Published in final edited form as:
Proc IEEE Int Conf Acoust Speech Signal Process. 2008 ; 2008: 4545–4548. doi:10.1109/ICASSP.2008.4518667.

FINE-GRAINED PITCH ACCENT AND BOUNDARY TONE LABELING WITH PARAMETRIC F0 FEATURES

Sankaranarayanan Ananthakrishnan and **Shrikanth Narayanan**
Speech Analysis and Interpretation Laboratory, Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089

Abstract

Motivated by linguistic theories of prosodic categoricity, symbolic representations of prosody have recently attracted the attention of speech technologists. Categorical representations such as ToBI not only bear linguistic relevance, but also have the advantage that they can be easily modeled and integrated within applications. Since manual labeling of these categories is time-consuming and expensive, there has been significant interest in automatic prosody labeling. This paper presents a fine-grained ToBI-style prosody labeling system that makes use of features derived from RFC and TILT parameterization of F0 together with a *n*-gram prosodic language model for 4-way pitch accent labeling and 2-way boundary tone labeling. For this task, our system achieves pitch accent labeling accuracy of 56.4% and boundary tone labeling accuracy of 67.7% on the Boston University Radio News Corpus.

Index Terms

prosody; pitch accent; boundary tone; ToBI; RFC; TILT

1. INTRODUCTION

Over the past couple of decades, linguistic theories of prosodic categoricity have assumed a position of some importance. The basic premise of these theories is that prosodic events such as pitch accents and boundary tones are intrinsically discrete in nature, and can be described by a language-dependent symbolic alphabet. One of the most popular standards for categorical annotation of prosodic events is ToBI (Tones and Break Indices) [1], which was developed in the early 1990s. A typical ToBI annotation of an utterance consists of four inter-related tiers:

1. the *orthographic tier*, which provides a plain-text transcription of the utterance.
2. the *tone tier*, which provides a symbolic transcription of prosodic events, mainly pitch accents and boundary tones.
3. the *break index tier*, which indicates the degree of separation (on a 0–4 scale) between successive words in the utterance.
4. the *miscellaneous tier*, which is used for comments, or to annotate non-linguistic phenomena such as disfluencies, laughter, etc.

The most important components of a ToBI-style annotation are the tone tier and the break index tier. The tone tier marks various categories of pitch accents, the most common among them being H*, !H*, L*, and L+H*. These represent *high*, *downstepped*, *low* and *rising peak* accents,

ananthak@usc.edu, shri@sipi.usc.edu.

respectively. Boundary tones are categorized as L-L%, L-H%, H-L% and H-H%, representing different combinations of rising and falling tones, of which the first two are the most common. Boundary tones usually correspond to higher break index values (3 or 4). On the other hand, a break index value of 0 indicates no separation between the words (cliticization).

The basic difficulty in large scale adoption of categorical prosody models in spoken language systems is the expense associated with producing annotated corpora. Manual annotation of ToBI-like labels is time-intensive and laborious. Hence, automatic labeling of prosodic categories is of significant interest to those working in this area. However, the majority of previous work on prosody labeling [2,3,4] has ignored fine prosodic categories while focusing on binary detection (presence vs. absence) of prosodic events such as pitch accents and boundary tones.

While knowledge of the presence or absence of prosodic events is quite useful for many applications, some systems can benefit from a more detailed description of these events. For instance, text-to-speech (TTS) systems can use these labels to generate human-like prosody, while dialog systems may find them useful for identifying different types of speech acts such as questions, declarative statements, and exclamations. Fine prosodic categorization is a difficult proposition even for human labelers. A study by Syrdal et al. [5] shows that pairwise inter-annotator agreement for the pitch accent categories of interest to us is of the order of 60%. Agreement levels for boundary tones is of the order of 75% for the majority tones L-L% and L-H%. To our knowledge, the only previous work on automatic fine-grained ToBI labeling is that of Ross et al. [6], who performed 3-way pitch accent identification (H*, !H* and L*) and 3-way boundary classification (L-L%, H-L% and L-H%) using a decision tree classifier. They obtained pitch accent classification accuracy of 72.4% (vs. 71.8% chance) and boundary tone classification accuracy of 66.9% (vs. 61.1% chance). However, this study was quite limited as it used data from only one speaker.

In this paper, we present a fine-grained ToBI labeler for 4-way classification of pitch accent (H*, !H*, L* and L+H*) and 2-way classification of boundary tones (L-L% and L-H%) using features derived from the rise-fall-connection (RFC) and associated TILT parameterization [7] of the F0 contour. We also evaluate the usefulness of a *n*-gram prosodic language model for these tasks. The remainder of this paper is organized as follows. Section 2 describes the data corpus we use for our experiments. Section 3 gives a description of our acoustic-prosodic features and classifier. Section 4 introduces the prosodic language model and presents the scheme for combining acoustic and lexical models for prosody labeling. Section 5 gives details of the experimental setup and presents a summary of the results. Section 6 concludes this paper with a brief discussion of our contributions and suggests future directions for research in this area.

2. DATA CORPUS

We used the Boston University Radio News Corpus (BU-RNC) [8] for our prosody labeling experiments. This corpus consists of about 3 hours of read broadcast news speech from 6 speakers (3 male, 3 female) with ToBI-style pitch accent and boundary tone annotations. The size of the usable corpus for pitch accent labeling was approximately 28,300 words and for boundary labeling, about 29,800 words. Based on analysis of the distribution of various pitch accent and boundary tone labels in the corpus (Table 1), we limited our pitch accent categories to 4 types, namely !H*, H*, L+H*, L* and 2 boundary tone categories, namely L-H% and L-L%. The remaining categories constituted an insignificant fraction of the corpus as compared to the above labels and were hence discarded. We note that, overall, approximately 14,343 (50.7%) of the words carried any of the above types of pitch accent, while about 5,615 (18.8%) of the words carried any type of the listed boundary tones. For both pitch accent and boundary

classification tasks, we created 10 training and testing sets for cross-validation by randomly splitting the dataset with approximately 80% of the data in the training partitions.

3. ACOUSTIC-PROSODIC MODEL

The acoustic-prosodic model uses raw acoustic correlates of prosody to classify pitch accents and boundary tones. Since the target labels are established chiefly on the basis of the shape of the F0 contour in the vicinity of the event, we only use features derived from the F0 contour for this task.

3.1. F0 Parameterization

Key to the task of fine prosodic categorization is a method for parameterization of the F0 contour that preserves its shape. While we used F0 ranges, differences and averages in the past [4] for establishing presence vs. absence of prosodic events, these features do not provide the discriminatory power to enable identification of different types of pitch accents and boundary tones. Although curve-fitting algorithms are often used to parameterize F0 contours, one popular and simple parameterization is provided by rise-fall-connection (RFC) analysis [7]. This model is particularly well-known in the speech synthesis community and is used for generating F0 contours from a small number of parameters.

RFC analysis treats each prosodic event (e.g. pitch accent or boundary tone) as being comprised of two parts - a rise component followed by a fall component. Each component is described by two parameters - an amplitude (*rise_amp*, *fall_amp*) and a duration (*rise_dur*, *fall_dur*). In addition, the RFC model records the peak value of F0 for the event (*f0_height*) as well as the position of the event (*position*) within the utterance for a total of six parameters that describe the shape of the local contour. Figure 1 illustrates these parameters for a sample prosodic event. In this paper, we assume that the locations of the prosodic events (but not their fine categories) are already known either by manual annotation or by automated techniques described in previous work. Hence, we discard the *f0_height* and *position* parameters and retain only the rise-fall amplitudes and durations.

The TILT model is closely related to the RFC model and describes local F0 contours using three parameters: *amplitude*, *duration* and *tilt* - which are derived from RFC parameters using simple algebraic operations as described in [7]. In our experiments, we compared the effectiveness of TILT parameters versus RFC parameters for fine prosodic categorization. All acoustic features were derived in a speaker-independent fashion.

3.2. Acoustic-prosodic classifier

The acoustic-prosodic classifier is based on the maximum *aposteriori* (MAP) principle as shown in Eq. 1.

$$\mathbf{P}^{*}=\arg\max_{\mathbf{P}} p(\mathbf{P}|\mathbf{A_p}) \tag{1}$$

where **P** stands for the prosody labels of interest and $\mathbf{A_p}$ represents the acoustic-prosodic features. This classifier was implemented as a multi-layer perceptron (MLP) that maps the acoustic-prosodic features derived from RFC and TILT analysis to the target labels. The pitch accent classifier was trained with 8 hidden nodes and 4 output nodes (one for each type of pitch accent). The boundary tone classifier was trained with 8 hidden nodes and 2 output nodes. We used softmax activation for the output nodes because it allowed us to interpret the MLP outputs as posterior probabilities of the corresponding classes. This was useful for integration with the

prosodic language model. The network weights were trained using the scaled conjugate gradient algorithm.

4. PROSODIC LANGUAGE MODEL

Previous work has demonstrated the usefulness of lexical and morphological (part of speech) models for automatic detection of prosodic events in speech [4], where they were shown to outperform a classifier based purely on acoustic-prosodic features.

On the other hand, the shape of the local F0 contour is the primary indicator of fine prosodic categories. In order to determine whether there is a relationship between lexical or morphological items and the various types of pitch accents and boundary tones, we built a model $p(\mathbf{P}|\mathbf{L})$ that attempts to predict prosody labels conditioned on these features. This was implemented as a factored *n*-gram model with trigram context.

We use the MAP framework in Eq. 2 to combine acoustic-prosodic and lexical evidence for classification.

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} p(\mathbf{P}|\mathbf{A_p}, \mathbf{L}) \tag{2}$$

In order to make the model more tractable, we invoke Bayes' rule and make the simplifying assumption that the lexical features are conditionally independent of the acoustic-prosodic features given the prosody labels. The classification equation is then given by Eq 3.

$$\begin{aligned} \mathbf{P}^* &= \arg\max_{\mathbf{P}} p(\mathbf{A_p}, \mathbf{L}|\mathbf{P}) p(\mathbf{P}) \\ &\approx \arg\max_{\mathbf{P}} p(\mathbf{P}|\mathbf{A_p}) p(\mathbf{L}|\mathbf{P}) \end{aligned} \tag{3}$$

This enables us to separate the joint term in Eq. 3 into a product of the acoustic-prosodic model and a prosodic language model $p(\mathbf{L}|\mathbf{P})$. This model was also implemented as a factored *n*-gram with trigram context. Posterior probabilities for the acoustic-prosodic model were obtained from the outputs of the neural network classifier.

5. EXPERIMENTAL RESULTS

We split the BU-RNC data into 10 random training and test partitions as described in Section 2. Experiments were performed on all 10 sets and the results reported are averages across the 10 test partitions. All performance improvment figures quoted in this section are statistically significant at the $p \leq 0.002$ level.

RFC analysis of the smoothed and interpolated pitch contours was carried out using the Edinburgh Speech Toolkit [9], and the corresponding TILT parameters were computed. Acoustic-prosodic models were trained using these two sets of features. We used the Stanford University maximum-entropy tagger, which uses the Penn Treebank tag set, to automatically predict POS tags from the orthography. Two variants of the prosodic language models were built - one using words and the other using POS tags - in order to determine which representation was more useful for prosody labeling.

Table 2 summarizes classification results obtained under various configurations for prosody labeling. In the case of 4-way pitch accent labeling, the chance level baseline accuracy of 54.0%

was obtained by assigning all pitch accents the most frequent label, H*. The TILT features did not perform significantly better than chance. However, using the RFC features resulted in a performance improvement of 2.4% over the baseline. The prosodic language models performed significantly worse than chance, indicating that short-range lexical and morphological context is not useful for predicting fine pitch accent categories. Due to the poor performance of the language models, the performance of the integrated classifiers was also below chance level. Table 3 shows the class-confusion matrix for the best performing classifier (acoustic-prosodic classifier with RFC features). Rows indicate the true labels, while columns give the predicted labels. We note that the minority class L* is hardly ever detected, while the majority of L+H* pitch accents are confused with the dominant H* pitch accent, a phenomenon also reported in [5].

For 2-way boundary tone classification, the chance level of 60.2% was obtained by setting all boundary tones to the most frequently occurring one, L-L%. Acoustic-prosodic classification based on TILT features resulted in a 4.6% improvement in accuracy, while RFC features performed even better, with 7.4% improvement. The prosodic language model based on words beats the baseline by 6.1%, although the model based on POS tags performs only marginally better than the baseline. This indicates that short-range lexical context can help predict boundary tone categories. The POS-based LM is not effective due to two factors: a) errors introduced by the automatic tagger and b) lower granularity of POS tags vis-a-vis words. The best performing system was the combination of RFC features and the word-based prosodic language model, which beat the baseline by 7.5%. The confusion matrix for the best performing classifier (integrated classifier with RFC features and word-based prosodic LM) reveals that the majority of boundary tone classes were correctly identified.

6. DISCUSSION AND FUTURE WORK

In this paper, we described a system that uses a simple, low-dimensional parameterization of the F0 contour based on RFC and TILT analysis to identify pitch accent and boundary tone categories in a speaker-independent fashion based on a neural network classifier. We also tested the performance of short-range prosodic language models using both words and automatically generated POS tags for labeling.

For both pitch accent and boundary tone labeling, we found the RFC features to be more useful for classification than the transformed TILT parameters. This is due to the fact that the transformation process results in a lower dimensional feature set, which does not retain the complete information contained in the RFC parameter set. While the language models did not improve pitch accent labeling performance, the word-based prosodic language model significantly improved boundary tone labeling accuracy. When combined with the acoustic-prosodic model, the integrated classifier gave the best results for boundary tone classification (7.5% improvement over chance level). Our results are not directly comparable to those of Ross et al. [6] due to differences in test conditions, including size of the corpus, chance levels and number of speakers. In the future, we plan to explore models based on long range lexical and syntactic features (derived from a syntactic parse of the orthography) for fine-grained prosody labeling.

7. REFERENCES

1. Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J. ToBI: A standard scheme for labeling prosody. Proceedings of the International Conference on Spoken Language Processing; 1992. p. 867-869.
2. Wightman C, Ostendorf M. Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 1994;2(4):469–481.

3. Chen, K.; Hasegawa-Johnson, M.; Cohen, A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. International Conference on Acoustics, Speech and Signal Processing; 2004. p. 509-512.

4. Ananthakrishnan, S.; Narayanan, S. Automatic prosody labeling using acoustic, lexical and syntactic evidence. IEEE Transactions on Speech and Audio Processing; 2007.

5. Syrdal, A.; McGory, J. Inter-transcriber reliability of ToBI prosodic labeling. Proceedings of the International Conference on Spoken Language Processing; Beijing, China. 2000. p. 235-238.

6. Ross K, Ostendorf M. Prediction of abstract prosodic labels for speech synthesis. Computer Speech and Language 1996;10:155–185.

7. Taylor, P. The TILT intonation model. Proceedings of the International Conference on Spoken Language Processing; 1998. p. 1383-1386.

8. Ostendorf M, Price P, Shattuck-Hufnagel S. The Boston University radio news corpus. 1995

9. King, S.; Clark, R.; Black, A.; Richmond, K.; Strom, V. The Edinburgh Speech Tools Library. The University of Edinburgh; http://www.cstr.ed.ac.uk/projects/speechtools

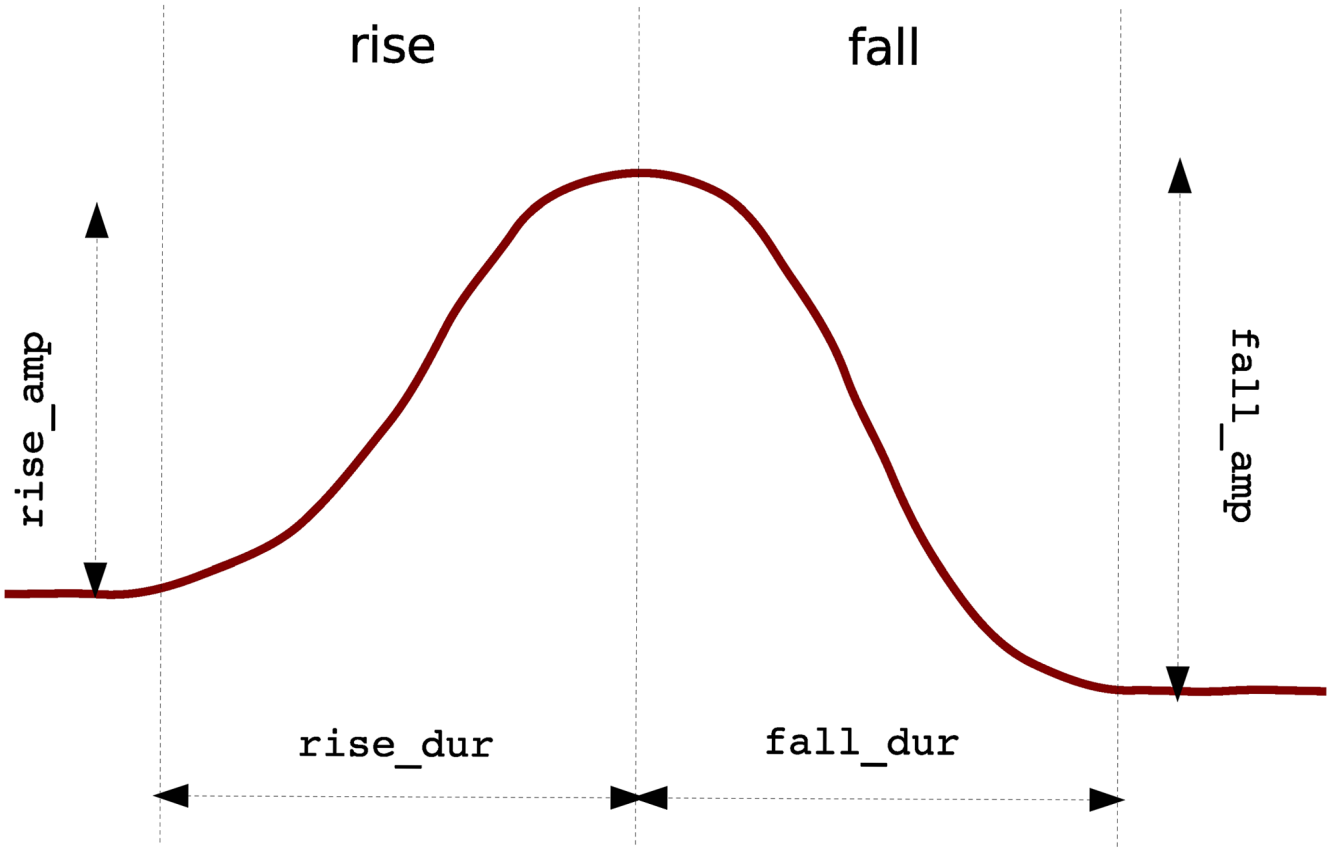

Fig. 1.
Illustration of RFC parameters

Table 1

Prosodic event distribution

Label	Train	Test
!H*	2,863	714
H*	6,203	1543
L*	464	114
L+H*	1,957	485
L-H%	1,782	437
L-L%	2,734	662

Table 2

Prosody labeling accuracy

Method	Accent	Boundary
Chance (baseline)	54.0%	60.2%
Acoustic (TILT)	54.1%	64.8%
Acoustic (RFC)	**56.4%**	67.6%
LM (words)	50.5%	66.3%
LM (POS)	50.5%	60.5%
RFC + word LM	50.6%	**67.7%**
RFC + POS LM	49.4%	67.0%

Table 3

Confusion matrices

	!H*	**H***	**L***	**L+H***
!H*	208	505	0	1
H*	136	1402	1	4
L*	37	76	0	1
L+H*	21	462	0	2

	L-H%	**L-L%**
L-H%	229	209
L-L%	146	515