AN AUTOMATIC PROSODY RECOGNIZER USING A COUPLED MULTI-STREAM ACOUSTIC MODEL AND A SYNTACTIC-PROSODIC LANGUAGE MODEL

S. Ananthakrishnan, Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory
Integrated Media Systems Center
Department of Electrical Engineering (Systems)
Viterbi School of Engineering
University of Southern California
http://sail.usc.edu
ananthak@usc.edu, shri@sipi.usc.edu

ABSTRACT

Automatic detection and labeling of prosodic events in speech has received much attention from speech technologists and linguists ever since the introduction of annotation standards such as ToBI. Since prosody is intricately bound to the semantics of the utterance, recognition of prosodic events is important for spoken language applications such as automatic understanding and translation of speech. Moreover, corpora labeled with prosodic markers are essential for building speech synthesizers that use data-driven approaches to generate natural speech. In this paper, we build a prosody recognition system that detects stress and prosodic boundaries at the word and syllable level in American English using a coupled Hidden Markov Model (CHMM) to model multiple, asynchronous acoustic feature streams and a syntactic-prosodic model that captures the relationship between the syntax of the utterance and its prosodic structure. Experiments show that the recognizer achieves about 75% agreement on stress labeling and 88% agreement on boundary labeling at the syllable level.

1. INTRODUCTION

By prosody, we usually refer to the broad category of supra-segmental information contained in a spoken utterance, including intonation, which is manifested in variations of the observable pitch, stress patterns, timing and pauses. These non-lexical cues carry key linguistic information that supplements the actual text of the utterance, which is obtainable from a conventional automatic speech recognizer (ASR). However, the main difficulty associated with utilizing this information source is that there is no clear relationship between these cues and the "meaning" we seek.

Standards for annotation of prosodic events provided a step in the right direction. Perhaps the most popular such framework is ToBI (Tones and Break Indices) [1]. Corpora annotated with prosodic markers using the ToBI (or similar) convention can be a useful starting point for building language understanding systems and high-quality speech synthesizers. However, it can be expensive and time-consuming to perform this labeling manually; hence, an automatic labeling procedure is desirable. An automatic prosody labeler, built from a small amount of manually labeled data, could be used to annotate large corpora very quickly.

Previous work on prosody labeling [2,3] has usually been centered on extracting static features from syllable-level units and possible phrase boundaries and classifying them using a probability model (GMM) or, more commonly, a decision tree. Some work has been done on incorporating prosodic features within a time-series modeling framework [4], but these models are designed at the phone level, and cannot be used for detecting events at higher linguistic and temporal levels.

We believe that the acoustic correlates of prosodic events consist of multiple streams of information that are correlated but are not always synchronous. For example, a stressed syllable may result in an increase in the local energy of the utterance, along with a lengthening of the vowel nucleus and exaggerated pitch movements. Local energy is a frame-level feature, but there may be only one or two distinct pitch movements within a syllable, and there is only one nucleus for each syllable. We therefore need a modeling framework that permits asynchrony between multiple feature streams, but retains the ability to capture the correlation between them.

There is also a very close relationship between the prosodic and syntactic structure of an utterance. We exploit this fact by incorporating a part-of-speech based syntactic-prosodic language model in our recognizer.

In the next section, we describe our approach to building the prosody recognizer, including the acoustic and language model components. In section 3, we provide details on the corpus used and on training methods. Section 4 describes our experiments and summarizes the results. The final section gives a brief outline of future directions.

2. PROSODY RECOGNIZER DESIGN

The basic structure of our automatic prosody recognizer is very similar to that of a regular ASR system. We define the most likely sequence of prosodic events as

$$P^* = \arg\max_P p(A \mid P) \cdot p(P)$$

where $p(A \mid P)$ is the acoustic model and $p(P)$ is the probability of a given sequence of prosodic events i.e. the language model (A represents a set of acoustic features, and P a candidate sequence of prosodic events). The most likely sequence can be found using a standard Viterbi search. The following subsections describe each of these components in more detail.

2.1. Acoustic Features

Since stress is defined to occur on syllables, we choose to perform feature extraction and prosody recognition at the syllable level. The prosody recognizer uses a number of features derived from the acoustic data. These include:

- *Intensity*: the frame-level energy normalized with respect to the average energy of the utterance. A larger intensity is often indicative of stress.
- *Pitch*: a piecewise linear fit to the pitch contour, the duration of each linear segment, and the distance of the center of each segment from the utterance-average pitch. Stress and prosodic boundaries are often accompanied by larger pitch movements.
- *Duration*: the normalized vowel nucleus duration of the current syllable, and the pause duration after each word-terminal syllable. A stressed syllable generally has a nucleus of longer duration, while a pause after a word-terminal syllable is an indication of a prosodic phrase boundary.

These features are split into three streams of information, since they evolve at different rates.

2.2. Acoustic Models

Our acoustic models need to be able to capture the correlation between the multiple streams of acoustic features discussed in the previous section, and at the same time, need to be able to handle the asynchrony between them. One approach to modeling such feature streams is to use coupled Hidden Markov Models (CHMMs). CHMMs are a special type of multi-stream HMMs that permit a different number of states in each stream. Moreover, transitions from one state to another in any given stream are determined not only by the previous state in that stream, but also in every other stream. CHMMs have been successfully used in recent audio-visual research [5] to jointly model speech and video information, which are correlated but asynchronous.

The acoustic models are built at the syllable level. All syllables are assumed to fall under one of four prosodic categories: stressed (s), unstressed (u), stressed boundary (sb) and unstressed boundary (ub). Only syllables that occur at the end of a word can be assigned to one of the "boundary" categories. In addition, we define *short* and *long* variants for each type; syllables with two or fewer phones are designated *short*, and all others as *long*. We therefore have a total of eight acoustic units, each one represented by a CHMM. The CHMMs representing the *short* syllables are designed with fewer states than those modeling the *long* ones.

2.3. Language Model

There is a very strong correlation between the prosodic events in an utterance and its lexical structure. Syntax, in particular, has been shown to be a very accurate predictor of prosody [6]. A very useful syntactic feature that can be obtained automatically is the part-of-speech. We can then relate the prosodic events labeled in the training corpus to these parts of speech and build a language model (see Sec. 3.2) that supplies the joint probability of a sequence of stress and boundary tone patterns and part-of-speech tags. For example, we can estimate the probability of a stressed proper noun following an unstressed determiner. One of the main features of such a language model is that it can be accurately estimated even from very small amounts of training data, since the vocabulary is quite limited. Such a model can provide very good prior estimates of stress patterns using only lexical information.

3. TRAINING: CORPUS AND METHOD

We used a subset of the Boston University (BU) Radio Speech Corpus [7], which is probably the only widely available corpus with prosodic annotations in the ToBI standard, for training and testing the prosody recognizer. Specifically, we chose a pool of 120 utterances (approximately an hour's worth of speech) by speaker 'f2b' as our training set. Each utterance is annotated with prosodic tags, including stress and boundary labels, part-of-speech tags generated by an automatic parser, pitch marks, and ASR-generated time alignment information at

the word and phone levels. Thus, none of our features rely on manual processing; all of them can be automatically extracted from the speech signal and from ASR output. We obtained the features described in section 2.1 from these annotations. A piecewise linear fit was obtained for the pitch marks using the least-squares criterion. The energy and vowel nucleus duration features were normalized using statistics computed from the training utterances. Since we needed alignments at the syllable rather than at the phone level for feature extraction, we used the NIST syllabification tool to split each utterance into syllables and obtained alignments for these syllables using the phone level alignments provided in the corpus.

The training procedure consisted of estimating the acoustic models (CHMMs) and the syntactic-prosodic language model. Each of these is described below.

3.1. Training the Acoustic Models

In order to be able to estimate parameters for the coupled HMMs using the standard Baum-Welch re-estimation procedure, we transformed the CHMM structure to a regular HMM but with additional state transitions and tied probability density functions using the procedure detailed in [5]. We implemented the acoustic models in HTK [8], since it supports user-defined feature sets, multiple information streams and parameter tying.

The input features were split into three streams as described earlier. The first stream consisted of just the normalized energy. The second stream was composed of three pitch related features: the slope of the linear fit segments, the duration of each linear segment, and the distance of the center of each segment from the utterance-average pitch. The third stream was made up of durational features, including the vowel nucleus duration and, for word-terminal syllables, the duration of the pause, if any, after that word.

For the *short* acoustic units, we chose 3 states for the energy stream, 2 states for the pitch stream, and just 1 state for the duration stream (degenerate case). For the *long* units, we built the CHMM with 5 states for the energy stream, but did not change the number of states for the other streams. We modeled the state conditional probability density functions as Gaussian mixtures, the number of mixtures for the energy, pitch and duration streams being 4, 4 and 5 respectively. The transformed HMMs consisted of 6 states for the *short* units and 10 states for the *long* units. We began by training these units using a flat-start approach, followed by Baum-Welch re-estimation. The probability density functions of certain states within each stream were tied according to the description in [5], and the estimation process was repeated a few more times to obtain the final acoustic models.

3.2. Training the Language Model

The syntactic-prosodic language model was trained from the utterance text tagged with part-of-speech information obtained from an automatic parser. For each word, the part-of-speech was determined along with its stress and boundary labels. A corpus was constructed from these part-of-speech/stress label pairs and was used to train a back-off trigram LM using the SRILM toolkit [9]. The vocabulary size for the LM was 156 (number of unique PoS tags times four stress/boundary categories), while the training corpus contained about 9,500 tokens in all.

4. EXPERIMENTS AND RESULTS

We used 45 utterances from the BU corpus spoken by 'f2b' as our test set. Acoustic features were extracted from automatically generated alignments and annotations after syllabification and were normalized in the same fashion as with the training set. Since, in our work, we assume that the text and part-of-speech tags for the test utterances are available (these can be obtained from ASR and automatic parsers), we were able to significantly reduce the search space for decoding by building a probabilistic FSG (PFSG) using prosodic variants of the test PoS sequence, with transition probabilities obtained from the syntactic-prosodic language model.

For each word in the test utterances, sub-word (syllable-level) PFSGs were built for each prosodic variant. Each sub-word PFSG was integrated at the appropriate location within the word-level PFSG constructed from the LM. Each node in the sub-word PFSG corresponded to a single syllable of the word, and consisted of one of the 8 acoustic units described earlier. The transition probabilities for the sub-word PFSGs were obtained by evaluating the acoustic features against each CHMM. Finally, the best path through this composed PFSG was determined through Viterbi search.

We conducted experiments using different combinations of components of the prosody recognizer and compared the results. We evaluated prosody recognition performance at the word and at the syllable level. Stress and boundary chance levels were estimated at the word and syllable level from the training data. These were used label the test data without the use of any models in order to obtain a baseline. The recognizer was then tested using only the language model to see how well prosody can be predicted from text alone. In this case, there was no measure of labeling performance at the syllable level. We then tested the system with only the acoustic models. Finally, we used both acoustic and language models for recognition. The labeling results obtained with our prosody recognizer for these cases are shown in Table 1.

	Stress Agreement		Stress False Positive		Boundary Agreement		Boundary False Positive	
	Word	Syllable	Word	Syllable	Word	Syllable	Word	Syllable
Chance	51.49	56.07	53.45	33.68	56.46	76.79	31.55	17.40
Syntax Only	79.70	N/A	24.25	N/A	82.10	N/A	12.93	N/A
Acoustics Only	72.03	73.97	28.49	17.38	77.32	86.01	17.81	9.50
Syntax + Acoustics	79.50	74.84	13.21	17.34	80.88	87.98	15.98	8.73

Table 1. Stress and boundary labeling results at the word and syllable level (all percentages)

These figures are encouraging, given that the average inter-transcriber agreement for manual annotators is 80-85% for stress labeling and 85-95% for boundary labeling. Moreover, all the ToBI accent types (H*, L*, L*+H, etc.) were mapped to a single stress label and all boundaries (H-, L-, L-H%, L-L%, etc.), including intermediate phrase boundaries, were mapped to a single boundary category.

It is clear from the results that the use of acoustic and/or language models boosts labeling performance much higher than using just the chance level observed in the training data. The acoustic model exhibits good performance at the syllable level, considering that we used a very low-dimensional acoustic feature vector (only 6 features). The combined model performs better than the acoustic-only model on all tasks, but produces slightly poorer results as compared to the syntax-only model on word-level labeling of stress as well as boundaries. This is because of the way the sub-word PFSGs are constructed; for the stressed variant of a word, we do not know which syllable has the stress, and we provide paths through the PFSG for both stressed and unstressed variants of each syllable. Acoustic confusion may then cause the incorrect path to be chosen by the search algorithm, producing errors at the word level.

5. DISCUSSION AND FUTURE DIRECTIONS

In this paper, we described an automatic prosody recognizer that used coupled HMMs to model asynchronous, multiple acoustic feature streams, augmented by a prosodic language model trained on parts-of-speech. The recognizer performs quite accurately even at the acoustic level, and this is further improved by the language model (except for the case discussed in the previous section). The labeling performance can be further improved by using information from a pronunciation lexicon that defines canonical stress patterns for each word. These can be used to determine which syllable within a word is usually stressed; this information can be used to reduce the search space of the sub-word PFSGs.

In this work, we assigned hard labels (stressed or unstressed, etc.) to words and syllables. Drawing such categorical boundaries is difficult at the best of times. Another idea might be to adopt a fuzzy labeling system, which assigns each word/syllable a strength score for presence of stress or boundary, but this again raises the question of how such scores should be interpreted.

6. REFERENCES

[1] M. Beckman and G. Elam, *Guidelines for ToBI Labeling*. http://www.ling.ohio-state.edu/~tobi/.

[2] C. Wightman and M. Ostendorf, "Automatic Labeling of Prosodic Patterns", *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 469-481, 1994.

[3] K. Chen, M. Hasegawa-Johnson and A. Cohen, "An Automatic Prosody Labeling System Using ANN-Based Syntactic-Prosodic Model and GMM-Based Acoustic-Prosodic Model", *Intl. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, pp. 509-512, 2004.

[4] K. Chen, M. Hasegawa-Johnson and S. Borys, "Prosody Dependent Speech Recognition with Explicit Duration Modeling at Intonational Phrase Boundaries", *Proc. Eurospeech*, 2003.

[5] S. Chu and T. Huang, "Audio-Visual Speech Modeling Using Coupled Hidden Markov Models", *Intl. Conf. on Acoustics, Speech and Signal Proc.*, vol. 2, pp. 2009-2012, 2002.

[6] S. Arnfield, "Prosody and Syntax in Corpus Based Analysis of Spoken English", Ph.D thesis, The Univ. of Leeds School of Computer Studies, 1994.

[7] M. Ostendorf, P. Price and S. Shattuck-Hufnagel, *The Boston University Radio News Corpus*, 1995.

[8] G. Evermann, S. Young, et. al., *HTK: Hidden Markov Modeling Toolkit*. http://htk.eng.cam.ac.uk/.

[9] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit", *Proc. Intl. Conf. Spoken Language Proc.*, Sept. 2002.