10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Prosody Prediction from Syntactic, Lexical, and Word Embedding Features

Rose Sloan[1], *Syed Sarfaraz Akhtar*[1], *Bryan Li*[1],
Ritvik Shrivastava[1], *Agustín Gravano*[2,3], *Julia Hirschberg*[1]

[1]Department of Computer Science, Columbia University, New York, USA
[2]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[3]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

rsloan@cs.columbia.edu, ssa2184@columbia.edu, b.li@columbia.edu,
rs3868@columbia.edu, gravano@dc.uba.ar, julia@cs.columbia.edu

Abstract

Accurate prosody prediction from text leads to more natural-sounding TTS. In this work, we employ a new set of features to predict ToBI pitch accent and phrase boundaries from text. We investigate a wide variety of text-based features, including many new syntactic features, several types of word embeddings, co-reference features, LIWC features, and specificity information. We focus our work on the Boston Radio News Corpus, a ToBI-labeled corpus of relatively clean news broadcasts, but also test our classifiers on Audix, a smaller corpus of read news, and on the Columbia Games Corpus, a corpus of conversational speech, in order to test the applicability of our model in cross-corpus settings. Our results show strong performance on both tasks, as well as some promising results for cross-corpus applications of our models.

Index Terms: prosody, text-to-speech, ToBI, syntax, word embeddings

1. Introduction

Generating accurate prosody can lead to more natural text-to-speech synthesis. However, most corpora used to train TTS systems are unfortunately **not** prosodically labeled. In order to generate accurate TTS output for novel sentences encountered by a TTS system, it is necessary to predict prosodic information from text. The problem of predicting prosody from text has been studied for over three decades. However, earlier work primarily focused on fairly simple feature sets, such as word position in sentence (distance from beginning and end) and sentence length, part of speech tags, punctuation, and word repetition. This is likely due to the difficulty in extracting more complex features at the time. While more recent work does make use of more complex features, it has often focused on smaller feature sets or on only one type of feature, such as incorporating syntactic features or only word embeddings.

Today, we have tools that can quickly and accurately extract a wide variety of linguistic features from text. In this paper we present results of work incorporating a number of these features, including syntactic features, word embeddings, co-reference information, LIWC features, and specificity information in order to improve prosody assignment from text. We focus our work on the Boston Radio News Corpus (BURNC), a corpus of relatively clean radio news speech, that has been labeled using the ToBI (Tones and Break Indices) labeling system [1, 2, 3]. We additionally test our models on Audix, a smaller ToBI-labeled corpus of news speech, and the Columbia Games Corpus, a corpus of conversational speech created to investigate the relationship between prosody and co-reference, in order to examine the performance of our model in cross-corpus and cross-domain settings.

We focus on two tasks: predicting *pitch accents*, words that are produced with more intonational prominence than others, and *phrase boundaries*, pauses or intonational breaks between words. We treat both tasks as binary classification tasks on the text in our corpora, predicting whether a given word is accented or not and whether or not a word is followed by a phrase boundary. To do this, we treat all '*' tone labels for Standard American English ToBI (such as H*, L* or !H* accents) as belonging to the positive *'pitch accent'* category, and all '4' and '4-' break labels in the ToBI phrase labeling scheme as belonging to the positive *'phrase boundary'* category.

2. Previous Work

The problem of predicting prosody from text has been studied for over three decades. Most of the earliest research used simple features that are relatively easy to extract from text, such as part of speech, a word's position in a sentence or paragraph, punctuation, and whether a word has previously appeared in the text [4, 5, 6]. Somewhat later work incorporated more complex syntactic features, showing that syntactic *supertags* and information about syntactic constituency improve prosody prediction [7, 8].

Much of the more recent work on prosody prediction has focused on incorporating a more complex set of syntactic features into the earlier feature sets. Ingulfsen performed phrase boundary detection on BURNC and examined a large set of syntactic features, showing that shallow features pulled from constituency and dependency parses outperformed more complex forms of syntactic chunking [9]. Chen et al. similarly used syntactic features to perform both phrase boundary and pitch accent detection, again on the BURNC data, using a neural model that incorporated parts of speech as well as information about words at the boundaries of syntactic constituents [10]. Tepperman and Nava also experimented on BURNC, showing that a model built using parse tree transducers outperforms a simple n-gram based model for both phrase boundary and pitch accent detection [11]. More recently, in 2015, Mishra et al. demonstrated that a non-lexicalized model built of part of speech tags and dependency features performs comparably to a lexical model containing the same features on phrase boundary prediction, making it more applicable in cross-corpus situations [12]. In the same year, Obin and Lachantin incorporated a rich set of syntactic features, pulled from syntactic and constituency parses as well as tree-adjoining grammar parses, showing that these were useful for predicting both breaks and accents in read and spontaneous speech [13].

10.21437/SSW.2019-48

More recent work has focused instead on using word embeddings for prosody prediction. Rendel et al. examined the use of a number of pre-trained embeddings in prosody prediction, showing that using GLoVe pre-trained embeddings and a continuous bag-of-word model trained on Google news data could provide sizeable improvement in pitch accent prediction and a slight improvement in phrase boundary prediction [14]. Similarly, Klimkov et al. used pre-trained word2vec embeddings along with part of speech and dependency parse features to perform phrase break detection on a corpus of audiobook speech [15].

In research on cross-corpus evaluation of prosody prediction, Rosenberg et al. investigated phrase boundary prediction and detection across five different corpora from different domains, including both BURNC and the Columbia Games corpus, showing not surprisingly that the performance drops in cross-corpus applications, particularly on the spontaneous speech of the Games corpus [16]. More recently, Rendel et al. trained a model on a corpus of professionally recorded speech specifically designed for building a TTS system and tested it on BURNC, showing that precision remained similar but that recall dropped significantly in the cross-corpus scenario [17].

In the work we present here we employ a new set of features — new syntactic features, several types of word embeddings, co-reference features, LIWC features, and specificity information — along with features known to be useful to classify pitch accent and phrase boundary prediction on the BURNC corpus. Many of these features have never been used before, and many of those that have have not been used in conjunction. We then test models trained on BURNC on the Audix Corpus of read radio news to examine cross-corpus prediction in a similar genre. We also test on a larger version of the Columbia Games Corpus to explore cross-corpus prediction on a different genre, spontaneous conversation.

3. ToBI Labeling for Standard American English

The labeling scheme our corpora were annotated in is the ToBI (Tones and Break Indices) system [2, 3], developed in the 1990s by a large number of linguists and computational linguists to enable the sharing of prosodically labeled data across multiple labs. The ToBI labeling scheme consists of four *tiers*: an *orthographic tier* for time-aligned transcripts; a *tone tier* for pitch accents and phrase accents; a *break index tier* where degrees of juncture between words are marked from 0 (no break) to 4 (an intonational phrase break); and a -miscellaneous tier in which other phenomena such as self-repairs, laughter, and filled pauses may be marked if desired. The ToBI system for Standard American English defines five pitch accent types marking simple H* and L* accents as well as complex tones that combine the two in different ways. It also defines two types of phrasal accents: intermediate phrases that are associated with a level 3 break index and intonational phrases that are associated with a level 4 break, usually accompanied by some degree of pause. The AuToBI system [18] was developed to classify these phenomena automatically by training on prosodically labeled and transcribed speech. In our work on prosody prediction from text, we collapse pitch accent types into a binary decision: accented or *deaccented* (not accented) and phrase breaks so level 4 or 4- (slightly lesser boundary) or other. Lower level boundaries are very difficult to annotate or to produce in a TTS system.

4. Corpora

4.1. BURNC

The major focus of the work we present here has been done on the BURNC (Boston University Radio News Corpus) [1] data of read news broadcast speech. BURNC was compiled from over seven hours of professional read radio newscast speech by Mari Ostendorf, Patti Price and Stephanie Shattuck-Hufnagel. Their primary objective in creating BURNC was to generate clean speech data conducive to prosody research. Speakers consist of three female and four male professional radio news announcers. Due to the high quality of speech, the corpus contains very few disfluencies or prosodic "irregularities". Orthographic transcription of the data was performed by hand, and part of speech labels were generated automatically and then hand-corrected.

A portion of the corpus was manually labeled with ToBI labels, including some data from each of the female speakers and two of the male speakers. We have labeled the remainder of the corpus using an AuToBI model [18], a tool that automatically generates prosodic labels from audio, trained on the prosodically labeled portions of BURNC, but we found that including these data slightly decreased the performance of our classifiers and therefore ultimately included only the hand-labeled data in our experiments. This is likely due not only to some inaccuracies in the AuToBI output, but also to the nature of the prosodically unlabeled portion of the corpus, which contains some unscripted interview questions; these are more likely to contain disfluencies.

4.2. Columbia Games Corpus

To examine how our classifier performs in cross-domain contexts, we also examined the Columbia Games Corpus, a corpus of spontaneous conversational speech created to examine the relationship between the *givenness* of a mentioned item and the prosody the item was produced with. It was hypothesized that the *deaccenting* of items which had been previously mentioned was related to the distance of the item from its previous mention, the number of times the item had been previously mentioned, the syntactic function (part of speech) of the item, and the number of "given" items in the current utterance. The Columbia Games Corpus consists of a collection of twelve spontaneous task-oriented dialogues from six male speakers and seven female speakers [19]. The subjects played two computer games in pairs that required verbal communication to achieve joint goals, one that required matching cards containing pictures of multiple objects on their screens (which were not visible to their partner) and another that required describing an object's locations on the describer's screen so that the listener could place the same object on their screen in exactly the same location (again, each subjects' screens were not visible to the partner). For both games, a different set of objects with varying sizes and colors appeared on each player's screen; successful completion of the games required players to describe these objects in the Cards games and to describe the location of objects in the Objects games. Subjects received points for each successfully completed subtask, and they were paid additional money for the earned points.

The entire corpus was manually transcribed and time-aligned. All of the Objects portion of the corpus and roughly one third of the Cards portion, for a total of slightly over 5 hours, were manually labeled with ToBI labels. The remainder of the corpus was labeled using an AuToBI trained on the manually labeled portion. The entire corpus was used in our

experiments. Note that this has resulted in a larger corpus than previously used in prosody prediction research but also one with some automatic prosodic labels as well as the manual labels.

4.3. Audix

The Audix corpus consists of ten news stories recorded by a female professional newscaster in laboratory conditions and comprises approximately thirty minutes of speech. The stories were selected from the AP newswire and were produced largely as written, with few disfluencies. The corpus is introduced in [6], which used classification and regression trees (CART) to detect pitch accents. We use six of these ten stories (as the remaining four did not have usable ToBI lablels). These six represent 17 minutes of speech and 2833 total words. As this is the smallest of our corpora, we primarily use it to evaluate models trained on other corpora.

5. Classification Model

Our models for prosody prediction were Random Forest estimators, each of which fits 200 decision tree classifiers on subsets of the data. (Due to the relatively small size of our corpora, we chose not to use deep learning models.) To assess the performance of the predictive models in our experiments, both within- and cross-corpus, it would not be satisfactory to use a regular leave-one-speaker-out cross-validation scheme, because the three corpora have different numbers of speakers and varying amounts of speech per speaker. Thus, to make these comparisons as fair as possible, we select one speaker each from BURNC and Games (speakers f3a and 101, respectively) to serve as the test sets. For BURNC, we use all remaining speakers as the training set. For Games, in order to avoid any potential effects of *entrainment* (the tendency of conversational partners to speak like one another), we also exclude all data from those sessions in which speaker 101 participated, and we use the data from the 10 remaining sessions as our training data.

The set of features we included in our models are presented below. All of these features were tested first on the BURNC corpus. In our cross-corpus experiments, we excluded all word embedding features, as these were created specifically from the vocabulary of BURNC. Additionally, as there was no punctuation or proper names present in Games, punctuation and named entity features were excluded from the cross-corpus experiments that included Games.

5.1. Positional Features

For all corpora, the length of the current sentence and the word's position in the sentence were included as features. In the BURNC and Audix corpora, punctuation was already present in the transcripts. In the Games corpus, as there was no punctuation present in the manual transcripts, these features were extracted using a rule-based sentence segmentation script, which assigned sentence breaks based on the length of silence between words, as well as the parts of speech occurring before and after a potential sentence break.

5.2. Syntactic Features

We used a wide variety of syntactic features in our model. For BURNC, gold standard labels for the current and next word's parts of speech were provided with the corpus. These labels were initially automatically generated and then were hand corrected, so they are fairly similar to the output of automatic taggers like the Stanford CoreNLP tagger but are presumably more accurate. For the Games and Audix corpora, these features were extracted using Stanford CoreNLP's part of speech tagger.

The Stanford parser was used to obtain dependency and syntactic parses for all corpora. From the dependency parse, we obtained each word's syntactic function, the label of the dependency relationship that has the given word as the head. The current and next words' syntactic functions were included as features.

The depth and width of the syntactic parse trees were included as features, as well as the depth of the current word in the tree. We also included the width and depth of the smallest constituent containing the current word, as well as this constituent's root label and the position of the current word with in the constituent. Finally, we included the width, depth, and root label as features and the minimal spanning tree containing the leaf nodes for the current word and the next word, as we believed these features would be helpful in determining the presence of a phrase boundary between the two words.

Lastly, in addition to part of speech tags, we include supertags, which are tags that provide more specific information about a word's syntactic role [20]. Based on the Tree-Adjoining Grammar formalism, supertags consist of a portion of a syntax tree, which can capture information about a word's arguments, as well as its part of speech. For example, a supertag can distinguish between a noun in a subject position and one in object position. Previous work has shown that supertags can improve prosody prediction [7]. So we extracted supertags to use as features with a newer bi-LSTM based supertagger pre-trained on the Wall Street Journal [21].

5.3. Word-Level Features

We included a number of word-level features in our model. Most simply, we included the number of syllables in each word. For BURNC and Audix, we also included the punctuation appearing after each word.

Using Stanford CoreNLP, we ran named entity recognition (NER) on both BURNC and Audix. (Note that the Games corpus did not include any named entities.) We included the NER tags of the current and next words as features.

Lastly, we used Linguistic Inquiry and Word Count (LIWC) [22] dimensions as features. LIWC is a system that uses a dictionary to categorize words into 73 categories pertaining to emotions, thinking styles, social concerns, and parts of speech. For example, the word *trying* belongs to the categories of cognitive processes, drives (e.g. needs and motives), verbs, tentative, and achievement.

5.4. Co-Reference Features

The presence and frequency of a word's previous co-references can have an effect on its prosody. In particular, it is believed that a new word is more likely to be accented than a given word that has previously appeared. In all three corpora, we identified co-reference using Stanford CoreNLP's deterministic CorefAnnotator [23]. Using these groups of co-references, we then extracted a set of co-reference features, specifically the number of previous mentions of the current word, the distance between the current word and its most recent mention, the part of speech of the most recent mention, and the syntactic function of the most recent mention. We also separately included the distance, part of speech, and syntactic function of the most recent explicit (identical) mention and the most recent implicit (non-identical) mention. In cases where co-references were multi-

word phrases, the head word of the phrase was used to generate these features.

5.5. Word Embedding Features

Word embeddings often augment the performance of NLP models by quantifying measures of syntactic and semantic relations in language, and they often can capture elements of syntax and usage that cannot be captured by more transparent linguistic features. Previous studies have shown that word embeddings can be a useful feature in predicting prosody [14, 15]. In our experiments, we looked at both word embeddings and sentence embeddings, which are generated by adding the embeddings of each word in a given sentence. In order to integrate word embeddings into our Random Forest model, we performed clustering on these embeddings using the k-means clustering algorithm [24].

We trained word embeddings on the BURNC corpus directly, as well as experimenting with a variety of pre-trained embeddings. To begin with, we trained 200d embeddings on BURNC directly, using the Word2vec skipgram model with negative sampling [25], with a window size of 4 words. Furthermore, because syntactic structure has a strong relationship with prosody, we also used the dependency parses to generate another set of 200d embeddings from BURNC by using word2vec with a word's lexical dependencies instead of a linear context window. We also used a set of pre-trained 300d dependency-based vectors trained on Wikipedia data by Levy and Goldberg [26].

For pre-trained embeddings, we used a set of embeddings trained from Google news [27], as the Google news dataset is similar in domain to BURNC. We also used a set of pre-trained 300d gender-neutral embeddings based on GloVe embeddings [28], as any gender bias present in our data is unlikely to have an effect on prosody. Finally, we used an algorithm to map words between two models of the same language [29] to adapt pre-trained GLoVe embeddings to BURNC.

In most cases, we assigned each word embedding to one of five clusters and each sentence embedding to one of twenty based on our k-means clustering results. (These values were determined empirically based on model performance.) These clusters were able to capture grammatical and syntactic properties. For example, cluster 1 contained many proper nouns and concrete nouns, cluster 4 contained many modifiers and comparatives, and cluster 5 contained many function words. For each embedding, we included the current word and sentence's embeddings, as well as the embeddings for the sentences and words two before and two after the current ones.

5.6. Speciteller

Speciteller [30] is a tool for determining how specific a given sentence is. Based on the words present in the sentence, it assigns a specificity score ranging from 0 (most general) to 1 (most detailed). Sentences with pronouns and general terms will have lower scores, such as *"Estimates vary widely on how much money could be saved"*, which has a score of 0.0186, whereas sentences with more proper nouns and specific terms, such as *"Quincy based Arbella Mutual Liability can now take over American Mutual's forty thousand car and home owner's policies"*, will have higher scores – in this case 0.872.

Feature Set	Accuracy	F1
Best Model	81.9%	0.844
Best Model without LIWC	80.9%	0.832
Best Model without Embeddings	80.2%	0.826
Best Model without Syntax	79.1%	0.817
Baseline	50.4%	

Table 1: *Performance scores for **pitch accent** detection*

Feature Set	Accuracy	F1
Best Model	93.4%	0.810
Best Model without Punctuation	92.5%	0.774
Best Model without Syllables	92.0%	0.770
Best Model without Syntax	85.5%	0.575
Baseline	82.8%	

Table 2: *Performance scores for **phrase boundary** detection*

6. Results

6.1. BURNC

The results for pitch accent detection are displayed in Table 1. Our best model achieves an accuracy of 81.9%, which is higher than Tepperman and Nava's model, which achieved an accuracy of 76.83% [11] and only slightly below Chen et al.'s accuracy of 82.7% [10]. This indicates that our model is quite strong but still has some room for improvement.

The best model for this task included all syntactic and word-level features, as well as gender-neutral embeddings, coreference features, and Speciteller. Syntactic features, word embeddings, and LIWC features had the strongest effect on the model. In particular, the two most heavily weighted features include the LIWC dimension for function words and the number of syllables in the current word, as function words and short words are less likely to be accented. Other heavily weighted features include the sentence embedding features, the parse tree's width and depth, and the Speciteller score. These results, along with the high performance of syntactic and embedding features on this task, indicate that context is highly important in determining pitch accent, particularly among longer content words. The performance of the gender-neutral embedding features in particular seem to indicate the importance of context, as all seven of the sentence embedding features were weighted among the top 15 individual features in the random forest model, whereas none of the word embedding features were. The fact that the only helpful embeddings for this task were gender-neutral embeddings, which were the set of embeddings least adapted to the news domain but also the only ones that removed semantic bias, also seems to indicate that general context plays a more important role in determining pitch accent than a particular word's semantic content.

The results for phrase boundary detection are displayed in Table 2. Our best model achieves an accuracy of 93.4% and an F1 of 0.810. This is a notable improvement over Rosenberg et al.'s model, which had an accuracy of 90.5% and an F1 of 0.781 [16].

The best model for this task included all syntactic features, all word-level features except for LIWC dimensions, the word embeddings trained directly on BURNC, and Speciteller scores. Syntactic features had by far the biggest impact on the model's

Training Set					
			Games	BURNC	Naive
Test Set	Games	Accuracy	73.5%	70.5%	65.5%
		F1	0.566	0.630	
	BURNC	Accuracy	70.2%	80.9%	52.8%
		F1	0.562	0.776	
	Audix	Accuracy	69.1%	80.2%	50.9%
		F1	0.557	0.766	

Table 3: *Accuracy and F1 scores for cross-corpus evaluation of pitch accent.*

Training Set					
			Games	BURNC	Naive
Test Set	Games	Accuracy	87.4%	84.2%	83.7%
		F1	0.565	0.442	
	BURNC	Accuracy	70.2%	92.6%	52.8%
		F1	0.137	0.783	
	Audix	Accuracy	79.4%	89.8%	78.5%
		F1	0.079	0.732	

Table 4: *Accuracy and F1 scores for cross-corpus evaluation of phrase boundaries.*

performance, followed by the number of syllables in the current word and punctuation following the word. In particular, heavily weighted features included features tied to the constituency parse, including the width of the parse tree, the width of the spanning tree, and the position of the current word in its smallest syntactic constituent, as well as the supertag t3, which consists of nouns appearing at the end of noun phrases. These results are unsurprising, as periods and commas generally indicate phrase boundaries, and phrase boundaries often occur at the end of syntactic constituents. However, the results of combining these features with our new features are most encouraging, as our results are improved by including more semantically oriented features, such as Speciteller, word embeddings, and NER features. (In fact, the Speciteller score is a heavily weighted feature in our trained model.) Furthermore, the performance of supertag features, which alone improve our model's accuracy by over 1%, indicates that more complex syntactic features can be useful in predicting prosody along with more simple constituency information.

6.2. Cross-Corpus Evaluation on the BURNC, Games and Audix Corpora

Tables 3 and 4 show the cross-corpus results for pitch accent and phrase boundary prediction, as well as the results on these corpora when tested with a naive model that predicts the majority class in all cases.

As these results show, the models perform well when tested on new corpora within the same domain, as the results of training on BURNC and testing on Audix are only mildly worse than the results of training and testing on BURNC for both tasks. Additionally, on the pitch accent prediction task, the model trained on BURNC seems to perform adequately in cross-domain applications, as that model has a higher F1 (albeit a lower accuracy) than the model trained and tested on Games and performs noticeably better than the majority class baseline. However, neither model performs particularly well, so it is clear that a different set of features is necessary to achieve strong performance on the Games corpus.

On the other hand, while our models apply well across corpora of the same domain on the phrase boundary prediction task, they perform very poorly in cross-domain situations — in this case when models trained on read speech are tested on conversational data or vice versa. The model trained on BURNC performs only slightly better than the baseline when tested on Games, and the models trained on Games and tested on BURNC and Audix perform extremely poorly, having a recall of less than 0.1. This is most likely because of the presence of punctuation in the news corpora, which is not present in Games. Furthermore, the news data contains longer, more structured sentences, whereas Games contains many more fragments and disfluencies. Therefore, phrases, and consequently phrase boundaries, are very different across the corpora.

7. Conclusions

In this paper, we have presented a text-based model that predicts binary ToBI labels for pitch accent and phrase breaks from text and can be applied across corpora of scripted news data. Our model performs very well when trained and tested on BURNC, particularly on the phrase boundary prediction task, where it noticeably outperforms previous text-based work on the same corpus. Performance on pitch accent prediction is also stronger than or comparable to prior work. It also shows relatively strong performance in cross-corpus applications within the same domain of broadcast news, but it performs less well in cross-domain applications of spontaneous conversation. Improving cross-domain performance by examining the prosodic differences between read and spontaneous speech and determining which features are similar between the two will be a focus of future study.

Additionally, while we have shown that our model successfully predicts ToBI labels, the success of prosody realization within text-to-speech systems can be much more subjective. Future work will focus on incorporating our models into a text-to-speech system to determine if they do in fact improve naturalness. Furthermore, by focusing on news data, and even within the Games corpus, this work focuses primarily on neutral speech presenting factual information. However, in many new text-to-speech applications, such as voice assistants, there is a need for producing not just conversational but also emotional speech, which involves major differences in prosody. Future work will extend our models into these domains.

8. Acknowledgements

This work was supported by the National Science Foundation under Grants IIS 1548092 and 1717680. The research was also supported by CONICET, ANPCYT PICT 2014-1561, and the Air Force Office of Scientific Research under award no. FA9550-18-1-0026.

9. References

[1] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[2] M. E. Beckman and J. Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[3] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A stan-

dard for labeling English prosody," in *Second International Conference on Spoken Language Processing*, 1992.

[4] J. Hirschberg and P. Prieto, "Training intonational phrasing rules automatically for English and Spanish text-to-speech," *Speech Communication*, vol. 18, no. 3, pp. 281–290, 1996.

[5] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

[6] J. Hirschberg, "Pitch accent in context predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, no. 1-2, pp. 305–340, 1993.

[7] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[8] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, "Improving intonational phrasing with syntactic information," in *2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2000, pp. 1289–1290.

[9] T. Ingulfsen, "Influence of syntax on prosodic boundary prediction," University of Cambridge, Computer Laboratory, Tech. Rep., 2004.

[10] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic-prosodic model," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–509.

[11] J. Tepperman and E. Nava, "Where should pitch accents and phrase breaks go? A syntax tree transducer solution," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[12] T. Mishra, Y.-j. Kim, and S. Bangalore, "Intonational phrase break prediction for text-to-speech synthesis using dependency relations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4919–4923.

[13] N. Obin and P. Lanchantin, "Symbolic modeling of prosody: from linguistics to statistics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 588–599, 2015.

[14] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5655–5659.

[15] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading tts: exploiting text structure information," in *Proceedings of Interspeech 2017*, 2017, pp. 1064–1068.

[16] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Phrase boundary assignment from text in multiple domains," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[17] A. Rendel, R. Fernandez, Z. Kons, A. Rosenberg, R. Hoory, and B. Ramabhadran, "Weakly-supervised phrase assignment from text in a speech-synthesis system using noisy labels," in *Proceedings of Interspeech 2017*, 2017, pp. 759–763.

[18] A. Rosenberg, "AuToBI – A tool for automatic ToBI annotation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[19] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[20] A. K. Joshi and B. Srinivas, "Disambiguation of super parts of speech (or supertags): Almost parsing," in *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, ser. COLING '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 154–160.

[21] J. Kasai, B. Frank, T. McCoy, O. Rambow, and A. Nasr, "Tag parsing with neural networks and vector representations of supertags," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1712–1722.

[22] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," The University of Texas at Austin, Tech. Rep., 2015.

[23] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[24] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[26] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2014, pp. 302–308.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[28] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *arXiv preprint arXiv:1809.01496*, 2018.

[29] S. S. Akhtar, A. Gupta, A. Vajpayee, A. Srivastava, M. G. Jhawar, and M. Shrivastava, "An unsupervised approach for mapping between vector spaces," *arXiv preprint arXiv:1711.05680*, 2017.

[30] J. J. Li and A. Nenkova, "Fast and accurate prediction of sentence specificity." in *AAAI*, 2015, pp. 2281–2287.