ISCA Archive
http://www.isca-speech.org/archive
8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Significance of word-terminal syllables for prediction of phrase breaks in Text-to-Speech systems for Indian languages

Anandaswarup Vadapalli[1], *Peri Bhaskararao*[1], *Kishore Prahallad*[1]

[1] Speech and Vision Lab, IIIT Hyderabad, India

anandaswarup.vadapalli@research.iiit.ac.in, bha.peri@iiit.ac.in, kishore@iiit.ac.in

Abstract

Phrase break prediction is very important for speech synthesis. Traditional methods of phrase break prediction have used linguistic resources like part-of-speech (POS) sequence information for modeling these breaks. In the context of Indian languages, we propose to look at syllable level features and explore the use of word-terminal syllables to model phrase breaks. We hypothesize that these terminal syllables serve to discriminate words based on syntactic meaning, and can therefore be used to model phrase breaks. We utilize these terminal syllables in building models for automatic phrase break prediction from text and demonstrate by means of objective and subjective measures that these models perform as well as traditional models using POS sequence information. Thus the proposed method avoids the need for POS taggers for prosodic phrasing in Indian languages.

Index Terms: Phrase Breaks, Word-Terminal Syllables, Text-to-Speech

1. Introduction

Phrase break prediction plays an important role in the context of speech synthesis. It is known that phrase breaks have a non-linear relationship with syntactic breaks [1]. It is also known that phrase breaks are specific to a speaker [2] [3].

Phrase breaks are manifested in the speech signal in the form of several acoustic cues like pauses as well as relative changes in the intonation and duration of syllables. Acoustic cues such as pre pausal lengthening of rhyme, speaking rate, breaths, boundary tones and glottalization also play a role in indicating phrase breaks in speech [4], [5], [6]. However, representing these non pause acoustic cues in terms of features is not easy and not well understood [2]. In this paper we restrict ourselves only to pauses in speech, and limit our phrase break models to predicting the locations of pauses while synthesizing speech. This is the approach followed in [7] and [8].

Traditional methods of phrase break prediction have used linguistic resources like part-of-speech (POS) sequences generated by POS taggers or shallow parsers to model phrase breaks. Many different machine learning algorithms have been applied to phrase break prediction; for example, decision trees [9], [10]; n-gram models [1], [11]; finite state transducers [12] and memory based learning [13]. However, regardless of the machine learning technique used, the primary feature used by the classifier has been POS tags.

In all the approaches mentioned above, POS tags are directly used as input into the phrase break classifier. In Parlikar and Black [7] they are used to construct grammar based parse trees, which in turn provides features for a decision tree based phrase break predictor. Previous work therefore suggests that POS tagging is a necessary first step in phrase break prediction.

All these traditional methods assume the availability of hand labeled training data, or high quality POS taggers/shallow parsers which can generate POS tags for the training data with a high level of accuracy. As a result, these methods can not be used for languages where the necessary linguistic resources are not readily available, and manual annotation of data is expensive and time consuming.

In view of the above limitations, there has recently been a lot of interest in unsupervised methods of inducing word representations which can be used as surrogates for POS tags, in the phrase break prediction task. Parlikar and Black [8] used the Ney-Essen clustering algorithm [14] to automatically induce POS tags. These induced POS tags are automatically generated from text using the frequency analysis of the words. However this approach faces an issue when applied to Indian languages, which are agglutinative in nature. In these languages words are formed by joining morphemes together. Moreover due to the postpositional nature of these languages, syllable level suffixes get attached to the ending of words. These suffixes give specific syntactic meaning in terms of tense, gender etc. These characteristics of Indian languages result in an increase in vocabulary size, i. e. the number of words. Thus it is hard to work with scripts of Indian languages using a word level representation.

A better solution may lie in dealing with sub-word units like syllables or multi-syllable units [15], [16], [17]. In [15], a set of *morpheme tags* units were manually identified and used to model phrase breaks. The *morpheme tags* consist of one or two syllables, typically found at the end of the word. The experiments were conducted on Telugu. Manual identification of this set of morpheme tags is hard and may require sufficient linguistic knowledge.

In the current work, we propose to look at syllable level units, and explore the terminal syllables of a word to model phrase breaks. A terminal syllable is the last syllable in the word. We hypothesize that these terminal syllables serve to discriminate words based on syntactic meaning, and that these terminal syllables can be used to model phrase breaks. We automatically identify the set of terminal syllables which could be used to model phrase breaks. We experiment our approach with six Indian languages, and report results on the automatic prediction of phrase breaks from the text using these terminal syllables. Finally, we incorporate the proposed phrase break model in a Text-to-Speech system, and demonstrate its usefulness with listening tests.

2. Database used in this study

In our study we look at two different corpora that have speech in different styles.

The ***IIIT-MCIT (Lenina)*** corpus is a corpus developed at IIIT Hyderabad, which was used to build a synthetic voice in Telugu. The corpus consists of text prompts taken from a set of popular children's stories in Telugu, and the corresponding recordings recorded in a story telling style in a clean studio environment. The corpus has 4043 text prompts and the audio size is about 6 hours. The style of the corpus is "story telling".

The ***IIIT-H Indic*** [18] database consists of text and speech data in Telugu, Hindi, Kannada, Tamil, Malayalam, Marathi and Bengali. Each language in the database consists of a set of 1000 text prompts selected from Wikipedia articles in the corresponding language, selected in such a way as to cover the 5000 most frequent words in the corresponding languages. The corresponding recordings were recorded by a native speaker of each language in a clean studio environment. On an average the size of the audio is about 1.5 hours for each language. The style of this corpus is "isolated sentences".

2.1. Annotation of phrase breaks

As we do not have a corpus with hand annotated phrase breaks, we derive the location and duration of the phrase breaks from the speech data. In order to derive the locations and durations of pauses introduced by the speaker, we force align the speech with the corresponding text prompts using the HMM tool in Festvox [19]. This gives us the locations of pauses introduced by the speaker while recording the utterances.

A question is what should be the duration of a pause to consider it as a phrase break? To answer the above question, we analyzed the durations of the pauses introduced by the speaker for both the *Lenina* and *IIITH-Indic* databases. Figure 1 shows the histogram plots of silence durations for all the six languages.

An analysis of the histogram plot shows that for Hindi the majority of the pauses are less than 80 ms in duration, for Telugu (IIITH-Indic) the majority of the pauses range from a few milliseconds to 480 ms, for Kannada and Tamil most of the pause durations range from a few milliseconds to 480 ms and for Bengali the pause durations range from a few milliseconds to 640 ms. In the case of the *Lenina* database, the pause durations range mainly from 80 ms to 640 ms.

We thus observe that the pause durations vary over a significant range within a language and also between languages. We experiment with different thresholds above which a pause is marked as a phrase break. We experiment with thresholds of 25 ms, 50 ms and 80 ms, whereby we mark all pauses with durations greater than the threshold as phrase breaks. We also experiment with the case where we mark all pauses as phrase breaks regardless of their duration.

3. Phrase Breaks vs. Syntactic Breaks

A question that is often asked is whether there is any relation between phrase breaks and syntactic breaks. While it is known that there is some correspondence between syntax and prosody, the relationship between them is not formally defined [1], [3]. We illustrate this by means of two examples.

Consider the Telugu sentence (represented in ITRANS transliteration scheme) shown in Table 1, taken from the *Lenina Database*, which has been annotated with the location of prosodic and syntactic breaks.

In a similar fashion consider the Hindi sentence shown in Table 2 taken from the *IIITH-Indic* database, which has also been annotated with the location of phrase and syntactic breaks.

From the above examples it is clear that while there is some correspondence between the phrase and syntactic breaks of an utterance, the relationship between them is not linear.

3.1. Syntactic breaks used in the study

Syntactic breaks were derived from text using the shallow parser [20] developed at IIIT Hyderabad. This tool uses conditional random fields (CRF) and transformational based learning (TBL) to perform chunking and POS tagging of text. In [20] the authors report accuracies of 77.37%, 78.66% and 76.08% for the chunking task and 79.15%, 80.97% and 83.74% for the POS tagging task, for the three languages Telugu, Hindi and Bengali respectively.

The location of syntactic breaks was derived by running the shallow parser on the text data. The tool parsed the text into syntactic constituents, and the end of each constituent was taken as a syntactic break.

3.2. Correlation between Phrase breaks and Syntactic breaks

In order to calculate the correlation between phrase and syntactic breaks, we conducted the following experiment for all languages under consideration: Telugu, Hindi, Kannada, Tamil and Bengali. For every word in each language, a binary feature which indicates the presence or absence of a break after that word, was derived. The presence of a break was indicated by 1 and the absence of a break by -1.

Let $\mathbf{S} = [s_1, ..., s_w, ..., s_N]$ denote the sequence of binary features derived for the words in the database using syntactic break information, where N denotes the total number of words in the database.

Let $\mathbf{P} = [p_1, ..., p_w, ..., p_N]$ denote the sequence of binary features derived using phrase breaks (pauses in speech), where N denotes the total number of words in the database. These breaks were derived from phrase break annotation described in 2.1.

The correlation coefficient between **S** and **P** is calculated using the following equation.

$$c(\mathbf{S}, \mathbf{P}) = \frac{\sum_{w=1}^{N}(s_w - \bar{s})(p_w - \bar{p})}{\sqrt{\sum_{w=1}^{N}(s_w - \bar{s})^2}\sqrt{\sum_{w=1}^{N}(p_w - \bar{p})^2}}$$

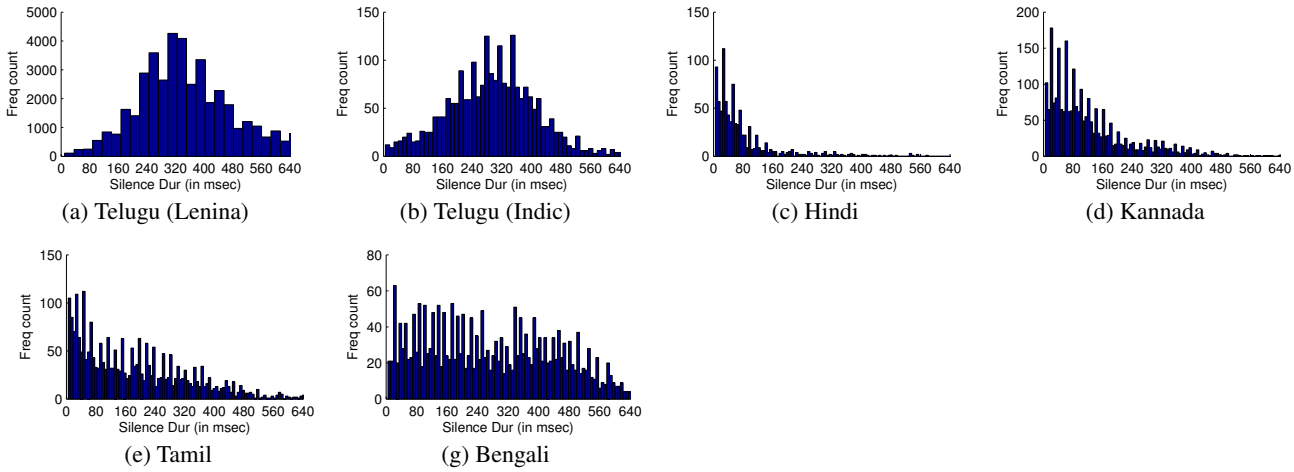
where $\bar{s}$ and $\bar{p}$ denote the mean values of **S** and **P** respectively.

Table 3 shows the correlation coefficients between syntactic and phrase breaks for the six languages.

Language	Correlation Coefficient
Telugu (Lenina)	0.26
Telugu (Indic)	0.27
Hindi	0.12
Kannada	0.29
Tamil	0.18
Bengali	0.20

Table 3: Correlation coefficients between syntactic breaks and phrase breaks for the six languages

An observation of the values in Table 3 shows that the values of correlation coefficients between syntactic breaks and phrase breaks, for all the languages, does not exceed 0.3. This indicates that there is a significant variation between syntactic and phrase breaks, in all the languages under consideration.



(a) Telugu (Lenina) (b) Telugu (Indic) (c) Hindi (d) Kannada

(e) Tamil (g) Bengali

Figure 1: Histograms of silence durations in all 6 languages.

"brahmadattud:u #B(260ms) kaashiiraajyaanni paripaalin:chei kaalan:loo #B(440ms) aa nagaran:loo #B(410ms) dhanikud:aina #B(360ms) oka goppavartakud:un:d:eivaad:u"
"brahmadattud:u | kaashiiraajyaanni | paripaalin:chei | kaalan:loo | aa nagaran:loo | dhanikud:aina | oka goppavartakud:un:d:eivaad:u"

Table 1: An example sentence in Telugu annotated with locations of phrase and syntactic breaks (the word-terminal syllables have been underlined) where *#B* denotes a phrase break and the numerical value in brackets denotes the break duration in milliseconds and | denotes a syntactic break

"san:bhava hai ki #B(80ms) isakaa #B(65ms) aavishhkaara #B(10ms) isasei bhii bahuta pahalei huaa hoo"
"san:bhava | hai | ki | isakaa aavishhkaara | isasei bhii | bahuta pahalei | huaa hoo"

Table 2: An example sentence in Hindi annotated with locations of phrase and syntactic breaks where *#B* denotes a phrase break and the numerical value in brackets denotes the break duration in milliseconds and | denotes a syntactic break

4. Correlation between word-terminal syllables and breaks

In order to model phrase breaks, we look at syllable level features. An examination of the Telugu example from Section 3 shows that a phrase break has occurred after two words ending in the syllable *loo*, and after a word ending in the syllable *d:u*. The utterance ending break has also occurred after a word ending in the syllable *d:u*. This motivated us to look at word-terminal syllables as a feature set which can be used to model phrase breaks. We performed an analysis, whereby the correlation between the word-terminal syllables and phrase breaks was studied. As part of this analysis we computed the conditional probability $p(\text{break} \mid \text{terminal syllable})$ as follows

$$p(\text{break} \mid \text{terminal syllable}) = \frac{N(\text{break, terminal syllable})}{N(\text{terminal syllable})},$$
$$\forall N(\text{terminal syllable}) > 50$$

Our analysis, showed that for a few terminal syllables, the probability of a word ending in that terminal syllable, preceding a phrase break, is high. That is, for a few terminal syllables the value of the conditional probability $p(\text{phrase break} \mid \text{terminal-syllable})$ is high. The value of this conditional probability tapers off beyond these top few terminal syllables. Table 4 shows the top terminal syllables, derived from this analysis for each of the six languages. The values in the parentheses are the values of $p(\text{phrase break} \mid \text{terminal-syllable})$ for those particular syllables.

As can be observed from the Table 4 we can see that the value of $p(\text{phrase break} \mid \text{terminal-syllable})$ for the top syllables are in the range 0.6 - 1.0. As a result, we hypothesize that these word terminal syllables are good candidates for a feature set to predict phrase breaks from text.

5. Prediction of phrase breaks from text

Prediction of prosodic phrase breaks from text can be achieved by building a phrasing model. Typically as a first approach a punctuation based phrasing model is used. The output of this model is then refined by using models built using POS tags and other linguistic information. However, text in Indian languages very rarely has any punctuations (except for sentence endings). Hence when dealing with Indian languages a simple punctuation based phrasing model will not work and more sophisticated phrasing models are required. These model can either be a set of heuristic rules or a machine learning model trained on features extracted from the text. Generally, the first step in building such phrasing models involves annotating the text with phrase breaks, which has been described in Section 2.1. This annotated text can be used in several ways. The text can be used to derive a set of heuristic rules, which can be used to derive the location of phrase breaks in the text. We can also extract several features from this text to train a machine learning model, which can be

Telugu(Lenina)	nai(0.98), buu(0.97), jaa(0.92), chchu(0.90), du(0.87), vaa(0.87), yyaa(0.85), daa(0.84), stei(0.83), yi(0.80), tei(0.80), di(0.79), chchi(0.79), t:ei(0.79), mmaa(0.78), d:u(0.78), llaa(0.75), chii(0.74), ppi(0.73), chi(0.73), ...
Telugu (Indic)	d:i(0.84), mu(0.82), loo(0.78), yi(0.77), du(0.76), di(0.70), nu(0.67), san:(0.65), llaa(0.64), d:u(0.63), vu(0.61), ru(0.58), d:aa(0.57), ki(0.54), ni(0.54), gaa(0.53), lu(0.53), ku(0.51), ran:(0.50), chi(0.48), ...
Hindi	hai(0.75), hain:(0.74), thaa(0.64), sha(0.43), da(0.40), ei(0.34), koo(0.26), yaa(0.24), nd~a(0.24), la(0.23), pa(0.23), kaa(0.22), ga(0.22), va(0.21), sa(0.21), na(0.20), ra(0.19), ti(0.18), kha(0.17), bhii(0.17), ...
Kannada	de(0.97), ki(0.93), lli(0.69), ru(0.67), re(0.65), nnu(0.62), l:u(0.61), gi(0.57), ge(0.57), da(0.54), ya(0.54), du(0.52), na(0.51), tra(0.50), ti(0.43), ga(0.43), vu(0.42), ttu(0.42), ka(0.41), ra(0.40), ...
Tamil	kum(1.00), n~ar(0.99), llai(0.96), lam(0.94), r:r:i(0.77), chan~(0.73), ng~kal:(0.72), thu(0.69), rai(0.67), than~(0.61), kal:(0.57), rkal:(0.54), ntha(0.53), n~r:u(0.48), yil(0.47), ththil(0.46), ththu(0.35), ya(0.35), ka(0.32), kku(0.32), ...
Bengali	hay(0.91), chhi(0.81), nya(0.78), sa(0.77), chhe(0.70), da(0.66), ja(0.60), naa(0.59), sha(0.58), ban:(0.57), i(0.57), ba(0.56), be(0.53), ke(0.51), na(0.51), t:a(0.50), re(0.50), nd~a(0.48), le(0.48), ga(0.48), ...

Table 4: Top word-terminal syllables for all the languages. The figure in brackets is $p(\text{phrase break} \mid \text{terminal syllable})$

used to predict phrase breaks from text.

We experiment with three different approaches and report the results of phrase break prediction from text, for both *Lenina* and *IIITH-Indic* databases. In our first approach we derive a simple rule which utilizes the syntactic break location (obtained from the shallow parser (Section 3)) along with terminal syllable information to derive the location of the phrase breaks in text. Our second approach utilizes terminal syllable information, which we extract from the text, to build a machine learning model for phrase break prediction. In our third approach, we use POS tag sequence information, (which we obtain from running the shallow parser over the text (Section 3)) to build a machine learning model for phrase break prediction.

As the text in these databases has already been annotated with prosodic phrase breaks, a ground truth to compute the performance of our approaches is available. We report the performance of our approaches in terms of the F-measure [21] which is defined as the harmonic mean of the precision and recall. F-measure values range from 0 to 1, with higher values indicating better performance.

5.1. Simple rule based phrase break prediction

We derive a simple rule to give us the location of phrase breaks in text. This rule uses syntactic break locations along with the terminal syllable information to derive the locations of the phrase breaks in text.

From our analysis of the correlation between terminal syllables and phrase breaks in Section 4 we observed that the top 50 word terminal syllables in each language have a high correlation of occurrence along with phrase breaks. We combined our knowledge of the syntactic break locations (derived from the shallow parser (Section 3)) with this observation to develop the following heuristic rule for phrase break prediction from text.

"If the word ending of a word in the text has been marked as a syntactic break and the last syllable of the word (the terminal syllable) is among the list of the top 50 terminal syllables for that language (derived from our analysis), then that syntactic break is also a phrase break."

We use this rule to derive the locations of the phrase breaks in text for both the *Lenina* and *IIITH-Indic* databases.

Table 5 displays the F-measures for our heuristic rule based prediction of phrase breaks, for all six languages. An analysis of the numbers shows that, with the exception of Hindi, the rule based system perfroms with F-measures ranging from 0.45 to 0.75.

Language	F-Measure
Telugu(Lenina)	0.62
Telugu(Indic)	0.57
Hindi	0.24
Kannada	0.55
Tamil	0.47
Bengali	0.49

Table 5: F-Measure for rule based prediction of phrase breaks in text

5.2. Phrase break prediction using terminal syllables in a machine learning model

We use the terminal syllable information, as features in a Classification and Regression Tree (CART) framework, to build a model (henceforth referred to as TS model) for predicting phrase breaks from text. As we are using syllable level features in this model, we experiment with different syllable level contexts in order to incorporate contextual information. We use 90% of the text in each language as training data while the remaining 10% was held back for testing. As it is a trivial task to predict breaks at utterance endings, we remove the examples corresponding to utterance ending breaks from the training data.

As an initial experiment, we considered the case where word boundaries that coincide with pauses greater than 80ms are marked as phrase breaks, while all other word boundaries are marked as non breaks. We generated example vectors of both phrase breaks and non breaks, using the terminal syllables along with contextual information. As this was a binary classification task, we also ensured that the number of training vectors of each of the classes (break and non break) were the same. As the number of non breaks were more than the number of breaks in the data, this was achieved by removing examples of non breaks till the total number of example vectors of non breaks and breaks were the same. These example vectors were then used to train the CART model.

We also experiment with different pause thresholds for marking phrase breaks, keeping the context the same in all cases. For the purpose of this experiment we consider the contextual information provided by the previous two syllables and the next two syllables immediately adjacent to the terminal syllable. As before, we take 90% of the text in each language as training data while the remaining 10% was held back for testing and breaks corresponding to utterance endings were omitted from the training data. In this case also, we ensured that the number of training vectors of both classes (breaks and non breaks) were the same.

Language	TS	POS	-1C, TS	-1C, POS	-1C, TS, +1C	-1C, POS, +1C	-2C, TS	-2C, POS	-2C, TS, +1C	-2C, POS, +1C	-2C, TS, +2C	-2C, POS, +2C
Telugu(Lenina)	0.63	0.62	0.69	0.63	0.72	0.69	0.69	0.64	0.72	0.75	0.74	0.75
Telugu(Indic)	0.47	0.56	0.49	0.56	0.53	0.59	0.47	0.52	0.53	0.60	0.52	0.59
Hindi	0.09	0.10	0.09	0.10	0.09	0.09	0.09	0.09	0.09	0.15	0.06	0.18
Kannada	0.37	0.41	0.40	0.42	0.40	0.41	0.39	0.42	0.40	0.44	0.39	0.43
Tamil	0.43	0.39	0.42	0.43	0.39	0.44	0.43	0.44	0.41	0.47	0.42	0.45
Bengali	0.41	0.56	0.34	0.53	0.47	0.53	0.37	0.53	0.49	0.56	0.49	0.54

Table 6: F-Measures for different contexts and setting SSIL $>$ 80ms for phrase breaks, for all six languages where -1C and -2C represents one and two units context respectively to the left and +1C and +2C represents one and two units context respectively to the right

Language	SSIL >50 ms taken as breaks		SSIL >25 ms taken as breaks		all SSIL taken as breaks	
	-2C, TS, +2C	-2C, POS, +2C	-2C, TS, +2C	-2C, POS, +2C	-2C, TS, +2C	-2C, POS, +2C
Telugu(Lenina)	0.72	0.75	0.73	0.75	0.73	0.75
Telugu (Indic)	0.54	0.60	0.54	0.62	0.54	0.62
Hindi	0.37	0.24	0.38	0.28	0.47	0.31
Kannada	0.46	0.51	0.47	0.50	0.55	0.59
Tamil	0.46	0.54	0.55	0.57	0.58	0.61
Bengali	0.51	0.59	0.53	0.59	0.53	0.58

Table 7: F-Measure for different silence thresholds, for all six languages where -2C represents two units context to the left and +2C represents two units context to the right

5.3. Phrase break prediction using POS tag sequence in a machine learning model

We use the same experimental setup as in Section 5.2 changing only the features used. We use the POS tag sequence information, as features in a Classification and Regression Tree (CART) framework, to build a model (henceforth referred to as POS model) for predicting phrase breaks from text. As the POS tag sequence information is a word level feature, we experiment with different word level contexts in order to incorporate contextual information. All the experiments in Section 5.2 are repeated using the POS tag sequence information as features

5.4. Analysis of results

Table 6 shows the performance of both the TS model and POS model, in predicting phrase breaks from text, for all six languages, when word boundaries greater than 80ms are marked as phrase breaks. An observation of the table shows that the performance of both the models for Hindi is poor. In this case, as observed in Section 2.1 the majority of the pause durations are less than 80ms. As as result the number of example vectors for phrase breaks are very few, resulting in a poorly trained model. Also in case of Hindi the sentences are short, and so there are few pauses in the middle of the sentences.

Table 7 shows the prediction accuracy for different pause thresholds for marking phrase breaks, keeping the context same in all cases.

An analysis of the F-measure numbers obtained from all experiments shows that, for automatic prediction of phrase breaks from text, models built using terminal syllables perform nearly as well as models built using traditional features like part-of-speech (POS) sequences.

	% Preference
No Phrasing model	10%
POS model	75%
No Preference	15%

Table 8: AB Test Results for No Phrasing model vs POS model

	% Preference
No Phrasing model	12%
TS model	73%
No Preference	15%

Table 9: AB Test Results for No Phrasing model vs TS model

6. Subjective evaluation of phrasing models

We perform subjective listening tests for Telugu, to compare utterances synthesized by incorporating the TS model and POS model with utterances synthesized with no explicit phrasing model. The listening tests were set up as an ABX task, for native speakers of Telugu. Two phrasing models were compared at a time. An utterance was synthesized by incorporating both models and both versions were presented to the participants in a randomized order, and the participants were asked to mark the version they preferred. They also had an option of no preference if they could not pick one utterance over the other.

In the first listening task, we compared utterances synthesized by incorporating the POS model with utterances synthe-

	% Preference
POS model	35%
TS model	34%
No Preference	31%

Table 10: AB Test Results for POS model vs TS model

sized with no explict phrasing model. For the second listening task, we compared utterances synthesized by incorporating the TS model with utterances synthesized with no explicit phrasing. Finally, we performed a third listening task where we compared utterances synthesized by using the POS model with utterances synthesized using the TS model. All the listening tasks were performed by 10 native speakers of Telugu, who evaluated 15 samples picked randomly from the test set. Tables 8, 9 and 10 show the results of these listening tests.

An examination of the results in Tables 8 and 9 shows that perceptually there is a marked preference for utterances synthesized with both the POS model and the TS model over utterances synthesized with no explicit phrasing. Table 10 shows that when the POS model and the TS model are compared with each other, perceptually there is no significant preference for one model over the other.

7. Conclusions

In this paper we describe phrase break prediction for Text-to-Speech systems in Indian languages. We look at syllable level units and explore the use of terminal syllables to model phrase breaks. We demonstrate the correlation between these terminal syllables and the acoustic breaks found in the speech signal. We also demonstrate that there is a nonlinear relationship between syntax and prosody, and that there are significant variations between syntactic breaks and phrase breaks.

We utilize these terminal syllables in building models for phrase break prediction from text in six Indian languages and demonstrate by means of objective and subjective measures that models built using these terminal syllables perform as well as traditional models built using part-of-speech (POS) sequence information.

The advantage of these terminal syllables, is that they can be directly derived from the text under consideration, thus eliminating the need for additional linguistic resources like shallow parsers or POS taggers, while also eliminating the need to model phrase breaks by computationally expensive unsupervised models.

The samples used for the listening tests are available online at `http://ravi.iiit.ac.in/~speech/SSW8/samples.html`.

In the future we wish to explore the use of Amazon Mechanical turk (MTurk) to conduct the listening evaluations. This would enable us to be able to conduct listening tests with more number of subjects to evaluate the models.

8. Acknowledgements

This work is partially supported by MCIT-TTS consortium project funded by MCIT, Government of India. We gratefully acknowledge the contributions of Prof. Hema Murthy, IIT Madras for the fruitful discussions on the subject. The authors would also like to thank all the volunteers who participated in the perceptual evaluations.

9. References

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

[2] K. Prahallad, E. V. Raghavendra, and A. W. Black, "Learning speaker-specific phrase breaks for text-to-speech systems," in *Proceedings of ISCA Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 2010, pp. 148–153.

[3] K. Prahallad, E. V. Raghavendra, and A. W. Black, "Semi-supervised learning of acoustic driven prosodic phrase breaks for text-to-speech systems," in *Proceedings of 5th International Conference on Speech Prosody (Speech Prosody 2010)*, Chicago, Illinois, May 2010.

[4] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 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] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 2149–2152.

[8] A. Parlikar and A. W. Black, "Data-driven phrasing for speech synthesis in low-resource langauges," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[9] M. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech and Language*, vol. 6, pp. 175–196, 1992.

[10] E. Navas, I. Hernez, and I. Sainz, "Evaluation of automatic break insertion for an agglutinative and inflected language," *Speech Communication*, vol. 50, no. 11-12, pp. 888–899, 2008.

[11] H. Schmid and M. Atterer, "New statistical methods for phrase break prediction," in *Proceedings of 20th international conference on Computational Linguistics,COLING '04*, Stroudsburg, PA, USA, 2004.

[12] A. Bonafonte and P. Agüero, "Phrase break prediction using a finite state transducer," in *Proceedings of 11th International Workshop on Advances in Speech Technology*, 2004.

[13] B. Busser, W. Daelemans, and A. van den Bosch, "Predicting phrase breaks with memory-based learning," in *Proceedings of 4th ISCA Speech Synthesis Workshop*, 2001.

[14] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.

[15] N. S. Krishna and H. A. Murthy, "A new prosodic phrasing model for Indian language Telugu," in *INTERSPEECH-2004-ICSLP*, vol. 1, Oct 6-11 2004, pp. 793–796.

[16] A. Bellur, K. Narayan, K. Krishnan, and H. Murthy, "Prosody modeling for syllable-based concatenative speech synthesis of Hindi and Tamil," in *Proceedings of 2011 National Conference on Communications*, 2011, pp. 1–5.

[17] S. C. Pammi and K. Prahallad, "POS tagging and chunking using decision forests," in *Proceedings of the IJCAI-07 workshop on Shallow Parsing in South Asian Languages*, Hyderabad, India, 2007.

[18] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H Indic Speech Databases," in *Proceedings of Interspeech*, Portland, Oregon, USA, 2012.

[19] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proccedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. 853–856.

[20] P. Avinesh and K. Gali, "Part-of-Speech tagging and chunking using conditional random fields and transformation based learning," in *Proceedings of the IJCAI-07 workshop on Shallow Parsing in South Asian Languages*, 2007.

[21] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.