

# PROSODY RECOGNITION FROM SPEECH UTTERANCES USING ACOUSTIC AND LINGUISTIC BASED MODELS OF PROSODIC EVENTS

*Alistair Conkie, Giuseppe Riccardi, and Richard C. Rose*  
*AT&T Labs-Research, Shannon Laboratory*

## ABSTRACT

A system for automatic recognition of prosodic events in speech utterances has been developed and applied to recognizing accent tones as defined by the tone and break index (ToBI) prosodic labeling standard. Both the acoustic and syntactic modeling portions of the system are described in the paper. The acoustic modeling portion of the system involves representation of ToBI labeled events using hidden Markov models (HMMs) that are defined over a set of prosodic features. The syntactic modeling component involves the prediction of prosodic events based on a stochastic finite state model defined over input labels obtained from a part-of-speech (POS) tagger. The system was evaluated in terms of its ability to recognize pitch accents in a single speaker read speech corpus when the orthographic transcription of the utterance was assumed to be known. It was shown to improve average labeling accuracy over a baseline text-only prosodic labeling system from 84.8% to 88.3%.

## 1 INTRODUCTION

Issues relating to robust and reliable descriptions of prosodic events in speech utterances have resulted in several prosodic labeling methods including the ToBI standard for English prosody [8]. The work described in this paper is motivated by the desire to develop automatic techniques for recognizing prosodic events for use in text-to-speech (TTS) synthesis, natural language understanding (NLU), and automatic speech recognition (ASR) systems. TTS systems could be more rapidly configured for new speakers and new task domains if an automatic prosody recognizer could replace or improve the efficiency of human labelers. The performance and efficiency of ASR and NLU systems could be improved if labeled prosodic events are directly incorporated in acoustic, language, or understanding models. A stochastic maximum likelihood based approach to recognizing prosodic events using both acoustic and syntactic modeling techniques is presented. It will be demonstrated here that high accuracy recognition of ToBI labeled accent tones can be obtained using these automatic techniques.

Prosody recognition is treated here as a problem of finding the most likely string of accent events,  $\hat{T}$ , given a sequence of acoustic prosodic features,  $A$ , and word sequence (including punctuation),  $V$

$$\hat{T} = \arg \max_T P(T|A, V) \quad (1)$$

$$= \arg \max_T P(A|T, V)P(T, V). \quad (2)$$

The first term in Equation 2 corresponds to the probability of the acoustic data stream given HMM models which are assumed to be dependent on the accent event sequence but independent of the word sequence. The

second term in Equation 2 corresponds to the joint probability of the accent event and word sequences. We learn the parameters of this linguistic component of the model from the multi-tiered ToBI labeled speech corpus.

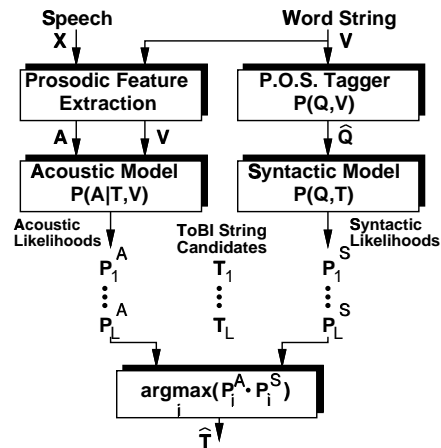


Figure 1: Block diagram of combined acoustic/syntactic prosodic label recognizer

Our combined acoustic/syntactic based system for recognizing prosodic events in speech utterances is illustrated by the block diagram in Figure 1. The acoustic portion of the system has two parts. These include a prosodic feature extraction component which obtains a set of fundamental frequency and energy based acoustic features  $A$  from the sampled speech waveform  $X$ , and an acoustic modeling component which performs maximum likelihood (ML) classification of utterance segments with respect to prosodic event based hidden Markov models (HMMs). Since the acoustic portion of prosodic event classification is performed using hypothesized word and syllable boundaries, it provides a set of candidate prosodic event strings that maximize the acoustic probability  $P(A|T, V)$ .

The syntactic portion of the system also has two parts. These include a part-of-speech (POS) tagger which obtains a sequence of POS labels  $Q$  from the input word sequence and a syntactic model which predicts the prosodic labels  $T$  from  $Q$ . The use of a POS tagger is motivated primarily by the fact that POS tags are known to be good predictors of prosodic events, and also by the fact that it would take an extremely large corpus to robustly estimate the parameters of a direct mapping from the accent event sequence to the word sequence. The output from the syntactic component is a set of candidate prosodic event strings that maximize an approximation to the syntactic probability  $P(T, V) \approx P(V, \hat{Q})P(\hat{Q}, T)$ , where  $\hat{Q} = \arg \max_Q P(V, Q)$  is obtained from the POS tagger. The most likely prosodic event string,  $\hat{T}$ , obtained by the combined system in Figure 1 corresponds to

the candidate string  $T_i$  with the highest combined acoustic/syntactic probability.

There has been previous work involving combined acoustic and syntactic models of prosodic events with application to both NLU and TTS [3, 10]. However, it is very difficult to draw performance comparisons between the different techniques because they were generally applied to different tasks, different corpora, and different standards for prosodic labeling. The performance of the system described in Figure 1 will be compared with that of a text-only prosodic labeling procedure on the speech corpus described in Section 2. The acoustic and syntactic modeling components of the prosody recognition system will be described in Sections 3 and 4 respectively. Section 5 presents the results of the experimental study, and Section 6 presents concluding remarks.

## 2 TOBI LABELED SPEECH CORPUS

### 2.1 The ToBI Labeling Scheme

ToBI is a system for transcribing the intonation patterns and other aspects of the prosody of English utterances [8]. Using this system utterances are labeled with tones and with break indices. The *tones* are symbolic labels for distinctive pitch events. They are transcribed as a sequence of high ( $H$ ) and low ( $L$ ) tones. They can represent either pitch accents or the edges of intonationally marked prosodic units. In this paper we focus on tones and in particular on tones associated with accents. These pitch accents include  $H^*$  and  $L^*$  for simple high and low pitch contours,  $L^* + H$  and  $L + H^*$  for low to high contours. They also include corresponding units  $H+!H^*$ ,  $L^*+!H$ ,  $L+!H^*$ , and  $H^*$  where the “!” symbol indicates a compression of the pitch range.

The *break-indices* mark the prosodic grouping of words in an utterance. A label is inserted at the end of each word to indicate the subjective strength of association with the next word. A scale of 0 (strongest association) to 4 (minimal association) is used. For example, a 4 frequently represents a sentence boundary.

### 2.2 The Corpus

The speech corpus consists of a substantial number of recordings of a single female speaker of American English. The total speech used in this study is approximately 1.5 hours. The recordings are of three broad types and there are approximately equal amounts of each type. All are of read text, and the recordings adhere closely to the text. The text for the first type of recording is a set of sentences that were designed to be diphone-rich. The sentences themselves are meaningful English sentences. The second text set can be classified as newspaper style. The third is of interactive prompt-style utterances.

The raw recordings were annotated in various ways. The first step was to run an automatic alignment program to identify word, syllable and phoneme boundaries. The automatically generated data was then verified manually.

Once this was done, the data was automatically labeled with tones and breaks. This was achieved by running a TTS system with the text version of the utterances. The resulting labels were corrected manually. The automatically produced labels are much less accurate than for the segmental alignment stage and the labels need to be substantially adjusted.

The experimental study described here involves recognition of eight pitch accents which are associated with stressed syllables of prominent words. In addition to the eight pitch accents mentioned previously ( $H^*$ ,  $L^*$ ,  $L^* + H$ ,  $L + H^*$ ,  $L^* + H$ ,  $H+!H^*$ ,  $L^*+!H$ ,  $L+!H^*$ , and  $H^*$ ) an additional symbol, “0”, is used to represent the “no-accent” case.

A total of 818 labeled utterances were used for training the acoustic and syntactic portions of the prosodic label recognition system described in Sections 3 and 4. These utterances contained 11846 words where 6455 of these words were associated with a pitch accent of some kind. Label recognition results were reported on a test set of 90 utterances containing 1166 words where 658 or 56% of these words are associated with a pitch accent.

## 3 ACOUSTIC MODELING FOR PROSODY

This section describes the acoustic feature extraction and acoustic model estimation components of the prosody recognition system. The goal is to train a set of continuous observation density hidden Markov Models (HMMs)  $\Lambda$  to represent the pitch accents in the ToBI-labeled speech corpus described in Section 2. These models underly the acoustic component of the accent recognition system described in Figure 1. The acoustic modeling aspects of the system including feature extraction, estimation of prosodic HMM models, and prosodic event classification are discussed separately.

### 3.1 Prosodic Feature Extraction

There are a large number of acoustic features that are thought to be correlates of prosodic events including fundamental frequency, energy, spectral slope, and segmental duration. In this work, the acoustic features were limited to those that could be derived from fundamental frequency (F0) and energy (E). There are two steps in the prosodic feature extraction process. First, raw fundamental frequency and energy values are estimated from the utterance and smoothed. Second, various empirical normalization procedures are performed to reduce variability among prosodic events.

The need for smoothing of F0 estimates that are to be used for intonational modeling has been raised in previous work in prosodic event recognition [3],[10]. Raw F0 estimates are typically updated at ten ms. intervals and contain discontinuities associated with unvoiced regions in speech. They also include many local perturbations, or micro-prosodic events. In order to deal with these events, a pitch extraction algorithm which produces a smoothed F0 contour was exploited [1].

Both the energy and F0 levels were normalized. Energy was normalized so that the maximum energy level over the entire utterance was constant. F0 was normalized so that the range of F0 values over a breath group was constant. First and second difference parameters were also computed from the smoothed and normalized F0 and E sequences. The prosodic features used in recognition were then F0,  $\Delta F0$ ,  $\Delta\Delta F0$ , E,  $\Delta E$ , and  $\Delta\Delta E$  update over ten ms. intervals.

### 3.2 Training Acoustic Prosodic Event Models

A stochastic HMM approach was investigated in an attempt to deal with the very high degree of variability exhibited by the prosodic features. An attempt was made to manage the effects of acoustic variability by smoothing and normalizing acoustic features, by exploiting knowledge of hypothesized syllable boundaries, and by using context dependent acoustic models to reduce the variability caused by the influence of different surrounding prosodic events. Managing variability in this way resulted in good acoustic classification performance even though the HMM model structure itself does directly reflect the dynamics of the underlying feature space.

Context dependent, continuous Gaussian mixture density hidden Markov models defined over the prosodic features described in the previous section were trained for each prosodic event class. These models have a left-to-right topology with a maximum of eight mixtures per HMM state. Since accent tones are generally associated

with a syllable, a given HMM unit is associated with an accent tone on a syllable in the context of accent tones on the surrounding syllables. The influence of surrounding context is especially apparent, for example, in anecdotal observations of the differences between the acoustic realizations of an unaccented syllable occurring in the context of a preceding high contour,  $H^*$ , as opposed to a low contour,  $L^*$ .

The location of the time instance for the peak associated with the prosodic event was obtained from human labelers according to the ToBI labeling standard [8]. Models are trained from speech segments that are centered on this peak event and with duration normalized to  $M$  frames.

### 3.3 Prosodic Event Classification

It is interesting to look at the performance of the acoustic prosody model in predicting the prosodic label for a given word. First, however, it is necessary to deal with the fact that even though the modeling problem as described by Equation 2 involves predicting a word level prosodic event  $\hat{T}$ , acoustic pitch accents are generally considered to be associated with a syllable. To deal with this, we took advantage of the fact that an accent will generally be associated with the syllable that is assigned primary stress in the word. Furthermore, we assumed that the principal stressed syllable can be obtained from a lexicon. Since it is assumed that the word and syllable boundaries in the utterance are given in advance or are obtained from a probabilistic alignment with the original utterance, we can identify the time segment in the word where the accent event is most likely to occur and estimate the likelihood of the accent event HMM for that interval.

A cost,  $C(m, n)$ , was computed for each segment  $m$  with respect to each prosodic event model  $\lambda_n$ ,

$$C(m, n) = -\log P(\vec{x}_{t_b^m}, \dots, \vec{x}_{t_e^m} | \lambda_n). \quad (3)$$

The time instants  $t_b^m$  and  $t_e^m$  are the beginning and end times for the sequence of prosodic feature vectors  $\vec{x}_t$  associated with segment  $m$ . The most likely prosodic event,  $\hat{T} = \arg \max_T P(A|T, V)$ , is obtained by finding the lowest cost prosodic model,  $\hat{\lambda} = \arg \min_{\lambda_n} C(m, n)$ .

## 4 STOCHASTIC TRANSDUCERS FOR TOBI LABEL PREDICTIONS

### 4.1 Stochastic Sequential Transducers

In this section we describe the stochastic model for accent prediction based on the syntactic features. This model is based on stochastic sequential transducers learned through an automatic algorithm [6, 7]. Finite state transducers  $\tau$  recognize strings from an input language (e.g. set of all possible sequences of part-of-speech tags) and map into strings of an output language (e.g. set of all possible sequences of ToBI labels). Stochastic transducers are powerful tools for representing the association between two information sources and measuring it with a joint probability (e.g.  $P(Q, T)$ ).

In general a stochastic transducer  $\tau$  maps the input sequence  $Q = q_1, \dots, q_M$  into one or more ToBI label sequences  $T_i = t_1^i, \dots, t_M^i$  ( $i = 1, \dots, N$ ). Consequently, the finite state machine is called deterministic ( $N = 1$ ) or non-deterministic ( $N > 1$ ) respectively. A stochastic transducer  $\tau$  assigns to each pair  $(Q, T_i)$  a probability  $P(Q, T_i)$ . The probability  $P(Q, T)$  is then computed as the sum over  $T_i$ :  $P(Q, T) = \sum_i P(Q, T_i)$  There are three issues related to learning stochastic finite state transducers:

- The generation of the state space.

- The next-state transition function (mapping one state into the next one).
- The computation and estimation of the probability  $P(Q, T_i)$ .

Our algorithm automatically defines the state space on the basis of all  $n$ -tuples of pairs  $(q_i, t_i)$  observed in the training set [6]. The next-state transition function depends on the context length  $n$  (order of the model) of  $(q_j, t_j^i)$  pairs used to compute  $P(Q, T_i) = \prod_j P(q_j, t_j^i | q_{j-n+1}, t_{j-n+1}^i, \dots, q_{j-1}, t_{j-1}^i)$ . Moreover the non-determinism in the transition function is implemented in order to account for unseen word pairs in the training set. The probabilities  $P(q_j, t_j^i | q_{j-n+1}, t_{j-n+1}^i, \dots, q_{j-1}, t_{j-1}^i)$  are estimated with discounted Maximum Likelihood techniques so that a non-zero probability is guaranteed overall the space of POS-ToBI pairs [6].

### 4.2 Application to ToBI label prediction

In this study we have considered part-of-speech tags as the only linguistic information for accent prediction. We have a total of 58 part-of-speech tags and combination thereof<sup>1</sup>. We have used this information to train the joint probabilities  $P(q_j, t_j^i | q_{j-n+1}, t_{j-n+1}^i, \dots, q_{j-1}, t_{j-1}^i)$ . An input word sequence is tagged with a stochastic trigram part-of-speech tagger whose accuracy, as measured on the Wall Street Journal database is 95% ([2]). We tested the performance of the POS tagger on our database and found slight performance degradation. The part-of-speech tagging operation in Figure 1 is performed both on the training and test set of word sequences. The POS tag stream is aligned with the ToBI sequences and the pair sequences  $\dots, q_j, t_j^i, \dots$  are generated as input to the self-organizing algorithm in [6].

We map any part-of-speech tag into a non-accented or accented prosodic class. We select the best hypothesis  $\hat{T}_i$  by maximizing the probability  $P(Q, T_i)$  with the Viterbi algorithm:

$$\hat{T}_i = \arg \max_{T_i} P(Q, T_i) \quad (4)$$

In Table 1 we report on the correct classification rate for the accented and non-accented label class for different order  $n$  of the transduction. The decrease in performance as we increase the model order for the accented class is partly attributable to a severe data sparseness problem. As a result, the first order model gives reliable prediction as supported by the previous work in [4]. However, there is a 20% classification error rate improvement for the non-accented class for the 4-th order statistical model. Furthermore, a higher order predictor will be more effective in conjunction with the acoustic model predictor.

Order	Accent(%)	No-Accent(%)
1	89.8	79.5
2	86.5	80.9
3	83.9	81.5
4	86.5	81.7

**Table 1: Correct classification rates for different order stochastic transducers.**

<sup>1</sup>For example *can't* is first tokenized and then tagged as *MD+RB*, where *MD* and *RB* are tags for modal auxiliary and adverbs.

## 5 EXPERIMENTAL RESULTS

This section describes the evaluation of the combined acoustic/syntactic prosody recognition system in Figure 1 on the evaluation speech corpus described in Section 2. The input to the system is a speech utterance,  $X$ , and an orthographic word transcription,  $V$ , that is assumed to be known. The implementation of the integrated system is based on finite state machine based representations which were briefly discussed in Section 4.

The problem of combining the outputs of the acoustic and syntactic components is facilitated by representing the set of all possible ToBI string candidates for a given utterance as a finite state machine (FSM) [5]. String lattices are a type of FSM that are used here to represent all possible sequences of ToBI accent labels that could be produced for a given utterance. Each arc in the acoustic lattice,  $\lambda_A$ , is associated with a symbol representing one of the accent tones and an acoustic cost as given by Equation 3. Each arc in the syntactic lattice,  $\lambda_S$ , is also associated with a symbol representing one of the accent tones and syntactic cost as given by  $-\log P(q_j, t_j | q_{j-2}, t_{j-2}, q_{j-1}, t_{j-1})$  from Section 4. The intersection, or Hadamard product, of these two lattices

$$\lambda_{AS} = \lambda_A \circ \alpha \lambda_S, \quad (5)$$

where  $\alpha$  is an empirical weighting, results in a network containing all sequences of ToBI strings that are common to  $\lambda_A$  and  $\lambda_S$  [5]. The path cost for a given string of ToBI labels in  $\lambda_{AS}$  is equal to the weighted sum of the path costs for that string in  $\lambda_A$  and  $\lambda_S$ . The optimum ToBI string candidate shown as  $\hat{T}$  in Figure 1 corresponds to the lowest cost path through lattice  $\lambda_{AS}$ .

The results obtained using several different systems for the above prosody recognition scenario are given in Table 2. The entries in the table represent percent correct recognition for a two-class, accented word versus unaccented word, accent recognition problem. The accent tones correspond to those tones described in Section 2. The performance for the unaccented and accented words are displayed separately in the Table and were obtained from 90 test utterances containing 1166 total words with 508 of those words being unaccented. The first row of Table 2 displays the performance of the baseline text-only, classification and regression tree based system developed by Hirschberg [4], and adapted to this task by Syrdal [9]. The average accuracy of this baseline system was 84.8%.

The second and third rows of Table 2 display the performance of the acoustic and syntactic portions of the system described in this paper. The independent acoustic and syntactic systems obtain 82.8% and 84.0% accuracy respectively, with both systems obtaining far greater accuracy in classifying accented words compared to unaccented words. The last row of the table displays the performance of the combined acoustic/syntactic system. It is important to note that the average performance of the combined system represents a 23% reduction in average error rate over the baseline system and a 44.7% reduction for the accented words in particular.

System	Accent(%)	No-Accent(%)	Average
Baseline	86.6	82.5	84.8
AC	85.5	79.5	82.8
SYN	86.5	80.9	84.0
AC + SYN	92.6	83.0	88.3

**Table 2: Percent correct two class accent recognition performance. Acoustic (AC) and syntactic (SYN) performance are shown separately.**

## 6 CONCLUSIONS

A prosodic event recognition system has been developed consisting of both a stochastic HMM based acoustic modeling component and a stochastic sequential transducer based syntactic modeling component. The system was applied to recognizing ToBI accent tones in utterances taken from a single speaker read speech corpus. All acoustic and syntactic model parameters in the system were trained using approximately 800 labeled utterances consisting of less than 12000 words. The accent tone recognition performance for a two class accent/no-accent task represents a 23% error rate reduction over the performance of a text-only baseline system.

The larger goal of this work is to integrate the automatic prosody labeling system into the framework of TTS, ASR, and NLU systems. The eventual impact of incorporating prosodic events on the overall system performance for these applications is dependent on many issues. These include the underlying representation of prosodic events as well as the performance of an automatic system in recognizing these events. However, we consider the performance that was demonstrated in this work begins to approach that which might be usefully applied toward these larger domains.

## 7 ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Ann Syrdal for her many helpful suggestions and advice. They would also like to thank Srinivas Bangalore for supplying the POS tagger used in this work.

## REFERENCES

- [1] P. C. Bagshaw, S. M. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. *Proc. European Conf. on Speech Communications*, pages 1003-1006, September 1993.
- [2] S. Bangalore. Unpublished work. 1998.
- [3] W. Hess, A. Batliner, A. Kiessling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt, and V. Strom. Prosodic modules for speech recognition and understanding in Vermobil. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*. Springer, 1997.
- [4] J. Hirschberg. Pitch accent in context: predicting intonational context from text. *Artificial Intelligence*, 63:305-340, 1993.
- [5] F. Pereira, M. Riley, and R. Sproat. Weighted rational transductions and their application to human language processing. *Proc. ARPA Workshop on Human Language Technology*, 1994.
- [6] G. Riccardi, R. Pieraccini, and E. Bocchieri. Stochastic automata for language modeling. *Computer Speech and Language*, pages 265-293, December 1996.
- [7] R. C. Rose, H. Yao, G. Riccardi, and J. Wright. Integration of utterance verification with statistical language. *Proc. ICASSP*, May 1998.
- [8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling English prosody. *ICSLP*, 2:867-870, 1992.
- [9] A. Syrdal. Personal communication. 1998.
- [10] H. Wright and P. Taylor. Modeling intonational structure using hidden Markov models. *Proc. ESCA Workshop on Intonation*, September 1997.