(19)

Recognizing Intonational Patterns in English Speech

by

Erin Marie Panttaja

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 1998

© Massachusetts Institute of Technology 1998. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
May 8, 1998

Certified by .....
Professor Justine Cassell
AT&T Career Development Professor of Media Arts & Sciences
Thesis Supervisor

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

JUL 14 1998
Eng.

Recognizing Intonational Patterns in English Speech

by

Erin Marie Panttaja

Submitted to the Department of Electrical Engineering and Computer Science
on May 8, 1998, 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 parsing of intonation is vital in the interpretation of meaning in human speech. The words a speaker uses may be ambiguous, either because of limitations of transmission (a loud room or imperfect microphone), or due to the fact that some discourse information may be found only in the intonation. The ability to interpret intonational information should enable computer systems to be more responsive to human users. The algorithms described in this thesis differentiate yes/no questions from statements in spoken English. The most effective of these, a slope-based algorithm, can recognize 69% of yes/no questions and 81% of statements correctly.

Thesis Supervisor: Professor Justine Cassell
Title: AT&T Career Development Professor of Media Arts & Sciences

Acknowledgments

Thank you to Justine Cassell, Scott Prevost, Kris Thórisson, and Janet Cahn. Thank you Sola, for injecting a bit more sanity into my life when I needed it, and a bit less when I needed *that*. And to GNL+, especially Deepa. Also thank you to Steve, Seph, Josh, Alex and Joseph for being there when I needed them. And thanks to Mom and Dad for teaching me my first language.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People engaged in spoken discourse exchange information via several different conversational modalities. In addition to words, interlocutors may represent meaning with gestures, facial expressions, and prosodic stresses. These modes can be used in many different ways. [PH90] [McN92] People may use intonation to call attention to particular words by lengthening them or making them louder — people may even negate the meaning of a phrase by shaking their heads or grimacing. They may use a gesture to amplify the meaning of an utterance, for instance to show the manner in which someone left the room, or indicates a particular painting on the wall. In this thesis, I examine the use of intonation in a multimodal system and evaluate several algorithms for extracting intonational patterns from the speech stream.

1.1 Discourse Processing

A primary impetus for the study of computational discourse is the desire to create a truly intuitive computer interface. Most generally available computer systems use text-based or graphical interfaces with keyboard and mouse input. While efficient for experienced users, these interfaces are often cryptic and difficult to learn. On the other hand, most people gain proficiency with conversation long before they use computers, and thus might be more comfortable with — and better capable of understanding information from — a computer that could interact in a more natural conversational

manner. In addition, some circumstances make traditional interfaces such as typing infeasible, for example, user disabilities or features of the working environment. On the other hand, there will always be applications which require the precision and expressive power of computer languages.

Toward this end, many researchers are beginning to examine the use of multimodal conversational agents. These are systems which are capable of communicating with a user via the medium of a character who exists as a picture on the screen. These systems use video and sometimes audio output, and use some combination of video, audio, mouse, keyboard, and motion sensing for input.

These agents extract much of their information by performing Natural Language Processing (NLP) on the speech input from a microphone. There are several commercial systems doing successful speech recognition, such as BBN's Hark system and Dragon's Naturally Speaking. [Bol93] [hd97] However, the field is still open to improvement: current systems look only at the speech stream and ignore the newer fields of gesture recognition and facial expression analysis. [Pie93] Within the speech stream, current systems treat the phonemes with speech recognition and the words with natural language processing but tend to ignore the information contained in the utterance level prosody, the patterns of stress which go with different types of propositional content and different situations.

Prosody provides a great deal of information in a conversational setting. It can give clues to the contextual framework of the utterances and discourse or show the beliefs and goals of a speaker in ways that speech alone cannot. [CP98] Intonation can easily show ambivalence or even disbelief in a statement, and can give gradations of intent which take much longer to express with words alone.

1.2 Goals of this Research

Current speech recognition systems collect the entire acoustical signal. They extract the speech stream, but discard much of the intonational data. Analyzing intonational cues provides information relevant to understanding the discourse. In addition, into-

nation can convey information which is redundant with the speech and can therefore increase the speed of understanding or allow for error correction in a multimodal conversational agent. In general, the addition of intonational parsing can help to make a conversational agent more robust.

A computer should be able to use intonational information in similar ways. In this thesis, I will explore ways in which a conversational agent can get information from intonation without reference to the words used. To narrow the focus of my work, I will restrict the domain of this research to differentiating yes/no questions from propositional statements in English using only intonational information. I will implement an algorithm which takes in an utterance and returns an indication of whether the utterance was a question or statement.

1.3 Evaluation

The two algorithms I developed were evaluated by running them on a dataset of statements and questions from the TRAINS corpus, a collection of natural telephone dialogue by American speakers. [ASF+94] In this study, pairs of subjects discussed plans for shipping boxcars of fruit. I will compare the results of my algorithm to those of having humans transcribers listen to the data and decide whether a particular sentence is a question or a statement. In order to ensure that the computer and speaker are using the same information, the human trials will be done using utterances in which the words have been obscured by using a band-pass filter.

1.4 Outline of this Thesis

In Chapter 1, I have introduced the uses of intonational parsing. Chapter 2 presents an overview of the current state of research on intonation recognition, including the linguistic theory behind prosodic feature recognition, the prosodic labelling itself, and the previous research on automated intonation labelling systems.

In Chapter 3, I discuss the problems and issues my thesis deals with, and in

chapter 4 describe my implementation of possible solutions. In Chapter 5 I discuss my evaluation metrics and testing methodology, as well as my results. Finally, in Chapter 6 I present an evaluation of my research and discuss possibilities for future work.

Chapter 2

Background

Multimodal interaction is a rich topic for research, as it includes such diverse facets of conversation as intonation, speech stream, gesture, and facial expression. Intonation can make significant contributions to the quality of speech interpretation. In this chapter I will discuss some of the work on multimodal characters and how intonation can be used in these systems. In addition, I will discuss intonation in general, including specific patterns found in English intonation, meanings that can be attributed to them, and how they can be recognized.

2.1 Vocabulary

Many terms used in the discussion of intonation in speech are somewhat obscure, or defined differently by different authors. Here I will provide the definitions used in this thesis.

Discourse is the exchange of words and other information between two or more people. It may involve speech, gesture, intonation, written communication, or other communicative media. The time-frame involved may extend from a few moments to a few months, and the participants may or may not be temporally or physically copresent.

A *prosodic phrase* is a cluster of words which illustrate a single idea and contribute meaning to discourse. Prosodic phrases are generally demarcated by intonational

cues. Several prosodic phrases spoken by a single speaker may combine to form an *utterance.* Any utterance can be described in terms of its *propositional content,* or total meaning, apart from the words.

In relation to the propositional content, an utterance may be divided between *theme* and *rheme.* The theme of an utterance is the given topic — this represents the established background assumed by the speaker.[HH76] The rheme is the new and focused information. In Figure 2-1 (from Hiyakumoto, Prevost, and Cassell), "the stupid programmer wrote" is the theme; it recapitulates information from the question. [HPC97] "The slow algorithm" is the rheme, as it represents the salient information in the utterance. In this paper, theme and rheme are discussed as being on the propositional level rather than being on the utterance level.

> I know the SMART programmer wrote the SPEEDY algorithm,
> but WHICH algorithm did the STUPID programmer write?
> (The STUPID programmer wrote) (the SLOW algorithm.)

Figure 2-1: Theme and Rheme

Theme and rheme can be present in the syntax of a sentence, but they are also often indicated by intonation and other prosodic features. The *fundamental frequency* of speech (also called *F0*) is a measure of the underlying pitch contour of an utterance. This is separate from the *intensity,* which is a measure of the energy used in producing the speech. *Duration* describes the lengths of phonemes, particularly of vowels, while *tempo* is a measure of the overall speed of the speech. *Loudness* is the volume of the speech. *Accent* describes how the speaker emphasizes different parts of an utterance. This emphasis may be instantiated by changes in loudness, duration, or by the use of *pitch accents. Pitch accents* are areas of the speech in which the fundamental frequency exhibits particular patterns, generally utterance minima and maxima.

Prosody, according to Kent and Reed, refers to the non-phonetic acoustical properties of speech, such as fundamental frequency, intensity, duration, tempo, and loudness. [KR92] These phenomena are called *suprasegmental* because they can cover units larger than a syllable. A pitch accent, for example, is a change in pitch applied

to the stressed syllable of a word. A *phrase accent* extends over a whole phrase. A *boundary tone* affects the end of a phrase.

Even more complex types of stress are found in the intonational contours of longer utterances. A statement, for example, may contain a *continuation rise* — a high boundary tone indicating that the speaker is not finished.

Intonation is defined by the Oxford English Dictionary as "manner of utterance of the tones of the voice in speaking; modulation of the voice; accent." [Pre] Intonation is is a subset of prosody. It is limited to the frequency, intensity, duration, and stress, without tempo and rhythmic considerations.

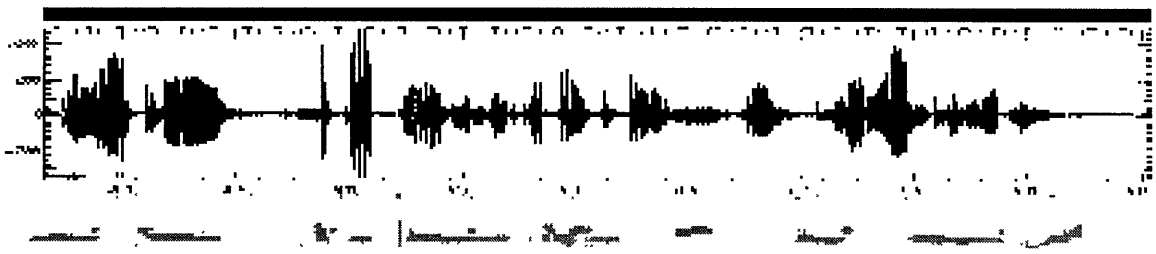

Figure 2-2: "Do you want to hook up the engines to the box cars ... uh ... so one to each?"

The Figure 2-2 shows a question: "Do you want to hook up the engines to the box cars ... uh ... so one to each?" The top portion of the diagram is the speech signal itself. From this it is possible to see pauses and the loudness of the speech. The lower portion shows the fundamental frequency of the utterance, which only appears in voiced segments of the speech. Note that the fundamental frequency rises in the last half second of the utterance, indicating that it is a question.

2.2 What is Intonation?

There are two different varieties of stress: lexical stress and phrasal stress. These two types of stress perform different roles in conversation, and thus each contributes different elements to the meaning of a discourse.

In English, the lexical stress of a word is a part of the word; it does not change

according to the context of the utterance in which it is found. For example, in the word "backgammon," the first syllable is always stressed (/'bak-.gam-*n /). Nearly every word in the English language has a set lexical stress. However, there are a very few cases when the lexical stress of a word may be affected by the words around it. For example, the word "Massachusetts" has its stress on the third syllable (/.mas-(*-)'chu:-s*t /), but in "Massachusetts Legislature," some people put the stress on the first syllable. In spite of these few irregularities, lexical stress is very predictable, and can be collected and stored in dictionaries. [SH92]

The phrasal stress of an utterance, on the other hand, changes depending on the context of the utterance. Pierrehumbert describes pitch accents, phrase accents and boundary tones, which occur within and at the ends of utterances. The meaning of an utterance can be changed by altering these accents and tones.

For example, in American English, sentences which end in a rise in pitch are usually questions. In many cases, this is also clear from the syntactic content of the sentence. The sentence in Figure 2-3 is clearly a question, regardless of the intonation. [Pie80]

Have you finished your thesis yet?

Figure 2-3: Question

In Figure 2-4A, "He went to lab yesterday," is clearly a statement. However, the same words can be uttered as a question with a stress on "lab" and a pitch rise toward the end of the sentence. Try reading the sentence "he went to lab yesterday" in two different contexts.

A. He had lots of work to do.
He went to lab yesterday.

B. It's the day after Thanksgiving.
He went to LAB yesterday?

Figure 2-4: Pitch

The intonation found in the two sentences is different. Given the intonation, the meanings can be differentiated even without looking at the previous sentences, which may provide contextual meaning.

Even a statement may be uttered with different intonations. The intonational changes may not alter the truth value, but will change the emphasis. The intonation used can tell the hearer which information is salient within the discourse, and and which is less relevant.

A. Would you like strawberries or apples?
I want strawberries AND apples.

B. What do you want with your ice cream?
I want STRAWberries and APples.

Figure 2-5: Contrastive stress

In 2-5A, the speaker is emphasizing the contrast between the offer and what she wants. She wants not just one thing, but two. In 2-5B, on the other hand, the speaker is introducing new rhematic information. She is emphasizing the specific fruits she wants by using a pitch accent to ensure that the hearer understands that it is important.

Prosody can also distinguish between two different senses of an utterance.

A. John called Sam a Republican and then he insulted him.

B. John called Sam a Republican and then HE insulted HIM.

Figure 2-6: Examples from *Lakoff 1973*

These sentences contain the same words, but have nearly opposite meanings. The meanings come from the stresses on he and him. In Figure 2-6A, John calls Sam a Republican, then John insults Sam. In Figure 2-6B, however, the emphasis on he and him demonstrates contrastive stress. Thus, in contrast to Figure 2-6A, 'he' refers to Sam, and him refers to John. This equates calling someone a Republican with insulting him.

The prosodic features most commonly analyzed in linguistic data are the duration, the relative loudness, and the pitch relative to the rest of the utterance. They are used because they can by determined computationally and they tend to correlate with features recognized by human hearers. [RP96]

In Scottish English, Brown, Currie, and Kenworthy note that questions on new topics begin with a high pitch, while questions on established topics begin with a low pitch. [BCK80] This enables the hearer to to know when to look for context in the rest of the conversation. In American English, the pitch range generally narrows at the end of the topic, and a new topic may be recognized by a significant expansion in the pitch range.

2.3 Coding Intonation

Pierrehumbert created a system of pitch accents to represent the intonation found in American English. [Pie93] She divide the accents into high and low tones (H and L). These tones are defined with respect to a user's fundamental frequency (f0). They can be combined to make the pitch accents: H*, L*, L*+H, and L+H*. Figure 2-7 from the ToBI training data shows the speech curve, fundamental frequency, and pitch accents associated with "Let's see I need oregano 'n marjoram 'n some fresh basil okay?" [BA97]

Pierrehumbert and Hirschberg classified these intonational patterns associating acoustical properties of an utterance with features of its meaning, such as salience and speaker beliefs. [PH90] In Figure 2-7, the accents on marjoram, oregano, and basil show that the speaker is making a list. From these patterns, one can determine some information about the meaning of a phrase. For example, once a hearer knows the theme and rheme of an utterance, he knows what information the speaker is looking for.

In Figure 2-8A, the stress is on "California." This indicates a contrast between California wines and other wines. In Figure 2-8B, on the other hand, the contrastive stress is on "wines," indicating a preference for California wines over, say, California

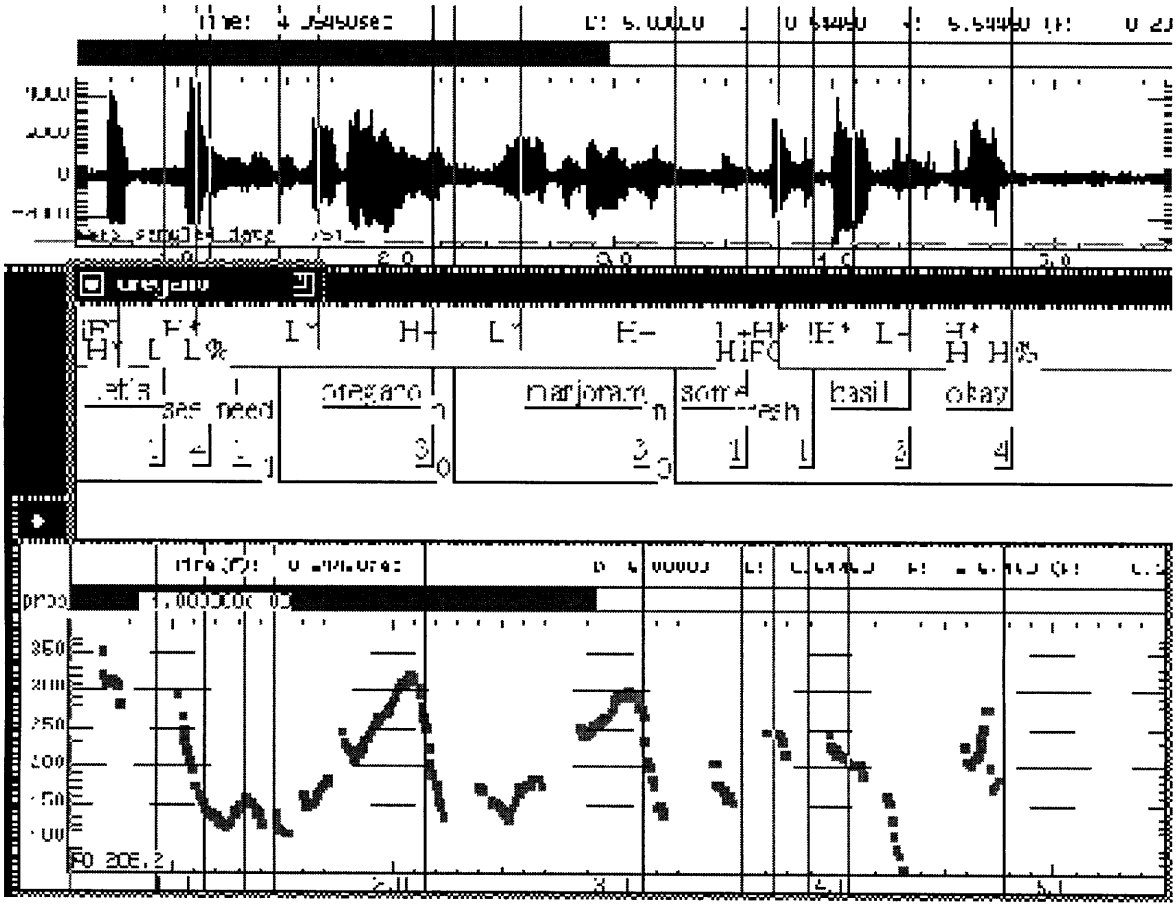

Figure 2-7: "Let's see I need oregano 'n marjoram 'n some fresh basil okay?"

A. I really like CALIFORNIA wines.
B. I really like California WINES.

Figure 2-8: Noun versus Adjective Stress

beers.

Pierrehumbert & Hirshberg's system enables one to make generalizations about the intonation found in an utterance. Without some sort of codified description of the intonation, it can only be described with a list of fundamental frequencies and phonemes. Their more generalized classification of contours allows researchers to investigate similarities in interpretation among instances of each contour.

A high tone is represented by H. An L, on the other hand, represents a low tone and a pitch lower than the average for the utterance. H and L accents may occur either alone or in combinations with other accents; in combinations, an * indicates the accent which maps to the lexical stress of the accented word; the other tone is either a leading or trailing tone, and generally has a shorter duration and less extreme pitch. In Figure 2-7, "marjoram" receives a L* accent, which means there is a strong pitch lowering in its first syllable. The word "fresh," on the other hand, has a pitch level which begins low and ends high, with the stress of the 'e' a high tone.

Pitch accents are also affected by an overall phrase accent, which may be either high or low. A phrase accent takes effect immediately after the pitch accent which precedes it, and lowers or raises the frequency. Several intermediate phrases may be put together to form an intonational phrase which has a boundary tone. The boundary tone describes general slope of the intonation. These two tones are represented by a pair of letters. For instance, a low pitch accent followed by a high boundary tone would be written L-H%. In Figure 2-7, "Let's see" is a subsegment of its own with a phrase accent and boundary tone. (L-L%) [BA97]

Different prosodic contours may give different meanings to an utterance. In general, when the main pitch accent is H*, the accented information is represented as new to the discourse, and is intended to be added to the hearer's picture of the world.

[PH90] A low (L) accent generally discusses information which the hearer should already know.

The terminal contour gives other information about the meaning of an utterance. A steep rise at the end of an utterance (as in L* L-H%) generally indicates a yes-no question. Final lowering often indicates an utterance intended to be informative.

This transcription system has been codified into a standard called ToBI (Tones and Boundary Indices). [BA97] ToBI is a method of transcription which codifies the rules first described by Pierrehumbert in a manner which increases the ability of coders to agree on a single transcription for a given utterance.

2.4 Automatic Recognition of Intonation

Much of the existing analysis of intonation is empirical in nature, and depends on an enormous amount of painstakingly transcribed speech data. In the past few years, researchers have begun to investigate the automatic transcription of words, intonation, and prosodic information from audio recordings. Such transcription could be used to create corpora for further research. Even partial automation would lead to significant time savings in transcription.

As we have discussed, prosodic information highlights the groupings of words in speech and contributes meaning to utterances. Ostendorf uses prosodic constituent structure to do automatic mapping of sound features to syntactic, semantic, and discourse structure. [Ost97] These cues are useful to a computer because the computer system has a much less detailed semantic representation of the information in the utterance than a person would. For example, a decrease in the pitch range used by a speaker generally indicates the end of a topic. Without this prosodic information, the computer is handicapped in trying to interpret the utterance.

Wightman et. al. explored segmental lengthening, which often indicates a prosodic boundary. [WSHOP92] They discovered that lengthening is limited to the syllable before the boundary, and that there are different kinds of boundary tones, which also may be distinguished automatically. Stifelman used only prosodic information to find

the structure in talks. [Sti95] High pitch and energy tend to mark points of emphasis in the speech, as well as segment boundaries.

Pierrehumbert suggests that language and intonational data need to be analyzed together in order to be useful. [Pie93] However, in the Grunt system, researchers used prosodic cues such as pause length and pitch accent to determine when users understood spoken directions. [DS89] Without resorting to words, the system was able to tailor the directions to the hearer by determining when the hearer could not understand the directions from the feedback she gave. Not analyzing the words neglects important information, but also limits the data to be analyzed to a feasible amount. It should eventually be possible to resynchronize the two data streams to match the intonational content with the words. This combined stream could then be fed into a natural language parser.

Ostendorf and Wightman et. al., attempted to label prosodic patterns in a read-speech dataset of radio newscasts. They looked at fundamental frequency, duration, and energy levels in the speech stream, and use them to recognize symbolic phrase boundaries. [WO94] This involves using both a decision tree and a Markov model to find patterns in the data, then using linear superposition to find an intonation boundary. This superposition is important because it deals with the suprasegmental nature of intonational information; it allows for the fact that a pitch accent or boundary tone may change the intonation several words away from the actual accent. Their algorithm works relatively well: in 71% of the cases it agreed with human scorers on the classification of a boundary tone. It detected false pitch prominences in only 2% of the cases, while in 29% of the cases it predicted boundaries, but disagreed with human coders on their classification. This is important because it means the algorithm can detect suprasegmental features. Knowing the segment boundaries can make parsing the speech easier.

Prosodic information may include information about affect as well. Roy and Pentland explored classification of approving and disapproving utterances. [RP96] They achieved recognition approximating that of humans without examining the words (65-85% versus 69-76% for humans)

Computer generated speech can also use prosody to indicate affect. Janet Cahn looked at giving affect data to the user through the use of intonation with a text to speech system. [Cah90] She associated various emotional states with different acoustical correlates. The system was successful in conveying some emotional states.

One difficult aspect of computational language recognition is the inherent ambiguity in spoken language. A given utterance may have several possible interpretations. Any reasonable grammar of the English language contains some ambiguity. Prosodic analysis can cut down on the ambiguity remaining in the system without losing resolution in the number of choices. Impossible options can be discounted, leaving more time to examine possible parses.

A. I saw the man with the telescope.
B. I SAW the man...with the telescope.

Figure 2-9: Theme and Rheme

Figure 2-9A is ambiguous. The phrase "with the telescope" can refer to either "I saw," or "the man." However, with the pause before "with the telescope," the prepositional phrase must be modifying "see."

Intonation can be used in a natural language understanding system to clarify ambiguities in speech recognition data. Automatic interpretation of intonation can help a system to reduce the number of hypotheses it needs to consider. Veilleux proposes two mappings between syntax and prosody, which can be used to remove either word- or sentence-level ambiguities in speech data. [Vei94]

At this point, we know that intonational information is important in interpreting discourse. It can relieve ambiguities and change the propositional content of an utterance. However, it is still not clear how much of this information can be processed in real time from the speech stream. Some of the information is highly subjective, and thus may be difficult to determine by a computer. However, researchers have already managed to automatically recognize boundary data, affect, and prominence data in the speech stream. Through analysis of acoustic correlates, it should be possible to

extract even more meaning.

2.5 Multimodality

Prosodic analysis can be very useful in a multimodal system. Using a combination of different modalities increases the information available to a system. For example, using both keyboard and mouse input to a computer allows a user to use whichever form of input is more appropriate in a given situation. This is also true in a conversational system. Different types of information will be used in different situations. An ability to understand multiple modalities will allow the system to take in more of the data. In addition, in cases in which the information is redundant or partially redundant, the system can compare data from multiple sources to check processing for errors. If information gleaned from two different modalities agrees, it is probably more trustworthy than that for which different modalities conflict.

The first level of prosodic analysis is a simple switch, which tells whether the speaker is speaking or not speaking. This type of analysis is helpful in providing information, not because it generates a new type of information, but because it can provide information faster than the speech recognizer can. [Bol93] This analysis is used in Kris Thórisson's Ýmir architecture, which incorporates a combination of different modalities to create appropriate conversational flow and turn-taking behaviors. [Tho96] Thórisson's implementation, a computer character named Gandalf, talks with a user about the solar system. Gandalf then raises his eyebrows to indicate to the user that she has been heard. Research on Gandalf has shown that users are more willing to endure long delays if they have an indication that they have been heard. This non-verbal response also helps to regulate the conversational flow of interactions, which in turn decreases Gandalf's level of confusion.

Gandalf used prosodic information to determine when the user had stopped speaking. A more sophisticated system could work in concert with a speech recognition system to determine whether an utterance is a question with more certainty, or at least with more speed. In addition, automatic recognition of intonation can help a

computer to understand discourse information such as ends of topics and the speaker's opinion of her words.

Chapter 3

Statement of Problem

Most previous research on intonation has focused on working with read text for the creation of large corpora of intonational data. The work which has been done to add non-lexical intonational data to real-time systems has been very limited.

Intonational information can be used in a real-time multimodal conversational system to help it to converse more smoothly. For example, questions and statements need to be processed entirely differently. A question requires an answer (and often a database search). A statement, on the other hand, implies that an action is required. This may be putting some new information in a database or moving to a new room in a house. Determining which path must be followed as early as possible will save computation in the long run.

This thesis focuses on bridging the gap between the simple speech switch that determines whether the user is speaking and more complicated, but still real-time, analysis. The best evaluation for this is to determine the difference between a yes/no question and a propositional statement. This is one of the simpler distinctions to make, as a yes/no question displays a L* H-H% contour in American English, and a statement will have a H* L-L% contour. It should be possible to recognize this distinction from the speech stream alone. In fact, some utterances of this type cannot be differentiated from the words alone, as seen in Figure 3-1.

Having this information will help a system to respond to a user's wants, as it will know, in this case, whether the user is asserting a fact. This changes what type of

You did your problem set.
You did your problem set?

Figure 3-1: Question

response the system should give.

Chapter 4

Implementation

My goal in this project was to create a system which can differentiate between yes/no questions and statements without looking at word choice or syntax. It needed to work quickly enough to be used in a real-time conversational system for collecting data. I will discuss several different implementations here, beginning with a somewhat more detailed overview of the intonation processing in the Gandalf system. I will then discuss the additions I made to Gandalf as well as the two systems I implemented outside of Gandalf, an average-frequency algorithm and a slope-based algorithm.

There are two primary directions in intonation recognition research. One is the probabilistic model, using Hidden Markov Models, the other is rule based systems. Both have significant strong points. However, a rule based system has the advantage that it can be more easily analyzed to shed light on theories about the ways in which people process language, because the rules can be evaluated one at a time. A Hidden Markov Model, on the other hand, may provide good behavior, but be too complex for decomposition into a theory about the human mind.

Here, I will discuss my initial implementation algorithm, then the two algorithms I tested intensely. All of them are rule-based systems for the detection of intonational patterns without the use of words.

4.1 Gandalf

The first implementation I built was within the existing Ýmir architecture, the underlying architecture used in Gandalf. It processed the speech and sent the current status to the system's main module.

4.1.1 Existing System

The Ýmir intonation code runs under Macintosh Common Lisp on a Quadra. It uses MIDI input to determine whether voiced sound is coming in, and uses this information to report every 30 milliseconds whether the user is speaking.

4.1.2 Improvements

My improvement to this intonation system involved calculating the slope of the intonational data. Every 30 milliseconds the system computes the current slope of the intonation by looking at the pitches of the two most recent samples. It classifies the slope as up, down, or flat, then compares the current slope to the previous slope. By discarding slopes which are not continuous, this system avoids problems with spurious data. If the current slope is different from the previous slope, it sends the previous slope to Gandalf; if the raw data has two slopes in a row which are the same, it sends the new slope. This system does not examine the exact slope; it instead looks at the direction of the slope, with some correction to allow for spurious data.

The intonational data is aligned with the words by assuming that the two data streams start at roughly the same time. From that point, the next time the intonation is turned off should match the end of a phrase. Because the clocks are running at the same rate, this offset can be used to determine the time at which the speech ended, and thus the last word of the utterance. While this does not deal with the issue of unvoiced letters at the beginning or end of an utterance, this should not be a problem in a preliminary implementation. By aligning the two data streams, it is possible to associate a particular intonational pattern with a set of words. This will also enable the system to associate pitch accents with particular words.

4.1.3 Results

While this approach is relatively straightforward, it has the drawback that it requires more computational power than the Gandalf system could handle. Speed issues of the Lisp code and the complexity of the Gandalf system made it impossible to test the algorithm. The primary problem was that the algorithm required more arithmetic operations than the Quadra could handle in real time. For the algorithm to run, it would have had to either run in less than real time or look at the data with less detail.

4.2 Standalone algorithms

The second pair of algorithms I wrote and tested work outside of a full conversational system. They compute the fundamental frequency of an utterance and use that without the words to decide whether an utterance is a yes/no question or a statement. They take in a sound file sampled at a rate of 16 kHz. Both algorithms then run the sound files through some commercial signal processing code to segment remove silence from the ends and compute the fundamental frequency.

4.2.1 Software Overview

These algorithms were written in C on a Silicon Graphics O2. I did much of my processing using Xwaves and ESPS, two programs from Entropic Systems. [Ent97]

Xwaves is a program for graphically displaying and manipulating sound files. I used it primarily as a research tool to aid me in finding patterns in the intonational data.

ESPS is a signal processing tool. I used it directly in my algorithm to transform the simple speech stream into frequency data. In order to remove whitespace from the beginning and end of a segment, I used find_ep, which uses frequency and silence thresholds to determine where the beginning and end of an utterance are. It required extensive modification of the arguments to fine-tune find_ep for the acoustics of the test data. This adjustment was done once for all of the testing and was not redone

for individual speakers.

In order to find the fundamental frequency curves, I used get_f0, which uses a normalized cross correlation function to calculate the fundamental frequency. [Tal95] In addition, it uses dynamic programming processing when deciding whether a given sample is voiced or unvoiced.

4.2.2 Average Frequencies Algorithm

The first algorithm I implemented was one which compared average frequencies for different parts of an utterance. While looking at a small dataset, I discovered that most questions have a higher average frequency at the end than the total average. This follows the patterns discussed in the literature.

This algorithm computes average frequencies for the whole utterance and the last quarter of the utterance. If the average for this last quarter is higher than the total average, the utterance is classified as a question; otherwise, as a statement.

4.2.3 Slopes Algorithm

The second algorithm I implemented fits a line to the fundamental frequency data. If the total slope of the utterance is greater than zero, the utterance is a question. Otherwise, it is a statement. The fact that this algorithm fits the line helps it to work in spite of occasional spurious data. Depending on where the data is sampled, get_f0 can get doubled or halved frequency measurement for periodic data. This spurious data may be a consequence of noise in the recording or an artifact of get_f0.

Chapter 5

Evaluation

I compared the performance of my algorithms to the performance of humans on similar data to evaluate their performance. To determine the information people can glean from intonation alone, I had people listen to utterances without words and determine whether they were statements or questions. In order to test the algorithms, I ran them on similar test data. I then compared the performance of the humans to that of the two algorithms. This enabled me to examine what information is actually in the intonation and what must be found elsewhere.

5.1 Dataset

The TRAINS Spoken Dialog Corpus is data collected in an attempt to create a conversationally adept agent which can help a user to construct a manufacturing and shipping plan for shipping fruit around England. [ASF+94] The set of conversations I used involved ten human speakers in five conversational pairs. While the TRAINS corpus contains overlapping speech and back-channel responses, I used only single utterances as input, as the algorithm is not yet sufficiently robust for overlapped utterances.

I created two sets of questions and statements for use in testing. I used one of them for the design process, and the other for final testing. The design dataset contains 16 yes/no questions and 37 statements. The final test set has 49 yes/no questions and

119 statements. The prevalence of statements in the datasets is due to the fact that they are much more common in the TRAINS corpus, as well as in ordinary speech.

5.2 Human Testing

Human beings are able to get information from the intonation associated with speech. To determine how much information people can extract from yes/no questions using only the speech stream using only intonational data, I had four native speakers of American English listen to a set of 100 sentences. In order to remove syntax and semantics from this evaluation, I ran the speech through a band pass filter. The subjects then determined which utterances were statements and which were yes/no questions.

5.2.1 Procedure

The band-pass filter removed all frequencies below 120 Hz and all frequencies above 600 Hz using a linear-phase finite-impulse response filter generated by ESPS. This made the words unrecognizable, but left the fundamental frequency contour untouched. The few changes which do appear are due to problems with the voicing estimation. ESPS removes f0 information for records which are unvoiced. I played the 100 filtered utterances for all four subjects, and each decided whether each utterance was a question or statement.

Figure 5-1 shows the fundamental frequencies exhibited by first the unfiltered then the filtered data stream. There are some areas in which the fundamental frequency is zero in the original data but not in the filtered data. This is due to oddities in the fundamental frequency detection algorithm which do not affect human ability to hear the frequency. These alterations do not change the overall curve of the fundamental frequency. "Oh it can only pull three loaded boxcars" is a statement, and as such, the slope of the fundamental frequency is flat toward the end.

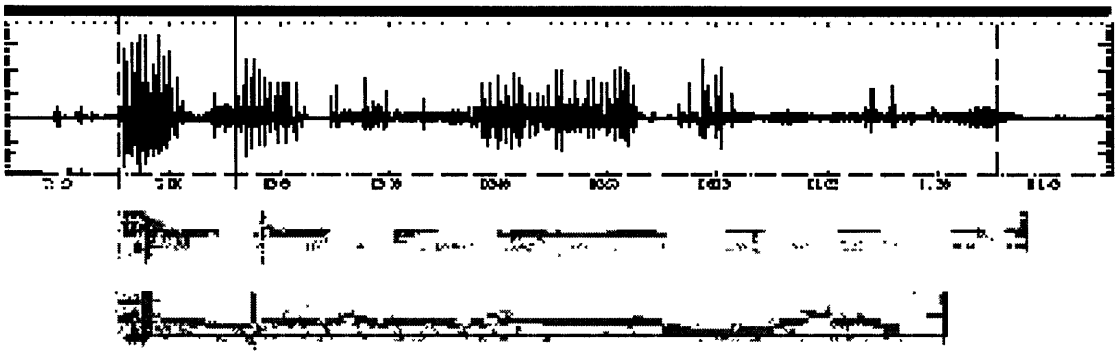

Figure 5-1: "Oh it can only pull three loaded boxcars."

5.2.2 Results

In deciding which utterances were questions and which were statements, I listened to all 100 utterances with the words intact. I classified things as question or statement if both the intonation and syntax agreed. I removed all cases I felt were ambiguous from the data.

	ev. as Y/N	ev. as S
yes/no question	144	36
statement	52	168
yes/no question	80%	20%
statement	24%	76%

Table 5.1: Confusion Matrix for People

Table 5.2.2 shows a confusion matrix for the people's responses. In the matrix, rows represent utterances which I evaluated as yes/no questions and statements, respectively. The columns, on the other hand, represent the filtered utterances as evaluated by the human listeners. Thus the listeners recognized 80% of yes/no questions correctly, and misrecognized 20% of the yes/no questions as statements. The listeners correctly identified 78% of the total utterances in this test. See Appendix A for the actual testing results.

In most cases, the subjects agreed on whether the speech was a statement or yes/no question. In those cases in which the subjects disagreed with my assessment, they generally all had the same conclusion. This disagreement often occurred in cases

in which the utterance exhibited a very small pitch range.

5.3 Algorithms

I tested both of my main algorithms using the initial TRAINS corpus. I used the information from testing to adjust the algorithms for better performance. Only the Slopes Algorithm merited further testing, and I used the final dataset to do further evaluation on it.

5.3.1 Average Frequencies Algorithm

Table 5.3.1 shows the results for the averages algorithm. It recognizes statements very well, but does no better than chance at recognizing questions.

	ev. as Y/N	ev. as S
yes/no question	8	8
statement	1	36
yes/no question	50%	50%
statement	3%	97%

Table 5.2: Confusion Matrix for Average Frequency Algorithm Results on Initial Dataset

This test set shows one of the major problems in the evaluation of this data. Statements are much more common than questions in this sort of data. In spite of the fact that this algorithm did not reasonably evaluate questions, it correctly identified 83% of the utterances.

The primary problem with this algorithm lies with the fact that a quarter of the way through the utterance is a measure which has nothing to do with the structure of the utterance. Arbitrarily choosing a point halfway or a quarter of the way through the utterance ignores its internal structure.

5.3.2 Slope Algorithm

The slope algorithm performed much better. It was able to recognize correctly 69% of the questions and 89% of the statements. (see Table 5.3.2) It correctly identified 83% of the utterances, as did the averages algorithm. However, it did significantly better than the 70% which could be produced by applying an algorithm that declares everything to be a statement.

	ev. as Y/N	ev. as S
yes/no question	11	5
statement	4	33
yes/no question	69%	31%
statement	11%	89%

Table 5.3: Confusion Matrix for Slope Algorithm Results on Initial Dataset

The final dataset included more test cases than the initial, and was not used in adjusting the algorithms. The Slope Algorithm results on the final dataset, (Table 5.3.2), are not quite as impressive, but are still statistically significant. The algorithm continued to correctly evaluate 69% of the questions. In this case, it evaluated 81% of the statements as such. It correctly evaluated 77% of the data. This is nearly as effective as the human listeners.

	ev. as Y/N	ev. as S
yes/no question	31	14
statement	23	96
yes/no question	69%	31%
statement	19%	81%

Table 5.4: Confusion Matrix for Slope Results on Final Dataset

5.4 What worked

People were able to classify most sentences correctly. This indicates that even without the words, the intonation contains enough information to differentiate between these two contours. People listen to changes in fundamental frequency and pitch level to

tell the difference between questions and statements. This is an important realization in the analysis of the content available solely from intonation, as it means intonational analysis can reasonably be done in parallel with speech recognition.

5.4.1 Average Frequency Algorithm

With very preliminary data (20 test sentences), the average frequency algorithm produced the correct result for over 90% of the test cases. With the final dataset, however, the results were much worse. I suspect this is due at least partially to the fact that the initial test set was of read text, and the final test set was not. Read speech generally has a clearer fundamental frequency pattern. Natural speech (as in the TRAINS corpus), on the other hand, contains many false starts. These will significantly alter the intonational contour, as they may change the direction of the utterence.

5.4.2 Slopes Algorithm

The slope algorithm showed very uneven performance. It was very good at predicting questions, but tended to overpredict them, particularly in the final data set.

5.5 What didn't work

Both people and the algorithm had problems with utterances which were in a monotone or very low voice. This is because in these cases, the slope changes were very limited, and difficult to detect. In addition, some questions are recognizable as such by syntax, but lack any sort of final rise. Both people and the algorithms had problems with these utterances.

There were several attempts at algorithms which should have improved performance, but did not. For the Average Slopes algorithm, I experimented with comparing slopes for different segments of the utterance. Looking at the last quarter of the speech worked the best. This algorithm should work better if it took into account some of the structure of an utterance. One good augmentation would be to look for

the points of maximum change in fundamental frequency, and use the last of them as a border. It could then compare the average frequency on either side of the border point.

Another version of the Slopes Algorithm computed the slope of the last segment of the utterance. This worked less well than computing the slope of the entire utterance. This is because the boundary tone takes effect from the beginning of the last pitch accent. An algorithm which found the last pitch accent, then computed the slope from that point to the end should be more effective, though significantly more effective. In addition, an algorithm could compute slopes for several different segments, then determine which is the most optimal fit.

In studying intonational contours, it is clear that the question contour can be represented by a curve more complex than a line. However, such an algorithm would need to be significantly more sophisticated than those presented here.

Chapter 6

Conclusions and Future Research

In this thesis, I examined the use of intonational cues in a multimodal system, and the application of several algorithms for extracting the intonational data. Intonation can be used to give a system more information about the user of a system, and may be able to increase the speed of a system. An algorithm which looks only at the slope of the fundamental frequency can correctly differentiate between yes/no questions and statements 77% of the time.

Future research should explore other information which can be gleaned from the speech stream, and work on integrating prosodic information with speech recognition.

Another consideration is the addition of a probabilistic HMM to an analytic algorithm. While this would remove some of the linguistic research benefits gained from the rule based approach, it may increase the performance of this system. A judicious combination of probabilistic and rule-based systems is likely to be the best approach. This combination could be created by using probabilities in the triggering of the rules, which would make it easier to combine different sources of data.

In addition, future research should work toward creating algorithms that can be incorporated into a larger, multimodal agent in order to evaluate the effectiveness of intonational data to enhance language recognition and to assist the agent in generating natural conversational behaviors. Both propositional and nonpropositional data should be useful in improving the overall interaction.

Appendix A

Results of Testing on Humans

This Table contains the results of the testing on human subjects. For each utterance, this table contains the actual classification, the classifications by each of the four subjects, and the classification given by the slopes algorithm.

utterance	actual	s1	s2	s3	s4	slopes
1	Q	Q	Q	S	Q	Q
2	S	S	Q	S	S	S
3	Q	Q	S	Q	Q	Q
4	S	Q	S	S	S	S
5	S	S	S	S	S	Q
6	Q	S	S	S	S	Q
7	S	Q	Q	Q	Q	Q
8	S	Q	S	Q	S	S
9	S	S	S	S	S	S
10	S	S	S	Q	S	S

utterance	actual	s1	s2	s3	s4	slopes
11	Q	Q	Q	Q	Q	S
12	Q	Q	Q	Q	Q	Q
13	S	S	S	Q	S	S
14	S	S	Q	S	S	S
15	S	S	S	S	S	Q
16	Q	Q	Q	Q	Q	Q
17	S	Q	S	Q	S	S
18	Q	Q	Q	Q	S	Q
19	Q	Q	Q	Q	Q	S
20	S	S	S	S	S	S
21	S	Q	S	S	S	Q
22	Q	Q	Q	Q	S	S
23	S	S	S	Q	S	S
24	S	Q	S	Q	S	S
25	S	S	S	S	S	S
26	S	S	Q	S	S	S
27	S	S	Q	S	S	S
28	S	Q	Q	Q	Q	S
29	S	S	Q	Q	Q	S
30	S	S	Q	Q	S	S
31	S	Q	S	Q	S	S
32	S	S	S	S	S	S
33	S	S	S	S	S	S
34	S	S	S	S	S	S
35	S	S	S	S	S	S

utterance	actual	s1	s2	s3	s4	slopes
36	S	S	S	S	S	S
37	S	S	S	S	S	S
38	S	Q	Q	Q	Q	S
39	S	S	S	S	S	S
40	Q	Q	Q	Q	Q	S
41	S	Q	S	S	S	S
42	S	Q	Q	Q	Q	S
43	S	S	S	S	S	S
44	S	S	S	S	S	S
45	S	Q	Q	Q	Q	Q
46	S	S	S	S	S	S
47	S	S	S	S	S	S
48	S	Q	S	S	S	S
49	S	S	S	S	S	S
50	Q	Q	Q	Q	Q	S
51	Q	Q	Q	Q	Q	Q
52	Q	Q	Q	Q	Q	Q
53	S	S	S	Q	S	S
54	Q	Q	S	Q	Q	S
55	Q	Q	Q	Q	Q	S
56	S	S	S	S	S	S
57	Q	Q	Q	Q	Q	S
58	Q	S	S	Q	Q	Q
59	S	Q	S	Q	S	S
60	Q	Q	S	Q	Q	S

utterance	actual	s1	s2	s3	s4	slopes
61	S	S	Q	S	S	S
62	Q	Q	Q	Q	Q	S
63	S	S	S	S	S	Q
64	Q	Q	Q	Q	Q	Q
65	Q	S	S	S	S	S
66	Q	Q	Q	S	S	S
67	Q	Q	Q	Q	Q	Q
68	S	S	S	S	S	Q
69	Q	Q	S	Q	S	Q
70	Q	Q	Q	Q	Q	Q
71	S	Q	Q	Q	Q	S
72	S	S	S	S	S	Q
73	Q	Q	Q	Q	Q	S
74	Q	Q	S	Q	Q	Q
75	Q	S	Q	Q	Q	S
76	Q	Q	Q	Q	S	Q
77	Q	Q	Q	Q	S	S
78	Q	Q	Q	Q	S	Q
79	S	S	S	S	S	Q
80	Q	Q	Q	Q	Q	Q
81	Q	Q	Q	Q	Q	Q
82	S	S	S	S	S	S
83	Q	Q	Q	Q	S	S
84	S	S	S	S	S	S
85	Q	S	S	S	S	S

utterance	actual	s1	s2	s3	s4	slopes
86	S	S	S	S	S	S
87	Q	Q	Q	Q	Q	S
88	Q	Q	Q	Q	Q	Q
89	Q	Q	Q	Q	Q	Q
90	Q	Q	Q	Q	Q	Q
91	S	S	S	Q	S	S
92	Q	Q	Q	Q	Q	Q
93	Q	Q	Q	Q	S	Q
94	Q	Q	Q	Q	Q	Q
95	Q	Q	Q	Q	Q	Q
96	Q	Q	Q	Q	S	Q
97	Q	S	S	S	S	S
98	S	S	Q	S	S	S
99	S	S	S	S	S	S
100	S	S	S	S	S	S

Bibliography

[All87] J. Allen. *Natural Language Understanding.* Benjamin/Cummings Publishing Co. Inc., Reading, Massachusetts, 1987.

[ASF+94] James F. Allen, Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Traum. The trains project: A case study in building a conversational planning agent. Technical report, University of Rochester Computer Science, 1994.

[BA97] Mary E. Beckman and Gayle A. Ayers. Guidelines to tobi labeling. http://ling.ohiostate.edu/Phonetics/E_ToBI/etobi_homepage.html, March 1997.

[BCK80] Gillian Brown, Karen L. Currie, and Joanne Kenworthy. *Questions of Intonation.* Croom Helm Linguistics Series. Croom Helm, London, 1980.

[Bol72] Dwight Bolinger. Accent is predictable (if you're a mind reader). *Language*, 48:633–644, 1972.

[Bol89] Dwight Bolinger. *Intonation and its Uses.* Stanford University Press, Stanford, California, 1989.

[Bol93] Bolt, Beranek & Newman, Inc. Speech and Natural Language Processing Department. *BBN HARK Recognizer Release 1.1 Beta*, document number 100-1.1 edition, 1993.

[Cah90] Janet Cahn. Generating expression in synthesized speech. techreport, Massachusets Institute of Technology Media Laboratory, 1990.

[CP98] Justine Cassell and Scott Prevost. Embodied natural language generation: A framework for the generation of speech and gesture. *unpublished*, 1998.

[Cru86] A. Cruttenden. *Intonation.* Stanford University Press, Stanford, California, 1986.

[DS89] Jim Davis and Christopher Schmandt. The back seat driver: Real-time spoken driving instructions. *Proceedings of IEEE Vehicle Navigation and Information Systems Conference*, pages 146–150, September 1989.

[Ent97] Entropic Research Laboratory, Inc. *ESPS/waves+ with EnSig 5.2*, 5.2 edition, 1997.

[hd97] http://www.dragon dictate.com/. SMI voice recognition experts. webpage, December 1997.

[HH76] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English.* Longman, 1976.

[HPC97] Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. Semantic and discourse information for text-to-speech intonation. *unpublished*, 1997.

[KR92] Ray D. Kent and Charles Read. *The Acoustic Analysis of Speech.* Singular Publishing Group, Inc., San Diego, California, 1992.

[Lak73] Robin Lakoff. Questionable answers and answerable questions. In Braj B Kachru et al., editor, *Issues in Linguistics: Papers in Honor of Henre and René Kahane*, pages 453–67. University of Illinois Press, Urbana, Illinois, 1973.

[McN92] David McNeill. *Hand and Mind.* Chicago: University of Chicago Press, 1992.

[Ost97] M. Ostendorf. Prosodic boundary detection. *unpublished*, 1997.

[PH90] Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In Cohen, Morgan, and Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusets, 1990.

[Pie80] Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD dissertation, Massachusets Institute of Technology, 1980.

[Pie93] Janet Pierrehumbert. Prosody, intonation, and speech technology. In Madeline Bates and Ralph Weischedel, editors, *Challenges in natural language processing.* Cambridge University Press, 1993.

[Pre] Oxford University Press, editor. *Oxford English Dictionary.* Oxford University Press, second edition.

[Pri81] Ellen Prince. Toward a taxonomy of given-new information. In Cole, editor, *Radical Pragmatics.* 1981.

[PS94] Scott Prevost and Mark Steedman. Specifying intonation from context for speech synthesis. In *Speech Commuinication*, pages 139–153, 1994.

[RP96] Deb Roy and Alex Pentland. Automatic spoken affect analysis and classification. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1996.

[SH92] S. Shattuck-Hufnagel. Stress shift as pitch accent placement. In *Proceedings of the International Conference on Spoken Language Processing*, 1992.

[Sti95] Lisa Stifelman. A discourse analysis of structured speech. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation.* Stanford University, March 1995.

[Tal95] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W.B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis.* New York:Elsevier, 1995.

[Tho96] Kristinn Thorrison. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills.* PhD dissertation, Massachusets Institute of Technology Media Laboratory, July 1996.

[Vei94] Nanette Marie Veilleux. *Computational Models of the Prosody/Syntax Mapping for Spoken Language Systems.* PhD dissertation, Boston University College of Engineering, 1994.

[Wai88] A. Waibel. *Prosody and Speech Recognition.* Morgan Kaufman, 1988.

[WO94] C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Trans. Speech and Audio Processing*, 2(4):469–481, October 1994.

[WSHOP92] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Accoustical Society of America*, March 1992.