Automatic Generation of Fundamental Frequency for Text-to-Speech Synthesis

by

Aaron Seth Cohen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Bachelor of Science in Electrical Engineering and Computer Science

and

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1997

Author .....
Department of Electrical Engineering and Computer Science
May 16, 1997

Certified by.....
Victor Zue
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Theses

Automatic Generation of Fundamental Frequency for Text-to-Speech Synthesis

by

Aaron Seth Cohen

Submitted to the Department of Electrical Engineering and Computer Science
on May 16, 1997, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The problem addressed by this research is the automatic construction of a model of the fundamental frequency (F_0) contours of a given speaker to enable the synthesis of new contours for use in Text-to-Speech synthesis. The parametric F_0 generation model designed by Fujisaki is used to analyze observed F_0 contours. The parameters of this model are used in conjunction with linguistic and lexical information to form context based prototypes. The success of the F_0 generation is evaluated using both objective error measures and subjective listening tests.

Thesis Supervisor: Victor Zue
Title: Senior Research Scientist

Acknowledgments

The work for this thesis was carried out at IBM's T. J. Watson Research Center in Yorktown Heights, New York. I would like to thank my direct supervisor, Robert Donovan, for providing many ideas and keeping me focussed throughout the project. I would also like to thank my manager, Michael Picheny, who allowed me to pursue this research independently. Many other members of the speech group at IBM have given me quite a bit of help, especially Raimo Bakis, Ramesh Gopinath, Mukund Padmanabhan, Lalit Bahl, and Adwait Radnaparkhi.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis presents work on the automatic generation of fundamental frequency (F_0) contours for text-to-speech synthesis. The F_0 contour conveys a great deal of information about the meaning of a sentence, and without an appropriate one an utterance can be perceived to be quite unnatural. Researchers have been attempting to generate adequate contours throughout the history of the study of text-to-speech synthesis. In an overview of the technology of twenty years ago [1], Allen indicated that very little was known about generating F_0 contours besides several linguistic theories. In another extensive survey of text-to-speech technology ten years later [26], Klatt presented several generation algorithms, although the rules that determined the F_0 contours were all created by hand. More recently, there has been a greater trend toward creating speech synthesis systems using automatic training. For example in [8], representative segments of a speech waveform are chosen from a training database and combined using the Pitch Synchronous Overlap and Add (PSOLA) algorithm [9]. There have also been some studies of generating F_0 using trainable models. An example of applying this trend to F_0 generation appears in [43], which breaks up the F_0 generation problem into two parts, predicting abstract prosodic labels from text and generating the F_0 contour from the labels. One major difference between the current work and previous research is that a hand labeled prosodic database is not needed here, only a reliable F_0 extraction algorithm.

1.1 Problem Statement

The objective of this research is to predict fundamental frequency (F_0) for text-to-speech synthesis. Traditionally, this was achieved through the use of rules, which were designed by hand to capture the most important aspects of F_0 generation from text. More recently, trainable approaches have been attempted in which the rules are created by statistically analyzing a training database. A trainable system is preferable for two reasons. Firstly, the properties of the text that are important for predicting F_0 can be derived from the data instead of being manually imposed. Secondly, the system should capture and reproduce the training speaker's speaking style. For example, in [43], a sequence of Tones and Break Indices (ToBI) labels were predicted from text and a F_0 contour generated from the ToBI labels. In [36], the relationship between parameters of the Fujisaki model and linguistic features were statistically analyzed using regression trees.

In the current research, a trainable model is constructed to predict accents of the Fujisaki model [21]. Initially the Fujisaki model is used to analyze observed F_0 contours, using a two-stage searching procedure with linguistic constraints (see Section 3.3). Models for word-level and phrase-level F_0 effects are built by statistically analyzing the parameters of the Fujisaki model in relation to the linguistic and lexical contexts in which they occur (see Sections 3.4 and 3.5). In synthesis, the most likely sequence of accent prototypes is chosen using a dynamic programming algorithm (see Chapter 4).

The success of the generation of F_0 contours for new text is measured by comparing those contours with observed ones using a set of sentences different from the ones used for training. In this research, this evaluation is carried out using both objective measures and subjective listening tests (see Chapter 5). The two steps of training and testing approximate the way in which a F_0 generation model might be used in a text-to-speech system that tries to emulate the voice of a speaker. In the initial phase, that speaker would read sentences presented to him and his observed F_0 contour would be analyzed. Then, when the text of a new sentence is given to the system, an

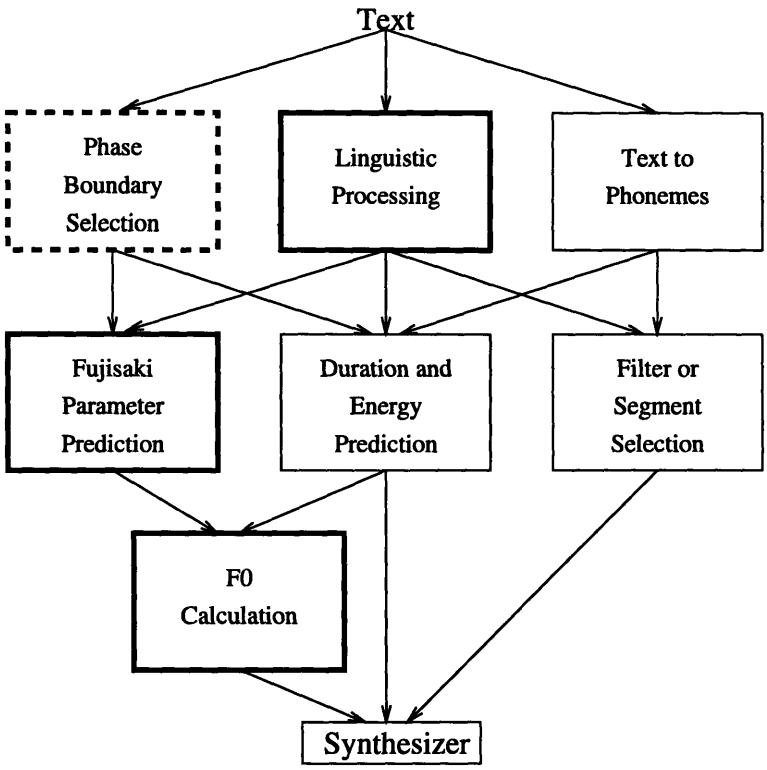

Figure 1-1: Overall structure of text-to-speech synthesis system

appropriate F_0 contour can be generated.

A block diagram of the overall text-to-speech system assumed for this research is presented in Figure 1-1. The modules with heavier lines are the ones which are implemented for this study. The dashed line around phrase boundary selection indicates that this is not accomplished automatically for this work, while the processes in the other boxes are. In the first layer, the text of a sentence to be synthesized is processed to generate more specific information about the text. The second layer involves taking that information and producing abstract information about the speech. Finally, the synthesizer combines the F_0 contour, the duration and energy, plus either filters that describe the spectral characteristics or prototypical segments of actual speech to create the synthetic speech.

1.2 Outline of Thesis

The organization of the remainder of this thesis will be as follows. Chapter 2 describes background information on previous research into the generation of F_0 contours. Research on several model types will be presented, including target and parametric methods for describing the contour. The linguistic, lexical, and other types of information that similar research has used to predict F_0 will be discussed. The final section in this chapter will describe some evaluation methods that other research has employed.

Chapter 3 describes the training phase of the system, in which prototypes are created for later use in synthesis. The linguistic and lexical information annotated to the database of text of speech is discussed in this chapter. Then, the method for extracting the parameters of the F_0 model that describe the observed F_0 contour of an utterance are described. Finally, there is a description of how prototypical parameters for the generation model are created to represent linguistic and lexical contexts.

Chapter 4 describes the synthesis phase of the system, in which F_0 contours are generated for new sentences. The selection of parameters to describe the F_0 contour for the new sentence is accomplished by a dynamic programming algorithm which is detailed in this chapter.

In Chapter 5, descriptions and results of both objective error measurements and subjective listening tests are presented. Some example F_0 contours are shown to further illustrate the way the system works. Finally, Chapter 6 concludes the thesis with a summary of the new ideas presented and some directions for further research.

Chapter 2

Background

This chapter will describe research that has already been conducted in the generation of fundamental frequency (F_0) contours. The first section will cover several types of models of F_0 generation and some of their applications. The second section will discuss information that has been used to predict the contours. Finally, methods that have been used to evaluate the success of F_0 generation systems will be presented.

2.1 Models of F_0 Generation

This section will introduce models of F_0 generation that have been published in the scientific literature. There are two main types of models, target and parametric. A specific parametric model described by Fujisaki will be discussed in some detail. Finally, some other models that do not fall into either category will be discussed.

2.1.1 Target Models

As proposed in [40], it is possible to abstract a F_0 contour into a series of high and low targets. This idea was recently formalized into the Tones and Break Indices (ToBI) system [53], which provides a standard way to annotate prosodic phenomena. This system contains two tiers, tonal and break indices. The first tier consists of pitch accents denoted by high or low markers combined with directional identifiers, which

attempt to encode word-level prosodic events. The second tier is a seven point scale that quantifies breaks between words, modeling the phrasal structure of the utterance. There have been several other proposed labeling schemes [31], all of which attempt to annotate prosody for computer databases of speech.

Given that some abstract prosodic labeling exists, the problem of generating F_0 from text can be broken up into two steps. The first step uses the text to produce the prosodic labels. For example, in [43], classification and regression trees were used to predict ToBI labels from linguistic information about the text, using a hand labeled database for training. In another example [42], a discourse-model was used to produce ToBI labels based on the context of the sentence and previous sentences in the conversation, attempting to provide distinctions in prosodic contrast between new and old words in the conversation.

Once the labels have been predicted, they can be used to produce the actual F_0 contour. In [24], quantitative prosodic labels of stressed syllables were used to form piece-wise linear F_0 contours. In [44], a dynamical system was described that takes ToBI labels and produces an F_0 contour, using a combination of an unobserved state vector and a noisy observation vector of F_0 and energy. In [6], ToBI labels were combined with stress and syllable positions in a linear manner to produce three F_0 values per syllable, which were smoothed to form a contour.

Some research has pursued automatic generation of such labels from both the text and the spoken utterance. This is a difficult task, because even with hand labeling, the level of agreement between labelers does not exceed 90% [41]. In one example [45], a stochastic model that used acoustical information and phrase boundary locations was used to predict ToBI labels from an utterance. With known phrase boundary locations on an independent test set, this system achieved 85% accuracy on syllable accent level in relation to the hand labeled database. In another example [56], acoustical information was used to predict pitch movements, another prosodic labeling scheme quantitatively describing rises and falls in the F_0 contour.

2.1.2 Parametric Models

Another form of abstraction of fundamental frequency (F_0) contours is through parameterization. In these methods, an explicit numerical model is used. The inputs to the model attempt to quantitatively describe prosodic phenomena, in a similar way to that in which targets qualitatively describe them. Therefore, any F_0 generation system using a parametric model contains the same two steps described for target models. Again, the first step consists of generating the parameters of the model from the text. However, the second step is straightforward, since these generated parameters directly produce an F_0 contour. One of the most commonly used parametric models is that described by Hirose and Fujisaki [21]. It uses the sum of two sets of inputs and filters to produce a F_0 contour. See Section 2.1.3 for a detailed description of Fujisaki's model. Another parametric model is the rise/fall/connection (RFC) model [54], in which F_0 is modeled as a sequence of quadratic and linear functions with varying amplitude and duration. For this model, after finding the best model parameters, an RMS error of between 4 and 7 Hz was achieved between the generated and actual F_0 contours. Another use of the RFC model has been as an analytical tool to locate pitch accents in spoken utterances [55].

2.1.3 Fujisaki's Model

Fujisaki's model is one of the most commonly used parametric models. This section contains a detailed mathematical and graphical description of the model and a survey of its previous uses for analysis and synthesis.

Description

The Fujisaki model proposes that the logarithm of fundamental frequency (F_0) can be modeled as the sum of the output of two filters and a constant. See Figure 2-1 for a block diagram of the model. Each of the filters has two poles at the same location. One filter takes a series of Dirac impulse functions, and its output models phrase level effects. Each impulse and its associated output will be referred to as a phrase accent.

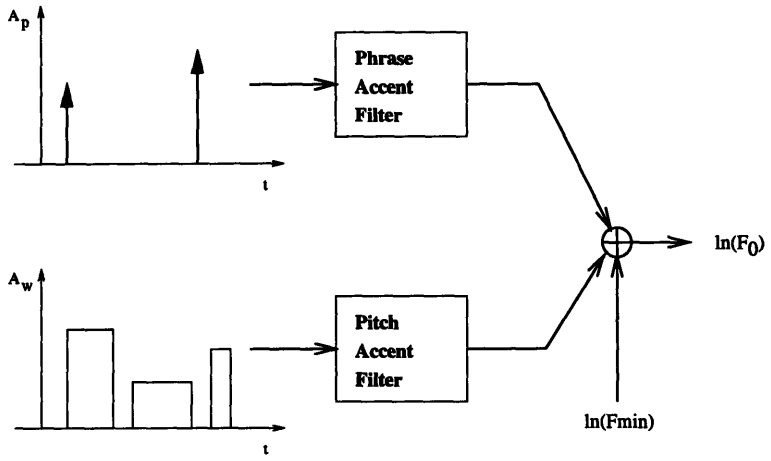

Figure 2-1: Block diagram of Fujisaki's model

The other filter uses step functions as input in order to model the word level effects on the F_0 contour. Each step function and its associated output will be referred to as a word or pitch accent. Note that this model creates a continuous F_0 contour while most text-to-speech systems only require F_0 at discrete times. The desired information can be obtained by sampling the generated contour at the appropriate times. The logarithm of F_0 is given by

$$\begin{aligned} \ln F_0(t) &= \ln F_{\min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - t_{pi}) \\ &\quad + \sum_{j=1}^{J} A_{w_j} \left\{ G_{w_j}(t - t_{wj}) - G_{w_j}(t - (t_{wj} + \tau_{wj})) \right\}, \end{aligned} \tag{2.1}$$

where

$$G_{p_i}(t) = \alpha_i^2 t e^{-\alpha_i t} \cdot u(t) \quad \text{and,} \tag{2.2}$$

$$G_{w_j}(t) = (1 - (1 + \beta_j t) e^{-\beta_j t}) \cdot u(t) \tag{2.3}$$

are the phrase level output and the word level output respectively. The variables in the above equations are described in Table 2.1.

Variable	Description
$F_{\min}$	Additive constant, minimum fundamental frequency
I, J	number of phrase and pitch accents, respectively
A_{p_i}, A_{w_j}	amplitude of the ith phrase accent and jth pitch accent
t_{pi}, t_{wj}	onset time of ith phrase and jth pitch accents
τ_{wj}	length of jth pitch accent
α_i, β_j	rate of decay of phrase and pitch accents, respectively
$u(t)$	step function, equal to zero for $t < 0$, equal to one otherwise

Table 2.1: Description of parameters used in Fujisaki's model

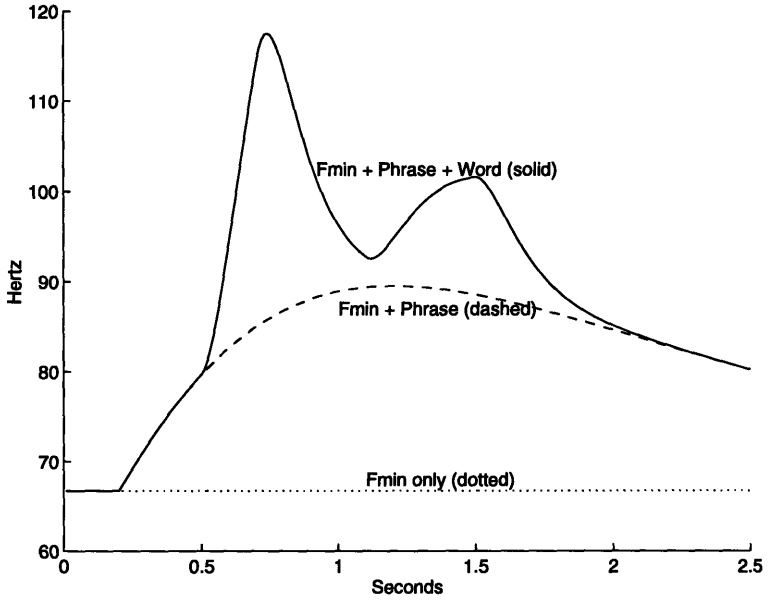

Figure 2-2: Example output from Fujisaki's model

Example Output

Figure 2-2 contains an example output from Fujisaki's model with one phrase accent and two pitch accents. This contour was created by specifying the baseline F_0 ($F_{\min} = 4.2$ ln(Hz)), the three parameters of the phrase accent ($t_p = 0.2$ sec, $A_p = 0.8$, and $\alpha = 1$ sec^{-1}), and the four parameters of each of the pitch accents ($t_{w1} = 0.5$ sec, $\tau_{w1} = 0.2$ sec, $A_{w1} = 0.5$, $\beta_1 = 10$ sec^{-1}, $t_{w2} = 1.1$ sec, $\tau_{w2} = 0.4$ sec, $A_{w2} = 0.15$, and $\beta_2 = 10$ sec^{-1}).

Previous Uses

Most of the early research on Fujisaki's model focused on proving that it was capable of modeling prosodic phenomena. This was shown by demonstrating an ability to find model parameters such that generated contours were adequately close to observed contours. This capability has been shown in many languages: Japanese [21], German [36, 37, 38], French [4], Chinese [11], Spanish [16], Swedish [30], and English [14].

More recently, research has focused on using the model to generate F_0 contours, but mostly in a rule based manner. In [37], the model parameters were set by rules based on linguistic features (e.g., sentence modality and word accent). In [38], the relationships between the same linguistic features as in the previous study and the model parameters were statistically analyzed, but generation was still rule based. In [36], decision trees based on syntactic information were used, but there was no comparison of predicted and actual F_0 contours for an independent test set. In [4], parameters were generated using rules based on a prosodic structure comprised of syntax and rhythm.

Fujisaki's model has been used analytically, for example to find phrase boundaries and examine the characteristics of the parameters in comparison to other features. The model has been used to detect phrase boundaries in Japanese, by finding the locations of the Fujisaki phrase accents [39]. In [20], Fujisaki's model was similarly used to predict phrase boundaries resulting in correct prediction two-thirds of the time. In [13], the parameters were compared with several linguistic features of a sentence, and correlations were found between the parameters and lexical stress, syntactic structure and discourse structure. More recently, in [18], the relationship between targets such as ToBI labels and parameters of the Fujisaki model was investigated. In [17], Fujisaki's model was used to analyze the difference between dialogue and reading styles of speech. It was found that the model parameters during dialogue exhibit more variation than when similar speech was read.

2.1.4 Other Models

Various other techniques have also been used to model fundamental frequency (F_0). Hidden Markov Models have been used to generate F_0 contours for isolated words [29]. Researchers have attempted to use neural networks to generate F_0. For example, in [47], three F_0 values were predicted in each phrase based on phrase characteristics, such as number and type of surrounding phrases. And in [10], neural networks were used to model the physical properties of the body (e.g., vocal folds and thyroarytenoid) that control F_0. In another example [7], six neural networks were used to produce durations, means, and shapes for generation of F_0 contours in Mandarin, at the syllable level. In [33], a combination of neural networks and decision trees was used to predict F_0 as two linear functions for every phoneme. In the $\log_2$(Hz) domain, the average absolute value error between the predicted and the realized F_0 was 0.203. In a more data driven approach [32], F_0 patterns for each syllable were extracted from a prosodic database based on an independence measure that was generated by rule from information such as grammatical categories and stress positions. In [48], prosodic information was predicted from a conceptual representation of a sentence. In [2], linear models were created for several levels, from sentences down to syllables. For synthesis, these models were superimposed, based on linguistic information taken from the new sentence. In [49], a F_0 contour was generated for a new sentence by extracting parameters that describe the F_0 contour of a similar sentence in a database, and modifying those parameters slightly.

2.2 Information Used for Prediction

The information that is used to predict the F_0 contour is just as important as choosing an appropriate model. Without the salient information, prediction will fail. It is usually considered that part-of-speech information is the most important in determining accent, although it can not completely determine the accent level [22]. In [46], mutual information between ToBI labels and many linguistic features was measured. It was shown that lexical stress, part-of-speech, and word class (e.g., content, function, or

proper noun) contain the most information. All of this information has been shown to be relatively easy to produce automatically from text. Some research [43] has used discourse models to find out which words were new to the conversation, indicating that these words should be stressed. However, in [22], it was shown that intonational prominence can be modeled well without "detailed syntactic, semantic, and discourse-level information." Intonational phrase boundaries were also important in determining the F_0 contour. Some research has attempted to automatically predict phrase boundaries using a stochastic parser [52], decision trees [58], part-of-speech trigrams [50], and even using F_0 information [39]. Furthermore, in [25], descriptive phrase accents were automatically generated using a very large training corpus of text with associated speech. However, none of these methods has proven to be completely successful.

2.3 Evaluation Techniques

Comparisons between the predicted F_0 contour and the actual contour of a sentence in the test set can be made both objectively and subjectively. The ultimate test of the success of an F_0 synthesis algorithm is how acceptable the synthesized contour sounds to a human listener. For such subjective tests, the test set could contain only new text, although in practice, it may be useful to assess the predicted F_0 contour in relation to the a human reading. However, objective tests are useful to measure progress throughout the course of a project and are necessary for comparing with published results. Common objective tests that have been used include RMS error or average absolute value error in Hz or the logarithm of Hz [6, 33, 44, 54]. The latter is preferred because the perceptual dynamic range of F_0 is expressed well in the log domain. When doing objective tests, there are two facts that one must consider; the smallest difference in Hz that can be perceived is approximately 1 Hz and that F_0 perception is relative [27]. Thus, some small errors or a general shift in the F_0 contour might be imperceptible. A possible subjective test [23, 37] is to synthesize the same utterance with only different F_0 contours, and then have subjects rate them

on a scale of one to five. After doing this for many sentences, a mean opinion score can be calculated from the ratings of all the subjects to compare the competing F_0 generation algorithms.

Some other less straightforward evaluation techniques have been proposed. In [35], the measure "resemblance to human speech" (RHS) was proposed to test the acceptability of manipulations of the prosody of human speech. Another way of measuring F_0 manipulations is to ask a listener to estimate an utterance's liveliness, with more lively speech preferred [57]. In [5], several criteria commonly used to assess synthetic speech were presented, including intelligibility, quality, and cognitive load.

Chapter 3

Training

This chapter describes the training phase of the system. During this phase, a model of fundamental frequency (F_0) generation is chosen to analyze a training database of text and corresponding spoken utterances. Information is added to the database about its lexical, temporal and linguistic structure, before the salient parameters of the F_0 model are extracted in an intelligent manner. Finally, models are created to capture both the word-level and the phrase-level prosodic events in a sentence.

Figure 3-1 shows a block diagram of the training process. The four boxes in the first layer of the figure represent the database annotation of Section 3.2, with the dashed line for phrase boundary selection indicating that it is the only module not implemented automatically. The second layer contains the automatic Fujisaki model parameter extraction described in Section 3.3. The final layer is the building of the pitch and phrase models described respectively in Sections 3.4 and 3.5.

3.1 Basic Tools

This section introduces some of the basic tools that are used in this research. A review of the Fujisaki F_0 generation model is presented followed by details of the databases that will be used for both training and testing.

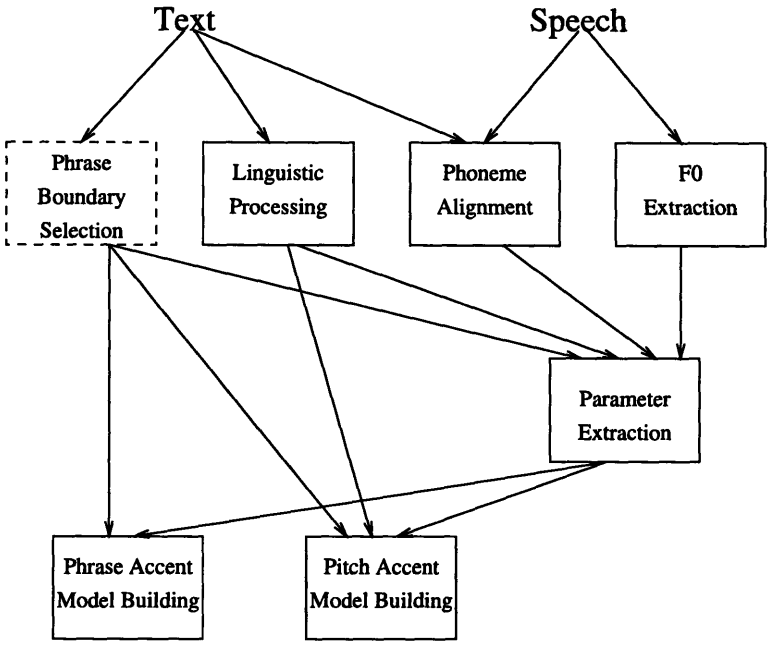

Figure 3-1: Block diagram of training process

3.1.1 F_0 Generation Model

Fujisaki's model, as described in Section 2.1.3, is one of the most commonly used F_0 generation algorithms. Due to its ability to model prosodic phenomena, along with its flexibility, this model is ideal for use in this study. Fujisaki's model has almost all of the functionality of the target models discussed in Section 2.1.1. With the phrase accents providing a declining baseline, the pitch accents can be inserted with appropriate amplitudes to mimic both high and low targets. The use of Fujisaki's model allows the relationship between linguistic phenomena and the F_0 contour to be determined automatically and statistically. Using a labeling system like ToBI (see Section 2.1.1) to express that relationship would require the use of a manually specified standard and the training database would have to be hand labeled.

3.1.2 Databases: Training and Testing

The database that will be used to evaluate the system is taken from the DARPA 1995 Hub 4 radio database, taken from the Marketplace program. Specifically, it consists

of the subset of those utterances that are all spoken by the anchorperson, David Brancaccio, without any music in the background. Mr. Brancaccio is a veteran newscaster with over 10 years of experience in broadcasting [34]. The number of sentences spoken by Mr. Brancaccio is approximately 400, each about 20 words long, totaling about one hour of speech. There are many advantages to using this database. The speaker is mostly reading from a script, but he is doing so in a prosodically interesting way. In fact, radio broadcasters attempt to convey as much information as possible through their intonation [12], which makes their F_0 contours more meaningful. There are very few disfluencies that make analysis difficult, while there is useful information to be gained by looking at the fundamental frequency. Several recent studies have also used radio corpora [12, 22, 43]. The amount of data from a single speaker is representative of that used in similar recent research [32, 33, 44, 58], allowing meaningful analysis to take place. Due to the declination effect from the phrase accents, Fujisaki's model is usually applied to declarative sentences [16]. Most of the sentences in the database are declarative, but they are somewhat longer than those usually analyzed with Fujisaki's model. Nevertheless, this research attempts to apply Fujisaki's model to these prosodically rich sentences. Sentences are analyzed that have been spoken by four other announcers, two male and two female. There is not enough data to train the models for each new speaker. These sentences are used as test sets to measure the F_0 generation model's ability to capture information about the F_0 contours used in the domain of business radio news. If these tests give poor results, then a model that was trained on a particular speaker represents only that speaker.

Another database, taken from the Wall Street Journal database, is used for comparison. Approximately 1150 sentences from one male speaker, each about 15 words long, are used for training and testing. These sentences should be less prosodically interesting than the ones from the Marketplace database, because the speaker is not a professional radio announcer, but rather a volunteer. As a result, the F_0 contours of these sentences should be easier to analyze.

3.2 Database Annotation

The database must be annotated with prosodic, linguistic and lexical information before the F_0 generation model can be trained. The extraction and stylization of the fundamental frequency is described in Section 3.3.1. Other information to be obtained includes a phone-level alignment, part-of-speech information, lexical stress, and phrase boundary locations. This section presents the methods by which this information is obtained and describes where it is used.

3.2.1 Phone-level Alignment

The phone-level alignment is provided by supervised recognition, using the development system from the speech group at IBM's T. J. Watson Research Center. The recognizer is given both the words and the speech of the sentence. The information returned is the list of phones that occur in the utterance, together with start and end times for each phone. In Section 3.3.3, the location of the vowel phones will be used to determine the location of pitch accents. In addition, regions of silence are marked where there is no speech present. This information is used in locating phrase boundaries (see Section 3.2.4) and as part of the context in the decision tree questions (see Section 3.4.3).

3.2.2 Part-of-Speech Information

The database is annotated with part-of-speech information for each word in a sentence, so that questions can be asked about the linguistic context when the decision tree is built for pitch accents (see Section 3.4.3). The 36 part-of-speech categories, as defined by the Penn Treebank Project [51], are divided into 14 subsets plus a category for silence. The subsets were created such that similar part-of-speech tags are grouped together, and so that each subset will have a significant number of words in the training data associated with it. See Table 3.1 for a complete list of all categories. These part-of-speech tags are obtained using a stochastic parser that was developed at IBM Research by Adwait Ratnaparkhi.

Number	Name	Tags Used
1	Noun	NN, NNS
2	Proper Noun	NNP, NNPS
3	Base Form Verb	MD, VB
4	Participle	VBD, VBG, VBN
5	Present Tense Verb	VBP, VBZ
6	Adjective	JJ, JJR, JJS
7	Adverb	RB, RBR, RBS, WRB
8	Determiner	DT, WDT
9	Pronoun	PRP, PRP$, WP, WP$
10	Preposition	IN, RP
11	To	TO
12	Conjunction	CC
13	Number	CD
14	Other	EX, FW, LS, PDT, POS, SYM, UH
15	Silence	

Table 3.1: Part-of-speech categories. The acronyms in the right hand column refer to labels used in the Penn Treebank project [51].

3.2.3 Lexical Stress

Each syllable is annotated with a lexical stress of primary, secondary, or unstressed, which is used as a question in building the pitch accent decision tree (see Section 3.4.3). This information is obtained by looking up each word in the COMLEX English Pronouncing Dictionary [28], assuming that the first pronunciation is correct. If a word is not in the (approximately 90,000 word) dictionary, then a stress pattern of primary followed by a sequence of unstressed, secondary stress pairs is assumed. A word was not found in the dictionary approximately 1.5% of the time in the Marketplace database. The dictionary was designed from several databases, including the Wall Street Journal database, and therefore all words in that database were defined in the dictionary. The same dictionary provides information about classes of words. The function word designation, as defined by COMLEX, is used as a question for decision tree building. A function word is a word that does not provide any semantic meaning, such as "the," "am," "anyhow," and "but".

3.2.4 Phrase Boundaries

Phrase boundaries are hand labeled based on the lexical transcription of a supervised recognizer (see Section 3.2.1). These phrase boundary locations are used both in the pitch accent model, as questions in building the decision tree (see Section 3.4.3), and in the phrase accent model, as the basis for the location of the accents (see Section 3.5). This is the only stage in database annotation that is not done automatically. This task is reported to have been accomplished automatically in recent literature [39, 52, 50, 58], but none of these results are duplicated here.

Three types of phrase accents have been designated so that their behavior can be modeled separately. A type 1 accent is at the beginning of a sentence, and usually only occurs at the start of the utterance. Types 2 and 3 occur during the middle of the sentence at boundaries between clauses or phrases. Type 3 accents are associated with a silence, because the speaker usually pauses before beginning a new phrase. Type 2 accents are not associated with a silence, but are placed in locations where phrase boundaries might occur. Note that the only information taken from the spoken utterance is the silence locations, otherwise the phrase boundaries are determined solely by the structure of the text. Heuristics used in assigning phrase boundaries include that they should be evenly spaced, if possible, and should occur approximately every eight to ten words.

3.3 Parameter Extraction

This section describes how, given the speech and the text of an utterance, parameters for the Fujisaki model are extracted that best describe that utterance. Initially the F_0 contour is extracted, and stylized to remove any outliers (see Section 3.3.1). In Section 3.3.2, a search technique based on the structure of Fujisaki's model is employed. During the search, the parameters are constrained both to associate them with relevant linguistic events and to make the search more efficient (see Section 3.3.3). Finally in Section 3.3.4, some results are presented using the techniques described here.

3.3.1 F_0 Extraction and Stylization

The fundamental frequency (F_0) for each ten millisecond frame of the original utterance is extracted using an autocorrelation method. This method convolves the speech waveform with itself and finds the biggest peak away from the origin. A frame could also be found to contain silence or unvoiced speech, neither of which has a valid F_0 value. In such a case the frame is assigned a F_0 value of 0 Hz. Since this computed F_0 contour contains a few outliers, a median filter technique is used to remove them. The median is taken of the five F_0 values taken from the current frame and the two previous and two next frames. If the current frame has been assigned a "positive" F_0 value and the difference between that value and the median is greater than 15 Hz, then the computed F_0 value is changed to the median. See Section 3.3.4 for some results on how often F_0 values were changed in this manner. Those frames that have F_0 values of 0 Hz are assigned positive values by linear interpolation. The new F_0 value for one of these frames, f_c, at a particular time t_c is given by

$$f_c = ((t_n - t_c) \cdot f_p + (t_c - t_p) \cdot f_n)/(t_n - t_p) \tag{3.1}$$

where f_p and t_p are the closest previous F_0 value and time and f_n and t_n are the closest next F_0 value and time. The final step in F_0 stylization is a low pass filter. This stage further smooths the contour and reduces the number of extrema that are used later to find the best phrase accents (see Section 3.3.2).

3.3.2 Search Technique

The objective is to find Fujisaki model parameters that generate a F_0 contour that is as close as possible to the observed (stylized) contour. Although Fujisaki's model has been used extensively, there are very few specific descriptions of an algorithm for performing this task. One suggested method [19] uses a left-to-right search to determine each successive phrase and pitch accent. Even without descriptions, most researchers agree that the parameters should be chosen such that the mean squared distance between the observed and the generated contours should be minimized. In

this study, this error is only computed at those frames that were originally assigned a valid F_0 value. There is no closed-form solution to this problem, thus the search for these parameters is performed iteratively, using a gradient descent algorithm. This algorithm requires starting locations for all of the parameters and partial derivatives of the error with respect to each parameter.

Searching for all of the parameters simultaneously proved to be cumbersome, and therefore the search is divided into two parts: finding the best phrase accents and finding the best pitch accents. The phrase accents are intended to describe the baseline of the F_0 contour. Therefore, the error to be minimized is not computed at every frame with a valid F_0 value. Instead, the minima of the stylized contour are found, and the phrase accents are chosen such that the error between the generated and observed contour at those points is minimized. Using the optimal phrase accents that have been computed, the pitch accents are found that minimize the error between the generated and observed F_0 contours. For this stage, the error is calculated using all of the frames with valid F_0 values, since the phrase and pitch accents comprise the entire generated contour.

3.3.3 Parameter Constraints

To perform the search described in Section 3.3.2, a specific number of phrase and pitch accents must be decided upon. This number has to be limited, otherwise any optimization algorithm will overfit the realized F_0 contour and the extracted parameters will have less significance. Each accent should be related to some linguistic phenomenon so that deriving new accents from new text will be possible. A similar idea of parameter extraction based on linguistic context was introduced in [15].

To achieve both limitation and linking of accents, the times of the phrase accents (t_{p_i}) and the end times of the pitch accents ($t_{w_j} + \tau_{w_j}$) will be determined by information present in the database. The times of the phrase accents are determined by the phrase boundaries, marked by hand as described in Section 3.2.4. The end times of pitch accents are placed in the vowel phones of stressed syllables, as determined by the supervised alignment described in Section 3.2.1. This reduction in freedom of

	Marketplace (MP)	Wall Street Journal (WSJ)
Number of cells	17	18
Total frames per cell	790.2	637.2
Voiced frames per cell	579.7	323.7
Number of replacements	10.8	5.3
Number of minima	22.2	19.9
Number of phrase accents	2.8	2.1
Number of pitch accents	26.8	16.2

Table 3.2: General information about the database subsets used for comparing search techniques (average values)

the search space does not significantly increase the error when the optimal parameters are found (see Section 3.3.4). With these limitations, each phrase accent has one parameter (amplitude) and each pitch accent has two parameters (length and amplitude) which need to be varied to find the optimal accent parameters.

Another constraint that will be used is that the global parameters, α, β, and $F_{\min}$, will remain constant after finding suitable values. It has been shown that these global parameters are relatively constant for a particular speaker [16, 21, 38, 39]. The global parameters are also chosen in an iterative manner. A representative sample of sentences are picked and the search for the optimal accent parameters is carried out. This procedure is repeated, changing only the global parameter values, until a minimum is found. The decay rate for the phrase accents (α) and the baseline F_0 value ($F_{\min}$) are chosen such that the error between the generated and observed F_0 contours at the minima of the stylized contour is smallest. Then, the decay rate of the pitch accents (β) is varied while searching for the optimal pitch accent parameters. The β which corresponds to the the smallest error between original and generated F_0 contours is chosen.

3.3.4 Searching Results

To compare the results of various searching techniques, twenty sentences were chosen at random from each of the two databases to be analyzed. Information about these subsets is provided in Table 3.2. Of those sentences, two from the Wall Street

Journal (WSJ) comparison database and three from the Marketplace (MP) comparison database had to be discarded. For these sentences, the first searching condition on Table 3.3 reached a local minimum, and the error was approximately five times greater than than for the other searching conditions. The total frames in the cell represents the number of ten millisecond frames from beginning to end of the utterance, including initial and final silences. The voiced frames per cell represents the number of frames that were judged to have a valid F_0 value, based on silence and voicing tests. The ratio of voiced frames to total frames is higher for the MP data because there are shorter pauses between words and because in the WSJ data there is a tendency to trail off at the end of words, causing silence to be marked incorrectly. The number of replacements is the number of times a F_0 value was replaced by the median of the F_0 values of its surrounding frames in F_0 stylization (see Section 3.3.1). Note that this correction was only invoked for less than 2% of the voiced frames, indicating that the output of the original F_0 extraction algorithm is fairly consistent. The number of minima is the number of points found in the stylized F_0 contour that will be used in the first part of the two-part searching style (see Section 3.3.2). The number of phrase and pitch accents is the average number of each in the sentences used for comparison. Note that the ratio of pitch accents to phrase accents is greater for the MP data. The number of pitch accents is determined automatically, but the number of phrase accents is determined by hand. The different ratios might reflect an inconsistency in the hand-labeling of the two databases.

Table 3.3 presents results from several search conditions. In all conditions, the same number of phrase and pitch accents are used. The numbers in the table represent the average of the mean squared error between the stylized and generated F_0 contours of all the sentences in the database subsets. The units are ln(Hz) per frame, with the first four results calculated over all frames with valid F_0, and the last two results calculated only for the minima of the stylized contour. The searching conditions refer to the techniques and parameter constraints discussed in Sections 3.3.2 and 3.3.3. Search style is either a simultaneous search where all parameters are found at the same time, or a two-part search where the phrase and pitch accent parameters are

	Database		Searching Conditions			
Result	MP	WSJ	Search Style	α and β	Accent loc.	Stage
1	0.0774	0.0395	Simultaneous	Variable	Variable	Final
2	0.0733	0.0392	Simultaneous	Fixed	Variable	Final
3	0.0579	0.0422	Two-part	Fixed	Variable	Final
4	0.0903	0.0614	Two-part	Fixed	Fixed	Final
5	0.1468	0.0771	Two-part	Fixed	Variable	Intermediate
6	0.1861	0.0769	Two-part	Fixed	Fixed	Intermediate

Table 3.3: Results from parameter extraction for various search conditions, RMS error in ln(Hz) per frame between extracted and observed F_0 contours

found in separate stages. The global parameters α and β can be variable so that the search algorithm can find the optimal values for each sentence or fixed to some globally optimal values for all sentences. Note that the global parameter $F_{\min}$ is fixed for all search conditions. The locations of the phrase and pitch accents can be either variable or fixed as discussed in Section 3.3.3. Finally, the first four results (final stage) contain errors computed for the entire generated contour while the last two results (intermediate stage) are errors just for the phrase accent component of the two-part search.

For both of the databases, the results indicate that restricting the search space does not lead to a serious disadvantage for finding parameters of the Fujisaki model that adequately describe the observed F_0 contour. When α and β are allowed to vary, the result is just as good as when they are fixed to one optimal value for all sentences. These variables have the largest gradient with respect to the error. Therefore, the gradient descent algorithm effectively optimizes α and β first, and sometimes these local minima do not allow the other parameters to reduce the error further. The continued restriction of the search space by using a two-part search has surprising results. For the WSJ database, results 2 and 3 are not very different. However for the MP database, the two-part search performs significantly better than the simultaneous search, with all other conditions equal. These results indicate that both breaking up the search and finding the phrase accents based on minima are good ideas. The biggest

loss of performance occurs when the accent locations are fixed to values specified by phrase boundary and lexically stressed syllable locations (result 4). However, the small degradation does not outweigh the advantage of having each accent directly linked to a word or phrase. Results 5 and 6 behave differently for each database. For WSJ, it shows that fixing the locations of the phrase accents has negligible effect on the error between the generated intermediate contour and the minima of the stylized contour. However for MP, the trend observed in results 3 and 4 that fixing accent locations is detrimental to performance is observed again. Generally, the error for the MP database is higher than for the WSJ database. This is due to the higher variability of the F_0 contour for trained radio commentators than for average people reading hundreds of sentences into a recorder.

3.4 Pitch Accent Model

This section describes the model of the word-level pitch accents. In order to enable an accent to be specified in a particular context at synthesis time, the context/pitch-accent pairs are clustered using binary decision trees, [3]. During synthesis (see Chapter 4), the new contexts are dropped down the trees to determine from which leaf the accents should be generated. The trees enable both seen and unseen contexts to be handled during synthesis. Such a decision tree is built using questions and data designed specifically for this task. Although the leaves are created so that similar data are clustered together, there still is a degree of variance within each leaf. Therefore, several prototypes are created at the leaves of this tree, so that during synthesis a typical accent will be used for each leaf. Finally, probabilities are calculated for use in choosing the best prototypes to use during synthesis.

3.4.1 Decision Tree Description

A binary decision tree is built in order to cluster similar pitch accents together, by asking binary questions about context at each node in the tree. For this decision tree, similarity is measured by assuming a Gaussian distribution for the data in each node,

and maximizing the log likelihood of the data. A further assumption made is that each element (i.e., length or amplitude of a pitch accent) of the data is independent of all other elements.

The following procedure is followed at each node in the tree which has not yet been split into two nodes. This current active parent node is divided into two children nodes, henceforth known as left and right children. The division is accomplished using a binary yes-or-no question that is asked about the context of each data point. Therefore, each data point in the parent node goes into exactly one of the children nodes. For example, X_p, X_r, and X_l are the sets of data at the parent, right child and left child nodes respectively. Then,

$$X_r \cup X_l = X_p, \text{ and} \tag{3.2}$$

$$X_r \cap X_l = \emptyset. \tag{3.3}$$

At each of these three nodes, the mean and variance of the data in that node is calculated. For each set X_s, the mean is $\boldsymbol{\mu}_s$ and the standard deviation is $\boldsymbol{\sigma}_s$, and for each dimension d the mean and standard deviation are μ_s^d and σ_s^d. Due to the assumption of independence of each dimension, the likelihood of element x_i in set X_s is given by a product of one-dimensional Gaussian probability density functions,

$$p_s(x_i) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_s^d} e^{\frac{(x_i^d - \mu_s^d)^2}{2(\sigma_s^d)^2}}, \tag{3.4}$$

where D is the total number of dimensions (the number of parameters associated with each accent). The total log likelihood of set X_s is given by

$$L_s = \sum_{x_i \in X_s} \ln(p_s(x_i)). \tag{3.5}$$

From all of the possible questions, the one that is used to split the data in that node must meet the following two criteria. The first is that the size of both of the children nodes, $|X_r|$ and $|X_l|$, must be greater than a set threshold. For all of the questions

Context		Data
Previous	Current	Intensity
Noun	Verb	1
Verb	Noun	3
Noun	Noun	2
Verb	Verb	4
Noun	Verb	2
Total Log Likelihood = -7.25		

	Left		Right		Total
Question	Data	Likelihood	Data	Likelihood	Likelihood
Previous = Noun	1, 2, 2	-2.11	3, 4	-1.64	-3.75
Current = Noun	2, 3	-1.64	1, 2, 4	-5.03	-6.67

Table 3.4: Simple decision tree example

that meet the first criteria, the one that maximizes the gain in total log likelihood, $L_r + L_l - L_p$, is chosen. A node does not produce any children if no questions satisfy the first criteria or if the gain in log likelihood does not exceed another set threshold. A node with no children is referred to as a leaf. Trees of various sizes can be grown by changing the two thresholds. See Appendix B for some example decision trees grown from actual data.

An example is illustrated in Table 3.4. In this contrived example, the current node contains five data points, the context consists of the previous and current part-of-speech, and the data is one-dimensional. Of the two questions that can be asked, clearly the first one divides the data into sets with more similar elements. This intuitive sense of similarity is reflected in the larger gain in log likelihood for the first question,

$$(-3.75 - (-7.25) = 3.5) > (-6.67 - (-7.25) = 0.58). \tag{3.6}$$

3.4.2 Decision Tree Data

As discussed in Section 3.3.3, the end time of a pitch accent is determined before the search for the optimal parameters takes place. With this information known, a pitch

accent can be completely described with just two values, its length and amplitude (τ_w and A_w). Another value that has proved useful in describing an accent is its intensity, which is the length multiplied by the amplitude. The intensity has been found to be a better quantitative measure of its effect on F_0 than either of the other parameters. Additionally, the intensity of a pitch accent is equal to the integral of the output of the pitch accent filter associated with that pitch accent.

Without intensity, any creation of prototypes through averaging will fail. For example, consider two accents, one with length 2 and amplitude 0.5, and the other with length 0.5 and amplitude 2. Both have intensity equal to 1, but a prototype accent with each dimension averaged separately will have length and amplitude equal to 1.25. This prototype's intensity is more than 50% greater than the original accents' intensities. Another concept introduced and used here is that of relative length. Instead of the absolute time between the start and the end of the accent, the relative length is this absolute time divided by the length of time between the end of the previous accent and the end of the current accent. This concept reduces variability and allows for similar accents at different speaking rates to have similar representations.

To exploit the regularity in intensity variation between accents, quantized intensity level are introduced. Once all of the accents in the training set are found, $M+1$ bin markers, $b_0, \ldots, b_M$, are found such that the number of accents from the training set in each bin is the same. A bin B_m contains those data points whose intensities are greater than $b_{(m-1)}$ and less than or equal to b_m. These bins are created to allow information about the variation of intensity to be quantified. An example of the predictability of intensity variation is taken from the marketplace training set with 14 bins plus one for end and beginning of sentence. Figure 3-2 presents this example, with a visual description of the intensity variation when transiting from one accent to another. Note particularly that accents with the lowest and highest intensities are very likely to be followed by accents with the highest and the lowest intensity levels, respectively. While the accents in the middle intensity levels tend to be more evenly distributed, there is still a general trend of greater probability along the reverse diagonal.

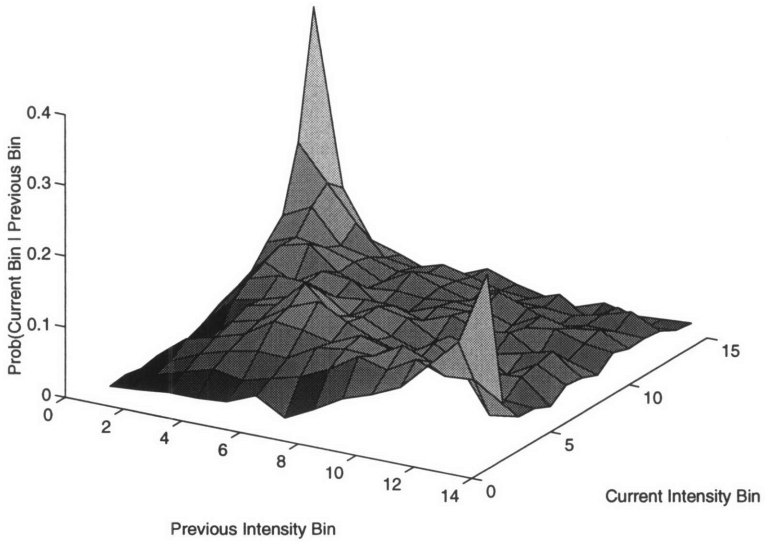

Figure 3-2: Graphical intensity bin variation probabilities

3.4.3 Decision Tree Questions

The questions that are used to build the tree essentially define the information that will be used to predict the level of the accents for new sentences. The information that other research has used for this task is described in Section 2.2. The questions that are used to grow the decision tree in this study are about the linguistic and lexical context of the text in which the accents originally occurred. There are five types of questions that can be asked. The first type is the part-of-speech category of the current word or the preceding and following three words. This information is obtained as described in Section 3.2.2. The second type of question asks whether or not the current word is a function word. The third type of question asks about the level of lexical stress, either primary or secondary, of the current syllable. Information used by these two questions is obtained as in Section 3.2.3. The fourth type of question asks how far the current accent is away from the previous phrase boundary. Phrase boundaries are located as in Section 3.2.4. The final type of question is about the intensity bin of the previous accent in the training data. The intensities are quantized into fourteen levels, as in Section 3.4.2, plus one for the beginning of

the sentence. At every node, all potential questions are asked, but only the one that produces the largest gain in log likelihood is used to split the node. The process of choosing the best question is described in Section 3.4.1. The average values of the intensity levels for both training databases in relation to the information in the first four types of questions is presented in Appendix A.1. The analysis presented there demonstrates that intensity does vary systematically with context, which establishes context clustering as a reasonable model building strategy. Additionally, Appendix B contains example decision trees with their associated questions that have been generated from actual data. These examples demonstrate that questions of all types are asked, especially those about previous intensity level, surrounding part-of-speech tags, and the function word designation.

3.4.4 Prototypes

The L leaves of the decision tree, $v_1, v_2, \ldots, v_L$, divide the data points into sets for each leaf, $Y_1, Y_2, \ldots, Y_L$, where Y_l, is the set of data points that match the context defined by the leaf v_l. For a particular leaf, v_l, at most J prototypes are created using a combination of k-means clustering and an iterative probabilistic clustering procedure. The procedures are carried out for each leaf individually.

K-means clustering is used to partition the N data points in the set Y_l of the leaf v_l into J regions. This process consists of arbitrarily choosing starting prototype locations, then iteratively placing each data point with the nearest prototype and recalculating the prototypes as averages of the data points associated with it. The final D-dimensional k-means prototypes, $\rho_1, \rho_2, \ldots, \rho_J$, are used as seeds to an iterative probabilistic procedure. These seed prototypes are further refined to produce the prototypes, $r_{1l}, r_{2l}, \ldots, r_{Jl}$, necessary for synthesis of the F_0 contour. This iterative probabilistic procedure consists of several steps, which are described below. The first step is similar to part of the k-means clustering process. Each data point (a vector),

$x_1, x_2, \ldots, x_N$, is placed into a set $X_{j'}$, where

$$j' = \underset{j}{\operatorname{argmin}} \ \mathcal{D}(x_n, \rho_j), \tag{3.7}$$

$$\mathcal{D}(x_n, \rho_j) = \sum_{d=1}^{D} \left(\frac{x_n^d - \rho_j^d}{\sigma_0^d} \right)^2, \tag{3.8}$$

and σ_0^d is the standard deviation of the dth dimension of the data in Y_l. The initial prototype is the mean and standard deviation of the data in each set, $r_j^{(0)} = \{\mu_j, \sigma_j\}$. (Parenthetical superscript refers to the iteration.) For each iteration $m \in \{1, \ldots, M\}$ and for each prototype $j \in \{1, \ldots, J\}$, the following sums are calculated,

$$s_0 = \sum_{n=0}^{N} p_j(x_n), \tag{3.9}$$

$$s_1 = \sum_{n=0}^{N} p_j(x_n) x_n, \text{ and} \tag{3.10}$$

$$s_2 = \sum_{n=0}^{N} p_j(x_n) (x_n)^2, \tag{3.11}$$

where the likelihood $p_j(x_n)$ is the Gaussian probability density function defined in Equation 3.4 with mean and variance defined by $r_j^{(m-1)}$ and $(x_n)^2$ in Equation 3.11 is a vector with each element of x_n squared. These values are used to calculate the prototype for the next iteration as follows:

$$\mu_j^{(m)} = (s_1/s_0), \tag{3.12}$$

$$\sigma_j^{(m)} = \sqrt{\frac{s_2}{s_0} - \left(\frac{s_1}{s_0} \right)^2}, \text{ and} \tag{3.13}$$

$$r_j^{(m)} = \left\{ \mu_j^{(m)}, \sigma_j^{(m)} \right\}. \tag{3.14}$$

After this process is complete the prototype calculated during the final iteration, $r_j^{(M)}$ is referred to as r_{jl}, the jth prototype of the leaf v_l.

3.4.5 Probabilistic Model

The three sets of probabilities that will be needed for the decoding stage of synthesis (see Section 4.1) are calculated at this stage. The first is the probability of the current leaf given the previous leaf and prototype

$$\delta_{jl_1l_2} = \Pr[v^{(i)} = v_{l_2} \mid r^{(i-1)} = r_{jl_1}], \tag{3.15}$$

where $v^{(i)}$ and $r^{(i)}$ are the leaf and prototype associated with the ith pitch accent of the sentence. (Parenthetical superscripts in this section refer to the number of the accent in the sentence.) The second is the probability of the current prototype given the current leaf

$$\lambda_{jl} = \Pr[r^{(i)} = r_{jl} \mid v^{(i)} = v_l]. \tag{3.16}$$

The final one is the probability of the current intensity level given the previous intensity level

$$\eta_{jk} = \Pr[x^{(i)} \in B_k \mid x^{(i-1)} \in B_j]. \tag{3.17}$$

The first probability, $\delta_{jl_1l_2}$, is calculated by dividing the sum of the probabilities of all the points in the current leaf (v_{l_1}) being in the current prototype (r_{jl_1}) and whose next leaf is v_{l_2} by the sum of the probabilities of all the points in the current leaf being in the current prototype. To calculate $\delta_{jl_1l_2}$, an intermediate probability, θ_{jn}, and a function, $\mathcal{N}_l(x_n)$ are used. θ_{jn} describes the probability of the prototype r_{jl} given the data point x_n. The Gaussian likelihoods used to calculate θ_{jn} use the means and variances of the prototypes in the current leaf. Mathematically, they are

$$\theta_{jn} = \Pr(r_{jl} \mid x_n) = \frac{p_j(x_n)}{\sum_{j'=1}^{J} p_{j'}(x_n)}, \text{ and} \tag{3.18}$$

$$\mathcal{N}_l(x_n) = \begin{cases} 1 & \text{if } v_l \text{ is the next leaf after } x_n \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

Using these variables,

$$\delta_{jl_1l_2} = \frac{\sum_{x_n \in Y_{l_1}} \theta_{jn} \mathcal{N}_{l_2}(x_n)}{\sum_{x_n \in Y_{l_1}} \theta_{jn}}, \tag{3.20}$$

where Y_l is the set of all data points in the leaf v_l.

The second probability, λ_{jl}, has been defined in two ways, and surprisingly the simpler of the definitions has given better results when used in prediction. See Section 5.1.3 for some results using the two definitions. The first method,

$$\lambda_{jl} = \frac{|X_j|}{|Y_l|}, \tag{3.21}$$

is based only on the number of data points closest to each prototype. X_j is determined by Equation 3.7 and is a subset of Y_l. The second method,

$$\lambda_{jl} = \frac{\sum_{x_n \in Y_l} \theta_{jn}}{\sum_{j'=1}^{J} \sum_{x_n \in Y_l} \theta_{j'n}}, \tag{3.22}$$

is a probabilistic method similar to that used to calculate $\delta_{jl_1l_2}$ in Equation 3.20.

The final probability, η_{jk}, is based solely on the number of times the observed behavior occurred in the training data. Note that this calculation takes place over all of the leaves, unlike that for the first two values, which are for particular leaves. B_m is the mth intensity bin, as discussed in Section 3.4.2. The conditional probability is expressed as

$$\eta_{jk} = \frac{|B_k \cap B_j^{-1}|}{|B_j^{-1}|}, \tag{3.23}$$

where B_j^{-1} is the set of all accents whose previous accent has intensity between b_{j-1} and b_j. This value represents the probability that the current prototype has a particular intensity, given the previous prototype's intensity. These probabilities will be used in the intensity variation model described in Section 4.1.2.

3.5 Phrase Accent Model

This section describes the phrase accent model that is used in this study. A phrase accent is described by its time and amplitude. The time is determined by the location of the phrase boundary it is associated with (as determined in Section 3.2.4). It is therefore only necessary to model the amplitudes of the accents. A model will

be constructed that is a linear combination of the information about the temporal context in which the accent was produced. A decision tree model like that used for the pitch accent model will not be used because of lack of data. Instead, a separate linear model is constructed for each of the three types of phrase accents (described in Section 3.2.4).

3.5.1 Prediction Information

The only information that will be used for predicting phrase accents is based on the location of the accent and its neighboring accents. Specifically, the information to be used is the number of the current phrase accent (i.e., the first/second/third accent in the utterance), the time until the previous and next phrase accents, and the number of words until the previous and next accent. If there is no next accent, then the time and number of words until the next accent are measured until the end of the utterance. For type 1 accents, the measures dealing with the previous accent are not used. See Appendix A.2 for details of the extracted phrase accents in relation to some of the prediction information. The relationships there demonstrate a fair correlation between the information and the phrase accent data.

3.5.2 Calculation of Coefficients

The problem of calculating appropriate coefficients for the phrase accent model is posed in a linear algebra framework. The coefficients are chosen in order to minimize the mean squared error between the actual phrase accent amplitudes and the amplitudes that would have been produced with the calculated coefficients. Coefficients are calculated for each type of accent using identical algorithms for each type. However, some information is not available for type 1 accents, like the number of words from the previous accents, due to the fact that type 1 accents always occur at the beginning of the sentence.

Data about N phrase accents are placed into the matrix A and the vector b. A is an $N \times 6$ matrix, and the nth row contains information about the context in which

Column	Definition
1	1 (to provide a constant bias for every accent)
2	Number of accent in sentence (i.e., first, second, etc.)
3	Number of words until next accent
4	Number of centiseconds until next accent
5	Number of words from previous accent
6	Number of centiseconds from previous accent

Table 3.5: Definition of elements of the nth row of the A matrix used for calculating the phrase accent model parameters

the nth phrase accent in the training data occured. Specifically, the elements in the nth row are defined in Table 3.5. b is a $N \times 1$ vector, with b_n equal to the amplitude of the nth phrase accent in the training data. The coefficients of the model are placed into the 6×1 vector x such that

$$x = \underset{\hat{x}}{\operatorname{argmin}} \ ||A\hat{x} - b||_2. \tag{3.24}$$

Note that $||z||_2$ is the 2-norm of the $M \times 1$ vector z, namely

$$||z||_2 = \sqrt{\left(\sum_{i=1}^{M}(z_i)^2\right)}. \tag{3.25}$$

3.5.3 Calculated Coefficients

The phrase accent model parameters calculated from both training databases are presented in Table 3.6. The set of coefficients built from the Marketplace training database used 1099 phrase accents (367 type 1, 327 type 2, and 239 type 3). The model built from the Wall Street Journal training database used 2348 phrase accents (999 type 1, 538 type 2, and 811 type 3). Comparing the constant coefficients of the different models, type 1 accents have the greatest amplitude, while type 3 accents are slightly larger than type 2. This agrees with the general view that the beginning of a sentence has a large accent, while intermediate phrase boundaries after silences

		Marketplace			Wall Street Journal		
		Type 1	Type 2	Type 3	Type 1	Type 2	Type 3
i	Information	x_i (Coefficients)			x_i (Coefficients)		
1	Constant	94.5	-23.0	15.4	11.6	-13.3	-4.30
2	Number of accent		2.76	-6.69		-1.24	-0.468
3	words to next acc.	1.81	-3.41	1.79	0.847	0.681	0.372
4	frames to next acc.	0.0185	0.228	0.0919	0.00269	0.0414	0.0518
5	words from last acc.		-0.775	-3.40		-0.0286	-0.432
6	frames from last acc.		0.118	0.141		-0.0098	0.0223

Table 3.6: Phrase accent coefficients for the linear model calculated from the Marketplace and Wall Street Journal training databases

produce larger accents than those without a silence. The other coefficients have less impact on the magnitude of the generated phrase accents, but at least they are fairly consistent between the two training databases.

Chapter 4

Synthesis

This chapter describes the synthesis stage of the fundamental frequency (F_0) generation system. To generate accent prototypes for a new sentence, the same linguistic processing that was performed during training (see Section 3.2) must be performed for the new sentence to generate the locations and the contexts for each accent. The decision trees and probabilities described in Sections 3.4.1 and 3.4.5 are used in a dynamic programming procedure to determine which of the pitch accent prototypes to use. The linear phrase accent model of Section 3.5 is used to determine the phrase accent amplitudes.

4.1 Pitch Accent Prototype Selection

This section describes the algorithm used to select the pitch accent prototypes to best represent the new sentence to be synthesized. First, the linguistic and lexical context information is obtained. Part-of-speech and lexical stress information is obtained in exactly the same manner as in training (see Sections 3.2.2 and 3.2.3). However, the locations of the stressed syllables can not be obtained via forced alignment as in Section 3.2.1, because the original spoken utterance will not be available during synthesis. Instead, the phone-level timing will be provided by another module of the synthesizer. However, in this work the spoken utterance is available, and the alignment will therefore be used to determine the pitch accent locations. The other

context information needed for determining the pitch accent prototypes is the intensity level of the previous accents produced, which is tracked dynamically as the generation progresses.

The notation used in the upcoming sections is as follows. A parenthetical superscript is a time or sequence index, while a subscript will refer to a particular element in a set. Thus, $v^{(i)}$ is the leaf used at the ith time step, while v_l is the lth leaf out of all of the leaves. Each prototype (r_{jl}) is inherently associated with a specific leaf (v_l), i.e., prototype j of leaf l. All of the probabilities calculated below are dependent upon the unique linguistic context of the sentence to be synthesized.

4.1.1 Dynamic Programming

Given the probabilities computed in Section 3.4.5, the prototypes are chosen such that the probability of the sequence of prototypes selected is greater than any other sequence. This selection involves two assumptions, which are described below. The probability of the sequence of prototypes, S, for the n pitch accents in the sentence can be written

$$\Pr(S) = \Pr(r^{(1)}, r^{(2)}, \ldots, r^{(n)}). \tag{4.1}$$

Using the definition of conditional probability, this expression can be written as a product of conditional probabilities,

$$\begin{aligned} \Pr(S) \;=\; & \Pr(r^{(n)} \mid r^{(n-1)}, \ldots, r^{(1)}) \cdot \Pr(r^{(n-1)} \mid r^{(n-2)}, \ldots, r^{(1)}) \cdot \\ & \cdots \Pr(r^{(2)} \mid r^{(1)}) \Pr(r^{(1)}). \end{aligned} \tag{4.2}$$

Each term of the above expansion can be written as

$$\Pr(r^{(i)} \mid r^{(i-1)}, \ldots, r^{(1)}) = \Pr(r^{(i)} \mid v^{(i)}, r^{(i-1)}, \ldots, r^{(1)}) \cdot \Pr(v^{(i)} \mid r^{(i-1)}, \ldots, r^{(1)}). \tag{4.3}$$

Now we make the two assumptions. First, we assume that the probability of the

current prototype ($r^{(i)}$) depends only on the identity of the current leaf ($v^{(i)}$), hence

$$\Pr(r^{(i)} \mid v^{(i)}, r^{(i-1)}, \ldots, r^{(1)}) = \Pr(r^{(i)} \mid v^{(i)}). \tag{4.4}$$

Second, we assume that the probability of the current leaf ($v^{(i)}$) depends only on the identity of the previous prototype ($r^{(i-1)}$), hence

$$\Pr(v^{(i)} \mid r^{(i-1)}, \ldots, r^{(1)}) = \Pr(v^{(i)} \mid r^{(i-1)}). \tag{4.5}$$

Therefore, the probability of the entire sequence of prototypes can be expanded to

$$\begin{aligned} \Pr(S) &= \Pr(r^{(1)} \mid v^{(1)}) \cdot \Pr(v^{(2)} \mid r^{(1)}) \cdot \Pr(r^{(2)} \mid v^{(2)}) \cdot \\ &\quad \cdots \Pr(v^{(n)} \mid r^{(n-1)}) \cdot \Pr(r^{(n)} \mid v^{(n)}). \end{aligned} \tag{4.6}$$

The probabilities on the right side of equation 4.6 are exactly the probabilities calculated earlier in Section 3.4.5,

$$\delta_{jl_1l_2} = \Pr(v^{(i)} = v_{l_2} \mid r^{(i-1)} = r_{jl_1}), \text{ and} \tag{4.7}$$

$$\lambda_{jl} = \Pr(r^{(i)} = r_{jl} \mid v^{(i)} = v_l). \tag{4.8}$$

With this formulation of the probability of a sequence of prototypes, dynamic programming is used to calculate the best sequence of prototypes. This process uses two variables that describe the best sequence of prototypes through accent i that ends in prototype j of leaf l. $\phi_{jl}(i)$: the probability of that sequence, and $\psi_{jl}(i)$: the prototype at accent $i-1$, of that sequence. These two variables, plus $\delta_{jl_1l_2}$ and λ_{jl} of Equations 4.7 and 4.8, are all that is needed to decide which sequence of prototypes is most probable, given the sentence to be synthesized.

The first variable,

$$\phi_{jl}(i) = \max_{S^{(1,\ldots,i)}} \Pr(r^{(i)} = r_{jl}), \tag{4.9}$$

is the probability of the optimal sequence of prototypes out of all the possible sequences of prototypes, $S^{(1,\ldots,i)}$ through accent i ending in prototype j of leaf l.

This variable can be expressed recursively using the probabilities calculated in Section 3.4.5,

$$\phi_{jl}(i) = \max_{1<j'<J,\, 1<l'<L} (\phi_{j'l'}(i-1) \cdot \delta_{j'l'l}) \cdot \lambda_{jl}, \tag{4.10}$$

where L is the total number of leaves and J is the maximum number of prototypes in a leaf (for a particular leaf there might be less than J prototypes). The context that specifies which leaf occurs at which accent location includes lexical and linguistic information and the intensity level of the previous accent. The previous intensity level is the only unknown context information before the sentence is synthesized. The intensity level of the previous prototype (j') completes the context information needed to determine the identity of the next leaf (l). Therefore, there is only one leaf that could follow each prototype, and the calculation of $\phi_{jl}(i)$ only involves searching over those prototypes that would cause leaf l to be produced as the next leaf. For the first iteration,

$$\phi_{jl}(1) = \lambda_{jl}, \tag{4.11}$$

but only for the leaf that is specified by the context, with the previous intensity level specified as beginning of sentence. For all other leaves, $\phi_{jl}(1)$ is zero.

The second variable used in the dynamic programming process,

$$\psi_{jl}(i) = \operatorname*{argmax}_{1<j'<J,\, 1<l'<L} (\phi_{j'l'}(i-1) \cdot \delta_{j'l'l}), \tag{4.12}$$

keeps track of the previous prototype in the optimal sequence of prototypes through accent number i ending at prototype j of leaf l. Since $\psi_{j'l'}(i-1)$ contains information about the prototype at accent number $i-2$ and the prototype that it points to contains information about its previous prototype and so forth, all of the prototypes in the path are known.

Once the dynamic programming procedure has been carried out through accent N, then the final prototype of the optimal path can be found,

$$\left\{j^{(N)}, l^{(N)}\right\} = \operatorname*{argmax}_{1<j'<J,\, 1<l'<L} (\phi_{j'l'}(N)) \tag{4.13}$$

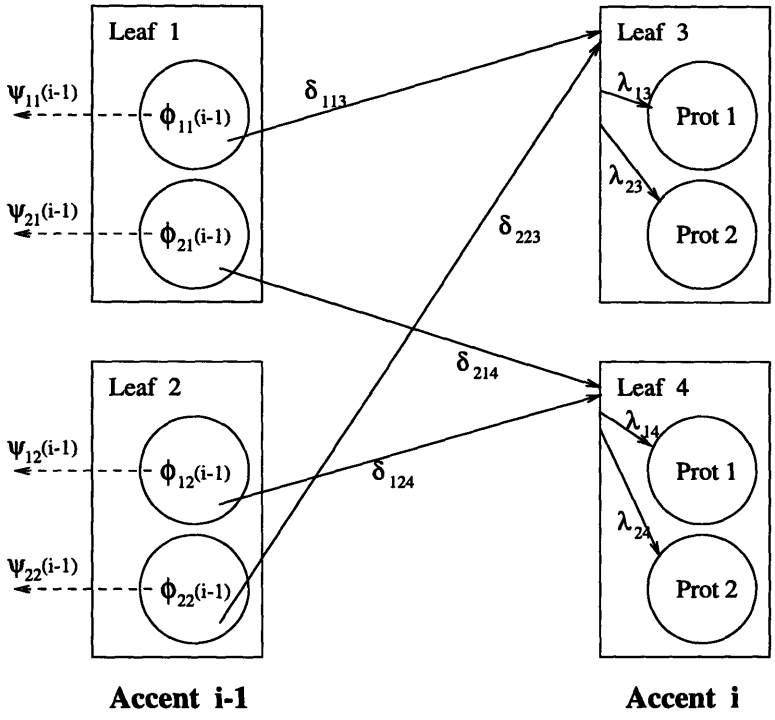

Figure 4-1: Dynamic programming process before stage i

$$r^{(N)} = r_{j^{(N)}l^{(N)}}. \tag{4.14}$$

Then, the rest of the prototypes can be computed iteratively, as follows:

$$\left\{j^{(i)}, l^{(i)}\right\} = \psi_{j^{(i+1)}l^{(i+1)}}(i+1) \tag{4.15}$$

$$r^{(i)} = r_{j^{(i)}l^{(i)}}, \tag{4.16}$$

with i going from $N-1$ down to 1.

Figures 4-1 and 4-2 give a graphical example of the dynamic programming process. In these figures, rectangles are leaves and circles are prototypes directly associated with the surrounding leaf. Solid arrows indicate probabilities, while dashed arrows are from one prototype to the previous one in its optimal path. Before stage i takes place, each of the prototypes of accent $i-1$ has a pointer back to the previous prototype in its optimal path (ψ), the cumulative probability of its optimal path (ϕ), and the probability (δ) of going to the leaf which must follow it in accent i. Additionally, each

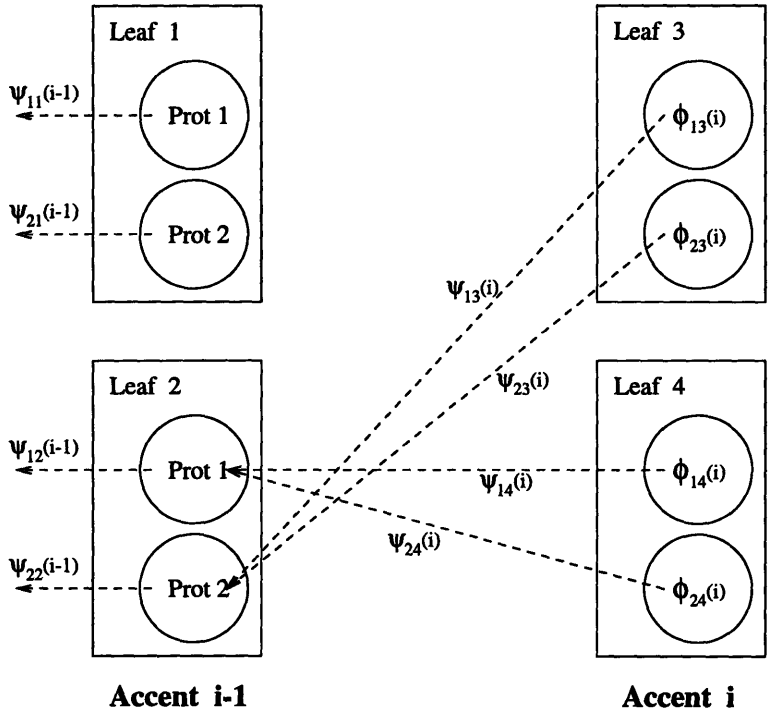

Figure 4-2: Dynamic programming process after stage i

of the leaves of accent i has probabilities (λ) for each of its prototypes occurring, given that the leaf has occured. Notice that each prototype in accent $i-1$ only has one leaf in accent i to which it has a non-zero transition probability. This is because the intensity of the prior prototype and the linguistic and lexical information about the sentence allows the decision tree to choose which leaf should be next. Notice that δ only interacts with the previous prototype and the current leaf, and that λ only interacts with the current prototype and leaf, just as in the two assumptions. After stage i in the dynamic programming procedure (Figure 4-2), the best previous prototypes have been chosen (ψ) and the cumulative probabilities have been calculated (ϕ) for the prototypes in the current accent. Notice that all prototypes in the same leaf point back to the same prototype in the previous accent. The computation of ψ and ϕ continues until the last accent is reached. At the end, starting at the prototype with the highest cumulative probability (ϕ), the pointers (ψ) are followed back to the beginning of the sentence to determine the best sequence.

4.1.2 Intensity Variation Model

As noted in Section 3.4.2, the variation of intensities of pitch accents is an important part of producing natural F_0 contours. Besides including the previous intensity in the context, the intensity variations can also be included in a probabilistic method. The calculation of the probabilities of moving from one intensity level to another was described in Section 3.4.5. A Markov assumption is made that the current intensity bin is only influenced by the intensity bin of the previous accent. Another probability of the sequence of prototypes can be calculated that is based solely on the intensity variation component as follows:

$$\begin{aligned} \Pr_{\text{int}}(S) &= \Pr(r^{(n)} \in B_{j_n} \mid r^{(n-1)} \in B_{j_{n-1}}) \Pr(r^{(n-1)} \in B_{j_{n-1}} \mid r^{(n-2)} \in B_{j_{n-2}}) \\ &\quad \cdots \Pr(r^{(2)} \in B_{j_2} \mid r^{(1)} \in B_{j_1}) \Pr(r^{(1)} \in B_{j_1}) \qquad (4.17) \\ &= \eta_{j_{n-1}j_n} \eta_{j_{n-2}j_{n-1}} \cdots \eta_{j_1 j_2} \eta_{j_0 j_1}, \qquad (4.18) \end{aligned}$$

where η_{jk} is defined in Equation 3.23. As in Section 3.4.2, B_j is all of the data points whose intensities are between b_{j-1} and b_j, with $b_0, \ldots, b_M$ intensity levels so that the data points are evenly distributed.

The probabilities of the sequence of prototypes can be found in two ways. The first method, described in Equation 4.2 in Section 4.1.1, uses the probabilities of transferring between leaves and prototypes. The new method, Equation 4.18, is discussed above. A better estimate of the most probable sequence of prototypes than either of them individually might be obtained if these two probabilities were combined. Their dynamic ranges might be different or one might be a better estimate than the other, therefore they were combined with a weighting factor. This is similar to the way some automatic speech recognition systems combine language model and acoustic model probabilities. Therefore, the new optimal probability ($\Pr^*$) of a sequence of prototypes, S, can be expressed as

$$\Pr^*(\mathrm{S}) = \Pr(\mathrm{S}) \cdot \Pr_{\text{int}}(\mathrm{S})^{\kappa}, \qquad (4.19)$$

where κ is an arbitrary constant weighting factor that is computed experimentally, so that the two probabilities are combined optimally.

This paradigm extending the probabilistic model fits easily into the dynamic programming procedure described above. Only three of the equations in Section 4.1.1 need to be changed so that the procedure calculates Pr*(S) instead of Pr(S). The calculation of the probability of the optimal path (Equation 4.10) becomes

$$\phi_{jl}(i) = \max_{1<j'<J,\, 1<l'<L} (\phi_{j'l'}(i-1) \cdot \delta_{j'l'l}) \cdot \lambda_{jl} \cdot (\eta_{B(j',l')B(j,l)})^{\kappa}, \tag{4.20}$$

where $B(j,l)$ is the intensity bin to which prototype j of leaf l belongs to. The initial condition (Equation 4.11) becomes

$$\phi_{jl}(1) = \lambda_{jl} \cdot (\eta_{oB(j,l)})^{\kappa}, \tag{4.21}$$

where $\eta_{oB(j,l)}$ is the probability that the intensity bin of the jth prototype of the lth leaf occurs on the first accent in the sentence. Furthermore, finding the previous prototype in the optimal path (Equation 4.12) becomes

$$\psi_{jl}(i) = \operatorname*{argmax}_{1<j'<J,\, 1<l'<L} (\phi_{j'l'}(i-1) \cdot \delta_{j'l'l}) \cdot (\eta_{B(j',l)B(j,l)})^{\kappa}. \tag{4.22}$$

4.2 Phrase Accent Amplitude Calculation

Determining the phrase accent amplitudes requires knowing the location and types of the phrase boundaries and the coefficients of the linear model. The locations and types are determined as in Section 3.2.4. Again, this is the only stage that is not implemented automatically. There have been some efforts to predict phrase boundaries from text, as discussed in Section 2.2. Some of these prediction algorithms could be incorporated into future systems for F_0 generation (see Section 6.2). The coefficients for the linear model are calculated as in Section 3.5.2. To get the amplitudes, the numerical information about the locations of the surrounding accents is simply plugged into the equations of the linear model.

Chapter 5

Results

This chapter will present some results obtained using the testing databases of both databases. Both objective (see Section 5.1) and subjective (see Section 5.2) evaluation techniques are used to measure the success of the generated contour. Along with the results of these tests, some example F_0 contours will be presented (see Section 5.3).

5.1 Objective Tests

This section will describe the objective tests used for this study to measure the success of the fundamental frequency (F_0) generation. The exact measures of error will be described. Additionally, the databases that were used for training and testing will be discussed. Finally, the results of these tests will be presented.

5.1.1 Error Measure

The objective measure is used to compute the effectiveness of the prediction of F_0 contours. It compares the contour produced by the original speaker to the contour generated by the model. Although both the generated (f_G) and the stylized (f_S) F_0 contours can be evaluated at all times in entire utterance, the difference between the two will only be measured at those frames when the silence and voiced tests judged the original speech to have a F_0 value (as opposed to being silence or unvoiced speech).

The measure that will be used is the RMS error between the two F_0 contours, defined here as,

$$\mathcal{E}_{\text{rms}} = \sqrt{\left(\sum_{i \in F} (f_S(i) - f_G(i))^2\right) / N}, \tag{5.1}$$

where F is the set of all frames with valid F_0 and $N = |F|$. This error measure can be computed with the F_0 contour in Hz ($\mathcal{E}^{\mathcal{H}}$) or in ln(Hz) ($\mathcal{E}^{\mathcal{L}}$).

5.1.2 Test Databases

The Marketplace database contains a main training and testing speaker plus a smaller number of sentences from four additional speakers. The main speaker, David Brancaccio, has 330 sentences for training and 37 sentences for testing. The other speakers are John Dimsdale (52 sentences), Sarah Gardner (60), Susan Goodman (24), and George Lewensky (37). For Wall Street Journal, the main speaker, a male, has 999 sentences for training and 141 sentences for testing.

The main test will be to train the system using the training databases, and to test it using the same speaker. In addition, the system trained on each of the main speakers will be tested on each of the additional speakers, to measure the level of speaker dependence. For each of the additional speakers, the global Fujisaki parameters (α, β, F_{min}) will be found that best describe that speaker. To measure how well the system captures domain specific information, the system that is trained on Marketplace will be tested on Wall Street Journal, and vice versa.

The tests will attempt to find the optimal size of the decision tree and the number of prototypes per leaf. Additionally, there will be tests to compare the result of calculating λ_{jl} using a hard assignment criterion (Equation 3.21) to that obtained when calculating it probabilistically (Equation 3.22). Finally, there will be tests to measure the effectiveness of the intensity variation model of Section 4.1.2.

5.1.3 Objective Test Results

Tables 5.1 and 5.2 present results from a variety of testing conditions, with the main test databases from the Marketplace (MP) and Wall Street Journal (WSJ) data. The numbers are the RMS error between the observed stylized F_0 contour and the contour generated using the testing conditions specified. The error rates obtained with MP are generally about twice the size of those obtained with WSJ. The average standard deviation of the observed F_0 contours in the test set is 31.52 Hz for MP and 17.96 Hz for WSJ, demonstrating that MP varies twice as much as WSJ. The average RMS error obtained when comparing all the test sentences to the same global mean is 33.85 Hz for MP and 18.39 Hz for WSJ.

The first set of results demonstrates the effects of probabilistic procedures on creating the prototypes and the probabilities needed for synthesis. As discussed in Section 3.4.5, the probability of the current prototype given the current leaf (λ_{jl}) can be calculated using the counts of the number of data points closest to each prototype or using a probabilistic equation. In Section 3.4.4, prototypes are created using the output of the k-means algorithm as seeds to a probabilistic procedure. Test 1 uses the more basic procedure for both the probability and the prototypes. Test 2 shows the improvement when the probabilistic procedure is used to create the prototypes. Test 3 gives a surprisingly bad result when λ_{jl} is calculated probabilistically. This final result might occur because the probabilities are closer to a uniform distribution and do not distinguish prototypes that represent outliers from those that represent a majority of the data.

The second set of results shows the effects of the choice of data used to build the decision tree for the pitch accent model (see Section 3.4.2). In test 4, the tree is built with one-dimensional data vectors containing only the intensity of the relevant pitch accent. In test 5, the tree is built with two-dimensional data vectors containing both intensity and relative length. The tree built without length information gives the better result (for the WSJ data, the difference is negligible).

The third set of results (tests 6–12) uses the intensity variation model of Sec-

Test	Conditions		$\mathcal{E}^{\mathcal{H}}_{\text{rms}}$	
	λ_{jl}	Prototypes	MP	WSJ
1	Hard counts	k-means	32.67	16.56
2	Hard counts	k-means + probabilistic	31.71	15.94
3	Probabilistic	k-means + probabilistic	35.59	17.41
4	Build tree **without** length information		31.71	15.94
5	Build tree **with** length information		32.46	16.09
	Intensity variation model tests		MP	WSJ
6	$\kappa = -1$		33.22	16.51
7	$\kappa = 0$		31.71	15.94
8	$\kappa = 0.25$		31.84	15.87
9	$\kappa = 0.5$		31.75	15.86
10	$\kappa = 1$		31.69	15.88
11	$\kappa = 2$		31.84	15.88
12	$\kappa = 4$		32.67	16.06
	Phrase Accents	Pitch Accents	MP	WSJ
13	Extracted	Extracted	12.33	6.51
14	Predicted	Extracted	18.58	9.03
15	Extracted	Predicted	31.71	15.94
16	Predicted	Predicted	33.94	15.86
	Cross Testing		MP	WSJ
17	Predicted	Extracted	24.75	29.04
18	Extracted	Predicted	34.40	20.75
19	Predicted	Predicted	43.93	38.87
	Phrase Model Tests		MP	WSJ
20	One Global Phrase Accent		22.0	12.4
21	Linear Model		21.3	11.8
22	Extracted Accents		18.0	10.8

Table 5.1: Results for objective tests on the Marketplace and Wall Street Journal databases with several test conditions

tion 4.1.2 with several weighting factors, κ. The model is slightly helpful for small positive values of κ, but begins to carry too much weight when κ reaches 4. The fact that performance degrades when κ is equal to -1 demonstrates that the intensity variation model is at least consistent with the main pitch accent model. For this value of κ, the probabilities relating to transferring from one intensity bin to another are divided instead of multiplied.

The fourth set of results shows the variation of the objective measure as each of the testing conditions used in the listening tests are used (see Section 5.2). Each of the phrase and pitch accents can be either extracted from the observed F_0 contour or predicted using their respective models. Tests 14 and 15 suggest that having the correct pitch accents is more important than having the correct phrase accents, or that phrase accents are easier to predict. For the WSJ data, the predicted phrase accents gave slightly better results than the extracted ones when using predicted pitch accents (tests 15 and 16). The results when both types of accents are predicted (test 16) can be compared to results from other research. Using the Boston University FM Radio corpus, a 33 Hz RMS error was obtained by Ross and Ostendorf [44] and a 34.8 Hz RMS error was obtained by Black and Hunt [6]. The database is similar to Marketplace, except that the speaker is female, thus the average standard deviation is slightly higher. One important difference between these two results and the results presented here is that the other studies predicted the F_0 contours from human-labelled ToBI labels, while this study used mostly information that can automatically be extracted from the text and speech of the database.

The fifth set of results uses the models trained on the other training database (i.e., the MP results use the phrase and pitch accent models trained with the WSJ training data). The remarkable degradation in performance indicates that these models are at least very domain specific, if not speaker specific. One interesting result is that for MP the error for test 17 is lower than for test 18, while for WSJ, the reverse is true. This suggests that the pitch accents are more important for MP, while the phrase accents are more important for WSJ.

The final set of results in Table 5.1 tests the phrase accent model described in

		Number of Leaves							
		Marketplace				Wall Street Journal			
		5	8	12	23	6	9	12	22
Prots Per Leaf	1	31.41	31.46	31.48	31.40	16.28	16.15	15.88	15.98
	2	31.41	31.60	31.52	31.41	16.30	16.20	15.94	16.07
	3	32.04	31.88	31.80	32.08	16.27	16.16	16.14	16.32
	5	31.48	31.75	31.71	32.41	17.39	16.30	15.94	16.55
	8	32.21	32.17	32.16	41.26	16.51	16.69	16.16	16.63

Table 5.2: Results of objective tests, varying both tree sizes and number of prototypes per leaf. (In RMS Hz.)

Section 3.5. The error for each of these tests in only calculated at the local minima of the stylized original contour, which is the criteria for finding the optimal phrase accent parameters as in Section 3.3.2. Additionally, the error is computed after the phrase accents have been generated, but before the pitch accents are added to the synthetic F_0 contour. In test 20, one global accent is calculated by averaging all of the relevant data, and is used for every accent. Test 21 uses the linear model, while test 22 uses the parameters that were originally extracted. These tests demonstrate that the linear model provides some improvement over using the same accent everywhere. However, this improvement is not substantial, especially for MP.

Table 5.2 presents results where the decision tree and the number of prototypes per leaf for the pitch accent model are allowed to vary. Extracted phrase accents are used in order to isolate the effects of the decision tree parameters. There is very little variation between the results, but several conclusions can be drawn. One disappointing result is that the prediction works as well or better when there is one prototype per leaf as when there are several. With only one prototype per leaf, most of the probabilistic modeling is not used. Additional results were obtained using only one leaf and one prototypes (i.e., the same accent at every stressed syllable), with average RMS errors of 32.9 Hz for MP and 17.1 Hz for WSJ. These final results indicate that both the decision tree and dynamic programming reduce the objective error significantly.

Table 5.3 presents tests performed on Marketplace testing speakers using a system

Speaker	sent.	μ_F	σ_F	$\mathcal{E}^{\mathcal{H}}_{\text{rms}}$	$\mathcal{E}^{\mathcal{L}}_{\text{rms}}$	F_{min}	α	β
David Brancaccio	37	131.3	31.31	33.94	0.2471	90.0	1.1	11.0
John Dimsdale	52	136.8	33.29	34.22	0.2444	94.6	1.2	14.0
Sarah Gardner	60	194.7	49.21	53.58	0.2716	134.3	1.2	10.0
Susan Goodman	24	195.4	39.57	42.04	0.2160	141.2	0.7	11.0
George Lewensky	37	101.7	23.55	26.07	0.2374	70.1	1.0	6.0

Table 5.3: Information about and objective test results from Marketplace testing databases

trained on the training database spoken by David Brancaccio. For each of the new testing speakers, the best global parameters (F_{min}, α, and β) are found such that when the phrase and pitch accent prototypes are plugged into the Fujisaki model, the generated and observed F_0 contours have the smallest error. The first two columns contain the name of the speaker and the number of sentences spoken. The second two columns are the mean and standard deviation of the observed, stylized F_0 contour. The third two columns contain the RMS error in Hz and in ln(Hz). The final three columns are the global parameters used for each sentence for that speaker. The results indicate that the system trained on David Brancaccio does equally well for the other Marketplace commentators. The RMS error in Hz is always a little higher than the standard deviation, and the RMS error in ln(Hz) is fairly consistent for all of the speakers. The optimal baseline value (F_{min}) is strongly correlated with the average F_0 value (μ_F), while the two decay rates (α and β) are very similar for all speakers.

5.2 Listening Tests

Listening tests have to be performed to evaluate the success of the F_0 generation algorithm. The objective tests performed in Section 5.1 compares several generated contours with one realization of a contour from a spoken utterance. However, for any sentence, every person who speaks the sentence uses different prosody. Thus, the generation algorithm could be judged to be natural even if it does not agree with the realization of the F_0 contour used for comparison. This section provides a description

Condition	Pitch Accents	Phrase Accents
1	**Original**	
2	Extracted	Extracted
3	Extracted	Predicted
4	Predicted	Extracted
5	Predicted	Predicted

Table 5.4: Testing conditions for listening tests

of the listening tests used in this study and some results from those tests.

5.2.1 Description

The sentences for these tests were generated using a straightforward linear prediction coefficient (LPC) synthesizer. The LPC filters used were obtained from the original utterance every frame (10 msec), along with the energy level. The F_0 contour being tested was used to generate the input to the filters. The input was either white noise, for unvoiced phones, or impulses at the desired fundamental frequency, for voiced phones. The output was scaled such that the energy was equal to the energy of the original utterance for each frame.

For each sentence, five test utterance were generated to measure the effectiveness of each part of the system. The test conditions are summarized in Table 5.4. The first uses the original stylized F_0 contour, using the LPC synthesizer (not the original speech itself). The other four use the combinations provided by using either predicted or extracted parameters for phrase and pitch accents. Extracted parameters are those found which best describe the original stylized F_0 contour, as in Section 3.3. Predicted parameters are the pitch accents found using the probabilistic models as in Section 4.1 and the phrase accents derived from the linear model as in Section 4.2. These conditions were chosen to measure how well the extraction of Fujisaki model parameters works (conditions 1 and 2), which type of accent is most important (conditions 3 and 4), and the overall success of the prediction (condition 5).

The sentences which were synthesized were chosen randomly from the test sets

	Database	
Condition	Marketplace	Wall Street Journal
1	4.3	3.9
2	3.7	3.4
3	3.4	3.7
4	2.7	2.4
5	3.0	2.6

Table 5.5: Mean Opinion Scores from listening tests

of both the Marketplace and the Wall Street Journal databases. For each database, the size of the decision tree and the number of prototypes at each leaf was optimized using the remainder of the test set.

All test conditions for each sentence were presented to the listener in a random order. The listener was only given one opportunity to listen to each test, and was asked to provide an evaluation for each test immediately after hearing it. The evaluation was to rate the naturalness of the prosody on a scale of one to five (unsatisfactory, poor, fair, good, excellent). The listener was initially presented with all five conditions of one sentence, without having to rate them, in order to become familiar with the range of quality. The averages of the evaluations of all of the listeners was used to form Mean Opinion Scores [23] for each of the test conditions.

The people who listened to the tests were drawn from the Human Language Technologies group at IBM research. Those selected have knowledge of speech processing, but none of them worked on this speech synthesis project. Another criteria for selection was that the subjects must be native speakers of American English, so that they could better judge the naturalness of the F_0 contours.

5.2.2 Listening Test Results

Table 5.5 presents the results from the listening tests, with the conditions defined above in Table 5.4. The results mostly conform to expectations, but there are a few surprises. Firstly, there is a decreased perception of naturalness through the conditions, which corresponds to the increase in automation in generating the F_0

contour. Secondly, the sentences from the Marketplace database were judged to be better than those from the Wall Street Journal database, which could be expected since the radio announcers were trying to use prosody to make their speech more interesting. Thirdly, the use of the predicted phrase accents model does not change perceived naturalness compared to using extracted accents. In the two pairs of conditions where the only difference was the change from extracted to predicted phrase accents (conditions 2 and 3 and conditions 4 and 5), the results are very similar, with the predicted phrase accents slightly outperforming the extracted ones in three out of four cases. These results confirm some earlier work by Möbius and Pätzold [37], in which listeners prefered original F_0 contours over extracted contours, and preferred extracted F_0 contours to predicted contours.

5.3 Examples

This section will present a few example F_0 contours, in order to illustrate some of the capabilities of this system. In these figures, three different F_0 contours are shown. The observed (stylized) contour is presented as a series of circles only at those ten millisecond frames that were judged to have a valid F_0 value. The contour generated by the Fujisaki model using the parameters extracted from the original contour is shown as a solid line. The dashed line is the predicted F_0 contour, using the model to generate both the pitch and phrase accents. Additionally, word boundaries are marked with solid vertical lines and phone boundaries are marked with dotted vertical lines.

Figure 5-1 shows the constrained Fujisaki model can successfully approximate the observed F_0 contour. The predicted contour is quite far away from the desired amplitude, but the patterns of rise and fall are somewhat similar in all contours. Even when forcing the pitch accents to occur during vowels in lexically stressed syllables (/ow/ in "new," /ay/ in "Chrysler," /aa/ in "car," and /iy/ in "Neon"), the fitted contour follows the observed contour very closely in the first two words, and almost exactly in the remaining words.

Figure 5-2 demonstrates the predicted contour being very similar to the observed

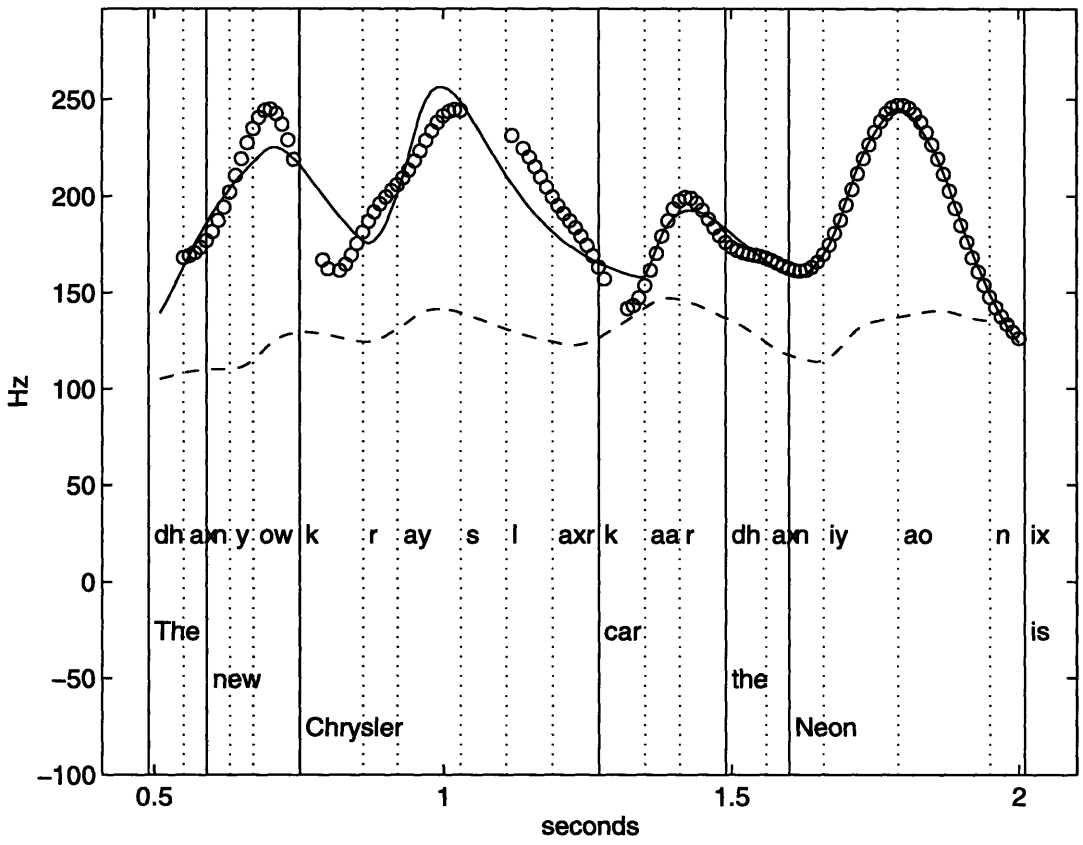

Figure 5-1: Partial F_0 contour of Marketplace sentence, "The new Chrysler car the Neon is just showing up in show rooms and it's being recalled, again."

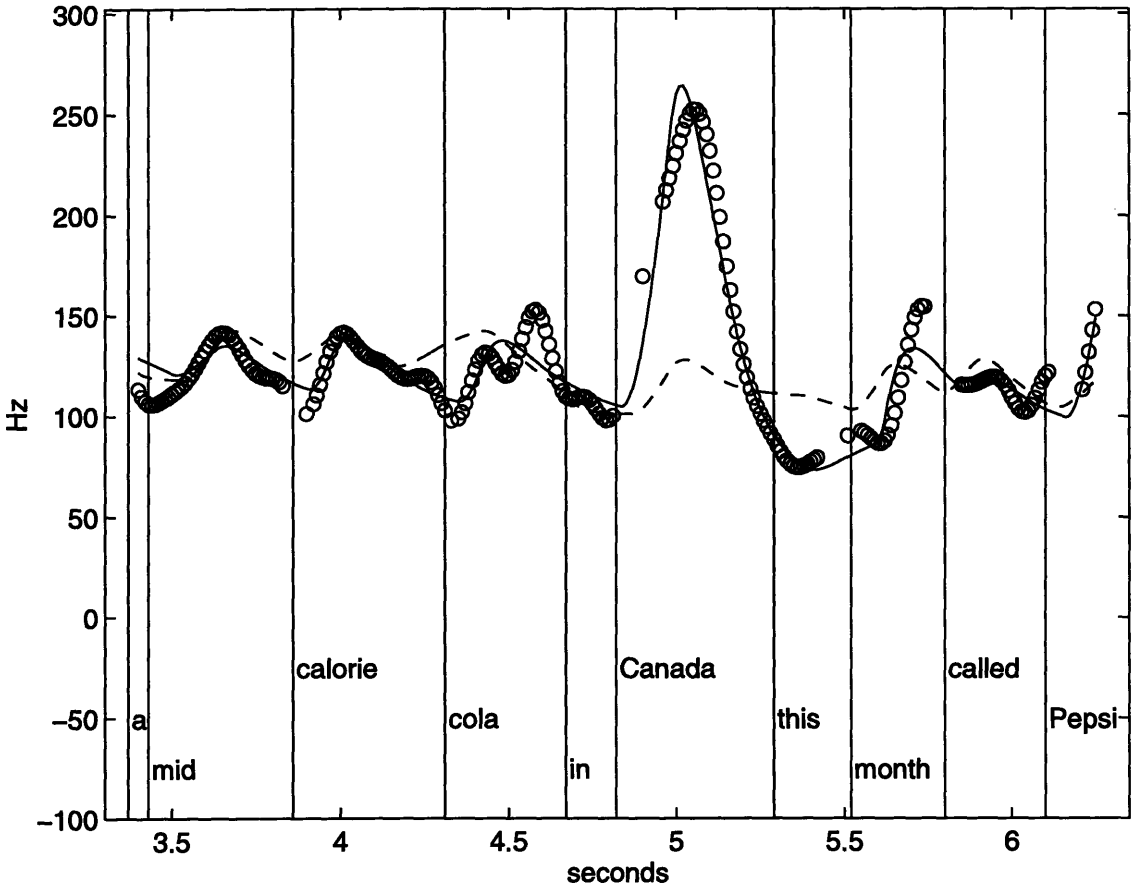

Figure 5-2: Partial F_0 contour of Marketplace sentence, "According to the San Francisco Examiner today, Pepsi is launching a mid calorie cola in Canada this month called Pepsi Max, which has about 50 calories a serving instead of 160."

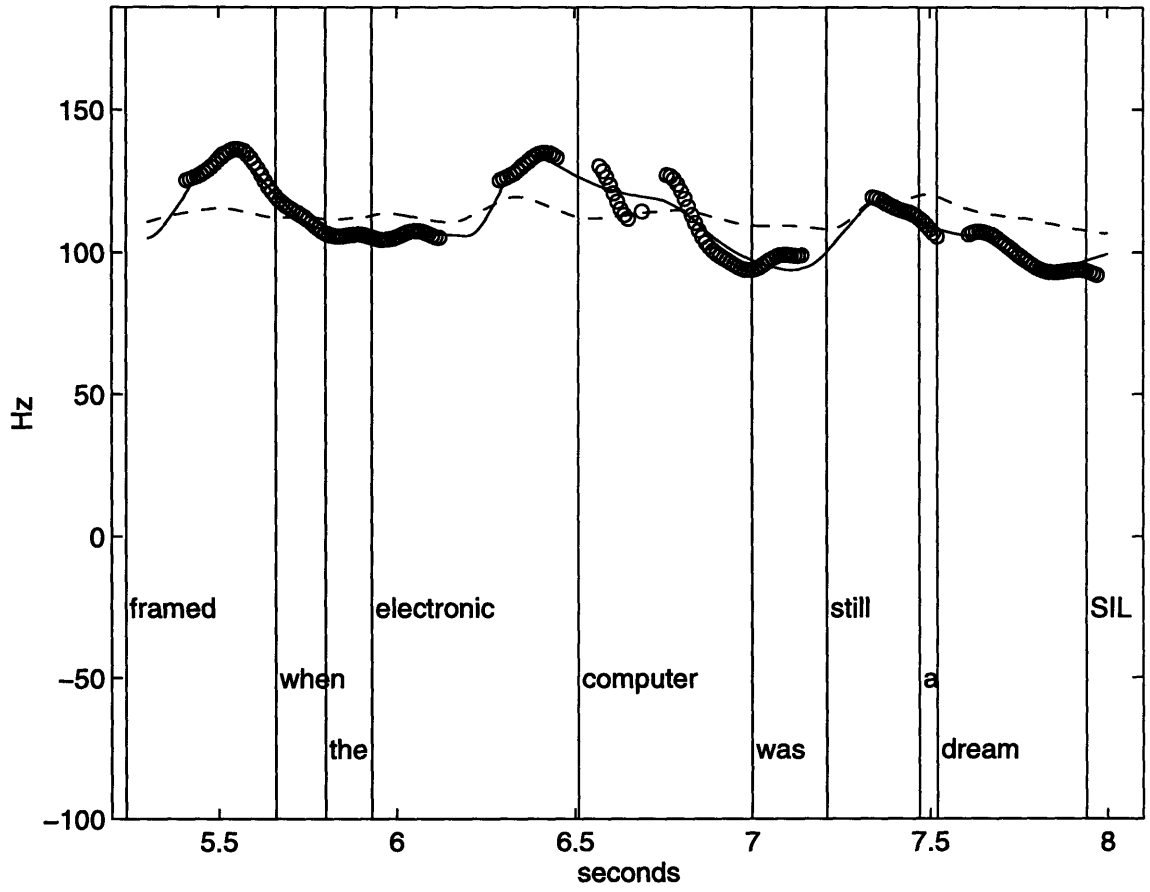

Figure 5-3: Partial F_0 contour of Wall Street Journal sentence, "Our current policy is still based on the Communications Act of 1934, framed when the electronic computer was still a dream."

contour for most of this segment, only differing at the word "Canada," where a very large pitch accent occurs in the spoken sentence. The problem occurs because it is very difficult to predict that the pitch accent for "Canada" will be so large. Accents in similar linguistic contexts, or even the accent in the same sentence spoken at a different time, will not necessarily have the same magnitude. This system can only hope to find those linguistic contexts where such large accents are more likely to occur, and to provide a contrast with those contexts where they are less likely to occur.

Figure 5-3 demonstrates some important differences between the Wall Street Journal and the Marketplace databases. In the first two F_0 contour examples (Figures 5-1

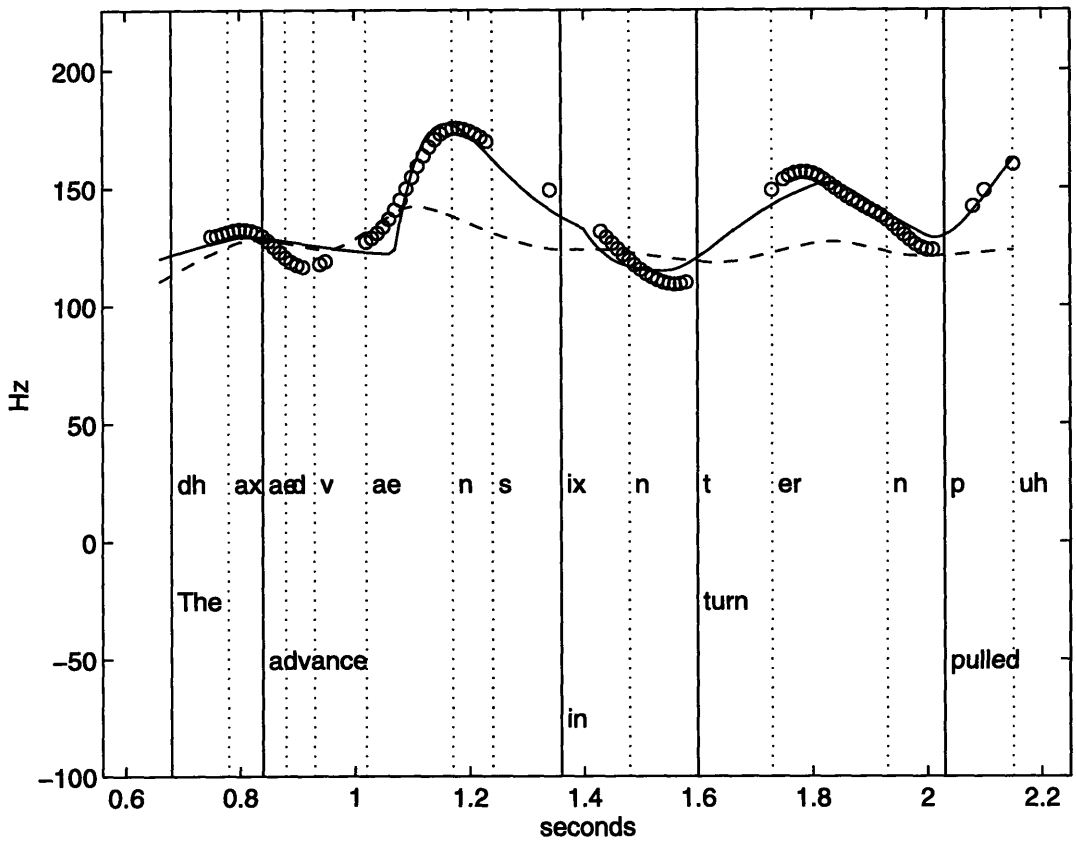

Figure 5-4: Partial F_0 contour of Wall Street Journal Sentence, "The advance in turn pulled up prices of delivery months representing the new crop season which will begin August first."

and 5-2), the variability is typical of the contours in the Marketplace database, while Figure 5-3 shows the relative lack of variability in the pitch contour that is typical of the Wall Street Journal sentences. Additionally, the predicted contour is able to approximate the observed contour more closely, showing that the accent prototypes are easier to predict for this database.

The pitch contours in Figure 5-4 are from a segment of another Wall Street Journal sentence. Again, the variability of the F_0 contour is much lower than from the Marketplace database, even at the beginning of the sentence. This example is another demonstration of how the constrained Fujisaki model can successfully approximate the observed contour. Two positive accents in "advance" and "turn" and one negative accent in "in" are all that is needed.

Chapter 6

Conclusions

In this final chapter, the new concepts which were presented in this thesis are summarized and some lines for further research are suggested.

6.1 New Concepts

Several new concepts have been introduced in this thesis which had not been documented previously. This section will outline these concepts and give references to detailed discussions about each.

6.1.1 Fujisaki Parameter Searching

The extraction of parameters of the Fujisaki model that best describe the F_0 contour of a spoken utterance uses two new ideas. The first is to search for phrase accent and pitch accent parameters consecutively instead of simultaneously (see Section 3.3.2). Furthermore, the phrase accent parameters are chosen so that the generated F_0 contour matches only the minima of the stylized F_0 contour, because the phrase accents are intended to describe a baseline for the contour. The second idea is to constrain the parameters using lexical and linguistic information (see Section 3.3.3). The ends of pitch accents are only allowed to occur during syllables which have known lexical stress, and phrase accents are placed only at phrase boundaries. These ideas have

been proven to be valid by the results presented in Section 3.3.4.

6.1.2 Pitch Accent Prototype Creation

In this study, pitch accent prototypes are created statistically so that the accents are most representative of the training data. Decision trees are built to define linguistic contexts where similar pitch accents occur (see Section 3.4.1). The parameters from Fujisaki's model are used as data for the tree building, with the new ideas of intensity and relative length used for normalization (see Section 3.4.2). Finally, a probabilistic procedure is used to actually create the prototypes and probabilities which describe them (see Sections 3.4.4 and 3.4.5).

6.1.3 Pitch Accent Prototype Selection

Out of all of the possible sequences of pitch accent prototypes, the one that is judged the most probable is chosen. In order to find the most likely sequence, an iterative dynamic programming procedure is used (see Section 4.1.1). As an addition to this procedure, an intensity variation model is introduced which attempts to make the intensities of the sequence of pitch accent prototypes vary in the same manner that was observed in the training data (see Section 4.1.2).

6.2 Further Research

In order for this F_0 generation system to be completely automated, the selection of phrase boundaries must be an unsupervised module of the system. There have been several studies which have successfully accomplished this [25, 39, 50, 52, 58]. Some which might be particularly useful for this work include predicting phrase boundaries using a stochastic parser [52] or using part-of-speech triples [50]. Both of these methods use information that is already being used in this work.

Another avenue of research that might prove fruitful would be to have more information available for questions to grow the pitch accent decision tree. This information

could either be more global (i.e., discourse model) or more local (i.e., phone-level information). Additionally, one or several words in a sentence could be flagged which should be assigned large pitch accents. This extra information would help separate accents which have similar linguistic contexts, but different accents. The paradigm of creating speech from purely text would have to be modified slightly, but it should not be difficult for a user to indicate which words should be emphasized.

One problem that arises in the creation of prototypes is that the extreme accents (very small or large intensities) are not often represented. Even with several prototypes per leaf, their probabilities are usually so small that they are not often used. Consequently, most of the prototypes are close to the global average. This might be avoided by creating one prototype for each intensity bin. At each leaf and for each sentence, the distribution of data points in intensity bins could be stored. To choose which prototypes to use for a new sentence, the probabilities of the prototypes given the linguistic contexts and of the total number of each prototype in the entire sentence could be optimized simultaneously. Like actual speech, this method would allow similar linguistic contexts to behave differently and increase the likelihood that a few extreme accents occur in each sentence.

Appendix A

Database details

This appendix presents the parameters of the Fujisaki model along with the information that will be used to build the pitch accent decision trees and to compute the coefficients in the linear phrase accent model. The accents are taken from the training set of the Marketplace (MP) and Wall Street Journal (WSJ) databases. The results confirm some basic assumptions and highlight some differences between the two databases. The major difference is that the dynamic range of the MP parameters is much greater than those from WSJ. This result is in line with the assumption that the Marketplace database is more interesting prosodically. Each of the tables presented here demonstrates some trend associated with the information. The combination of all of these trends is what enables both pitch and phrase accents to be predicted successfully.

A.1 Pitch Accents

Table A.1 contains the number and average intensity of pitch accents by the part-of-speech category that the pitch accents occur in. The part-of-speech categories are defined explicitly in Table 3.1. The percentages of each part-of-speech category are very similar for the two databases, with two exceptions. There is a much higher percentage of proper nouns (type 2) in the MP data, while there is a much higher percentage of numbers (type 13) in the WSJ data. The average intensity values

Part-of-Speech Category	Marketplace		Wall Street Journal	
	Number	Average	Number	Average
1	2566	6.26	5048	2.34
2	826	7.29	1203	3.61
3	375	5.93	875	2.40
4	602	6.32	1530	2.95
5	405	3.16	826	1.70
6	827	7.21	1675	3.88
7	363	7.94	755	3.66
8	922	-1.19	1709	0.84
9	278	1.70	690	4.14
10	1039	1.03	1909	0.99
11	186	-0.12	458	1.75
12	205	3.72	488	2.33
13	323	9.69	1089	3.23
14	49	9.87	76	3.52
Total	8966	4.80	18331	2.47

Table A.1: Average intensity of pitch accents by part-of-speech category

indicate that the words conveying the most meaning are more accented. Determiners, prepositions, and the word "to" (types 8, 10, 11) are consistently significantly below average, while proper nouns, adjectives, adverbs, and numbers (types 2, 6, 7, 13) are consistently above average. In a contrast between the two databases, pronouns (type 9) are above average for WSJ data, but below average for MP data.

Table A.2 presents the number and average intensity for pitch accents in four conditions, using all permutations of primary and secondary lexical stress and function

		Marketplace		Wall Street Journal	
		Number	Average	Number	Average
Meaning	Primary	5281	7.44	10845	3.22
	Secondary	642	2.53	1383	0.89
Function	Primary	3029	0.69	6081	1.50
	Secondary	14	0.28	22	0.70

Table A.2: Average intensity of pitch accents by primary/secondary stress and meaning/function word categories

Distance from Phrase	Marketplace		Wall Street Journal	
	Number	Average	Number	Average
1	988	7.43	2348	5.13
2	981	4.83	2317	3.37
3	972	4.01	2276	2.81
4	971	2.71	2192	1.90
5	942	4.49	2088	1.86
6	884	3.39	1884	1.70
7	779	4.14	1615	1.60
8	678	5.40	1272	1.76
9	549	4.63	966	1.55
10	419	6.14	662	1.32
11	318	5.84	388	0.94
12	227	6.02	198	0.81
13	139	7.74	87	1.14
14	71	6.98	28	-0.94
15	49	8.57	10	4.95

Table A.3: Average intensity of pitch accents by distance from previous phrase boundary (in accents)

or meaning word (a meaning word is one which is not a function word). The most important result from this table is that accents in meaning words on primary stressed syllables have much higher intensity values than any other type of accent. Interestingly for MP data, the intensity of secondary stressed syllables in meaning words is higher than primary stress in function words, while in WSJ data, the opposite is true. Finally, Table A.2 shows that secondary stressed syllables occur very infrequently in function words, because most function words are only one syllable.

Table A.3 contains the number and average intensity values of pitch accents, indexed by how many accents they are away from the previous phrase boundary. The main observation to be drawn from this table is that the plot of the average intensities is similar to an inverted phrase accent (see Figure 2-2 for a phrase accent plot). This trend is especially true in the MP data, where the fourth accent after the phrase boundary is much lower, and the accents immediately after and considerably after are much higher.

Intensity Level	Marketplace		Wall Street Journal	
	Number	Average	Number	Average
1	607	14.31	1190	6.64
2	591	8.00	1087	3.72
3	614	5.54	1106	3.07
4	630	5.30	1204	2.58
5	635	5.10	1205	2.38
6	627	4.81	1229	1.87
7	633	5.45	1268	1.99
8	621	4.31	1268	1.58
9	622	4.38	1286	1.64
10	615	3.75	1295	1.61
11	611	4.17	1293	1.69
12	611	2.90	1303	1.40
13	603	0.09	1298	1.46
14	617	-5.59	1300	-1.20
None	330	13.64	999	8.67

Table A.4: Average intensity of pitch accents by intensity level of previous pitch accent

Table A.4 presents the number and average intensity of pitch accents, based on the intensity level of the preceding accent. The concept of intensity levels is discussed in Section 3.4.2. This table further demonstrates that the higher the intensity of a pitch accent, the lower the intensity of the next accent. For the MP data, this is most evident in the first three and last three intensity levels, with the middle levels being more even. When there is no preceding accent (i.e., the first accent in the sentence) the intensity is very high. However, this value might be skewed by the fact that the starting time of the first accent can extend into the time before the sentence begins. Some of the intensity of such accents is used to perturb the generated contour before the sentence begins, where the generated contour is not evaluated.

A.2 Phrase Accents

Table A.5 presents information about phrase accents from both databases. The two left-most columns refer to the number of accent in the sentence (i.e., first or

Accent Number	Accent Type	Marketplace			Wall Street Journal		
		Number	Amplitude	Words	Number	Amplitude	Words
1	1	358	169.1	8.4	999	18.82	7.7
	2	9	92.2	4.2	0		
	3	0			0		
	All	367	167.2	8.3	999	18.82	7.7
2	1	28	111.9	10.9	0		
	2	196	99.7	8.9	384	1.79	7.7
	3	103	121.6	8.9	435	3.79	8.1
	All	327	107.7	9.1	819	2.85	8.0
3	1	17	128.9	8.8	0		
	2	94	95.8	8.6	121	-1.25	7.3
	3	128	115.1	9.1	313	1.70	7.5
	All	239	108.5	8.9	424	0.88	7.4
4	1	7	112.0	9.6	0		
	2	53	98.6	8.4	33	-4.31	6.1
	3	60	114.6	9.6	60	-1.31	6.2
	All	120	107.4	9.0	93	-2.38	6.2
≥ 5	1	2	105.2	8.5	0		
	2	21	115.0	8.9	0		
	3	23	85.5	8.4	3	-5.58	6.0
	All	46	99.8	8.6	3	-5.58	6.0
All	All	1099	127.4	8.7	2348	9.07	7.7

Table A.5: Average amplitude of phrase accents by number in sentence and type of accent

third) and the type of accent as defined in Section 3.2.4, respectively. The remaining columns contain the number of accents in that category, the average amplitude of those accents, and the number of words until the next phrase accent. The major difference between the two databases is that the accents in the MP data are much higher than those in the WSJ data. This effect is largely due to the larger $F_{\min}$ which was found to be optimal for the WSJ data. The reason for the difference in $F_{\min}$ values is that there is a greater declination effect in the MP data, so large phrase accents model the data better. After the first accent in the sentence, the minor accents associated with a silence (type 3) are larger than those not associated with a silence (type 2). This makes sense, because a silence is a clear indication that the speaker considers that point to be a phrase boundary, while a purely structural phrase boundary might not be a location where the speaker will place a phrase accent. Another important observation is that the first accent in the sentence is clearly larger than all other categories of accents. The WSJ speaker took more pauses, thus there is a larger percentage of type 3 accents in that database. This is also a reason why phrase accents occur more frequently (inverse of the average number of words between accents) overall in the WSJ database.

Appendix B

Example Decision Trees

This appendix presents two decision trees that were grown automatically in order to create the prototypes for the pitch accent model. The two trees and their associated questions are generated from the data of the Marketplace and Wall Street Journal training databases. The trees are not complete, but highlight the questions that affected the largest amount of data. Each node in the decision tree contains a bold node number and the number of data points present at the node. The node numbers are cross-referenced in the tables containing the questions. For a particular piece of data, if the answer to the question associated with the node is yes, then that data follows the left arrow leaving the node. If the answer is no, then the data follows the right arrow. A description of how the decision trees are built can be found in Section 3.4.1 and the types of questions that can be asked at each node are described in Section 3.4.3. Note that there are fifteen intensity bins, numbered from lowest intensity to highest.

For the Wall Street Journal database, the graphical representation of the decision tree is Figure B-1 and the associated questions are in Table B.1. For the Marketplace database, the same information can be found in Figure B-2 and Table B.2. The questions that are asked in both of these examples demonstrate further that clustering the pitch accent data using decision trees is a reasonable approach. The questions include many about the previous intensity bin and about the location of the surrouding silences. Part-of-speech context is used with questions asked about

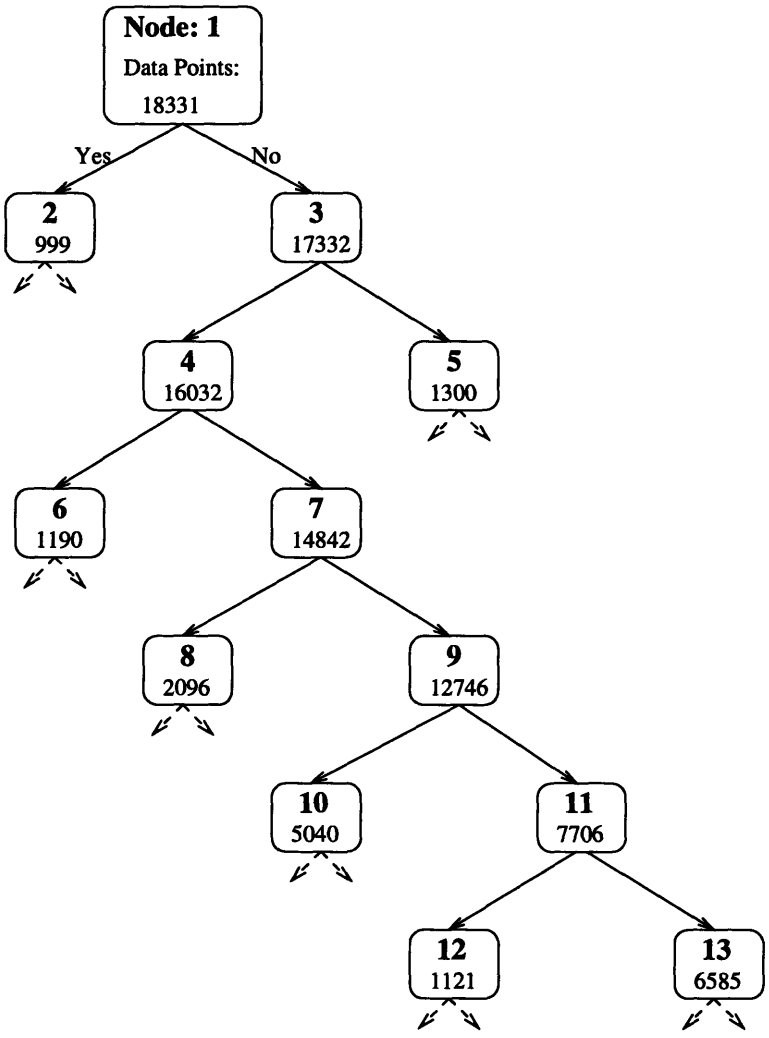

Figure B-1: Pitch accent model decision tree for the Wall Street Journal training database

nouns, determiners, pronouns, conjunctions and function words. In the Marketplace decision tree, questions about the location of the next phrase accent are also used. All of these questions fulfill their purpose of discriminating between different types of pitch accents.

Node	Question
1	Previous Intensity Bin = 15
2	Current Part-of-Speech = Pronoun
3	Previous Intensity Bin $\leq$ 13
4	Previous Part-of-Speech = Regular Noun
5	Current Word = Function
6	Previous Tag = Silence
7	Tag after next = Silence
8	Next Tag = Silence
9	Current Word = Function
10	Previous Intensity Bin $\leq$ 10
11	Tag before previous = Silence
12	Previous Tag = Noun or Verb
13	Previous Intensiy Bin $\leq$ 5

Table B.1: Decision tree questions for the Wall Street Journal training database

Node	Question
1	Previous Tag $\neq$ Silence
2	Previous Intensity Bin $\leq$ 12
3	Previous Intensity Bin $\leq$ 2
4	Previous Intensity Bin = 1
5	Current Word = Function
6	Previous Intensity Bin = 15
7	Next Phrase Accent $\leq$ 8 Words Away
8	Current Word = Function
9	Previous Intensity Bin = 14
10	Previous Part-of-speech = Preposition
11	Current Part-of-speech = Determiner
12	Next Phrase Accent $\leq$ 7 Words Away
13	Current Lexical Stress = Primary
14	Previous Part-of-speech = Noun
15	Next Part-of-speech = Noun
16	Previous Intensity Bin = 11
17	Previous Intensity Bin $\leq$ 6
18	Previous Part-of-Speech = Conjunction

Table B.2: Decision tree questions for the Marketplace training database

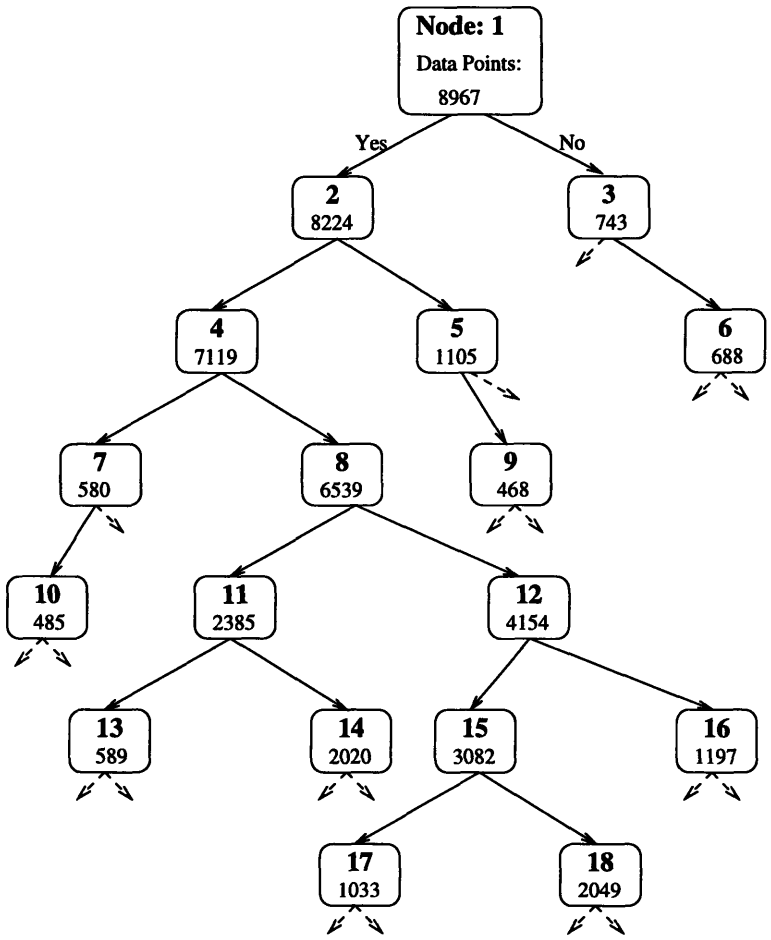

Figure B-2: Pitch accent model decision tree for the Marketplace training database

Bibliography

[1] Jonathan Allen. Synthesis of speech from unrestricted text. *Proceedings of the IEEE*, 64(4):433–442, April 1976.

[2] Véronique Aubergé. Developing a structured lexicon for synthesis of prosody. In Gérard Bailly, Christian Benoît, and T. R. Sawallis, editors, *Talking Machines*, pages 307–321. Elsevier Science Publishers, 1992.

[3] Lalit Bahl, Peter de Souza, P. S. Gopalakrishnan, and Michael Picheny. Context dependent vector quantization for continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 632–635, Minneapolis, 1993.

[4] Gérard Bailly. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8:137–146, 1989.

[5] Christian Benoît and Louis C. W. Pols. On the assessment of synthetic speech. In Gérard Bailly, Christian Benoît, and T. R. Sawallis, editors, *Talking Machines*, pages 435–441. Elsevier Science Publishers, 1992.

[6] Alan Black and Andrew Hunt. Generating F_0 contours from ToBI labels using linear regression. In *Proceedings of the International Conference of Spoken Language Processing*, Philadelphia, October 1996.

[7] Sin-Horng Chen, Shaw-Hwa Hwang, and Chun-Yu Tsai. A first study of neural net based generation of prosodic and spectral information for Mandarin text-to-speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 45–48, San Francisco, 1992.

[8] Robert Donovan. *Trainable Speech Synthesis.* PhD thesis, University of Cambridge, 1996.

[9] T. Dutoit and H. Leich. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, 1993.

[10] Glenn Farley. Control of voice F_0 by an artificial neural network. *Journal of the Acoustical Society of America*, 96(3):1374–1379, September 1994.

[11] Hiroya Fujisaki, Keikichi Hirose, and Haitao Lei. Prosody and syntax in spoken sentences of standard Chinese. In *Proceedings of the International Conference of Spoken Language Processing*, pages 433–436, Banff, Canada, October 1992.

[12] Hiroya Fujisaki, Keikichi Hirose, Noboru Takahashi, and Hiroyoshi Morikawa. Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and television announcers. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2039–2042, Tokyo, 1986.

[13] Hiroya Fujisaki and Hisashi Kawai. Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 663–666, New York, 1988.

[14] Hiroya Fujisaki and Sumio Ohno. Analysis and modeling of fundamental frequency contours of English utterances. In *Proceedings of Eurospeech*, pages 985–988, Madrid, September 1995.

[15] Hiroya Fujisaki and Sumio Ohno. Prosodic parameterization of spoken Japanese based on a model of the generation process of F_0 contours. In *Proceedings of the International Conference of Spoken Language Processing*, Philadelphia, October 1996.

[16] Hiroya Fujisaki, Sumio Ohno, Kei-ichi Nakamura, Miguelina Guirao, and Jorge Gurlekian. Analysis of accent and intonation in Spanish based on a quantitative model. In *Proceedings of the International Conference of Spoken Language Processing*, pages 355–358, Yokohama, September 1994.

[17] Hiroya Fujisaki, Sumio Ohno, Masafumi Osame, Mayumi Sakata, and Keikichi Hirose. Prosodic characteristics of a spoken dialogue for information query. In *Proceedings of the International Conference of Spoken Language Processing*, pages 1103–1106, Yokohama, September 1994.

[18] Hiroya Fujisaki, Sumio Ohno, and Osamu Tomita. On the levels of accentuation in spoken Japanese. In *Proceedings of the International Conference of Spoken Language Processing*, Philadelphia, October 1996.

[19] Edouard Geoffrois. A pitch contour analysis guided by prosodic event detection. In *Proceedings of Eurospeech*, pages 793–796, Berlin, 1993.

[20] Toshio Hirai, Norio Higuchi, and Yoshinori Sagisaka. Automatic detection of major phrase boundaries using statistical properties of superpositional F_0 control model parameters. In *Proceedings of Eurospeech*, pages 1341–1344, Madrid, September 1995.

[21] Keikichi Hirose and Hiroya Fujisaki. Analysis and synthesis of voice fundamental frequency contours of spoken sentences. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 950–953, Paris, 1982.

[22] Julia Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63:305–340, 1993.

[23] IEEE. IEEE recommended practice for speech quality measurements. *IEE Transaction on Audio and Electroacoustics*, AU-17(3):225–246, September 1969.

[24] M. E. Johnson. Synthesis of English intonation using explicit models of reading and spontaneous speech. In *Proceedings of the International Conference of Spoken Language Processing*, Philadelphia, October 1996.

[25] Andreas Kießling, Ralf Kompe, Anton Batliner, Heinrich Niemann, and Elmar Nöth. Automatic labeling of phrase accents in German. In *Proceedings of the International Conference of Spoken Language Processing*, pages 115–118, Yokohama, September 1994.

[26] Dennis Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

[27] John Laver. *Principles of Phonetics.* Cambdrige Textbooks in Linguistics. Cambridge University Press, 1994.

[28] Linguistic Data Consortium, University of Pennsylvania. *COMLEX English Pronouncing Lexicon*, release 0.2 edition, 1995.

[29] Andrej Ljolje and Frank Fallside. Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(3):225–246, October 1986.

[30] Mats Ljungqvist and Hiroya Fujisaki. Generating intonation for Swedish text-to-speech conversion using a quantitative model for the F_0 contour. In *Proceedings of Eurospeech*, pages 873–876, Berlin, 1993.

[31] Joaquim Llisterri. Prosody encoding survey. URL:http://www.lpl.univ-aix.fr/projects/multext/CES/CES2.html, 1996.

[32] E. López-Gonzalo and L. Hernández-Gómez. Automatic data-driven prosodic modeling for text to speech. In *Proceedings of Eurospeech*, pages 585–588, Madrid, September 1995.

[33] F. Mana and S. Quazza. Text-to-speech oriented automatic learning of Italian prosody. In *Proceedings of Eurospeech*, pages 589–592, Madrid, September 1995.

[34] Marketplace Web Site. URL: http://www.usc.edu/marketplace/, 1996.

[35] Nobuaki Minematsu, Seiichi Nakagawa, and Keikichi Hirose. Prosodic manipulation system of speech material for perceptual experiments. In *Proceedings of the International Conference of Spoken Language Processing*, Philadelphia, October 1996.

[36] Hansjörg Mixdorff and Hiroya Fujisaki. A scheme for a model-based synthesis by rule of F_0 contours of German utterances. In *Proceedings of Eurospeech*, pages 1823–1826, Madrid, September 1995.

[37] Bernd Möbius and Matthias Pätzold. F_0 Synthesis based on a quantitative model of German intonation. In *Proceedings of the International Conference of Spoken Language Processing*, pages 361–364, Banff, Canada, October 1992.

[38] Bernd Möbius, Matthias Pätzold, and Wolfgang Hess. Analysis and synthesis of German F_0 contours by means of Fujisaki's model. *Speech Communication*, 13:53–61, 1993.

[39] Mitsuru Nakai, Harald Singer, Yoshinori Sagisaka, and Hiroshi Shimodaira. Automatic prosodic segmentation by F_0 clustering using superpositional modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 624–627, Detroit, 1995.

[40] Janet Pierrehumbert. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70(4):985–995, October 1981.

[41] John Pitrelli, Mary Beckman, and Julia Hirschberg. Evaluation of the prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference of Spoken Language Processing*, pages 123–126, Yokohama, September 1994.

[42] Scott Prevost and Mark Steedman. Using context to specify intonation in speech synthesis. In *Proceedings of Eurospeech*, pages 2103–2106, Berlin, 1993.

[43] Ken Ross. *Modeling of Intonation for Speech Synthesis.* PhD thesis, Boston University, 1995.

[44] Ken Ross and Mari Ostendorf. A dynamical system model for generating F_0 for synthesis. In *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis*, pages 131–134, New Paltz, New York, September 1994.

[45] Ken Ross and Mari Ostendorf. A dynamical system model for recognizing intonation patterns. In *Proceedings of Eurospeech*, pages 993–996, Madrid, September 1995.

[46] Ken Ross, Mari Ostendorf, and Stefanie Shattuck-Hufnagel. Factors affecting pitch accent placement. In *Proceedings of the International Conference of Spoken Language Processing*, pages 365–368, Banff, Canada, October 1992.

[47] Yoshinori Sagisaka. On the prediction of global F_0 shape for Japanese text-to-speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 325–328, Albuquerque, 1990.

[48] Shinsuke Sakai and Kazunori Muraki. From Interlingua to speech: Generating prosodic information from conceptual representation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 329–332, Albuquerque, 1990.

[49] Naohiro Sakurai, Takemi Mochida, Tetsunori Kobayashi, and Katsuhiko Shirai. Generation of prosody in speech synthesis using large speech database. In *Proceedings of the International Conference of Spoken Language Processing*, pages 747–750, Yokohama, September 1994.

[50] Eric Sanders and Paul Taylor. Using statistical models to predict phrase boundaries for speech synthesis. In *Proceedings of Eurospeech*, pages 1811–1814, Madrid, September 1995.

[51] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn treebank project, March 1991.

[52] Richard Sharman and Jerry Wright. A fast stochastic parser for determining phrase boundaries for text-to-speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 357–360, Atlanta, May 1996.

[53] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference of Spoken Language Processing*, pages 867–870, Banff, Canada, October 1992.

[54] Paul Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15:169–186, 1994.

[55] Paul Taylor. Using nerual networks to locate pitch accents. In *Proceedings of Eurospeech*, pages 1345–1348, Madrid, September 1995.

[56] Louis F. M. ten Bosch. Automatic classification of pitch movements via MLP-based estimation of class probabilities. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 608–611, Detroit, 1995.

[57] Hartmut Traunmüller and Anders Eriksson. The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations. *Journal of the Acoustical Society of America*, 97(3):1905–1915, March 1995.

[58] Michelle Wang and Julia Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196, 1992.

5466 - 6vl