I : 0.05 (105/0.23) [59/0.4]
|  |  |  |  SylPosinWord1 = M : 0.09 (243/0.39) [110/0.39]
|  |  |  |  SylPosinWord1 = F : -0.25 (153/0.32) [69/0.46]
|  |  |  |  SylPosinWord1 = Single : 0 (6/0.14) [7/0.08]
|  |  |  POSw+1 = ADJ : 0.01 (382/0.31) [190/0.39]
|  |  |  POSw+1 = ADV : -0.06 (140/0.33) [78/0.23]
|  |  |  …
|  |  PosofWordMajor = F : -0.14 (731/0.14) [338/0.11]
|  cluster-s1_major = cl : -0.22 (525/0.54) [260/0.45]
|  cluster-s1_major = cl-1
|  |  NumofSylToFolMinorBreak < 1.5 : 0.2 (202/0.51) [71/0.33]
|  |  NumofSylToFolMinorBreak >= 1.5 : -0.43 (947/0.76) [481/0.83]
Accent = cl
|  POSw+1 = PRON : 1.88 (27/0.77) [11/1.07]
|  POSw+1 = NOUN : 1.99 (569/0.79) [270/0.77]
|  POSw+1 = TELL : 1.9 (15/1.14) [12/1.72]
|  POSw+1 = QUES : 2.67 (75/1.02) [44/1.12]
|  POSw+1 = NONE : 1.59 (89/0.37) [39/0.31]
|  …
Accent = cl-1
|  SylStruct = NC : -2.02 (118/0.9) [69/1.13]
|  SylStruct = ON
|  |  cluster-s1_minor = NONE : -1.52 (104/0.38) [63/0.39]
|  |  cluster-s1_minor = cl0 : -1.61 (247/0.49) [115/0.69]
|  |  cluster-s1_minor = cl : -1.97 (354/0.62) [186/0.7]
|  |  cluster-s1_minor = cl-1 : -1.41 (38/0.23) [17/0.29]
|  SylStruct = ONC
|  |  cluster-s1_minor = NONE : -1.2 (42/0.18) [20/0.13]
|  |  cluster-s1_minor = cl0 : -1.49 (118/0.37) [71/0.46]
|  |  cluster-s1_minor = cl : -1.74 (192/0.43) [101/0.47]
|  |  cluster-s1_minor = cl-1 : -1.12 (22/0.12) [11/0.12]
|  SylStruct = N : -2.38 (161/1.86) [87/1.37]
|  SylStruct = ONCC : -1.48 (5/0.21) [4/0.1]
|  SylStruct = OONC : -1.16 (4/0.41) [3/0.1]
|  SylStruct = OON : -1.5 (2/0.09) [1/2.25]
|  SylStruct = NCC : -0.8 (1/0) [0/0]
|  …
```

Figure 7-14: Regression tree obtained for Experiment Set 1 using training set.

Resultant regression tree obtained in Experiment Set 2 for training dataset is given in **Figure 7-15**. The tree has five splitting levels. Attributes ate each split are as follows:

1. *cluster-s1_major* (most significant attribute)
2. *Accent*, and *SylPosinWord1*
3. *PosofWordMajor*, *SylPosinWord1*, *Accent*, and *Break*
4. *POSw+1*, *SylStruct*, and *SylNoinWord*
5. *SylPosinWord1*, and *Accent*

Rest of the attributes do not play any role in the resultant regression tree therefore they are irrelevant to accent prediction using decision trees.

```
cluster-s1_major = NONE : 0.98 (350/1.19) [197/1.24]
cluster-s1_major = cl0
|  Accent = cl0
|  |  PosofWordMajor = I : 0.62 (188/0.66) [94/0.45]
|  |  PosofWordMajor = M
|  |  |  POSw+1 = PRON : 0.06 (32/0.49) [15/0.87]
|  |  |  POSw+1 = NOUN : 0.05 (847/0.32) [418/0.31]
|  |  |  …
|  |  |  |  SylPosinWord1 = I : 0.05 (105/0.23) [59/0.4]
|  |  |  |  SylPosinWord1 = M : 0.09 (243/0.39) [110/0.39]
|  |  |  |  SylPosinWord1 = F : -0.25 (153/0.32) [69/0.46]
|  |  |  |  SylPosinWord1 = Single : 0 (6/0.14) [7/0.08]
|  |  |  POSw+1 = ADJ : 0.01 (382/0.31) [190/0.39]
|  |  |  POSw+1 = ADV : -0.06 (140/0.33) [78/0.23]
|  |  |  …
|  |  PosofWordMajor = F : -0.14 (731/0.14) [338/0.11]
|  Accent = cl_cl-1
|  |  SylPosinWord1 = I : -1.03 (286/2.68) [132/1.94]
|  |  SylPosinWord1 = M : 0.45 (303/3.27) [186/3.61]
|  |  SylPosinWord1 = F : 1.71 (960/2.01) [469/1.97]
|  |  SylPosinWord1 = Single : 0.65 (87/2.9) [40/2.59]
cluster-s1_major = cl_cl-1
|  SylPosinWord1 = I
|  |  Accent = cl0 : -0.07 (423/0.34) [195/0.22]
|  |  Accent = cl_cl-1
|  |  |  SylStruct = NC : -1.88 (84/1.86) [52/1.96]
|  |  |  SylStruct = ON : -1.47 (344/1.51) [179/1.4]
|  |  |  …
|  SylPosinWord1 = M : -0.49 (913/1.26) [445/1.48]
|  SylPosinWord1 = F
|  |  Break = SI : -0.06 (0/0) [0/0]
|  |  Break = M
|  |  |  SylNoinWord < 2.5
|  |  |  |  Accent = cl0 : -0.36 (116/0.5) [68/0.55]
|  |  |  |  Accent = cl_cl-1 : 0.71 (119/3.71) [46/3.6]
|  |  |  SylNoinWord >= 2.5 : -0.83 (196/1.47) [108/1.78]
|  |  Break = SF : 0 (40/0.7) [24/1.29]
|  |  Break = F : 0.74 (154/1.81) [61/2.56]
|  |  Break = I : -0.7 (0/0) [1/0.47]
|  |  Break = I/F : -0.06 (0/0) [0/0]
|  SylPosinWord1 = Single : -0.28 (211/1.99) [110/2.4]
```

Figure 7-15: Regression tree obtained for Experiment Set 2 using training set.

When the two regression trees are compared, it is observed that the first splits are different in each of them. In the former, the *Accent* attribute turns out to be the most prominent attribute while in the latter *cluster-s1_major* attribute occurs at the first split thus it is the most prominent attribute.

7.2.2 Improving Accent Prediction

The best performance in accent prediction experiments is obtained by means of binary prediction, that is, the case where positive and negative accents are merged into a single *accented* class. The evaluations on the test set show 74.56% and 75.89% correct prediction for Experiment Set 1 and 2, respectively. When the two set of experiments are examined in detail, it becomes obvious that the Experiment Set 1 predicts the *no-accent*s better while Experiment Set 2 predicts accented syllables better.

Results are much more promising then our previous attempts on classifying pitch accents by means of clusters associated syllable pitch contours using k-means partitioning algorithm, however they still need improvement. It is observed that the selected threshold value set in accent assignment procedure results in a highly selective algorithm. Therefore, the algorithm rejected some of the prominent accents which have slope values below the selected threshold. A typical case is illustrated in **Figure 7-16**. In the given example, the sentence ‘dövizde yapılan analizlerde ciddi bir sıçrama beklenmiyor yıl sonuna kadar’ (serious changes are not expected in the analysis made over the currency till the end of the year) is examined. It can be observed that the accent assignment algorithm missed the negative accent on the first syllable of the word ‘ciddi’ (serious) due to the threshold value for determining candidates for accented syllables. Capturing the negative slope on the demonstrated syllable is necessary since it constitutes the local minimum of the pitch contour. However, in order to capture such multimodalities, the slope threshold should be set to a very low value.

Another critical point in accent assignment algorithm is encountered in the syllables whose syllable pitch contours show multimodalities. In slope calculation, only the initial and final F0 values are considered to determine the slope of the intended syllable. So, syllables hving multimodal pitch contours are discarded. In fact, although the syllables showing multimodalities attain a rather considerable peak (valley) F0, the declination (inclination) afterwards causes slope computation to assign a comparatively low slope value to the syllable. A typical case is shown in **Figure 7-17** where the sentence ‘yavuz

işe gitti ancak cihan çarşıya çıkmadı' (yavuz went to work but cihan did not go to shopping) is examined. The negative pitch accent on the syllable 'cid' of the word 'ciddi' (serious) can not be captured by the accent assignment algorithm due to selected threshold value. These accents can be captured using rather low threshold values.

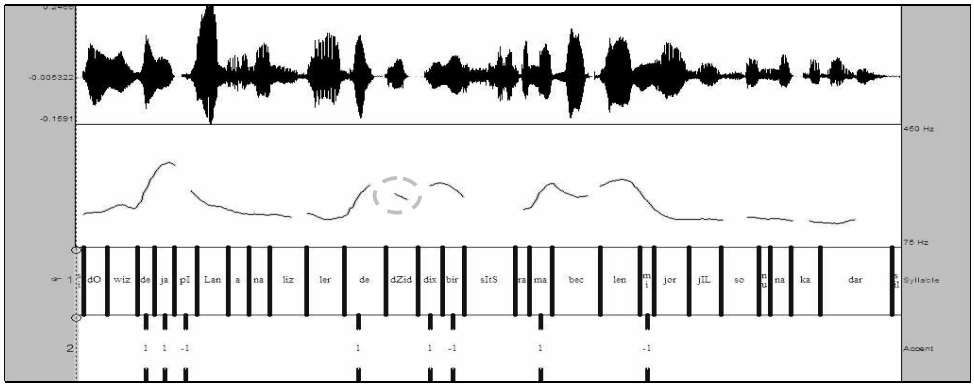

Figure 7-16: Sound waveform, pitch contour, syllable labels and pitch accents of the sentence 'dövizde yapılan analizlerde ciddi bir sıçrama beklenmiyor yıl sonuna kadar'

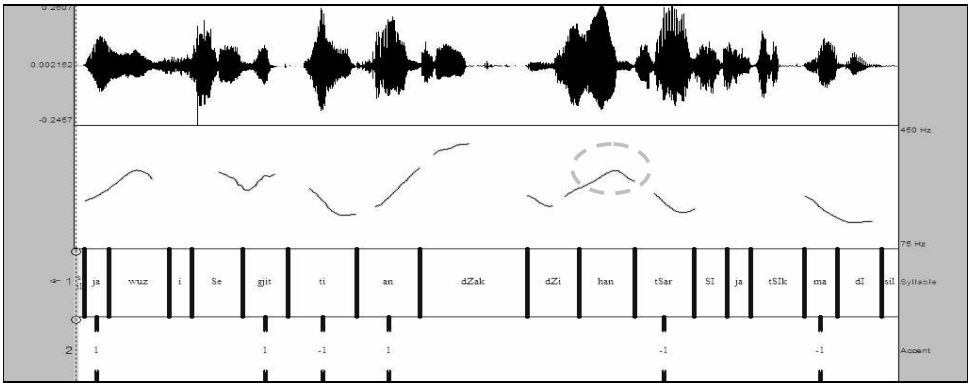

Figure 7-17: Sound waveform, pitch contour, syllable labels and pitch accents of the sentence 'yavuz işe gitti ancak cihan çarşıya çıkmadı'

There are also some limitations due to the accent assignment algorithm. The algorithm is constraint to assign one accent of each type to each word. However, in the database, there are words that have more than one accent of either accent type within its limits. **Figure 7-18** illustrates a misplacement in the accent assignment process observed in the sentence 'ancak savunanlar da hayli fazla deniliyor' (however it is said that the defenders are also too many). The first word of the sentence 'ancak' (however) has two prominent

positive slopes: first accent implies the lexical stress of the word and the second implies continuation at phrase boundary. As shown in the figure, the accent assignment algorithm misses second accent since the algorithm is limited to assign single accent of each type to each word. Such misplacements may result in reduced decision tree performance.

The accent assignment algorithm is modified to tackle the above mentioned problems.

1. A rather small threshold value is assigned to determine the candidate words.
2. One accent per word constraint is removed.

Figure 7-19 - **Figure 7-21** illustrate pitch accents of the example sentence given in **Figure 7-18** after modifications. **Figure 7-19** illustrates corresponding accents for the example sentence using a threshold of 90 for positive slopes and 100 for negative slopes. Using a threshold value of 100 for positive slopes allows the algorithm to capture the rising portion of the pitch contour on the second syllable of the word 'ancak' (however). However, current assignment is not sufficient enough to represent both the rise and fall on the syllable at the same time since it still misses the negative slope. The algorithm assigns only one accent per syllable and never assigns both positive and negative accents on the same syllable. Future studies will consider syllables with rise/fall patterns as turning points. For example, the resulting assignment inform that the pitch contour continue rising till the end of the word. An approximation for the resulting contour after reconstruction (dotted line) is also shown in **Figure 7-19**. The pitch on the word 'ancak' (however) is perceived differently from the original contour. Therefore, such syllables are not assigned to any accent.

Two different applications involving different threshold values are given in **Figure 7-20** and **Figure 7-21**. In **Figure 7-20**, the accent assignment method where the threshold value for positive accents is set to 90 as previously and the negative accent is set to 150 is depicted. **Figure 7-21** presents the case in which the positive threshold value is set to 100 and the negative threshold value is set to 150. According to the figures given, increasing (decreasing) threshold values results in selecting more (less) accents per sentence. From our discussions about multimodal syllables and the accent assignment methods given in figures, the best choice for accent thresholds is 100 for positive and 150 for negative accents.

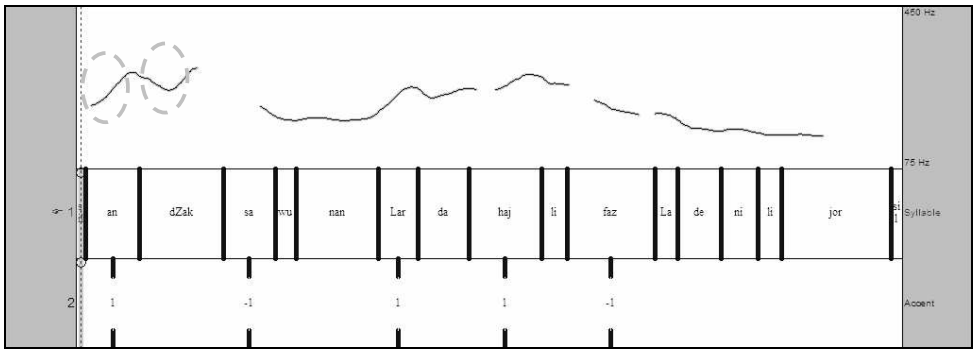

Figure 7-18: Pitch contour, syllable labels and pitch accents of the sentence ‘ancak savunanlar da hayli fazla deniliyor’.

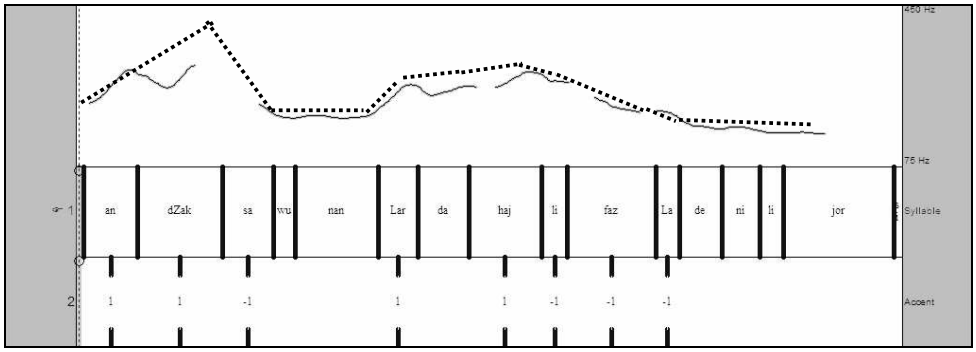

Figure 7-19: Original (continuous) and reconstructed (dotted) pitch contours, syllable labels and pitch accents of the sentence ‘ancak savunanlar da hayli fazla deniliyor’.

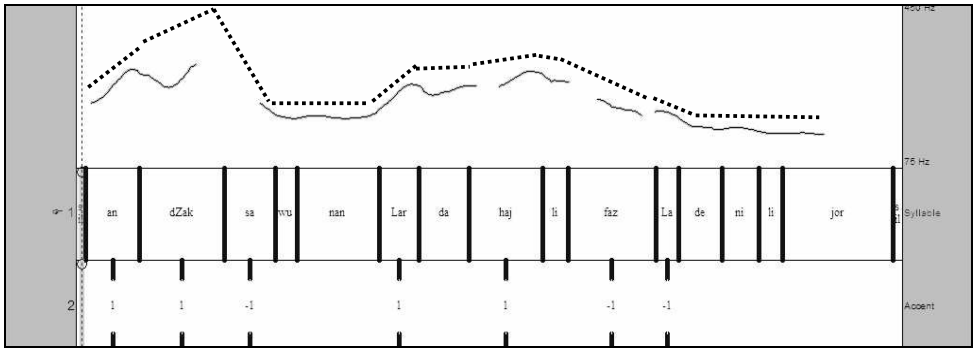

Figure 7-20: Original (continuous) and reconstructed (dotted) pitch contours, syllable labels and pitch accents of the sentence ‘ancak savunanlar da hayli fazla deniliyor’

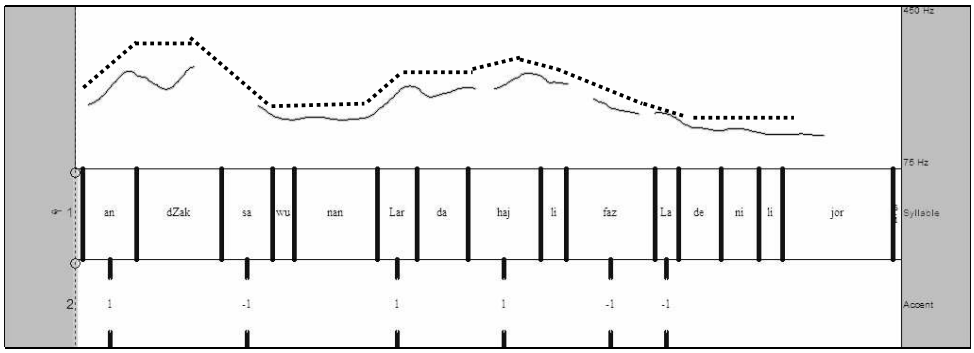

Figure 7-21: Original (continuous) and reconstructed (dotted) pitch contours, syllable labels and pitch accents of the sentence 'ancak savunanlar da hayli fazla deniliyor'

7.2.2.1 Accent Prediction

Accent location prediction is performed employing two experimental setups as discussed in Section 7.2.1.1: Experiment Set 1 and 2. Performances of the decision trees are evaluated using **Training**, **Test** and **CrossVal** methods.

Statistical observations given in tables **Table 7-21** through **Table 7-27** belong to the accent prediction experiments. For both of the experiment sets, **Test** and **CrossVal** performances are lower than **Training** performances however they are more reliable.

Table 7-21: Correct and incorrect classification rates for Experiment Set 1 & 2.

			Number of Syllables	**Percentage**
Experiment Set 1	Training	Correctly Classifed Syllables	9920	79.4681%
		Incorrecly Classified Syllables	2563	20.5319%
	Test	Correctly Classifed Syllables	2126	62.8251%
		Incorrecly Classified Syllables	1258	37.1749%
	CrossVal	Correctly Classifed Syllables	9859	62.1352%
		Incorrecly Classified Syllables	6008	37.8648%
Experiment Set 2	Training	Correctly Classifed Syllables	10074	80.7018%
		Incorrecly Classified Syllables	2409	19.2982%
	Test	Correctly Classifed Syllables	2234	66.0165%
		Incorrecly Classified Syllables	1150	33.9835%
	CrossVal	Correctly Classifed Syllables	10669	67.2402%
		Incorrecly Classified Syllables	5198	32.7598%

According to **Table 7-21**, best correct classification rates are observed in the cases where evaluations are performed on training set in both experiment sets. These results illustrate the upper limits of the decision trees. It is observed that correct classification rates obtained in Experiment Set 2 is slightly better than those of Experiment Set 2.

Kappa coefficients for both experiment sets are given in **Table 7-22**. Although correct classification rates of Experiment Set1 are slightly worse than Experiment Set 2, corresponding Kappa coefficients of Experiment Set 1 are better than those of Experiment Set 2. Evaluations using training sets approach 0.7 which is regarded as good statistic correlation. Best Kappa coefficient is observed in Experiemtn Set 1.

Table 7-22: Kappa statistics of Experiment Set 1 & 2.

		Kappa Statistics
Experiment Set 1	Training	0.67
	Test	0.38
	CrossVal	0.38
Experiment Set 2	Training	0.61
	Test	0.32
	CrossVal	0.34

The confusion matrices of Experiment Set 1 & 2 are given in **Table 7-23** and **Table 7-25**, respectively. Diagonal entries of the tables correspond to correct predictions while off-diagonals correspond to false predictions. Confusion matrix of Experiment Set 1 (**Table 7-23**) shows that decision trees cannot discriminate accented syllables from the not-accented syllables. However, they perform a better discrimination in between positive and negative accented syllables.

In order to compare performances of correct class predictions of both experiment sets, confusion matrix of Experiment Set 1 (**Table 7-23**) is converted to two-class confusion matrix given in **Table 7-24**. The converson is performed by merging the statistics of positive and negative classes. Comparison of the cofusion matrices of both experiment sets (**Table 7-24** and **Table 7-25**), it is observed that prediction of *accented* class is better in Experiment Set 2 but Experiment Set 1 predicts *no-accent*s better. Approximately 11% improvement is achieved in correct prediction of accented syllables.

Table 7-23: Confusion matrices observed in Experiment Set 1 (*positive*, *negative* and *no-accent*).

		Classified as:	no-accent	positive	negative
Experiment Set 1	Training	no-accent	5327	444	336
		positive	765	2639	159
		negative	677	182	1954
	Test	no-accent	1273	287	203
		positive	320	527	66
		negative	290	92	326
	CrossVal	no-accent	5679	1220	971
		positive	1668	2418	390
		negative	1364	395	1762

Table 7-24: Confusion matrices observed in Experiment Set 1 (*accented* vs *no-accent*).

		Classified as:	no-accent	accented
Experiment Set 1	Training	no-accent	5327	780
		accented	1442	4934
	Test	no-accent	1273	490
		accented	610	1011
	CrossVal	no-accent	5679	2191
		accented	3032	4965

Table 7-25: Confusion matrices observed in Experiment Set 2 (*accented* vs *no-accent*).

		Classified as:	no-accent	accented
Experiment Set 2	Training	no-accent	4728	1379
		accented	1030	5346
	Test	no-accent	1106	657
		accented	493	1128
	CrossVal	no-accent	5100	2770
		accented	2428	5569

TP rates, FP rates, Precisions, Recalls and F-Measures for Experiment set 1 & 2 are given in **Table 7-26** and **Table 7-27,** respectively. It is observed that best TP rates are observed for *no-accent*s in Experiment Set 1 while accented class is predicted better in Experiment Set 2.

Table 7-26: TP rate, FP rate, Precision, Recall and F-measures of Experiment Set 1.

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Experiment Set 1	Training	0.872	0.226	0.787	0.872	0.827	no-accent
		0.741	0.07	0.808	0.741	0.773	positive
		0.695	0.051	0.798	0.695	0.743	negative
	Test	0.722	0.376	0.676	0.722	0.698	no-accent
		0.577	0.153	0.582	0.577	0.579	positive
		0.46	0.101	0.548	0.46	0.5	negative
	CrossVal	0.722	0.379	0.652	0.722	0.685	no-accent
		0.54	0.142	0.6	0.54	0.568	positive
		0.5	0.11	0.564	0.5	0.53	negative

Table 7-27: TP rate, FP rate, Precision, Recall and F-measures of Experiment Set 2.

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Experiment Set 2	Training	0.774	0.162	0.821	0.774	0.797	no-accent
		0.838	0.226	0.795	0.838	0.816	accented
	Test	0.627	0.304	0.692	0.627	0.658	no-accent
		0.696	0.373	0.632	0.696	0.662	accented
	CrossVal	0.648	0.304	0.677	0.648	0.662	no-accent
		0.696	0.352	0.668	0.696	0.682	accented

As a result, it can be said that accented syllables are predicted better with the second experiment set where we collected accented syllables into a single category. But the performance in predicting not-accented syllables degrades when the second experiment set is involved.

Improved accent assignment algorithm (ref Section 7.2.2) outputs are used in the last six experiments. Them main motivation is to improve the performance of the former set of experiments. **Table 7-28 - Table 7-30** demonstrates the performances of the evaluations on training and test databases for comparison purposes. The performances using the 10-fold cross validation statistics are not considered since the remaining two evaluations clarify current state.

The classification rates of the evaluations on train and test data are given in **Table 7-28**. As revealed by the table, former experiments' results are better than experiments involving new assignments for pitch accents. The performance of the decision tree reduced with the modified database. Therefore, it can be concluded that the resulting

improvements in pitch accent assignment does not result in a performance improvement in decision tree performance.

Table 7-28: Correct classification rates of Accent schemes before (Original) and after modification (Modified).

			Classification Rates	
			Correct	**Incorrect**
Original	Experiment Set 1	Train	84.23%	15.77%
		Test	74.56%	25.44%
	Experiment Set 2	Train	84.58%	15.42%
		Test	75.89%	24.11%
Modified	Experiment Set 1	Train	79.47%	20.53%
		Test	62.83%	37.17%
	Experiment Set 2	Train	80.70%	19.30%
		Test	66.02%	33.98%

Table 7-29 demonstrates the overall performance obtained from the original database and modified database. The columns correspond to TP rate, FP rate, precision, recall, and F-measures, respectively. The prediction statistics also show that the former set of experiments involving original database is better than the current set of experiments for not-accented syllables. The accented syllables are better predicted with the current set f experiments involving modified database. In fact, for the binary case, the accented syllables are even better predicted than the not-accented syllables with the modified database.

Table 7-30 demonstrates the confusion matrices of the decision trees using original and modified databases, all together. Although, the number of correctly classified accented syllables increased in the experiment set involving modified database, the overall performance is not improved further.

Considering all the statistics of both experiment sets involving original and modified accent values given in the three tables, it can be said that the latter experiment set predict accented syllables better. However, the prediction performance for the not-accented syllables falls with the modified accent values. So, there is a trade-off in the selection of the methods. If former decision tree is used, the prediction performance on the accented syllables is rather low. But, if latter is used, then, the prediction performance of the not-

accented syllables falls. As a matter of fact, performance improvement in predicting accented syllables is accomplished in the latter set of experiments.

Table 7-29: TP Rate, FP Rate, Precision, and F-measure before and after modification.

			TP Rate	FP Rate	Precision	F-Measure	Class
Original	Experiment Set 1	Train	0.926	0.28	0.855	0.889	no-accent
			0.726	0.031	0.84	0.779	positive
			0.658	0.038	0.783	0.715	negative
		Test	0.849	0.417	0.801	0.824	no-accent
			0.594	0.081	0.604	0.599	positive
			0.488	0.057	0.63	0.55	negative
	Experiment Set 2	Train	0.916	0.28	0.854	0.884	no-accent
			0.72	0.084	0.828	0.77	accented
		Test	0.837	0.396	0.807	0.822	no-accent
			0.604	0.163	0.653	0.627	accented
Modified	Experiment Set 1	Train	0.872	0.226	0.787	0.827	no-accent
			0.741	0.07	0.808	0.773	positive
			0.695	0.051	0.798	0.743	negative
		Test	0.722	0.376	0.676	0.698	no-accent
			0.577	0.153	0.582	0.579	positive
			0.46	0.101	0.548	0.5	negative
	Experiment Set 2	Train	0.774	0.162	0.821	0.797	no-accent
			0.838	0.226	0.795	0.816	accented
		Test	0.627	0.304	0.692	0.658	no-accent
			0.696	0.373	0.632	0.662	accented

One major drawback of the modified accent assignment is that some of the lexically stressed syllables accented in the original database are no longer accented because of the newly set threshold values. An example is shown in **Figure 7-22**. In **Figure 7-22**, the accents obtained after the modification is depicted in the third window. Accents associated before modification are given in the fourth window. When both schemes are examined, it is seen that the original assignment method obeys rules of lexical stress assignment better than the modified method. In order to preserve lexical stresses as well as newly added accents, both outputs can be used.

Table 7-30: Confusion matrices before and after modification.

			Classified as:	no-accent	positive	negative
Original	Experiment Set 1	Train	**no-accent**	7411	260	335
			positive	577	1685	59
			negative	677	60	1419
		Test	**no-accent**	1906	196	144
			positive	219	344	16
			negative	256	30	273
	Experiment Set 2	Train	**no-accent**	7335	671	
			accented	1254	3223	
		Test	**no-accent**	1881	365	
			accented	451	687	
Modified	Experiment Set 1	Train	**no-accent**	5327	444	336
			positive	765	2639	159
			negative	677	182	1954
		Test	**no-accent**	1273	287	203
			positive	320	527	66
			negative	290	92	326
	Experiment Set 2	Train	**no-accent**	1106	657	
			accented	493	1128	
		Test	**no-accent**	5100	2770	
			accented	2428	5569	

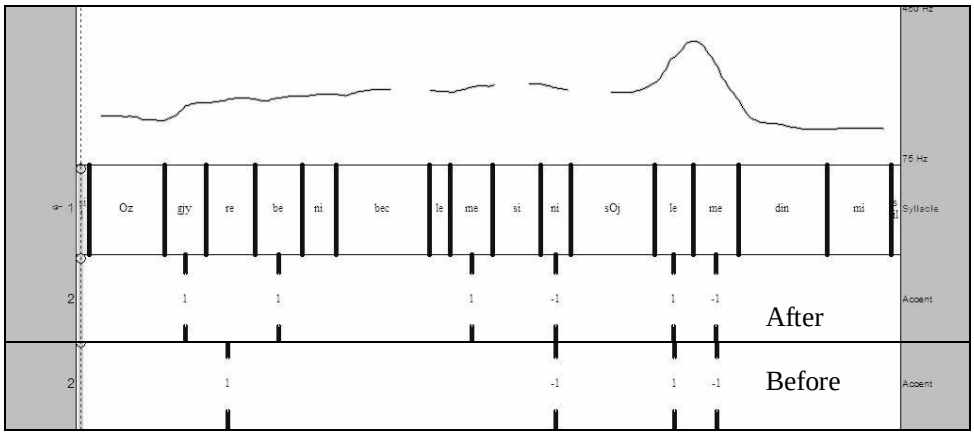

Figure 7-22: Pitch contour and syllable labels and accent states before and after modification of the sentence ‘özgüre beni beklemesini söylemedin mi’

7.2.2.2 Slope Prediction for Accented Syllables

As with the case discussed in Section 7.2.1.2, we perform slope amplitude prediction involving two sets of experiments: Experiment Set 1 and 2.

For each set of experiments, three evaluation methods are used: evaluation using training data (**Training**), evaluation using test data (**Test**), and evaluation using 10-fold cross validation method (**CrossVal**).

Table 7-31 demonstrates the quantitative performance measures for slope prediction after modifying accent assignemtn algorithm. Baseline performances using mean slope values are also given in the table.

According to the table, both experiment sets outperforms baseline model. Performance improvement in MAE with respect to the baseline in Experiment Set 1 is around 54.2% while it is around Performance improvement with respect to baseline in 27.3% in Experiment Set 2.

The table also shows that overall performance of Experiment Set 1 is better than that of Experiment Set 2. Best performances in both experimental setups are observed in evaluations using training dataset. Test dataset and 10-fold cross validation evaluation performances are almost similar.

Table 7-31: Performance statistics of the the baseline, Experiment Set 1 & 2.

		CC	**MAE**	**RMSE**
Baseline	Training	~0	0.99	1.9
	Test	~0	0.92	1.66
	CrossVal	~0	0.97	1.85
Experiment Set 1	Training	0.91	0.42	0.59
	Test	0.88	0.45	0.62
	CrossVal	0.88	0.45	0.64
Experiment Set 2	Training	0.67	0.67	1.02
	Test	0.60	0.69	1.04
	CrossVal	0.59	0.73	1.11

Comparison of the performances of current set of experiments with those given in **Table 7-20**, it is bserved that slope amplitude prediction is improved after modification of accent assignment algorithm for Experiment Set 1 but the performances decrease in

Experiment Set 2 after modification. Overall performance improvement in MAE is around 15% for Experiment Set 1. Experiment Set 2 performs slightly worse than slope former corresponding experiment set given in **Table 7-20**. The overall performance reduction in MAE is around 6%.

Considering all the comparisons, we can say that modified accent assignment improves performance in slope prediction using three accent categories. However, with two accent case, performance measures of the former experiments were better.

Considering all results, using the original accent assignment algorithm seems to provide better performances. Therefore, corresponding results are used in pitch reconstruction phase.

7.2.3 Pitch Contour Reconstruction

As discussed previously, a three-step procedure is followed for modeling pitch contours. First step involves pitch accent placement, second step incorporates regression trees for the prediction of accent slopes. In the last step, slopes estimates are used to reconstruct syllable pitch contours which are used in developing resultant pitch contour estimate.

In the first two steps of the pitch contour modeling, statistical corpus based methods are employed. The classification task in the first step is handled by using the decision tree algorithm (J48) of WEKA package. The second part involves numeric prediction; regression tree algorithm (REPTree) of the WEKA package is used at this step.

For accent prediction, two different approaches are conducted. In the first one, the accent of a syllable is predicted as one of the three accent classes. In the second approach, the accented syllables are merged to construct a single class for accented syllables. Together with the not-accented syllables, the accented syllables constitute the dependent variable of the decision tree learner. Then, accent types of each syllable are predicted among the two classes.

Actual accent types of each syllable are considered as separate independent variables for predicting accented syllables' slopes in the second step.

For each prediction task, the database is split into two subsets: training and test datasets. Training dataset is used to develop an appropriate classification/regression tree while test dataset is used to evaluate the performance of induced tree. Training dataset

consists of 12483 samples of the syllable database and the test dataset consists of the remaining 3384 samples of the syllable database. These values correspond to 78.67% and 21.33% of the syllable database, respectively. The prediction performance is examined on training and test datasets as well as by 10-fold cross validation.

When overall statistics given in **Table 7-28 - Table 7-30** are considered, it is observed that the performance of the decision tree is better when the former experiment set using original accent assignment algorithm is used. Therefore, for pitch contour reconstruction purposes, we will mainly rely on the corresponding results.

For accent prediction, two different frameworks are provided: First set of experiments involve three accent classes (class1, class-1, and class0) for slope predictions while the second set use two accent classes (class1_class-1 versus class0). Among the two experiment sets, the latter is better in performance than the former. However, their performances are still comparable. When the latter case is taken into account, one more decision task should be performed, to discriminate the negative sloped accents from the positive sloped accents, which may result in performance reduction for the overall case. Therefore, three accent classification results are used since they classify each syllable of the corresponding test data into one of three accents.

Results of the slope prediction experiment, with *Accent* related attribute having three categorical values (positive, negative, and no-accent), given in Section 7.2.1, are used in pitch contour reconstruction.

It should be mentioned that, although evaluations using training dataset, test dataset, and 10-fold cross validation method are provided, results of the first two are considered during pitch contour reconstruction process. The results on training dataset are better since the same database is used to grow and test the decision tree. However, evaluation using training data does not reveal much information about the performance of the decision/regression trees on new data. Therefore, results of the test data are focused in general.

For slope amplitude prediction, it is assumed that accent of the syllables in the database are predicted previously, so they can be used as independent attributes in regression tree development. Same assumption holds for pitch reconstruction phase also. So, for reconstruction purposes, slope prediction results are directly employed, assuming that the accent status of the syllables in the train database can be estimated with

approximately 75% correct prediction rate. The slope values used in the learning phase are drawn from the normalized pitch contours. The slope histogram of the overall database, train and test database are given in **Figure 7-11** - **Figure 7-13**.

In the reconstruction phase, we use slope values associated to each syllable as well as initial F0 value of each sentence. The slope value is used in combination with syllable duration. For each syllable, the corresponding pitch contour is computed using the previous syllable's final F0, syllable duration and associated syllable slope. For making slope computations to present a more realistic framework, we set the estimated slope values to zero for not-accented syllables (modified estimates). Both slopes are demonstrated in the reconstructed contours fro comparison purposes.

For each syllable in the test set, we select ten time points that are equally spaced. Then, for each time point, corresponding F0 value is computed as follows:

$$F0(t_i) = F0(t_{i-1}) + (t_{i-1} - t_i) * m \tag{7-11}$$

where $\{t_i\}_{i=1}^{10}$ corresponds to one of the ten time points belonging to the syllable, m is the slope estimate of the syllable and $F0(t_{i-1})$ is the previous F0 value computed at time $\{t_{i-1}\}_{i=1}^{10}$. For the sentence initial F0 ($F0(t=0)$), the original sentence start F0 is used. Future studies incorporate regression trees to estimate sentence initial F0.

Since, slopes as well as the sentence initial F0 values are drawn from the normalized pitch contours, the resultant reconstructions correspond to the normalized pitch contours. The resultant pitch contours are shown in **Figure 7-23** - **Figure 7-33**. All the contours given in the figures are generated using three slope values: original slopes, estimated slopes, and estimated slopes with slope values corresponding to not-accented syllables set to zero. We also provide the results of the modified dataset (**predict1**) for making comparisons.

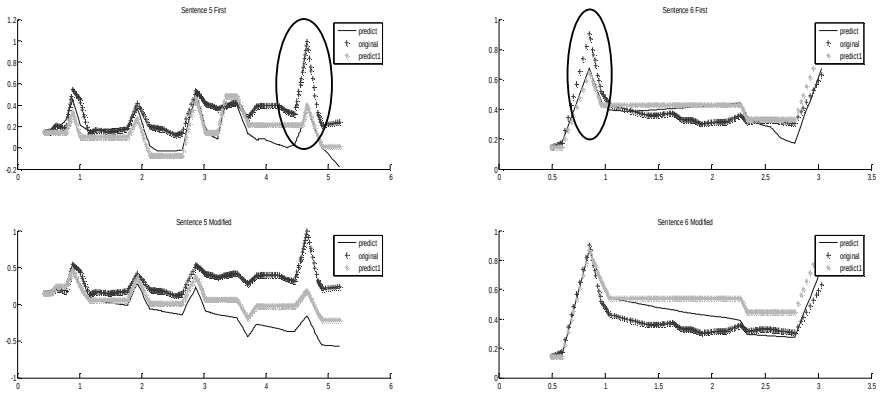

Figure 7-23: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets. Contours are generated using sentence initial F0 and three slope values: original slopes (bold line), estimated slopes (slim line), and modified estimates (gray line) for the sentence.

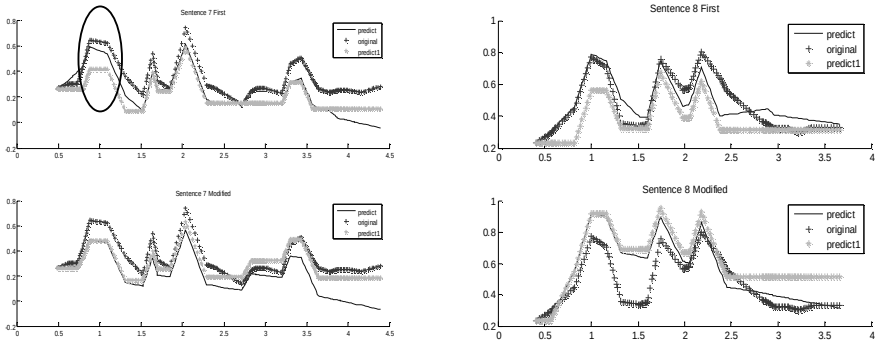

Figure 7-24: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

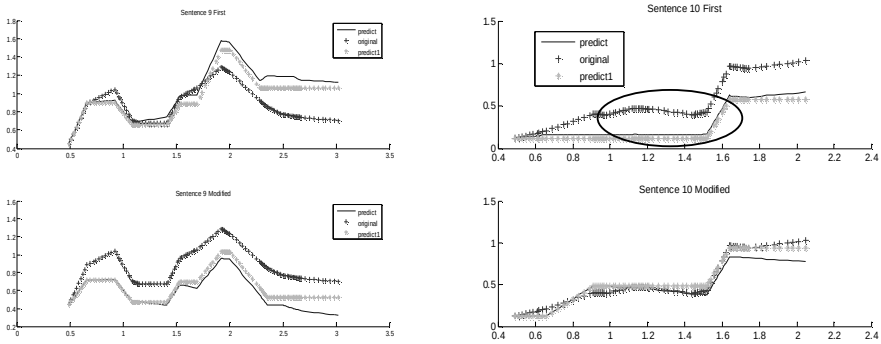

Figure 7-25: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

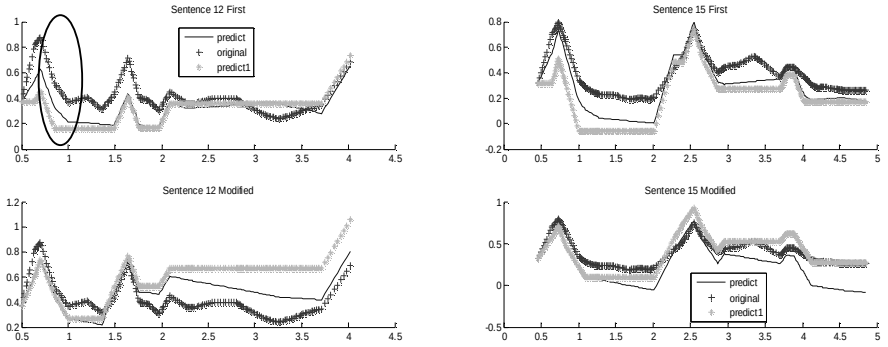

Figure 7-26: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

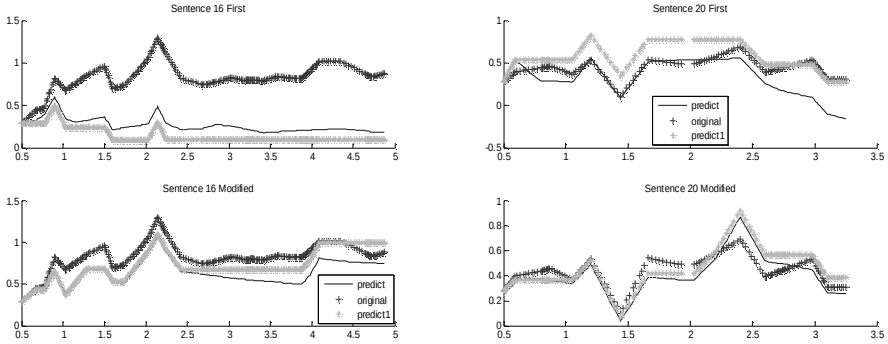

Figure 7-27: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

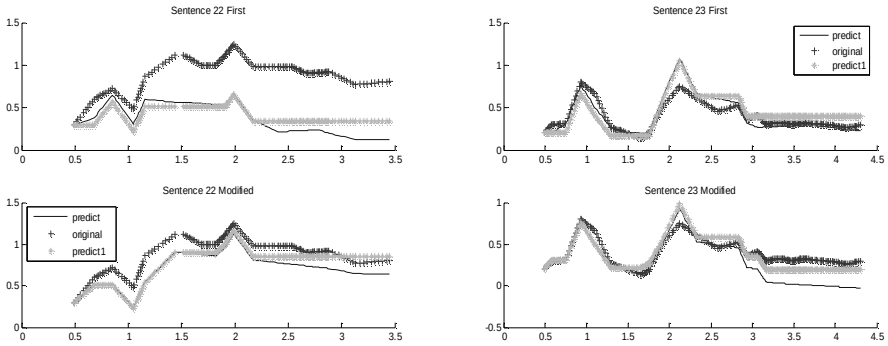

Figure 7-28: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

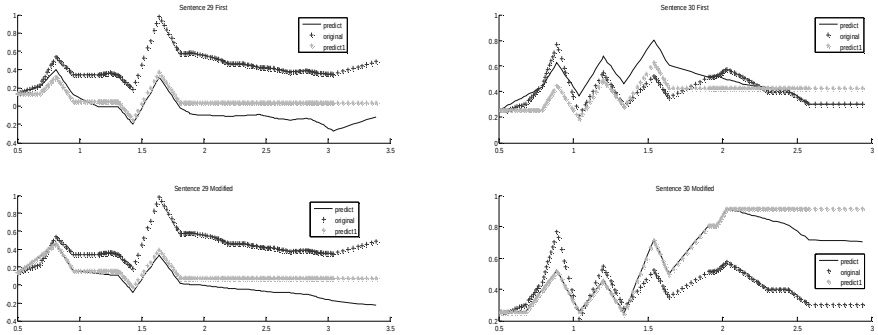

Figure 7-29: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

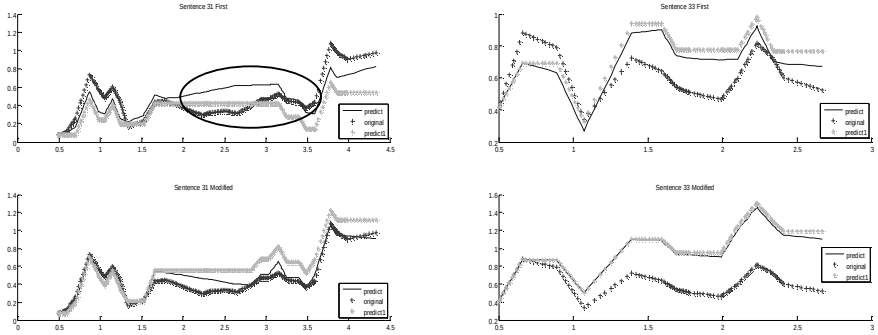

Figure 7-30: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

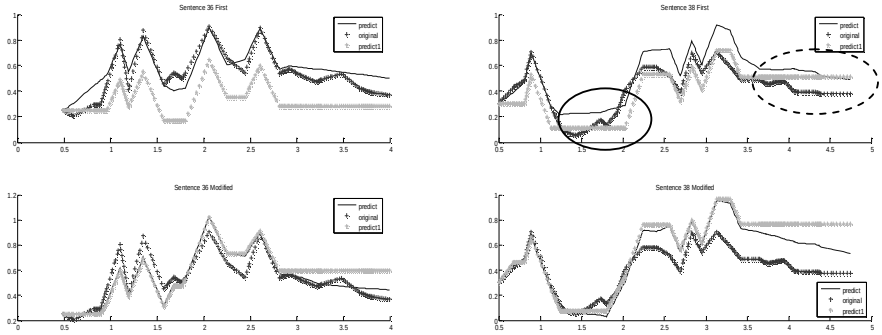

Figure 7-31: Reconstructed pitch contours using original (upper window) and modified (lower window) datasets.

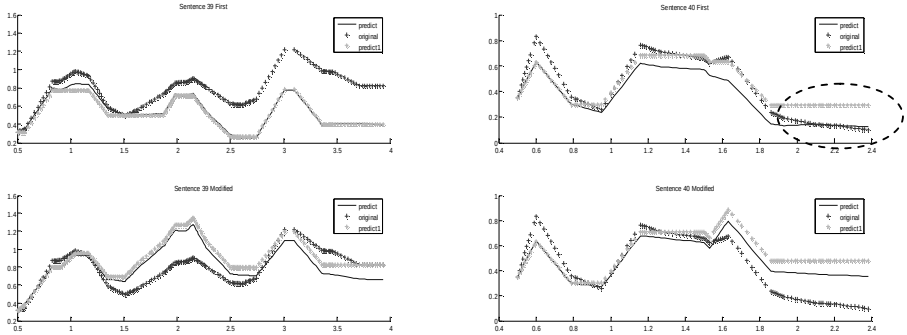

Figure 7-32: Reconstructed pitch contours using original slopes (+), estimated slopes (-), and modified estimates (*) for the sentence ‘’ ().

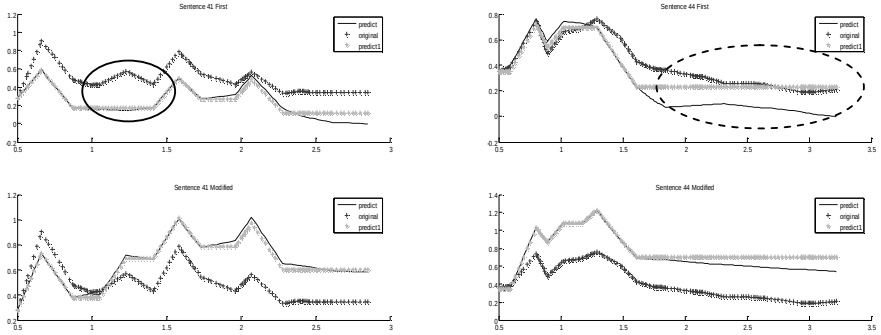

Figure 7-33: Reconstructed pitch contours using original slopes (+), estimated slopes (-), and modified estimates (*) for the sentence ‘’ ().

In some of the cases, it is observed that using modified dataset instead of original dataset provides better pitch contour estimates whereas in some cases, the reconstructed pitch contours corresponding to the regression tree derived from the original database provide better results.

For almost all cases, the peak locations are estimated satisfactorily. But, the amplitudes of the peaks can not be attained (marked with circles in **Figure 7-23**, **Figure 7-24**, and **Figure 7-26**). One major reason of this peak difference between actual and predicted contours arises because of sudden jumps on several syllables of the sentences. When the database is considered, it is observed that sudden jumps are not encountered frequently, thus, the corresponding regression tree produces smoother slopes for the corresponding jumps.

Another discrepancy observed on the estimated pitch contours is the inefficiency in tracking the syllables with smoother slopes (marked with circles in **Figure 7-30**, **Figure 7-31**, and **Figure 7-33**); however, current accent assignment algorithm does not take into account smoother slopes.

For most of the not-accented syllables, the pitch contour where the actual contour is very smooth, the pitch contour shows a declination pattern. However, since we directly set the not-accented syllables' slope to 0, the resultant pitch contours can not have declination on not-accented syllables (marked with dashed circles in **Figure 7-31**, **Figure 7-32**, and **Figure 7-33**).

As a result, when triple accent classification and corresponding slope prediction algorithms are considered, the resultant pitch contours of randomly selected test sentences are estimated quite satisfactorily. The prediction accuracy of the classification trees can not catch up with the state of the art prosody modeling systems but it is believed that the results can be further improved by increasing the size of the training data and by providing more appropriate annotation schemes.

CHAPTER 8

SUMMARY AND CONCLUSIONS

Prosody plays an important role in speech communication. It is related to the suprasegmental aspects of spoken language such as tonal quality, stress, intention, emphasis and speaker's characteristics. In natural speech, prosody of an utterance depends on semantics, context, syntax, intended audience, and emotional or physical state of the speaker. The three mathematically tractable components of prosody generally cited are: Pitch; segment duration and intensity. In this study, pitch contour and phoneme durations are modeled to serve as a basis for Turkish speech and linguistic research. Steps of our modeling efforts are summarized in the following sections.

8.1 Summary

First chapter introduces a brief definition of prosody and its components. Objectives and motivations are discussed in this chapter.

Focusing on the most influencing research, an overview of different approaches to intonation and duration modeling is given in the second chapter. Intonation modeling studies are discussed under two broad categories: Phonological and phonetic modeling approaches. Examples of phonological and phonetic intonation models are introduced. Consequently, rule-based and recent corpus-based duration models are reviewed.

Third chapter introduces text and speech databases developed during the progression of the thesis studies. The text database is designed to provide phonetic and prosodic balance. A set-covering algorithm is used to select sentences from a larger set to guarantee phonetic coverage. Resultant phonetically balanced set is modified syntactically to attain prosodic coverage. Designed text is recorded by a native female speaker in a soundproof booth. Phonetic transcription and alignment is provided for the resultant speech corpus.

Chapter 4 introduces prosodic attributes incorporated in modeling phoneme durations. Attributes used for phoneme duration modeling involves phoneme (segment) identity, preceding/following phoneme identities, lexical stress, and positional attributes for segment, syllable, and word. Individual effects of durational attributes on phoneme durations are examined in terms of statistical measures such as mean value, standard deviation and coefficient of variance. Some of the observations on phoneme durations and durational attributes are given below:

- It is observed that lexical stress does not play an important role in Turkish phonemic structure as in other languages such as English; however, phrase-final lengthening is observed in Turkish (ref **Table 4-5**).
- Studies on voiced and unvoiced consonants reveal that differences between voiced and voiceless consonants are very significant, i.e. in the order of 30-40 ms; voiceless consonants are longer in duration than their voiced counterparts (ref **Table 4-4**).
- It is also observed that when followed by a voiced consonant, phoneme duration increases except for vowel + voiced-plosive combination. Moreover, voiced fricative followers influence voiceless phoneme durations (~30 ms) more than voiced plosive (~12 ms) and affricate (~14 ms) followers (ref **Table 4-6**).
- When phoneme durations with respect to syllable position are examined, it is observed that voiced consonants are slightly longer when they occur in *coda* position. Besides, affricates, nasals, plosives and liquids occurring at onset are significantly longer in duration (around 20-30 ms) than the ones occurring at coda (ref. **Table 4-11**).
- Studies on phoneme duration with respect to syllable type showed that phonemes have shorter durations in open syllables than in closed syllables (ref. **Table 4-12** and **Table 4-13**).
- Phoneme durations are also affected by position-of-parent-syllable-in-parent-word: phonemes of word-initial and word-final syllables are longer and phonemes of single-syllable-words attain the maximum average duration (ref **Table 4-15**).
- According to parent-word-position-in-sentence, phonemes occurring at last words of the sentences are longer than the ones occurring at sentence-initial or sentence-medial words. The percentage lengthening is approximately 20% (ref **Table 4-17**).

- Average phoneme duration is shortened as the number of syllables in parent word increases (ref. **Table 4-19**). Phoneme duration is 41% longer in single-syllable-words than in words having ten syllables.
- Phoneme durations do not show a characteristic change with respect to the number-of-words-in-sentence (ref. **Table 4-20**).
- Average segment duration increases as the number-of-words-from-preceding-phrase-break increases since the probability of encountering a new phrase break increases (ref. **Table 4-23**).
- Words immediately followed by a phrase break attain maximum average phoneme durations (ref. **Table 4-23**).

The chapter ends with a discussion about attribute dependencies using mutual information criterion. Mutual information matrix showed that there is a stronger relation between phoneme identity and contextual attributes.

Attributes identified in Chapter 4 are used for phoneme duration modeling in Turkish. Corresponding results and discussions about phoneme duration modeling are given in Chapter 5. Forward selection method is used to determine the set of durational attributes that best models phoneme duration. Performances of the resulting models are quantitatively analyzed. Best correlation coefficient and root mean square error is obtained with the attributes *phoneme-identity*, *left/right*, *lexical-stress-of-parent-syllable*, *syllable-type-of-parent-syllable*, *Part-of-Speech-of-parent-word*, *phrase-information* and *number-of-words-to-following-phrase-break* attributes. Corresponding correlation coefficient and root mean squared error are 0.78 and 20.05 ms, respectively.

To improve duration prediction performance several modifications, duration quantization, modification of attribute values, outlier analysis, and shift and/or scale modification, are proposed. Duration quantization provides a dimension reduction in the duration values. Before modification, there are 242 distinct duration values. Quantization is performed using 54 quantization levels. Prediction performances are slightly worse than original duration values. However, results showed that phoneme durations can be modeled using fewer amount of data.

Another modification is performed in the selection of attributes. Using identities of preceding and following phonemes (Left/Right) requires a larger database for modeling purposes. It is observed that the speech database used in phoneme duration modeling does

not cover all representatives of all possible triphones. Hence, instead of directly incorporating identities of preceding and following phonemes, their manner of articulations are used in phoneme duration modeling. Keeping all other attributes, utilization of manner of articulations resulted in a slight improvement in prediction performances. Percentage improvements in correlation coefficient and root mean squared error are approximately 2% and 3%, respectively. Furthermore, manner of articulation values used are enlarged by adding the phonemes that effect phoneme duration significantly. Significance is evaluated by means of coefficient of variance. However, proposed modification does not improve the prediction performance obtained using original manner of articulations.

The numerical durational attributes and the maximum values that an attribute can attain are given as follows: *syllable-position-in-word* (10), *word-position-in-sentence* (19), *length-of-word-in-syllable-units* (10), *length-of-sentence-in-word-units* (19), *position-of-syllable-in-sentence* (45), *number-of-words-from(to)-preceding(following)-phrase-break* (8 for each) and *number-of-syllables-from(to)-preceding(following)-phrase-break* (27 for each). Their cross-product should be spanned by the database to be used in modeling. Hence, 10x19x10x19x45x8x8x27x72 = 75792672000 phonemes are required to represent all combinations of numerical attributes. However, this is not possible in general with the available speech databases. Therefore, two modifications are proposed for reducing the size of numerical attributes: normalization and symbolic representation. Proposed modifications are evaluated incorporating syllable-position-in-word attribute. The sample attribute attains values changing from 1 to 10. Symbolic representation involves coding of the attribute with respect to its relative position. For syllable-position-in-word attribute, possible attribute values for symbolic coding are {I, for word-initial-syllables, F for word-final-syllables, S for one-syllable-words and M for others}. Normalization involves length information as well. Resulting values are real valued lying in the closed range of 0 (for word initial and single syllable words) and 1 (for word final syllables). Using normalized attributes may increase the attribute span; however they eliminate utilization of length based attributes. Performances of the proposed modifications are evaluated keeping all other attributes fixed and using modified attributes. According to the results obtained, attribute value modification does not result in significant performance improvement however slightly better results are obtained.

Another modification for performance improvement in phoneme duration modeling is performed by excluding extreme duration data. The data range used in modeling is determined considering phoneme duration statistics. The duration data is widely spread in 2ms – 295ms range. Mean, standard deviation and median values of duration data are 63 ms, 57 ms and 31 ms, respectively. Besides, 91.3% of the data lies in the 22 ms – 117 ms range in the overall database. Considering duration statistics, durations outside 10 - 150 ms range (approximately 1.7% of overall dataset) are set as extreme duration values. Rest of the duration values are modeled using the durational attributes described in Chapter 5. Resulting prediction performances are improved significantly yielding a correlation coefficient of 0.75 and an RMSE of 18.6 ms. Best correlation coefficient reported in literature is around 0.9 [Venditti and van Santen, 1998]. The performance difference between two data sets points that although manual correction is performed on phoneme boundaries, there are still segmentation errors in the database.

Shift and/or scale modification is another modification applied to improve the prediction performance in phoneme duration modeling. Predicted durations and corresponding RMSEs are redefined using shift and/or scale parameters. Shift and/or scale parameters are found so that corresponding RMSE is minimized. Parameters are trained on training set RMSE and applied on test set predictions. Best correlation coefficient and RMSE values obtained are 0.79 and 19.5 ms, respectively. Resulting modification improves correlation coefficient and RMSE 2.6% and 4.4%, respectively.

Chapters 6 and 7 present pitch contour modeling studies. For pitch contour modeling, syllables are selected as the basic units. Attributes that are used for pitch contour modeling are defined in Chapter 6. Almost all attributes involved are defined in literature. However, *NegFlag*, *Sentence-type* and *POSRoot* attributes have not been used in previous studies. *NegFlag* is a binary attribute that represents whether current syllable comprise a negation suffix or not. *Sentence-type* attribute is coded with 4 categorical attributes that corresponds to the parent sentence structure described in Chapter 3. *POSroot* attribute is used to capture parent word's original morphemic constitute, i.e. noun, adjective, verb, and etc. Turkish is a highly agglutinative language. There are derivational suffixes as well as inflectional suffixes. Words can appear in their derived forms, for example, an adverb can be obtained using the derivational suffix –yIp from a verb. POSroot attribute holds parent word's original constitution. Chapter ends with a discussion about the relation of

attributes and pitch contour parameters using information gain, gain ratio and symmetrical uncertainty measures.

Pitch contour modeling studies are presented in Chapter 7. Two different methods are proposed for pitch contour modeling. One method can be associated to phonetic modeling methods introduced in Chapter 2. The other method can be considered as a phonological model since it captures the prominence of syllables. Both proposals aim at describing syllable pitch contours with a limited set of symbols.

One method generates a codebook of syllable pitch contours and uses corresponding codewords in pitch contour prediction. Codebook generation is performed by means of vector quantization of syllable pitch contours. Codebooks of various sizes are generated but pitch contour modeling is performed using 24-codebooks. Prediction performances are given in Section 7.1.2. All codebook entries are not sufficiently represented in the database (**Table 7-4**). Besides, there are similar patterns in the codebook that may be counted as wrong classifications. Hence, resulting correct classification percentage is rather low, around %27. Best TP rate is obtained for codeword 2, which is the most frequent codeword observed in the database, as 0.79. Worst TP rate is obtained for codeword 12, which is one of the rarest codewords in the database.

Codewords are used to assign accent status to syllables depending on two criteria: multimodality and dynamic range. Former experimental results show that centroids with slight level differences cause lower performances. Therefore, level differences are removed from syllable pitch contours. Resultant contours are vector quantized in two stages. In the first stage, a codebook of 100 centroids is generated from the syllable pitch contours. In the second stage, codewords that are generated in the first stage are used to generate a codebook of 25 elements. Among the resultant codewords, the ones having multimodalities are associated to pitch accents. Binary prediction is performed to decide whether a syllable is accented or not using previously described prosodic attributes. Corresponding decision tree performance is given in Section 7.1.3. The percentage of syllables that are correctly classified is around 80.6%. Although overall prediction performance of binary classification is very good when compared to 24 codeword classification, TP rate of the accented syllables is rather low, around 43%. Examination of resultant predictions yield that what determines pitch accent is not multimodality but the dynamic range.

Accent assignment is revisited using dynamic range information. Observations on perceptual tests reveal that prominence is perceived on abrupt changes of pitch contour. Therefore, second approach uses dynamic range rather than contour shapes. Codewords with dynamic ranges greater than a predefined threshold are associated to pitch accents. Two experiments are performed using different thresholds. In the first experiment, the threshold value is set to 108 Hz. Corresponding binary classification predicts approximately 80.9% of syllables correctly. Overall performance of the decision tree is improved slightly however accented syllable prediction performance is worse since 108Hz threshold causes most of the accented syllables to be eliminated. In the second experiment, the threshold is lowered to 40 Hz. Corresponding decision tree predicts 80.9% of the syllables correctly. The TP rate of accented syllables is improved significantly (81%) at the cost of lowered TP rate for unaccented syllables (~60%). According to the resultant predictions, it is observed that TP rate of the less frequent dependent variable is lower than that of the frequent ones.

Other approach relies on the definition of pitch accent for Turkish. Pitch accents correspond to perceptual prominence and are mainly aligned with lexically stressed syllables of the words. Accented syllables are associated with syllables having sudden and large pitch movements. A pitch accent assignment algorithm is developed to describe the accent status of syllables with respect to slope values and a predefined slope threshold. Rising patterns are associated to positive accents while falling patterns are associated to negative accents. Smoother contours are associated to no accent. Rising and falling patterns are combined to produce accented syllables. Accent prediction is performed within two frameworks. Three accent states (positive, negative and no accent) and two accent schemes (accented versus not-accented) are predicted using decision tree learning. Corresponding slope predictions are performed using regression trees. Resultant performances yield that two-accent prediction performs slightly better than three-accent prediction. The utmost correct classification rate obtained in three-accent classification is around 84.2%. Binary prediction performs slightly better with a correct classification rate of 84.5%. However, binary prediction requires one more step to map syllables to three-accent scheme. Predicting triple accent from binary predictions can lower the performance due to generalization of decision tree learning. Hence, syllable-pitch-contour prediction studies are based on triple accent classifications results. Best TP rates for not-accented, positive and negative accented syllables are 92.6%, 72.6% and 65.8%,

respectively. Not-accented syllables comprise approximately 64% of the database; that is the main reason why best performance of triple accent classification corresponds to not-accented syllables.

After accent prediction, slopes are predicted for the corresponding accented syllables. Accent states of the syllables are incorporated in slope prediction. Best performance in slope prediction is obtained using triple accents. Resultant correlation coefficient and root mean squared error are 0.86 and 0.69, respectively. For comparison purposes, correlation coefficient and root mean squared error corresponding to average slope values are given as -3.6e-16 and 1.9, respectively. Involving regression trees improves predictions using average slope values approximately 63.5%. Slope predictions are used to reconstruct sentence pitch contours.

Considering resultant sentence pitch contours, accent assignment algorithm is modified. Former version of the algorithm is constraint on selecting only one accented syllable per word. However, consecutive syllables may show rising/falling patterns in some words. The algorithm is improved to catch up all syllables that show rising and falling patterns. With this improvement, best accent classification with triple and two-accent classification is performed with 79.5% and 80.7% accuracy, respectively. The TP rates for triple accent classification are 0.87, 0.74, and 0.7 for not-accented, positive accented and negative accented syllables. Corresponding TP rates in two-accent classification are 0.77 and 0.84 for not-accented and accented syllables, respectively.

Using modified accent assignment scheme, best slope prediction is obtained via incorporating triple accents in learning. Corresponding correlation coefficient and root mean squared error are 0.91 and 0.59, respectively. Slopes predicted using regression trees provide an improvement of approximately 68.95% over average slope values. Sentence pitch contours are generated using accent information and predicted slopes. It is observed that slopes follow original pitch contour patterns. However, there are level differences between reconstructed sentence contours and the original contours. This level shift is mainly due to syllables having multimodalities. Multimodal syllables can not be modeled accurately with the current accent assignment algorithm since only one rising or falling slope per syllable can be assigned. However, multimodal syllables require more complex patterns. Future improvements to handle this problem are revisited in the subsequent section. Another point related to the resultant sentence-pitch-contour predictions is that predicted contours can not reach the maxima observed in the original

contour. The main reason for this phenomenon is such maxima are rarely observed in the database, at most one syllable per sentence. And they generally correspond to the lexically stressed syllable of the intended word (focus) in the sentence. Currently, focus information is not incorporated in decision tree learning. As discussed in the following section, using focus information is considered as a future work.

Both proposals are different from previous intonation modeling studies. Almost all intonation modeling systems rely on syllable units however their approaches differ in the way they utilize syllabic information. From the point of view of describing fundamental frequency contours, one of the proposed approaches can be considered as phonetic and the other as phonological. Both depend on phonetic analysis; however, the latter relies on describing accent scheme for Turkish. Both approaches rely on syllable pitch contours to predict pitch contours. Ten equidistant F0 values are used to develop codebooks or to assign pitch accent to the syllables. Based on the resulting scheme, predictions are performed. As introduced in Chapter 2, non-parametric methods also rely on raw F0 values however proposed methods differ in the way they utilize raw F0 values. Main differences are summarized in the following paragraphs.

Vector quantization is used in different areas of intonation modeling studies. However, they differ from the proposed approach. Most of the studies incorporating vector quantization in pitch contour modeling provide parametric representations considering all syllables or only accented syllables. Sigmoids, Bezier functions, polynomial extensions are used to represent pitch contours parametrically. Function coefficients are vector quantized using minimum distance criterion. Some others use a set of F0 values together with some duration parameters and perform vector quantization afterwards. Proposed approach performs vector quantization of all syllable pitch contours. Ten equidistant F0 values for each syllable are used as input to the vector quantization algorithm. Resulting codewords are used in pitch contour prediction. Codewords are also used in the determination of accented syllables. Hence, a mapping from phonetic description to phonological entities is performed. Consequently, pitch prediction is carried out by means of binary prediction. By means of two-level vector quantization, codebook inventory is pruned so that identical centroids are merged. Syllables are associated to accent status depending on the pruned centroids taking into account dynamic range and shape of the centroids. Binary prediction is performed to determine accent status of the syllables.

Other proposed approach uses phonetic analysis of pitch contours to assign accent labels to syllables. From the ten F0 values, mean and slope values for each syllable are calculated. Developed accent assignment algorithm associates accents considering slope and mean values of the syllables. A syllable is accented if it has a slope value that is greater than a determined threshold. Threshold is determined experimentally. However, an optimization can be performed on the threshold value. Optimization can be performed by means of analysis-by-synthesis. Threshold can be optimized by means of perceptual listening tests. Labeled database is used to predict whether a syllable is accented or not accented.

Other methods that employ non-parametric methods for intonation modeling predict every F0 value independently or using vector regression trees. Vector regression trees resemble our first proposal in the sense that F0 values are predicted considering a minimum distance criteria, usually Mahalonobis distance.

8.2 Future Directions on Turkish Prosody

Perceptual Evaluation of Performance: Phoneme durations and pitch contour modeling for Turkish is accomplished. Performances of developed models are evaluated quantitatively. However, prosody is meaningful perceptually. Hence, perceptual evaluations can be carried out to evaluate model performances as a future work.

Sentence Pitch Contour Modeling: Pitch contour modeling studies can consider sentence pitch contours, not syllables, for locating pitch accents. Syllables can be associated with pitch accents accordingly. Currently, slope computation considering syllable pitch contour is not robust such that each syllable is assigned at most one slope value. Assigning only one slope per syllable can not capture multimodal pitch patterns in some of the syllables. Moreover, slopes are computed considering syllable initial and final pitch values. Hence, multimodalities observed on syllable pitch contours do not yield significant slope values. Those syllables which have sharp peaks or valleys are not associated to pitch accents. **Figure 8-1** shows speech waveform, pitch contour and accents associated to syllables of the sentence ‘ancak savunanlar da hayli fazla deniliyordu’. On the syllable ‘cak’, there is a fall-rise pattern which corresponds to a local valley. Depending on the chosen threshold, accent assignment algorithm can assign only one type of accent although there are falling and rising patterns. Every prominent peak and valley as well as rise and fall can be captured considering overall sentence pitch

contours. Sentence contours can be inspected to find prominent pitch events (rise/fall, peak/valley). Slope values can be calculated considering the initial and final positions of corresponding events. However, how this information will be incorporated into the current study is still an open issue. There are several limitations with the current study: 1) Syllables either have positive accents, negative accents or no-accent however, fall-rise and rise-fall patterns comprise both accent schemes. Solution to this problem 2) Let us assume that the accent inventory is enlarged as in ToBI annotation scheme. Then, timing will be another problem. There may be early rises/falls or late rises/falls or they may appear right in the middle. 3) During reconstruction, predicted syllable slopes are used. Hence,

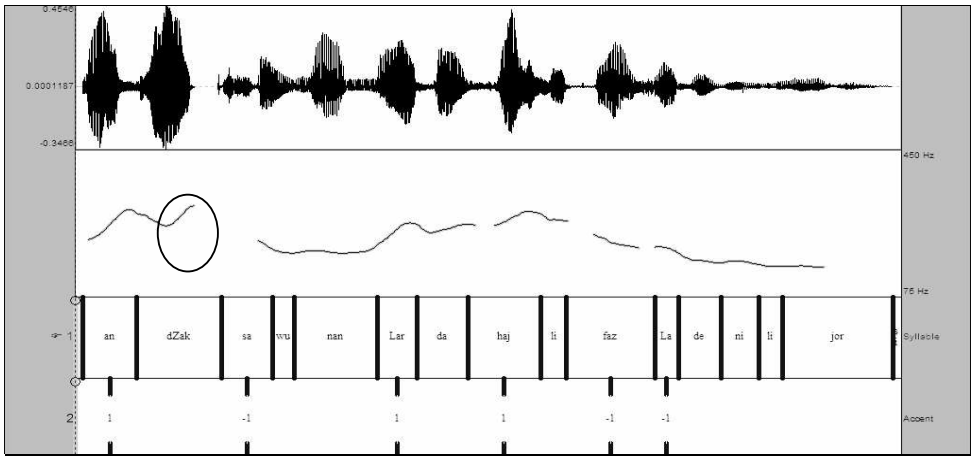

Figure 8-1: Speech waveform, corresponding smoothed and interpolated pitch contour, syllable labels and pitch accents of the sentence ‘ancak savunanlar da hayli fazla deniliyor’ (however it is said that the defenders are also too much).

Codebook Generation: Codebook generation algorithm can be revisited. In this study, pitch contours of every syllable is taken into account. Vector quantization is performed over all syllables to generate a syllable-pitch-contour codebook. However, sentences are not composed of successive pitch events. Events are separated by smooth contours that are not perceptually significant. Codebook generation does not rely on this fact. Codebook generation can be performed considering only prominent pitch events as a future study. This way, prominent syllables can be represented in detail and excess information related to smooth contours can be discarded.

Acoustic Segmentation of Syllables: For pitch contour modeling, syllable boundaries derived from phoneme boundaries are considered for each word taking into account orthographic word form. However, Turkish native speakers concatenate successive words if one ends with a consonant and other starts with a vowel. This phenomenon is known as *liaison*. Liaison is not applicable if there is a break between two successive words. Such a case is shown in **Figure 8-2**. The words 'analiz' and 'edilmek' obeys liaison rules and syllable boundaries can be assigned accordingly. The syllable boundaries considering single words are given in the figure. Considering liaison effect, the words act as a single word and syllables can be segmented as follows: 'a', 'na', 'li', 'ze', 'dil', 'mek'. With this modification, rise-fall pattern enclosed within the syllable 'liz' can be partitioned into rise and fall pattern enclosed in syllables 'li' and 'ze'.

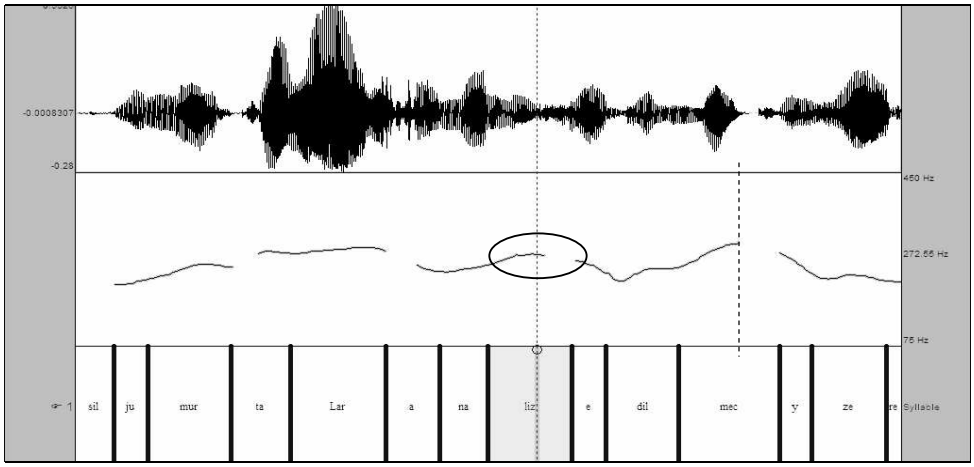

Figure 8-2: Speech waveform, corresponding pitch contour, and orthographic syllables of the sentence 'yumurtalar analiz edilmek üzere …'. Using syllable boundaries considering speech waveform, rise-fall pattern squeezed in the orthographic syllable 'liz' can be split into rise and fall patterns corresponding to acoustic syllables 'li' and 'ze'.

More on Speech Corpus: Phonetic transcription and alignment of the developed corpus are provided within the scope of this thesis. Phonetic alignment is performed automatically. Approximately 70% of the corpus is corrected manually and used in phone durations and pitch contour modeling. 30% of the corpus will be manually corrected in future.

Designed text database used in building speech corpus has not been annotated with punctuation marks. The main motivation in not using punctuation marks is to set the speaker completely free. With this way speaker uttered corresponding text in the way she thought to be correct. However, punctuation helps the speaker to impose regular intonation patterns. Without punctuation, although correct, unexpected intonation patterns can be observed. Punctuation marks also help in determining possible breaks in the speech. A full-stop and a comma generally corresponds a long and a shorter pause in speech. So utilization of punctuation marks provides speaker and model developer certain facilities. Therefore, punctuation marks can be provided to the designed text and re-recorded. Currently, phrase breaks are obtained perceptually. Perceptual phrase break assignment results in a more accurate break scheme than regarding only punctuation marks. Though, punctuation marks can be used to verify perceptual phrase breaks.

Lexical Stress Assignment for Complex Structures: Lexical stress assignment for compounds and phrases can be handled as a future work. Within the course of this thesis, a stress assignment algorithm has been generated considering Turkish stress rules. However, this algorithm considers words one-by-one and assigns lexical stress accordingly. Word stress pattern can be altered by compounding and phrasing in Turkish. In **Figure 8-3**, the sentence 'dövizde yapılan analizlerde ciddi bir sıçrama beklenmiyor yıl sonuna kadar' with its pitch contour is given. The phrase 'dövizde yapılan analizlerde' acts as a single word and the syllable 'de' of the word 'dövizde' is the lexically stressed syllable of the phrase. Currently, stress assignment algorithm handles each word independently and assigns corresponding lexical stresses. Accordingly, three syllables are lexically stressed: 'de' of dövizde', 'lan', and 'de[4]' of 'analizlerde'. To improve performance, stress assignment algorithm can be revised to handle compound words and phrases. However, it should be noted that the challenge lies in detecting compound words and phrases not in assigning lexical stress to them. Lexical stress rules apply compounds and phrases almost same if one can capture the quantifier of the compound/phrase.

Utilization of Focus Information: Focus is an important aspect of speech. It is observed that words that are focused reaches maximum pitch wherever it is located in the sentence. In our modeling studies, focus information has not been incorporated yet. However, incorporating focus information can greatly improve prediction performance.

[4] It should be noticed that, the rise in the second 'de' indicates that sentence will continue after the phrase.

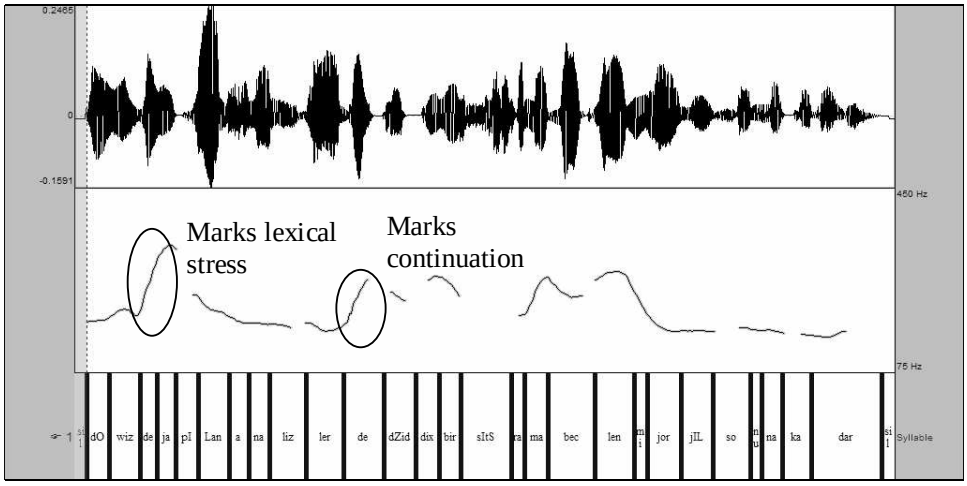

Figure 8-3: Speech waveform, smoothed and interpolated pitch contour, and orthographic syllable boundaries of the sentence ‘dövizde yapılan analizlerde ciddi bir sıçrama beklenmiyor yıl sonuna kadar’. The phrase ‘dövizde yapılan analizlerde’ acts as a single word and the syllable ‘de’ is the lexically stressed syllable of the phrase.

8.3 Discussions

Phoneme duration modeling:

For duration modeling, attributes are selected sequentially so that each new attribute increases prediction performance. Best prediction performance is obtained with *Phoneme Identity*; hence it is selected as the best predictor. It is observed that newly added attributes after seventh attribute do not provide further improvement. The seven attributes that best predicts phoneme duration are *Phoneme Identity*, *Left/Right*, *Lexical Stress*, *Syllable Type*, *Word Part-of-Speech*, *Phrase Break*, and *Number of Words to Following Phrase Break*.

It is also observed that the *Left/Right* attribute makes the best contribution to duration modeling. Hence, *Phoneme Identity* and *Context* are the two most influential attributes in duration modeling. Depending on the observations that *Phoneme identity* and immediate *left* and *right* context play an important role in duration prediction, it can be inspected that using larger contextual windows in duration modeling may improve performance. However, larger contextual windows require large databases since the number of units to be covered increases multiplicatively. For example, for a window of three phonemes

(left-current-right), the total number of possible combinations is 43x42x43 (= $0.077x10^6$) while for a window of 5 phonemes (left2-left1-current-right1-right2), the total number of combinations is 43x43x42x43x43 (= $0.14x10^9$). So, increasing the window size to include one more contextual phoneme results in a dimension increase of 43x43 (~99%). Hence, the problem can be investigated for a couple of phonemes only. Phoneme selection can be performed depending on the frequency of possible contextual windows. A speech database can be constructed to include possible combinations for the considered phonemes. Effect of larger contextual windows on phoneme duration can be revealed by means of quantitative analysis methods.

However, it is almost impossible to analyze effects of larger contextual windows on duration modeling for all phonemes with a limited database. Experiments performed using phonetic class instead of phoneme identity for left/right context show that reducing the dimension of contextual window does not result in a reduced performance but slightly better performance. Hence, the dimension problem encountered in contextual window/phoneme duration dependency can be handled using phonetic class instead of phonetic identity. Phonemes can be classified depending on their discriminative characteristics such as vowels versus consonants; voiced versus unvoiced; or more specifically depending on how they are produced.

Vowels are classified by the highest point reached by the tongue both in vertical and horizontal directions. These directions are split into two parts: High/Low; and Front/Back. Vowels are also split into two depending on their lip shaping: Round/Unround. Consonants involve constrictions, or gestures that narrow the vocal tract at a particular point. Consonants are discriminated with respect to their place of articulation: Bilabial/Labio-dental/Dental/Alveolar/Velar/Glottal; and manner of articulation: Plosive/Affricate/Stop/Nasal/Fricative/Approximant. Selection of features to be used in duration modeling as a predictor is another problem. Generally, manner of articulations are used in phoneme duration modeling studies. In our studies, instead of identity, manner of articulations for consonants and backness/frontedness for vowels are used. However, feature combinations or other aspects of features may be more appropriate. Hence, in order to determine the features to be used, experiments involving limited databases in which different aspects of consonants and vowels are handled can be formed.

Another point of interest is the effect of word/phrase/sentence (for paragraphs) boundaries in context study. Increasing the size of contextual windows is a step towards using word-sized or even larger units. When larger contextual windows are selected, the probability that a word boundary is enclosed within the selected window increases. Hence, one may wonder even if the context is same, is phoneme duration affected by existence of a word boundary. Effect of word boundary on context hence phoneme duration can be investigated using same contextual windows but with and without word boundaries. The sentences 'balık aldırdım' and 'balı kaldırdım' comprise a pair of such constitutions. Both sentences have the same phonetic sequence hence whatever context size is chosen the elements of the window will be the same. Effects of word boundary can be uncovered considering the phonemes at boundaries.

One other factor that may have impact on duration modeling is the speaking rate.

> *"…Results showed that the consonant and vowel durations were all significantly influenced by speaking rates and utterance units. At five kinds of speaking rates, the durations of vowels changed more than those of consonants. Fricative durations expanded more than the durations of other consonants, while the duration of unaspirated stops kept constant at five speaking rates. Vowels in monosyllable had longer duration than those in other utterance units…."* [Jing 2004]

As stated in Jing, consonant and vowel durations are affected by local speaking rate. Generally in duration modelling studies, this phenomenon is underestimated. The main reason for this is that duration modelling studies use recordings of a speaker with a normal style. However, most models depend on recordings that last over days. Hence, speaker shows variations during in his/her speech. Consequently, one may wonder how duration modelling performance is affected by changes in speaking rate. To observe the effects of speaking rate on duration modelling, first effects of speaking rate on phoneme durations are to be investigated. To this aim, a small speech database comprising normal, fast and slow speech can be constructed and average phoneme durations for each speaking rate can be determined. If there is a relationship between phoneme durations and speaking rate, then speaking rate can be incorporated as a predictor in duration modeling studies.

Discussions about phoneme duration modeling studies revealed the question that how well we model phoneme durations? Are the attributes used in duration modeling studies are sufficient? Or can the algorithms used in duration modeling studies develop duration models from the given attribute set adequately? Let us assume that the attributes are well

selected and the algorithms produce perfect models. Average RMSE between observed and predicted durations are around 20 ms in literature, that is, there is a considerable error in duration prediction. What is the reason for this much of RMSE in duration prediction? Is there randomness in duration modeling issues? In fact, phonemes are produced by a physiological system. Like every system, the product of a physiological system may have certain randomness while producing same phoneme even in the same context. Another important fact about duration modeling is that it relies on either manual or automatic labeled speech databases. Even in manual labeling, there may be inconsistencies which may result in randomness.

Segmentation inconsistencies are generally encountered in labeling vowel-voiced boundaries. In order to avoid inconsistencies, one may suggest using a labeling standard however such a standard may also be insufficient in determining the segment boundary especially in between voiced-semivowel transitions. Another solution may be considering larger units for such cases. Some of the duration modeling studies relies on syllable-sized units however in Turkish; syllables may be inadequate to capture such transitions. For example in the sentence 'roma jazarm1S', which can be rewritten in syllables as 'ro-ma-ja-zar-m1S', the syllables 'ma' and 'ja' can not be segmented accurately because of a vowel-semivowel-vowel transition. Therefore in Turkish, larger unit concept does not point out single type of units such as syllables but a unit with variable size.

In order to detect such units, visual cues may be incorporated in segmentation. Lip, tongue and chin motion can be used in segmentation of speech into consistent speech units. Modeling phoneme duration can then be transformed into modeling lip, tongue and chin timings which can be determined more accurately than phoneme durations.

Pitch contour modeling:

Pitch contour modeling studies generally rely on describing an intermediate representation of pitch contours. The purpose of using intermediate representations is to decrease the complexity of relation between linguistic attributes and the pitch contour, which is continuous in nature. ToBI and Fujisaki approaches are two extremes of intermediate representations. ToBI is a phonological model that represents pitch contours as a sequence of discrete symbols whereas Fujisaki's model interprets pitch contours as the superposition of three waveform components: baseline, local and global pitch excursions.

A major drawback of ToBI-like representations is the need for expert manual labeling which is too much time-consuming. Labeling consistency among different labelers is another problem; for example, one labeler may mark a syllable as accented while the other may not or one may mark a syllable as H+L*, the other as L*, and so on. Hence, such models are highly influenced by human factors.

Labeler dependency is eliminated by means of automatic pitch contour modeling methods such as Fujisaki or Tilt model. Both models analyze pitch contours automatically. However, Tilt model incorporates ToBI labels in the analysis procedure therefore it may be viewed as a semi-automatic model. Fujisaki model decompose pitch contours into three components: base F0, local excursions (pitch accents) and global excursions (phrase accents). Tilt model assigns a tilt value for each accent or boundary tone. Both models rely heavily on pitch contours themselves in order to extract the parameters necessary to synthesize the pitch contours.

Pitch contour extraction is performed in two ways. In the first approach, a laryngograph is used in recording. In the second approach, pitch contour is extracted from the speech signal itself. Laryngograph signal is more appropriate for pitch contour modeling studies than the pitch contour extracted from speech signal since it is more reliable. In this study, pitch contours are extracted from the speech signals and it is observed that although the performance of the algorithm used to extract pitch contours is quite good, it is almost impossible to avoid errors. Because of the timing considerations, the errors are not manually inspected within the framework of the study assuming that they can be compensated by smoothing and interpolation processes. The assumption holds generally however there are gross errors and hence manual tuning is required.

However, using laryngograph signal may also be inadequate for pitch contour modeling studies that rely on parametric representations. Tilt and Fujisaki model parameters are derived from continuous and smoother pitch contours as opposed to original contours. Original pitch contours reveal discontinuous patterns that also exhibit perturbations due to segmental effects, i.e. microprosody. The discontinuities are encountered at the unvoiced regions of the speech where there is no F0. Consonants, especially plosives, result in smaller perturbations such as sudden peaks in the pitch contour. Vowels also contribute to the pitch contour by means of their intrinsic pitch values, i.e. the intrinsic pitch values of high vowels are higher than those of low vowels; hence, same type of pitch accents show different pitch patterns.

Microprosody is eliminated by interpolation and smoothing in most of the studies. However, our studies show that interpolation and smoothing are not so powerful in reducing the microprosodic effects: Interpolation avoids the discontinuities due to unvoiced regions; smoothing eliminates short-time perturbations; however larger deviations are still present and cannot be handled automatically. It is obvious that segmental effects change the shape of the pitch contours hence attempts to handle microprosodic effects can be understood. However, preprocessing of pitch contours may change intonation. Interpolation does not affect the pitch contour shape since during synthesis time the unvoiced regions will be handled appropriately. However, smoothing has a considerable effect on pitch contours that cannot be reversed. Depending on the smoothing filter used, it is possible to smooth almost all details. Therefore, effects of smoothing on pitch contour modeling can be further studied. The degree of smoothing that will not change modeling performance can be determined by means of experiments. Besides, as mentioned previously, larger microprosodic effects are still present in pitch contours after preprocessing. Hence, their effects on pitch contour modeling can be analyzed considering perceptual tests, and algorithms to remove these perturbations can be developed.

One factor that affects pitch contour modeling performance is the speech corpus incorporated in modeling. As opposed to the most of the studies in pitch contour modeling, a collection of various kinds of isolated sentences is used in this study. This resulted in a decreased performance in our studies since each sentence type has not been sufficiently represented in the database. It is believed that increasing the size of the speech corpus by adding sufficient representatives of each sentence type increases prediction performance. Besides, most studies address pitch contour studies incorporating only one type of sentences such as declaratives. However, in order to develop a pitch contour model that serves as a tool for speech synthesis applications all sentence types should be modeled.

It is observed that modeling pitch contours of different sentence types at a time results in low prediction performance. Hence, pitch contour modeling for different sentence types should be handled as separate problems. Each sub-problem may require different set of attributes depending on the sentence type. For example, question sentences may need an extra attribute indicating the type of the question sentence: polar, inverted, or wh-question. Hence, for each sub-problem, different attribute sets can be generated and the

most influential attributes can be used to develop pitch contour model for the corresponding sentence type.

One issue that is not addressed frequently in pitch contour modeling studies is the utilization of speech databases that contains paragraph sentences. In our studies, isolated sentences are used to develop pitch contour models. Therefore, resulting models are appropriate for isolated sentences. Then, the question remains as what if one chooses to use paragraph sentences in pitch contour modeling. Like the downstepping in isolated sentences, pitch contours of paragraph sentences may show gradually decreasing patterns. There may even be a relationship between the pitch contours of paragraph sentences and isolated sentences that can be described in terms of sentence position in the paragraph. This relationship, if exists, can be revealed by forming a speech database of sentences at different locations of paragraphs.

Concludingly, pitch contour and duration modeling studies are far from complete; there are even untouched ideas that may increase the naturalness and quality of synthetic speech.

REFERENCES

[1] Abdullahmeşe, E., 1998, "*Fundamental Frequency Contour Synthesis for Turkish Text-to-Speech*", Master Thesis, Boğaziçi University.

[2] Adalı, O., (1979), "*Turkish Morphemes*", Turkish Language Association, Ankara

[3] Agüero, P. D., Wimmer K., Bonafonte A., (2004), "*Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters*", in Proceedings of International Conference on Spoken Language Processing (ICSLP'04), Jeju Island, Korea, October 4 - 8

[4] Aksan, D., (1995), "*With Every Aspects of Language: The Fundamentals of Linguistics*", Turkish Language Association, Ankara

[5] Arslan, L. M., Talkin, D., (1997), "*Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum*", Proceedings of the EUROSPEECH 1997, Rhodes, Greece, Vol. 3, pp. 1347-1350.

[6] Arslan, L. M., (1999), "*Speaker Transformation Algorithm Using Segmental Codebooks*", Speech Communication, vol. 28, pp. 211-226.

[7] Atabay, N., Özel, S., Çam, A., (1981), "*Turkish Syntax*", Turkish Language Association, Ankara

[8] Auran, C., "*Momel-Intsint v10.3*", <http://aune.lpl.univ-aix.fr:16080/~auran/english/ index.html>, Last accessed: October 2005.

[9] Nart, B. A., Oflazer, K., Say, B., (2003), "*The Annotation Process in the Turkish Treebank*", in Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC), April 13-14, 2003, Budapest, Hungary

[10] Banziger, T., Scherer, K. R., "*The role of intonation in emotional expressions*", Speech Communication, vol. 46, pp. 252-267

[11] Barker, C., (1989), "*Extrametricality, the Cycle, and Turkish Word Stress*", UCSC qualifying paper.

[12] Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., Nöth, E., (2001), "*Prosodic Models, Automatic Speech Understanding, and Speech Synthesis:Ttowards the Common Ground*", in Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark) volume 4 pp. 2285-2288.

[13] Batusek, R., (2002), "*A Duration Model for Czech Text-to-Speech Synthesis*", in Proceedings of Speech Prosody 2002, France,

[14] Bayer, A. O., (2005), "*A Study on Language Modeling for Turkish Large Vocabulary Continuous Speech Recognition*", Master Thesis, Middle East Technical University.

[15] Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S., (2004), "*The Original ToBI System and the Evolution of the ToBI Framework*", in S.-A. Jun (ed), Prosodic Models and Transcription: Towards Prosodic Typology, Oxford University Press, Chapter 2, pp. 9-54.

[16] Black, W. A., Hunt, J. A., (1996), "*Generating F0 Contours From ToBI Labels Using Linear Regression*", in Proceedings of Third International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, USA

[17] Black, A. W., Taylor, P. A., (1997), "*Assigning Phrase Breaks From Part-Of-Speech Sequences*", in Proceedings of Eurospeech97, vol. 2, pp. 995-998, Rhodes, Greece

[18] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, P. J., (1984), "*Classification and Regression Trees*", Wadsworth and Brooks

[19] Brinckmann, C., Trouvain, J., (2003), "*The Role of Duration Models and Symbolic Representation for Timing in Synthetic Speech*", in International Journal of Speech Technology, vol. 6, pp. 21-31.

[20] Botinis, A., Granström, B., Möbius, B., (2001), "*Developments and paradigms in intonation research*", in Speech Communication, vol. 33, pp .263-296.

[21] Bulyko, I., Ostendorf, M., (2002), "*A Bootstrapping Approach To Automating Prosodic Annotation For Limited-Domain Synthesis*", in Proceedings IEEE Workshop on Speech Synthesis.

[22] Buhmann, J., Vereecken, H., Fackrell, J., Martens, J. P., Coile, B. V., (2000), "*Data Driven Intonation Modeling Of 6 Languages*", in Proceedings 6th conference ICSLP, China Military Friendship Publish, Vol. 3., pp. 179-182.

[23] Büyük, O., Erdoğan, H., Oflazer, K., (2005), "*Konuşma Tanımada Karma Dil Birimleri Kullanımı ve Dil Kısıtlarının Gerçeklenmesi*", in Proceedings of IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı.

[24] Campbell, N., (2000), "*Timing in Speech: A Multi-Level Process*", in M. Horne (ed) Prosody: Theory and Experiment, Kluwer Academic Publishers, Dordrecht

[25] Campione E., Flachaire E., Hirst D.J. and Véronis J. (1997), "*Stylization and Symbolic Coding of F0, a Quantitative Approach*", in Proceedings of ESCA Tutorial and Research Workshop on Intonation, Athens, Greece

[26] Campione E., Véronis J., (1998a), "*A Multilingual Prosodic Database*", in Proceedings of ICSLP98, Sidney, Australia.

[27] Campione E., Véronis J., (1998b), "*Towards a Reversible Symbolic Coding of Intonation*", in Proceedings of ICSLP98, Sidney, Australia.

[28] Campione, E. Veronis, J., (1998c), "*A Statistical Study Of Pitch Target Points In Five Languages*" ICSLP'98, Sidney, Australia.

[29] Campione E., Hirst D.J. and Véronis J., (2000), "*Automatic Stylisation and Symbolic Coding of F0: Implementations of the INTSINT Model*", in A. Botinis (ed.) Intonation. Research and Applications. (Kluwer, Dordrecht),

[30] Carmichael, L., (2003), "*Intonation: Categories and Continua*", in Proceedings Northwest Linguistics Conference, Canada.

[31] Cassidy, S. and Harrington, J., (1996), "*EMU: an Enhanced Hierarchical Speech Data Management System*", Paper presented at the Speech Science and Technology Conference Adelaide, Australia.

[32] Chen, S. H., Hwang, S. H., Wang, Y. R., (1996), "*A Mandarin Text-to-Speech System*", in Computational Linguistics and Chinese Language Processing, vol.1, no.1, pp. 87-100

[33] Chen, S. H., Lai, W. H., Wang, Y. R., (2003), "*A New Duration Modeling Approach for Mandarin Speech*", in IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 4.

[34] Chung, H., (2002), "*Duration models and the perceptual evaluation of spoken Korean*", in Proceedings of Speech Prosody, pp. 219-222

[35] Clark, R. A. J., (2003), "*Generating Synthetic Pitch Contours Using Prosodic Structure*", Ph.D. Dissertation, University of Edinburgh.

[36] Cohen, W. W., (1996), "*Learning Trees and Rules with Set-Valued Features*", in Proceedings of 14th National Conference on Artificial Intelligence

[37] Cordoba, R., Vallejo, J. A., Montero, J. M., Gutierrez-Arriola, J., Lopez, M. A., Pardo, J. M., (1999), "*Automatic Modeling of Duration in Spanish Text-to-Speech System Using Neural Networks*", in ESCA workshop, Budapest, Hungary

[38] Cordoba, R., Montero, J. M., Gutierrez-Arriola, J, Vallejo, J. A., Enriquez, E., Pardo, J. M., (2002), "*Selection of the Most Significant Parameters for Duration Modeling in a Spanish Text-to-Speech System Using Neural Networks*", Computer Speech and Language, Vol. 16, pp 183-203

[39] Çarkı, K., Geutner, P, Schultz, T., (2000), "*Turkish LVCSR: Towards Better Speech Recognition for Agglutinative Languages*", in Proceedings of IEEE International Conference of Acoustics, Speech and Signal Processing, vol. 3, pp. 1563-1566.

[40] Çilingir, O., (2003), "*Large Vocabulary Continuous Speech Recognition*", Master Thesis, Middle East Technical University.

[41] Çiloğlu, T., Çömez, M. and Şahin, S., (2004), "*Language Modeling for Turkish as an Agglutinative Language*", in Proceedings of IEEE 12th Signal Processing and Communications Applications Conference (SİU'04), Kuşadası.

[42] Çömez, M. A., (2003), "*Large Vocabulary Continuous Speech Recognition for Turkish Using HTK*", Master Thesis, Middle East Technical University.

[43] Demircan, Ö., (2001), "*Turkish Prosody*", Yıldız Teknik University, İstanbul

[44] d'Alessandro, C. and Mertens, P., (1995), "*Automatic Pitch Contour Stylization Using a Model of Tonal Perception*", in Computer Speech and Language 9, 257-288.

[45] Dobnikar, A., (1996), "*Modeling Segment Intonation for Slovene TTS System*", in Proceedings of ICSLP96.

[46] D'Imperio, M., (2000), "*The Role of Perception in Defining Tonal Targets and Their Alignment*", PhD. Dissertation, the Ohio State University

[47] Dunham, M., (2003), "*Data Mining Introductory and Advanced Topics*", Prentice Hall.

[48] Dusterhoff, K. E., Black, A. W. and Taylor, P. A., (1999), "*Using Decision Trees Within the Tilt Intonation Model to Predict F0 Contours*",.in Proceedings of Eurospeech'99, Budapest, Hungary.

[49] Dusterhoff, K., (2000), "*Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*", Ph.D. Dissertation, University of Edinburgh.

[50] Febrer, A., Padrell, J., and Bonafonte, A., (1998), "*Modeling Phone Duration: Application to Catalan TTS*", in Proceedings of 3rd ESCA Workshop on Speech Synthesis, Australia

[51] Fidan, D., (2002), "*Türkçede Ezgi Örüntüleri*", Master Thesis, Ankara University.

[52] Fotinea, S. E., (1999), "*Sentence-Level Prosodic Modeling of the Greek Language with Applications to Text-To-Speech Synthesis*", Ph.D. Dissertation submitted to Department of Electrical and Computer Engineering National Technical University of Athens-NTUA.

[53] Frid, F., (2001), "*Prediction of intonation patterns of accented words in a corpus of read Swedish news*", Working Papers 49, pp. 42-45, Dept. of Linguistics, Lund University

[54] Fujisaki, H. and Nagashima, S., (1969), "*A Model for the Synthesis of Pitch Contours of Connected Speech*", Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, Vol. 28, pp. 53-60

[55] Fujisaki, H. and Hirose, K., (1984), "*Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese*", Journal of the Acoustical Society of Japan, Vol. 5, pp. 233-242

[56] Fujisaki, H., (2003), "*Prosody, Information, and Modeling - with Emphasis on Tonal Features of Speech –*", in Proceedings of Workshop on Spoken Language Processing, Mumbai, India (Invited Keynote Paper).

[57] Goldsmith, J., (1999), "*Dealing with Prosody in a Text-to-Speech System*", in Proceedings of International Journal of Speech Technology, vol. 3, pp. 51–63, Kluwer Academic Publishers

[58] Grice, M., Ladd, D. R., Araniti, A., (2000), "*On the Place of Phrase Accents in Intonational Phonology*", Phonology, vol. 17, pp. 143-185.

[59] Hatiboğlu, V., (1972), "*Turkish Syntax*", Turkish Language Association, Ankara

[60] Hirshberg, J., Pierrehumbert, J., (1986), "The *Intonational Structuring Of Discourse*", in Proceedings of 24th, Annual Meeting of the Association for Computational Linguistics.

[61] Hirst D.J., Di Cristo A. and Espesser R (2000), "*Levels of Representation and Levels of Analysis for the Description of Intonation Systems*", in M. Horne (ed) Prosody: Theory and Experiment, Kluwer Academic Publishers, Dordrecht

[62] Hirst, D. J., Ide, N., and Veronis, J., (1994), "*Coding Fundamental Frequency Patterns for Multilingual Synthesis with INTSINT in the MULTEXT Project*", in Proceedings of 2nd ESCA Workshop on Speech Synthesis, New York, USA

[63] Hirst, D. J., (2001), "*Automatic analysis of prosody for multilingual speech corpora*" in E.Keller, G.Bailly, J.Terken & M.Huckvale (ed) Improvements in Speech Synthesis, Wiley.

[64] Hirst, D. J. (2000), "*ProZed: a Multilingual Prosody Editor for Speech Synthesis*", in Proceedings of IEE Workshop State of the Art in Speech Synthesis, London.

[65] Hirst, D. J., Di Cristo, A., (1998), "*A Survey of Intonation Systems*" , in Hirst & Di Cristo (eds). Intonation Systems: A Survey of Twenty Languages, pp. 1-44, Cambridge University Press, Cambridge.

[66] Hirst, D. J., (1998), "*Intonation in British English*", in Hirst & Di Cristo (ed). Intonation Systems: A Survey of Twenty Languages, pp. 56-77, Cambridge University Press, Cambridge.

[67] Huang, A., Acero A., Hon, H., Ju, Y., Liu, J., Meredith S., Plumpe, M., (1997), "*Recent Improvements on Microsoft's Trainable Text-to-Speech Synthesis System-Whistler*" in Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97),Vol. 2.

[68] Iida, A., Campbell, N., (2001), "*A Database Design for a Concatenative Speech Synthesis System for the Disabled*", in Proceedings of SSW4.

[69] Jilka, M., Möhler, G., Dogil, G., (1999), "*Rules for the Generation of ToBI-based American English Intonation*", in Speech Communication, vol. 28, pp. 83-108.

[70] Jing, J., (2004), "*The influence of speaking rates and utterance units on segmental duration of Mandarin speech*", The Journal of the Acoustical Society of America, Vol. 116, Issue 4, pp. 2628.

[71] Keller, E., Werner, S., (1997), "*Automatic Intonation Extraction and Generation for French*", 14th CALICO Annual Symposium, ISBN 1-890127-01-9, West Point. NY.

[72] Kenney, Ng., (1998), "*Survey of Data-Driven Approaches to Speech Synthesis*", MIT Area Exam Paper, October

[73] King, S., Taylor P. A., (2000), "*Detection Of Phonological Features In Continuous Speech Using Neural Networks*", in Computer Speech and Language, 14(4), pp. 333-353.

[74] Klatt H. D., (1987), "*Review of Text-to-Speech Conversion for English*", in Journal of the Acoustical Society of America, vol. 82, pp. 737--793.

[75] Kochanski, G., Shih, C., Jing, H., (2003), "*Quantitative measurement of prosodic strength in Mandarin*", in Speech Communication, vol. 41, 625–645

[76] Kornfilt, J.,(1997), "*Turkish*", Routledge, London.

[77] Koutny, I., Olaszy, G., Olaszi, P., (2000), "*Prosody Prediction from Text in Hungarian and its Realization in TTS Conversion*", in International Journal of Speech Technology, vol. 3, pp. 187–200, Kluwer Academic Publishers.

[78] Krishna, N. S., Talukdar, P. P., Bali, K., Ramakrishnan, A. G., (2004), "*Duration Modeling for Hindi Text-to-Speech Synthesis System*", in Proceedings of International Conference on Spoken Language Processing (ICSLP'04), Korea

[79] Krishna, N. S., Murthy, H. A., (2004), "*Duration Modeling of Indian Languages Hindi and Telugu*", in Proceedings of 5th Speech Synthesis Workshop.

[80] Lee, S.and Oh, Y. W., (1999a), "*Tree-Based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems*", in Speech Communication, Vol. 28, pp 283-300

[81] Lee, S.and Oh, Y. W., (1999b), "*CART-Based Modeling of Korean Segmental Duration*", in Proceedings of Oriental COCOSDA WORKSHOP.

[82] Lee, S. and Oh, Y. W., (2001), "*Tree-Based Modeling of Intonation*", in Computer Speech and Language, Vol. 15, pp 75-98

[83] Lee, L. S., Tseng, C. Y., Ouh-Young, M., (1989), "*The Synthesis Rules in a Chinese Text-to-Speech System*", in IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37. No. 9, pp. 1309-1320.

[84] Lees, R. (1961) "*The Phonology of Modern Standard Turkish*", Uralic and Altaic Series 6, Bloomington: Indiana University.

[85] Lemmetty, S., (1999), "*Review of Speech Synthesis Technology*", Master Thesis, Helsinki University of Technology, March 1999

[86] Louw, J. A., Barnard, E., (2004), "*Automatic intonation modeling with INTSINT*", in Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa, pp. 107-111, Grabouw.

[87] METU-Sabanci Turkish Treebank, "*METU-Sabancı Turkish Treebank*", <http://www. ii.metu.edu.tr/~corpus/treebank.html>, Last accessed: October 2005.

[88] Mixdorff, H., (2001), "*MFGI, a Linguistically Motivated Quantitative Model of German Prosody*", In Improvements in Speech Synthesis, E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (Ed.), Wiley Publishers, pages 134-143, UK.

[89] Mixdorff, H., (2000), "*A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters*" in Proceedings of ICASSP 2000, vol. 3, pages 1281-1284, Istanbul, Turkey.

[90] Mixdorff, H., Jokisch, O., (2001), "*Implementing and Evaluating an Integrated Approach to Modeling German Prosody*", in Proceedings of the 4th ISCA Workshop on Speech Synthesis, Perth Atholl Palace Hotel, Scotland, pp. 211-216.

[91] Monaghan, A. I. C., (1992a), "*Heuristic Strategies for the Higher Level Analysis of Unrestricted Text*", In G. Bailly & C. Benoit (ed), Talking Machines, pp. 143-161. Amsterdam: Elsevier.

[92] Monaghan, A. I. C., (1992b), "*Generating Synthetic Prosody: Means & Ends*", Actes du Seminaire Prosodie, Aix-en-Provence, pp. 9-24, October (Invited keynote paper).

[93] Monaghan, A., (1997), "*The Role of Prosody in NLP*", Keynote paper, in Proceedings of FRACTAL 1997

[94] Möbius, B. and van Santen, J. P. H., (1996), "*Modeling Segmental Duration in German Text-to-Speech Synthesis*", in Proceedings of International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, USA, October 3-6, Vol. 4, pp 2395-2398

[95] Möhler , G., (1998), "*Describing Intonation with a Parametric Model*", in Proceedings of ICSLP98, Sydney.

[96] Möhler, G., Conkie, A., (1998), "*Parametric Modeling of Intonation Using Vector Quantization*", in Proceedings of Third International Workshop on Speech Synthesis.

[97] Möhler, G., (1999), "*Comparing Two Different Principles of Parametric F0 Modeling*", in Proceedingsof the Joint ASA/DAGA Meeting Berlin.

[98] Oflazer, K., (1994), "*Two-level Description of Turkish Morphology*", Literary and Linguistic Computing, Vol. 9, No: 2, 1994.

[99] Oskay, B., Salor Ö., Özkan, Ö., Demirekler M. and Çiloğlu T., (2001), "*Türkçe Tümceler için Metinden Ezgi Belirlenmesi ve Uygulanması*", in Proceedings of IEEE 12th Signal Processing and Communications Applications Conference (SİU'01), Gazimagusa

[100] Oskay, B., (2002), "*Automatic Modeling of Turkish Prosody*", Masters Thesis, Middle East Technical University, September 2002

[101] Özge, U., (2003), "*A Tune-Based Account of Turkish Information Structure*", Master Thesis, Middle East Technical University.

[102] Pierrehumbert, J., (1983), "*Automatic Recognition of Intonation Patterns*", in Proceedings of the 21st Annual Meeting on Association for Computational Linguistics.

[103] Pierrehumbert, J., (2000), "*Tonal Elements and Their Alignment*", in M. Horne (ed) Prosody: Theory and Experiment, Kluwer Academic Publishers, Dordrecht

[104] Pirker, H., Alter, K., Rank, E., Matiasek, J., Trost H., Kubin, G., (2997), "*A System of Stylized Intonation Contours in German*", in Proceedings of Eurospeech-97, Vol.1, pp.307-310, Rhodes.

[105] Boersma, P., Weenink, D., "*Praat: doing phonetics by computer* (Version 4.3.01)", <http://www.praat.org/>, Last accessed: October 2005.

[106] Wood, S., "*Praat for beginners*", <http://www.ling.lu.se/persons/Sidney/praate/>, Last accessed: October 2005.

[107] Riedi, M. P., (1998), "*Controlling Segmental Duration in Speech Synthesis Systems*", Ph.D. Dissertation, Swiss Federal Institute of Technology, Zurich

[108] Roiger, R., Geatz, W. M., (2002), "*Data Mining A Tutorial Base Primer*", Addison Wesley.

[109] Ross, K. N., (1995), "*Modeling of Intonation for Speech Synthesis*," Ph.D. dissertation, Boston University College of Engineering.

[110] Sakurai, A., Hirose, K. and Minematsu, N., (2003), "*Data-Driven Generation of F0 Contours Using a Superpositional Model*", Speech Communication, Vol. 40, pp 535-549

[111] Salor, Ö., Pellom, B., Çiloglu, T., et.al., (2002a), "*New Corpora and Tools for Turkish Speech Research*", in Proceedings of IEEE Signal Processing And Communications Applications Conference (SİU'02), Pamukkale

[112] Salor, Ö., Pellom, B., Çiloglu, T., et.al., (2002b), "*On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language*", in Proceedings of the International Conference on Spoken Language Processing (ICSLP'02), Denver, USA, September 2002

[113] Salor, Ö.and Demirkler, M., (2004), "*Speech Conversion Using MELP Speech Coding Algorithm*", in Proceedings of IEEE 12th Signal Processing and Communications Applications Conference (SİU'04), Kuşadası

[114] Sezer, E., (1981), "*On Non-Final Stress in Turkish*", Journal of Turkish Studies, Vol. 5, pp. 61-69.

[115] Shih, C., Kochanski, G., (2002), "*Prosody and Prosodic Models*", in Proceedings of ICSLP 2002, Denver Colorado

[116] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., (1992), "*ToBI: A Standard For Labeling English Prosody*" in Proceedings of the 1992 International Conference on Spoken Language Processing, vol. 2, pp. 867-870.

[117] Sreenivasa, K. R., Yegnanarayana, B., (2004), "*Modeling Syllable Duration in Indian Languages Using Neural Networks*" in *Proceedings Int. Conf. Acoust., Speech Signal Processing*, Montreal, Quebec, Canada, pp. 313-316.

[118] Sun, X., (2002a), "*The Determination, Analysis, and Synthesis of Fundamental Frequency*", PhD. Thesis, Northwestern University.

[119] Sun, X., (2002b), "*F0 Generation For Speech Synthesis Using A Multi-Tier Approach*", in Proceedings of ICSLP02.

[120] Syrdal, A., Moehler, G., Dusterhoff, K., Conkie, A, and Black, A. (1998), "*Three Methods of Intonation Modeling*", in Proceedings of 3rd ESCA Workshop on Speech Synthesis, pp. 305-310, Jenolan Caves, Australia,

[121] Şahin, S., (2003), "*Language Modeling for Turkish Continuous Speech Recognition*", Master Thesis, Middle East Technical University.

[122] Taylor, P. A., (1992), "*A Phonetic Model of English Intonation*", Ph.D. Dissertation, University of Edinburgh.

[123] Taylor, P. A., Isard, S.D., (1992), "*A New Model Of Intonation For Use With Speech Recognition And Synthesis*", in International Conference on Spoken Language Processing, Banff, Canada

[124] Taylor, P. A., (1995), "*Using Neural Networks To Locate Pitch Accents*", In Proceedings Eurospeech '95, Madrid.

[125] Taylor, P. A., (1998), "*The Tilt Intonation Model*", in Proceedings of ICSLP98.

[126] Taylor, P. A., (2000), "*Analysis and synthesis of Intonation Using the Tilt Model*", Journal of the Acoustical Society of America, 107(3), pp. 1697-1714.

[127] Taylor, P. A., (1995), "*The Rise/Fall/Connection Model of Intonation*", Speech Communication, vol. 15, pp. 169-186.

[128] Taylor, P. A., (2000), "*Concept-to-Speech by Phonological Structure Matching*", Philosophical Transactions of the Royal Society, Series A.

[129] The Ohio State University Department of Linguistics, "*ToBI*", <http://www.ling.ohio-state.edu/~tobi/>, Last accessed: October 2005.

[130] Türk O. and Arslan L. M. (2002), "*Subband Based Voice Conversion*", Proceedings of the ICSLP 2002, Vol. 1, pp.289-292, September 2002, Denver, Colorado, USA.

[131] Türk, O. and Arslan, L., (2004), "*Robust Voice Conversion Methods*", in Proceedings of IEEE 12th Signal Processing and Communications Applications Conference (SİU'04), Kuşadası

[132] Underhill, R., (1976), "*Turkish Grammar*", Cambridge, MA: MIT Press.

[133] University of Cambridge, (2005), "*HTK Speech Recognition Toolkit*". <http://htk.eng.cam.ac.uk/>, Last accessed: October 2005.

[134] van Santen, J. P. H., Sproat, R. W. S., Olive, J. P., Hirschberg J. (1997) "*Progress in Speech Synthesis*", Springer-Verlag New York, Inc.

[135] Vegnaduzzo, M., (2003), "*Modeling Intonation for the Italian Festival TTS Using Linear Regression*", Master Thesis, University of Edinburgh.

[136] Venditti, J. J., van Santen, J. P. H., (1998), "*Modeling Vowel Duration for Japanese Text-to-Speech Synthesis*", in Proceedings of the International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia

[137] Véronis, J., Di Cristo, P., Courtois, F., Chaumette, C., (1998), "*A Stochastic Model of Intonation for Text-To-Speech Synthesis*", Speech Communication, 26(4), pp. 233-244.

[138] Violaro, F., Böeffard, O., (1998), "*A Hybrid Model for Text-to-Speech Synthesis*", IEEE Transactions On Speech And Audio Processing, Vol. 6, No. 5.

[139] Vural, E. and Oflazer, K., (2004), "*A Prosodic Turkish text-to-Speech Synthesizer*", in Proceedings of IEEE 12th Signal Processing and Communications Applications Conference (SİU'04), Kuşadası

[140] Wells, J.C., "*SAMPA for Turkish*", <http://www.phon.ucl.ac.uk/home/sampa/turkish.htm>, Last accessed: October 2005.

[141] Witten, H. I. and Frank, E., (1999), "*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*", Morgan Kauffman Publishing

[142] Wright, H., Taylor, P., (1997), "*Modeling Intonational Structure Using Hidden Markov Models*", in ESCA workshop on Intonation: Theory Models and Applications, Athens, Greece.

[143] Xu, Y., Wang, Q. E., "*Pitch targets and their realization: Evidence from Mandarin Chinese*", Speech Communication 33 (2001) 319±337

[144] Yapanel, Ü., (2000), *"Garbage Modeling Techniques For A Turkish Keyword Spotting System*", Master Thesis, Boğaziçi University.

[145] Yılmaz, C., (1999), "*A Large Vocabulary Speech Recognition System for Turkish*", Master Thesis Bilkent University.

APPENDIX A

SYLLABLE PITCH CONTOUR CODEBOOK

In this part of the Appendix, cluster centroids and cluster members described in Section 7.1.2 is given.

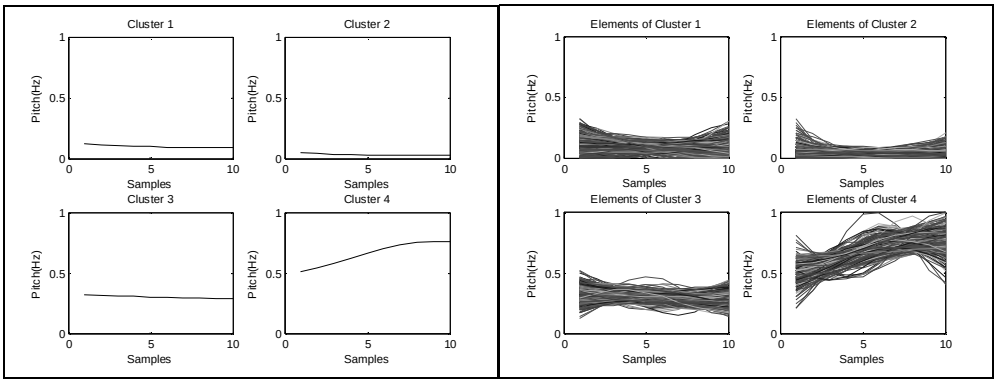

Figure A-1: Cluster centroids (left) and cluster members (right).

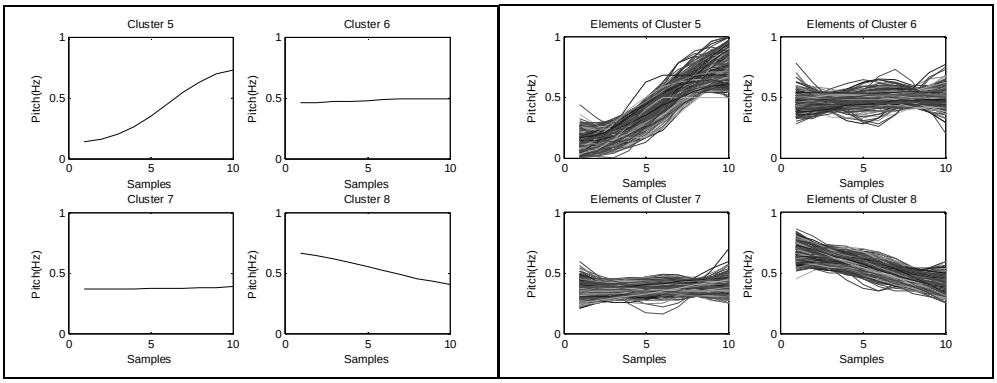

Figure A-2: Cluster centroids (left) and cluster members (right).

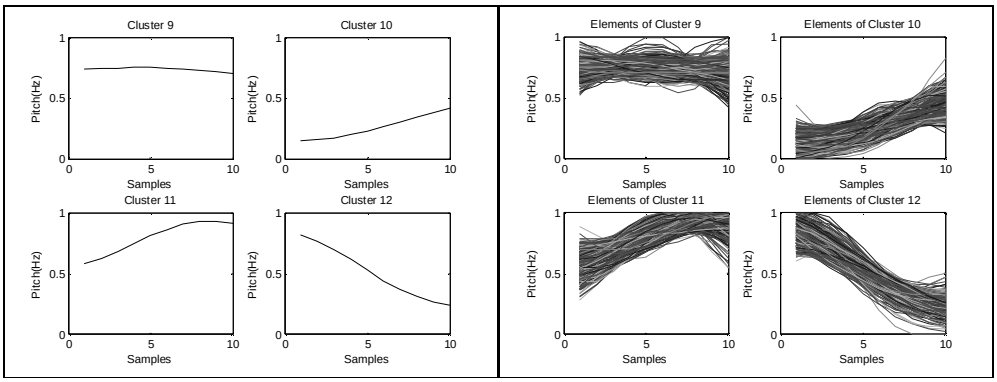

Figure A-3: Cluster centroids (left) and cluster members (right).

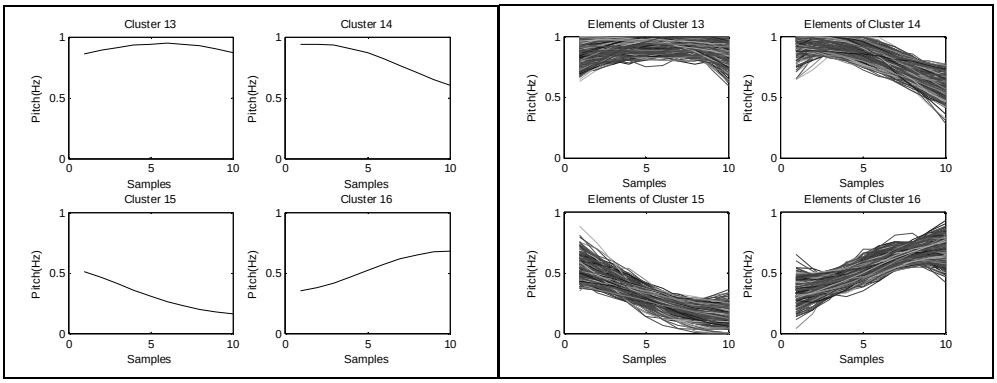

Figure A-4: Cluster centroids (left) and cluster members (right).

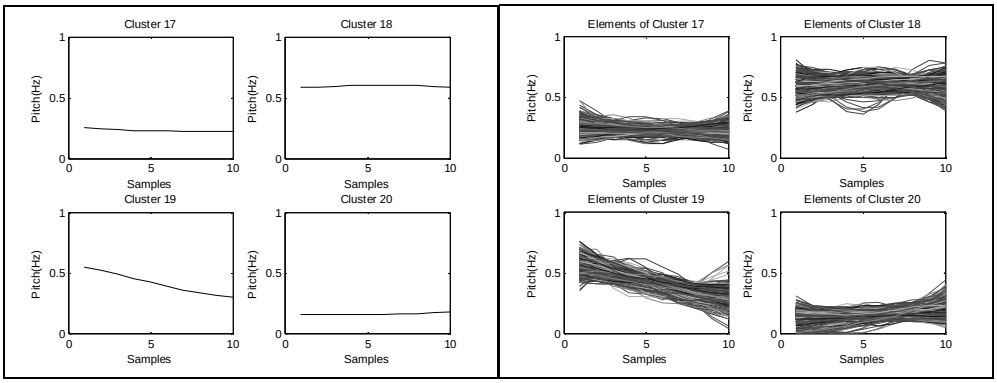

Figure A-5: Cluster centroids (left) and cluster members (right).

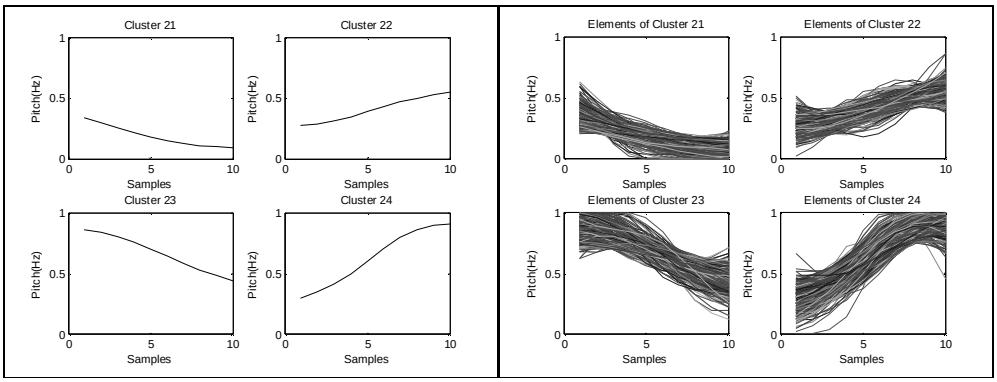

Figure A-6: Cluster centroids (left) and cluster members (right).

APPENDIX B

ACCENT ASSIGNMENT USING SYLLABLE PITCH CONTOUR CODEWORDS

In this part of the Appendix, cluster centroids and members obtained using two-level k-means algorithm described in Section 7.1.3 is presented. Cluster size is reduced

1) by eliminating clusters with centroids representing levels or pure rises and falls

2) by merging clusters of the same shape (determined by 25 cluster centroids) into single clusters. Eliminated cluster centorids are marked with X.

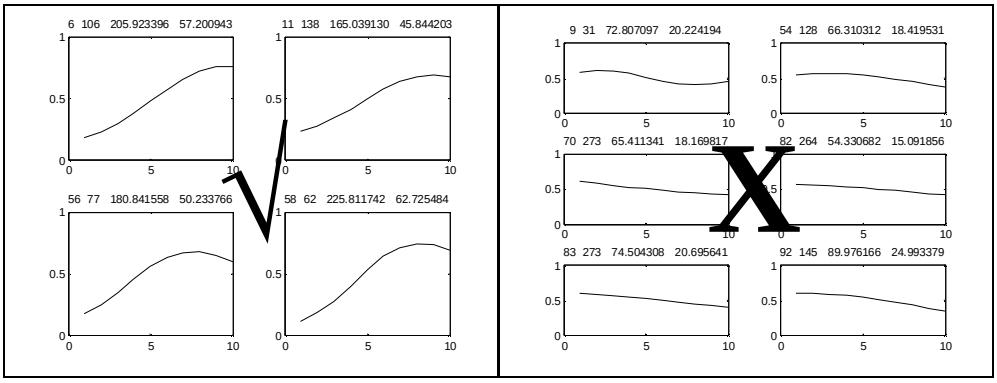

Figure B-1: Cluster centroids: numbers represent centoid's ID, frequency of pitch contours represented by this centroid, dynamic range of the centroid with respect to constant $F0_{min}$ and $F0_{max}$, and percentage of the dynamic range.

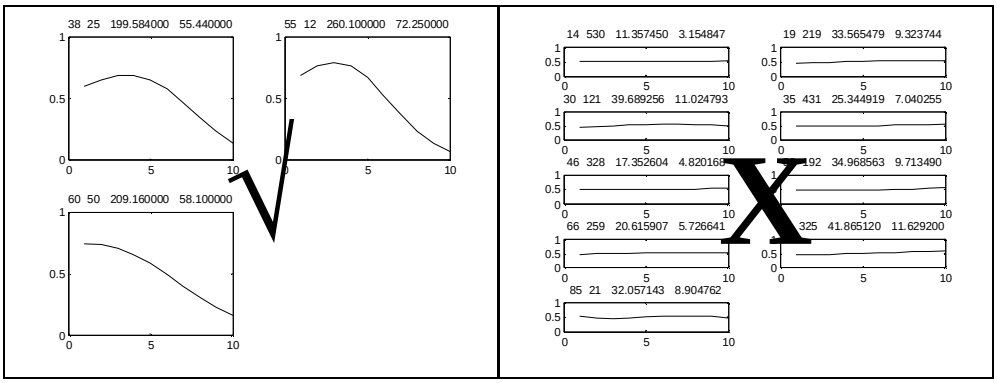

Figure B-2: Cluster centroids

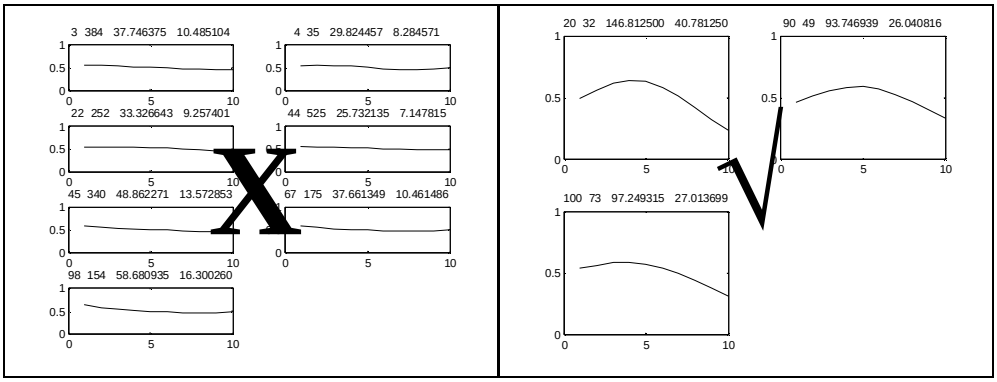

Figure B-3: Cluster centroids

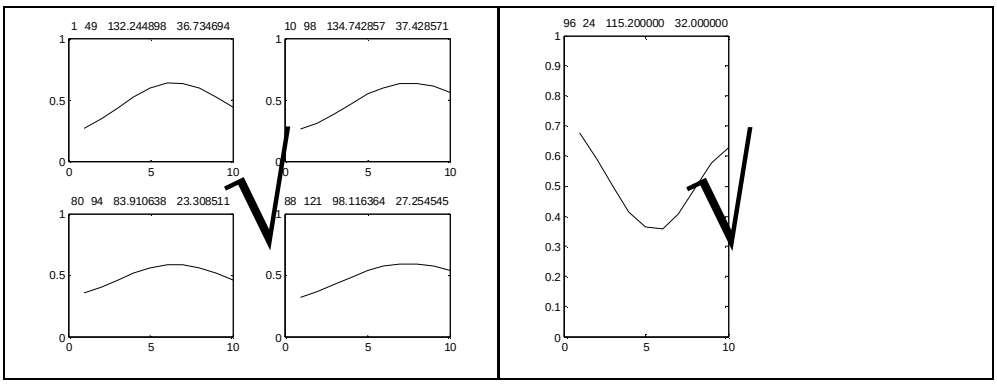

Figure B-4: Cluster centroids

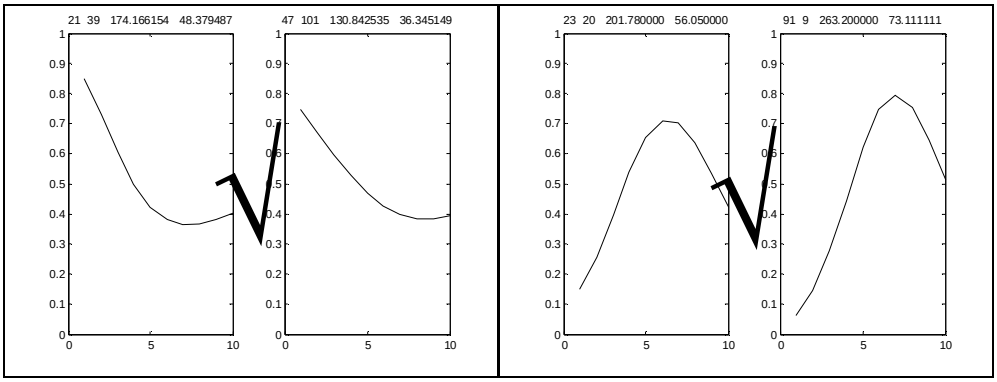

Figure B-5: Cluster centroids

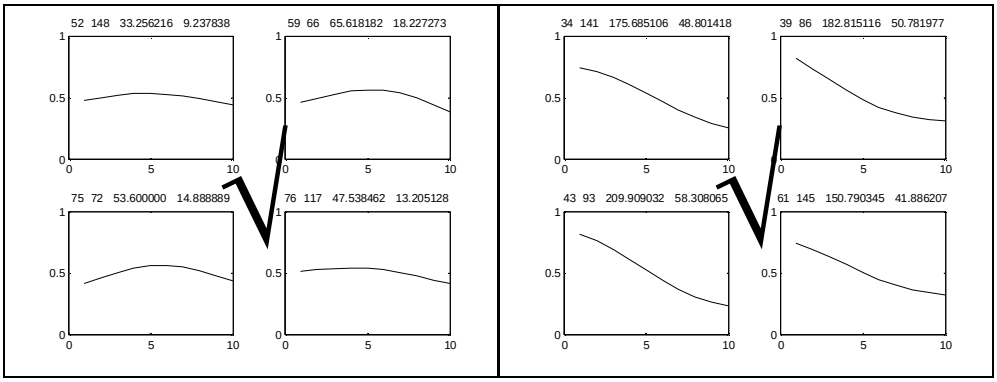

Figure B-6: Cluster centroids

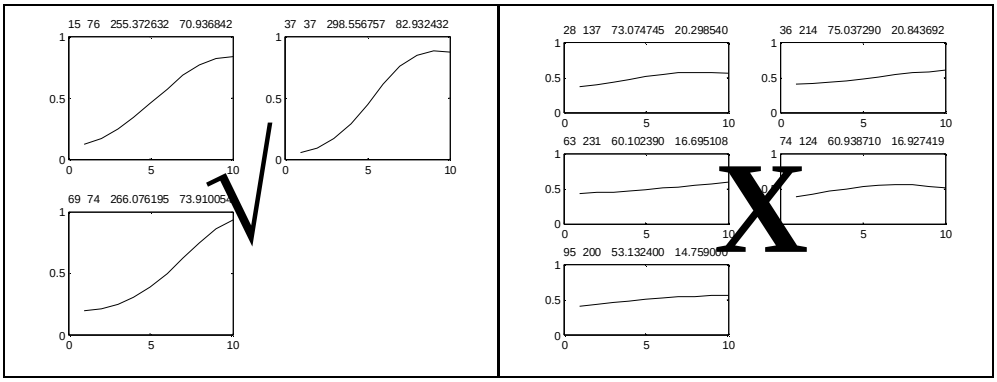

Figure B-7: Cluster centroids

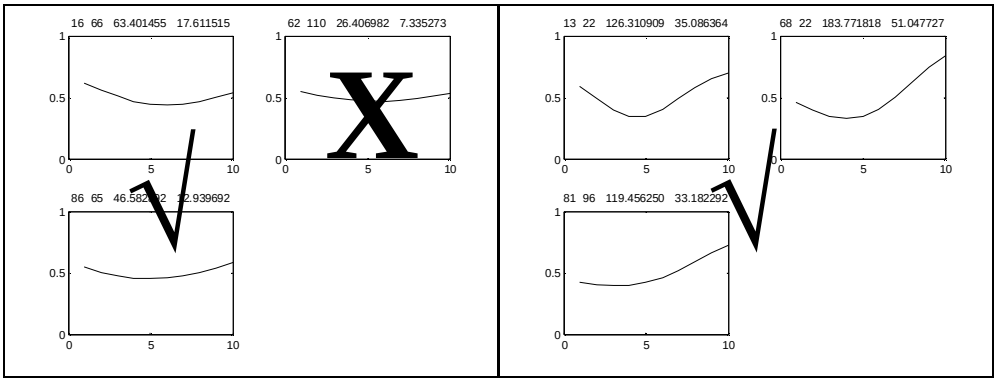

Figure B-8: Cluster centroids

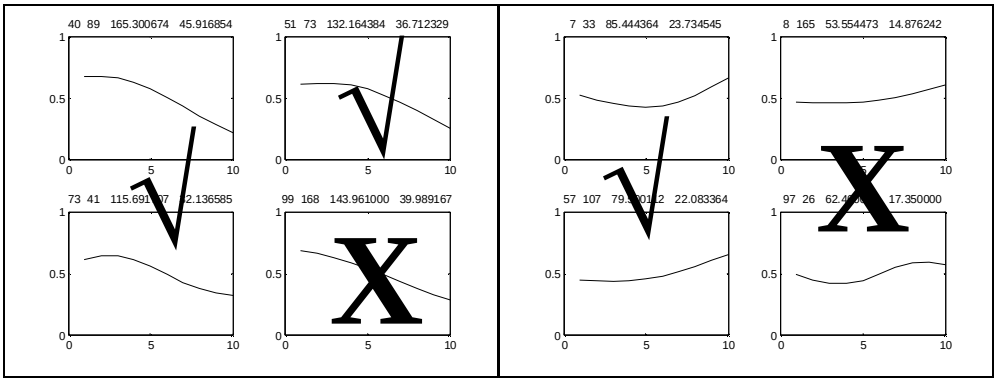

Figure B-9: Cluster centroids

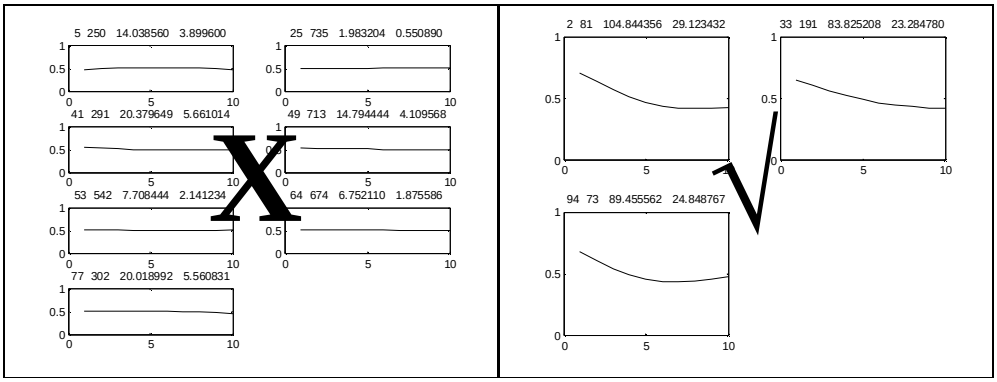

Figure B-10: Cluster centroids

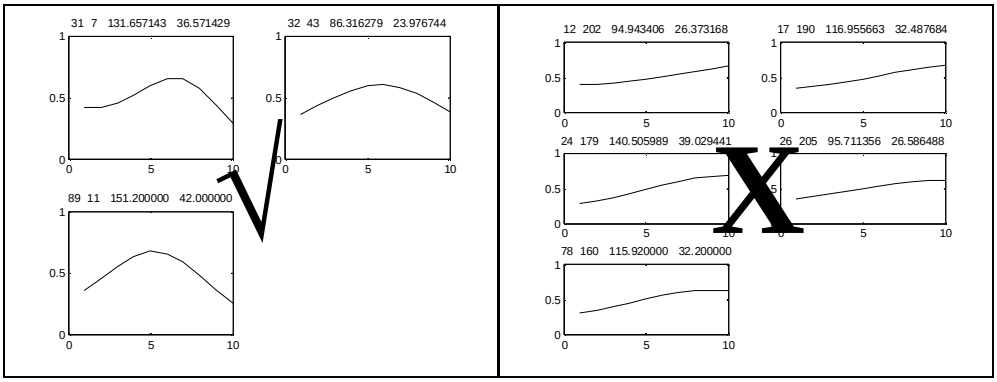

Figure B-11: Cluster centroids

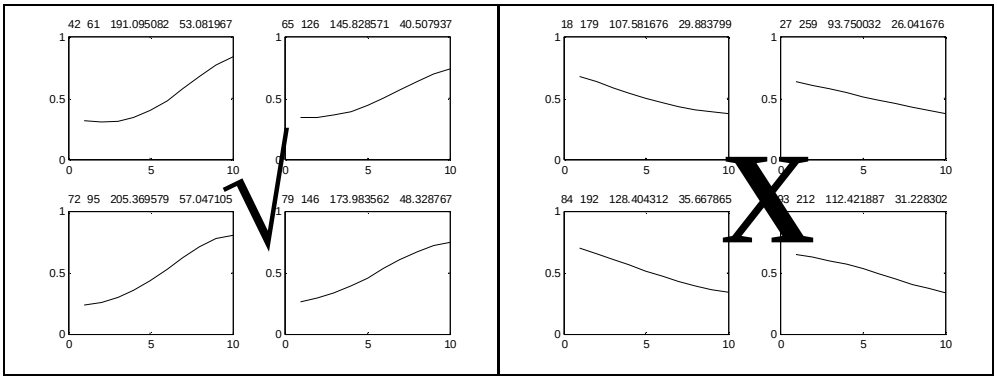

Figure B-12: Cluster centroids

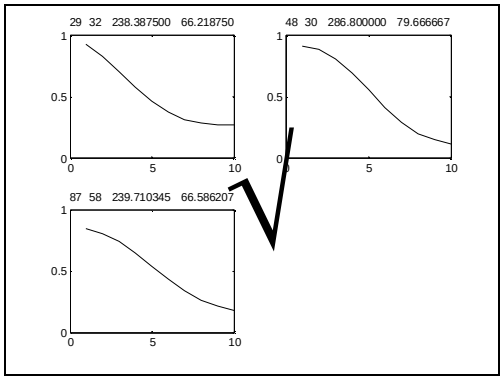

Figure B-13: Cluster centroids

VITA

Özlem Öztürk was born in Ankara, Turkey in 1973. She received her B.Sc. and M.Sc. degree from Middle East Technical University (METU), Department of Electrical and Electronics Engineering in 1995 and 1999, respectively. She is working towards Ph.D. degree at METU, Department of Electrical and Electronics Engineering since 1999.

She worked as a research and teaching assistant at METU from August 1995 to June 2001. Then she has joined Dokuz Eylül University (DEU), Department of Electrical and Electronics Engineering, where she is currently employed as a research and teaching assistant since June 2001. She has worked as a researcher at the TCTS Labs, Mons, Belgium, from August 2001 to October 2001. Her areas of interest are signal processing, speech synthesis, text-to-speech, and duration and pitch contour modeling.

Her publications are:

[1] Özkan, Ö., (1999), "*Neural Algorithms for Blind Seperation of Sources*", Masters Thesis, Middle East Technical University, Ankara.

[2] Özkan, Ö., Ünver, Z., (1999), "*Blind Separation of Mixtures of Speech and Music Burried in Noise*", in Proc. of the 8th Electrical Electronics and Computer Engineering Symposium, Gaziantep (Turkey), September 6-12, (in Turkish).

[3] Oskay, B., Salor, Ö., Özkan, Ö., Demirekler, M., Çiloğlu, T., (2001), "*Prosody Modeling For Turkish Sentences and Its Application*", in Proc. of the 9th Signal Processing and Applications Symposium, Gazimağusa (Turkish Republic of Northern Cyprus), April 25-27, (in Turkish).

[4] Öztürk, Ö., Çiloğlu, T., Demirekler, M., Bozşahin, C., (2002), "*Design of a Verb Lexicon For Turkish Text-to-Speech System*", in Proc. of the 10th Signal Processing and Applications Symposium, Denizli (Turkey), (in Turkish).

[5] Öztürk, Ö., Bozkurt, B., Çiloğlu, T., Dutoit, T., (2003), "*Database Development TTS Speech Corpus Building Using a Greedy Selection Algorithm*", in Proc. of the 11th Signal Processing and Applications Symposium, İstanbul (Turkey), (in Turkish).

[6] Hacioğlu, K., Pellom, B., Çiloğlu, T., Öztürk, Ö., Kurimo, M., Creutz, M., (2003), "*World Splitting for Turkish LVCSR*", in Proc. of the 10th Signal Processing and Applications Symposium, İstanbul (Turkey).

[7] Bozkurt, B., Öztürk, Ö., Dutoit, T., (2003), "*Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection*", EUROSPEECH 2003, Geneva (Switzerland).

[8] Hacioğlu, K., Pellom, B., Çiloğlu, T., Öztürk, Ö., Kurimo, M., Creutz, M., (2003), "*On Lexicon Creation for Turkish LVCSR*", EUROSPEECH 2003, Geneva (Switzerland).

[9] Öztürk, Ö., Çiloğlu, T., (2004), "*Segmental Duration Modeling for Turkish TTS*", in Proc. of the 11th Signal Processing and Applications Symposium, İzmir (Turkey), (in Turkish).

[10] Öztürk, Ö., Çiloğlu, T., (2004), "*Phonetization and Lexical Stress for Turkish TTS Systems*", in Proc. of the 11th Signal Processing and Applications Symposium, İzmir (Turkey), (in Turkish).

[11] Öztürk, Ö., Salor, Ö., Çiloğlu, T., Demirekler, M., (2004), "*Duration Modeling For Turkish Text-to-Speech Synthesis System*", LREC 2004, Lisbon (Portugal).