INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Weakly-Supervised Phrase Assignment from Text in a Speech-Synthesis System Using Noisy Labels

Asaf Rendel [a], Raul Fernandez [1], Zvi Kons [2], Andrew Rosenberg [1], Ron Hoory [2], Bhuvana Ramabhadran [1]

[1]IBM TJ Watson Research Center, Yorktown Heights, NY – USA
[2]IBM Haifa Research Lab, Haifa – Israel

fernanra@us.ibm.com, zvi@il,ibm.com, amrosenb@us.ibm.com

Abstract

The proper segmentation of an input text string into meaningful intonational phrase units is a fundamental task in the text-processing component of a text-to-speech (TTS) system that generates intelligible and natural synthesis. In this work we look at the creation of a symbolic, phrase-assignment model within the front end (FE) of a North American English TTS system when high-quality labels for supervised learning are unavailable and/or potentially mismatched to the target corpus and domain. We explore a labeling scheme that merges heuristics derived from (i) automatic high-quality phonetic alignments, (ii) linguistic rules, and (iii) a legacy acoustic phrase-labeling system to arrive at a ground truth that can be used to train a bidirectional recurrent neural network model. We evaluate the performance of this model in terms of objective metrics describing categorical phrase assignment within the FE proper, as well as on the effect that these intermediate labels carry onto the TTS back end for the task of continuous prosody prediction (i.e., intonation and duration contours, and pausing). For this second task, we rely on subjective listening tests and demonstrate that the proposed system significantly outperforms a linguistic rules-based baseline for two different synthetic voices.

Index Terms: speech synthesis, phrasing, prosody modeling, recurrent neural networks

1. Introduction

Intonational phrasing is used in spoken language to decompose utterances into meaningful sub-units that convey syntactic, semantic, and other discursive functions. In the case of synthetic speech, the proper segmentation of an input text string into such units, and their corresponding acoustic realization, is of importance to produce speech that is natural, and aids the listener's comprehension by signaling the structure of the text. The FE of a TTS system is therefore often tasked with assigning such phrase structure from text, and with making that information available to the back-end to exploit in the creation of prosody models. The development of such a phrasing system can rely on either linguistic knowledge, or infer phrasing rules directly in a data-driven way. Since resources bearing prosodic annotations for supervised learning are few, and their development costly, we seek to exploit a methodology that relies on already-existing resources and, without the need for creating brand-new hand-labeled annotations, augments acoustic unlabeled corpora with (potentially noisy) labels for training a data-driven system. We accomplish this by merging a set of heuristics to arrive at a ground truth for training (Sec. 2) and present a modeling recurrent neural-network approach that focuses on exploiting lightweight or learnable features (Sec. 3). The phrase predictions thus obtained are shown to make an impact in the back end of the TTS system by producing output that is judged to have better phrasing in a perceptual listening test (Sec. 5).

2. Resources and Ground Truth Labeling

To create a training resource for supervised learning in this work, we rely on a phonetically-aligned corpus (including silences) of approximately 150K words recorded for use in a unit-selection TTS system by a female professional native speaker of North American English. The phonetic transcriptions were obtained by forced alignment against the transcripts using 3-state-per-phone hidden Markov models (HMM). The audio recordings were additionally analyzed using the AuToBi tool to automatically label break indices (using the conventions of the ToBi framework) with pre-trained, freely-available models that had been derived from independent data sources (specifically, the Boston University Radio Corpus [1], and the Boston Directions Corpus [2]). Details of this background work may be found in [3] and [4]. The pre-trained model used in this work is version 1.5 released with the AuToBi tool and available from [5].

Modeling phrase boundaries was treated as a word-level binary classification task, where a word was labeled as preceding a boundary whenever any of the following criteria was met:

1. the word is followed by a silence longer than 125ms in the phonetic transcripts,
2. the word is followed by a silence of 80ms which coincides with a phrase break, as predicted by linguistic rules in the text-analysis FE,
3. the word is followed by a Break Index of 4, as predicted by the AuToBi tool, and the word simultaneously exhibits substantial phrase-final lengthening $phr_l > 175ms$ (where the details of the phr_l calculation are described below).

Criteria 1 and 2 are purely unsupervised criteria that appeal, respectively, to the co-occurrence of major intonational phrase boundaries with acoustic pauses (#1) and/or with the syntactical constituency that the rules-based phrase predictor is exploiting (#2). The third criterion seeks to compound evidence from a data-independent prediction of a major intonational boundary concomitantly with evidence of the well-studied and -documented phrase-final lengthening effect [6]. Given the data mismatch, a Break Index of 4 is likely to be less accurate on this target corpus than on the data sources it was trained on [4].

The lengthening heuristic phr_l in Criterion 3 seeks to strengthen that evidence via the following metric. Given

[a] This paper is dedicated to the memory of our colleague Asaf Rendel, formerly with IBM Research - Haifa, who recently passed away.

http://dx.doi.org/10.21437/Interspeech.2017-487

the k^{th} word, let l_1 be the maximum amount of *lengthening* Δ_l computed over any 3 consecutive HMM-state window $\max_j \sum_{i=j}^{i=j+3} \Delta_l(s_i)$ where the index j ranges over the s_i states in the second half of the k^{th} word. Let l_2 be the cumulative lengthening over the start of the following word w^{k+1}, $l_2 = \sum_n \Delta_l(s_n)$ where the index n ranges over the first 3 states of that word. The amount of lengthening Δ_l is defined as the difference between the *actual duration* of a state-sized fragment of speech (as determined by the acoustic alignments), and a model-based *predicted duration* for that unit based on a recurrent neural-network prosody model (see Section 3 for details on prosody models and architectures). Finally, let l_3 be the duration of any silence that follows the k^{th} word, then $phr_l = l_1 + l_2 + l_3$.

After applying these criteria, the corpus exhibits a phrase-break rate of 14.3%, 3.5% of which occur without an associated punctuation mark (the punctuation rate of the script is 11.8%). As expected, these figures demonstrate the need to not rely exclusively on punctuation as a predictor of phrasing since we observe punctuation marks in the text that do not co-occur with an acoustically realized phrase break (e.g., serial commas) as well as phrase boundaries occurring without an orthographic mark.

3. Modeling Approach

The modeling approach adopted is that of bidirectional recurrent neural networks (BiRNN), a class of models that has proved to be state-of-the-art in various prosodic prediction tasks for TTS [7]. The inputs to the model are various numerical and categorical features described as follows, where all categorical features have been one-hot encoded.

3.1. Predictive Features and Targets

The following base features are defined for each of the word tokens w_i in the input text:

- The type of punctuation following w_i.
- Unigram log probability $p(w_i)$.
- The Normalized Pairwise Mutual Information (NPMI) with respect to the previous and following words: $NPMI(w_i, w_{i-1})$ and $NPMI(w_i, w_{i+1})$, where:

$$NPMI(x, y) = \log \frac{p(x)p(y)}{p(x, y)} / \log p(x, y)$$

.

- The identity of the 80 most common words (primarily function words) appearing in the training corpus.
- The number of *normalized* words associated with w_i extracted from an alignment between the raw and normalized text strings (a naïve measure of the word's "complexity"). For instance, for the input token \$4.85 →*four dollars and eighty five cents*, this feature is 6.
- Learned (150-dimensional) embeddings predictive of a related task (see details in Sec. 3.2).

The probabilities to compute the features above are derived from a smoothed bigram language model trained on 350 million words [8].

The models are trained in a multi-task framework where, in addition to the binary phrase-break prediction primary task, we define the following targets for the auxiliary tasks:

- a normalized and clipped silence duration target $d_{sil} = \min(dur/400ms, 1.0)$ when a phrase boundary follows w_i and 0 otherwise; dur is the absolute duration of a post-lexical silence as determined by the alignments,
- a binary target signaling whether w_i is part of a short multi-adjective sequence.

The use of this last auxiliary target is to better model the role of punctuation and motivated by the fact that commas within adjectival sequences rarely co-occur with full phrase boundaries (unlike nominal comma-separated lists, where minor or even full phrase breaks with a continuation rise are common).

3.2. Learned Features

Whereas the supervised training corpus contains around 150K words, much larger amounts of unlabeled text are available, which could increase a model's generalization. To exploit unlabeled resources, we explore the approach of *feature learning*: training a model discriminatively to perform some presumably related task (e.g., language modeling), and then extracting the activations from one of the model's internal hidden layers and using them as input features for our main task(s). The related task we use is punctuation restoration, or simply a "comma predictor" since commas correlate with syntactical boundaries and phrasing. To create data to train such a "comma predictor" we use the Gigaword corpus and randomly drop out 80% of the commas (the remaining 20% are used as development for diagnosing convergence) to predict a 3-dimensional target encoding whether a word is followed by a comma, an em-dash, or neither. While aware that this methodology follows a standard target-driven supervised framework, the task can be treated as "unsupervised" with respect to the goal of the work since the trainable features are learned from a corpus lacking any phrasing labels, and is limited only by the amount of text data available provided it follows standard conventions for casing and punctuation.

The input features we use for the related task are 50-dimensional word embeddings and a punctuation feature (as in the baseline feature set above). We use a vocabulary of 70K words, extracted with respect to their frequency in the Gigaword corpus [9], with the following properties: (i) it is case sensitive, (ii) all its digits are mapped to the symbol #, effectively tying all digits strings of a common length (e.g., $123 \equiv 890 \equiv \#\#\#$), (iii) it includes special unknown-word symbols encoding the casing of the initial character, plus a 3-letter suffix with the word's ending, tying words that may have low frequency of occurrence, but a common morphological structure (e.g., *Einsteinium*≡*Mendelevium*≡UNK-<upper>-<*ium*>, and (iv) it includes a generic unknown-word symbol to assign to words not covered by the previous categories.

The model for the related-task prediction consists of a 4-layer recurrent bidirectional network using Long Short-Term Memory (LSTM) activation units, with the following structure: each layer is composed of a forward (F_n) and a backward layer (B_n), such that (i) the input to F_n is a concatenation of that composite layer's input with the output of B_n, and (ii) F_n provides the entire layer's output. The output dimensionality of the layers are 200, 150, 100, and 50, and that of the B layers is half (respectively). The 50-dimensional embeddings (inputs) are initialized with pre-computed substitute-based word embeddings available from [10] (and introduced by [11]), and are updated as part of the training. The activations at the second layer (i.e., output of F_2) are used to provide the learned features.

3.3. Final Model and Predictions

The structure of the final model consists of a standard 2-hidden-layer LSTM BiRNN with a multi-output layer to jointly predict the targets previously described. The input to this model are the features defined in section 3.1 (including the learned activations from the related-task network). The dimensionality of the forward hidden recurrent layers are 90 and 45 in each layer respectively, with the dimensionality of backward layers being half of those. Once the final model is trained, the following post-processing rules are used to turn the posterior scores output by the network ($BreakScore$) into a final set of phrase-break and pause-duration assignments:

```
if $w_i$ is followed by a terminal punctuation mark then
    → Break=1   Pause=400ms
else if $w_i$ is followed by non-terminal punctuation then
    if $w_i$ is part of a date pattern or location pattern then
        → Break=0   Pause=1ms
    else if $w_i$ is part of an adjectival sequence then
        → Break=0   Pause=50ms
    else if $BreakScore > PunctThresh$ then
        → Break=1   Pause=Predicted
    else
        → Break=1   Pause=50ms
    end if
else if $w_i$ is not followed by punctuation then
    if $BreakScore > HighThresh$ then
        → Break=1   Pause=Predicted
    else if $BreakScore > LowThresh$ then
        → Break=0   Pause=1ms
    else
        → Break=0   Pause=0ms
    end if
end if
```

Date patterns are strings such as *July 22nd, 2010* or *July, 2010* which were identified by matching. Punctuated location patterns are strings like *Alphen aan den Rijn, The Netherlands*, and their identification is based on data mined from the Geonames database [12]. The predicted duration above refers to normalized duration target predicted by the network times 200ms. The threshold values tested are as follows: $PunctThresh = 0.25$, $LowThresh = 0.65$ and $HighThresh = 0.75$, where the low and high thresholds were selected to correspond to approximately 80% and 90% precision on the test set respectively. These rules attempt to combine predictions with some heuristics in order to (a) delimit the raw scores output by the classifier in some meaningful way, such as by inspecting how they correlate with precision and recall on a held-out set, (b) refine the predictions of pause durations in an attempt to introduce *minor breaks* (not directly available in the training data) and (c) make allowances for more complex categories of punctuated strings, such as those containing geographic entities, which may not be observed enough in the training corpus.

4. Previous and Related Work

There is a rich literature on the relation between syntactical and prosodic structure spanning several decades of work [13, 14], with a number of contributions on computational approaches that attempted to predict prosodic structure (and phrasing in particular) from text (earlier proposals employed such modeling techniques as decision-tree and hidden Markov models [15, 16, 17]), and some with intended applications to speech synthesis [18, 19]. One characteristic of this line of work is the focus on syntactical information, and the dependency on annotated corpora for supervised model development: some of this paper's authors already explored the use of syntax on a phrasing task [20], and in [21] we find an example of recent work that focuses on phrasing for TTS exploiting features derived from a dependency parser. The proposal in this paper goes in an alternative direction in that it (i) focuses on a more lightweight approach that dispenses with syntactical and part-of-speech features for the sake of expediency in the TTS FE, (ii) makes extensive use of deep-learning techniques both as a final modeling strategy, and as a tool for extracting intermediate learnable representations (going beyond the lexical embedding representations already explored in [22]), and (iii) tries to alleviate the dependency on high-quality annotator labels for fully supervised learning.

5. Evaluation

The proposed approach was evaluated in terms of both objective performance metrics and impact on speech-synthesis quality against a baseline system that consists exclusively of linguistic rules. A detailed description of the full set of rules implemented by this baseline system, and ways in which they interact, is beyond the scope of this paper, but we summarize some of the main differences between the two approaches in Table 1.

Table 1: *Summary of differences between approaches.*

Feature	Baseline	Proposal
Break at unpunctuated boundary?	No	Yes
Pause duration at non-terminal breaks	Either 50ms or 150 ms pause	Variable predicted duration
Minor breaks?	No	1ms pauses
Breaks *within* dates and location patterns	May add a break and 50ms pause	Treats as units
Pauses around adjectival sequences?	No	50-ms pause

The baseline generally shows a stronger dependency on punctuation. The proposal, on the other hand, is able to predict more boundaries in the absence of punctuation, and more variability in post-phrasal pause durations. It is important to remark that, although at its core the proposal is a statistical system, it also subsumes the heuristics we have already explained to deal with dates and geographical locations, which are essentially corrective rules applied to compensate for perceived deffficiencies in the generalization of the statistical model. (In the evaluation, however, only a very small number of test cases (5 out of 195) were affected by these last 2 rules.)

5.1. Preliminary Objective Evaluation

To get an initial sense of how the model performs, we focused on non-punctuated boundaries since those constitute the most difficult case, and, because of their relatively low frequency of occurrence, can be overshadowed by the majority (punctuated-boundary) class (recall the training corpus showed only a 3.5% rate on non-punctuated class). By applying different thresholds to the model's posterior on the boundary class, we observed a reasonable behavior on the precision of the model, ranging from 80%-90%, with a corresponding drop in recall from 40%-30%; that is to say, an F1-measure ranging from 45%-53%. In contrast, the overall F1-measure of the model is around 90% when all boundaries (punctuated and not) are considered.

To look at generalization outside the domain of training, we considered the Boston Radio University Corpus and evaluated non-punctuated boundary performance for the one speaker (out of six) with the largest amount of labeled data (i.e., speaker F2). Although this corpus shows a rather different phrasing rate (around 22%), the metrics are somewhat comparable: precision in the 80%-90% range with a respective drop in recall from 45%-25% (i.e., an F1-measure ranging from 39%-57%).

5.2. Subjective Evaluation of Impact on Speech Synthesis

Since the final objective of this work is to improve the overall quality and naturalness of speech synthesis, and because different types of phrasing errors may be judged differently from a perceptual point of view, we turned to a formal listening test to compare the performance of the proposal against the baseline phrasing system. These models are evaluated under the IBM unit-selection speech-synthesis system with Bidirectional RNN prosody-target models, an architecture similar to the one described in [7] where the prosody of the post-search selected units are modified to their predicted targets.

The differences between the baseline and statistical phrasing models will be due to primarily three contributions: (1) the duration of the post-break pauses already described, (2) the location of the boundary, as well as (3) any effect that the categorical boundary may carry on to the target prosody model via features that summarize phrasal structure (examples of these are basic counts such as *number of* {*phones, syllables, words*} *to the* {*beginning, end*} *of a phrase*). In addition, matching phrase boundaries to the speaker style and to the original recording helps the system select longer contiguous segments from the original speech, potentially leading to higher-quality output.

Phrase-sensitive prosody models were built for two separate North American English synthetic voices: a female voice (FV) matching the 150K-word script already described (i.e., the training materials for phrase prediction), and a male voice (MV) reading an independent script of approximately 143K words. The test material were 195 sentences from three different domains: *general news*, as well as two statement categories useful in the context of argumentation: *expert evidence* and *study evidence*. These correspond to relatively long segments of text summarizing, respectively, the opinions of experts and the contents of a study, and which are used to back up a position on a topic. The choice of these categories stems from ongoing work on having a synthetic voice capable of engaging in persuasive conversation (details may be found in [23]). A crowd-sourced mean-opinion-score (MOS) test was carried out to rate each text under the two models by asking subjects for 5-point scale judgments on two attributes: overall and phrasing quality. The questions and scales were as follows: *I-Rate the overall quality and naturalness of this sample. (1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent)* and *II-Rate the phrasing of this sample in terms of its use of pauses, rhythm, flow, etc. (1=Bad: Severe phrasing issue(s), 2=Poor, 3=Fair, 4=Good, 5=Excellent: No perceivable phrasing issues)*. Although we are interested in isolating the effect of our approach on phrasing quality, we include a question about overall naturalness to ensure there is no trade-off between these two attributes, resulting in potential perceptual artifacts impacting overall quality. Table 2 summarizes results for the two systems and voices, and p-value of statistical significance assessed via the non-parametric Mann-Whitney U-test. We see that the proposal leads to a significant improvement in the perceived phrasing quality for both synthetic voices, and that this gain comes at no loss in overall naturalness.

Table 2: *Overall MOS scores (and standard deviation) for two synthetic voices and systems.*

Voice	Attribute	Baseline	Proposal	p-value
FV	Naturalness	3.75 (.83)	3.78 (.8)	.154
	Phrasing	3.76 (.92)	3.86 (.86)	**.0031**
MV	Naturalness	3.69 (.83)	3.75 (0.84)	.002
	Phrasing	3.71 (.89)	3.78 (.85)	**.0071**

In table 3 we have segregated the above results according to the genre of the material. In this case, we observe that different genres contribute differently to the overall gains, and that there is complementary behavior for the 2 different voices: for FV the *news* and *study-evidence* are significantly improved whereas for MV the gains are reflected in the *expert-evidence* sentences. For all the voices and genres, we again confirm that naturalness does not degrade as a result of the proposed phrasing model.

Table 3: *MOS scores (and standard deviation) for two synthetic voices and systems by genre (GN=general news; EE=expert evidence; SE=study evidence).*

Genre	Voice	Att.	Baseline	Proposal	p-value
GN	FV	Nat.	3.60 (.90)	3.69 (.83)	.073
		Phr.	3.61 (1.01)	3.83 (.90)	**.0016**
	MV	Nat.	3.61 (.83)	3.63 (.87)	.324
		Phr.	3.63 (.91)	3.63 (.89)	.450
EE	FV	Nat.	3.77 (.82)	3.78 (.81)	.366
		Phr.	3.84 (.89)	3.83 (.85)	.337
	MV	Nat.	3.66 (.84)	3.75 (.84)	.035
		Phr.	3.70 (.89)	3.83 (.82)	**.0047**
SE	FV	Nat.	3.81 (.77)	3.84 (.77)	.275
		Phr.	3.77 (.87)	3.91 (.84)	**.0025**
	MV	Nat.	3.77 (.81)	3.82 (.80)	.117
		Phr.	3.75 (.87)	3.82 (.82)	.080

6. Future Extensions and Conclusions

We have described and evaluated an approach for building trainable phrase-prediction models for use in a TTS FE showing significant perceptual differences in phrasing across two distinct North American English voices. The main motivators for this approach were: (i) the reliance on features that can be easily extracted (or learned) directly (to reduce dependency on parsing and tagging), and (ii) the use of a ground truth for supervised learning that minimized the need for hand-labeled annotations by using, instead, a set of combined predictive heuristics. The nature of some of those heuristics, however, assume either the existence of linguistic knowledge, or a prosodic acoustic labeler (AuToBi), which in turn has been trained on hand-labeled annotations. To make this work extensible, we would need to refine the heuristics and explore ways in which we can relax language dependency. One possible avenue is to look at the feasibility of using pre-trained acoustic prosodic labelers across languages (and deploy the English AuToBi labeler on other datasets to produce one candidate estimate of phrasing), or to rely more heavily on cues of phrasing that show robustness across languages (e.g., silence; see [24, 25] for some precedent research). This direction remains the topic of future work.

7. References

[1] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, “The Boston University radio news corpus,” Boston University, Tech. Rep. ECS-95-001, 1996.

[2] C. Christine 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, pp. 106–112.

[3] A. Rosenberg, “AuToBI - a tool for automatic ToBI annotation.” in *Interspeech*, Tokyo, 2010, pp. 146–149.

[4] ——, “Modeling intensity contours and the interaction of pitch and intensity to improve automatic prosodic event detection and classification,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, Dec 2012, pp. 376–381.

[5] “AuToBi,” http://eniac.cs.qc.cuny.edu/andrew/autobi/, 2015, [Online].

[6] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, “Segmental durations in the vicinity of prosodic phrase boundaries,” *JASA*, vol. 91, no. 3, pp. 1707–17, April 1992.

[7] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, “Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system,” in *Proc. Interspeech*, Dresden, 2015, pp. 1606–1610.

[8] S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, “Advances in speech transcription at IBM under the DARPA EARS program,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[9] D. Graff and C. Cieri, “English Gigaword,” http://catalog.ldc.upenn.edu/LDC2003T05, 2003.

[10] O. Melamud, D. McClosky, S. Patwardhan, and M. Bansal, “The role of context types and dimensionality in learning word embeddings,” in *NAACL-HLT*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1030–1040. [Online]. Available: http://www.aclweb.org/anthology/N16-1118

[11] M. A. Yatbaz, E. Sert, and D. Yuret, “Learning syntactic categories using paradigmatic representations of word context.” in *EMNLP*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 940–951.

[12] “Geonames,” http://http://www.geonames.org/, 2016.

[13] J. Bachenko and E. Fitzpatrick, “A computational grammar of discourse-neutral prosodic phrasing in english,” *Computational Linguistics*, vol. 16, pp. 155–170, 1990.

[14] N. M. Veilleux, “Computational models of the prosody/syntax mapping for spoken language systems,” Ph.D. dissertation, Boston University, Boston, MA, USA, 1994.

[15] M. Q. Wang and J. Hirschberg, “Predicting intonational boundaries automatically from text: the ATIS domain,” in *HLT ’91: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1991, pp. 378–383.

[16] ——, “Automatic classification of intonational phrase boundaries,” *Computer Speech and Language*, vol. 6, no. 2, pp. 175–196, 1992.

[17] M. Ostendorf and N. Veilleux, “A hierarchical stochastic model for automatic prediction of prosodic boundary location,” *Computational Linguistics*, vol. 20, no. 1, pp. 27–54, 1994.

[18] 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.

[19] P. Koehn, S. Abney, J. Hirschberg, and C. M., “Improving intonational phrasing with syntactic information,” in *ICASSP*, Istanbul, Turkey, 2000, pp. 1289–1290.

[20] A. Rosenberg, R. Fernandez, and B. Ramabhadran, “Phrase boundary assignment from text in multiple domains,” in *Interspeech*, Portland, 2012, pp. 2558–2561.

[21] T. Mishra, Y. Kim, and S. Bangalore, “Intonational phrase break prediction for Text-to-Speech synthesis using dependency relations,” in *ICASSP*, April 2015, pp. 4919–4923.

[22] 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 *ICASSP*, March 2016, pp. 5655–5659.

[23] R. Rinott, L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, and N. Slonim, “Show me your evidence - An automatic method for context dependent evidence detection,” in *EMNLP*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 440–450.

[24] A. Rosenberg, , E. Cooper, R. Levitan, and J. Hirschberg, “Cross-language prominence detection,” in *Speech Prosody*, Shanghai, 2012.

[25] V. Soto, E. Cooper, A. Rosenberg, and J. Hirschberg, “Cross-lanugage phrase boundary detection,” in *ICASSP*, Vancouver, Canada, 2013, pp. 8460–8464.