AuToBI – A Tool for Automatic ToBI annotation

Andrew Rosenberg

Department of Computer Science, Queens College / CUNY, USA
andrew@cs.qc.cuny.edu

Abstract

This paper describes the AuToBI tool for automatic generation of hypothesized ToBI labels. While research on automatic prosodic annotation has been conducted for many years, AuToBI represents the first publicly available tool to automatically detect and classify the breaks and tones that make up the ToBI annotation standard. This paper describes the feature extraction routines as well as the classifiers used to detect and classify the prosodic events of the ToBI standard. Additionally, we report performance evaluating AuToBI models trained on the Boston Directions Corpus on the Columbia Games Corpus. By evaluating on distinct speakers domains and recording conditions, this evaluation represents an accurate representation of the performance of the system when applied to novel spoken material.

Index Terms: prosody, automatic prosody annotation, tools

1. Introduction

The ToBI prosodic annotation standard [1] was developed to phonologically describe the intonation of Standard American English (SAE). Almost as early as its introduction, researchers have been experimenting with ways to automatically detect ToBI labels from the speech signal [2, 3]. ToBI has distinguished itself as a useful system for describing the intonational content of English speech by enabling researchers to identify correlates between ToBI tone sequences and other communicative phenomena including focus [4], topicality [5], contrast [6], discourse acts [7], information status [8], turn-taking behavior [9], and charisma [10].

The ToBI Standard describes SAE intonation in terms of **break indices** describing the degree of disjuncture between consecutive words, and **tones** which are associated with phrase boundaries and pitch accents. Pitch accented words are prominent from the surrounding utterance. Five types of pitch accents – pitch movements that correspond to perceived prominence of an associated word – are defined in the standard: H*, L*, L+H*, L*+H, H+!H*. In addition to these five, high tones (H) can be produced in a compressed pitch range indicated as !H.

Two levels of prosodic phrasing are defined: the intermediate phrase and the intonational phrase. The presence of a prosodic phrase boundary is indicated by perceived disjuncture between two words. Intonational phrases boundaries are defined by the highest degree of disjuncture, and are often associated with silence. Each intonational phrase is comprised of one or more intermediate phrases. The level of disjuncture between words is indicated on the BREAKS tier. Each word boundary has an associated “break index”, which can take a value from 0 to 4, indicating increased disjuncture. Break indices of ‘4’ indicate intonational phrase boundaries, while ‘3’ indices indicate intermediate phrase boundaries. Typical word boundaries have a break index of ‘1’. Each intermediate phrase has an associated phrase accent, describing the pitch movement between the ultimate pitch accent and the phrase boundary. Phrase accents can have High (H-), downstepped High (!H-) or low (L-) tones. Intonational phrase boundaries have an additional boundary tone, to describe a final pitch movement. These can be high (H%) or low (L%). Since each intonational phrase boundary also terminates an intermediate phrase, intonational phrase boundaries have associated phrase accents *and* boundary tones. Each intermediate phrase must contain at least one pitch accent.

AuToBI is a system to automatically hypothesize the presence and type of prosodic events that are present in a spoken utterance. Automatic generation of ToBI labels consists of six tasks: 1) detection of pitch accents, 2) classification of pitch accent types, 3) detection of intonational phrase boundaries, 4) detection of intermediate phrase boundaries, 5) classification of intonational phrase ending tones, and 6) classification of intermediate phrase ending tones. In the current version, the system requires an input segmentation of the signal into words. Initially, accents and phrase boundaries are detected. Then the type of pitch accent is hypothesized, and the phrase ending tones are classified. Each component detection or classification module was trained using the weka machine learning toolkit. AuToBI performs the requisite feature extraction from the speech signal and generates predictions for each element of the ToBI standard using these stored models. The AuToBI feature extraction and prosodic event detection and classification models are freely distributed for non-commercial use under the GNU GPL. As of the publication date, the most recent version of AuToBI can be downloaded from

`http://eniac.cs.qc.cuny.edu/andrew/autobi/`

The rest of this paper is structured as follows. In Section 2, we describe related prior work on detecting and classifying ToBI annotations. We describe the architecture of the AuToBI system in Section 3, and the component detection and classification modules in Section 4. In Section 5, we describe the performance of the system on the Columbia Games Corpus. We conclude and describe future work in Section 6.

2. Related Work

There are two user studies that measure the annotator reliability of ToBI labelers [11, 12]. These studies determine the upper bound of the performance of automatic systems for these tasks. Pitch accents are detected with 9̃1% accuracy and classified at 6̃1%. Intonational and intermediate phrases are detected with 9̃3% and 50% respectively. Intonational phrase internal Phrase accents show 40% agreement, while intonational phrase final phrase accent/boundary tones pairs have agreement of 85%.

There has been significant work in the detection and classification of prosodic events. Wightman and Ostendorf [13] used decision trees and HMMs to detect and classify prosodic event sequences. Veilleux and Ostendorf in examining the interaction between prosodic phrasing and syntactic parsing explored the automatic detection of phrase boundary detection and classification [14]. Sun [15] achieved 92% accuracy in the detection of pitch accents on the speech of a single speaker using Boosted decision trees. Ananthakrishnan et al. [16] explored the use of Coupled HMMs for pitch accent detection and classification. Levow [17] demonstrated the importance of contextual information in the detection of pitch accents. Sluijter et al. investigated the role of spectral balance in acoustic prominence [18]. These represent a small fraction of the research on the acoustic correlates and automatic detection and classification of prosodic events, but represent some of the more influential research that has been incorporated into the described system.

3. System architecture

AuToBI requires three inputs: 1) a wave file containing the speech signal, 2) a TextGrid file containing word segmentation and 3) previously trained classifiers for prosodic event detection and classification tasks. AuToBI operates by first extracting pitch, intensity and spectral information from the speech signal. These acoustic contours are then aligned to the word-defined regions. For each prosodic event detection and classification task, the features required by the corresponding classifier are generated for each word from the aligned acoustic contours.

The extraction of pitch (f0) is performed by a Java implementation of the Praat [19] "Sound to Pitch (ac)..." function. Similarly, intensity extraction is performed by a Java implementation of Praat's "Sound to Intensity..." function. In every task, both raw and speaker normalized pitch and intensity information is used by the classifier. Speaker normalization is performed using z-score normalization, where a data point x is transformed to $z_i = \frac{x_i - \mu_i}{\sigma_i}$ where μ is the mean value in the training data from speaker i and σ_i is the standard deviation of the extracted feature. If available speaker normalization parameters can be stored externally, and loaded at runtime. This allows normalization information calculated over a large amount of spoken material to be reused when running AuToBI on a single utterance. AuToBI includes a utility to construct and store speaker normalization parameters from a batch of wav files. If stored parameters are unavailable, the mean and standard deviation of pitch and intensity are calculated over the current utterance file.

As described in Section 4, the different classification tasks require the construction of distinct feature sets. The AuToBI feature extraction system is structured in such a way that only feature extraction routines that will be used for a classification task will be constructed. At initialization, the system registers feature extractor classes, and the names of the features that will be extracted by running the associated extraction routine. Each classification task has an associated feature set. These feature sets describe the feature names that are required by the associated classifier. When AuToBI extracts features for a given classification task, feature extractor registered for each required feature is executed. This modular structure allows new feature extraction routines to be easily added to the system without fear that they will add overhead to the existing operation of the system: if a feature set does not require the new feature, the new feature extraction routine will never be run.

We run each of the six classifiers – pitch accent detection and classification, intonational and intermediate phrase detection, phrase accent classification and boundary tone/phrase accent classification – on every word. This introduces an inefficiency to the system. Pitch accent types are hypothesized for words where no pitch accent was detected. Similarly phrase ending tones are hypothesized for phrase internal tokens. When generating the final hypothesized output, only hypothesized tones with coincidental accents or phrase boundaries persist. While this introduces some inefficiency, it allows the detection and classification routines to execute independently, with their hypotheses resolved before generating output.

The detection and classification of prosodic events is accomplished using classifiers trained using the weka machine learning toolkit [20]. Weka's implementation in java allows a tight integration between feature extraction and classification components of the AuToBI system. A schematic of the AuToBI system can be found in Figure 1.

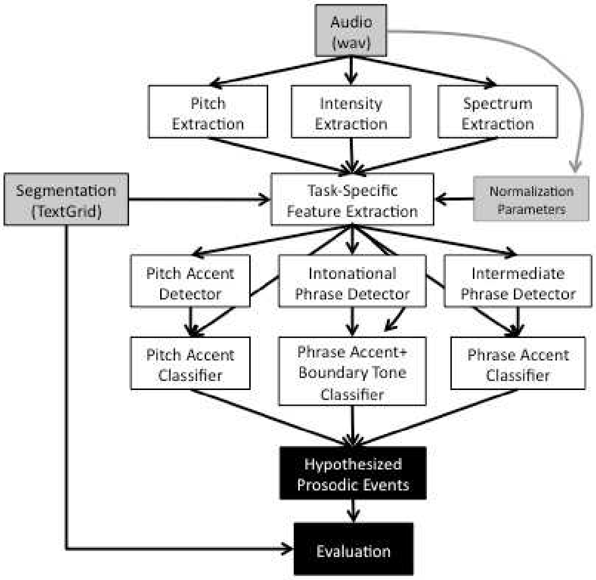

Figure 1: *AuToBI Schematic. (Grey boxes represent user input. Black boxes represent system output. White boxes are internal modules.)*

4. Automatic Prosodic Event Detectors and Classifiers

In this section, we describe the individual detection and classification modules that make up the AuToBI system. The experiments that led us to use the described feature sets and classifiers are described in [21].

The current version of AuToBI includes two sets of classifiers one trained on the read subcorpus of the Boston Directions Corpus (BDC) [22] and the other trained on the spontaneous subcorpus. The BDC consists of spontaneous and read speech from four native speakers of SAE, three males and one female. Each speaker performed a series of nine direction giving tasks. This elicited spontaneous speech was subsequently transcribed manually, and speech errors were removed. Subjects later returned to the lab and read transcripts of their spontaneous monologues. The corpus was then ToBI [1] labeled and annotated for discourse structure. The read subcorpus contains approximately 50 minutes of speech and 10818 words. The spontanous subcorpus contains approximately 60 minutes of speech over 11627 words.

4.1. Pitch Accent Detection

The presence of pitch accents in SAE are typically characterized by excursions in pitch, intensity [23], and duration as well as increased high-frequency emphasis characterized as spectral tilt [24, 25]. In order to best detect these excursions acoustic information needs to be extracted relative to its surrounding context [17, 26]. To capture these qualities, we extract mean, minimum, maximum, standard deviation and z-score of the maximum of raw and speaker normalized pitch and intensity contours and their slopes. Moreover, we use z-score normalization to identify the normalized mean and maximum value relative to eight word-based context windows. These include zero, one or two previous words and zero, one or two following words. We also extract these features over a contour of two spectral contours: 1) the energy contained in the frequency region between 2 and 20 bark[1] and 2) the ratio of the energy in this frequency region to the total energy in the frame. We do not speaker normalize spectral tilt values. This spectral region was identified as the most robust predictor of pitch accent in [27]. Pitch accent detection is performed using Logistic Regression classifiers. Using a similar feature set, this classifier was able to detect pitch accents with 82.90%±0.509 accuracy when evaluated on BDC-spontaneous [21] .

4.2. Pitch Accent Classification

While the presence of a pitch accent is recognized by the acoustic prominence of a word or syllable relative to its surrounding context, the tones associated with the pitch accent – the **type** of pitch accent – is determined by the shape and timing of the pitch contour during the excursion itself. Thus, we extract acoustic information only from the loudest syllable region for the classification of pitch accent types. We identify syllable regions within words using an implementation of an acoustic pseudo-syllabification technique by Villing et al. [28]. We select the pseudo-syllable which contains the maximum intensity in the word as the representative syllable for classifying pitch accent type. We capture the shape of the contour within this region by extracting the minimum, maximum, mean, standard deviation and z-score of the maximum of the raw and speaker normalized pitch and intensity contours as well as the contour slopes. We also include the pseudosyllable duration in the feature set. We explored more heavily engineered features to capture pitch contour shape during a pitch accent, however, we did not find them to improve classification performance [21]. We classify pitch accents using a confidence weighted combination of ensemble sampled [29] SVMs This approach yields a Combined Error Rate of 0.284 on BDC-spontaneous [21] .

4.3. Phrase Detection

Phrases are determined by the amount of disjuncture between two words. This disjuncture is associated with the presence of silence, pre-boundary lengthening and acoustic reset. We extract representations of silence both as a binary variable and the length of silence before the next word. As in the pitch accent detection feature set, the feature vector includes minimum, maximum, mean, standard deviation and z-score of the maximum of raw and speaker normalized pitch and intensity contours and their slopes extracted from the word preceding a candidate boundary. In addition to these, for each feature we calculate the difference between the feature value on a given word and the following word. These features are able to capture the acoustic reset across the a candidate boundary. Pre-boundary lengthening is represented by including the duration of the word preceding each candidate boundary. The same feature set is used in detection of intonational and intermediate phrase boundaries. Intonational phrase detection is performed using AdaBoost with one split decision trees. Intermediate phrase boundary detection is performed using Logistic Regression. On BDC-spon material, the accuracy and f-measure of these two classifiers is 93.13%±0.798 (F_1=0.810±0.022) for intonational phrase boundaries, and 91.65%±0.459 (F_1=0.541±0.021) for intermediate phrase boundaries [21].

4.4. Phrase Ending Classification

ToBI describes the phrase ending intonation using phrase accents tone at the end of intermediate phrases and a boundary tone at intonational phrase boundaries. Since every intonational phrase boundary is also an intermediate phrase boundary, intonational phrase boundaries are associated with both a phrase accent and boundary tone. Since it is difficult to disentangle the influences of phrase accents and boundary tones, AuToBI classifies these simultaneously at intonational phrase boundaries. This leads to an inventory of pairs: L-L%, L-H%, H-L%, !H-L%, H-H%.

As phrase ending tones are realized immediately prior to phrase boundaries, acoustic features are extracted from the final 200ms of phrase final words. We extract the following features to represent the acoustic behavior in this region: minimum, maximum, mean, standard deviation and z-score of the maximum of the raw and speaker normalized pitch and intensity contours and their slopes. These feature sets with support vector machines with linear kernels are able to classify intonational phrase final tones with 54.95%±2.44 accuracy, and intermediate phrase ending tones with 68.6%±1.66 accuracy [21].

5. Evaluation on Columbia Games Corpus

In this section, we describe results evaluating the performance of AuToBI. These evaluation experiments were carried out on the Columbia Games Corpus (CGC) which is described in Section 5.1. The results of these experiments are presented in Section 5.2. Since the CGC contains spontaneous speech, we used the BDC-spontaneous models for the evaluation.

5.1. Columbia Games Corpus

The Columbia Games Corpus (CGC) [9] is a collection of 12 spontaneous task-oriented dydactic conversations between native speakers of Standard American English (SAE). In each session, two subjects played a set of computer games requiring verbal communication to goals of identifying or moving images on a screen. Critically, neither subject could see the other participant, or the other player's screen. The sessions included two types of games, three instances of the CARDS game and one instance of the OBJECTS game. In the first phase of the CARDS game, one subject describes the image of a card, while the other subject searches through a deck for a card that matches the described card. In the second phase, the cards each had three images on them. Moreover the cards available to match with the more complicated cards were limited. Points were scored by the amount of matching objects between the target and selected card. This added complexity was meant to encourage discussion. In the OBJECTS game, both players were presented with a mostly white screen containing a tableau of iconic images. On

[1] 20 bark is above the Nyquist rate of the training and evaluation data.

one player's screen, one of the objects was blinking. The other player's task was to move the object on their screen to the location in which it appeared on the describing player's screen. In both games, both subjects took the roles of describing the card or tableau of icons an equal number of times. Further details of the two games are available in [9].

All of the OBJECTS games and five sessions of the CARDS games have been annotated with the ToBI standard by trained annotators. This comprises approximately 320 minutes of annotated dialog comprising 49,972 words. While each session was annotated by a single labeler, training sessions included all labelers, and difficult, borderline and ambiguous cases were discussed by the group. 51.2% of all words in the CGC are accent bearing. Intonational phrases are approximately 3.57 words long, while intermediate phrases are approximately 2.74 words long. It is notable that the accent rate of the BDC-spontaneous subcorpus is similar to the CGC at 49.5%, though the phrase lengths in the dialog speech of the CGC are shorter than the intonational (5.32) and intermediate phrases (3.73) of the spontaneous monologue speech.

5.2. AuToBI Evaluation

We generate hypothesized prosodic annotation on the manually annotated portion of the CGC and evaluate the performance on each of the component tasks.

Pitch accents are detected with 73.5% accuracy. Pitch accent types are classified with xxx% accuracy and a Combined Error Rate of xxx%. It is worth noting that the pitch accent detection accuracy is significantly below the best previously published results on this task. However, this evaluation scenario differs significantly from previous results. All previous results train and evaluate prosodic event detection and classification systems only within the same corpus. When evaluated in a comparable fashion, the techniques used by AuToBI generate state-of-the-art results [21]. Our intention in this evaluation is to establish a reasonable expectation for users when applying AuToBI to new data sets. Intonational phrase boundaries are detected with 90.8% accuracy, representing an F_1 of 0.812. As found in previous studies the precision of this detection task, 0.941, is higher than the recall, 0.715. This is due to the quality that it is uncommon for silence to occur when there is no intonational phrase boundary. This leads to high precision classifiers using a single feature. Detection of phrase boundaries which are not indicated by silence is substantially more difficult. Intermediate phrase boundaries that do not also intonational phrase boundaries are difficult to detect, AuToBI detects these phrase boundaries on the CGC with 86.33% accuracy and a corresponding F_1 of 0.181 (p:0.536, r:0.109). Intonational phrase final tones are classified with 35.34% accuracy, while intermediate phrase ending phrase accents are classified with 62.21% accuracy.

6. Conclusion and Future Work

In this paper we describe AuToBI, a tool to perform automatic ToBI annotation. AuToBI is distributed as an open-source java project. The system includes modular feature extraction routines and prosodic event detection and classification models trained on BDC-read and BDC-spontaneous material. The system includes six classification tasks: 1) pitch accent detection, 2) pitch accent classification, 3) intonational, and 4) intermediate phrase detection, 5&6) classification of phrase ending tones at both levels of phrasing. We also evaluate AuToBI on the Columbia Games Corpus. We find some substantial effects of genre on the performance, but we believe these results represent an accurate representation of the expected system performance on unseen data.

AuToBI is a new system leaving open many avenues of future work to improve its performance and usefulness. The first augmentation we will make is the distribution of models trained on the Boston University Radio News Corpus, and the Columbia Games Corpus. Currently AuToBI requires word segmentation to be delivered by the user as input. However, a pseudosyllabification module is already included in the package. We will extend the capabilities of the system to operate on hypothesized syllable regions when no word segmentation is given by the user. AuToBI has been designed so that each feature extraction and classification task can run independently of one another, though in this version AuToBI runs on a single thread. Support for multi-threading processors will significantly improve the runtime of AuToBI. Finally, we will continue to investigate and distribute new techniques and features to improve the performance on each task.

7. Acknowledgements

The authors would like to thank Julia Hirschberg and Agustín Gravano.

8. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[2] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," *Comput. Linguist.*, vol. 20, no. 1, pp. 27–54, 1994.

[3] C. Wightman, N. Veilleux, and M. Ostendorf, "Use of prosody in syntactic disambiguation: An analysis-by-synthesis approach," in *HLT*, 1991, pp. 384–189.

[4] J. Gundel, "On different kinds of focus," in *Focus: Linguistic, Cognitive and Computational Perspectives*. Cambridge University Press, 1999.

[5] N. Hedberg, "The prosody of contrastive topic and focus in spoken english," in *Workshop on information structure in context*, 2003.

[6] S. Prevost, "A semantics of contrast and information structure for specifying intonation in spoken language generation," Ph.D. dissertation, University of Pennsylvania, 1995.

[7] A. W. Black, "Predicting the intonation of discourse segments from examples in dialogue speech," in *ESCA/Aalborg University*, 1995, pp. 197–200.

[8] M. Grice and M. Savino, "Can pitch accent type convey information status in yes-no questions," in *Concept to Speech Generation Systems*, 1997.

[9] A. Gravano, "Turn taking and affirmative cue words in task-oriented dialog," Ph.D. dissertation, Columbia University, 2009.

[10] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, pp. 640–655, 2009.

[11] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the tobi framework." in *ICSLP*, 1994.

[12] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic tobi prediction and alignment to speed manual labeling of prosody," *Speech Communication*, vol. 33, no. 1–2, pp. 135–151, January 2001.

[13] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processig*, vol. 2, no. 4, October 1994.

[14] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring based on prosodic phrasing," in *HLT '91: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1992, pp. 429–434.

[15] X. Sun, "Pitch accent predicting using ensemble machine learning," in *ICSLP*, 2002.

[16] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *ICASSP*, 2005.

[17] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Interspeech*, 2005.

[18] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 503–513, 1997.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9-10, pp. 341–345, 2001.

[20] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," in *ICONIP/ANZIIS/ANNES International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192–196.

[21] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[22] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[23] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *Journal of the Acoustic Society of America*, vol. 118, no. 2, pp. 1038–1054, August 2005.

[24] M. Heldner, E. Stragert, and T. Deschamps, "Focus detection using overall intensity and high frequency emphasis," in *ICPhS*, 1999.

[25] A. Rosenberg and J. Hirschberg, "Detecting pitch accent using pitch-corrected energy-based predictors," in *Interspeech*, 2007.

[26] ——, "Detecting pitch accents at the word, syllable and vowel level," in *HLT-NAACL*, 2009.

[27] ——, "On the correlation between energy and pitch accent in read english speech," in *Interspeech*, 2006.

[28] R. Villing, J. Timoney, T. Ward, and J. Costello, "Automatic blind syllable segmentation for continuous speech," in *ISSC*, vol. 2004. IEEE, 2004, pp. 41–46. [Online]. Available: http://link.aip.org/link/abstract/IEECPS/v2004/iCP506/p41/s1

[29] R. Yan, Y. Liu, R. Jin, and A. Hauptmann, "On predicting rare cases with svm ensembles in scene classification," in *ICASSP*, 2003.