10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Word-based Neural Prosody Modeling with ToBI

Hee Hwang, Kristine M. Yu

University of Massachusetts, Amherst

hhwang@cs.umass.edu, krisyu@linguist.umass.edu

Abstract

We present a neural model of American English intonation using the discrete tonal transcription system MAE-ToBI. The model uses the words and tonal sequences of the MAE-ToBI annotated portion of the Boston University Radio Speech Corpus. We took as a starting point Dainora's probabilistic finite-state grammar of the tonal sequences of two speakers in the corpus. We extended Dainora's grammar to cover all six speakers in the corpus and found that bigram probabilities and distinctions in distribution of tones over pre-nuclear and nuclear intermediate phrases showed the same patterns as Dainora's results. To expand beyond her work, we built a word-based Long Short-Term Memory (LSTM) neural model that predicts the MAE-ToBI sequence within an intermediate phrase. We used both randomly initialized vector word embeddings and pre-trained word embeddings from BERT, a bidirectional transformer. BERT achieved 80.58%, 99.79%, and 90.74% accuracy in detecting pitch accent, intermediate, and intonational phrase boundaries. The results demonstrate that it is possible to predict prosody given only texts and discrete prosodic labels without acoustic information. The improvement with BERT demonstrates how the addition of unannotated text data to a small prosodically annotated corpus could be leveraged for prosodic modeling in low-resource languages.

Index Terms: prosody, phonetics, natural language processing, computational linguistics, recurrent neural network, attention

1. Introduction

1.1. Motivation

The connection between linguistic theory on sentence-level prosody and its interfaces, on the one hand, and sentence-level prosody modeling in automatic speech recognition (ASR) and text-to-speech (TTS), on the other, has steadily decreased over the years. Over the years up to the current time, linguistic theories on intonation and its interfaces with syntax and semantics/pragmatics have been defined over abstract prosodic categories such as pitch accents and prosodic boundaries [1, 2, 3, 4, 5]. However, prosodic features in ASR and TTS have moved away from being introduced as these prosodic categories towards being introduced as raw acoustic features [6, 7, 8, 9, 10]. Moreover, attempts at incorporating highly specific contextual morphosyntactic/semantic/pragmatic features [11, 7] has given way to incorporating such features via opaque embeddings in neural nets [9, 10]. The move towards embeddings in neural parsing and speech synthesis with the raw speech signal has resulted in remarkable gains in performance, without the need for expert-annotated prosodic category labels or contextual rules. This comes at the cost of facilitating crosstalk between ASR/TTS and linguistic theory for deepening our scientific understanding of prosodic interfaces, especially in under-resourced languages where there isn't sufficient data available to adequately train neural nets.

In this paper, we work towards bridging this disconnect between ASR/TTS and linguistic theory. We present a proof-of-concept demonstration of how neural nets can be used to facilitate the prediction of the most likely discrete strings of prosodic tones associated with an input word sequence in American English, even with a limited set of prosodically labeled training data. We use the prosodically annotated portion of the Boston University Radio Speech Corpus [12] for training and testing. We take as a starting point Dainora's [13] probabilistic model trained on a portion of this corpus, which in turn was based on Pierrehumbert's finite-state acceptor of English intonational sequences [1] that became the basis for the prosodic annotation system of Mainstream American English ToBI. Dainora augmented Pierrehumbert's finite state grammar to include some context from preceding tones in her first and second-order Markov models over MAE-ToBI label sequences.

But Dainora's model still lacks any contextual information from the sentence, pointing to a disconnect that exists within linguistic work, between generative grammars for prosodic strings [1] and generative models mapping between sentences and prosody [4]. Namely, Pierrehumbert's finite state grammar (and Dainora's elaboration thereof) is an acceptor over pitch accents and boundary tones, rather than a transducer, and does not include any contextual information from the sentence (and minimal contextual information from preceding tones). Theories of the syntax-prosody interface define transductions between syntactic trees and prosodic trees, which express prosodic constituency, while not specifying pitch accents that appear in Pierrehumbert's finite state grammar. How can we combine theories of these two domains together to map from a sequence of words to a sequence of tones? Here, we try using word embeddings.

Bengio et al.[14] introduced a neural probabilistic language model that used so-called word embeddings to encode the meaning of words. A word embedding is a vector of real numbers derived from a word. This representation could hold synonymity and antonymy. For example, $vector(\text{“}cat\text{”}) - vector(\text{“}kitten\text{”})$ is very similar to $vector(\text{“}dog\text{”}) - vector(\text{“}puppy\text{”})$. NLP tasks such as text generation, sequence classification, question answering, and named entity recognition perform extremely well with this concept. Word embedding is also useful for parsing[15, 9] and sentiment analysis[16].

Here, we investigate using word embeddings to predict prosody. Prosodic annotations are expensive, especially for low resource languages. There are several transcription models[17, 18, 19] that predict the fundamental frequency contour. Unlike these models, ToBI[20] uses abstract discrete symbols H and L to represent high and low tones. This simplification comes in handy when generalizing over tonal patterns. This paper combines word embeddings with ToBI annotations as a first step. Any system of discrete prosodic symbols, e.g., one derived from [21], could also be used.

10.21437/SpeechProsody.2020-208

1.2. Problem Definition

Our goal is to predict tonal patterns from sentences. Given input word sequences, our model classifies the phrases into ToBI tunes, i.e., a sequence of tone sequences, where each word is aligned with its corresponding tone sequence. After getting the tone sequences, we conduct four different tasks. First, we test how well the model predicts the exact tone sequences. After getting the tune, we identify the pitch accent in the intermediate phrase. Finally, we predict boundary tones in both intermediate phrases and intonational phrases.

1.3. Our Approach

We used ToBI-transcribed data from the Boston University Radio Speech Corpus (BU)[12]. We aligned words with ToBI symbols to employ a neural network based on word embeddings. These word embeddings can be pre-trained and can possibly contain rich contextual information. Our research is based on two assumptions. First, prosody is stochastic and context-dependent. The context consists of adjacent words and tones, syntax, and semantic information. Second, even among diverse English speakers, prosody has shared/common patterns. To prove these assumptions, we tested previous work on prosody modeling with ToBI, and the results were confirmed/validated by tests using six speakers of Boston University Radio News Corpus.

We tested the model with prosodic phrases and corresponding ToBI labels using a Long Short-Term Memory (LSTM) recurrent neural network[22] as well as BERT[23], a bidirectional transformer that uses pre-trained contextual embedding. After getting the prediction from the model, we used this output to perform different tasks. First, we checked if the sequence had a pitch accent. For instance, H*L-L% has one pitch accent, H*. Second, we inspected the output to see if the tone sequences have intermediate phrase boundary tones. Third, we did a similar test to detect the IP phrase boundary tone.

1.4. Main Contributions

We re-examined the previous work on probabilistic prosody modeling and tested it on extensive data. Through the use of word embedding and neural networks, we created LSTM and BERT models that predict pitch accents and boundary tones using English sentences and corresponding ToBI labels. Using this prediction model would help researchers to unravel underlying linguistic patterns inside the tune sequences. In the rest of this paper, we further discuss the replication of the probabilistic prosody model based on Dainora (Section 2), word-based tone representation (Section 3), and neural models (Section 4). Code will be released at `https://github.com/heeh/neural_prosody`

2. Probabilistic Prosody Model

In this section, we verify Dainora's probabilistic grammar approach based on two speakers in the BU corpus with all six speakers in the corpus, because (1) we believe general tune patterns exist across multiple speakers and that (2) the tune patterns must show consistency within a larger dataset. We verify that Dainora's reported patterns of nuclear tune distribution, differences in pitch accent distribution in non-final and final positions, and counts of intonational and intermediate phrases, tokens, tunes, and pitch accents generalize to the larger data set.

2.1. Data

The BU Corpus includes sound files, words, part of speech tagging, and ToBI labels. Each file contains 3-8 seconds of speech. The corpus contains speech from three women and four men and two types of news, live and pre-recorded. Dainora used these two types from Speakers F1A and F2B. We replicate her findings and expand this model to six speakers using both styles. We collected files that have ToBI transcription and removed two ill-formed files, f2bs11p2 and m3bprlp4. Table 1 summarizes the data used.

Speaker	**|IPs|**	**|ips|**	**|tokens|**
$f1a$	807	1204	4385
$f2b$	2808	3913	12776
$f3a$	440	739	2809
$m1b$	771	1226	5062
$m2b$	658	968	3608
$m3b$	291	439	1935
sum	5771	8489	30575

Table 1: *Details of BU corpus data used*

2.2. Distribution of nuclear ip tunes

We collected nuclear pitch accents over the whole dataset and compared it with Dainora's distribution. The order of the five most frequent nuclear ip tunes remains consistent, although frequency differs.

Nuclear Tune	**Occurrences** ($Dainora$ / All)	**Frequency(%)** ($Dainora$ / All)
H*L-L%	398 / 3220	33 / 45
H*L-H%	276 / 2017	23 / 28
L+H*L-H%	158 / 590	13 / 8
L+H*L-L%	116 / 538	10 / 7
L*L-H%	75 / 263	6 / 4

Table 2: *Top five most frequent nuclear ip tunes*

2.3. Number of pitch accents in ips

Table 3 shows pitch accent frequency in intermediate phrases. We confirmed that the difference between two kinds of intermediate phrases, nonfinal and final, generalized to the six speakers. Dainora used 568 nonfinal and 1207 final intermediate phrases while we used 2900 and 6058, respectively.

D: Dainora's / A: All

*	Non-final ip(D/A)		Final ip(D/A)	
	Occur	Freq(%)	Occur	Freq(%)
1	303 / 1394	53 / 52	387 / 1983	32 / 34
2	201 / 936	35 / 35	531 / 2435	44 / 42
3	52 / 254	9 / 9	210 / 982	17 / 17
4	9 / 71	2 / 3	58 / 266	5 / 5
5	2 / 26	<.5/ 1	13 / 75	1 / 1
6	1 / 2	<.5/ <.5	8 / 25	1 / <.5
7	0 / 2	0 / <.5	0 / 7	0 / <.5

Table 3: *Frequency distribution of number of sequential pitch accents within nonfinal and final intermediate phrases*

2.4. Nuclear and prenuclear accents

Table 4 shows the distribution of pitch accent types in nuclear and prenuclear accent positions. Dainora's results showed that H* tones constitute 74% of pitch accents in non-final intermediate phrases but only 60% of pitch accents in final intermediate phrases. However, our results produce 77% and 72% respectively, narrowing the gap between the two kinds of ips.

D: Dainora's / A: All

	Non-final ip(D/A)		Final ip(D/A)	
	Occur	Freq(%)	Occur	Freq(%)
H*	899 / 5462	74 / 77	730 / 5459	60 / 72
L*	31 / 228	3 / 3	94 / 304	8 / 4
L+H*	214 / 1063	18 / 15	299 / 1564	25 / 20
L*+H	10 / 33	1 / <.5	6 / 12	0 / <.5
H+!H*	60 / 327	5 / 5	78 / 248	6 / 3

Table 4: *Frequency distribution of pitch accent types within nonfinal and final intermediate phrases*

2.5. Summary and Limitations

Dainora's probabilistic model clearly shows consistency using a larger data set. However, her model concerns the distribution of only tone sequences and doesn't reveal the relationship between tunes and linguistic structure from syntax, semantics, and pragmatics. This limitation signals that we need to find a way to link such contextual information to tone sequences.

3. Word-based Tone Representations

3.1. Definition

A first step towards linking up such contextual information is via words. A word can carry both a tone sequence and contextual information and is a rich yet manageable unit of information. The BU Corpus provides timestamps for syllable, word, part of speech (POS), and tone sequences. We labeled each word with ToBI tones based on the timestamps. By applying this procedure to the whole dataset, we essentially convert the problem of prosodic prediction into a word sequence tagging task; we acquire tonal information from a given word sequence. To illustrate, consider the example below:

Marianna	made	the	marmalade
H*	O	O	H*L-L%

Notice that "made" and "the" lack tones while marmalade has a tone sequence comprised of two tones, H* and L-L%. When words do not have tones, we mark them as 'O.'

This word-based tone representation has several advantages over the tone-only model. Our tonal sequence can carry words that contain actual meaning and linguistic information, such as a part of a speech tag. Furthermore, the word-based alignment enables us to directly apply a state-of-the-art neural model that uses pre-trained contextual word embedding. The contextual word embedding means that a word representation depends on the sentences or documents which contain it. For example, the word 'apple' has a different representation in "I ate an apple" and "This is an Apple computer."

4. Neural Model

We showed that the probabilistic tonal sequence model shows consistency when scaling up. We introduced word embedding, which contains contextual information as well as tonal sequences. These two ideas allow us to build a neural model to do sequence prediction tasks, where the inputs are word sequences corresponding to intermediate phrases, and the outputs are corresponding ToBI labels, as schematized in Figure 1.

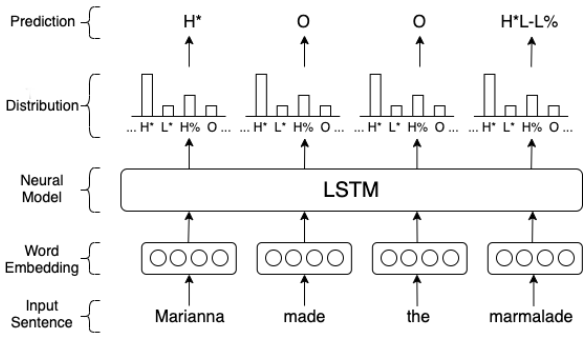

Figure 1: *Neural Model*

4.1. Data

In training, we aligned each word with its tone sequence and grouped the resulting tunes by intermediate phrase. We made all tone sequences equal length by padding them when necessary. The data breakdown in terms of train, test, and dev fell into a 3:1:1 ratio. Below are the number of intonational phrases (IPs), intermediate phrases (ips), and tokens, and the mean and standard deviation of the ip length.

	|IPs|	**|ips|**	**|tokens|**	**mean(|ip|)**	**stdev(|ip|)**
train	3461	5070	18570	3.64	1.88
test	1121	1705	5970	3.50	1.83
dev	1193	1714	6035	3.51	1.77

Table 5: *Data splits for BU corpus*

Since the vocabulary of words came only from the training set, we also checked the vocabulary hit rate for the dev and test sets and confirmed that it was reasonably high: close to 90% for both.

	Total words	**hit**	**miss**	**rate(%)**
test	5970	5338	632	89.41
dev	6033	5251	782	87.03

Table 6: *Vocabulary Hit Rate*

4.2. LSTM

LSTM uses a word embedding to process the input word and output tone sequence. The word embedding is a numeric representation obtained by assigning a word or tone sequence to a distinct integer. Thus, the input and output of the model become arrays of integers. The LSTM predicted the presence of pitch

accents well; however, the model is not as good at predicting the presence of boundary tones.

Test	precision	recall	F1
pitch accent(*)	76.68	74.22	**75.43**
ip boundary tone(-)	100.00	47.16	**64.09**
IP boundary tone(%)	100.00	47.81	**64.70**

Table 7: *LSTM Pitch Accent and Boundary Tone Detection*

		True Label	
		pitch accent(*)	No Label
Prediction	accent(*)	2197	668
	No Label	763	2342
	Total	2960	3010

		True Label	
		ip bound(-)	IP bound(%)
Prediction	Correct	804	536
	Incorrect	901	585
	Total	1705	1121

Table 8: *LSTM Confusion Matrix on Test Set*

4.3. BERT

BERT[23] is a state-of-the-art pre-trained deep bidirectional representation from unlabeled text that can offer context. While LSTM produces the same word embedding for a particular word, BERT uses different word embedding for each sentence. These embeddings are pre-trained by English text outside our corpus. This model performed better than the LSTM did, achieving 80.58%, 99.79%, and 90.74% in predicting the presence of pitch accent and boundary tones. The improvment over the LSTM clearly demonstrates that prosody modeling is improved in a context-rich environment.

Test	precision	recall	F1
pitch accent(*)	82.24	78.99	**80.58**
ip boundary tone(-)	100.0	99.59	**99.79**
IP boundary tone(%)	100.0	83.05	**90.74**

Table 9: *BERT Pitch Accent and Boundary Tone Detection*

		True Label	
		pitch accent(*)	No Label
Prediction	accent(*)	2338	505
	No Label	622	2505
	Total	2960	3010

		True Label	
		ip bound(-)	IP bound(%)
Prediction	Correct	1698	931
	Incorrect	7	190
	Total	1705	1121

Table 10: *BERT Confusion Matrix on Test Set*

5. Comparison

In comparing the two models, we focus on the detection of pitch accents as well as intermediate and intonational boundaries. In all categories, BERT outperformed LSTM. We hypothesize that BERT predicted boundary tones so well because BERT is highly focused on endings via the [SEP] token, a separator between sentences, according to Clark[24]. Predicting exact tunes turned out to be challenging for both models since the model predicts more than 100 tone sequence labels, while data is limited to only 30,000 tokens. In principle, the number of possible tone sequences on a word is unbounded; for instance, it is possible to observe tone sequences as long as L+H*L+H*L+H*H-, from an acronym WBUR.

Test	LSTM	BERT
Pitch accent(*)	75.43	**80.58**
ip boundary(-)	64.09	**99.79**
IP boundary(%)	64.70	**90.74**
Exact Tune	26.75	**39.03**

Table 11: *F1 score, LSTM vs. BERT*

Table 12 shows predicted tone sequences over words by each model. The LSTM shows a high frequency of H* tones. This shows the source of LSTM's low accuracy. BERT's predictions also skewed towards tunes with H*,but were more widely distributed across different tone sequences.

Tune	LSTM	BERT
H*	1976	1265
H*L-L%	600	529
H*H-	171	341
L+H*	0	293
L-L%	31	200
L-H%	231	183
H*L-H%	29	151

Table 12: *Tone sequences predicted*

6. Conclusions

We discussed Dainora's probabilistic prosody model of the BU corpus and verified its consistency with a bigger dataset. We induced word-based tonal sequences using word embeddings with LSTM and BERT models. The better performance of the BERT model demonstrates that context affects the prediction of tone sequence and an inexpensive way to introduce context in prosodic modeling. We believe that neural modeling has the capacity to reveal properties of the relation between tunes and syntactic/semantic/pragmatic structure such as inversion and fronting. Moving forward in this work also invites many methodological questions in designing neural models for prosodic prediction.

7. Acknowledgements

Jean Joyce and Nicholas Monath from the UMass CIIR department helped us to retrieve Boston University Radio News Corpus. Tu Vu provided us with helpful comments on BERT as well.

8. References

[1] J. B. Pierrehumbert, "The phonology and phonetics of english intonation," *Ph.D. dissertation, Massachusetts Institute of Technology*, 1980.

[2] R. D. Ladd, "Phonological features of intonational peaks," *Language*, vol. 59, pp. 721–759, 1983.

[3] J. Pierrehumbert and J. B. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in communication*, 1990, pp. 271–311.

[4] E. Selkirk, "The syntax-phonology interface," in *The Handbook of Phonological Theory*, J. Goldsmith, J. Riggle, and A. C. L. Yu, Eds. Wiley-Blackwell, 2011, pp. 435–484.

[5] S. Jeong, "Intonation and sentence type conventions: Two types of rising declaratives," *Journal of Semantics*, vol. 35, no. 2, pp. 305–356, 2018.

[6] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*. Springer, 2004, pp. 105–114.

[7] J. Hirschberg, "Speech synthesis, prosody," *Encyclopedia of language & linguistics*, vol. 7, pp. 49–55, 2006.

[8] J. K. Pate and S. Goldwater, "Unsupervised dependency parsing with acoustic cues," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 63–74, 2013.

[9] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 69–81. [Online]. Available: https://www.aclweb.org/anthology/N18-1007

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[11] N. M. Veilleux, "Probabilistic model of acoustic/prosody/concept relationships for speech synthesis," in *Concept to Speech Generation Systems*, 1997.

[12] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "Boston university radio speech corpus," *Linguistic Data Consortium*, no. LDC96S36, 1996.

[13] A. Dainora, "An empirically based probabilistic model of intonation in American English," Ph.D. dissertation, University of Chicago, Chicago, IL, 2001.

[14] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944966

[15] R. Socher, J. Bauer, C. Manning, and A. Ng, "Parsing with compositional vector grammars," *ACL Conference*, 2013.

[16] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," *EMNLP*, 2013.

[17] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[18] P. A. Taylor and A. W. Black, "Synthesizing conversational intonation from a linguistically rich input," in *SSW*, 1994.

[19] D. Hirst, A. Di Cristo, and R. Espesser, *Levels of Representation and Levels of Analysis for the Description of Intonation Systems*. Dordrecht: Springer Netherlands, 2000, pp. 51–87.

[20] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi:a standard for labeling english prosody," *International Conference on Spoken Language Processing*, 1992.

[21] J. Cole and S. Shattuck-Hufnagel, "New methods for prosodic transcription: Capturing variability as a source of information," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 7, no. 1, 2016.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv e-prints*, p. arXiv:1810.04805, Oct 2018.

[24] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.