



Recent Improvements to Automatic Speech Recognition Using Prosodic Features

CSC 2528 Term Project

Christopher Collins
Department of Computer Science, University of Toronto

May 6, 2002

Abstract

Automatic speech recognition and understanding (ASR) is a technology of increasing importance as audio data becomes more prevalent and embedded computers become more commonplace. Traditional approaches use acoustic data to create word hypotheses and language models to evaluate the hypotheses. This work examines recent enhancements to ASR using acoustic data related to segments larger than phones (prosody) for several applications. Relevant literature on the use of prosody in detection of errors, disambiguation, word recognition, and boundary detection/topic segmentation are reviewed.

DISCARD THIS PAGE: CITATIONS INCLUDED TO INDUCE *et al.*
on FIRST APPEARANCE IN DOCUMENT

(Hess, Batliner, Kiessling, Kompe, Nöth, Petzold, Reyelt and Strom,
1997) (Stolcke, Shriberg, Hakkani-Tür, Tür, Rivlin and Sönmez, 1999) (Stolcke,
Shriberg, Hakkani-Tür and Tür, 1999) (Stolcke, Shriberg, Bates, Ostendorf,
Hakkani-Tür, Plauche, Tür and Lu, 1998) (Swerts, Litman and Hirschberg,
2000) (Krahmer, Swerts, Theune and Weegels, 1999) (Silverman, Beckman,
Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert and Hirschberg, 1992)

Contents

1	Introduction	1
1.1	Automatic Speech Recognition	1
1.2	Prosody	3
2	Detection of Errors	6
2.1	Identifying Misrecognition and Recognizing Corrections	7
2.2	Identifying Disfluency	11
3	Disambiguation	13
4	Word Recognition	18
5	Topic Segmentation	21
6	Conclusion and Future Directions	24

1 Introduction

1.1 Automatic Speech Recognition

Speech recognition by humans is an extremely complex process, drawing on both acoustic signals and a vast knowledge base. Emulating this behavior with a computer has long been a goal of computational linguistics and engineering research, and one assumed to be a future possibility by everyone from scientist Alan Turing to film-maker Stanley Kubrick. The realization of this goal has proved much more difficult than anticipated. The task is traditionally divided into two subtasks — automatic speech recognition (ASR) and automatic speech understanding (ASU), although these tasks recently have been linked in the use of semantic information to clarify recognition uncertainties. This combined ASR/ASU will be referred to as ASR throughout this work. The ultimate goal is to create a computer which incorporates large-vocabulary, speaker-independent continuous recognition and understanding.

There are many motivations for adding speech as a new modality for human-computer interaction. Speech is the most natural human-human communication means, and is the most efficient. People speak at rates of 120-250wpm but type on average 100-150wpm (Ainsworth, 1988). Speech interaction is hands-free, an important factor for drivers, equipment operators, the physically challenged, and anyone who cannot interact via the traditional

keyboard and mouse. The VERBMOBIL project (Hess et al., 1997) set an exciting goal of speech to speech language translation. Every improvement in ASR brings us closer to this goal for a technology which would improve international communication.

Since the early 1970s, work in Large-Vocabulary Continuous Speech Recognition (LVCSR) has focused on sentence segmentation techniques and the training and use of Hidden Markov Models (Jurafsky and Martin, 2000). Applications of dynamic programming to HMM decoding, such as Viterbi and A* decoding further advanced the usefulness of HMMs for automatic speech recognition. Both decoding algorithms take as input feature vectors drawn from the data, likelihoods (*i.e.*, from Gaussian mixtures or maximum likelihood) and priors (*i.e.*, from n-gram statistical models) and output a string of recognized words.

Kompe (1997) outlines a general modular structure for ASR systems. The components are feature extraction, word recognition, syntactic analysis, semantic/pragmatic analysis, dialogue control, answer generation, and speech synthesis. The last two modules are specific for a speech-in/speech-out interactive system. The dialogue control module is important in any interactive situation involving spontaneous speech. It tracks history and resolves anaphora, deals with spontaneous speech phenomena such as disfluencies, and copes with barge-in (when a user interrupts the system unexpectedly).

Recently, Greenberg (1996) has suggested that instead of a detailed spectral signal, information can be gained through temporal analysis of coarser acoustic signals. Studies of the hearing abilities of the hearing impaired, as well as acoustic signal analysis, show that this type of analysis would be more robust in the face of reverberations and noise. Current research has turned to using additional acoustic features such as these, known as prosodic features, to elicit additional information from speech data. Kompe (1997) claims that all of the modules for a general ASR system can be enhanced using this prosodic knowledge. This work reviews recent ASR research in using prosodic features for detecting errors and problems in recognition, for disambiguating syntax and meaning, and for segmenting a speech sample into sentences and topics.

1.2 Prosody

Prosodic information in a speech signal is of growing importance in speech recognition research. This area was virtually ignored until the late 1980s, and now many of the talks at speech technology conferences relate to prosody. Many definitions of this term have been posited, from very broad to very specific and narrow. Broadly, prosody is accepted as properties of speech related to segments larger than phones (Jurafsky and Martin, 2000). These properties include pitch, pauses, relative duration, intonation, voice quality, and energy, which more generally can be considered to compose ac-

centuation, phrasing, and pauses. Ladd and Cutler (1983) further refine this definition in terms of concrete features, such as acoustic parameters of pitch, duration, and intensity, and abstract features of phonological organization on the suprasegmental level. Fujisaki (1997) states that prosody only exists when a message is produced as a coherent string of sounds, thus defining prosody as the organization of various language units into an utterance which conveys linguistic, paralinguistic, and non-linguistic information. Linguistic information is the string of words or symbols and their relations. Paralinguistic information is (unconventionally) defined as information not inferable by a written text but added by the speaker to convey additional linguistic information, such as speech act categorization as a question or assertion. Non-linguistic information is defined as that prosodic information which characterizes the age, gender, emotional state, etc. of the speaker.

Prosody varies greatly between languages. Prosody can be thought of as the most unique footprint of a language (Hess et al., 1997), thus applications of prosody must be language specific. Most of the work reviewed for this paper was carried out in English, German, and Japanese, but results of considering prosody in speech recognition can be found for Swedish, Italian, Spanish, Mandarin, and Russian, among others.

Prosody has much potential for application in ASR. For example, homographs and homophones (*e.g.*, *permit* vs. *permit*) in English can be disambiguated by pitch accent. In German, these accents are rather fixed

and therefore their observed position can be used to aid selection between candidate words in a word-recognition lattice (Kompe, 1997). Speech parsing for understanding is syntactically more difficult than text understanding because punctuation is not supplied. Prosody can be used to detect phrasal and sentential boundaries, sometimes to a greater extent than punctuation does. Semantic and pragmatic analysis can be aided by use of accentuation (through prosodic phrasing and pitch range reset). For example, information on focus (Beckman, 1997) can disambiguate scope ambiguities (*e.g.*, “[old men] and women are holding a protest.” vs. “[old men and women] are holding a protest.”). Dialogue analysis can be aided by prosodic information in the classification of dialog acts (*e.g.*, “fifteen.” vs. “fifteen?”.)

Despite the great potential, there are also a variety of difficulties that arise when prosody is added to an ASR system (Kompe, 1997). For example, pitch accent can have at least two meanings — as emphasis or indicating a question. Different speakers have been found to realize prosodic events through different means. Prosodic information may be redundant if semantic and syntactic analysis can disambiguate an utterance, leaving the benefit of additional information debatable. One negative result of multiple prosodic realizations of events is that it is difficult (and thus-far impossible) to define a functional mapping between prosodic boundaries and syntactic boundaries.

Many difficulties in prosody research are related to the use of “lab speech” vs. spontaneous speech. Lab speech — speech corpora contain-

ing many repetitions of utterances designed to elicit a desired prosodic contour — can be useful in examining prosodic phenomena if selected carefully. Beckman (1997) and others note that lab speech can be characterized by distributions of accent types (such as the “flat hat” contour seen in F0¹ in American English recitations) which are not representative of spontaneous speech. However, if one recognizes these difficulties and designs the corpus to elicit contrasts as required (*e.g.*, by emphasizing words in a script or inducing desired prosody through providing short dialogues to set up context), lab speech can be quite useful. Spontaneous speech corpora are often plagued by lower quality acoustic data, which may be inappropriate for preliminary research. However, sometimes spontaneous speech is preferred, such as in the study of disfluencies, emotional indicators, repairs, and filled pauses.

The variety of applications for prosody in ASR is tremendous, and this is an open research area. Several applications of prosody to speech recognition have been chosen from this vast collection and will be presented in the following sections.

2 Detection of Errors

Speech recognition by both computers and humans is imperfect. Humans have a great capacity to distinguish when they have been unable to recognize

¹F0 is the first formant frequency, a frequency region of high energy in a spectrogram of an utterance.

what was said and to request clarification, or to accept a correction from a speaker when they misrecognize an utterance. Humans also have a great ability to recognize when a speaker has made a disfluent utterance, a filled pause, or a self-corrective repair. These abilities are essential for a complete ASR system to support successful interaction, and this section will examine prosodic enhancements to ASR which address these issues.

2.1 Identifying Misrecognition and Recognizing Corrections

Speech recognition traditionally is carried out such that paths through a word graph are assigned probabilities (based on acoustic confidence scores) and the most likely is chosen (Jurafsky and Martin, 2000). If none are above a threshold likelihood, then the system responds with a “rejection error” — a statement of failure and request for clarification. In some cases the string which is selected may be incorrect, even given an acceptable confidence level. It becomes crucial then to recognize cues from the input which indicate a user is not responding to a current question, but rather correcting a past system response which was implicitly assumed correct. The following example illustrates the difference between explicit and implicit verification:

1. User: I want to go to Toronto.
2. Explicit:
 - (a) System: Do you want to go to Oahu?

(b) User: No, I want to go to Toronto.

3. Implicit:

(a) System: When do you want to go to Oahu?

(b) User: I want to go to Toronto.

The potential for system confusion in the implicit example is obvious. There is a trade off in system efficiency which prefers implicit verification in some cases due to a reduction in number of turns, but this makes corrective statement detection and interpretation more difficult. In addition, it has been found that corrective statements take a different style of speech, so characterization of these differences is important in order to improve recognition accuracy of corrective utterances.

Krahmer et al. (1999) frame this problem in terms of information grounding. In this theory, communication occurs in two phases: a presentation phase in which the current speaker conveys information to a listener, and an acceptance phase in which the listener confirms or denies that the information was conveyed successfully. In that work, non-prosodic information, such as number of words, new information, etc. are studied for correlation with acceptance and rejection classification of user utterances.

It has been shown (Swerts et al., 2000) that corrections are often made in a “prosodically marked” speaking style. This analysis is supported by results of Levow (1998) for work on the SpeechActs corpus of interactive

desktop computer control. Corrections in both studies tend to be hyperarticulated (slower and louder speech with wider pitch excursions and longer pause durations). This speech style is confirmed by examination of mean and maximum formant frequency (F0) values, mean and maximum energy values, total duration, length of pause preceding turn, speaking rate, and length of silence within the turn. These features are combined into vectors of differences between values for labeled corrective and non-corrective turns. These vectors confirm that there are statistically significant differences in corrective and non-corrective speech. Levow (1998) report a 16% failure rate for utterances following recognized utterances and a 44% failure rate for utterances following failures. Thus hyperarticulation, which is especially effective in human-human communication, can thus contribute to compounding the recognition errors as it differs markedly from training data for an ASR system.

Corrections of misrecognitions are more markedly different than repetitions prompted by system failures (corrections of rejection errors), and therefore are more likely to be misrecognized themselves (Levow, 1998). This suggests that a lower tolerance for acoustic uncertainty, resulting in more rejection errors and fewer misrecognition errors, may help the overall performance and usability of ASR systems by inducing fewer misrecognition chains. In work on the TOOT corpus (Hirschberg, Litman and Swerts, 1999; Swerts et al., 2000) of human-computer telephone dialogues in the domain of train

schedules, the different correction strategies (repetition, paraphrase, omission, and addition) are shown to have different effectiveness. Therefore, systems which guide users to correction strategies with higher recognition rates and smaller deviations in speaking style are expected to have better performance. The most effective strategy is for the system to explicitly confirm all user turns. Recognition rates and user satisfaction are higher, even though the average interaction requires a greater number of turns.

Reports on the TOOT corpus report that prosodic differences between corrective and non-corrective turns increase with number of turns from the original error. Unexpectedly, more distant corrections are better recognized. No explanation for this contradiction is given, which weakens the claim of a significant relationship between deviation in speaking style and correction recognition accuracy.

These studies have focussed on detecting corrective speech and comparing the efficiencies of corrective strategies. The recommendations to design systems to direct users into corrective strategies less likely to induce problematic hyperarticulated speech is an acceptable temporary measure, but future work should pro-actively focus on how to adjust recognition modules to automatically detect and cope with hyperarticulated speech, rather than requiring users to adopt an unnatural method of speaking.

2.2 Identifying Disfluency

Spontaneous speech is often riddled with disfluencies, repairs, and filled pauses. It is important for a speech recognition system to detect these disfluencies and speaker self-corrections and deal with them appropriately, in order to have a better understanding of what is being said. Shriberg, Bates and Stolcke (1997) pioneered work in prosodic detection of disfluencies, using the SwitchBoard corpus of human-human telephone dialogues and decision trees based on vectors of extracted prosodic data. This work was then combined with n-gram language models (Stolcke et al., 1998) following recommendations of (Heeman and Allen, 1997) that boundary detection, speech repairs, and discourse markers must be resolved together. Speech repairs, for example, are often signaled by both prosodic means (pauses) as well as syntactic anomalies, therefore should not be considered in terms of prosody or language modeling independently.

In (Stolcke et al., 1998), relevant prosodic information (such as pauses, F0 contours, and signal-to-noise ratios) were extracted in vectors of values and added to word transcripts, alignments, and hand-labeled disfluency annotations to form a set of training data for the models. This prosodic information is relatively independent of the word-based cues which are used in the associated n-gram language model, thus it was expected their combination would improve accuracy on this task.

The study was interested in both disfluencies and in linguistic events at

word boundaries. Data were divided into six classes representing the events of interest:

- Sentence boundary *e.g.*, He went out * The store was closed
- Filled pause *e.g.*, I uh * hate jazz
- Repetition *e.g.*, I * I hate jazz
- Deletion *e.g.*, I was * I hate jazz
- Repair *e.g.*, He * she likes it
- Else/fluent *e.g.*, she * likes it

Models for prosodic decision trees and event language models with and without segmentation information were trained and tested independently on the task of classifying language events. Each classifier has performance significantly greater than the baseline of 81.8%, obtained by labeling every event as “Else/fluent”. Each model gives a probability of an event given the input (words or prosodic features). These probabilities were later combined in a linear model interpolation, as well as independent model combination, and joint modeling (word-based posterior frequencies added to feature vectors for a prosodic decision tree). Linear interpolation using an empirically optimized combination gives the best result. It should be noted that by including segmentation information (turn-boundaries and pause durations) in the language model, the authors violate the independence assumptions

of their models. They claim any negative effect of the independence assumption violation is outweighed by the 4% relative improvement of event detection using the interpolated model with segmentation information on unseen test data.

Sentence and phrase boundary detection is an important area of prosody research in itself. Further applications of prosody to boundary detections are discussed in section 5.

3 Disambiguation

Applications in natural language understanding often have to deal with disambiguation problems, for example attachment ambiguities, scope ambiguities, and word sense ambiguities. Tasks involving speech data must resolve these disambiguation problems, plus the additional problems of finding word and sentence boundaries, dialogue act classification, and recognizing the words themselves. This section will present investigations in prosodic semantic and syntactic disambiguation indicators. Many of works in this area use corpora prosodically labeled with the ToBI prosody transcription system (Silverman et al., 1992) and its extensions to other languages.

Syntactic ambiguity resolution with prosodic information has had varied success. Hunt (1997) reports 76% accuracy on the task of resolving various syntactic ambiguities in professionally read English speech. Success in cross-linguistic work on prosodic indicators for syntactic and seman-

tic disambiguation carried out for English, Spanish, and Italian (Avesani, Hirschberg and Prieto, 1995; Hirschberg and Avesani, 1997) have been much lower. Hirschberg and Avesani (1997) report many possible relationships for semantic (scope of negation, focus-sensitive operators) disambiguation for English, but no significant correlations were found for syntactic disambiguation (PP, adverbial, and relative clause attachment) in either English or Italian. The differences in the successes of these two studies on syntactic disambiguation may be contributed to differences in the corpora and evaluation methods used for the experiments. In addition, the goals of these works were different — (Hunt, 1997) seeks to use syntactic and prosodic information in several models to automatically train models to evaluate binary syntactic ambiguities, where (Hirschberg and Avesani, 1997) seeks to evaluate several proposed correlations for statistical significance and similarity cross-linguistically. The possibility exists, therefore, that the models of (Hunt, 1997) discover and use different relationships using different features than (Hirschberg and Avesani, 1997).

In (Hunt, 1997), three models are presented for the resolution of syntactic ambiguities with prosodic information. Models are trained on sets of syntactic (hand-parsed link grammar labels) and acoustic (prosodic features of duration and energy, counts of phonemes per syllable, etc.) data. Models are tested on resolving syntactic ambiguities at word boundaries. The exact forms of the ambiguities in question are not given nor are they broken down

into classes as in (Hirschberg and Avesani, 1997). It seems they can include any of a range of link grammar attributes, including attachments and sentence boundaries. The baseline is reported as being random chance (50%). The first model used, break index linear regression, requires hand-labeling of break indices in the training data and achieves 76% accuracy on test corpora. Two other models (canonical correlation analysis and linear discriminant analysis) do not require any prosodic break labeling, yet achieve 74% accuracy, suggesting it may be possible to train syntactic-prosodic models on large corpora of unlabeled data.

The corpora of (Hunt, 1997) (the radio news corpus and the ambiguous sentence corpus) are based on recordings of professional news readers. In the case of the ambiguous news corpus, the news readers were given two interpretations of each of 35 sentences to read. In contrast, the corpora of (Hirschberg and Avesani, 1997) are based on ambiguous sentences which the readers (both in English and Italian) were allowed to interpret themselves, using the context of a surrounding paragraph. The interpretations made by the readers were tested to ensure they were as intended by asking several questions of the readers after they made the recordings. In most cases, the interpretations intended were correctly elicited by the surrounding paragraphs. This method, it could be argued, produces a corpus more similar to spontaneous speech, even if some of the example paragraphs were slightly artificial-sounding. Results of a search for correlations based on

hand-labeling of utterances using the ToBI transcription standard showed significant disambiguation of syntactic ambiguities by neither English nor Italian speakers. The difference here may be due to the use of only the ToBI labels in the search for patterns, whereas (Hunt, 1997) uses 10 acoustic features automatically labeled from the data. In addition, it has been claimed (Hirschberg and Avesani, 1997; Kompe, 1997) that ambiguous sentences which are disambiguated by context will show less significant prosodic markers than those (such as in the radio news and ambiguous sentence corpora) which are not. The reason given is that the speaker recognizes (perhaps subconsciously) the lack of real ambiguity and therefore does not provide additional information to the listeners. If true, this type of uncertainty in the reliability of prosodic markers brings into question their usefulness for syntactic disambiguation in ASR for spontaneous speech.

The results for semantic disambiguation were somewhat more promising. Comparing English and Italian speakers, Hirschberg and Avesani (1997) see that scope of negation and focus-sensitive operators were reliably distinguished by both groups. Scope of negation ambiguities were disambiguated by phrasing differences. English speakers tend to include major or minor prosodic phrase boundaries before a subordinate conjunction in the narrow scope context and not in the wide scope context of sentences such as “William isn’t drinking because he’s unhappy”. Similar phrasing is seen in Italian, with all but one of the narrow scope utterances in the corpus be-

ing uttered in two intermediate phrases, while all wide scope versions are uttered in one phrase.

Focus sensitive ambiguities, such as the focus of *even* and *only* in sentences like “Harold even telegraphed the paper” are also consistently disambiguated by prosodic differences. In both English and Italian, when a sentence is disambiguated, it is through pitch accent. The pitch accent in these cases is associated with the focussed item. The item not in focus is de-accented. In Italian, the speakers had a tendency to also accent the focus-sensitive operator. Some cases from the Italian corpus exhibit accentuation on both focus candidates, but the focussed item was attributed a high degree of prominence through higher F0 values and/or longer vowel durations. In both languages, the focus of the operator represents the nuclear stress of the utterance.

The results reported in (Hirschberg and Avesani, 1997) show trends for speakers to use certain prosodic means to disambiguate semantic ambiguities. None of the indicators are definite, thus any application of these trends should be probabilistic, *i.e.*, a model calculates probabilities of attachment given a set of prosodic feature values.

Syntactic and semantic disambiguation with prosodic information holds much potential for automatic speech understanding systems. However, a review of the literature reveals conflicting results, a lack of consistently accepted evaluation standards, a lack of an accepted method for obtaining and

annotating corpora, and even a lack of a consistent framing of the problem in terms of the ambiguities most likely to be resolved through these means. Therefore, much research needs to be done before these features will be applicable to working ASR systems.

4 **Word Recognition**

Most applications discussed thus far in this work have considered the questions of adding syntactic, semantic, or pragmatic information to a representation of language beyond what is available in raw word texts. Many of these applications use hand-corrected transcripts as the training and test corpora and list testing with automatically-produced corpora as future work, presumably once these transcripts become reliable enough to not severely cloud any results. This then induces the question: can we use the same prosodic information to improve these transcripts? Can language models be constrained with prosodic cues? Several authors have attempted an answer.

Prosody is not currently widely used for word recognition, mostly because prosody is usually considered only relevant on structures above the word level. An approach generally considered promising, then, is to use prosody to model high level structures of an utterance, then evaluate a word sequence hypothesis in terms of the consistency of the word sequence and the prosodically-derived structure. Early work in this area worked on large lattices of possible word patterns, using prosody to evaluate possi-

ble parses, especially where word-boundaries were uncertain (Veilleux and Ostendorf, 1993). In a similar trend to the other areas of prosody-based research, more recent work relies on likelihood models and continuous prosodic feature values rather than labeled phonological features.

In an extension of disfluency detection work of (Stolcke et al., 1998), researchers at SRI have presented preliminary work using the hidden event n-gram language model (Stolcke, Shriberg, Hakkani-Tür and Tür, 1999). Basically, given a model of syntactic structure and its prosodic manifestations as well as a language model relating syntactic structure and word sequences, a word sequence hypothesis and its syntactic structure can be compared to the syntactic structure predicted by prosody for consistency. Formally, the best word hypothesis W^* is given by:

$$W^* = \underset{W}{\operatorname{argmax}} \sum_s P(W, S, F) P(A|W)$$

where W is a word sequence, S is a parse or structure of W , F is the set of prosodic features, and A is the non-prosodic acoustic data used to generate word hypotheses.

The model proposed in this case is an n-gram model evaluated with standard back-off and smoothing techniques. The word boundary events are as in (Stolcke et al., 1998). The n-gram sequence is thus a sequence of word-event pairs, where the “fluent” event is assumed when no event is explicitly given. For example:

- Right [S] I don’t [DEL] uh [FP] I’m not really sure...

Using an n-gram model and making a Markov assumption that prosodic features are dependent only on the n words surrounding the given boundary, the model becomes equivalent to a HMM where states are (word, event) pairs and output is vector of the prosodic features. Transition probabilities come from the n-gram model, and output probabilities from the prosodic model.

The prosodic model in this work was a decision tree trained to give the probabilities of prosodic features given a word and event. Thus the summation over all possible event sequences can be carried out through a forward evaluation through the HMM, and the best word hypothesis can be found using the equation above.

Although (Stolcke, Shriberg, Hakkani-Tür and Tür, 1999) report a low (2%) improvement in overall word error rate, an analysis of their results shows improvements in highly misrecognized words such as “I”, “the”, and “and”. These words are high-frequency and often phonologically reduced, thus good candidates for misrecognition. The prosodic model suppresses false recognitions of these words by correlating them with sentence boundaries and disfluencies. Therefore unless the prosodic information suggests a sentence boundary or disfluency, the probabilities of these words will be reduced. Since this model correlates hidden events (disfluencies and sentence boundaries) with prosody, and hidden events represent only 18% of word boundaries in spontaneous speech, the method is inherently limited. The authors suggest future work could include an expanded set of events, such

as discourse class markers.

5 Topic Segmentation

Methods for sentence boundary detection using prosodic features have already been discussed in section 2.2 as being best calculated in tandem with disfluencies and in section 3 as one of the possible syntactic features detectable by prosodic means. Prosodic information has been used in many works, including (Kompe, 1997), to disambiguate phrase boundaries and thus assist parsing of ambiguous syntactic structures. In this section we will discuss another exciting application of prosody to boundary detection in ASR: topic segmentation.

Topic segmentation is the task of defining boundaries in a stream of text or speech which separates two topically homogeneous blocks. Most work thus far in this area has focussed on lexical information, since most of the corpora were textually based. As topic segmentation is important for summarization and information retrieval applications (such as audio searching and browsing on the internet), its application to speech corpora will be very important in future as the amount of speech data available continues to grow. Prosodic information is relatively uncorrelated to word identity, thus is expected to provide additional segmentation information beyond purely lexical methods.

Stolcke, Shriberg, Hakkani-Tür, Tür, Rivlin and Sönmez (1999) suggest

a topic boundary model based on HMMs, using speech data from the Linguistic Data Consortium (LDC) 1997 Broadcast News Corpus. In this work, the first task is to chop the data into contiguous units which can be assumed to belong to one topic each. These basic units are closely correlated to sentences, although sentences can only be estimated with speech data. Each chopping boundary is then labeled as a “topic boundary” or a “non-topic boundary”. The task then becomes to find the best labeling of boundaries given the word sequences and the prosodic features of duration, energy, and intonation.

A prosodic model is built as in (Stolcke, Shriberg, Hakkani-Tür, Tür, Rivlin and Sönmez, 1999; Stolcke, Shriberg, Hakkani-Tür and Tür, 1999) by extracting a wide range of features thought to be potentially reflective of breaks in the temporal and intonational contour as expected at topic breaks. This use of the Classification and Regression Trees (CART) technique is reminiscent of similar applications to topic segmentation in (Hirschberg and Nakatani, 1998), which used a smaller corpus (the Boston Directions Corpus) with forced alignment to human transcriptions. Features were extracted from a window around a given chopping-boundary. (Previous work extracted features at each word-boundary). A decision tree is then trained using the known (from human labelers) boundary classifications and feature vectors at each position. This tree can then give probabilities for each class of chopping boundary given a set of features. As these features are only loosely

correlated to word identity, the authors expected this model to be robust to recognition errors. The feature set (and decision trees) were later pruned to avoid overfitting. The most salient features were found to be F0 differences across the boundary, pause duration, speaker change, and gender,

A language model was an HMM in which states are representative of topic clusters, the observations are chopped units. Two special states representing “topic boundary” and “non-topic boundary” were incorporated to allow the use of the prosodic information. Between sentences, the model must pass between one of these two states. These transition likelihoods were obtained from the decision tree posterior probabilities of the prosodic model.

This method produced a model which is very intuitive in its combination of lexical and prosodic information, yet not restrictive in the types of relationships it considers — a wide range of possible prosodic features are considered for the initial training of the decision tree. In addition, no significant loss in performance is seen when the model is applied to automatically-obtained transcriptions vs. human-transcribed forced-aligned text. Results show that using pause duration as the chopping criteria is very effective, so potential future work could combine the chopping and HMM steps using pause duration, the best chopping criteria and most important decision tree feature. Stolcke, Shriberg, Hakkani-Tür, Tür, Rivlin and Sönmez (1999) report an accuracy score of about 20% for all variants of this method, using

the Topic Detection and Tracking Phase 2 standard. Although they give no comparison or baseline data, this accuracy value is comparable to the recall of 79% reported for the best model of (Hirschberg and Nakatani, 1998).

6 Conclusion and Future Directions

Many areas of ASR have potential for enhancement using prosodic information. This work has presented a review of error detection, disambiguation, word recognition, and topic segmentation. Other areas of interest include using prosody to cope with barge-in, detect speaker identity, detect speaker emotion, detect dialogue act classification, and more. As this is a fledgling field of research, future work should focus on confirming preliminary results as well as continuing cross-linguistic work to discover if prosodic relations are language dependent or innate in human communication.

References

- Ainsworth, W. A. (1988). *Speech Recognition by Machine*, Vol. 12 of *IEEE Computing Series*, Peter Peregrinus Ltd., London.
- Avesani, C., Hirschberg, J. and Prieto, P. (1995). The intonational disambiguation of potentially ambiguous utterances in English, Italian, and Spanish, *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. 1, Stockholm, pp. 174–177.

- Beckman, M. E. (1997). *A Typology of Spontaneous Speech*, in Sagisaka, Campbell and Higuchi (1997), chapter 2.
- Fujisaki, H. (1997). *Prosody, Models, and Spontaneous Speech*, in Sagisaka et al. (1997), chapter 3.
- Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception, *Proceedings of the ESCA Workshop on the 'Auditory Basis of Speech Perception'*, Keele, U.K., pp. 1–8.
- Heeman, P. A. and Allen, J. F. (1997). Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog, in P. R. Cohen and W. Wahlster (eds), *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Somerset, New Jersey, pp. 254–261.
- Hess, W., Batliner, A., Kiessling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M. and Strom, V. (1997). *Prosodic Modules for Speech Recognition and Understanding in VERBMOBIL*, in Sagisaka et al. (1997), chapter 23.
- Hirschberg, J. and Avesani, C. (1997). The role of prosody in disambiguating potentially ambiguous utterances in English and Italian, *Proceedings of the ESCA Tutorial and Research Workshop on Intonation*, Athens.

- Hirschberg, J., Litman, D. and Swerts, M. (1999). Prosodic cues to recognition errors, *Proceedings of the Automatic Speech Recognition and Understanding Workshop*.
- Hirschberg, J. and Nakatani, C. (1998). Acoustic indicators of topic segmentation, in Mannell and Robert-Ribes (1998), pp. 1255–1258.
- Hunt, A. (1997). *Training Prosody-Syntax Recognition Models without Prosodic Labels*, in Sagisaka et al. (1997), chapter 20.
- Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*, 1 edn, Prentice Hall, New Jersey.
- Kompe, R. (1997). *Prosody in Speech Understanding Systems*, Vol. 1307 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin.
- Krahmer, E., Swerts, M., Theune, M. and Weegels, M. (1999). Problem spotting in human-machine interaction, *Proceedings of Eurospeech*, Budapest, Hungary.
- Ladd, D. R. and Cutler, A. (eds) (1983). *Models and measurements in the study of prosody*, Springer-Verlag, Berlin, chapter 2.
- Levow, G.-A. (1998). Characterizing and recognizing spoken corrections in human-computer dialogue, *Proceedings of COLING/ACL*, Montreal.

- Lu, Y. (ed.) (2000). *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing, China.
*<http://www.icslp.org>
- Mannell, R. H. and Robert-Ribes, J. (eds) (1998). *Proceedings of the Fifth International Conference on Spoken Language Processing*, Australian Speech Science and Technology Association, Incorporated (ASSTA), Sydney, Australia.
- Sagisaka, Y., Campbell, N. and Higuchi, N. (eds) (1997). *Computing Prosody*, Springer-Verlag, New York.
- Shriberg, E., Bates, R. and Stolcke, A. (1997). A prosody-only decision-tree model for disfluency detection, *Proceedings of Eurospeech*, Rhodes, Greece, pp. 2383–2386.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody, *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, Banff, Canada, pp. 867–870.
- Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani-Tür, D., Plauche, M., Tür, G. and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words, *in* Mannell and Robert-Ribes (1998), pp. 2247–2250.

- Stolcke, A., Shriberg, E., Hakkani-Tür, D. and Tür, G. (1999). Modeling the prosody of hidden events for improved word recognition, *Proceedings of Eurospeech*, Budapest, Hungary, pp. 307–310.
- Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G., Rivlin, Z. and Sönmez, K. (1999). Combining words and speech prosody for automatic topic segmentation, in J. Allan (ed.), *Proceedings of the DARPA Broadcast News Workshop*, Virginia, U.S.A.
- *<http://www.nist.gov/speech/publications/darpa99/>
- Swerts, M., Litman, D. and Hirschberg, J. (2000). Corrections in spoken dialogue systems, in Lu (2000).
- *<http://www.icslp.org>
- Veilleux, N. M. and Ostendorf, M. (1993). Prosody/parse scoring and its applications in ATIS, *Proceedings of ARPA HLT Workshop*, Plainsboro, NJ, pp. 335–340.