A Grammar Based Approach to Style Specific Phrase Prediction

Alok Parlikar and Alan W Black

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213. USA
{aup, awb} @cs.cmu.edu

Abstract

We present an approach to style specific phrasing for Text-to-Speech (TTS) systems. We formulate the problem of phrase break prediction (or phrasing) as generation of a sequence of breaks (B) and non-breaks (NB) after each word in a sentence. We use prosodic breaks in speech data to build shallow parses over corresponding text. We then learn a grammar that can predict these shallow prosodic parses from text. We then combine this prosodic phrasing information with other word level features in a CART tree to predict where phrase breaks should be inserted in new text. We show that a model built to target a specific reading style can predict phrase breaks more accurately than the standard generic model.

Index Terms: Speech Synthesis, Style-specific Phrase Breaks

1. Introduction

Predicting prosodic phrases (phrase breaks) is an essential step during speech synthesis, because other prosodic models depend on it. Phrase Break Prediction (PBP) models are typically trained on standard corpora. For example, the Festival[1] system uses a model[2] that is trained on the MARSEC[3] data. However, the same generic model gets used for all speakers and styles of speech. In addition, this model only uses three features to predict the breaks: the parts of speech of two previous words and one next word. While this quad-gram PBP is simple and efficient, it does not capture longer distance context which could be provided by a syntactic structure.

This work is an effort to investigate whether better models can be achieved by using phrase structure informaion. Although it has been pointed out that traditional syntactic phrase structure is not directly appropriate for prosodic phrasing[4], as prosodic phrases often cross over syntactic boundaries, there are still linguistic aspects that are useful in building prosodic phrasing models. We train and use our model on data containing different styles of speech, and show that longer distance information can have an impact on phrasing, and thus affect other prosody models positively.

We make direct use of a traditional part-of-speech (POS) tagger. We use the POS tagger that is built-in to Festival which is a standard statisistical tagger trained from the Penn Treebank[5]. For the grammatical structure part of the system we are **not** interested in traditional linguistic syntactic constituents. We are interested in prosodic phrase constituents. Thus we train a Stochastic Context Free Grammar for each of our styles, using a standard forward-backward algorithm[6], again part of the Festival suite of utilities. The prosodic phrase structure grammar is trained from bracketed POS tags. The bracketing is derived from the acoustic information derived when labeling the phonetic segments in our speech databases. Prosodic phrases are defined in this data as the words between pauses of length greater than 80ms. Thus the SCFGs that are trained may be thought of prosodic phrase chunking and do not necessarily relate to traditional syntactic phrases, but capture the tag chunking that is actually found in our speech data.

We train our models from the same data that is used to build a synthetic voice: a list of text prompts, and corresponding speech recordings. We phonetically label the speech using its transcripts. This labeling provides information about where the speaker inserted pauses when reading text, and how long the pauses were. In Section 2, we describe the data sets used in this work and in Section 3, we analyze the distribution of breaks in the respective styles. In Section 4 and 5, we describe our grammar based method in detail. We then provide objective and subjective evaluation of our approach in Section 7.

2. Corpora in different styles

We looked at five different corpora that have speech in different styles.

The *ARCTIC-A* corpus consists of the ARCTIC-A prompt set[7] recorded speaker AUP (an Indian English speaker). The style of this corpus is "short sentences". The corpus has 593 prompts with an average of 9 words per prompt and the audio size is about 30 minutes.

We took the *Europarl*[8] parallel corpus between English and Portuguese. This data contains proceedings of the European Parliament. We selected prompts from the English side of the corpus. These prompts were also recorded by speaker AUP. The style of this corpus is "parliament proceedings". The corpus has 595 prompts with an average of 14 words per prompt and the audio size is about 50 minutes.

The *F2B* corpus is from the Boston University Radio News Corpus[9]. The style of this corpus is "radio broadcast". The corpus has 464 prompts with an average of 19 words per prompt and the audio size is about 55 minutes.

The *Obama* corpus consists of public talks by the US President Barack Obama. Audio and transcripts of two of his public addresses were used to build this voice: (i) 2009 Presidential candidate speech "A more perfect Union", (Mar 2008, Philadelphia) and (ii) Address at the Military Academy (Dec 2009). The style of this corpus is "public address". The corpus has 465 prompts with an average of 18 words per prompt and the audio size is about 61 minutes.

The *Emma* corpus[10] is taken from an Audiobook (Emma, by Jane Austen) in the Librivox database. The book was recorded by a female volunteer. The style of this corpus is "audio book". The corpus has 9936 prompts with an average of 15 words per prompt and the audio size is about 1040 minutes.

Within the Festival[1] and Festvox[11] frameworks, we built a clustergen[12] voice on each of these datasets. In the process, we phonetically labeled these datasets, and created festival's utterance structures for the data.

3. Analysis of Phrase Breaks

When people read out text in different styles, they could be inserting phrase breaks differently. To understand the extent to which phrase break distributions differ across styles, we analyzed the different datasets at hand.

Using the festival utterance structures created when building the clustergen voices, we can extract information about where in every utterance the speaker had inserted a break, what its duration was, and in what context it occurred.

3.1. Break/Non-break Distribution

The different styles of speech appear to vary with respect to the global distribution of breaks versus non-breaks. We measured the percentage of word boundaries where a break was found in the original recordings for each dataset. Note that we excluded the breaks at the end of the utterances. Table 1 shows that while the ARCTIC-A, Europarl and F2B datasets have a similar proportion of breaks in them, the Obama and Emma data have more breaks. The table also shows how many breaks were globally predicted by festival's default phrasing model on each dataset. The numbers show that the default model is inserting more breaks than expected. To see why this may be the case, we looked at the MARSEC[3] data from which the default phrasing model is trained. That data has 14.15% of the word boundaries marked with breaks.

Table 1: Percentage of breaks in corpus

Dataset	Total Words	Actual Breaks	Default Predictions
ARCTIC-A	5313	6.25 %	8.96 %
Europarl	8066	6.48 %	11.28 %
F2B	9214	6.37 %	14.30 %
Obama	8402	9.21 %	14.50 %
Emma	158209	8.27 %	16.19 %

3.2. Duration of Breaks

It turns out that the styles we are looking at don't differ just in the proportion of breaks, but also the distribution of durations of the breaks. We looked at the histograms of break durations on the datasets and observed that breaks in ARCTIC-A and Europarl are of similar lengths, whereas Emma and F2B have longer breaks on average. The Obama corpus has many long breaks, and also a long tail of breaks that go well over half a second in duration. Table 2 summarizes the parameters of these distributions.

Table 2: Duration in seconds of pauses in recorded speech

Dataset	Mean	Stdev
ARCTIC-A	0.115	0.059
Europarl	0.111	0.067
F2B	0.273	0.099
Obama	0.391	0.311
Emma	0.180	0.162

4. Overview of Grammar Based Approach

Our Grammar based PBP method involves two models: (i) A CFG that can parse, or chunk given text, and (ii) A CART Tree that can actually predict the phrase breaks.

The Grammar introduces a bracketed structure over the text. The idea is to obtain prosodic phrase constituents within which a prosodic break is unlikely to occur. Section 5 describes how we learned and used the context free grammar.

The CART tree uses features produced by the phrase structure in addition to other features within a Festival utterance and predicts whether a word boundary should be marked as a break (B) or not (NB). This technique is described in Section 6.

Festival's PBP model uses two models: a "phrasing" model and a language model. The phrasing model provides candidate breaks, and the language model is used to conduct a Viterbi search for the best sequence of breaks. Our CART tree simply replaces the phrasing model. We still take advantage of the Viterbi search and use Festival's default language model for the purpose.

5. Grammar for Prosodic Parsing

The idea of Prosodic Parsing, or Prosodic Chunking is to identify constituents in given text within which a break is unlikely to occur. For a given sentence, there can be multiple ways to parse it. Take an example sentence: "There are five hundred students in the auditorium." If we use brackets to mark prosodic constituents, two possible prosodic parses of this sentence are: (i) "(There are) (five hundred students) (in this classroom)", and (ii) "(There are five hundred students) (in this classroom)". However, the parse "(There are five) (hundred students) (in this classroom)" seems unlikely. While on many ocassions linguistic constituents can form prosodic constituents, we are only looking at and using the latter. It has also been shown that prosodic phrasing can be significantly different from syntactic phrasing [13]. Prosodic constituents are easy to obtain, since we have speech labeled with breaks. They also may be more conforming to the style of the speech than constituents obtained using some other parsing grammar. In this section, we describe how we learned the grammar, and how we use it.

5.1. Training the Grammar

The basic step of training the Grammar is identical to the examples described above. We take 90% of our dataset for training, and use the phonetic labels to annotate prosodic constituents for each utterance. We then use Festival's SCFG utilities[6] that build the required grammar.

Unlike in the example described above, we do not use words as non-terminals of our grammar, as we do not have nearly enough data to train such models. Words are collapsed to a finite set of tags. We use POS tags of words as the terminal symbols. Festival's POS tagger uses the tagset from the Penn Tree Bank[5]. We can either use that full set, or further reduce it. We experimented with three different tagsets that Festival provides, described in Table 3. The first POS is the standard set from the Penn Tree Bank. The second is GPOS (or "guessed part of speech") which derives its tags from a trival look up table for function words and classifies all others as content word – this type of tagger is very easy to implement in new languages when no POS training data is available. The third set is a reduced Penn Tree Bank set that was automatically optimized for phrase break prediction in [2].

Table 3: Non-terminals in Grammar Training

Tag Set	Non Terminals
POS	Entire PennTreeBank set
GPOS	aux, cc, content, det, in, md, pps, to, wp
PHRPOS	cc, cd, dt, ex, in, j, md, n, of, pdt, pos, prp, punc, r, to, uh, v, wdt, wp, wrb

Apart from the set of terminals, another parameter to configure in grammar training is the number of non-terminals. We ran experiments with non-terminal counts of 5, 10 and 15.

The two main parameters in our grammar training: tagset, and non-terminal count, were each configured to have three possible values. That gives us nine different models to choose from. We built the nine models on the F2B corpus. Our preliminary results showed that the PHRPOS tagset with a non-terminal count of 10 gave optimal result, thus we use that combination for further experiments.

Once we have a grammar built, Festival has the necessary tools to use it and parse new text. This introduces a "Syntax" relation within its utterance, that provides us with the bracketed structure over text. This bracketing can be used to design features for CART training.

6. Phrase Break Prediction with CART

Once we have built a grammar from our style specific training data, we parse our entire training set with the that grammar. We then dump features for each word in the corpus, including features about their positions with respect to the prosodic phrase structure predicted by the grammar.

With the word-level features and the truth value of break/no-break, we train a CART classification tree using wagon. We use 80% of the speech data as training, 10% data for development and held out the remainder for testing. We optimize the trees for entropy, rather than classification accuracy. We empirically found that the most reasonable stop-value for CART training was "5" for the smaller datasets, however since the Emma dataset is quite large, the models did better with a stop-value of "50" (and is quicker to train).

Table 4 lists the features that we included in our CART training. To take context into account, we use these features for the current word, two previous words, and two next words. After building trees on all the datasets, we looked at the top features in the their respective trees. The features "has-punc", "end-brackets", "delta-brackets" and "gpos" seem to be the ones carrying a lot of information about the breaks.

The CART trees act as drop-in replacements of Festival's phrasing model. At synthesis time, we use the appropriate grammar to parse our utterance, and find out the probabilities of break/non-break at each word boundary. We divide this probability by the unigram probabilities of breaks and non-breaks in the style specific corpus during the Viterbi search.

Table 4: CART model features

Name	Description
pos	Part of Speech
gpos	Collapsed Part of Speech
has-punc	Is word followed by punctuation
lpunc	Current token is punctuation, but not next
token-in-quote	Does a single quote appear in this or previous token? (Disambiguate end-quote from possessive)
dist-to-eos	No. of words before sentence end
Grammar:	
end-brackets	Count end-brackets in prosodic parse
start-brackets	Count open-brackets in prosodic parse
delta-brackets	(scfg-end-brackets) − (scfg-start-brackets)
abs-delta-brackets	abs(scfg-delta-brackets)

7. Evaluation

A Phrase Break Prediction model can be evaluated by comparing the prediction it makes to the actual breaks identified in speech. Accuracy, as an objective measure can be used to compare the new grammar based model to the standard model in Festival. Because the phrasing model has an impact on other prosodic models, we can compare the synthesized waveforms and evaluate whether the new model produces a waveform better on an objective measure. Apart from this, we can conduct listening tests to assess if the new phrasing model makes any changes that perceptually improve speech quality. This section describes our results in these three evaluations.

7.1. Objective Evaluation

7.1.1. Accuracy of Break Prediction

Given the predictions of a PBP model on heldout data and the corresponding truth values, we can measure accuracy in two different ways. We could measure the percentage of breaks and non-breaks correct, or the overall accuracy. However, since the task basically involves predicting a break (non-breaks are there by default), we thought measuring the F-1 measure of breaks is a suitable measure of accuracy. Ideally, we want to predict all breaks actually present (high recall) and predict no other word boundaries as being breaks (high precision). We hence use the F1 metric. Table 5 shows our results.

All but the Emma datasets are small in size and the train/test split of the data might introduce high variance in the results. We hence performed a 10-fold cross validation on these data and present the average results here. The Emma dataset is large, and building models on it is computationally expensive, so it has not been cross-validated.

7.1.2. MCD of Synthesis

Mel-Cepstral distortion(MCD) is the objective metric often used to judge voice quality of synthesized speech. Calculation of MCD requires a time-alignment of the two speech samples, which can be done using Dynamic Time Warping (DTW). Similar to [13], we use the MCD to evaluate how the new phrasing model compares to the default model. Except for the Emma corpus, we perform 10 fold cross validation and report average MCD. Table 6 shows our results.

Table 5: Break Prediction Accuracy (F1 score)

DataSet	Base F1	New F1
ARCTIC-A	80.11	85.12
Europarl	70.42	77.67
F2B	66.17	73.67
Obama	66.41	63.80
Emma	69.98	82.94

Table 6: Comparison of MCD with Default Breaks Versus Learned Breaks

DataSet	Base MCD	New MCD
ARCTIC-A	7.47	7.18
Europarl	7.12	6.67
F2B	6.20	5.95
Obama	10.25	10.08
Emma	6.98	6.60

7.2. Subjective Evaluation

In addition to the objective evaluation, we conducted a listening test to compare the new model to Festival's default model. We only conducted this evaluation on the F2B corpus. Twenty randomly selected sentences from the held out test set were synthesized using the standard phrasing model and the new model. Ten fluent speakers of English (a mix of people native to the US, the UK and India) were the subjects of this experiment. For each test sentence, subjects listened to the two audio clips synthesized using the two models. They marked the version that they preferred. Test sentences were presented in random order, and the two audio clips for each sentence were played in random order as well. Excluding some responses that subjects could not submit due to technical reasons, we have 150 datapoints of system preferences. Figure 1 shows the outcome of this evaluation.

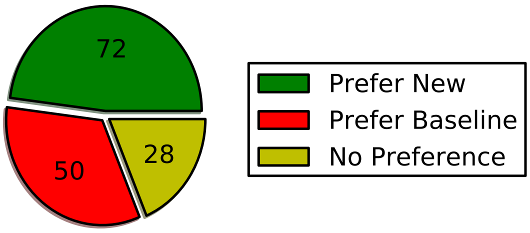

Figure 1: Subjective Preference of Phrasing Models

8. Conclusions and Future Work

We have shown that data used to build a synthetic voice can be used to build a customized phrase break prediction model that performs better in both objective and subjective evaluation. The improved model also improves the quality of synthesis. The model makes use of notion of prosodic grammar and features based on this syntax are quite useful in CART trees that predict phrase breaks.

The next step is to try building similar models on non-standard speech corpora. It has been shown[14] that improvements in phrasing could improve the synthesis of automatically translated text. It would be interesting to use this grammar based model on recordings of MT output and see if that improves intelligibility of the synthesis.

In this work, we only modelled the location of phrase breaks. However, our analysis shows that duration of the breaks themselves also relate to the underlying style and we need to extend our models to predicting the location and duration of breaks.

9. Acknowledgment

This work was supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

10. References

[1] A. W. Black and P. Taylor, "The festival speech synthesis system: system documentation," Human Communication Research Centre, University of Edinburgh, Tech. Rep., January 1997. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival

[2] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[3] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "Marsec: A machine-readable spoken english corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47–53, 1993.

[4] J. Bachenko, E. Fitzpatrick, and C. Wright, "The contribution of parsing to prosodic phrasing in an experimental text-to-speech system," in *Association for Computational Linguistics*, New York, New York, 1986, pp. 145–153.

[5] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1994.

[6] F. Pereira and Y. Schabes, "Inside-outside reestimation from partially bracket corpora," in *Association for Computational Linguistics*, Newark, Delaware, 1992, pp. 128–135.

[7] J. Kominek and A. W. Black, "CMU arctic databases for speech synthesis," in *SSW-5*, Pittsburgh, Pennsylvania, June 2004, pp. 223–224.

[8] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Machine Translation Summit*, Phuket, Thailand, September 2005, pp. 79–86.

[9] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep., March 1995. [Online]. Available: http://ssli.ee.washington.edu/papers/radionews-tech.ps

[10] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[11] A. W. Black and K. Lenzo, "Building voices in the festival speech synthesis system," 2002, http://festvox.org/bsv/.

[12] A. W. Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Interspeech*, Pittsburgh, Pennsylvania, September 2006, pp. 194–197.

[13] K. Prahallad, E. V. Raghavendra, and A. W. Black, "Learning speaker-specific phrase breaks for text-to-speech systems," in *SSW-7*, Japan, September 2010.

[14] A. Parlikar, A. W. Black, and S. Vogel, "Improving speech synthesis of machine translation output," in *Interspeech*, Makuhari, Japan, September 2010, pp. 194–197.