Semi-Supervised Learning of Acoustic Driven Prosodic Phrase Breaks for Text-to-Speech Systems

Kishore Prahallad[1 2], *E. Veera Raghavendra* [1], *Alan W Black* [2]

[1] International Institute of Information Technology, Hyderabad, India.
[2] Language Technologies Institute, Carnegie Mellon University, USA.

kishore@iiit.ac.in, raghavendra@iiit.ac.in, awb@cs.cmu.edu

Abstract

In this paper, we propose a semi-supervised learning of acoustic driven phrase breaks and its usefulness for text-to-speech systems. In this work, we derive a set of initial hypothesis of phrase breaks in a speech signal using pause as an acoustic cue. As these initial estimates are obtained based on knowledge of speech production and speech signal processing, one could treat the hypothesized phrase break regions as labeled data. Features such as duration, F0 and energy are extracted from these labeled regions and a machine learning model is trained to perform the classification of these acoustic features as belonging to the class of a phrase break or not a phrase break. We then attempt to bootstrap the machine learning model using unlabeled data (i.e., the rest of the data).

Index Terms: speech synthesis, acoustic driven phrasing, semi-supervised

1. Introduction

While it is generally agreed that there is some relationship between syntax and prosody, the relationship is not well defined. It is also known that the syntactic phrase structure is not identical to prosodic phrase structure of a given text. In the context of text-to-speech (TTS), it is essential to predict prosodic phrase breaks in the text. Prosodic phrase breaks are used by different modules such as F0 generation, duration and insertion of pauses which are known to affect the quality of synthesized speech. To build a Prosodic Phrase Break Predictor (PBP), which predicts prosodic phrase breaks in the text, the training data is typically hand annotated with a break symbol between words which are perceived as being phrase breaks. Moreover, PBP make use of either syntactic level information or Part-of-Speech (POS) tag sequences as features extracted from the text data to predict prosodic phrase breaks between the words.

The current techniques of building PBP has the following disadvantages: 1) In terms of preparation of training data to build a PBP, the process of hand annotation is laborious, time consuming and is hardly scalable to different languages. 2) Typically, a PBP built on a standard corpus (for example, in Festival, a PBP is built on Boston University Radio corpus data) is employed to predict breaks for all voices. Thus prosodic phrasing patterns are generalized across all voices while ignoring the speaker-specific phrasing patterns. 3) PBP assumes availability of syntactic parsers or POS taggers. The availability of such linguistic resources may be difficult for minority or resource poor languages.

In order to address the issue of preparing training data to build a PBP, our focus is on acoustic driven prosodic phrase annotation of text, where acoustic cues present in a speech signal are exploited to identify the phrase breaks in the corresponding text transcription. It is known that acoustic cues such as pauses, prepausal lengthening of rhyme, speaking rate, breaths, boundary tones, glottization are useful in indicating phrase breaks in the speech signal [1] [2] [3]. The use of acoustic driven prosodic phrase annotation of text circumvents the need of manual preparation of training data for PBP and also provide the scope for incorporation of speaker-specific phrasing patterns

Detection of prosodic phrase breaks based on acoustic cues has been attempted earlier. Wightman *et. al.,* [4] have used acoustic cues for automatic identification of phrase breaks using machine learning techniques based on supervised learning. In [4], the training data used to build an Acoustic Driven Prosodic Phrase Annotator (ADP) was obtained from Boston University Radio Corpus which is hand annotated with phrase breaks in the speech signal. Ananthakrishnan *et. al.,* [5] have used unsupervised techniques for prosody labeling, but relies heavily on lexical and syntactic features to improve the acoustic clustering. Our work in this paper differs from these earlier works in the following ways. In this work, we derive a set of initial hypothesis of phrase breaks in the speech signal using pause as an acoustic cue. As these initial estimates are obtained based on knowledge of speech production and speech signal processing, one could treat the hypothesized phrase break regions as labeled data. Features such as duration, F0 and energy are extracted from these labeled regions and a machine learning algorithm is used to build a ADP, which performs the classification of these acoustic features as belonging to the class of a phrase break or not a phrase break. We then attempt to bootstrap the ADP model using unlabeled data (i.e., the rest of the data). Thus our approach could be viewed as semi-supervised learning of acoustic driven phrasing.

This paper is organized as follows: Section II describes the nature of acoustic cues indicating the phrase break information in the speech signal. Section III describes the process of building acoustic driven phrasing models and the method of bootstrapping adapted in this work. Section IV discusses the results of acoustic driven phrasing on IIIT-LEN voice, which is a speech database in Telugu (an Indian language).

2. Nature of Acoustic Cues Indicating Prosodic Phrase Breaks

In order to illustrate the nature of the acoustic cues that indicate prosodic phrase breaks, a simple listening experiment was conducted using utterances from a story, where each utterance was one or two paragraphs long. These utterances were part of a story (Chapter 2 of EMMA by Jane Austen) recorded by a female speaker in Librivox database. The text of these utterances were marked with punctuation marks such as comma, fullstop and a semicolon thus providing sufficient hints to the reader about the possible prosodic or syntactic boundaries. The story was spoken in a story telling fashion with pauses wherever required. From the original recordings (referred to as set-A), a new set of utterances referred to as set-B was created by removing pauses in each of the utterances in set-A. These pauses were manually but carefully removed especially in the case of stops preceding or succeeding the pauses.

A set of 5 non-native speakers of English acted as listening subjects in this experiment. The subjects were asked to listen to each utterance in set-B on day one. They were given the text of the utterance with all punctuations and capital letters removed, and were asked to mark the punctuation wherever they perceived a break in acoustic signal. A day later, the same five subjects were asked to listen to the utterances in set-A. They were given the text with all punctuations and upper casing of the letters removed, and were asked to mark the punctuation wherever they perceived a break in acoustic signal. A sample utterance is shown below. *" Sorrow came (75:5:5) – a gentle sorrow (370:5:5) – but not at all in the shape of any disagreeable consciousness (550:4:5). Miss Taylor married (640:5:5). It was Miss Taylor's loss which first brought grief (550:5:5). It was on the wedding-day of this beloved friend that Emma first sat in mournful thought of any continuance (1290:5:5).... "*

At each punctuation mark i, the three numbers in succession denote 1) the duration of the pause in Milli seconds , 2) number of subjects thought they perceived a break in listening the utterance from set-B which is denoted by s_i^B and 3) number of subjects thought they perceived a break in listening the utterance from set-A which is denoted by s_i^A. The value of the pair (s_i^B, s_i^A) range from $(0, 0)$ to $(5, 5)$. In total there were 63 locations spread over all 5 utterances where subjects perceived a break.

A scatter plot of the pair of values (s_i^A, s_i^B), where $0 \leq s_i^A \leq 5$, $0 \leq s_i^B \leq 5$, and $i = 1..63$ is as shown in the Fig. 1. The value of s_i^A and s_i^B is referred to as score in Fig. 1. The scatter plot shown in Fig. 1 demonstrates a correlation of 0.82 between the values of s_i^B and s_i^A. Further analysis showed that 1) in 92% of the cases, at least two subjects (one during set-A, and another during set-B) agreed / perceived a break at the same location 2) in 33.3% of the cases, all the five subjects (during set-A and during set-B) perceived a break at the same location and 3) There was higher correlation (0.952) between the location of the percevied boundary and the existence of a punctuation mark in the original text. This also indicates that the punctuation marks acted as a guide to the speaker of the paragraphs to introduce boundaries during production process. The correlation of 0.82 between the values of s_i^B and s_i^A indicate that acoustic cues other than simple pause play a major role in indicating a phrase break in the speech signal. This is substantiated by the observation that in 92% of the cases, atleast two subjects (one during set-A, and another during set-B) agreed / perceived a break at the same location.

This experiment shows that acoustic cues other than pauses play a role in indicating prosodic phrase breaks. However, an enumeration of these non-pause cues is a difficult task. While studies have shown that acoustic cues such as pre-pausal lengthening of rhyme, speaking rate, breaths, boundary tones and glottization play a role in indicating the phrase breaks in a speech signal [1] [2] [3], the representation / parameterization of these complex acoustic cues is not well understood. Many of these complex acoustic cues are often represented by extraction of average duration, F0 and energy values [6]. In our work, we have also adapted the extraction of average duration, F0 and energy values to represent non-pause acoustic cues of phrase break as shown in Table 1.

Figure 1: Scatter plot of scores obtained for utterances in Set-A and Set-B

Table 1: Syllable level features extracted at phrase break

Break Features	Description
pause duration	Duration of the pause at the word boundary
vowel duration	Vowel duration in the syllable
f0_maxavg_diff	Diff. of max and avg f0
f0_range	Diff. of max and min f0
f0_avgmin_diff	Diff. of avg and min f0
f0_avgutt_diff	Diff. of syl avg and utterance avg f0
en_maxavg_diff	Diff. of max and avg energy
en_range	Diff. of max and min energy
en_avgmin_diff	Diff. of avg and min energy
en_avgutt_diff	Diff. of syl avg and utterance avg energy

3. Building an Acoustic Driven Prosodic Phrase Annotator

3.1. Building a baseline ADP (ADP-0)

The process involved in building a baseline ADP relies on hypothesis of phrase break regions based on pauses in speech signal. The steps involved in building the baseline ADP (referred to as ADP-0) are as follows.

- Identify the word level boundaries in the speech signal based on the forced-alignment of speech with the corresponding transcript. The forced-alignment technique used here provides an optional silence HMM between every word, and hence during Viterbi decoding if there exists a pause region then it will marked automatically.
- Identify the pause regions p in the speech signal. Based on the duration of pause regions p_d, the pauses are marked as B and BB. Here B denotes a type of phrase break, when $50\ ms \geq p_d \leq 150\ ms$, and BB denotes another type of phrase break when $p_d > 150\ ms$.

3.2. Bootstrapping an ADP model (ADP-1)

For speech databases in English language, we have an option of using supervised technique of building an ADP model. In [7], an ADP model built on F2B (Boston University Radio Corpus) was used as seed model and was bootstrapped with unlabeled EMMA speech database. Iterative training was performed to obtain the final ADP model for EMMA database. It was shown in [7], that such bootstrapping technique is useful to generate acoustic driven phrasing for unlabeled (EMMA) database and is an useful tool to improve the quality of synthetic voice. However, this technique of bootstrapping requires a hand labeled data which may be difficult to obtain for speech databases in other languages. We have also tried to use an ADP model built on F2B (English language) to bootstrap on IIIT-LEN (Telugu language). However, such cross-lingual bootstrapping did not perform well with respect to baseline systems. Also such process of bootstrapping from models built on different speakers may have bias in capturing speaker-specific phrasing patterns.

In this work, we build an ADP-1 model based on the phrase breaks regions identified by ADP-0 as described in Section 3.1. The steps involved in building the ADP-1 by the process of bootstrapping on phrase break regions identified by ADP-0 is as follows.

1. Extract duration, F0 and energy features from the phrase regions as identified by ADP-0 in Section 3.1. At each phrase break, a set of 10 features related to duration, F0 and energy features are computed for the last syllable (ν) in the word at the phrase break. Similar features are computed for two neighboring (one left and right) syllable of ν. The feature set computed for each syllable is shown in Table 1, and is based on the work in [6].
2. Build a CART model, where the predictee is phrase break level (B / BB / NB) and the predictors are duration and F0 features. Here NB denotes not a phrase break. The features for NB are obtained by considering the acoustic features of syllables in a word which is immediate previous to a word identified as phrase break (B / BB).
3. Use the CART model to (re)-label the speech data and classify each word boundary as belonging to one of the classes: B / BB / NB. This step will provide a new set of training examples for B / BB / NB classes.
4. Update / retrain the CART model with the new set of training examples.
5. Repeat steps 3 and 4 for 1-2 iterations.

3.3. Evaluation of ADP model

To evaluate an ADP model, the location of predicted phrase breaks could be compared with manually identified phrase breaks, and the accuracy of an ADP model could be reported in terms of precision and recall. However, such evaluation criteria would limit the purpose of building an ADP model for languages and speech databases which may not have such hand labeling done. An alternate method of evaluation is to incorporate the prosodic phrase breaks predicted by an ADP model in a text-to-speech system, and perform subjective and objective evaluations of synthesized speech to know whether the acoustic phrasing has provided any improvement to the quality of synthesized speech. To perform this evaluation, statistical parametric synthesis such as CLUSTERGEN [8] and HTS [9] is a better platform than unit selection synthesis, as the effect of phrase break dependent features such as duration are directly evident in statistical parametric synthesis. CLUSTERGEN is a statistical parametric synthesizer which predicts duration and F0 for each phone from the input text. Spectral parameters are generated for each phone based on its duration value and synthesis of the speech is performed using spectral parameters and voiced / unvoiced excitation based on F0 values.

The process followed to incorporate and evaluate the effectiveness of an ADP model in CLUSTERGEN is as follows:

- From ADP model, obtain the location of prosodic phrase break in the text of all utterances. We have used three levels of phrase breaks ($NB/B/BB$) in this work.
- Divide this annotated text into training set (R-set) and held out test set (T-set).
- Use R-set for building the synthesizer as done in CLUSTERGEN. The build process of CLUSTERGEN is modified to incorporate phrase break as one of the features in the clustering process.
- Synthesize utterances from T-set and perform an objective evaluation in comparison with original utterances as spoken by the native speaker. The process of objective evaluation computes spectral distortion between the original and synthesized utterance. However, due to variations in the durations of original and synthesized utterances, they are aligned first using dynamic programming and Mel-Cepstral Distortion (MCD) is computed between the aligned frames. The MCD measure between two Mel-cepstral vectors is defined as $MCD = (10/\ln 10) * \sqrt{2 * \sum_{i=1}^{25} (mc_i^t - mc_i^e)^2}$, where mc_i^t and mc_i^e denote the original and the synthesized Mel-Cepstra respectively. Lesser the MCD better is the synthesis quality. MCD is calculated over all the Mel-Cepstral coefficients, including the zeroth coefficient.
- Build the phone duration model using R-set and report the accuracy of the prediction model in terms of Root Mean Square Error (RMSE) and correlation values.

Table 2: Objective evaluation of IIIT-LEN voice using ADP. MCD scores indicate spectral distortion of original and synthesized speech and are measured in dB. The RMSE values indicate the performance of phone duration model. The Corr. values also indicate the performance of phone duration model.

	Baseline-0	Baseline-1	ADP-0	ADP-1
MCD	7.174	7.176	5.736	5.655
RMSE	0.783	0.783	0.775	0.769
Corr.	0.622	0.622	0.631	0.639

4. Experiments on Telugu Speech Database

The Telugu database referred to as (IIIT-LEN) used in this work is collected from a female native speaker of Telugu. A set of 3150 utterances is used as training set (R-set), and a set of 350 utterances is used as T-set. The duration of R-set is 8 hours and 24 minutes while the duration of T-set is 58 minutes. IIIT-LEN is a read speech and consists of utterances extracted from news bulletin. Each utterance could be as long as 15 words. As discussed in Section 3.1, ADP-0 and ADP-1 models were built for IIIT-LEN voice. As described in Section 3.3, ADP models were incorporated to build CLUSTERGEN voice for IIIT-LEN and the performance of IIIT-LEN voice was evaluated on T-set using MCD as shown in Table 2. In Table 2, *Baseline-0* refers to IIIT-LEN voice generated using default settings in CLUSTERGEN, and *Baseline-1* refers to IIIT-LEN voice where phrase breaks predicted based on punctuation marks (as obtained in the original text) are incorporated in build process of CLUSTERGEN. From Table 2, it can observed that the MCD scores of ADP-0 / ADP-1 performs significantly better than that of Baseline-1 and Baseline-0 suggesting that incorporation of acoustic driven phrasing improves the quality of synthetic speech. Informal listening experiments conducted on ADP-0 / ADP-1, showed that the synthesized speech has prosodic phrase breaks which has improved the perceptual as well as objective measures with respect to Baseline-0 and Baseline-1. The RMSE and correlation values shown in Table 2 also suggest that ADP-0 / ADP-1 performs better than Baseline-0 / Baseline-1 systems. From Table 2, we can also observe that ADP-1 (generated by bootstrapping from ADP-0) performs better than ADP-0.

In addition to objective evaluation, a subjective evaluation was also conducted where the native speakers of Telugu were asked to listen to an utterance synthesized from TTS voices using Baseline-0 and ADP-1. The subject was asked to state whether he / she preferred a particular voice or had no preference. A total of 6 subjects participated in the listening test, thus providing a set of 60 data points on 10 utterances. Table 2 summarizes the subjective listening test, and it could be observed that TTS voice built using ADP-1 was preferred for 43% of utterances and the Baseline-0 voice was preferred for only 8% of utterances.

Table 3: Subjective evaluation of IIIT-LEN voice using ADP.

	Baseline-0	ADP-1	No-preference
Baseline-0 vs ADP-1	5 / 60	26 / 60	29 / 60

5. Conclusion

In this work, we have developed a semi-supervised learning for building models for acoustic driven phrasing. The process consists of generating initial hypothesis of phrase break regions using silence as acoustic cue. Given these initial hypothesis, the bootstrapping was done on the unlabeled (rest of data) by extracting duration, F0 and energy features from the hypothesized phrase break regions. We have also shown that the incorporation of acoustic driven phrasing improves the quality text-to-speech systems. As the proposed technique do not make use of any hand labeled data or require language specific knowledge, it could be useful to build prosodic phrase models for minority and under resourced languages. We have also noted that a prosodic phrase break prediction (PBP) from text make use of either syntactic level information or POS tag sequences as features extracted from the text data, and the availability of such linguistic resources may be difficult for minority or resource poor languages. Our future work is related to development of unsupervised technique of extracting features from the text data which could be used as input for PBP.

6. References

[1] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1707–17, 1992.

[2] L. Redi and S. Shattuck-Hufnagel, "Variation in realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407–429, 2001.

[3] H. Kim, T. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of L- and L-L% in switchboard and radio news speech," in *Proceedings of Speech Prosody*, Dresden, 2006.

[4] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[5] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical and syntactic evidence for automatic unsupervised prosody labeling," in *Proceedings of INTERSPEECH*, Pittsburgh, USA., 2006.

[6] S. Ananthhakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical and syntactic evidence," *IEEE Transactions on Audio, Speech and Language*, vol. 16, no. 1, pp. 216–228, 2008.

[7] K. Prahallad, V. R. Elluru, and A. W. Black, "Significance of acoustic driven phrase breaks in Text-to-Speech systems," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (submitted)., 2010.

[8] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proceedings of INTERSPEECH*, Pittsburgh, USA, 2006.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, 2007.