INTONATIONAL PHRASE BREAK PREDICTION FOR TEXT-TO-SPEECH SYNTHESIS USING DEPENDENCY RELATIONS

Taniya Mishra, Yeon-jun Kim and Srinivas Bangalore

Interactions, 31 Hayward Street, Suite E, Franklin, MA 02038 *
[tmishra,ykim,sbangalore]@interactions.net

ABSTRACT

Intonational phrase (IP) break prediction is an important aspect of front-end analysis in a text-to-speech system. Standard approaches for intonational phrase break prediction rely on the use of linguistic rules or more recently, lexicalized data-driven models. Linguistic rules are not robust while data-driven models based on lexical identity do not generalize across domains. To overcome these challenges, in this paper, we explore the use of syntactic features to predict intonational phrase breaks. On a test set of over 40 thousand words, while a lexically driven IP break prediction model yields an F-score of 0.82, a non-lexicalized model that uses part-of-speech tags and dependency relations achieves an F-score of 0.81 with added feature of being more portable across domains. In this work, we also examine the effect of contextual information on prediction performance. Our evaluation shows that using a three-token left context in a POS-tag based model results in only a 2% drop in recall compared to a model that uses both a left and right context, which suggests the viability of using such a model for incremental text-to-speech system.

Index Terms— Intonational phrase, phrase breaks, IP prediction, prosody, text-analysis

1. INTRODUCTION

Intelligibility of a text-to-speech (TTS) system is highly correlated with the intonational phrasing inferred from the input text. We define an intonational phrase (IP) as a speech segment that spans a single prosodic contour, bounded by pauses on the edges, either for breathing or to separate information units in production. The pauses between such intonational phrases are called *Intonational Phrase Breaks*. As illustrated in the example from [1], *Bill does not drink because he is unhappy* has two distinct interpretations based on whether or not there is a phrase break between *drink* and *because*.

There have been many attempts to characterize the location of these breaks through rules, patterns and even lexicalized data-driven models. While the data-driven models overcome the brittleness of rule-based models, their lexical dependence encapsulate the nuances of the domain and hence they perform optimally only on the domain texts the IP models were trained on. Our goal in this paper is to explore syntactically based data-driven models for predicting such intonational phrase breaks from the input text and compare its performance to a lexically based model. The rationale for our approach is to develop a domain independent technique to intonational phrasing that overcomes the limitation of a heavily lexicalized model and explore the impact on accuracy when lexical information is not used in predicting IP boundaries.

Furthermore, we explore the possibility of predicting the phrase boundaries using only features from the left context of a word with the aim of developing a strictly left-to-right incremental TTS system. Incremental TTS systems [2] are meaningful for applications such as simultaneous interpretation where arbitrarily long text streams are to be synthesized at low latencies. The model we present in this paper demonstrates that incremental phrase boundary prediction is viable and is only marginally less accurate when compared to a model that uses both left and right context.

The rest of the paper is organized as follows. In Section 2, we establish connections to prior work. In Section 3.1, the data is described, in Section 3.2, the features used for prediction of intonational phrase boundaries in our text-to-speech system are presented. The model evaluations are presented and discussed in Section 4. Finally, in Section 5, conclusions are drawn based on our evaluations.

2. RELATION TO PRIOR WORK

Intonational phrase structure is a well studied problem in linguistics by phoneticians, psycholinguists and syntacticians alike [3–6]. The results of such explorations has been applied to text-to-speech synthesis in many instances [7–9]. It has been widely discussed as to how intonational phrase breaks correlate both with syntax (for example, conjunctions are either immediately followed or preceded by an IP break [10] and semantics (breaks between nouns are atypical, but when nouns occur in a long semantic list structure, prosodic breaks are more likely to exist [11]). There have been computational models to predict the intonational phrase breaks as well. These models rely on a classification paradigm using a variety of features including lexical features [1], while sub-

*Work done while at AT&T Research

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

sequent work has attempted to extend the features to include syntactic features as well [12, 13]. Due to their reliance on lexical features most of these approaches achieve optimal performance only on the domains that the model is trained on. We attempt to discern the impact of lexical features by building models that rely on syntactic and semantic features, with the expectation that such models are less likely to be affected by domain variation than the lexicalized models.

3. DATA AND FEATURES

3.1. Data

To build the intonational phrase break prediction models, the target output labels were ”learned” from a speech database consisted of approximately 50 hours (44,000 individual utterances) of speech obtained from a female speaker of American English. The speaker read a variety of textual material with good phonemic and prosodic coverage, a majority of which were presented as isolated sentences. The audio was recorded 16kHz, 16-bit in a studio environment.

For the recorded material, the speaker had a mean pitch of 219 Hz and standard deviation of 62 Hz. The speech material was labeled automatically with word, syllable and phoneme boundaries and silence information based on forced alignment of text and audio. There were two distinct clusters of silences identified in this database: first, long silences with an average duration of 193.63 ms, and second, very short silences with an average duration of 67 ms. On manual inspection, the very short silences were found to be related to physiological phenomena such as glottal stops between vowels, and the long silences to true intonational phrase breaks. A distribution of the silences identified in this speech database is shown in Figure 1.

For this project, in each read aloud sentence, the very short silences were thresolded out on the basis of the silence duration. Only the longer silences, assumed to indicate intonational phrase breaks were considered for developing the IP break prediction models. While a full manual inspection of the speech database was not performed, the assumption that the longer silences indicated intonational phrase breaks is reasonable considering that the database consisted of read speech rendered by a professional voice-talent where there would be few (or none) pauses for sentence-planning or disfluencies or hesitations. At the token level (i.e., word or punctuation), roughly 18% of the tokens were identied to immediately precede an intonational phrase break. The data was also automatically annotated with part of speech tags and dependency relations. No manual labeling was performed on the database for our experiments.

3.2. Features

Our goal was to develop data-driven models to predict where intonational phrase boundaries should occur in any given text input to a text-to-speech system during synthesis. In our approach, the prediction relies on lexical, syntactic and semantic features extracted from every token (word or punctuation) in the input text with a *n*-token left and right context. A detailed description of these features are presented below:

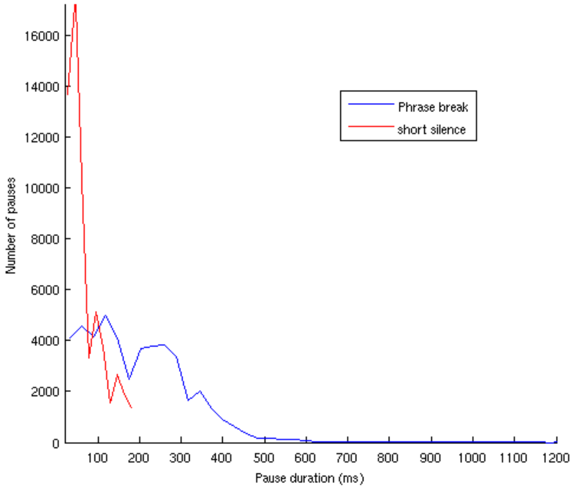

Fig. 1. *Distribution of silences in speech data*

- WI: Lemmatized word identiity. In our final IP prediction model, we do not want to use word identity as a feature because this feature does not generalize to unseen data. We have however used word identity as a feature for experimental comparison. To reduce sparsity, we lemmatized the word. *Lemmatization* is the process of replacing the inflectional and variant forms of a word by its *lemma* or base form. For example, “cats”, “catty”, and “cat-like” are each reduced to their lemma “cat” by lemmatization.

- POS: Part-of-speech tags. The part-of-speech tags were automatically generated using the Penn Treebank tagset using a discriminatively trained tagger [14].

- PYTPE: Punctuation type. If the token under consideration is a punctuation, it’s type is broadly classed as sentence-final (such as periods, colons, question marks) or sentence-medial (such as commas, semi-colons, and parenthetic dashes). All other punctuation types (such as quotation marks or parentheses) are each their own class because they may occur in either sentence initial, final or medial positions — it depends on context. This feature was extracted based on the observation that a speaker’s IP boundaries often align with sentence-final and sentence-medial punctuation in read speech.

- PDIS: Distance from punctuations. It is the distance of the current token from the last (and next) seen punctuation is computed in the number of syllables.

The next four features are dependency-relation-based features that were computed for every token w_i in the given text. The dependency relationships were automatically generated with a shift-reduce based dependency parser similar to [15]. The dependency relation label definitions are as defined in [16].

- D1: ishead(w_i) = {True, False} : Computes whether the word under consideration is a *head* in a dependency relation.
- D2: num_dependents(w_i) = {0, 1, 2, ...}: Number of dependents a head word has in the dependency parse tree. It is 0 if the token is not a head.
- D3: dep_rels(w_i) = {rel1_rel2_rel3_...}: It is the concatenated list of dependency relationships that the head word has.
- D4: max_dist_to_dependent(w_i) = {0, 1, 2, ...}: The maximum distance between a head word and its dependents in the dependency parse tree. The distance is computed in number of words.

4. MODEL EVALUATION

We modeled intonational phrase break prediction as a binary classification task, i.e., each token (word or punctuation) in a given text is classified as having an IP break immediately following it (1) or not (0). Five IP break prediction models were built using different subsets of the features outlined in Section 3.2. All models were trained using a logistic regression based binary classifier from the the LLAMA machine learning toolkit [17]. For every model, 90% percent of the data described in Section 3.1 was used for training while 10% was held out as test set. The test and training samples were randomly selected without replacement. For each token, w_i, each of these features was computed over a window of three tokens to the left and to the right: w_{i-j}, w_i, w_{i+j}; $j = \{1, 2, 3\}$. Only true words comprised the w_i token; no punctuation tokens were included; however, punctuation tokens may be present in the left and right context.

We also modeled a second set of intonational phrase break prediction models that only considered a left context window. These models were developed with an eye towards detecting intonational phrase breaks for incremental TTS. Such models — at least for English — cannot use the dependency relation based features that we outlined in Section 3.2) because most dependency trees have the head preceding its dependent(s), although some head-final dependencies do occur.[1] For these models, we only considered POS tags and the Word Identity features in model building, and each of these features was computed over a token w_i, and three tokens preceding it: w_{i-j}, w_i, $j = \{1, 2, 3\}$.

Each of the aforementioned models was trained and tested on the same test and training set split of the data described in Section 3.1.) The token-level classification performance of each model on our test set of 46221 samples is presented in Table 1. In the test set, 84.7% of the samples do not have an immediately following IP break (target-label=0) and 15.3% of the samples have an immediately following IP break (target-label = 1). Thus, the baseline accuracy for this set, obtained by assigning the majority class (target-label=0) to all test examples is 0.847. The F-score of this baseline model is of course 0 since no true positives are recognized by such a system. This is considered the baseline system and all relative improvements are computed compared to this baseline.

Model	Accuracy	Precision	Recall	F-score
All features	0.947	0.858	0.787	0.821
Word Identity	0.948	0.869	0.779	0.822
POS Tags	0.945	0.867	0.753	0.806
Dep. relations	0.937	0.845	0.723	0.779
All w/out WI	0.946	0.859	0.772	0.813
WI (only left)	0.946	0.872	0.760	0.813
POS (only left)	0.94	0.860	0.740	0.795

Table 1. *Results of IP break prediction models.*

We also plotted the change in F-score (computed on the test set) as we systematically changed the size of the context window from 0 to (+/-) 5 tokens, such that the left and right context size of a given token w_i were symmetric, for three models, the model trained on all features, the model trained only on word identity features, and the model trained on only POS tags as features. The plot is shown in the top panel of Figure 2. The change in F-score is primarily due to a change in recall, shown in the lower panel of the same figure. The change in precision across the context-window change was slight — mean precision of 0.87 with a standard deviation of 0.01 across the varying context window sizes. The same training and test splits were maintained across all models and all context window sizes.

4.1. Discussion

Through our evaluations, we wanted to answer three questions: Which features are most predictive of intonational phrase breaks? How much does contextual information help in the prediction of intonational phrase breaks? Finally, if we want to avoid using word-identity information to build IP break prediction models, can syntactic and semantic features compensate for the loss in predictive power?

To answer the question regarding the importance of context to IP break prediction, consider the plot shown in Figure 2. The plot indicates that the effect of context on in-

[1] In future work, we plan to explore an approach where the dependency parser and the phrase boundary prediction models are tightly coupled.

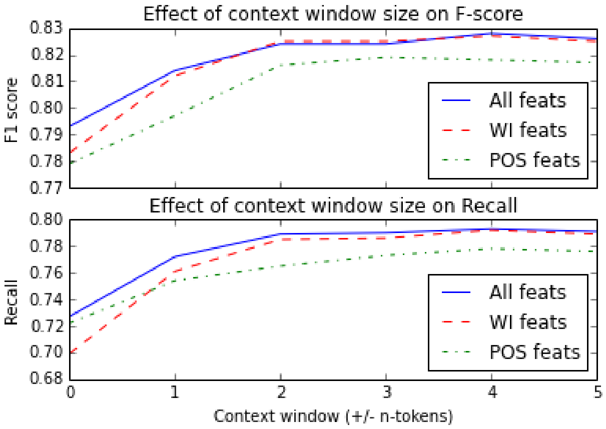

Fig. 2. *Effect of context window size*

tonational phrase break prediction is not particularly long-distance. Increasing the size of the context window increases the performance of all three models under investigation, but the F-score appears to plateau at 3 preceding and succeeding tokens; adding more context insignificantly changes performance. A likely explanation for these results is that an intonational phrase is roughly 5 to 6 words long as seen by analysis of the Boston University Radio News corpus in [18]. Based on this analysis, we used a context window of 3 preceding and 3 succeeding tokens for developing the seven models for feature comparison, whose performance evaluations are outlined in Table 1.

The first five rows of Table 1 show the evaluation results of models built with a 3-token left and right context window surrounding the token under consideration. The last two rows show the results of models built with only a 3-token left context window. These results indicate that that all subsets of features under investigation are predictive of intonational phrase breaks. The mean prediction accuracy across the seven models is approximately 0.94, which is significantly higher than the 0.847 baseline accuracy of the test set (baseline computed by assigning majority class to all test samples).

Word Identity is a more predictive feature compared to POS-tags or dependency relation based features; the F-score for the Word Identity based model is 0.82 while it is 0.80 for the POS-tag-based model and 0.77 for the dependency-relation based model, shown in rows 2, 3, and 4 of Table 1, respectively. Row 5 of Table 1, which shows the evaluation results of a model that use POS tags, punctuation and dependency relation based features, indicates that the predictive power lost when Word Identity is not in the model can be substantially compensated for by the use of features based on POS-tags, dependency relations and punctuation information. The implication of this result is as follows. Given that Word Identity is not a generalizable feature across multiple domains with varying vocabulary, the complementary use of syntactic and semantic features that can offer similar discriminative power is extremely attractive. We can either not use Word Identity as a feature in supervised domain-independent intonational phrase break models, or assign Word Identity a lower weight. We can further increase the precision of this non-lexically derived model by post classification threholding without a large loss in recall. Observe the plot in Figure 3, obtained by systematically changing the classification threshold of the non-lexical model defined in row 5 of Table 1. The precision and recall values computed on the test set shows that we can increase precision to 0.9, with recall still well over 0.7. Considering rows 6 and 7 in light of the results presented in the previous rows of Table 1, we see that the loss of right context causes approximately a 2% drop in recall values and none in precision. This suggests that such data-driven models for intonation phrase prediction could viably be used for incremental TTS systems.

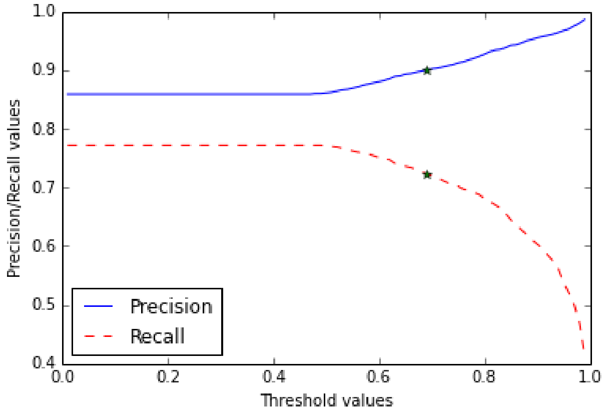

Fig. 3. *Post-classification thresholding*

5. CONCLUSIONS

Accurate prediction of phrase boundaries is imperative for naturalness and intelligibility of a text-to-speech system. While the presence of a phrase boundary depends on a variety of lexical, syntactic, semantic and pragmatic factors, in this paper, we present an approach that relies on local context to predict phrase boundaries at an F-score of 0.82. We demonstrate that the prediction problem can be modeled with features that are non-lexical with minimal loss in accuracy but with the potential of being more portable across domains. Furthermore, we show that a model relying only on left context features of a word is almost as accurate as a model that uses both left and right context, suggesting that phrase boundary prediction can be done incrementally for an incremental TTS system.

6. REFERENCES

[1] J. Hirschberg and P. Prieto, "Training intonation phrasing rules for english and spanish," *Speech Communication*, 1996.

[2] T. Baumann and D. Schlangen, "Evaluating prosodic processing for incremental speech synthesis," in *Proceedings of Interspeech*, Portland, USA, Sept. 2012, ISCA.

[3] D. Bolinger, *Intonation and its uses: Melody in Grammar and Discourse*, Edward Arnold, London, UK, 1989.

[4] E. Selkirk, "The Syntax-Phonology Interface," in *Handbook of phonological theory*, J. Goldsmith, J. Riggle, and A. Yu, Eds. Blackwell Publishing, 2011.

[5] C. Brierley, *Prosody resources and symbolic prosodic features for automated phrase break prediction*, Ph.D. thesis, University of Leeds, Leeds, UK, 2011.

[6] D. Watson and E. Gibson, "The relationship between intonational phrasing and syntactic structure in language production," *Language and Cognitive Processes*, vol. 19(6), pp. 713–755, 2004.

[7] J. Apel, F. Neubarth, H. Pirker, and H. Trost, "Have a break! modelling pauses in german speech," in *Konferenz zur Verarbeitung natrlicher Sprache (Konvens)*, Vienna, Austria, 2004.

[8] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion," in *Interspeech*, Florence, Italy, Aug. 2011.

[9] N. Braunschweiler and L. Chen, "Automatic detection of inhalation breath pauses for improved pause modelling in hmm-tts," in *Speech Synthesis Workshop*, Barcelona, Spain, 2013.

[10] A. Wennerstrom, *The Music of Everyday Speech*, Oxford University Press, 2001.

[11] A. W. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," in *Proceedings of Eurospeech 97*, Rhodes, Greece, 1997, pp. 995–998.

[12] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, "Improving intonational phrasing with syntactic information," in *Proceedings of ICASSP*, 2000.

[13] V. Keri, S. C. Pammi, and K. Prahallad, "Pause prediction from lexical and syntax information," in *International Conference on Natural Language Processing*, Hyderabad, India, 2007, pp. 45–49.

[14] S. Bangalore and P. Haffner, "Classification of large label sets," in *Proceedings of the Snowbird Learning Workshop*, 2005.

[15] J. Hall, J. Nivre, and J. Nilsson, "Discriminative classifiers for deterministic dependency parsing," in *Proceedings of the COLING/ACL on Main conference poster sessions*, Morristown, NJ, USA, 2006, pp. 316–323, Association for Computational Linguistics.

[16] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tackstrom, C. Bedini, N. Bertomeu Castello, and Jungmee Lee, "Universal dependency annotation for multilingual parsing," in *Proceedings of ACL*, 2013.

[17] P. Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.

[18] A. Rosenberg, *Automatic Detection and Classification of Prosodic Events*, Ph.D. thesis, Columbia University, New York, NY, 2009.