Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness

Andrew Rosenberg

Department of Computer Science, Queens College (CUNY), USA

andrew@cs.qc.cuny.edu

Abstract

In this paper, we explore the differences between direct and symbolic sequential modeling of prosody. We use sequential models to characterize speech in two tasks, classifying speaking-style and distinguishing native from non-native speech. We explore the use of a *spike-and-slab* model to directly model pitch contour data. We find in both of these tasks that sequences of symbolic prosodic events to lead to improved performance over approaches that model pitch contours directly. We also explore the use of hypothesized prosodic events in both tasks. We find the speaking-style results to be robust to automatic annotation, while, when classifying nativeness, the spike-and-slab model leads to better performance.

Index Terms: Prosodic Analysis, ToBI labeling, Speaking-Style, Genre, Nativeness Classification

1. Introduction

Symbolic representations of prosody, such as ToBI [1], aim to represent only the communicatively relevant components of prosodic variation. This approach to prosodic representation allows for prosody to be analyzed and discussed without representing the acoustics of the speech. Symbolic representations of prosody are independent of a particular speaker's pitch range, or the speaking-rate of a specific utterance. Symbolic representations of prosody allow for theories about the communicative impact of prosodic variation to be hypothesized and tested while remaining constant in the face of various specific acoustic realizations due to speaker idiosyncrasies, lexical choice or other intervening factors. The acoustic realization is not, of course, independent from its symbolic representation. Successful symbolic representations ignore irrelevant acoustic material while retaining communicative and perceptually salient information.

From a modeling perspective, the use of symbolic representations can be viewed as a dimensionality reduction decision, distilling the raw acoustics into a relatively small set of symbols. An alternative, commonly used modeling approach has been dubbed *direct modeling*. Under this approach, prosody is represented "directly" to a statistical modeling approach by acoustic measurements – pitch, intensity, duration. Using symbolic representations serves as an intermediate step between the acoustics and the modeling approach. Here the acoustic information is first represented by the symbolic representation of prosody, then the symbolic representation is modeled. From a statistical modeling perspective, the use of symbolic modeling hypothesizes that the target being modeled, t, is conditionally independent of the acoustics, a, given the symbolic representation of prosody, p, $t \perp\!\!\!\perp a|p$.

Language, generally, and prosody, specifically, are inherently sequential information streams. In this paper, we explore sequential models based on symbolic representations of prosody – symbolic models – and direct sequential models. We evaluate the performance of these models on two distinct tasks: classification of speaking style, or genre, and distinguishing native from non-native speech. The ToBI standard is used as the representation in all of the symbolic modeling tasks. A number of studies have examined the role of intonation in differentiating spontaneous and read speaking styles; Hirschberg [2] examined some symbolic representations, while Laan [3] only investigated direct representations of prosody. Teixeira et al. [4] explored the classification of speaker nativeness using direct prosodic modeling; we believe this is the first application of symbolic modeling of prosody to this task.

Symbolic modeling is used less often than direct modeling when incorporating prosodic information for spoken language processing tasks. Manual annotation of prosodic labels is time-consuming and expensive. Direct modeling approaches are particularly attractive when the amount of data to be processed makes manual annotation of prosody impossible. Along with manual ToBI labels, we explore the use of the AuToBI [5] system to generate hypothesized ToBI labels automatically. AuToBI is an open-source toolkit for the detection and classification of prosodic events. There has been significant experimentation incorporating prosodic information in spoken language processing tasks. Shriberg and Stolcke [6] summarized a number of successful applications of direct modeling of prosody, including sentence segmentation, speech act classification and speaker recognition. Eidelman et al. [7] recently found that part-of-speech tagging can be improved by incorporation of symbolic phrase boundaries. Chen et al. [8] demonstrated the use of prosodic categories to improve speech recognition.

We describe the ToBI standard and experimental material in Section 2. In Section 3, we present the details of the modeling approaches that are evaluated including a novel sequential pitch modeling approach. Experimental results are discussed in Section 4. Section 5 contains a concluding discussion and directions for future work.

2. Material

The ToBI (Tones and Break Indices) standard of prosodic annotation [1] describes prosody as a sequence of high (H) and low (L) tones associated with three prosodic events. These events are pitch accents, which indicate acoustic stress or prominence, and two levels of phrase boundaries. ToBI describes phrasing as a hierarchical structure where each intonational (full) phrase contains one or more intermediate phrases, which in turn contain one or more words. Each intermediate phrase boundary is associated with a phrase accent. Each intonational phrase boundary, which is by definition also an intermediate phrase boundary, is associated with a boundary tone. ToBI also includes annotation for catathesis (!H), when speech is produced in a compressed pitch range. In the case of pitch accents, high

and low tones can be used to construct complex tones describing pitch movement and alignment during the associated pitch accent. The full ToBI standard includes annotation of 8 types of pitch accents (H*, !H*, L*, L+H*, L+!H*, L*+H, L*+!H, H+!H*), 3 types of phrase accents (H-, L-, !H-) and 2 boundary tones (H%, L%). These ToBI annotations comprise the vocabulary of symbols that we will use in the symbolic modeling approaches described in Section 3.

To evaluate the use of symbolic and direct sequential modeling, we classify speaking style and nativeness. The four speaking-styles, READ, SPONTANEOUS, BROADCAST NEWS (BN) and DIALOG, are drawn from three corpora. The Boston Directions Corpus (BDC) [9] contains spoken material from four speakers delivering both spontaneous elicited monologues and reading their own spontaneous monologues approximately two weeks after their original production. Each BDC file represents a unique direction giving task. In total there are 50 minutes of read and 60 minutes of spontaneous speech. Both are spoken by the same four speakers. These two subcorpora will be used to represent READ and SPONTANEOUS speech. The Boston University Radio News Corpus (BURNC) [10] is a corpus of professionally read radio news data. A 2.35 hour subset from six speakers (three female and three male) has been annotated with the full ToBI standard. Each file represents a paragraph of BN speech. This material will be used to represent the BROADCAST NEWS or BN style of speech. The Columbia Games Corpus (CGC) [11] 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. Neither subject could see the other participant. Each file represents a transcript from a whole game. The corpus includes approximately 320 minutes of speech. The DIALOG material is drawn from the CGC. It is valuable to note that this "dialog" material is evaluated based on one side of the dialog at a time; only one speaker is present in any training or evaluation file.

In the nativeness classification experiment, we will compare NATIVE or L1 speech drawn from the BURNC, to NON-NATIVE or L2 speech productions of the same material. The material contains two news stories drawn from BURNC: **p** – computerized parole officers – and **r** – the Safe Roads Act. The non-native material was read by 4 native Mandarin Chinese speakers, between 25 and 30 years old, with 6 to 19 years of experience with English. The non-native material has been annotated using the ToBI standard. Each file in this corpus contains a full BURNC story. While we refer to this corpus as non-native or L2 speech, it is important to acknowledge that it only represents L2 productions by native Mandarin Chinese speakers rather than "non-nativeness" defined more broadly.

In two experiments, we evaluate the use of hypothesized ToBI tones in symbolic sequential modeling. All hypothesized ToBI tones were generated using AuToBI [5] with models trained on all of the BDC material, both read and spontaneous speech. To avoid any influence of the training data on hypotheses, when using hypothesized tones, BDC material is excluded from the experiment.

3. Sequential Modeling

To perform the symbolic sequential modeling, we use a tri-gram model over ToBI tone sequences. Within this model, each symbol represents the tone and type of prosodic event with which it is associated. These models are trained using SRILM [12] and Good-Turing smoothing. We explore two variants of this model, in one we include only ToBI tones, in the other we include a DEACCENTED annotation when a word does not bear pitch accent. These two modeling strategies are referred to as ToBI and ToBI-Deacc. In the ToBI setting, these sequences are made up of only pitch accents, phrase accents and boundary tones. The ToBI-Deacc also includes information about deaccented words. These prosodic event sequences are generated from manual annotations of prosody. To evaluate the use of automatic prosodic labeling in symbolic sequential modeling, we also hypothesize prosodic events using AuToBI [5]. We use identical modeling approaches when using the manual and automatic ToBI labels. The two models using automatically hypothesized events are referred to as AuToBI and AuToBI-Deacc.

We compare these symbolic approaches to three direct sequential models. Two of these direct approaches approximate a symbolic approach by calculating a discrete value for each word. We extract the mean pitch (log Hz) over each word and determine if it is above or below the speaker's mean pitch. This leads to two pitch values: HIGH and LOW. The mean value to treshold high and low values is calculated over each wave file as described in Section 2. This binary quantization loses a lot of information by considering values barely above or below the mean value to be as prosodically salient as those dramatically above or below. To be more sensitive to differences in pitch, we also use a four-way quantization of mean pitch in each word. Under this quantization scheme, there are two low and high values corresponding to values that are within one standard deviation of the mean, and values that are more than one standard deviation from the mean.

In addition to using aggregations of pitch extracted at the word level, we also experiment with a sequential model of pitch values, $p(x_i|x_{i-1})$, modeling the joint $p(x_i, x_{i-1})$ as a two dimensional Gaussian. To make sure that the sequential model is robust to speaker differences, we normalize each pitch (logHz) value. Pitch information is extracted from each audio file at 10ms intervals using AuToBI's implementation of Praat's Get Pitch...(ac) algorithm [13]. The speaker normalization is performed using z-score normalization based on mean and standard deviation values calculated over each file. However, this model of pitch values ignores the fact that there are regions of speech where no pitch information is available, either because the speaker is silent or because the current phone is unvoiced. Therefore, we use of a pitch model that represents regions where no pitch is extracted along with pitch information. This model incorporates the likelihood that a pitch value will immediately precede or follow an area with no available pitch. Rather than treating frames with no pitch as 0 logHz or some other imputed value and risk skewing the normalcy or variance of the logHz distribution, we allow empty pitch values to exist in a separate dimension. Thus, we model a pitch distribution as a mixture of a Gaussian distribution for valid pitch values and a point degenerate distribution for empty pitch regions. This *spike-and-slab* model is very closely related to the hierarchical model described by Ishwaran and Rao [14]. Under this model the likelihood of a pitch value is represented by

$$p(x) = \delta(x = \emptyset)\pi + \delta(x \neq \emptyset)(1 - \pi)N(x; \mu, \sigma^2)$$

where $\delta(\ldots)$ is the Kronecker delta function, evaluating to 1 when the proposition is true, and 0 otherwise, π is the mixture coefficient between the valid and empty pitch mixture components, and μ and σ^2 are the mean and variance of the logHz pitch distribution.

To construct a sequential model, we calculate the conditional distribution of each pitch point given the previous value

under this spike-and-slab mixture model. This sequential model is similar to the Multi-Space Probability (MSP) HMM model proposed by Tokuda et. al [15]. Where their model is state-conditioned, the spike-and-slab model directly models pitch as a markov chain, without a latent variable. Under this sequential model, entire sequences of frames without pitch are represented by a single null pitch point. Thus, zero probability mass assigned to the case where $x_i = \emptyset \wedge x_{i-1} = \emptyset$. The sequential probability is defined as follows,

$$p(x_i|x_{i-1}) = \frac{p(x_i, x_{i-1})}{p(x_{i-1})},$$

$$p(x_i, x_{i+1}) = \begin{cases} \pi(1-\pi)N_f(x_i; \mu_f, \sigma_f^2) & \text{if } x_{i-1} = \emptyset \\ \pi(1-\pi)N_p(x_{i-1}; \mu_p, \sigma_p^2) & \text{if } x_i = \emptyset \\ (1-\pi)^2 N(x_i, x_{i-1}; \vec{\mu}, \Sigma) & \text{otherwise} \end{cases}$$

where $N_f(x; \mu_f, \sigma_f^2)$ is a Gaussian model of values *following* empty regions $N_p(x; \mu_p, \sigma_p^2)$ is a model of values *preceding* empty regions, and $N(x_i, x_{i-1}; \vec{\mu}, \Sigma)$ is a two-dimensional normal distribution of pitch values.

To motivate the use of this model of pitch, we calculate sequential log-likelihood of the pitch contour extracted from the BDC-read audio file h2r1 modeled using the *spike-and-slab* model and a Gaussian distribution. To evaluate the generalization performance we also calculate the log-likelihood of another file, h2r2, read by the same speaker under the models trained on h2r1. These results can be found in Table 1. The spike-and-slab model has a higher training and generalization log likelihood than the Gaussian model. Not only does the spike-and-slab model generate better fits of pitch information, but it incorporates additional information about the pitch sequence by representing areas where no pitch is available.

	h2r1 (train)	h2r2 (eval)
Gaussian	-429.687	-541.262
Spike-and-Slab	210.708	-421.757

Table 1: $\log p(x)$ of Gaussian and Spike-and-Slab models.

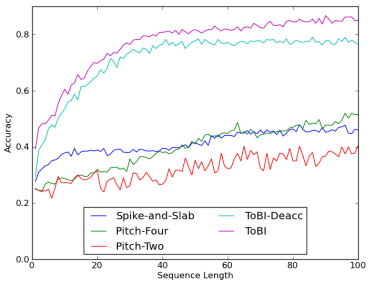
4. Experiments

In this section, we describe three sets of experiments. Each set of experiments uses the same framework. We divide the material into training and testing sets. We train sequential models based on the training material and then evaluate based on sequences drawn from the test material. In each experiment, we train one model for each speaking-style or nativeness class. To evaluate the impact of the sequence duration we extract sequences from $N \in [1, 100]$ words in length. The performance is evaluated by randomly selecting 500 distinct sequences of length N for each class and calculating the overall accuracy. These sequences may be extracted from overlapping regions. We take the sequential model which generated the highest log likelihood as the hypothesis.

We raise one caveat: these results may, in part, represent labeler idiosyncrasies as well as difference in the speech material. Each of the four corpora investigated were annotated by distinct ToBI labelers. While ToBI annotation shows high inter-annotator agreement [16], there is rarely perfect consensus.

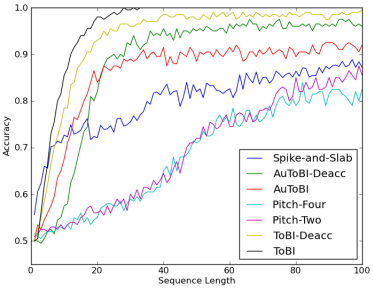
4.1. Four-way Speaking-Style Classification

We evaluate the use of symbolic and direct sequential modeling for the classification of four speaking styles: READ, SPONTANEOUS, BN and DIALOG. From each corpus, we identify a single male speaker for testing: h2 for the BDC material, m1b from BURNC and 101 from CGC. No training material is spoken by the evaluation speaker.



	Acc. @100
ToBI	85.0
ToBI-Deacc	76.5
Pitch-Two	40.5
Pitch-Four	51.5
Spike-&-Slab	46.2

Figure 1: Four-way Speaking-Style Classification Accuracy.



	Acc. @100
ToBI	100.0
ToBI-Deacc	99.5
AuToBI	92.0
AuToBI-Deacc	96.0
Pitch-Two	85.5
Pitch-Four	82.5
Spike-&-Slab	87.2

Figure 2: Binary Speaking-Style Classification Accuracy.

Figure 1 reports the classification accuracy from each of the five sequential modeling approaches while varying the sequence size from 1 to 100 words, highlighting the accuracy at 100 words . The sequential models are ToBI, ToBI-Deacc, Pitch-Two, Pitch-Four and Spike-and-Slab. In this experiment, we find that symbolic sequential modeling clearly outperforms each direct modeling approach. We find that including a NONE tone for non-accented words yields reduced performance on this task. This suggests that the difference in speaking-style is based more on the speaker's use of prosodic contour, rather than the rate or context of deaccented words.

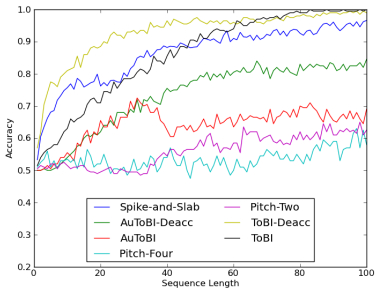
4.2. Binary Speaking-Style Classification

To evaluate the use of hypothesized prosodic event sequences in symbolic sequential modeling, we exclude the BDC material from the experiments. The AuToBI models used in the hypothesis generation were trained on BDC material. This leads to a binary classification task, distinguishing BN from DIALOG.

Figure 2 reports the classification accuracy from each of the seven sequential modeling approaches while varying the sequence size from 1 to 100 words. The sequential models are ToBI, ToBI-Deacc, AuToBI, AuToBI-Deacc Pitch-Two, Pitch-Four and Spike-and-Slab. In this experiment, we find again that symbolic sequential classification is a more reliable approach to speaking-style classification than direct modeling. At 33 words, the manual ToBI tone sequence can classify these two speaking-styles with 100% accuracy. We find that while the hypothesized tones do not perform as well as the manual tones, they also outperform all three direct modeling approaches when evaluated on sequences of more than 12-15 words. We also note that, the AuToBI-Deacc model out performs the AuToBI model, yet the ToBI model outperforms the ToBI-Deacc model. It is possible that AuToBI model consistently over- or under-predicts pitch accents in one of the two speaking styles.

4.3. Nativeness Classification

In this experiment, we examine the use of symbolic and direct sequential modeling for the classification of speech as spoken



	Acc. @100
ToBI	100.0
ToBI-Deacc	99.0
AuToBI	69.0
AuToBI-Deacc	84.5
Pitch-Two	62.5
Pitch-Four	58.0
Spike-&-Slab	96.6

Figure 3: Binary Nativeness Classification Accuracy.

by a native or non-native speaker. The training and evaluation material in this study is smaller than in the speaking style experiments. We use only two stories from the BURNC for training and evaluation. As usual, the evaluation material includes a single isolated speaker – speaker m1b from the BURNC and s03 from the Mandarin Chinese material. Figure 3 contains the classification accuracy from each of the seven sequential modeling approaches while varying the sequence size from 1 to 100 words. The sequential models are identical to those used in the binary speaking style experiment: ToBI, ToBI-Deacc, AuToBI, AuToBI-Deacc Pitch-Two, Pitch-Four and Spike-and-Slab. Again, we find that symbolic sequential modeling of prosodic sequences yield the best performance on this task. However, these models require the availability of manual ToBI annotations. The next best results are obtained by the spike-and-slab pitch model. This indicates that direct modeling can be successfully applied to the task of nativeness classification. The success of direct modeling here captures the different pitch behavior used by native American English speakers and native Mandarin Chinese speakers. Since Mandarin is a tonal language, it is unsurprising that its speakers use pitch differently. While the speakers also use prosody distinctly as reflected in the ToBI-Deacc results, these distinctions are not reflected in the AuToBI hypotheses to an extent sufficient to outperform the spike-and-slab model. It is common for Mandarin Chinese speakers of English to produce more pitch accents than native speakers [17]. This difference in accent rate explains why in this task the ToBI-Deacc and AuToBI-Deacc models outperform their counterparts that do not include a token for deaccented words.

5. Conclusion and Future Work

Sequential modeling of prosody is a simple and efficient technique for modeling long range trends in prosodic variation. In this paper, we show how these models can be capture qualities such as speaking-style and nativeness. In both of these tasks, we find that symbolic modeling with ToBI tones yields the best performance indicating that the symbolic representation is a valuable intermediate representation of prosody. We find that in classifying speaking-style, a symbolic sequential model based on hypothesized ToBI tones is still able to generate reliable results. While the *spike-and-slab* direct modeling approach performs poorly for speaking-style classification, it performs better than hypothesized prosodic events in classifying nativeness. Overall, we find that symbolic representations are a valuable compact representation of prosodically relevant acoustic information. Improved hypothesized prosodic symbols to may approach the classification performance of manually annotated ToBI tones on these two tasks.

Symbolic modeling is a less popular approach to prosodic analysis, largely due to the investment in time and resources to perform prosodic annotation. Direct modeling approaches are often used in situations where the amount of data makes manual annotation of prosody impossible. Through AuToBI, we believe symbolic modeling approaches can be evaluated on larger datasets without manual annotation. In the future, we will explore the use of symbolic modeling on segmentation, structural tagging, speaker recognition, and word recognition – tasks that have typically used only direct modeling due to a lack of manual annotation of categorical representations of prosody, and unreliable automatic automatic annotation procedures.

6. Acknowledgements

The author would like to acknowledge Julia Hirschberg for many constructive comments.

7. References

[1] K. Silverman, et al., "Tobi: A standard for labeling english prosody," in *ICSLP*, vol. 2, 1992, pp. 12–16.

[2] J. Hirschberg, "A corpus-based approach to the study of speaking style," in *Prosody, theory and experiment: studies presented to Gösta Bruce*, G. Bruce and M. Horne, Eds. Springer, 2000.

[3] G. P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, no. 1, pp. 43 – 65, 1997.

[4] C. Teixeira, H. Franco, E. Shriberg, K. Sönmez, and K. Precoda, "Prosodic features for automatic text-independent evaluation of nativeness for language learners," in *ICSLP*, 2000.

[5] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[6] E. Shriberg and A. Stolcke, "Direct modeling of prosody: an overview of applications in automatic speech processing," in *Speech Prosody*, 2004, pp. 575–582.

[7] V. Eidelman, Z. Huang, and M. Harper, "Lessons learned in part-of-speech tagging of conversational speech," in EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 821–831.

[8] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole, and J. Choi, "Prosody dependent speech recognition on radio news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 232–245, 2006.

[9] 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.

[10] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

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

[12] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *ICSLP*, 2002.

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

[14] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: Frequentist and bayesian strategies," *The Annals of Statistics*, vol. 33, no. 2, pp. 730–773, 2005.

[15] K. Tokuda, T. Masuko, N. Miyakazi, and T. Kobayashi, "Multi-space probability distribution hmm," *IEICE Transactions on Information and Systems*, vol. E85-D, no. 3, March 2002.

[16] A. Syrdal and J. McGory, "Inter-transcriber reliability of tobi prosodic labeling," in *ICSLP*, 2000.

[17] A. Rosenberg and J. Hirschberg, "Production of english prominence by native mandarin chinese speakers," in *Workshop on Prosodic Prominence: Perceptual, Automatic Identification*, 2010.