Learning Speaker-Specific 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

The objective of this paper is to investigate whether prosodic phrase breaks are specific to a speaker, and if so, propose a mechanism of learning speaker-specific phrase breaks from the speech database. Another equally important aspect dealt in this work is to demonstrate the usefulness of these speaker-specific phrase breaks for a text-to-speech system. Experiments are carried out on two different English voices as well as on a Telugu voice, and it is shown that speaker-specific phrase breaks improves duration as well as spectral quality of synthetic speech.
Index Terms: speech synthesis, speaker-specific phrase breaks, semi-supervised learning

1. Prosodic Phrase Breaks

In the context of TTS, it is essential to predict prosodic phrase breaks in the text [1] [2]. Prosodic phrase breaks predicted from the text are used by different modules such as F0 generation, duration and insertion of pauses. Modeling prosodic phrase patterns involves building a prosodic phrase break annotator (PBA) and a prosodic phrase break predictor (PBP). A PBA model annotates text/speech data with the location of prosodic phrase breaks. Often human operators act as PBAs - the annotation is done by listening to speech data. This could also be achieved by machine learning techniques, which use acoustic cues to locate prosodic phrase breaks. A PBP model predicts prosodic phrase breaks in the given text based on either a set of rules or machine learning techniques. The output of PBA - text data annotated with prosodic phrase breaks - is used to train a PBP model. Features related to syntactic level or part-of-speech sequence are extracted from text, and a machine learning model is built to predict a break or not-a-break between words.

Current techniques of modeling phrasing patterns – such as [2], suffer from the following limitations –

- A human annotator is used to annotate text with a break symbol between words which are perceived as being phrase breaks. This process of hand annotation is laborious, time consuming and is not scalable to multiple languages.
- Typically, a PBP model is trained on a standard corpus. For example, in Festival, a default PBP model for English is trained on Boston University Radio News corpus data and employed to predict breaks for all English voices. Thus the same prosodic phrasing pattern is used for all voices ignoring speaker-specific phrasing patterns.
- A PBP model assumes availability of syntactic parsers and/or part-of-speech taggers. The availability of such linguistic resources may be difficult for minority or resource poor languages. Such situations need solutions which extract a new set of features from the text. For example, these features could be based on frequency count of words. Typically, words with very high frequency count are functions words, and an unsupervised clustering of words can be done based on frequency counts. This leads to representation of words as a sequence of cluster numbers similar to part-of-speech sequence.

In the scope of this paper, the objective is to build a PBA model using machine learning techniques which make use of acoustic cues to locate prosodic phrase breaks. Such techniques make annotation faster and cheaper. At the same time, the ability to model phrasing patterns in a given speech database could bring in speaker-specific phrasing patterns. As a part of this investigation, we would like to know whether prosodic phrase breaks are specific to a speaker, and if so, propose a mechanism for learning speaker-specific phrase breaks. Another equally important aspect dealt with in this paper is to demonstrate the usefulness of these speaker-specific phrase breaks for a TTS system. Experiments are carried out on two different English voices as well as on a Telugu voice, and it is shown that speaker-specific phrase breaks improves duration as well as spectral quality of synthetic speech.

2. Are Prosodic Phrase Breaks Speaker-Specific?

In order to examine the correlation between syntactic phrase breaks and prosodic phrase breaks, an experiment was conducted as follows. A short story, from Emma by Jane Austin, Volume 1, Chapter1, spoken by four speakers – Sibella, Sherry, Moira and Elizabeth, from Librivox (www.librivox.org) was considered. This short story consisted of 54 paragraphs and around 3000 words. For every word in the story a binary feature was derived indicating whether there was a break or not after the word. The presence of a break was indicated by 1 and the absence of a break was indicated by -1.

Let S denote the sequence of features derived for the words in the short story using syntactic phrase breaks. These breaks were derived using the Stanford Parser [3] which parses the text in the form of a tree. The phrase breaks were assigned based on the noun phrase, verb phrase, adjective phrases, etc. Let R denote the sequence of features derived for the words in the short story using prosodic phrase breaks. These breaks were derived based on the duration of the pause after the word. A pause was considered as a prosodic phrase break, when its duration is greater than 150ms. Since the story was spoken by four different speakers, we derived the feature vectors R_s, R_h, R_m and R_e representing the prosodic phrase break sequences for Sibella, Sherry, Moira and Elizabeth respectively. The correlation coefficient between S and R was computed and is as

shown in Table 1.

From Table 1, we can observe that correlation coefficient between syntactic phrase breaks and prosodic phrase breaks varies from 0.26 to 0.33. These lower values indicate that the syntactic phrase breaks and prosodic phrase breaks differ significantly. From Table 1, it could also be observed that the correlation coefficient between the prosodic phrase breaks of any two speakers varies between 0.65 to 0.75. These values indicate that the correlation coefficient between the prosodic phrase breaks of any two speakers is higher than the correlation coefficient between the prosodic phrase break and the syntactic phrase break. At the same time, the correlation coefficient between the prosodic phrase breaks of any two speakers is lesser than 1, thus suggesting that the prosodic phrase breaks could be specific to a speaker. Thus we refer to the prosodic phrase breaks as speaker-specific phrase breaks. In the context of a typical text-to-speech system, we deal with speech data from a single speaker, and it is appropriate to learn speaker-specific phrase breaks from the given speech database and predict speaker-specific phrase breaks in the text during synthesis time.

Table 1: *Correlation between syntactic phrase breaks and prosodic phrase breaks of different speakers.*

	Syntactic	Elizabeth	Moira	Sherry	Sibella
Syntactic	1	0.29	0.30	0.23	0.31
Elizabeth		1	0.66	0.61	0.72
Moira			1	0.58	0.69
Sherry				1	0.62
Sibella					1

3. Cues Characterizing Phrase Breaks

Speaker specific phrase breaks are manifested in the speech signal in the form of pauses as well as relative changes in the intonation, duration of rhyme, glottalization etc. In order to illustrate the complex nature of the acoustic cues that indicate prosodic phrase breaks, a 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.

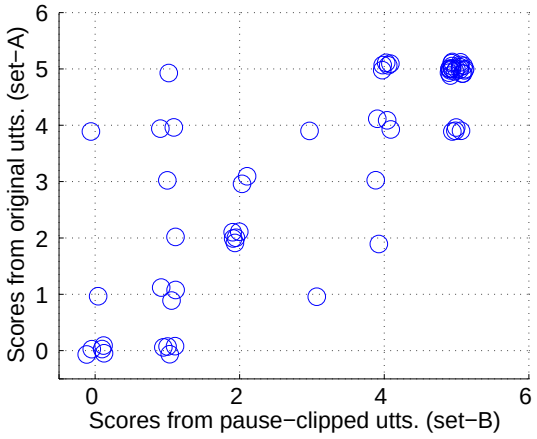

Figure 1: Scatter plot of scores obtained for utterances in Set-A and Set-B

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 [4] [5] [6], the representation / parameterization of these complex acoustic cues is not well understood. Many of these

Table 2: Syllable level features extracted at phrase break, adapted from [7].

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

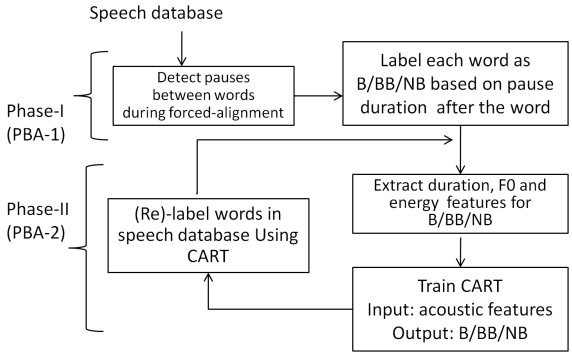

Figure 2: Flow chart of the proposed algorithm for learning speaker-specific phrase breaks.

complex acoustic cues are often represented by extraction of average duration, F0 and energy values [7]. 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 2.

4. Learning Speaker-Specific Phrase Breaks

To learn speaker-specific phrase breaks, we propose an unsupervised learning algorithm. Fig. 2 provides a schematic diagram of the proposed algorithm consisting of two phases.

In the first phase, 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. In the second phase, features such as duration, F0 and energy are extracted from these labeled regions and a machine learning algorithm is used to build to classify these acoustic features as belonging to the class of a phrase break or not a phrase break. We then attempt to bootstrap this machine learning model using unlabeled data (i.e., the rest of the data).

4.1. Phase 1: Using Pause as an Acoustic Cue

In phase-1, we hypothesize the phrase break regions based on pauses in speech signal. This phase is referred as to building a Phrase Break Annotator (PBA-1), and the steps involved in building PBA-1 is 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 $p_d > 150\ ms$.

4.2. Phase 2: Bootstrapping

In phase-2, we built a Prosodic Break Annotator (PBA-2), based on the phrase breaks regions identified by PBA-1. The steps involved in building the PBA-2 by the process of bootstrapping on phrase break regions identified by PBA-1 is as follows.

1. Extract duration, F0 and energy features from the phrase regions as identified by PBA-1 in Section 4.1. 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 2, and is based on the work in [7].
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.

4.3. Evaluation of PBA Models

To evaluate a PBA model, the location of predicted phrase breaks could be compared with manually identified phrase breaks, and the accuracy of a PBA model could be reported in terms of precision and recall. However, such evaluation criteria would limit the purpose of building a PBA 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 a PBA 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 a PBA model in CLUSTERGEN is as follows:

- From PBA model, obtain the location of prosodic phrase breaks in the speech signal. This process would classify each word boundary as ($B/BB/NB$) based on acoustic features. As a result, the text of all utterances is annotated with break markers. We used special punctuation symbols to denote B/BB in the text.
- Divide this annotated text into training set (T-set) and held out test set (H-set).
- Use T-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 H-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 T-set and report the accuracy of the prediction model on H-set in terms of z-scores of Root Mean Square Error (MSE).

5. Results and Discussion

To evaluate the usefulness of speaker-specific phrase breaks, single sentence recordings such as CMU ARCTIC may not be useful. The utterances in CMU ARCTIC database are short and often may not have any phrase break with in an utterance Hence we considered audio books which are multi-sentence and of story telling type. In this work, we have used the audio books of EMMA and WALDEN from Librivox.

The recordings of EMMA by Jane Austen are done by a female speaker. The duration of this audio book is 17.34 hours. We downloaded the associated text from Project Gutenberg, and the text was arranged into 2693 utterances, where each utterance is typically a multi-sentence paragraph. Given that the audio books have large audio files, suitable modifications were done to standard forced-alignment algorithm to obtain segment and word level boundaries for each utterance [10]. A set of utterances were used to build speaker-specific PBA-1 and PBA-2 models. The duration of the utterances used for training (T-set) is 15.67 hours and the duration of the utterances in held-out test set (H-set) is 1.66 hours.

The recordings of WALDEN are done by a male speaker. The duration of this audio book is around 14 hours, and the text was arranged into 1260 utterances. The duration of T-set used for training speaker-specific PBA-1 and PBA-2 is 12.72 and the duration of H-set used for testing is around 1.45 hours.

The Telugu database referred to as (IIIT-LEN) used in this work is collected from a female native speaker of Telugu. This database consisted of 3400 utterances. The duration of T-set is 8 hours and 24 minutes while the duration of H-set is 58 minutes.

As discussed in Section 4.1, PBA-1 and PBA-2 models were built for EMMA, WALDEN and IIIT-LEN speech databases. As described in Section 4.3, PBA models were incorporated to build CLUSTERGEN voices for EMMA, WALDEN and IIIT-LEN and the performance of these voices evaluated on their respective H-sets using MCD is as shown in Table 3. In Table 3, *Baseline* refers to CLUSTERGEN voices generated using default settings in CLUSTERGEN.

Table 3: Objective evaluation of synthetic voices using PBA. MCD scores indicate spectral distortion of original and synthesized speech and are measured in dB. The MSE values indicate the performance of phone duration model measured in terms of z-scores.

	EMMA		WALDEN		IIIT-LEN	
	MCD	MSE	MCD	MSE	MCD	MSE
Baseline	5.55	0.848	5.40	0.891	7.17	0.783
PBA-1	5.43	0.847	5.12	0.902	5.73	0.775
PBA-2	5.36	0.845	5.09	0.877	5.65	0.769

From Table 3, it can observed that the MCD scores of PBA-1 / PBA-2 performs better than that of the Baseline suggesting that the incorporation of speaker-specific phrase breaks improves the quality of synthetic speech. Informal listening experiments conducted on PBA-1 / PBA-2, showed that the synthesized speech has prosodic phrase breaks which has improved the perceptual as well as objective measures with respect to Baseline. The RMSE values shown in Table 3 also suggest that PBA-2 performs better than the Baseline system. From Table 3, we can also observe that PBA-2 (generated by bootstrapping from PBA-1) performs better than PBA-1.

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 and PBA-2. 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 4 summarizes the subjective listening test, and it could be observed that TTS voice built using PBA-2 was preferred for 43% of utterances and the Baseline voice was preferred for only 8% of utterances.

Table 4: Subjective evaluation of IIIT-LEN voice.

	Baseline	PBA-2	No-preference
Baseline vs PBA-2	5 / 60	26 / 60	29 / 60

6. Conclusion

In this work, we have shown that the prosodic phrase breaks differ from syntactic phrase breaks, and the prosodic phrase breaks are specific to a speaker. We have proposed a two phase algorithm to learn speaker-specific phrase breaks and demonstrated that the incorporation of these speaker-specific phrase breaks improves the quality of synthetic speech. In the scope of this paper, we have dealt with automatic building and evaluation of PBA models. By evaluating PBA models in the TTS framework on held-out test sets, we have essentially assumed a perfect PBP model (i.e., prediction of pauses from text). This was done primarily to highlight the significance of speaker-specific

phrase breaks for TTS systems and avoid any prediction errors that may arise from a PBP model trained using machine learning techniques. A potential future work is to demonstrate the usefulness of a PBP model trained on speaker-specific phrase breaks in comparison with a PBP model trained on a standard corpus as in [2].

7. References

[1] J. Bachenko and E. Fitzpatrick, "A computational grammar of discourse-neutral prosodic phrasing of English," *Computational Linguistics*, vol. 16, no. 3, pp. 155–170, 1990.

[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] D. Klein, D. Christopher, and D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[4] 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.

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

[6] 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.

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

[8] A. 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. Black, and K. Tokuda, "The hmm-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, 2007.

[10] K. Prahallad and A. W. Black, "Handling large audio files in audio books for building synthetic voices," in *SSW7 Workshop*, (submitted), 2010.