Using Prosody in ASR: the Segmentation of Broadcast Radio News

Sasha Calhoun

Supervisors: Mark Steedman and Stephen Isard

External Advisor: Helen Wright Hastie

Master of Science

in

Speech and Language Processing

Department of Theoretical and Applied Linguistics

University of Edinburgh

2002

Abstract

This study explores how prosodic information can be used in Automatic Speech Recognition (ASR). A system was built which automatically identifies topic boundaries in a corpus of broadcast radio news. We evaluate the effectiveness of different types of features, including textual, durational, F0, Tilt and ToBI features in that system. These features were suggested by a review of the literature on how topic structure is indicated by humans and recognised by both humans and machines from both a linguistic and natural language processing standpoint. In particular, we investigate whether acoustic cues to prosodic information can be used directly to indicate topic structure, or whether it is better to derive discourse structure from intonational events, such as ToBI events, in a manner suggested by Steedman's (2000) theory, among others.

It was found that the global F0 properties of an utterance (mean and maximum F0) and textual features (based on Hearst's (1997) lexical scores and cue phrases) were effective in recognising topic boundaries on their own whereas all other features investigated were not. Performance using Tilt and ToBI features was disappointing, although this could have been because of inaccuracies in estimating these parameters. We suggest that different acoustic cues to prosody are more effective in recognising discourse information at certain levels of discourse structure than others. The identification of higher level structure is informed by the properties of lower level structure. Although the findings of this study were not conclusive on this issue, we propose that prosody in ASR and synthesis should be represented in terms of the intonational events relevant to each level of discourse structure. Further, at the level of topic structure, a taxonomy of events is needed to describe the global F0 properties of each utterance that makes up that structure.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Sasha Calhoun*)

Acknowledgements

Thanks go firstly to Mark Steedman for inspiring my interest in this area and for his guidance during the project. Thank you also to Steve Isard for his gentle helpfulness, patience and useful comments. I am very grateful to Helen Wright Hastie for the kind support and practical input she has given to a student she has never met.

I am forever in debt to Simon King for his exemplary guidance, practical support and genius both in this project and throughout the rest of the MSc course. I would also have been lost without Michael Bennett to turn to for every tricky problem in the lab, as well as Cassie Mayo for her encouragement and Latex skills and the rest of the Common Room crew for answering various anomalous queries.

Many thanks finally to my parents for being there for me as always and for their proof-reading skills. And of course to the rest of my fantastic coursemates, particularly Sarah, who bizarrely made the lab an enjoyable place to be and who have kept me relatively sane this year.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

As our own experience and the wealth of material on the subject attests, prosody is vital to the full communication of everything we say, yet it has also proved one of the most intractable sources of information in the speech signal. Prosody is used to augment or alter the meaning of the words we use in the spoken signal from the word level itself, to the phrase level, right up to the level of global discourse structure. It is used to convey illocutionary force and the emotive content.

In order to really advance the state-of-the-art in intelligent Automatic Speech Recognition (ASR) and synthesis systems therefore, a reliable system for recognising and interpreting this prosodic information in the speech signal is needed. This is particularly true at the level above the word. In this study, we have chosen to investigate how acoustic cues can be used to recognise discourse information contained in prosody at the level of topic structure. We will build a system which can automatically segment radio news broadcasts into topics. We are interested in this level of discourse structure because it is relatively unstudied in the ASR literature and because it allows us to see the extent to which different acoustic cues contribute to the identification of different levels of discourse structure.

It is hoped that the results of this study may also be able to inform us on how humans identify topic structure in the speech signal. If it is true that if a certain acoustic feature is reliable in doing automatic recognition at a certain level of discourse structure, then humans also use that cue to identify that level of discourse structure, then we will have learnt something about human speech production and recognition. Similarly,

if a certain abstract representation of prosodic information is effective in recognising discourse automatically, then it is probable that that representation gives an insight into how humans process prosodic information.

We will begin by defining discourse structure and topic structure, two crucial concepts in this work. We will then go into reasons why we are interested in the topic segmentation task. Finally, we will outline the structure of this report in the following chapters.

1.1 Discourse Structure and Topic Structure

Throughout this work we will frequently be referring to discourse structure and topic structure. Given the relatively straight-forward structure of the texts in the corpus we will be using, it is not necessary to fully discuss the literature regarding the nature of discourse and topic structure. Instead, we will give a working definition sufficient for our purposes here broadly based on Grosz & Sidner's (1986) discourse structure theory.

According to this theory, discourse structure is hierarchical and composed of three separate but interrelated components: linguistic structure, intentional structure and attentional state. The linguistic structure describes the way linguistically definable units (in syntactic or phonological terms) naturally aggregate within the text. When we talk about finding levels of discourse structure, we mean recognising these linguistic units. These units aggregate together on different levels, successively forming larger units, from the word level, to the phrase level to the sentence level, to the topic level.

The intentional structure describes the structure of the text in terms of the purpose of each linguistic unit in the whole discourse. The theory goes that the natural aggregation of these units is motivated by their discourse purpose. Again, the intentional structure can be determined on multiple levels of the discourse structure. We are particularly interested in topic structure, an intentional structure unit, i.e. the segmentation of the text units based on what topic each part of the text is about.

It is very hard to come up with a full-proof definition of 'topic'. As it pointed out by Spark Jones (1999) in her review paper of automatic summarisation, what you view as

the topic of a stretch of discourse, and therefore how you segment discourse, is very subjective and depends on your purposes in summarising (or segmenting) the text. Fortunately, in the type of short-news item radio broadcast news texts we are looking at, topic boundaries can be identified in a straightforward way. The text is fairly clearly divided into items about completely different topics, as you can see in the following example taken from our corpus where <TOPIC> labels show topic boundaries and <S> labels show sentence boundaries (see Appendix A for the full text):

(1.1)
```
<TOPIC><S> Massachusetts Chancellor of Higher Education
 Franklin Jennifer is calling for a seven point seven percent
 tuition increase at state colleges and universities . brth </S>
<S> The increase would cost students between sixty and one
 hundred forty dollars a year . </S></TOPIC>
<TOPIC><S> Governor Dukakis met with environmentalists today ,
 who gathered  at the State House to push for open space
 legislation . brth </S>
<S> WBUR's David Barron reports . </S></TOPIC>
```

Indeed, a primary reason for choosing to work with these texts is so that this issue does not cloud the topic segmentation results.

The final component of Grosz & Sidner's (1986) discourse structure theory, attentional state, mostly relates to the focus of attention among the participants of a discourse in dialogue situations. Since we are only concerned with monologue texts, this is not so relevant here.

1.2 Applications of Topic Segmentation

Why are we interested in the problem of topic segmentation? It is interesting from two perspectives. The task itself is an important first step in many other forms of information extraction. Secondly, linguistic theory on the intonational marking of topic structure, on which ASR systems could be based, is currently reasonably underdeveloped and empirically untested.

After having identified words and sentences, topic segmentation represents the next major level of grouping in informative texts such as broadcast radio news, TV news or newspaper articles. In order to do topic identification (e.g. for the purpose of keyword searching or database creation), and automatic summarisation, it is first necessary to identify topic blocks within the texts. Although much work has been done on this problem for written texts, it is only in the past few years that a significant amount of effort has been made on the segmentation of spoken texts into topics. Spoken texts present unique challenges, as there is no formatting information such as paragraph breaks and headings; and there is less use of cue phrases, such as *notwithstanding* to indicate topic structure. There is, however, significant scope to see what use can be made of prosodic information - the spoken equivalent of such written cues - in the segmentation of spoken texts.

ASR systems have now become reasonably effective at identifying words on the basis of the acoustic signal, and achieve reasonable performance at identifying sentences. However, at the topic level, there has been much less success in accurately identifying boundaries. There are at least two reasons for this. The problem of topic segmentation is more complex than that of word or sentence segmentation. This is evidenced by the fact that human annotators find it hard to agree on topic boundaries whereas this is much less true for words and sentences. Therefore, it is much more important for the automatic recognition process to be guided by a substantiated linguistic theory. However, at present linguistic theory in this area is not fully developed and tested. To this end, the present study tests existing theories about the nature of intonational marking of topic structure in an empirical ASR study.

1.3 Overview

In Chapter 2, we examine the literature relating to the segmentation of discourse structure into topics using prosodic information and the prosodic marking of topics in discourse. We begin by reviewing the acoustic cues which are commonly used to characterise such prosodic information both in human perception and production and in automatic speech recognition and synthesis. We then review work in this area which is analysed as coming from two perspectives. The first group of studies proceed on the basis that information about topic structure can be derived directly by measuring

acoustic cues, specifically F0 levels, amplitude and durational features, in the speech signal. Evidence is presented that shows humans mark and perceive topics and topic boundaries using these cues. The second group of studies form part of a line of work which claims that intonation should be represented by a series of intonational events, specifically ToBI features. We present Steedman's (2000) theory, which says that the information status of events within a discourse, and therefore the discourse structure itself, can be derived from these intonational events. The contrast between these two approaches to the extraction of discourse information from prosodic information is a theme which recurs often in this work. Lastly, we explain the Tilt intonational event system, which allows intonational events to be identified automatically from the speech signal, and was therefore needed for this study as there was not sufficient ToBI annotated data available.

In Chapter 3 we present the task, topic segmentation, which will be used to test the effectiveness of the theories about intonational information which will have been established in Chapter 2. We review previous systems which have been built to automatically segment spoken texts into topics; and present various textual, as well as prosodic, cues which have been used to aid in this task. The purpose of this is two-fold. Firstly, we want to establish how useful intonational information is in such systems, i.e. is it worthwhile in terms of performance gain to include this intonation information, and if so, which acoustic indicators are most effective for this task? Secondly, relating back to the discussion in Chapter 2, we wish to compare the performance with textual, F0 and ToBI features to try to gain insight into how much information, if any, humans get about topic structure from these sources of information. In other words, what is the information cross-over between textual and acoustic information? Having established that, we can try to identify what other information the human listener uses to be able to do topic segmentation so effectively. Finally, we briefly explain the machine learning algorithms which will be used to create the topic segmentation system in Chapter 4.

Chapter 4 sets out the topic segmentation experiment that was carried out as suggested by the literature in Chapters 2 and 3. The aim was to assess the effectiveness of the various textual, acoustic and ToBI/Tilt features in identifying topic boundaries in the Boston University Radio News Corpus. The motivation for using this corpus, and its properties, are explained. We then detail how each of the established features was extracted from the spoken text in the corpus. The results of the experiment are presented. We outline findings both in terms of performance when using each of the different fea-

tures, and in terms of which features were most important with each of the statistical classifiers when all the features were included. We go on to look at performance when using different methods to extract ToBI features from the information in the corpus. Finally, we evaluate the results in terms of the aims of the topic segmentation task suggested in Chapter 3. We compare the performance of the system to previous attempts at this exercise, and then look at results in terms of the specific hypotheses laid out at the beginning of the chapter. We also briefly contrast the performance of the two machine learning algorithms used.

Chapter 5 continues the theoretical discussion of Chapter 2 in light of the experimental findings in Chapter 4. We consider the related issues of the most effective way to represent prosodic information in ASR systems; and how humans represent prosodic information. We begin by looking at which acoustic cues were most effective in determining topic structure and therefore to what extent such cues are reliable as direct indicators of topic structure. We go on to look at the performance of ToBI features in the system, and what this shows about the advantages and disadvantages of representing prosody in terms of intonational events like ToBI in ASR systems. Finally, we sketch out brief proposals about what a theory of prosodic marking of discourse structure should look like to be useful for ASR, keeping in mind the plausibility of such a theory in terms of human production and perception.

Finally, in Chapter 6 we summarise the major findings of the experiment and relate these to the conclusions of the discussion about the representation of intonational information.

Chapter 2

Intonation Theory

We will begin by reviewing current literature that relates to how humans mark topic structure with intonation. This will be used to inform the automatic topic segmentation system which will be discussed in the later chapters. It seems legitimate to begin with human processing of topic structure, as humans are the best known segmenters of topic structure.

Firstly, we must specify exactly what we mean by prosody. As we are interested in automatically deriving prosodic features from the speech signal, we will focus on prosodic features which have been found to have easily quantifiable acoustic correlates. We will discuss how these can be derived from the speech signal.

Once these acoustic correlates have been established, we can look at how they are used to convey discourse structure, in particular topic structure (see the Introduction for definition). Literature in this area can be divided into two groups. The first, mainly psycholinguistic or phonetic production studies, look at the extent to which different acoustic correlates of prosody are used to mark discourse structure (for review see Cutler, Dahan & van Donselaar 1997). Hence they come from the perspective that prosodic features of speech are a direct realisation of the intended discourse structure. They tend to concentrate on acoustic features which can be readily extracted and measured from the speech signal. The second group, coming from a more theoretical linguistic perspective, see prosodic events as having an abstract structure in their own right (Pierrehumbert & Hirschberg 1990, Silverman, Beckman, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg 1992); or, in the case of Steedman's Information

Structure theory (Steedman 2000), see abstract prosodic events as forming part of the abstract discourse structure of a text.

We will review the literature from both perspectives always mindful of how the experimental evidence presented or theories proposed can be used to automatically detect topic boundaries in a text on the basis of acoustic cues. To this extent, we will examine how useful each view of prosody is for the ASR community in terms of the trade off between being able to reliably extract the required information from the speech signal and capturing the true richness of meaning that prosody provides.

One of the major stumbling blocks of the integration of prosodic information in ASR systems has been the difficulty in recognising abstract intonational events in the speech signal. From this perspective, we will introduce Taylor's (2000) Tilt intonation theory, which encodes much of the same information as Steedman's theory, but can be derived automatically from the acoustic signal.

2.1 Prosody and its Acoustic Correlates

In their review of the role of prosody in sentence processing, Shattuck-Hufnagel & Turk (1996, p.196) provide a useful working definition of prosody:

> "we specify prosody as both (1) acoustic patterns of F0, duration, amplitude, spectral tilt, and segmental reduction, and their articulatory correlates, that can be best accounted for by reference to higher-level structures, and (2) the higher-level structures that best account for these patterns ... it is 'the organizational structure of speech'."

When attempting to use prosody in the processing of a speech signal, we tend to concentrate on the first part of the definition above and equate prosody with its acoustic correlates. However, it is important to remember the second part of the first definition: any one acoustic measure is influenced by many things, from the particular speech segment we are dealing with, to the global discourse structure, to the characteristics of the speaker.

That said, let us review the acoustic correlates of prosody cited in the literature. The most commonly discussed is fundamental frequency (F0). Speakers manipulate the

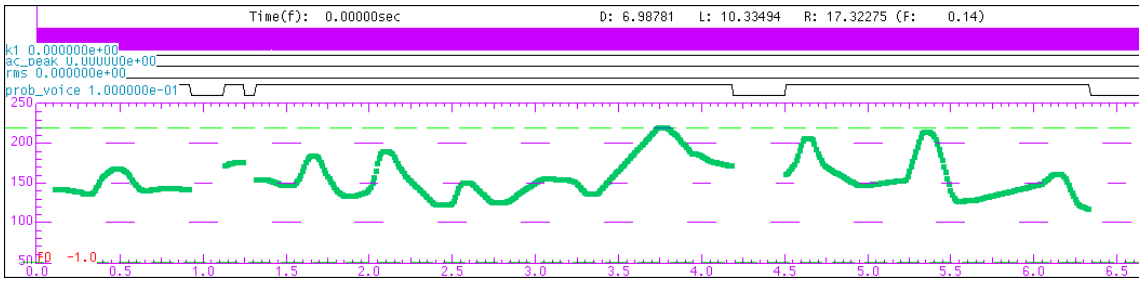

Figure 2.1: Example of F0 contour:
'Paula Gold says Alfa Romeo led the pack and was followed by Peugeot, Yugo Sterling and Range Rover'.

speed at which their vocal folds vibrate to convey a higher or lower pitch to their listeners. As is discussed by Warren (1999, p.157), F0 cannot be taken as an absolute indication of intended pitch. Different speakers have different F0 ranges, most obviously the female F0 range is much higher than the male. Our perceptual systems account for this. Secondly, our perception of the F0 range of one speaker is not linear, i.e. we do not perceive a doubling of F0 as a doubling in pitch. Thirdly, it is not correct to directly associate higher F0 with greater prominence, as speakers accord greater significance to F0 movement in the perception of prominence (Fry 1958). Lastly, F0 is affected by whether or not the segment in question is voiced, and by whether the preceding or following segments are voiced. Bearing all this in mind, F0 is a useful indicator of prosodic structure. This is not least because there are many commercially available programs to extract the F0 contour from an auditory speech signal.

The extracted F0 contour is generally used in two ways. Either, the shape of the contour over a sentence or paragraph is measured, or particular features - peaks or troughs - in the contour are picked out. Figure 2.1 shows the F0 contour of a sentence from our corpus. Distinct peaks and troughs can clearly be seen, as can breaks in the F0 contour in the unvoiced segments of the word *says* near the beginning of the sentence.

The second important acoustic factor reported in studies into prosody is duration, reported in milliseconds of segment or syllable durations. Evidently, this is again affected by the segment involved. Some sounds take longer than others to say and some, e.g. nasals as opposed to plosives, are more prone to lengthening than others, i.e. their

duration varies more. Also, the amount of the change in duration is affected by the overall speaking rate. Given that one can only compare similar segments, it can be difficult to use segmental durational differences in automatic recognition, not least because of the problem of exactly determining the boundaries of the segments in the first place. A simpler indicator can be pause duration, including filled pauses, such as *um* or *er*. The number and duration of pauses, however, decrease as speaking rate increases, so pause duration cannot be taken as an absolute measure across different speakers in different situations.

The final acoustic measure which is commonly used in the automatic recognition of prosodic structure is amplitude, the amount of energy present in a sound or sequence of sounds. Again, there is not a direct correspondence between amplitude and what we perceive as loudness. As Warren (1999) notes, open vowels typically have a greater amplitude than close vowels, although they are not perceived as such.

Shattuck-Hufnagel & Turk (1996) note that spectral tilt and segmental reduction, e.g. a difference in vowel quality, can be acoustic correlates of prosodic structure. Fougeron & Keating (1997) also show that prosodic boundaries can be marked by articulatory strengthening. However, to my knowledge there has been no successful attempt to automatically extract these sorts of acoustic features from a speech signal. We will therefore lay these acoustic correlates of prosody aside.

2.2 Studies of Prosody and Discourse

So what exactly is it that the acoustic indicators mentioned above are said to be marking? Although the studies below differ somewhat in the theoretical standpoint that they come from, it is generally agreed that speech can be divided into a hierarchical structure of nested groupings, from the prosodic foot to the topic boundary. Two such proposals are shown in Figure 2.2.

The exact nature of the various intonational groupings is in dispute, as is the regulation that the intonational hierarchy should be strictly nesting (see discussion in Ladd 1996, chap.6). However, for the purposes of the present study, we can say that there is an intonational phrasing that aligns more or less with the syntactic sentence,[1] and at a higher

[1]or full clause, see Shattuck-Hufnagel & Turk (1996) for discussion. In this study, we will be

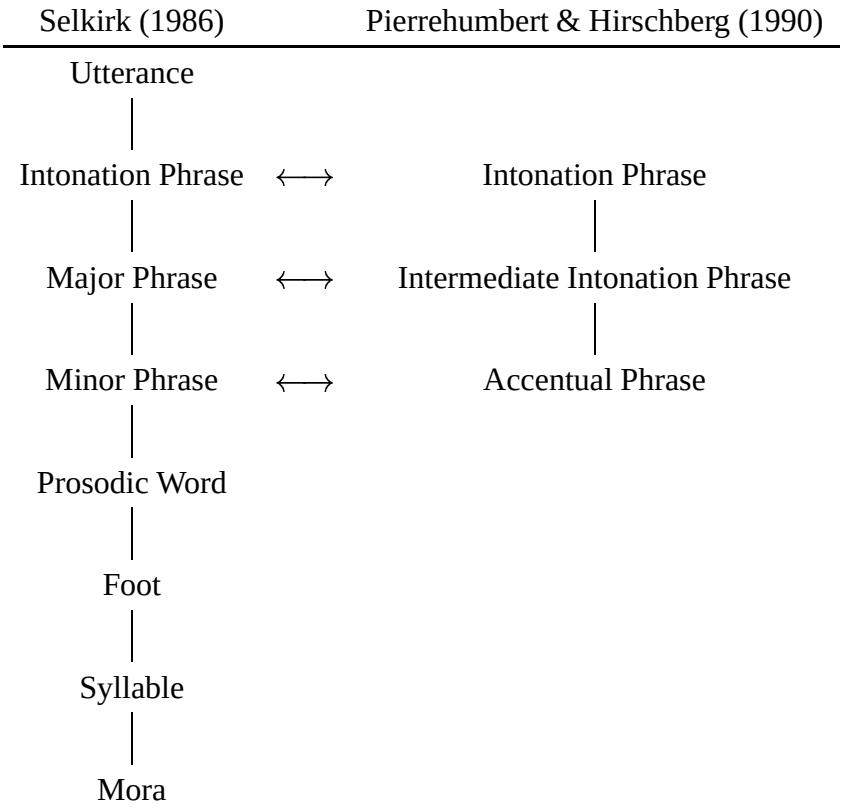

Figure 2.2: The Prosodic Hierarchy (adapted from Shattuck-Hufnagel & Turk 1996, p.206)

level with topic boundaries. That is, prosodic cues are used to mark the boundaries of sentences and topics in discourse. Again, we should be careful to note that they could also mark lower level phrase boundaries; and that different boundary markers might be used in different circumstances, for example, if the sentence were a question or a statement.

It is also generally agreed that acoustic indicators of prosody are used to make certain words more prominent than others in order to affect the way they are interpreted in the discourse context. Such markers are called pitch accents. According to the traditional phonological analysis, there is one such pitch accent, the nuclear accent, per intonational phrase. These pitch accents should be distinguished from lexical stresses (see Ladd 1996, p.46-51). Put simply, every content word in English has one syllable which receives primary lexical stress. If the word also has a pitch accent, it will be aligned

concerned only with syntactic sentences.

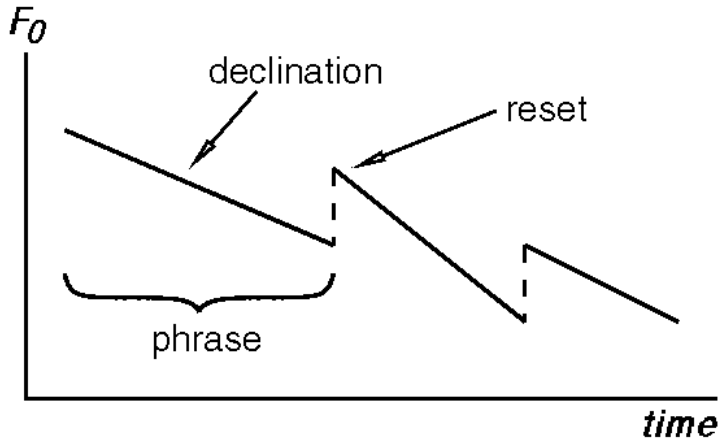

Figure 2.3: Pitch Declination over Topics and Intonational Phrases (King 2001)

with this syllable.[2] The acoustic correlates of lexical stress are very similar to those for pitch accents, and this is important when trying to automatically identify segments with pitch accents.

2.2.1 Topic Boundaries

Cutler et al. (1997) review a number of studies which show that speakers prosodically mark the boundaries of topics differently from the boundaries of sentences. Speakers have been shown to start new topics relatively high in their pitch range and finish by compressing their range (Brown, Currie & Kenworthy 1980, Venditti & Swerts 1996). Brown et al. (1980) also found that there was a rise in amplitude at the beginning of a topic, and a fall at the end. A similar pattern has been found over sentences, or intonational phrases, leading to the type of pitch declination pattern shown in Figure 2.3. The F0 peak in the first sentence in a topic is higher than the F0 peak in subsequent sentences.

Swerts & Geluykens (1994) investigated speakers' use of prosodic cues to discourse structure in a production study of Dutch instruction monologues. They found that the F0 peak and the mean F0 of the first sentence in a topic was higher than in subsequent

[2]There are exceptions to this, such as the Rhythm Reversal Rule (Vogel, Bunnell & Hoskins 1995), where the stress falls on a different syllable to the one that normally receives primary stress.

sentences. Further, they showed that speakers are more likely to use low boundary tones at the end of a topic (the giving of one instruction), as opposed to at the end of any other sentence. Boundaries were marked as either low or not low by trained intonation researchers after listening to the speech signal. Swerts & Geluykens go on to report that speakers always pause between topics, whereas they only sometimes pause between other sentences; and that they pause for longer. In a perception study associated with the experiment (Swerts & Geluykens 1993), listeners could successfully use the melodic and pausal information to predict topic boundaries when segmental information was removed from the signal.

Sluijter & Terken (1993) found similar results in their study of read multi-sentence texts in Dutch. They got speakers to read the same sentences in different positions in a topic unit (beginning, middle, end) and compared the acoustic parameters of each. They looked at the F0 contour for each sentence and manually extracted an F0 topline and baseline as well as the F0 maximum. They found that the onsets and offsets of both these lines and the F0 maximum got significantly lower over the course of the topic unit.

Swerts's (1997) study was carried out using spontaneous monologues in Dutch, out of concern that the types of speech materials used in the studies above might represent an unnaturally rigid topic structure. He got subjects to mark topic boundaries both on the basis of text alone and on hearing the speech signal. He found that the level of inter-annotator agreement as to the presence of a topic boundary was directly correlated to the presence at that point of the acoustic cues mentioned above. That is, the presence of a low boundary tone, greater F0 maximum than the previous sentence and a long pause between sentences were more likely to make annotators agree that there was a topic boundary. Moreover, these three factors worked together.

In a series of experiments based on Grosz & Sidner's (1986) theory of discourse structure, Grosz, Hirschberg and Nakatani again show that pitch range and contour, pause duration and amplitude are all used to mark the 'global' level of discourse structure (Grosz & Hirshberg 1992, Nakatani, Hirschberg & Grosz 1995, Hirschberg & Nakatani 1996). In a pilot study using broadcast news stories similar to Swerts', Grosz & Hirshberg (1992) found that the F0 maximum, average F0, amplitude and speaking rate (syllables per second) were all greater in sentences that the annotators had reliably identified as the beginning of a topic, as opposed to other sentences. On the

basis of these findings, they used Classification and Regression Tree Analysis (CART, see Chapter 4 for explanation) to build decision trees from these feature values. They found the intonation features could be used to reliably identify the beginning of a topic 91% of the time. Similar results were found using a corpus of both spontaneous and read direction-giving monologues (Nakatani et al. 1995, Hirschberg & Nakatani 1996). Beckman & Edwards (1992) found evidence in their study that speakers lengthen the final segment in different levels of intonational phrasing, with the amount of lengthening correlating to the level of intonational structure. Although they do not look at topic boundaries, it seems plausible that these could be lengthened more than sentence boundaries.

2.2.2 Pitch Accents

As well as using prosodic cues to mark the boundaries of topics, speakers use pitch accent to mark the information status of different words or concepts within a topic. Cutler et al. (1997) review a number of studies that show that speakers mark 'new' entities in a discourse with a pitch accent and that 'given' entities are deaccented. There is much dispute about the definition of 'given' and 'new'. However, in the studies below 'new' is taken to be the first mention of a particular concept and 'given' is all subsequent mentions. This is sufficient for the broadcast news paradigm we are concerned with here.

In a study of question-answer pairs (*Why is Ken smiling? (HE/he) won the (LOTTERY/lottery)*), Birch & Clifton (1995) found that listeners' judgements of prosodic appropriateness were higher when new information (*lottery*) was accented compared to old (*he*). Terken & Nooteboom (1987) found that subjects could more quickly determine whether a sentence correctly described a visual display (e.g., *the P is on the left of the K*) if the newly introduced entities were accented and the old ones were not. Donselaar & Lentz (1994) showed that listeners find it harder to process words if they are inappropriately accented. In a word-naming task, response times were slower when given words were accented than when they were not. Donselaar (1995) found a similar pattern when synonyms were used instead of the same noun to refer to an entity.

Fowler & Housum (1987) investigated this further, excising first and subsequent mentions of a concept in radio monologues. They found that second mentions of a concept

were in general shorter and had poorer vowel quality than first mentions. In an out-of-context recognition task they proved to be less intelligible than first mentions and listeners could reliably identify whether they were first or second mention. Hawkins & Warren (1994) related this finding to the fact that most first mentions are accented and second mentions unaccented. However, this just reinforces the findings discussed in the previous paragraph.

Given this evidence, a number of researchers in this field have claimed that the accenting of certain words affects how we fit them into our discourse model. Fowler & Housum claim that speakers reduce the intelligibility of repeated mentions of words in order to indicate that they refer back to a previous mentioned entity. Terken & Nooteboom (1987) suggest that the presence of an accent indicates to listeners that a new discourse entity must be constructed.

So how does this relate to our topic segmentation task? Hirschberg & Ward (1991) show that deaccentuation acts as a kind of anaphoric device, making it clear that a given entity is being referred to, and hence that the speaker is still talking about the same topic. Gernsbacher & Jescheniak (1995) propose the complement of this, they studied whether accentuation can act as a cataphoric, i.e. forward reference, device. They showed that key concepts (topics) in a discourse are accented so that they gain a special status in the mental representation of the listener, helping the listener to access that concept later in the discourse.

Nakatani et al. (1995), reported above, attempted to incorporate these types of ideas in their investigation of the correlation between prosodic cues and discourse structure. They recorded whether noun phrases in the directions corpus were ‘accentually reduced’, i.e. bore fewer pitch accents than the citation form. They found that the simple correspondence between deaccentuation and givenness did not hold, as accentuation interacted with other factors. However, in general, new entities were more likely to be accented, and given entities to be accentually reduced.

In terms of building a system for the automatic identification of topic boundaries (given that sentence boundaries have been identified), the studies reported in this section appear to be an ideal starting point. A range of quantifiable acoustic factors have been identified which are used by speakers and understood by listeners to indicate topic boundaries. Within a given sentence these are: the F0 peak, the mean F0 (or topline

and baseline onset and offset), the presence or absence of a low boundary tone, the peak intensity, the mean intensity, the length of the pause at the sentence boundary, the speaking rate (for example in syllables per second), and the amount of lengthening in the final syllable. In addition, the presence of pitch accents on new vocabulary items may indicate a new topic; and deaccentuation of old vocabulary items may indicate topic continuation.

2.3 ToBI and Intonation/Information Structure

The studies reviewed above measured the effect of each acoustic correlate of prosody in turn. The underlying assumption of this is that speakers manipulate each acoustic correlate (or rather its articulatory equivalent) independently to indicate various aspects of the hierarchical structure of the text. This may not be the theoretical standpoint of the researchers involved but it does fall out from the way these studies were conducted. However, as is pointed out in some of the aforementioned studies (Nakatani et al. 1995, Swerts 1997), the various correlates appear to work together. For example, a particularly low boundary tone may abrogate the need for a long pause to indicate a topic boundary. Further, there is a great deal of variation both between speakers and with the same speaker on different occasions as to how these acoustic indicators are used. Despite this, we as listeners can perceive the effects these acoustic measures are trying to capture.

For these among other reasons, it has become standard in the linguistic community to analyse prosody in terms of intonational events. That is, pitch accents and boundary tones combining in a linear fashion to make a prosodic structure (Shattuck-Hufnagel & Turk 1996, p. 229). In the past ten years, the most widely used system for such intonational analysis has been Tones and Break Indices (ToBI, Pierrehumbert & Hirschberg (1990), Silverman et al. (1992)). This system will be briefly explained. We will then reframe the evidence for prosodic marking of topic boundaries presented in the last section in terms of the ToBI system.

As was alluded to, the notion of the relationship between pitch accents and information status (given/new) reported in the studies above is very simplistic. There are many reasons why a word might receive a pitch accent in a given discourse. In fact, one of

the motivations for the development of the ToBI system was to recognise that there are different types of pitch accents, each of which has been claimed to have a different intonational meaning. We will review a line of research from Beckman & Pierrehumbert (1986) to Steedman (2000) which claims that tones form part of the semantic interpretation of a sentence. We will then discuss how this can be used in our topic segmentation problem.

2.3.1 ToBI

In the ToBI system the pitch contour is represented as a series of pitch accents and edge tones. Pitch accents can consist of a single H or L tone or a combination of two tones. In the combination accent, the central tone is 'starred' (hence H* or L*), and the other tone is said to lead or trail from it. The combination tones are meant to be used when there is a clear movement up or down in the F0 contour, as opposed to just a general rise to an H* peak or fall to an L* low. There is some disagreement as to exactly which combinations of H and L of all the possible ones are allowed (see discussion in Ladd 1996, chap.3), but the conventions in our corpus allow for L*, H*, L*+H and L+H* (Beckman & Hirschberg 1999). In addition to this, the second accent in a combination, or the second single accent in an intonation phrase, can be downstepped. This is indicated with a ! before the accent, e.g. !H* (more on this below).

Boundaries between prosodic groupings (see Figure 2.2) are marked on a scale from 0-4. Boundary markings between 0-2 are generally used for boundaries within phrases. Level 3 breaks (intermediate phrases) are marked with phrase accents, H- and L-. Full intonational phrases, level 4, are marked as either H% or L%. The system therefore analyses pitch contours as a finite state network of intonational events, which are either pitch accents or boundary tones, each of which is made up of H or L tones. This is represented diagrammatically in Figure 2.4. As was mentioned above, there is some dispute as to what exactly constitutes intermediate and intonational phrases. However, it is generally agreed that sentence boundaries align with intonational phrase boundaries (Shattuck-Hufnagel & Turk 1996). Since we are only working at the sentence level and above, we do not need to address the problem of the definition of lower-level constituents such as the intermediate phrase. There is at present no reliable system for automatically producing ToBI annotations so these are in general done manually by

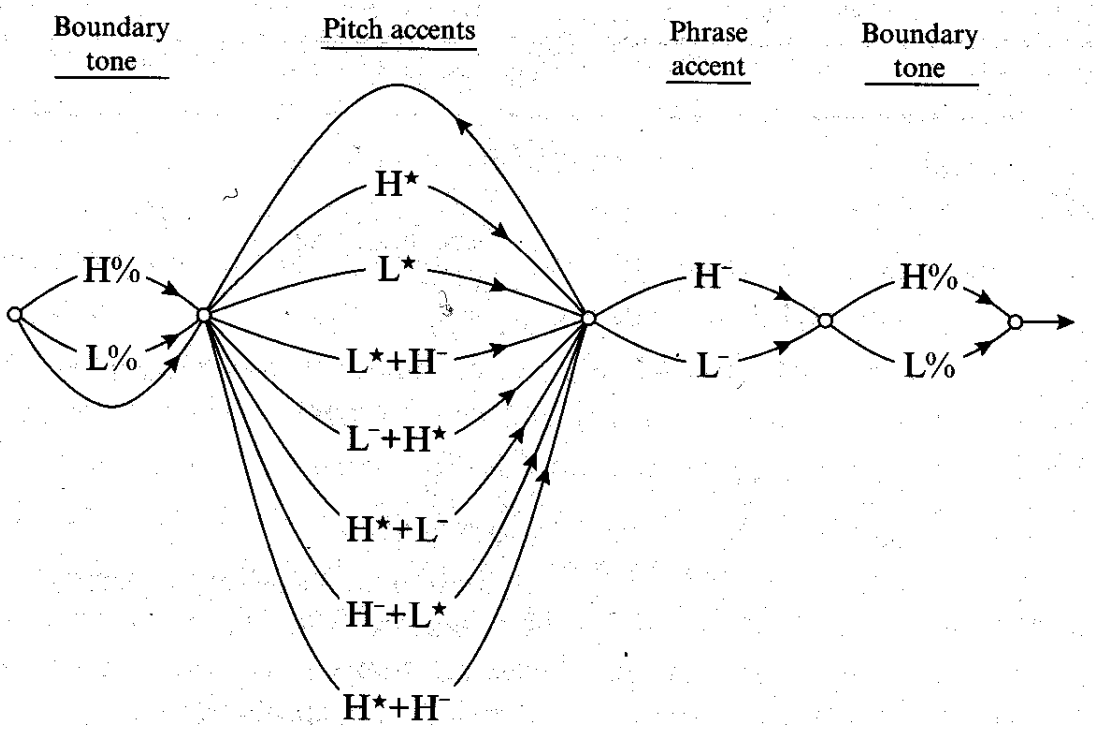

Figure 2.4: ToBI Finite State Network (Ladd 1996, p.81)

trained annotators.

If the ToBI intonation framework is valid it should be possible to reframe all the prosodic phenomena described above in terms of the ToBI system. The ‘low boundary tone’ feature would seem the easiest to start with. We can just say that sentences ending in L% are more likely to be topic endings. Or, perhaps more fruitfully (since L% is the most common boundary tone), that sentences ending in H% are likely to indicate topic continuations. This may turn out to be too simplistic, as it could be the ‘tune’ (the final pitch accent as well as the phrase accent and boundary tone) which is important in determining whether the phrase is perceived as a continuation or not. This will be further discussed in relation to Steedman’s theories below.

The representation of general F0 declination in the ToBI system has proved to be a problematic area for researchers. As Ladd (1996, p.280-283) discusses, the notion of a global decline in F0 does not fit well with the formulation of intonation as a series of (independent) tonal events. The strategy recommended to ToBI annotators has been to annotate successive F0 peaks within the same intonational phrase that show the general

F0 declination using the downstep (!) marker. However, this does not help to indicate the decline in F0 over a number of phrases. Further, since F0 declination is such a common phenomenon, it is often not heard by the human annotators. This led Beckman & Pierrehumbert (1986) to call general F0 declination a paralinguistic feature. This is problematic as F0 declination is not universal, for example, it does not occur in some question intonation patterns. The issue is not settled, which is unfortunate because it would be helpful to have a way of representing F0 declination that abstracts away from absolute F0 levels which are heavily speaker and situation dependent.

So why would we want to work with ToBI features? It is precisely because they derive from human perceptions of intonational events. They are thus able to go from acoustic measures which are tied to one speaker to find generalisations about intonational structure which will hold across all speakers. This is particularly evident in Steedman's Information/Intonation Structure Theory.

2.3.2 Steedman's Information Structure Theory

In his theory of grammar developed over the past ten years Steedman (1991, 1996, 2000, 2001) has claimed:

> "the Surface Syntax of natural language acts as a completely transparent interface between the spoken form of the language, including prosodic structure and intonational phrasing, and a compositional semantics. The latter subsumes quantified predicate argument structure, or Logical Form and discourse Information Structure." (Steedman 2001, p.1)

His claim is then that intonational tones, in the ToBI system, are, like words, part of the surface structure of a sentence, and are used to derive the semantic interpretation of the sentence. As we are not directly concerned with the derivation of sentence semantics here, we will concentrate on the parts of the theory that concern the identification of themes and rhemes, which are relevant to the segmentation of topic structure.

Steedman claims that different tunes are used to mark phrases in an utterance as either the theme or the rheme:

(2.1) Theme: L+H* LH%

Rheme: H* L and H* LL%

As Steedman (2000, p.672) discusses, these markers are not uncontroversial. In particular, there is much debate as to whether a reliable phonetic difference between the L+H* and H* pitch accents can be found (see Ladd & Schepman forthcoming, Calhoun 2002). However, accepting that a reliable phonetic different between 'theme' and 'rheme' tunes can be found, the theory works. Roughly speaking, the theme is 'what you're talking about', and the rheme is 'what you want to say about what you're talking about'. Take the following example (from Steedman (2000, p.654)):

(2.2) Q: I know who proved soundness.
But who proved COMPLETENESS?

A: (MARCEL) (proved COMPLETENESS).
H*L L+H* LH%

(2.3) Q: I know which example Marcel PREDICTED. But which result did Marcel PROVE?

A: (Marcel PROVED) (COMPLETENESS).
L+H* LH% H* LL%

In example 2.2, *completeness* is 'what we're talking about', therefore it is the theme, and *Marcel* is 'what we want to say about it', therefore it is the rheme. In example 2.3 it is the other way around. The theme-rheme distinction is similar to the given-new distinction described above, but not exactly the same. Steedman in fact splits the given-new distinction into two: the theme-rheme distinction and the background-focus distinction. This can be seen in example 2.4 (from Steedman 2000, p.659):

(2.4) Q: I know that Marcel likes the man who wrote the musical.
But who does he ADMIRE?

A: (Marcel ADMIRES) (the woman who DIRECTED the musical)

$$\underbrace{\underbrace{\text{Marcel}}_{\textit{background}}\ \underbrace{\text{ADMIRES}}_{\textit{focus}}}_{\textit{theme}}\ \underbrace{\underbrace{\text{the woman who}}_{\textit{background}}\ \underbrace{\text{DIRECTED}}_{\textit{focus}}\ \underbrace{\text{the musical}}_{\textit{background}}}_{\textit{rheme}}$$

$$\underbrace{\underbrace{\text{L}+\text{H}*}_{\textit{focus}}\ \underbrace{\text{LH\%}}_{\textit{background}}}_{\textit{theme}}\ \underbrace{\underbrace{\text{H}*}_{\textit{focus}}\ \underbrace{\text{LL\%}}_{\textit{background}}}_{\textit{rheme}}$$

Focus operates at the word level and belongs to the word which receives the pitch accent. Focus then draws attention to different elements within a theme or a rheme. For instance, above, the whole constituent *Marcel admires* is the theme, because it is what the speaker is talking about. However, *admires* is in focus because it is what was being directly asked about.

Steedman's theory is built within the framework of Combinatory Categorial Grammar (CCG). For the purposes of this work, it is sufficient to say that this system allows tones to be included in the surface structure of the sentence and hence for a combination of pitch accent and boundary tone to mark a sentence constituent as either the theme or the rheme.

As Steedman is at pains to point out, renditions such as the above would present a very careful pronunciation of the sentence, and in many cases the theme will be unmarked or the L- boundary may not be phonetically realised. Steedman claims that this does not present a weakness in his theory as this sort of under-production and ambiguity is rife in all aspects of language. However, as Ladd (1996, p.224) comments, this makes the theory hard to verify independently and recognise automatically.

In a study of different types of accents and tunes occurring with constituents of different information status in a televised political discussion show, Hedberg & Sosa (2001) conclude that although there may be systematic correlations between intonation and information structure categories, these are not as simple as the literature suggests. In particular, they found that the L+H*LH% tune was more likely to mark 'contrastive focus', i.e. a rheme that contrasts with something previous said, than to mark the theme. They did find that the L+H* accent was more likely to mark 'topics' (themes in Steedman's terms), they were also used to mark a significant number of 'foci' (rhemes). The H% boundary tone was also more likely to mark rhemes than themes. They do not present any specific findings about the H*L- or H*LL% tunes. However, they do find that the H* accent was fairly evenly distributed in all five categories of information status they used, although they did find that the LL% boundary was much more likely to mark rhemes. Again, this evidence is not fatal to the theory as long as phonetically distinct theme and rheme intonation events can be identified. One of the aims of this study is to assess Steedman's claims on a new set of data.

Presuming that Steedman's theory as it presently stands does have validity, how can we use it for the topic segmentation task? His information structure is a sentence internal concept, and is not the same as topic structure here. However, the information status of entities in a single utterance should relate to their status in the overall discourse. The most straight-forward application would seem to be that if an entire sentence has a L+H*LH% tune, i.e. is a theme, then the next sentence will be in the same topic. However, this is not likely to occur very often as most sentences represent complete information units within themselves, i.e. have themes and rhemes. Alternatively, we could look at the first pitch accent in a sentence, or the tune of the first intermediate phrase. If it is a marked theme, this is an indication that this is a continuation of a topic, i.e. giving more information about something that has already been talked about. If it is a rheme (has a H* accent), it is more likely that we are starting a new topic, as there is no established 'something we're talking about' yet. Another direction would be to look for chains of words where the first mention has a H* accent and the subsequent mentions have either L+H* accents or are unmarked. Where these chains break off would be likely to be topic boundaries. This could be more indicative than just looking for repeated words as it cuts out the 'noise' of repeated mentions of non-topic words.

2.4 Taylor's Tilt Intonation System

The principal problem with using ToBI features to try to identify topic boundaries in an automatic system is that at present they can only be identified by human annotators. This may be, however, because they capture human intuitions about intonation which have not yet been sufficiently described in terms of acoustic correlates. As will be further discussed in Chapter 4, our corpus is only partially labelled with ToBI labels. Therefore, the topic boundary system was built using Tilt features (Taylor 2000), which can be derived automatically. Tilt features have been successfully used by Hastie (2002) in the identification of utterance types in dialogue. The Tilt system will be described below, along with an explanation of how it can be used to approximate ToBI features.[3]

Like the ToBI system, Tilt characterises intonation as a series of linear intonational

[3]Tilt parameters can of course be used in their own right to do automatic discourse analysis, as was shown in Hastie's (2002). However, in this study we wish to use them to approximate ToBI features in order to evaluate the predictions about these features made above.

events. Again, these are pitch accents and boundary tones. The first stage in extracting Tilt features is the identification of these events. The speech is divided into evenly spaced frames and these are passed through a network of Hidden Markov Models (HMM) which identifies each frame as *a* a pitch accent, *b* a boundary tone, *ab* a combined pitch accent and boundary tone - where two distinct events cannot be identified, *c* a continuation - the speech between two intonational events, or *sil* silence. The HMM training used in this experiment will be further explained in Chapter 4. The Tilt model is essentially phonetic, so intonational events are defined as perceptible excursions in the F0 contour. There are no 'level accents', i.e. accents which may be phonologically real but have no phonetic realisation.

Once these events (*a*'s and *b*'s) have been identified, the F0 contour within each event is used to find the five Tilt parameters which describe it. Firstly, a peak-picking algorithm is used to describe each event as a rise, fall or combined rise-fall - based on whether there is a clearly discernible peak. The position of the peak measured from the beginning of the utterance is entered. The Tilt parameters are then described in terms of the amplitude in Hertz and the duration in seconds of the rise and the fall. The next two Tilt parameters, amplitude and duration, are the sum of the magnitudes of the amplitude and duration in the rise and fall (from Taylor 2000, p.14):

$$Tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{2.5}$$

$$Tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{2.6}$$

The last parameter is Tilt, an average of the amplitude Tilt and duration Tilt:

$$Tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})} \tag{2.7}$$

The Tilt of an event is a number between -1 and +1. As can be seen in Figure 2.5, different Tilts correspond to different shapes in the F0 contour. For example, a pure rise is +1, an even rise and fall is 0, and a long rise followed by a short fall is +0.5. The advantage of this system is that the Tilt parameter captures the shape of the F0 contour, whereas in ToBI a lot of different shapes must be put into the H* category.

How can we use the Tilt system to approximate ToBI features that we have identified

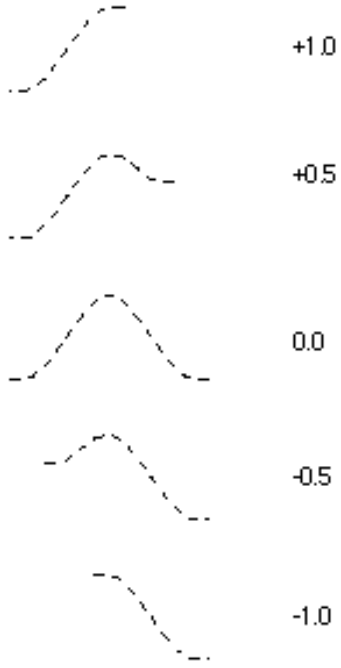

Figure 2.5: Five Events with Different Values of Tilt (Taylor 2000, p.15)

as relevant to the immediate task in the discussion above, i.e. H*, L+H*, L-, LH% and LL%? We can see that Tilt, like ToBI, characterises intonation in terms of a series of events. However, the only events it recognises are excursions in the F0 contour. Therefore, it will not recognise the L- and LL% boundary tones. The nearest approximation will be to look for sentences which do not end in a *b* event, i.e. that end in a *c* - continuation. On the other hand, the LH% boundary tone should be easier to identify, we look for a *b* event with positive Tilt. *ab* events with Tilt close to one would be a pitch accent followed by LH%, those close to zero or negative could possibly be a pitch accent closely followed by a L% boundary.

Identifying the difference between H* and L+H* is again a vexed issue. As Taylor (2000, p.24) himself points out, there is considerable overlap in the Tilt system between the H* and L+H* accents, as shown in Figure 2.6. He considers this as evidence against the current putative difference between them. However, for the purposes of the current system, we must assume it exists.

The best we can do then is to say that an *a* event with Tilt less than 0 is definitely H*,

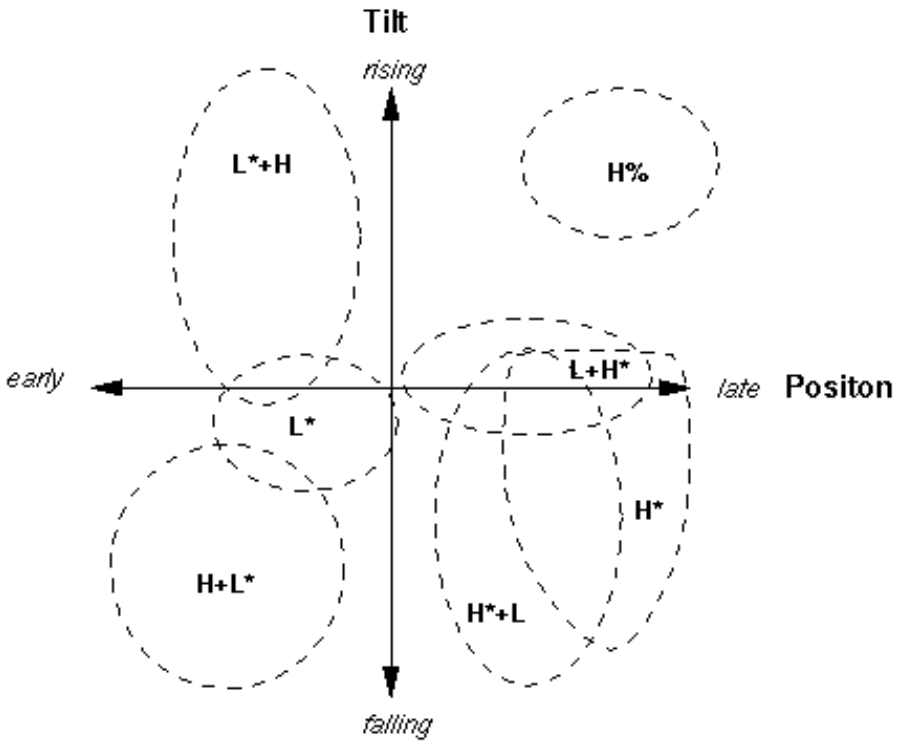

Figure 2.6: Overlap Between Tilt of H* and L+H* (Taylor 2000, p.24)

as the rise is either short or not there, but that an *a* event with Tilt greater than 0 could be L+H* or possibly H*. In this case, we will have to rely on the identification of the tunes discussed above, H* L(L%) versus L+H* LH%, to distinguish between themes and rhemes in terms of Steedman's theory. (In practice, it was found that using just the boundary tone was more effective with the corpus used in this study).

2.5 Conclusion

In this chapter we have reviewed intonational cues to topic structure from two different perspectives. As was stated earlier, according to the first view, we can identify topic structure by taking direct acoustic measures - including F0 levels, pause duration and speaking rate. In terms of building an ASR system that does topic identification, this would seem computationally the most straightforward. However, as was discussed, it is problematic because speakers use these different cues to indicate a variety of discourse phenomena and different speakers use different cues to different extents. Also, absolute measures of acoustic indicators have only limited use, as these vary from speaker to speaker and situation to situation.

According to the second view then, intonation can be viewed as a series of events, with certain events being indicative of topic boundaries. This is computationally more expensive and prone to error as there are two recognition processes to be undergone. The first is to accurately identify and classify the intonation events themselves, the second to map intonation events onto discourse structure. However, the separation of the two processes, if successful, can lead to a much more powerful system as inter-speaker and perhaps even inter-language variability can be dealt with in terms of the mapping between acoustic indicators and intonational events whereas the taxonomy of mappings between intonational events and discourse structure should remain constant or at least vary systematically between languages.

Chapter 3

The Topic Segmentation Task

We now turn to the task to which we will apply the theory on prosodic marking of topic structure discussed in the last chapter. Topic segmentation has been recognised as an important step in a number of information extraction tasks for some years now. In any kind of topic identification, it is essential that the boundaries of the topic are accurately identified (see Chapter 1 for further discussion). In the studies reported below and in this project, topic segmentation is treated as a statistical classification task. First the text is broken into chunks (in this case generally either a sentence or a chunk of a fixed length, say 20 words), and then each chunk is classified as being either at the beginning of a topic or in a topic. This is done on the basis of features about each chunk determined by the system designer which are used to form a statistical model of the data. We will begin by reviewing the features which have been used to date to find topic structure, before explaining the statistical models that will be used in this project.

Topic segmentation was a separate competition within the 1999 and 2000 NIST Topic Detection and Tracking Workshop competitions (Fiscus, Doddington, Garofolo & Martin 1999). The overall aim of that series of workshops was to advance the state of the art in systems designed to automatically recognise broadcast news stories over a certain time period and group related stories into topics. There were five entrants in this section in the 1999 competition: CMU, Dragon Systems, IBM, SRI and the University of Iowa; and one in the 2000 competition: Mitre.

As will be seen, the majority of the systems primarily use textual features which have been shown to be effective in identifying topic boundaries in written text. We will

review the textual features used, starting from seminal work in this area on text tiling (Hearst 1997), as well as the use of cue phrases and topic key words.

We will then move on to systems which used prosodic features to determine boundaries, detailing how they were used. We will also mention Hirschberg & Nakatani (1998), who while not in the NIST competition, have evaluated the use of prosodic cues for much the same task. We will also look at Silipo & Crestani's (2000) study, which showed that there is a high correlation between words which receive pitch accents and topic words.

Many of the systems presented in these competitions are principally concerned with the machine learning techniques used. This is not a central concern of this project. However, we will briefly review the machine learning algorithms used by these systems and go into greater detail about the machine learning algorithms used in this study.

Once we have established the textual and prosodic cues which have been used in the topic segmentation task so far, we will be in a position to evaluate the effectiveness of using these cues as compared to the more sophisticated prosodic cues suggested in the previous chapter. This evaluation could lead to two possible conclusions. We could ascertain the value of including higher level intonational features in such a system from an ASR perspective. From a linguistic perspective, we would have further evidence for or against the proposals about intonational structure made above.

3.1 Written Cues

There are three types of textual cues that are commonly used to identify topic boundaries in the written domain:

1. **Text Tiling:** The distribution of 'content' words is used to determine topic boundaries. The theory is that content words will tend to co-occur within topics but not over topic boundaries.

2. **Cue Phrases:** Certain words or phrases are used by writers to indicate topic boundaries or topic continuations and therefore tend to correlate more highly with them.

3. **Topic Key Words:** Words in a text are compared to lists that have previously been found to correlate with topics identified in the training data

We will go through each in turn.

3.1.1 Text Tiling

The text tiling technique was developed by Hearst (1997), originally to determine subtopic boundaries in expository texts (his experiment used five scientific magazine articles) for the purpose of information retrieval and summarisation. The algorithm in Hearst (1997) comprises three stages:

1. Tokenisation
2. Lexical Score Determination
3. Boundary Identification

In the tokenisation stage, the text is separated into words, which are converted to lower case. Stop words (closed-class words and other high frequency words, see Salton (1989)), are removed as this type of word was found to be unlikely to indicate the current topic by Hearst. The text is then divided into 'token sequences', pseudo-sentences of a predefined length (in this case 20 words) to allow for comparison between equal sized blocks of text.

In the lexical score determination stage, the gap between each adjacent topic sequence is assigned two different lexical scores to determine the similarity between two blocks of k token sequences.

The first score is based on the number of tokens which are in both the first and the second block. This is illustrated in Figure 3.1. "The lexical score is calculated by a normalized inner product: given two text blocks b_1 and b_2, with k token sequences, where $b_1 = \{token-sequence_{i-k}, ..., token-sequence_i\}$ and $b_2 = \{token-sequence_{i+1}, ..., token-sequence_{i+k+1}\}$,

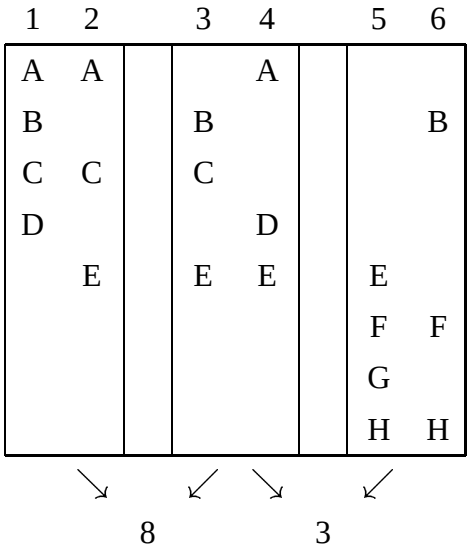

Figure 3.1: Lexical Score Determination - Continuation (Hearst 1997, p.44)

$$score(i) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}} \tag{3.1}$$

where t ranges over all the terms that have been registered during the tokenization step (thus excluding stop words), and $w_{t,b}$ is the weight assigned to term t in block b" (Hearst 1997, p.49). Here Hearst uses the frequency of the terms in their block as the weight. Formula 3.2 returns a number between 0 and 1. As should be clear from Figure 3.1, a number closer to 0 should indicate a topic boundary, a number closer to 1 should indicate a topic continuation (note that the figure gives the raw count).

The second score is based on the number of tokens that appear in the second block which do not occur in the first block. This is the reverse of the first score. As is shown in Figure 3.2, a high score comes from a high number of 'new' vocabulary items and hence should indicate a topic boundary. (Again, the figure does not show normalised scores). In the broadcast news genre, it is reasonable to assume that a vocabulary item which did not occur in the previous two sentences is new, as topics are quite short and sequential. A low number of new words should indicate a topic continuation. The lexical score is determined as follows,

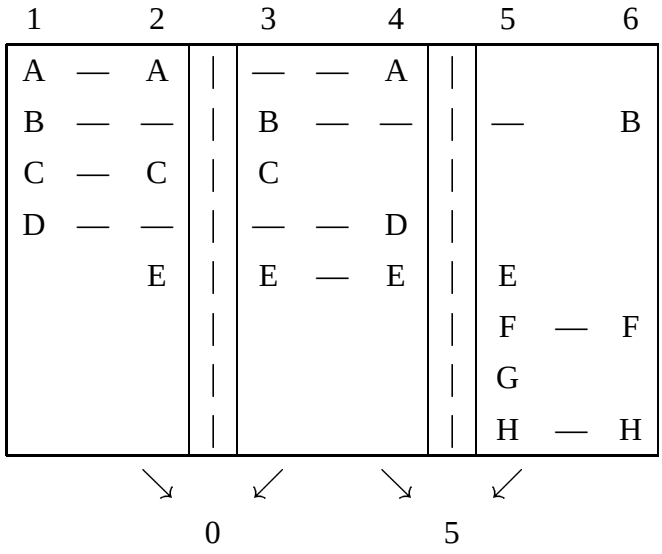

Figure 3.2: Lexical Score Determination - Introduction (Hearst 1997, p.44)

$$score(i) = \frac{NumNewWords(b_2)}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}} \tag{3.2}$$

where b_1, b_2 are defined as above and w is the total number of tokens.

We can see how both of these scores work in the following example from our corpus:

(3.3) <TOPIC><S> A new committee has been set up to look into the Boston police department's drug control unit . brth </S>
<S> Mayor Ray Flynn says the move is not in response to the case of accused police killer Albert Lewin brth and allegations of police misconduct . </S>
<S> Flynn says the commission comprised primarily of legal experts will investigate procedures at the DCU , brth including training and issuing of search ad arrest warrants . </S></TOPIC>
<TOPIC><S> While most people protest against the Internal Revenue Service tax time , brth the protest is coming from an unusual corner this year . brth </S>

Here the words *police* and *Flynn* are repeated through the first topic, while no content words are used both before and after the topic boundary. In terms of the new score, there are a number of new vocabulary items to indicate the beginning of a new topic, including *people*, *Internal* and *protest*.

In Hearst's original algorithm, the two lexical scores for each gap are used to identify subtopic boundaries by determining a 'depth score' at each interval. The depth score is computed by comparing each lexical score to the lexical score of the gaps immediately following and preceding that one. We will not go into the details of this here as our system uses the lexical scores directly as features. However, the theory is the greater the depth score, the more likely that there was a topic boundary. In a follow-up study, Hearst found that these scores did indeed accord with reader judgements of topic boundaries.

Lexical distribution was used as a feature in the IBM system (Dharanipragada, Franz, McCarley, Roukos & Ward 1999). Instead of using the normalisation equation laid out above to determine lexical similarity or difference in adjacent blocks of text, they used a term frequency/inverse document frequency (*tfidf*) score (see Dharanipragada et al. 1999, p.3). This metric was also used by the University of Iowa system (see Eichmann, Ruiz, Srinivasan, Street, Culy & Menczer 1999, p.6).

The MITRE system (Grieff, Morgan, Fish, Richards & Kundu 2000) computed essentially the same feature slightly differently. For each word, they compared the number of times that word occurs in the previous fifty words to the number of times it would occur in fifty words on average. This was used to detect topic change (see Grieff et al. 2000, p.2).

3.1.2 Cue Phrases

Writers use discourse and formatting conventions to signal topic structure to their readers. These are fairly genre-specific, and many, such as paragraphing and headings, do not apply to the spoken domain. However, it is still reasonable that we should be able to find words or short phrases such as *however* which signal topic continuation, or *In other news* which signal topic change. For example, in the following extract from our corpus, *meanwhile* signals topic change:

(3.4)
```
<S> McGovern opened public hearings on the state's fiscal
 nineteen-ninety budget today with a terse warning to those
 planning to testify . </S></TOPIC>
<TOPIC><S> Meanwhile , county sheriffs from across Massachusetts
 took to the state house today to demand more money . brth </S>
<S> The sheriffs say jails in eight counties will run out of funds
 by May first . </S></TOPIC>
```

This list of cue phrases could be formulated manually, for example by examining transcripts, but this approach would be fairly brittle. The IBM system (Dharanipragada et al. 1999) extracts such cue phrases automatically using a mutual information criterion. The likelihood of a unigram or bigram occurring either near a topic boundary or not near a topic boundary is computed according to formula 3.5,

$$MI(t,w) = log\frac{P(t,w)}{P(t)P(w)} \tag{3.5}$$

where t is either a topic boundary, or not topic boundary, w is the unigram or bigram count, and c is the total count.

Those unigrams or bigrams which are found to correlate highly either with a topic boundary or a continuation, but not both, are kept as 'cue phrases'. The presence of any of these cue phrases in a sentence is then taken as a feature in the machine learning algorithm. In this case of Dharanipragada et al.'s (1999) system this was a decision tree (see below for description). The MITRE system also used cue words as features, but used the *tfidf* metric to identify them (see Grieff et al. 2000, p.2).

The CMU system (Beeferman, Berger & Lafferty 1999) used two language models, effectively combining the textual phenomena behind both text tiling and cue words. They divided the entire training corpus into trigrams, which were used to train two language models. The short-term model was an HMM and was considered to be reasonably stable over a topic boundary. The long-term model was a maximum entropy model which used the trigrams as features. They then tested the ability of each model to predict the next word. The long-term model was more highly predictive but more likely to break down at topic boundaries. Therefore, a topic boundary was detected when the short-term model became better than the long-term one at predicting the next

word.

3.1.3 Topic Key Words

The final textual feature used by the NIST systems was the similarity between words found in one block and lists of words associated with different topics determined from the training set. If the words in one block were significantly more likely to have come from one topic than from another, on the basis of overlap with the lists of words for that topic, then this would be likely to be a new topic. This was the approach taken by both the Dragon system (Mulbregt, Carp, Gillick, Lowe & Yamron 1999) and the SRI system (Stolcke, Shriberg, Hakkani-Tür, Tür, Rivlin & Sönmez 1999), who constructed an HMM model for each topic in the training corpus.

Unfortunately, this feature would not be effective in this project as our corpus is not large enough to accurately identify a broad range of plausible topics. We will not go into this further here. In any case, this method always suffers in its ability to deal with unseen topics.

3.2 Prosodic Cues

While two of the systems in the NIST competitions did not use any non-textual features (CMU, Dragon), the rest of the systems made some attempt to incorporate prosodic cues to discourse structure. Most of the systems (IBM, Iowa, SRI and Mitre) used pause length as a feature. Some systems used the length of the non-speech events, i.e. including *umm*s and *er*s, etc instead of just silences. These were presumed to be longer at topic boundaries than other phrase boundaries (as discussed in the previous chapter).

3.2.1 Pitch features

The only system in the NIST competition to make use of F0 features was from SRI (Stolcke et al. 1999). They combined F0 features with the lexical and pause duration features described above using a decision tree (see below). They have since evaluated

the use of these features in similar tasks using different corpora (Shriberg, Stolcke, Hakkani-Tür & Tür 2000, Tür, Hakkani-Tür, Stolcke & Shriberg 2001). After having divided the speech signal into intonational phrases (primarily on the basis of pause duration), they computed the following as features of that phrase: the minimum, maximum and mean F0 in the 200ms immediately preceding and following the boundary, and the range of F0 in the phrase compared to that speaker's baseline. The relevance of these features to determining topic boundaries was discussed in the previous chapter. They found that their system performed slightly better when only prosodic features were included, as opposed to only lexical features. Combining lexical and prosodic features led to approximately a 4% reduction in the error rate. It is interesting to compare the performance of textual and prosodic features as this may give an indication of the information overlap between the two. That is, if combined performance with both types of features is not significantly better than with one or the other, then they may be capturing the same sort of information, and we are still missing a source of information from which humans determine topic boundaries. Pause duration was the single most important prosodic feature, although all the F0 measures combined were better than pauses at determining topic boundaries.

Hirschberg & Nakatani (1998) used much the same prosodic features in a decision tree trying to automatically determine topic boundaries in the Boston Directions Corpus and the NIST corpus. They also included the maximum, minimum and mean energy (in rms) as features. They concluded that it was feasible to determine topic boundaries in an audio signal solely using these prosodic features.

Silipo & Crestani (2000) studied the relationship between acoustic stress and the information content of words. In a corpus of monologues about a variety of subjects, they found that there was a high correlation between words that had pitch accents (as identified by a trained annotator) and the *tf.idf* score of the word. From this they concluded that the presence of a pitch accent is a good indicator that a word is a topic word.

3.3 Machine Learning Algorithms

The systems described above made use of a number of different machine learning algorithms, including HMMs (CMU, SRI, Mitre and Dragon), decision trees (IBM,SRI),

maximum entropy (CMU) and clustering (Iowa). In the NIST competition the CMU system performed best, then Dragon, IBM and SRI, with Iowa performing significantly worse (see Fiscus et al. (1999), Mitre was in the 2000 competition). It is difficult to separate out how much of each system's performance was due to the quality of the statistical techniques used and how much to the usefulness of its features. For this reason, the system built in this project makes use of two machine learning algorithms: decision trees and maximum entropy. As will be discussed, these methods are well-suited to classification tasks which involve a number of diverse features, the usefulness of which is not certain.

3.3.1 CART

Classification and Regression Trees (CART) models have been a standard method of building statistical models for many years now (Breiman, Friedman, Olshen & Stone 1984). They use the same theory as regression trees but are used to predicting categorical, as opposed to continuous, variables. They are therefore appropriate for this data set. The implementation used in this project is the *wagon* CART building program that is part of the University of Edinburgh's Speech Tools Library (Taylor, Caley, Black & King 1999). A trained CART tree asks a series of yes-no questions about features of a data set in order to classify it. In this case the features will be the textual and prosodic features described above and we wish to classify each sentence as either in a topic or at the beginning of a topic. The questions may be about discrete features of the data set, e.g. 'Does the word *however* occur in the sentence?'. Or they may be about continuous features. In this case the continuous values are 'binned' and then treated like discrete values, e.g. 'Is the pause before the sentence boundary greater than 50ms?'. Questions continue to be asked until a terminal node of the tree is reached at which point the sentence is classified.

The CART tree is built automatically from data in a training set, which is a set of feature vectors thought to be indicative of the testing set. At the beginning all the data is put at the root node of the tree. The program then asks all possible questions about the features of the data set, selecting the one that splits the data so that each new set has the least impurity (see below for definition). This continues until some stopping criterion is reached. In *wagon* this is when all the samples at one node have

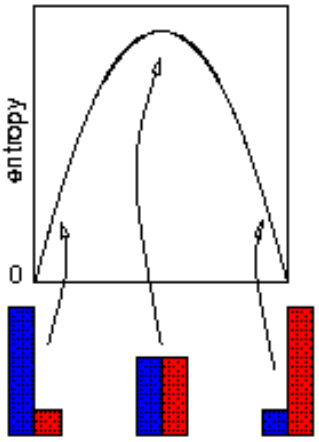

Figure 3.3: Entropy of Different Data Sets (King 2001)

the same classification or there are fewer than a certain number of samples at one node, whichever comes first. This stopping value is used so that the tree does not overfit to the training data. In the experiment reported below, a stopping value of eight samples was found by optimising over a held-out set when all features were included.

For discrete data, impurity is measured in terms of the entropy of each set multiplied by the number of data points. Entropy is calculated by formula 3.6,

$$H = \sum_{x} P(x) log(P(x)) \tag{3.6}$$

where $P(x)$ is the probability of a feature vector x (a sentence) having a certain label, which is then summed over all the xs in the set. As can be seen in Figure 3.3, which shows the entropy of three possible splits of a hypothetical data set, entropy is lower the more predictable a data set is. When entropy is 0, things are 100% predictable (for example, if the data set consisted entirely of entities of the same class, e.g. all sentences are at the beginning of a topic). For continuous data, the variance multiplied by the number of items in each data set is used. The split with the lowest total variance is taken.

CART is a good model to use in cases like these where there are lots of features, both

discrete and continuous, and it is not certain how useful each of them is. Another big advantage is that it builds trees which can be read by humans. It can therefore be used as a diagnostic tool to try to determine which features are the most important in the classification. Features which appear high up in the tree are likely to be important and those which do not appear at all are not. CART is fast to train and implement. However, it can have a problem dealing with sparse data as it is inherent in the system that smaller and smaller parts of the data set are used to make decisions as the tree is built.

3.3.2 Maximum Entropy

Maximum Entropy (Maxent) has become increasingly popular as a model for natural language processing applications in the past few years as it deals well with large numbers of potentially complex features and, unlike for example HMMs, makes no assumption of independence among the features the model takes into account (Manning & Schutze 1999, p.589). This makes it ideal for the kinds of features used in this study, for much the same reasons as were discussed in relation to CART models above. It has reached or surpassed the state-of-the-art in statistical classification tasks from part-of-speech tagging (Ratnaparkhi 1996) to text classification (Nigam, Lafferty & McCallum 1999).

The maximum entropy principle is summed up well by Berger, Pietra & Pietra (1996, p.3):

> "Intuitively, the principle is simple: model all that is known and assume nothing about that which is unknown. In other words, given a collection of facts, choose a model which is consistent with all the facts, but otherwise as uniform as possible."

The probability of a data point x in a log-linear model (like Maxent) is given by formula 3.7,

$$P(x|M) = \frac{1}{Z} exp(\sum_{i=1}^{n} f_i \lambda_i) \tag{3.7}$$

where Z is a normalising term,

$$Z = \sum_{y} exp(\sum_{i=1}^{n} f_i \lambda_i) \tag{3.8}$$

In other words, the probability of *x* given the model *M* is the expected value of all the features associated with it (f) multiplied by the weight associated with each feature (λ) to the power of *e*. Each feature has an associated weight, if the weight is 0 then that feature has no effect on the probability of *x*. If it is high, it means the presence of that feature makes *x* highly probable. If it is very negative, it means the presence of that feature makes *x* highly unlikely.

So how is the model arrived at? This is suggested by the principle above. The model is the probability distribution with the highest entropy that obeys the constraints of the training data. The constraints are the expected values of the features determined by the system designer in the training data. The weights are found automatically. In the implementation of Maxent used in this study, OpenNLP Maxent (Baldridge 2001), this is achieved through Improved Iterative Scaling, which is an Expectation-Maximisation algorithm (Osbourne 2000, p.3).

The OpenNLP framework allows discrete features to be entered and then the formulation of a Maxent model carried out automatically. It does not currently deal with continuous parameters so these had to be 'binned' before being entered into the system.

3.4 Summary

We are now in a good position to build a statistical classifier to predict topic boundaries in radio broadcast news. In this chapter we have identified a number of textual cues which can be used as features, including Hearst's lexical scores and the presence of cue words. We have also identified a number of prosodic cues which have been shown to be effective for the task, such as pause length and F0 mean and range. In the previous chapter, we suggested prosodic features based on a Tilt analysis of the intonational structure of each sentence. Finally, we have laid out two statistical models, CART and Maxent, with which the statistical classifier can be built.

As was alluded to at the beginning of this chapter, we are not only interested in achieving the best possible performance at the topic segmentation task. We also want to see the value of each of the different types of features used, and therefore the degree to which those features are actually used to indicate topic structure by speakers.

Chapter 4

Experiment

Over the last two chapters we have been building up the theoretical bases on which to construct a statistical classifier to do topic segmentation. We will now describe an experiment which was carried out to evaluate how effective the different cues to topic structure set out above are, in particular comparing the performance of textual, acoustic and ToBI features. Below we will describe how the task was set up, the corpus used and how the different features were calculated. We will then set out the results of the experiment and compare these to previous attempts at the topic segmentation task reported in the previous chapter.

4.1 The task

In this experiment, topic segmentation of broadcast radio news is treated as a binary classification task. The classifier's job is to decide if each sentence is at the beginning of a topic (BTOPIC), or in a topic (ITOPIC). For simplicity's sake, sentences are taken as marked in the transcript associated with our corpus.[1]

The classifier is trained on both textual and prosodic information associated with the corpus. The aim of the experiment is to assess the effectiveness of each type of feature (textual, acoustic and ToBI) in the topic segmentation task. In particular, we hypothe-

[1] Obviously, in a real system, both word recognition and sentence boundary detection would have to be carried out automatically. However, as these tasks are somewhat orthogonal to topic segmentation, it seems legitimate to imagine here that they are being carried out by a previous module in the system.

sise:

1. Results should improve with the addition of each set of features.
2. Textual features alone and acoustic features alone (durational and F0) should yield approximately equal performance but combined should lead to a significant improvement in performance.
3. Performance with ToBI features should equal or better performance with acoustic features.

In addition, as was discussed in the previous chapter, two statistical models will be used. This enables us to evaluate how effective each model is at performing this kind of task.

4.1.1 Boston University Radio Corpus

The corpus used in this study is the Boston University Radio News Corpus (Ostendorf, Price & Shattuck-Hufnagel 1994). It is a corpus of professionally read radio news data, including speech and accompanying annotations, designed for speech and language research applications. Relevant files provided with the corpus include:

- **Sound:** NIST SPHERE-format waveform files used by speech corpora in the LDC database (*Linguistic Data Consortium* 2002). The stories were digitised into paragraph units, typically about the size of one topic.
- **Text:** Orthographic transcripts were done by hand. These have then been manually annotated with XML-like sentence and topic boundary markers. (See Appendix A for an example).
- **F0:** SPS-format pitch files (Entropic-Research-Labs 1998).
- **Phones:** Time-aligned phone labels were generated automatically using a stochastic segment model constrained by the orthographic transcription of the sentence (Kimball, Ostendorf & Bechwati 1992). The TIMIT phonetic labelling system was used (Zue, Seneff & Glass 1990).

- **Word:** Word boundaries were found automatically from the phone files and transcriptions.

- **ToBI:** The first ten stories in this section were manually annotated with ToBI labels. These were used to train the Tilt models which were in turn used to approximate ToBI features.

One section of the corpus was used, that of a female speaker, *f1a*. The corpus was limited to one speaker to avoid issues of between-speaker variability. This section was chosen as it contained short news bulletins more suitable for topic segmentation (as opposed to longer, single story news items). There were 40 stories in all, comprising 52 minutes of speech and 610 sentences. The first sentence of each story was omitted because it was completely predictable that it would be BTOPIC, leaving a data set of 570 sentences. Of these, 168 sentences were BTOPICs, or 29.5% of sentences. Training and testing were carried out using five-fold cross validation. This method is said to lead to more reliable results than simply using one testing and training division, especially when the sample size is small (Bailey & Elkan 1993). The corpus was divided into five blocks of eight stories each. Each test was then run five times, each time with a different block as a testing set and the other four used to train the model. The results from the five runs were then averaged to get the final results. For some of the features reported below, a held-out set was needed to tune parameters. In this case, one of the four blocks in each training set was used as the held-out set for that block, and the parameters found for that block were used for that block only.

The features set out below were extracted automatically from all training and testing sentences using the methods described from the above files in the corpus.

4.1.2 Textual Features

Two types of textual features were extracted, Hearst's lexical scores and cue phrases. The Text files were used to get the relevant words. Lexical scores are fully described in section 3.1.1 above. A list of common words found on Hearst's Text Tiling website (Hearst 1995) was used to remove stop words. The normalised scores for continuation (words in both sections) and new words (number of new words in second section) were then entered as features.

Cue phrases were extracted from each training set using the mutual information criterion described in section 3.1.2. Bigrams were extracted using the CMU-Cambridge Language Modeling Toolkit (Clarkson & Rosenfeld 1997). Then the mutual information score of each unigram (word) and bigram in a sentence and the classification of that sentence (BTOPIC or ITOPIC) was calculated as described in section 3.1.2. Those unigrams or bigrams which had a higher mutual information score with BTOPIC than ITOPIC than vice versa were retained as cue phrases. However, it was found that this led to too many phrases being found, some of which were clearly topic words not cue phrases. So, a threshold difference between the mutual information score for BTOPIC and ITOPIC was employed. As well, a minimum mutual information score and minimum frequency for the bigrams and unigrams was used. All three values were found by optimising on a held-out set, so that the cue phrases found were more general, rather than specific to one story, or set of stories. Even still, some of the cue phrases are specific to this data set. For example, *Massachusetts* comes up as a cue phrase, as this must be a frequent enough word in a corpus of Boston radio news. Once the cue phrases have been found, there presence or absence in each sentence is entered as a feature.

4.1.3 Duration Features

Four different types of duration features were calculated, each trying to capture the generalisation made in section 2.2.1 above that speakers slow their speaking rate over the course of a topic. These were the length of the pause before the sentence, the speaking rate, the amount of final lengthening and the total duration of each sentence.

To find the length of the pause before the sentence, the Phone files were used to get the end time of all the phones in each story. Then the Word and Text files were used to get the times of the beginning and end of each sentence. All pauses or breaths (as transcribed in the Phone files) between the two sentences were found and their cumulative duration was entered as a feature on the second sentence. Unfortunately, the Sound files came divided into units shorter than one story. It was not clear whether information about pausing and breaths was retained when the files were divided. Instead the amount of ‘nothing’ (transcribed as H#) at the beginning and end of each Sound file was entered as the pause time. This may have led to some inaccuracies, particularly

since the majority of BTOPICs occurred at the beginning of Sound files.

The speaking rate was calculated in words per second. The Text file was used to find the length in words of each sentence. This was then divided by the duration of the sentence, taken as the length in seconds between the end of the first word in the sentence and the end of the last. Although this effectively omitted the first word from the rate, it significantly simplified the calculations because of how the annotation files were set up. Hirschberg & Nakatani (1998) used syllables per second in their calculation of speaking rate. This would probably be a more accurate measure as syllables vary in length considerably less than words. However, given that syllables were not annotated in the corpus, this would have been difficult to implement. Words per second is used successfully by some researchers in the speech synthesis community to vary speaking rate (e.g. Arons 1992).

To calculate the amount of final lengthening, the mean and standard deviation of the length of all sonorants and fricatives in the training set was found from the Phone files. Others sounds were not included as they do not vary in length very much compared to sonorants and fricatives. Vowels with primary stress were annotated differently from those without, so these were kept separate. The Text, Phone and Word files were then used to get the last three phones in each sentence. This was done to approximate the last syllable in the last word of the sentence. If any of these sounds was either a sonorant or a fricative, its length was recorded. This length was then normalised with reference to the average length of that phone in the training data according to formula 4.1:

$$Z_i = \frac{x_i - \overline{x_i}}{s_i} \tag{4.1}$$

where x_i is the length of the phone in seconds, $\overline{x_i}$ and s_i are the mean and standard deviation of the phone in the training set. The normalised length of all the relevant phones at the end of each sentence were then averaged. The score, giving the average amount of lengthening, was then entered as a feature.

The final measure of duration is the length in seconds of each sentence. Although it may seem a crude measure, Wright (2000) found the feature to be helpful in her work on automatically classifying dialogue acts.

4.1.4 F0 Features

Various features extracted from the F0 contour of each sentence are calculated using a program developed by Paul Taylor and Helen Wright Hastie (Wright 2000). The program works on entire F0 files and their corresponding energy values, so the Sound files first had to be divided into files of sentence length. The division was done using an esps tool (Entropic-Research-Labs 1998). An esps tool was also used to create an F0 file for each sentence length sound file.

The first set of features extracted by the program are global F0 features, attempting to capture the overall F0 level of the sentence. These are the maximum F0, as well as the mean F0 and energy over the utterance.

The second set of features try to capture the nature of the boundary of the sentence. The means of the F0 and energy of the final 200ms of the sentence, and the penultimate 200ms are taken. These values are normalised against the sentence means and entered as features. The figure of 200ms was taken from a similar study by Shriberg, Taylor, Bates, Stolcke, Ries, Jurafsky, Coccaro, Martin, Meteer & Ess-Dykema (1998). The ratios of the mean F0 and energy in the end and penultimate regions were also entered as features. The point of these six features was to measure whether the F0 at the boundary was falling or rising and how dramatically.

4.1.5 ToBI Features

As was discussed in section 2.4, we are using Tilt features (Taylor 2000) to approximate ToBI features in this study as only a small portion of the corpus had been annotated with ToBI features. Extraction was carried out using the system designed by Taylor which uses the HTK toolkit (Young, Kershaw, Odell, Ollason, Valtchev & Woodland 2000). This involves firstly the parameterisation of the F0 contour into a form suitable for HTK. The resulting series of feature vectors (F0, differentiated F0, energy and cepstral coefficients) is then segmented into intonational events using Viterbi recognition of HMMs of each event. Finally, Tilt parameters are extracted for pitch accents and boundary tones. Each set of Tilt parameters was then classified in terms of ToBI labels.

Taylor's system came with HMMs for the intonational events *a* (pitch accent), *b* (boundary tone), *ab* (combination accent+boundary), *c* (continuation) and *sil* (silence). These were trained on the DCIEM Map Task corpus (Bard, Sotillo, Anderson, Thompson & Taylor 1996). Unfortunately, it was found that the performance of these HMMs in picking out intonational events with *f1a* was not very good. Therefore, it was decided to retrain the HMMs using the HTK toolkit. The original HMMs were used as the starting point. Then the ten stories in the corpus which had been marked with ToBI labels were used to provide training data for new HMMs. Each of the ToBI labels in these ten stories was changed into a suitable intonational event label as follows:

- L* and L% accents became *c* as the Tilt system was designed only to recognise pitch excursions.
- H*, L+H*, L*+H and X* (unknown pitch accent) and all downstepped variants became *a*.
- H-, HL%, HH%, LH%, X- and X% (unknown boundary tones) became *b*.
- If an *a* event preceded a *b* event by less than 15ms, it was called an *ab*. This was decided partly by inspection of a number of the training files, and because each feature vector used for Tilt training represents 10ms of speech, so it would be unlikely to be able to recognise two events so close together.
- If two *a* or *b* events were more than 35ms apart, a *c* event was put between them. This figure was decided by inspection of a number of the training files.
- The label file was corrected so that there were never two *c*s in a row and a *sil* label was added to the beginning and end of each file (sentence length).

Unfortunately, the ToBI labels only gave the time for the peak of each accent, not the start and end points of each intonational event, so the times in the original files could not be used to do the HMM training. Therefore the label sequence obtained from the above process for each sentence was used to find the HMMs using embedded training. This was iterated a number of times.

The resulting HMMs were tested using Viterbi recognition on the same training set. The word accuracy rate algorithm was applied to accent recognition to evaluate the

results, so the label sequence in the training file was compared to the found label sequence. According to this measure, the number of iterations of embedded training, the language modelling scaling factor (the weight put on the word model (the event HMM) as opposed to the language model (bigram)), and the word insertion probability (controlling the number of loops as opposed to transitions between HMMs in the Viterbi training) were optimised. Although it is not good practice to test and train on the same data in this manner, it seemed expedient given the very small amount of training material available. The performance of the new HMMs was still somewhat disappointing, the accent detection accuracy score being only around 40%. However, it was the best that could be done with the present data.

Using the retrained HMMs, intonational events were identified for all the files in the corpus and Tilt parameters (start F0, amplitude, duration, Tilt and position, see section 2.4 for details) extracted for *a* and *b* events. As was discussed in section 2.4 with reference to Figure 2.5, the Tilt values were used to approximately identify ToBI features as follows:

- *a* events with Tilt less than 0 were called H*, as they have no discernible L-like rise.
- *a* events with Tilt greater than 0 were called X*, as they could be L+H*, L*+H or even H* (though the last is less likely).
- *ab* events with Tilt less than 0 were called X*L-, as they represent some sort of pitch accent followed by a fall in F0.
- *ab* events with Tilt greater than 0 were called X*H-, as they represent some sort of pitch accent followed by a rise in F0.
- *b* events with Tilt greater than 0 were called H-, as they show a rise in F0.

Since there was no way of distinguishing between intermediate phrase boundaries and intonational phrase boundaries, it was decided that H- could also be H% and so on. Obviously, this classification system is somewhat rough. This is particularly true for events with Tilt values around 0. To try to assess how much of an impact these approximations were having on the results, two further sets of values were tested. Firstly, the Tilt values themselves were entered as features, as will be detailed below. Secondly,

the performance of the actual ToBI features as annotated in the corpus was tested directly against the derived ToBI features for the ten stories for which annotations were available. For the purposes of this test, the stories were divided into seven for training and three for testing, approximately a 75-25% split of the data set of 135 sentences.

Finally, the entire data set with derived ToBI annotations was used to extract features for each sentence as are suggested by the literature presented in Chapter 2. These are as follows:

- **Pitch Accents:** Hearst's lexical scores described in section 3.1.1 were reformulated so that instead of using stop words to get rid of redundant information pitch accents were used. For the continuation score, only words which had a pitch accent (H*, X*, X*L- or X*H- or any other pitch accent in the original ToBI annotations) before a potential boundary and no pitch accent after the boundary were included. For the new words scores, only words that appeared after the boundary for the first time with a pitch accent were included. These scores were normalised in the same way as the other lexical scores. The resulting features were trying to capture the evidence presented about accenting and deaccenting of given and new information in section 2.2.2.

- **Words:** These lexical scores were again recomputed, this time following the theories about marking of given and new information within the ToBI system presented in section 2.3.2. Therefore, words were only included in the continuation metric if a word before the boundary had an H* pitch accent (plus X*L- in the derived annotations as this was highly likely to actually be H*, and !H* and H+!H in the original annotations) and the same word following the boundary had either no accent or an X* accent (or X*H-, as these could be L+H*) in the derived annotations and L+H* or L+!H* in the original annotations. For the new score metric, the new word had to have an H* accent (or variants as above).

- **Tunes:** Finally, the boundary tones said to indicate the beginning of a topic and the continuation of a topic at the end of section 2.3.2 were added as features. It was noted whether there was a H* (or variants) within the first three words of the sentence (which should indicate BTOPIC). Then whether there was a L+H* (or variants) within the first three words (which should indicate ITOPIC). It was recorded if there was an H-, the theme tune (or X*H- in the derived annotations,

and L-H% or H- in the original annotations) in the first seven words of the sentence (to approximate the intermediate phrase). Finally, it was noted if the last or second to last word in the previous sentence had a H- (or variants, plus H*, as this would probably be a misclassified boundary tone in this position), which should indicate the present sentence is ITOPIC. For the comparison with Tilt, the value of Tilt was entered as a feature in each of the three positions (pitch accent in first three words, boundary in first seven words, final boundary). If there was no accent in the relevant position, it was given a Tilt value of 1.1, being outside the Tilt range (so all the non-accents would be binned together).

4.1.6 Bins

All of the F0 and durational features, as well as Hearst's lexical scores, are continuous variables. Unfortunately, both CART and Maxent are designed to work with discrete features. (Although CART can handle continuous features, as is explained below). Therefore, the continuous variables had to be 'binned', i.e. divided into bins with approximately equal frequencies of values in each, so they could be treated as categorial.

The *wagon* CART building program, see section 3.3.1 above, is set up to deal with continuous variables, and provides a flag to change the number of bins into which each is divided. A script to do this had to be created for the Maxent program. The number of bins to be used was found for each classifier by optimising over a heldout set when all the continuous features were included.

Unfortunately, this is not an ideal solution as different variables would be optimised with different numbers of bins. However, it was decided that it was not a good idea to try to have a different number of bins for different variables. Firstly, this would loose information about variance within each bin which *wagon* uses to calculate impurity (as the categorisation would have to be done before the features were entered into CART). Secondly, the setup of the Maxent software seemed to lead to suspiciously low results when dealing with only one feature, in some cases getting zero F-scores (see below), so it would be difficult to determine what the optimum number of bins should be. Therefore the system is optimised as described above, with the caveat that scores may be effected by around 1-2% because of binning problems.

4.2 Results

We now present the findings of the experiment described above. Firstly, results are given in terms of the performance of the system with different features. Secondly, we briefly outline which features were most important in terms of weights/position in the tree for each classifier. Lastly, we present a comparison of the different ways of deriving intonational event features.

4.2.1 Performance

Tables 4.1 and 4.2 show results using the CART classifier and the Maxent classifier respectively. Results for each feature or group of features are given in the left-hand column, and the results when successively combining all the figures in the right-hand column. The performance of the acoustic features is grouped both together and into duration and F0 features. The performance of just the acoustic features is at the bottom of the left-hand column. The performance of the system when Tilt features representing tunes were entered directly is given separately before the performance with derived ToBI features. In addition, the combination of textual and F0 features, textual and Tilt and textual and ToBI features is given in the left-hand column.

The precision, recall, F-score are given for each result. These as used as they are thought to give a better indication than accuracy scores of the relative contribution of each feature because they are not swamped by the number of ITOPICs correctly identified (see Provost, Fawcett & Kohavi 1998), and we are generally more interested in finding the beginnings of topics than confirming that most sentences are within topics. Precision is a measure of the total number of BTOPICs the classifier correctly identified over the total number of BTOPICs it identified. Recall is the number of BTOPICs the classifier correctly identified over the total number of BTOPICs in the corpus. The F-score is a commonly used mixture of the two, calculated using formula 4.2:

$$Fscore = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.2}$$

	Separately			Cumulative		
Feature	Prec	Recall	F-score	Prec	Recall	F-score
Text Tiling	54.2	60.3	**57.1**	54.2	60.3	**57.1**
Cue Words	66.1	70.7	**68.3**	76.2	74.0	**75.1**
Textual	76.2	74.0	**75.1**	76.2	74.0	**75.1**
Pause Length	31.5	46.0	**37.5**	80.4	76.7	**78.5**
Speaking Rate	7.1	22.2	**10.8**	76.2	76.2	**76.2**
Final Lengthening	22.0	36.7	**27.5**	76.8	75.4	**76.1**
Sentence Duration	24.4	41.8	**30.8**	78.0	74.0	**75.9**
Duration	39.8	47.9	**43.5**	78.0	74.0	**75.9**
Global F0	73.8	73.8	**73.8**	75.6	78.4	**77.0**
Boundary Features	32.1	32.5	**32.3**	76.8	79.1	**77.9**
Pitch Accents	25.6	38.7	**30.8**	76.8	79.1	**77.9**
F0 Features	69.0	69.5	**69.3**	76.8	79.1	**77.9**
Textual+F0	80.4	79.9	**80.1**2	-	-	-
Acoustic	72.0	74.7	**73.3**	76.8	79.1	**77.9**
Tilt tunes	14.9	42.3	**22.0**	76.8	78.7	**77.7**
Textual+Tilt	76.8	74.1	**75.4**	-	-	-
ToBI tunes	0.0	0.0	**0.0**	76.2	80.5	**78.2**
Words	26.8	49.5	**34.7**	75.0	84.6	**79.5**
Textual+ToBI	78.6	75.9	**77.2**	-	-	-
ToBI	27.4	36.5	**31.3**	75.0	84.6	**79.5**
All	-	-	-	75.0	83.4	**79.0**

Table 4.1: Results using CART

	Alone			Cumulative		
Feature	Prec	Recall	F-score	Prec	Recall	F-score
Text Tiling	68.3	65.1	**66.7**	65.7	67.7	**66.7**
Cue Words	52.1	77.0	**62.1**	68.3	77.6	**72.6**
Textual	68.3	77.6	**72.6**	68.3	77.6	**72.6**
Pause Length	13.2	40.7	**19.9**	70.1	76.0	**72.9**
Speaking Rate	0.0	0.0	**0.0**	68.3	73.5	**70.8**
Final Lengthening	5.4	39.1	**9.5**	69.5	75.8	**72.5**
Sentence Duration	22.2	49.3	**30.6**	67.1	73.7	**70.2**
Duration	38.9	53.7	**45.1**	67.1	73.7	**70.2**
Global F0	77.8	74.2	**76.0**	81.4	76.0	**78.6**
Boundary Features	28.1	49.5	**35.9**	82.0	77.0	**79.4**
Pitch Accents	35.3	67.8	**46.5**	81.4	77.7	**79.5**
F0 Features	77.2	79.1	**78.1**	81.4	77.7	**79.5**
Textual+F0	80.2	77.9	**79.1**	-	-	-
Acoustic	77.8	79.8	**78.8**	81.4	77.7	**79.5**
Tilt tunes	31.1	48.6	**38.0**	80.8	77.1	**78.9**
Textual+Tilt	70.7	77.1	**73.8**	-	-	-
ToBI tunes	7.8	40.6	**13.1**	82.6	80.2	**81.4**
Words	26.9	60.0	**37.2**	81.4	79.1	**80.2**
Textual+ToBI	71.3	78.8	**74.8**	-	-	-
ToBI	40.7	57.6	**47.7**	81.4	79.1	**80.2**
All	-	-	-	82.0	80.1	**81.1**

Table 4.2: Results using Maxent

Group	CART	Signif	Maxent	Signif
Baseline	70.5	-	70.5	-
Duration	69.5	No (t=.37)	68.8	No (t=0.62)
ToBI	64.6	Yes (t=2.13)	70.6	No (t=-0.01)
Tilt	68.9	No (t=0.59)	66.4	No (t=1.49)
F0	81.9	Yes (t=-4.56)	85.8	Yes (t=-6.36)
Dur+F0	84.6	Yes (t=-5.79)	86.2	Yes (t=-6.56)
Textual	85.1	Yes (t=-6.02)	83.0	Yes (t=-5.05)
Text+Dur	85.4	Yes (t=-6.17)	81.2	Yes (t=-4.25)
Text+Tilt	85.3	Yes (t=-6.12)	83.4	Yes (t=-5.23)
Text+ToBI	86.3	Yes (t=-6.61)	84.2	Yes (t=-5.60)
Text+Dur+F0	87.2	Yes (t=-7.05)	86.1	Yes (t=-6.51)
Text+F0	88.2	Yes (t=-7.57)	86.0	Yes (t=-6.46)
All	88.2	Yes (t=-7.57)	87.4	Yes (t=-7.15)

Table 4.3: Results by Group: Accuracy

Table 4.3 shows the results of each grouping of features with the two classifiers in terms of accuracy. Accuracy is a measure of the total number of sentences the classifier identified over the total number of sentences, and is commonly reported in statistical studies. This score enables us to compare the performance of the system to the baseline, which is the result if we just had a system that chose the most likely tag (ITOPIC) for each sentence. This should be compared with a topline of 100% (since it is assumed human annotators would be completely accurate at performing topic segmentation with this corpus). Here we can see that the performance of the duration and the ToBI features alone are below or at the level of the baseline. These features are therefore not performing well. All other features perform significantly better than the baseline (using a two-tailed t-test at the 95% significance level).

There is no statistically significant difference in results between any other groupings of features for both the classifiers, although the performance with a combination of textual and F0 features is significantly better than for the F0 features on their own with CART (t=-2.99).

```
((utt_f0_mean < 169.923)
 ((cue_word_today is 0)
  (((BTOPIC 0.0300429) (ITOPIC 0.969957) ITOPIC))
  ((utt_f0_mean < 159.283)
   (((BTOPIC 0) (ITOPIC 1) ITOPIC))
   ((norm_end_nrg_mean < -0.549097)
    (((BTOPIC 0.25) (ITOPIC 0.75) ITOPIC))
    (((BTOPIC 0.777778) (ITOPIC 0.222222) BTOPIC)))))
 ((cue_word_Massachusetts is 0)
  ((cont_score < 0.0940721)
   ((cont_score < 0.0312348)
    ((cue_word_brth is 0)
     ((utt_f0_mean < 195.062)
      (((BTOPIC 0.25) (ITOPIC 0.75) ITOPIC))
      (((BTOPIC 0.909091) (ITOPIC 0.0909091) BTOPIC)))
     (((BTOPIC 0.888889) (ITOPIC 0.111111) BTOPIC)))
    ((cue_word_Boston is 0)
     (((BTOPIC 0.03125) (ITOPIC 0.96875) ITOPIC))
     (((BTOPIC 0.666667) (ITOPIC 0.333333) BTOPIC))))
   (((BTOPIC 0.08) (ITOPIC 0.92) ITOPIC)))
  (((BTOPIC 0.933333) (ITOPIC 0.0666667) BTOPIC))))
```

Figure 4.1: CART tree when all features are included

4.2.2 Features

Figure 4.1 shows the final cart tree produced when all features were included. `utt_f0_mean` refers to the mean F0 of the sentence. The various `cue_word`s refer to the presence or absence of that word in the sentence. `norm_end_nrg_mean` means the normalised mean energy in the final 200ms of the sentence. `cont_score` refers to Hearst's continuation lexical score. Below each feature is given the resulting classification from the features above it and the associated probability. For example, if it says `((BTOPIC 0.0300429) (ITOPIC 0.969957) ITOPIC)`, this means there is a 3.0% probability the data point

Feature Type	FUF (%)
utt_f0_mean	35.6
cue_words	35.6
cont_score	24.4
norm_end_nrg_mean	4.4

Table 4.4: Importance of Features by FUF

is BTOPIC, but a 97.0% probability that it is ITOPIC, therefore the system would classify it as ITOPIC.

Table 4.4 shows the importance of each of the features in this tree using the Feature Usage Frequency (FUF) metric. This is calculated by dividing the number of ways a data point could pass through a particular feature (a node) in the tree by the total number of ways it could pass through any node in the tree. Where a feature type appears more than once, the scores are added together.

Table 4.5 shows the features in the Maxent classifier that had weights of either less than -0.5 or more than 0.5 for either BTOPIC or ITOPIC along with their weights for each classification. A very negative weight means the presence of that feature makes the classification highly unlikely. A very positive weight makes the presence of the feature highly likely (see section 3.3.2 for further explanation).

Feature labels are as above, except that numbers beside the features refer to the number bin that that feature is in (see above for discussion on binning). `beg_a` refers to the Tilt value of the pitch accent in the first three words of the sentence. `f0_diff` refers to the difference in the F0 level in the penultimate and final regions of the F0 contour. `new` refers to Hearst's new lexical score. `cont_X` is the continuation lexical score, except using pitch accents to pick out the topic words. The other features should be self-explanatory.

Feature	BTOPIC	ITOPIC
beg_a=2	-1.53	0.13
f0_diff=6	-1.27	0.13
new=6	0.44	-1.22
new=6	-1.22	0.18
cue_word_says=false	-1.22	0.14
end_f0=3	-1.03	0.13
pen_f0=4	-0.99	0.08
pen_f0=1	-0.83	0.12
cue_word_be=true	-0.78	-0.21
cue_word_Massachusetts=true	0.24	-0.73
new=1	-0.73	0.19
rate=6	-0.71	0.08
max_f0=6	0.24	-0.70
pause=5	-0.70	0.18
new_H=1	-0.55	0.09
cont_X=6	-0.52	0.09
pen_f0=5	-0.51	0.11

Table 4.5: Important Features in Maxent Classifier

	CART				Maxent			
Whole Data Set								
Feature	Prec	Recall	F-score	Accur	Prec	Recall	F-score	Accur
Baseline	-	-	-	70.5	-	-	-	70.5
Tilt (tunes)	14.9	42.3	**22.0**	68.9	31.1	48.6	**38.0**	66.4
Tunes	0.0	0.0	**0.0**	74.2	7.8	40.6	**13.1**	82.6
Words	26.8	49.5	**34.7**	75.0	26.9	60.0	**37.2**	81.4
ToBI	27.4	36.5	**31.3**	75.0	40.7	57.6	**47.7**	81.4
ToBI annotated data - 10 stories								
Baseline	-	-	-	74.2	-	-	-	74.2
Tunes (auto)	0.0	0.0	**0.0**	67.7	0.0	0.0	**0.0**	67.6
Words (auto)	0.0	0.0	**0.0**	74.2	45.5	62.5	**52.6**	73.5
ToBI (auto)	12.5	25.0	**16.7**	67.7	54.5	66.7	**60.0**	76.5
Tunes (hand)	0.0	0.0	**0.0**	74.2	9.0	33.3	**14.3**	64.7
Words (hand)	0.0	0.0	**0.0**	74.2	18.2	100.0	**30.8**	73.5
ToBI (hand)	12.5	33.3	**18.2**	70.8	63.6	63.6	**63.6**	76.5

Table 4.6: Comparison of Intonation Feature Extraction Methods

4.2.3 Tilt Features

Table 4.6 shows two comparisons that attempt to establish how much the process of deriving ToBI values affects results. The first part of the table compares results using the entire data set when Tilt values are entered directly as tune features to performance with the derived ToBI values. As can be seen, Tilt features were worse than ToBI in terms of accuracy (this is significant only with the Maxent classifier, using a two-tailed t-test at the 95% significance level). They were, however, significantly better with both classifiers in terms of F-scores. The second part of the table compares performance on the ten stories that were annotated with ToBI data, firstly when features were found using the derived ToBI annotations, and secondly when the original ToBI annotations were used. As can be seen, results are slightly better with the original annotations, both in terms of accuracy and F-score. However, these results are not significant.

4.3 Evaluation

It is difficult to compare these results directly to those of the studies discussed in the preceding chapter, however, we will attempt to draw some general conclusions here. We will also briefly evaluate the results in terms of the hypotheses above. Finally, we will compare the performance of the two statistical classifiers on this task. In the next chapter we will discuss what these results mean for the discussion of different theories about intonation in Chapter 2.

4.3.1 Comparison to Previous Studies

The performance of the systems in the NIST competition reported above was measured in terms of 'segmentation cost'. This was a score based on the number of words the system missed the real topic boundary by. In that competition, sentence boundaries were not given and so some systems used other methods to divide the text into blocks initially. The results cannot therefore be compared directly to this study. In addition, the corpus used in that competition comprised about 60,000 news stories, obviously making the training a lot more reliable. It did contain multiple speakers, though, making use of acoustic features more difficult.

Stolcke et al. (1999) do, however, compare the use of textual and acoustic cues. They found that the performance of textual cues and prosodic cues alone was approximately the same, but that performance improved by approximately 4% when the two were combined. That indeed seems to correlate with the findings here, although the results are not statistically significant. The improvement in results when both types of features are combined is slightly less than Stolcke et al. (1999) found.

Stolcke et al. (1999) also report that the top feature on their decision tree when all the features were combined was pause length. This does not appear to have been a very reliable feature in our system. The final tree produced by the CART model (see Figure 4.1) does not even include pause length as a feature and it is relatively low down on the decision tree for acoustic features although it important in the Maxent classifier. Pause length on its own as a feature only leads to an F-score of 20-37%. This could be because of differences in the speaking styles in the two corpora. Or, it could also be because the pause length feature was not accurate enough given the way the sound

files were set up (see the discussion in section 4.1.3 above).

Hirschberg & Nakatani (1998) include only acoustic features, similar to those used in this study. They report performance in terms of precision and recall, getting an F-score of about a 40%. However, this is a measure of the number of 10ms frames of speech which the system identifies as being in a break between topics, which is a quite different and harder task. Nevertheless, we do confirm that such acoustic features are reasonably reliable in identifying topic boundaries.

4.3.2 Hypotheses

We can make some general remarks about the hypotheses stated above. The first hypothesis was that results should improve with the addition of each set of features. Although the performance of the combined system is better than with any one of the sets of features, the performance with both classifiers is sightly worse when the durational features are added to the textual features than with the textual features on their own. As discussed in the previous section, this may be because the way the durational features were calculated was not accurate enough, or because the features chosen did not capture the prosodic phenomena they were trying to sufficiently well. It may also be because pause and duration are more indicative of lower level structures such as sentences than topics. This will be discussed more in the next chapter.

The second point to note is that the boundary F0 features perform significantly worse than the global F0 features, and in fact worsened the performance of one classifier when just trained on F0 features. We will return to this point in the next chapter.

The second hypothesis was that textual and acoustic features alone should yield approximately equal performance, while combined they should lead to a significant improvement in performance. We confirm that textual and acoustic features are approximately as effective as each other in identifying topic boundaries. When both types of features are combined there is an improvement in performance, although it is only statistically significant with the CART classifier. Again, we will return to what this may show about how we signal topic structure and the extent to which listeners rely on both types of cues.

The third hypothesis was that performance with ToBI features should equal or bet-

ter performance with acoustic features. Here the results are less than encouraging. The combination of textual and ToBI features performs slightly worse than textual or acoustic ones with Maxent, although not significantly worse. The performance with CART is the same. The performance of the ToBI features alone is equal to or worse than the baseline in terms of accuracy, and the F-score is less than 50% with both classifiers. The results of this experiment seem to show that although acoustic features are reasonably accurate when used on their own to recognise discourse structure, ToBI features are not. As will be discussed in the next chapter, it is unclear whether this is as a result of faults in the theory relating ToBI features to discourse structure from the last chapter, because of problems with the classification of ToBI features or because of the many approximations that had to be made in deriving ToBI features in this system. Results were slightly better when using the original ToBI annotations on a smaller data set, but not markedly so, which may suggest problems with the theory itself. However, firm conclusions should not be drawn on this point as the data set was very small. Results in terms of F-score were also better when using Tilt values on their own to identify tunes, although performance was slightly worse when these features were combined with textual and acoustic features. It is difficult to tell what this means, although it may indicate that some important information is lost with classifying intonational events categorially, as required by ToBI.

4.3.3 Classifiers

The results above show that CART and Maxent achieve approximately equal levels of performance. CART models are in general more stable, and seem to lead to better performance with fewer features. The Maxent model varied considerably with seemingly small tweakings of different parameters, although it gave better performance with large numbers of features.

Both models seem to be useful for this type of task, where there are large numbers of features to deal with. However, unlike it is claimed in the literature, the performance of both appears to be affected by the inclusion of too many spurious features (as performance goes down when less effective features are added).

4.4 Summary

In this chapter we have evaluated the usefulness of the various textual, acoustic and ToBI features which may indicate topic structure in a statistical classification task using a corpus of broadcast radio news stories. We found that durational, Tilt and ToBI features alone are unreliable as indicators of topic structure. Textual and F0 features are equally effective on their own, and a combination of textual and F0, textual and ToBI or textual and Tilt are also effective, in decreasing order of effectiveness. In the final system including all features, only the utterance mean F0, cue words, the continuation lexical score and the energy in the last portion of the sentence were included as features by the CART classifier. In the Maxent classifier the Tilt value of the first pitch accent in the sentence was the most important feature, although cue words, boundary F0 features and new lexical scores using ToBI featured highly.

In the next chapter, we will discuss what these results mean in terms of the theories presented in Chapter 2. However, as will also be noted then, it will always remain unclear how much of the performance is due to the value of the features and how much to the accuracy (or lack of) with which they were able to be calculated.

Chapter 5

Discussion

We now return to the questions suggested at the end of Chapter 2. They are separate but impact upon each other. The first is to ask whether it is more useful for ASR to view the discourse information conveyed by prosody as extractable directly from acoustic cues quantified separately, or whether prosody should be viewed as a series of events (identified primarily by acoustic information) from which discourse information can be derived. In this case, how should these events be represented? The second question is to ask what humans do. Do we map acoustic cues in the speech we hear directly to determine the discourse structure intended by the person we are listening to? Or do we identify intonational events which in turn allow us to elicit discourse structure? The present study is concerned with the first question. However, the second question is important as it is our knowledge as speakers that informs, at least in part, the type of information we input as features to ASR systems. In return, it seems reasonable to assume that if a certain feature or set of features works well in an ASR system, then it captures an important feature or set of features of human speech. (The broader question of the extent to which ASR designers should seek to mimic human speech production and recognition processes and to what extent psycholinguists should be influenced by their findings remains open).

The experiment reported in the previous chapter can be viewed as a case study of a small part of this problem, i.e. deciding how intonation should be represented in automatic speech recognition (and synthesis). We were looking at the extent to which different prosodic cues to topic structure identified in the linguistic and psycholinguistic literature are helpful in finding topic boundaries in a corpus of broadcast radio news.

This is a good medium in which to test the effectiveness of using acoustic cues versus intonational events to identify topic structure. The types of acoustic indicators being measured have already been shown to mark lower level structure (e.g. words and sentences) so we can see if they are rich enough to accurately identify higher level topic structure as well. Although the theory of intonational marking of information status and discourse structure is reasonably well developed, it has not been tested in an application like this. In addition, it is at higher levels of discourse analysis that a motivating theory becomes more necessary, e.g. though a working definition of a word suitable for ASR is reasonably easy to find, it is not so for a topic. It would be helpful for ASR to have a phonologically motivated (and automatically recognisable) definition of topic boundary. (Though of course this must go hand-in-hand with a semantic definition).

We will firstly review which acoustic features were the most effective in identifying topic structure in the radio news corpus and draw some conclusions from this. We will then evaluate the performance of ToBI features and assess how useful the features as they stand are for ASR. We will also go over some of the practical reasons why performance with these features was not as good as could be hoped and the implications of this. Finally, we will make some suggestions about what a theory of the intonational marking of topic structure should look like.

5.1 Effective Acoustic Features

In Chapter 2 we said that we can represent the F0 contour in terms of its global properties, or as resulting from the linking of a series of intonational events. The results of this study seem to show that, in terms of marking topic structure, global F0 features are the most effective acoustic indicators. The performance of the system was much better when global indicators were included as features than when boundary features were used, whether other features (durational and textual) were included or not. Further, global F0 features were more prominent in both the final CART tree and Maxent classifier. As will be developed more below, this may indicate that global F0 features, specifically F0 declination, are the primary prosodic means used to mark topic boundaries by speakers and are therefore the most reliable indicators of such in an ASR system.

So what do we make of the poor results using F0 boundary features. Does this invalidate the general theory that topic ends are marked with a falling F0 contour and topic continuations are marked with a rising boundary? Not necessarily. It would be reasonable to deduce from the present results that boundary F0 features, and indeed durational features (which also exhibited poor results compared to what might be expected from the literature) are more reliable to distinguish discourse structure at the sentence or phrase level. It is still plausible that boundary tones mark the discourse status of an utterance or phrase in relation to the broader topic structure of the text, but evidently a more fine-grained categorisation than simply whether the F0 contour at the boundary is falling or rising is needed to tease this out.

Results were significantly worse (using a two-tailed t-test) when using pitch accents to try to identify topic words in order to calculate Hearst's text tiling scores than when all words (minus stop words) were included. This could be because the presence or absence of a pitch accent is too crude a measure of information status (given vs new). However (as will be discussed in the next section), when ToBI features (which according to the literature are a more accurate indication of information status) were included instead results were even poorer. Again, we could work on the idea that pitch accents mark the information status of discourse entities at the phrase or sentence level, and this information is then incorporated as a whole into the global topic structure of the text. If this were the case, it would not be effective for ASR to try to use pitch accents directly to identify topic threads and hence topic boundaries. Of course, there are other plausible reasons for this finding. No allowance was made for the use of anaphora, so subsequent non-identical references to the same entity would be missed (although this would affect the traditional Hearst scores in the same way). As was noted in Chapter 2, pitch accents mark the information status of phrases not words, so it may be that repeated references are missed because the wrong word is identified as indicating a new topic by its pitch accent. It may also simply be because the system used to automatically identify pitch accents was not good enough to be effective.

5.2 Advantages and Disadvantages of ToBI Features

As we saw in the previous chapter, results with ToBI features were disappointing. While F0 features on their own performed well, performance with only ToBI features

was below the baseline. Features derived from ToBI annotations did not appear as features at all in the final CART tree, and did not bear much weight in the Maxent classifier. However, when combined with textual features, performance with ToBI features was only marginally worse than performance with a combination of textual and F0 features.

In particular, the use of tunes to imply topic endings and continuations in the literature proved fruitless in this study. (In fact, we used boundary tones to indicate tunes but testing with the whole tune (accent plus boundary tone) on the hand-annotated data did not yield any better results). Using the whole corpus, tune features failed to predict any topic boundary with the CART classifier and achieved an F-score of only 13.1% with the Maxent classifier. Even when testing using the hand annotated data in the smaller part of the corpus, the best result was an F-score of 14.3% using the Maxent classifier. This would seem to show that the posited tunes do not indicate topic continuations and endings as was suggested in Chapter 2. We could conclude that phrasal tunes, along with pitch accents, should not be taken in themselves to indicate the discourse status of phrases at the topic level. This does not, however, invalidate the theme/rheme marking theory we discussed earlier. It works in well with the idea that intonational events help to mark the status of entities within sentences, and then this information is contributed to the higher topic structure.

These results could, however, be symptomatic of a broader problem with trying to apply Steedman's theory of intonational marking of information structure as laid out in Chapter 2. There are two potential problems for ASR applications. Firstly, certain intonational events (most crucially L boundary tones) are not represented by excursions in the F0 contour, they are therefore very hard to recognise automatically. Secondly, some intonational events, in particular the L+H* L-H% theme marker, are commonly not phonetically realised, as is acknowledged within that theory. It is thus very hard to distinguish this from any other word bearing no pitch marker, or any other two words with no boundary between them. A reliable solution to this problem will have to be found if Steedman's theories are to be used in an ASR system at either the phrase level or the topic structure level.

Again, it may be that the theory as it stands is valid but there were two many inaccuracies in the way in which ToBI features were extracted and evaluated in this study. Some of this is reasonably solvable. For instance, better HMMs to find intonation

events could have been used if more training data had been available. Also, it is within the capacity of existing technology for phrases, rather than just words, to be marked as having pitch accents, and these used in computing Hearst's lexical scores, or in trying to identify chains of connected references to the same topic. However, it remains true for the present that there are no reliable means to extract ToBI features automatically from the speech signal. We cannot define these events precisely enough in acoustic terms. We will probably not know if it is worthwhile and valid to represent intonation in terms of such events until these definitions are found. Similarly, it is difficult to determine if such events should be represented in categorial (like ToBI) or continuous (like Tilt) terms. It is easier and more accurate to extract continuous features from a continuous signal. This does not, however, guarantee this approach will be more reliable and it is a very difficult to test the question empirically.

5.3 Marking of Higher Level Structure

What can we conclude on the basis of the results presented here about the way prosodic information should be represented in order to best inform an ASR system that wishes to do topic segmentation? The only acoustic cue which can be manipulated directly to successfully gauge topic structure is the global F0 level of a sentence. All the other acoustic measures taken directly were not rich enough to indicate topic structure on their own, although they may be reliable indicators of lower level structure. Textual features, Hearst's lexical scores and cue phrases, were also reasonably effective in identifying topic boundaries. However, results when combining these two features only led to accuracy of around 87%. This might improve with more training data, but still falls short of the standard of humans, who would probably find this a very easy task. This suggests that humans, and any ASR system which wishes to reach the performance of humans, use prosodic cues in a more sophisticated way.

The proposal could look like this, humans use prosodic cues to help construct different levels of discourse structure, from the phone/word level up to the topic structure level. The lower level entities, along with further F0 cues, work together to inform the construction of higher level discourse structure. Therefore, different types of prosodic cues (as identified by different acoustic measures) may be used to recognise entities on more than one level of discourse, but any one cue will be more effective on some

levels than others.

Of the acoustic measures used in this study, it is likely that duration and local F0 features (peaks and troughs) primarily indicate discourse structure at the intonational phrase or sentence level. In terms of Steedman's theories of the marking of themes and rhemes, this means that intonational events (characterised by local F0 features and durational information) are used firstly to identify phrases and then to mark their informational status. This information, rather than the acoustic cues themselves, can then be fed into a system identifying higher level constituents.

We have not come much further here in deciding whether intonational events are necessary in themselves, or whether these acoustic cues can be used directly at multiple levels of discourse recognition. In fact, it might be more helpful in trying to identify a 'theme' accent or tune to start with acoustic properties rather than be tied to the, possibly misleading, ToBI representation. This route, however, may be problematic when dealing with the types of inter- and intra-speaker variation in acoustic cues discussed in Chapter 2. It still seems sensible to pursue an extra level of intonational event identification which subsumes variation problems, and then have a direct mapping from these events to discourse recognition. Intonation events are also more useful for speech synthesis, as they facilitate a direct mapping from an abstract intonation representation to the realisation of the F0 contour appropriate to the discourse status of the utterance involved.

Assuming that these events exist then, the results of this study give little support the marking of the information status of entities in a discourse suggested by Steedman's theories. As was discussed in the previous section, this is not conclusive, as the poor results could be caused by the process of estimating ToBI features with Tilt ones, or could be the fault of the identification of the ToBI features themselves. It could also be that the relationship between information structure at sentence level and topic structure level is more complex than we had presumed. In any case, this study does show that in order for this type of theoretical work to be of use to the ASR community, further work needs to be done to define intonational events precisely enough that they can be identified automatically by acoustic measures. It is an open question, upon which the results of this study cannot really assist, whether these events should be categorial or continuous. In terms of making a system derived from the speech signal, it would seem more straightforward to use continuous features unless clear evidence can be

found of categorial distinctions in human perception and production. Also, the theory of the relationship between intonational structure and information structure needs to be refined so that it is expressed only in phonologically realised terms.

Above the phrase level, a further level of intonational event annotation might be needed to represent the properties of the intonational phrase as a whole. It follows from this study that mean F0 and the peak F0 level would be good indicators of these meta-events. Since F0 declination over the course of topics, and not just utterances, appears to be such a reliable indicator of topic structure, it does not seem reasonable to continue to postulate it as a paralinguistic phenomenon (as it currently standard, see section 2.3.1). The recognition of such meta-events might also enable us to identify the function of phrases in contexts where general F0 declination is not so common, such as dialogues.

In practice, a system which recognised at least these two levels of discourse structure (entities and intonational phrases) at once would probably be more effective, as information from the two levels interacts. It would not be too difficult to build up a discourse model of the text from this. The added advantage of such a system is that it would be much more straightforward to allow the system to go further and identify the topic or even do summarisation.

5.4 Conclusion

To come back to the questions posed at the beginning of this chapter, the results of this chapter give some support to the notion that prosody gives information for the building of discourse structure at different levels, and it is too simplistic to expect to use all relevant acoustic cues directly to identify discourse structure at the topic level. The one exception to this may be that the global F0 properties of an utterance may give a reliable indication of that utterance's role in topic structure.

Further experimental work is needed to determine if it is more effective to use acoustic cues directly to build each level of discourse structure or whether intonational events should be identified first. If this is the case, a taxonomy of such events needs to be established which both captures human intuitions about intonational categories, and is recognisable automatically. Any theory relating these events to discourse semantics

needs to accord to these criteria as well to be useful for ASR.

Chapter 6

Conclusion

In this study we have built a system which automatically recognises topic boundaries in a corpus of broadcast radio news. We evaluated the effectiveness of various textual, durational, F0, ToBI and Tilt features which were suggested by the literature on topic segmentation to be helpful in carrying out this task. We found using solely durational, Tilt or ToBI features, topic boundaries cannot be reliably identified. However, using solely textual or F0 features, reasonable performance is reached. Using a combination of textual and F0 features, textual and ToBI and textual and Tilt features led to performance significantly above the baseline.

When all features were combined, the only features to be used by the CART classifier were utterance mean F0, cue words, Hearst's continuation lexical score and the energy in the last portion of the sentence. The Maxent classifier relied more strongly on the Tilt value of the first pitch accent in the sentence, as well as cue words, boundary F0 features and Hearst's new lexical score than on other features.

On the basis of these findings and the literature relating to the representation of prosodic information both in ASR and in human production and perception, we made several proposals about how such information should be represented in order to be useful to an ASR system wishing to do recognition of discourse structure. We concluded that each acoustic cue is more effective at determining discourse structure at certain levels of such structure than at others. For instance, the mean and maximum F0 in an intonational phrase is effective in determining topic structure while the degree of final lengthening is not. We suggested that the identification of lower levels of discourse

structure using prosodic cues would help in the identification of higher levels, although the recognition process could take place simultaneously. The lower levels would include the identification of the information status of entities within intonational phrases along the lines of Steedman's information structure theory.

We did not reach a final conclusion on whether acoustic cues can be used directly to do recognition at each level of discourse structure; or whether it is better to first identify intonational events. We also left open how such events should be characterised, if they exist, i.e. as continuous variables (like Tilt) or categorial ones (like ToBI). Further research needs to be done on this issue, although it was suggested that a well-motivated and automatically recognisable taxonomy of intonational events would be better able to deal with inter- and intra-speaker variation and would be more suitable for speech synthesis applications.

This work gives insight into the way prosodic information can be used in ASR systems to identify higher level discourse structure. On the basis of this study we can conclude that the line of research relating ToBI intonational events to discourse semantics is not currently well advanced enough to be of direct use in such a system. More studies like this one are needed to develop these theories so they can be of use to ASR. In particular, the phenomenon of global F0 decline needs to be dealt with within the intonational event theory as it has a discourse semantic value crucial to the identification of topic structure.

Appendix A

Example of Radio News Broadcast

Example of a radio news broadcast from section *f1a* of the Boston University Radio News Broadcast. Sentence boundaries are marked with <S> markers, while topic boundaries are marked with <TOPIC> markers.

```
<S> <BROADCAST> </S>
<TOPIC><S> I'm Irene Doyle . </S></TOPIC>
<TOPIC><S> Massachusetts Senate Ways and Means Committee Chair
 Patricia McGovern says she'll consider further spending cuts to keep
 the state's budget in balance next year . brth </S>
<S> McGovern opened public hearings on the state's fiscal nineteen-
 ninety budget today with a terse warning to those planning to
 testify . </S></TOPIC>
<TOPIC><S> Meanwhile , county sheriffs from across Massachusetts took
 to the state house today to demand more money . brth </S>
<S> The sheriffs say jails in eight counties will run out of funds by
 May first . </S></TOPIC>
<TOPIC><S> Nearly a year of wrangling over the future of Boston school
 superintendent Laval Wilson will likely end tonight as the school
 committee votes whether to renew his contract . brth </S>
<S> Wilson has been trying to negotiate a new pact with the panel for
 weeks . brth </S>
<S> His current term expires in June . </S></TOPIC>
```

<TOPIC><S> Massachusetts Governor Michael Dukakis says his Choose- A-
School proposal will strengthen inner city school systems rather then
luring away their brightest students . brth </S>
<S> The plan would allow parents to send their children to schools in
another district if space is available . brth </S>
<S> Dukakis says the plan is already working in two communities . </S>
<S> The Massachusetts Teachers Association opposes the plan , saying it
will widen the gap between wealthy and poor districts . </S></TOPIC>
<TOPIC><S> Massachusetts Chancellor of Higher Education Franklin
Jennifer is calling for a seven point seven percent tuition increase
at state colleges and universities . brth </S>
<S> The increase would cost students between sixty and one hundred
forty dollars a year . </S></TOPIC>
<TOPIC><S> Governor Dukakis met with environmentalists today , who gathered
at the State House to push for open space legislation . brth </S>
<S> WBUR’s David Barron reports . </S></TOPIC>
<TOPIC><S> The Boston Bruins face-off against the Buffalo Sabers
tonight at the Garden . </S>
<S> A victory tonight would move the B’s onto the second round of the
N.H.L. play-offs . brth </S>
<S> The Red Sox and the Celtics are off tonight . </S></TOPIC>
<S> </BROADCAST> </S>

Bibliography

Arons, B. (1992), Techniques, perception and application of time-compressed speech, *in* 'Proceedings of the American Voice I/O Society', pp. 169–177.

Bailey, T. L. & Elkan, C. (1993), Estimating the accuracy of learned concepts, *in* 'Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence', pp. 895–900.

Baldridge, S. (2001), *OpenNLP Maxent*, University of Edinburgh. *https://sourceforge.net/projects/opennlp*.

Bard, E., Sotillo, C., Anderson, A., Thompson, H. & Taylor, M. (1996), 'The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment', *Speech Communication* **20**, 71–84.

Beckman, M. & Edwards, J. (1992), Intonational categories and the articulatory control of duration, *in* E.-B. Tohkura & Y. Sagisaka, eds, 'Speech Perception, Production and Linguistic Structure', OHM Publishing Co. Ltd., pp. 356–375.

Beckman, M. & Hirschberg, J. (1999), 'The tobi annotation conventions', *http://www.ling.ohio-state.edu/tobiame_tobi/annotation_conventions.html*.

Beckman, M. & Pierrehumbert, J. (1986), 'Intonational structure in English and Japanese', *Phonology Yearbook* **3**, 255–310.

Beeferman, D., Berger, A. & Lafferty, J. (1999), 'Statistical models for text segmentation', *Machine Learning* **34**(1-3), 177–210.

Berger, A. L., Pietra, S. A. D. & Pietra, V. J. D. (1996), 'A maximum entropy approach to natural language processing', *Computational Linguistics* **22**(1), 1–36.

Birch, S. & Clifton, C. (1995), 'Focus, accent and argument structure: Effects on language comprehension', *Language and Speech* **38**, 365–391.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Brown, G., Currie, K. & Kenworthy, J. (1980), *Questions of Intonation*, Croom Helm, London.

Calhoun, S. (2002), On the H*/L+H* distinction. Term Paper for MSc in Speech & Language Processing, University of Edinburgh.

Clarkson, P. & Rosenfeld, R. (1997), Statistical language modeling using the CMU-Cambridge toolkit, *in* 'Proceedings ESCA Eurospeech'.

Cutler, A., Dahan, D. & van Donselaar, W. (1997), 'Prosody in the comprehension of spoken language: A literature review', *Language and Speech* **40**(2), 141–201.

Dharanipragada, S., Franz, M., McCarley, J., Roukos, S. & Ward, T. (1999), Story segmentation and topic detection in the broadcast news domain, *in* 'Proceedings of the DARPA Broadcast News Workshop', Herndon, Viginia.

Donselaar, W. v. (1995), Listeners' use of the 'information-accentuation' interdependence in processing implicit and explicit references, *in* 'Proceedings of the Fourth European Conference on Speech Communication and Technology', Madrid, pp. 979–982.

Donselaar, W. v. & Lentz, J. (1994), 'The function of sentence accents and given/new information in speech processing: Difference strategies for normal-hearing and hearing-impaired listeners?', *Language and Speech* **37**, 375–391.

Eichmann, D., Ruiz, M., Srinivasan, P., Street, N., Culy, C. & Menczer, F. (1999), A cluster-based approach to tracking, detection and segmentation of broadcast news, *in* 'Proceedings of the DARPA Broadcast News Workshop', Herndon, Viginia.

Entropic-Research-Labs (1998), '*xwaves* manual'. version 5.3.1.

Fiscus, J., Doddington, G., Garofolo, J. & Martin, A. (1999), NIST's 1998 topic detection and tracking evaluation (TDT2), *in* 'Proceedings of the DARPA Broadcast News Workshop', Herndon, Virginia.

Fougeron, C. & Keating, P. A. (1997), 'Articulatory strengthening at edges of prosodic domains', *Journal of the Acoustical Society of America* **101**(6), 3728–3740.

Fowler, C. & Housum, J. (1987), 'Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction', *Journal of Memory and Language* **26**, 489–504.

Fry, D. (1958), 'Experiments in the perception of stress', *Language and Speech* **1**, 126–152.

Gernsbacher, M. & Jescheniak, J. (1995), 'Cataphoric devices in spoken discourse', *Cognitive Psychology* **29**, 24–58.

Grieff, W., Morgan, A., Fish, R., Richards, M. & Kundu, A. (2000), Mitre TDT-2000 segmentation system, *in* 'Proceedings of the NIST Topic Detection and Tracking Workshop', Gaithersburg, Maryland.

Grosz, B. & Hirshberg, J. (1992), Some intonational characteristics of discourse structure, *in* 'Proceedings of the International Conference on Spoken Language Processing', Banff, Canada, pp. 429–432.

Grosz, B. & Sidner, C. (1986), 'Attention, intentions, and the structure of discourse', *Computational Linguistics* **12**, 175–204.

Hastie, H. (2002), 'Automatically predicting dialogue structure using prosodic features', *Speech Communication* **36**(1-2), 63–79.

Hawkins, S. & Warren, P. (1994), 'Phonetic influences on the intelligibility of conversational speech', *Journal of Phonetics* **22**, 493–511.

Hearst, M. (1995), 'Test tiling: Source code', *http://elib.cs.berkeley.edu/src/texttiles/common_words.list*.

Hearst, M. (1997), 'Texttiling: Segmenting text into multi-paragraph subtopic passages', *Computational Linguistics* **23**(1), 33–64.

Hedberg, N. & Sosa, J. M. (2001), The prosodic structure of topic and focus in spontaneous English dialogue, *in* 'Topic & Focus: A Workshop on Intonation and Meaning', LSA Summer Institute, Univerity of California, Santa Barbara.

Hirschberg, J. & Nakatani, C. (1996), A prosodic analysis of discourse segments in direction-giving monologues, *in* 'Proceedings of the Twenty-Fourth Annual Meeting of the Association for Computational Linguistics', Santa Cruz, pp. 286–293.

Hirschberg, J. & Nakatani, C. (1998), Acoustic indicators of topic segmentation, *in* 'International Conference on Spoken Language Processing', Sydney, Australia.

Hirschberg, J. & Ward, G. (1991), 'Accent and bound anaphora', *Cognitive Linguistics* **2**(2), 101–121.

Kimball, O., Ostendorf, M. & Bechwati, I. (1992), 'Context modeling with the stochastic segment model', *IEEE Transactions on Signal Processing* pp. 1584–1587.

King, S. (2001), Speech and language processing. Course Notes, University of Edinburgh.

Ladd, D. R. (1996), *Intonational Phonology*, Cambridge University Press, UK.

Ladd, R. D. & Schepman, A. (forthcoming), "sagging transitions' between high pitch accents in English: experimental evidence', *Journal of Phonetics* . to appear in 2002.

Linguistic Data Consortium (2002), *http://www.ldc.upenn.edu*.

Manning, C. D. & Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Mulbregt, P. v., Carp, I., Gillick, L., Lowe, S. & Yamron, J. (1999), Segmentation of automatically transcribed broadcast news text, *in* 'Proceedings of the DARPA Broadcast News Workshop', Herndon, Virginia.

Nakatani, C., Hirschberg, J. & Grosz, B. (1995), Discourse structure in spoken language: Studies on speech corpora, *in* 'Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation', Stanford, CA, pp. 106–112.

Nigam, K., Lafferty, J. & McCallum, A. (1999), Using maximum entropy for text classification, *in* 'IJCAI-99 Workshop on Machine Learning for Information Filtering', pp. 61–67.

Osbourne, M. (2000), Estimation of stochastic attribute-value grammars using an informative sample, *in* 'Proceedings of Coling 2000', Saarbrucken.

Ostendorf, M., Price, P. & Shattuck-Hufnagel, S. (1994), The Boston Unversity radio news corpus, Technical Report ECE-95-001, Boston University.

Pierrehumbert, J. & Hirschberg, J. (1990), The meaning of intonational contours in the interpretation of discourse, *in* P. Cohen, J. Morgan & M. Pollack, eds, 'Intentions in Communication', MIT Press, Cambridge, MA, pp. 271–312.

Provost, F., Fawcett, T. & Kohavi, R. (1998), The case against accuracy estimation for comparing induction algorithms, *in* 'Proceedings of the 15th International Conference on Machine Learning', San Francisco, CA, pp. 445–453.

Ratnaparkhi, A. (1996), A maximum entropy model for part-of-speech tagging, *in* 'Proceedings of Empirical Methods in Natural Language Conference'.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Selkirk, E. (1986), 'On derived domains in sentence phonology', *Phonology Yearbook* **3**, 371–405.

Shattuck-Hufnagel, S. & Turk, A. E. (1996), 'A prosody tutorial for investigators of auditory sentence processing', *Journal of Psycholinguistic Research* **25**(2), 193–247.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. & Tür, G. (2000), 'Prosody-based automatic segmentation of speech into sentences and topics', *Speech Communication* **32**(2), 127–154.

Shriberg, E., Taylor, P., Bates, R., Stolcke, A., Ries, K., Jurafsky, D., Coccaro, N., Martin, R., Meteer, M. & Ess-Dykema, C. (1998), 'Can prosody aid the automatic classification of dialog acts in conversational speech?', *Language and Speech* **41**(3-4), 439–487.

Silipo, R. & Crestani, F. (2000), Prosodic stress and topic detection in spoken sentences, *in* 'Proceedings of the SPIRE Conference', Coruna, Spain, pp. 243–252.

Silverman, K., Beckman, M., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), A standard for labelling english prosody, *in* 'Proceedings of the International Conference on Spoken Language Processing (ICSLP)', Vol. 2, Banff, pp. 867–870.

Sluijter, A. & Terken, J. (1993), 'Beyond sentence prosody: Paragraph intonation in Dutch', *Phonetica* **50**, 180–188.

Spark Jones, K. (1999), Automatic summarizing: Factors and directions, *in* I. Mani & M. T. Maybury, eds, 'Advances in Automatic Text Summarization', MIT Press, Cambridge, MA, pp. 1–14.

Steedman, M. (1991), 'Structure and intonation', *Language* **67**, 260–296.

Steedman, M. (1996), *Surface Structure and Interpretation*, MIT Press, Cambridge,MA.

Steedman, M. (2000), 'Information structure and the syntax-phonology interface', *Linguistic Inquiry* **31**(4), 649–689.

Steedman, M. (2001), *The Syntactic Process*, MIT Press, Cambridge,MA.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G., Rivlin, Z. & Sönmez, K. (1999), Combining words and speech prosody for automatic topic segmentation, *in* 'Proceedings of the DARPA Broadcast News Workshop', Herndon, Virginia.

Swerts, M. (1997), 'Prosodic features at discourse boundaries of different strengths', *Journal of the Acoustical Society of America* **101**, 514–521.

Swerts, M. & Geluykens, R. (1993), 'The prosody of information units in spontaneous monologue', *Phonetica* **50**, 189–196.

Swerts, M. & Geluykens, R. (1994), 'Prosody as a marker of information flow in spoken discourse', *Language and Speech* **37**(1), 189–196.

Taylor, P. (2000), 'Analysis and synthesis of intonation using the tilt model', *Journal of the Acoustical Society of America* **107**, 1697–1714.

Taylor, P., Caley, R., Black, A. W. & King, S. (1999), *Edinburgh Speech Tools Library: System Documentation*, 1.2 edn, University of Edinburgh, Edinburgh.

Terken, J. & Nooteboom, S. (1987), 'Opposite effects of accentuation and deaccentuation on verification latencies for given and new information', *Language and Cognitive Processes* **2**, 145–163.

Tür, G., Hakkani-Tür, D., Stolcke, A. & Shriberg, E. (2001), 'Integrating prosodic and lexical cues for automatic topic segmentation', *Computational Linguistics* **27**(1), 31–57.

Venditti, J. & Swerts, M. (1996), Intonational cues to discourse structure in Japanese, *in* 'Proceedings of the Fourth International Conference on Spoken Language Processing', Philadelphia, pp. 725–728.

Vogel, I., Bunnell, H. T. & Hoskins, S. (1995), The phonology and phonetics of the rhythm rule, *in* B. Connell & A. Arvanti, eds, 'Phonology and Phonetic Evidence: Papers in Laboratory Phonology', Vol. IV, Cambridge University Press, UK, pp. 111–127.

Warren, P. (1999), Prosody and language processing, *in* S. Garrod & M. Pickering, eds, 'Language Processing', Pyschology Press, UK, pp. 155–188.

Wright, H. (2000), Modelling prosodic & dialogue information for ASR. PhD Thesis, University of Edinburgh.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (2000), *The HTK Book.* version 3.0.

Zue, V., Seneff, S. & Glass, J. (1990), 'Speech database development at MIT:TIMIT and beyond', *Speech Communication* **9**, 351–356.