Pause Prediction from Lexical and Syntax Information

Venkatesh Keri
Language Technologies Research Center,
International Institute of Information Technology,
Hyderabad, India.
venkateshk@students.iiit.net

Sathish Chandra Pammi
Language Technologies Research Center,
International Institute of Information Technology,
Hyderabad, India.
sathishp@students.iiit.net

Kishore Prahallad
Language Technologies Institute,
Carnegie Mellon University, Pittsburgh, USA.
skishore@cs.cmu.edu

Abstract

This paper describes an approach for predicting the pauses in the text utterance which has to be synthesized so as to increase the naturalness of the synthesized voice. We propose that the pause in an utterance depends on both the language syntax and also on the lexical structure of the sentence. Lexical based approach uses sub-word information such as syllable sequence and other related features to predict the pauses. On the other hand syntax based approach uses linguistic information such as part of speech information. Here we will describe some experiments to predict pauses in a sentence based on both lexical and syntactic information by using statistical methods like Conditional Random Fields (CRF) and Classification and Regression Trees (CART). All these experiments were done on the Telugu corpus. Pause prediction performance on this corpus is 83.872% using CART and 84.178 % using CRF. We also provide examples and observations on the improved quality of the text to speech system by using this pause prediction module.

Index Terms- Pause Prediction, Lexical Model and Syntactic model, Text to Speech Systems.

1 Introduction

One of the most important stages in Prosody Modeling for Text to Speech system is Pause Prediction. As punctuations in the text help in understanding the correct meaning of the text utterance, similarly pauses increase the naturalness and intelligibility of the speech utterance. In addition to this most of the other modules of prosody modeling depend on the pause prediction. Previous work to assign the phrase structure to text used machine learning techniques such as Decision Trees, n-gram models, HMM and memory-based learning trained on syntactic features. In this paper, we have used Decision Trees and Conditional Random Fields (CRF) and compared both the methods on lexical, syntactic and hybrid models. If the silence in an utterance is so small that it is not perceivable by the human ear then it can be considered as non-pause. So in our experiments we have taken 150 msec of silence as the minimum threshold for a silence to be a pause.

There are wide areas of applications where the Pause Prediction has its role. In automated systems such as Dialog systems, Mail Readers etc. where intelligibility plays an important role to make the system more natural, pause insertion is one of the standard methods.

The paper is organized in the following way: We start by looking at the previous approaches used to model a pause prediction in Section 2. Next we will discuss the models such as lexical, syntactic and their hybrid to build a pause predictor in Section 3. Section 4 describes the data such as POS tags and sub-word level features used to train the model. Section 5 deals with the performance criteria used to evaluate all the above models, Section 6 deals with the various experiments conducted and the corresponding results and finally Section 7 concludes the paper.

2 Previous Approaches

Many algorithms have been proposed for pause prediction, ranging from simple to complex methods. Simplest of them being a Context-Free

Grammar based system in which rules are framed based on the POS tag information. However, these models cannot be generalized since syntax of language varies vastly from one language to another. Pause prediction in this method uses only local context rather than global context. For example: if the rule is "*pause after a conjunction"* which is true in many cases but there are still some sentences where this is false. Such sentences are shown below.

[Pradeep] and [Sandeep] are going to airport.

[My brother is at home] and [his brother is at college]

In the first example, there is a pause after "and" as it is between two words but in the second example, there is no pause after "and" as it is between two phrases. So this approach is not efficient as it is not reusable for other languages.

So, in order to overcome the issues mentioned above many automatic machine learning algorithms were proposed which can be mainly classified into linguistic and acoustic methods. Linguistic methods are based on modeling the language syntax information such as using part of speech [1] and acoustic methods are based on modeling features such as fundamental contour, accent, duration etc [3]. These acoustic features cannot be used in the initial stage of linguistic processing but only after unit selection. The main disadvantage of these approaches is that they require a huge amount of manually annotated data at linguistic level as well as at acoustic level, which is very costly and time consuming.

3 Basic Model

Some of the previous approaches were linguistic information, acoustic information. But between these two levels we can find lexical level of information. In our experiments we are trying to use linguistic (language syntax) and lexical information. In the process of building these two models another approach is also modeled which is a hybrid of these two models. Pause can be defined as a perceivable silence and depending on the length of duration of silence. There can be many types of pauses but in our experiments we have only two types of pauses i.e. pause if silence is greater then 150 msec and non-pause if it is smaller than 150 msec.

3.1 Lexical Model

This will model the pause prediction on the basis of the lexical structure of the sentence which is strongly bounded to the pronunciation. Lexicon is the unit of spoken language. It may be a phone, syllable etc. We conducted many experiments (discussed in section 6) and came to a conclusion that syllable is the best unit for lexical modeling. Syllable can be represented as C*VC* where C is a consonant and V is vowel. Theoretical proof for using syllable as lexical unit is that it is the only smallest unit which can be pronounced on its own and a pause can occur only after completing a unit but not in between.

So, we formally define the problem as between every pair of words in a sentence there is a *word boundary* which can be of pause type pause or non-pause. However, in principle any number of pause types is possible. The task of the algorithm can be seen as to predict the best sequence of boundary types for a given sentence.

With the above algorithm, we have built the model by using syllable as the basic unit. Feature set for every word boundary are previous syllable, next syllable, number of syllables in previous word, number of syllables in between the present word boundary and the previous pause and the type of boundary that is assigned by fixing a threshold on the duration of the pause (as 150 msec). This is modeled on the basis on local context of pause as the syllable information influences only shorter length of utterance. The advantage of this model over the Syntactic model is features are automatically extracted from the labels generated by the automatic segmentation of speech signal and are discussed in section 4.

Lexical model is trained using Classification and Regression Trees (CART) and Conditional Random Fields (CRF) with the same features mentioned above. In CART the tree is built with all the above features as data for training and the pause type as predictee. In CRF HMM is built where each state represents one of the pause types. In general this can be put into mathematical equation as

$$P(J|F) = P(F|J) * P(J)$$

Where J is the sequence of pauses and F is a trigram of present and previous two sets of features. $P(F|J)$ is the emission probability which

represents the basic probability of each juncture type at each point in the lexical context and *P(J)* a transition probability.

3.2 Syntactic Model

Pause prediction can also be modeled on the basis of language syntax (linguistic information) as mentioned in section 2. Syntactic model will basically model the syntax of the language based on the parts of speech information. So, the problem can be framed as between each pair of part of speech there is a *word boundary* and the type of boundary are pause and non-pause as mentioned above. This is an n-gram model of parts of speech junctions where at each junction it produces the best sequence of pauses.

$$P(J|C) = P(C|J) * P(J)$$

Where J is the sequence pauses and C is a trigram of present and previous two part of speech junctions. *P(C|J)* is the emission probability which represents local information only, i.e. the basic probability of each juncture type at each point in the syntactic context and a transition probability of *P(J)* which represents the global information.

Corpus used by Lexical model is taken to train and test this model, but as there was no POS tagged data available for this data it is tagged by using the POS tagger as mentioned in Section 4. This is modeled on the basis of global context of pause as the POS information influences longer length of utterance. But the complexity in this approach is the POS tag data required to model the POS tagger (in section 4) has to be created manually and has to be created separately for each language.

3.3 Hybrid Model

There are two reasons to go for a hybrid approach. Firstly lexical approach will model the lexical structure of the pause prediction but not the syntax of the language, where as the syntactic approach models the language syntax but not the lexical structure. So in order to take the advantages of both the models, a hybrid approach is proposed which will model both the models in a single model. Secondly in the above two approaches one models the local information and other global information. So, in order to make a hybrid model, the POS junction is also added into the feature set of the lexical approach and rest of the algorithm is same as the lexical approach.

This approach was also modeled using CART and CRF and the later outperforms the former. The Hybrid approach performs better when compared to the other approaches.

4 Database

The corpus used for both the models contains 1631 sentences. Training set contained 9000 words, 1523 pauses and 7477 non-pauses. Testing set contained 2660 words, 420 pauses and 2240 non-pauses.

Lexical features are obtained from the labeled data of the speech signal which uses HMM based phone labeler which not only labels the phone boundaries in the transcript but has an additional advantage of labeling the silence in the speech signal even though it is not mentioned in the text transcript. This additional functionality automates the process of getting the data for training the pause prediction for lexical approach.

Part of speech data for the Syntactic approach is manually annotated corpus of 30106 words tagged with 26 tags. With this tagged data a HMM based POS tagger is modeled with a window size of 3 using TnT (*Trigrams 'n' Tags*) Toolkit [7]. This can be mathematically represented as

$$P(T|W) = P(W|T) * P(T)$$

Where T is the sequence of tags and W is sequence of present and previous two words. *P(W|T)* represents the emission probability of tagging the word w_i as t_k and *P(T)* represents the transition probability of the sequence of POS tags. Training data for POS tagger contained 27565 words and the testing data contained 2541 and its accuracy was 72.34%.

5 Performance Criteria

Unfortunately, it is not easy to judge the performance of a phrase prediction algorithm. For a given corpus of utterances there will be more non-pauses than pauses, failure to predict a pause can be judged better than over-predicting a pause. A way of judging the performance is if the

	Tools	True Pause	False Pause	True Non-Pause	False Non-Pause
Lexical	CRF	123	297	2100	140
	CART	72	348	2136	104
Syntactic	CRF	20	400	2189	51
	CART	10	410	2216	24
Hybrid	CRF	145	275	2095	145
	CART	103	317	2128	112

Table 1: Confusion Matrix

pauses are massively over-predicted then we can say that the performance is low. So to solve this evaluation issue to some extent a confusion matrix [4] is built as shown in table 1, which contains all the values for number of true pauses, number of false pauses, number of true non-pauses and number of false non-pauses. True Pause is a count of all positions at which the pause is correctly predicted and marked as pause. True Non-Pause is a count of all positions at which the non-pause is correctly predicted and marked as non-pause. False Non-Pause is a count of all positions at which the non-pause is wrongly predicted and marked as pause. False Pause is a count of all positions at which the pause is wrongly predicted as Non-Pause. The performance of system is high if True Pause and True Non-Pause is high and False Pause and False Non-pause is low and vice versa. The other serious problem is that there can be different but valid ways of breaking an utterance. As the results are compared against actual examples they may differ in acceptable as well as unacceptable ways, and there is no easy way to compute this. Ostendorf and Veilleux [5] deal with this problem by having five different speakers reading each test utterance. Assignment is considered correct if the whole utterance matches any of the five samples. Since we did not have the resources to rerecord our database examples we could not conduct those experiments.

6 Experiments & Results

Performance of the pause prediction algorithm is measured by recall, precision and F-measure as shown in Table 2 and the corresponding graphs are shown in Figure 1 and Figure 2. This section gives the reports of our experimental results on above discussed models using CRF and CART respectively. As shown in table 3, the Lexical model gives overall accuracy of 83.54% and 83%. Syntactic model gives 83.01% and 83.68%. And finally, the hybrid model, which uses both lexical as well as syntactic information, gives best overall accuracy of 84.17% and 83.87%.

We also conducted human perception test by synthesizing 8 sentences for all three categories i.e. with out pause prediction module, and with pause prediction using CART and CRF and asking six native speakers to rank each of the utterance with a score of 1-5 (1 being very bad, and 5 being very good). The average score obtained across all utterances and speakers was found to be 3.42, 4.08 and 4.46 for with out pause prediction, with pause prediction using CART and CRF respectively which also coincides with the results in table 3. Comparing the overall performance of hybrid model with the perception test results we can infer that 83.87 % of CART performance corresponds to average perception improvement

	Recall Pause	Precision Pause	F-measure Pause	Recall Non-Pause	Precision Non-Pause	F-measure Non-Pause
Lex-CRF	29.29	5.53	9.30	93.75	94.47	94.10
Lex-CART	17.14	3.26	5.47	95.36	96.74	96.04
Syn-CRF	4.76	0.91	1.52	97.72	99.09	98.40
Syn-CART	2.38	0.45	0.75	98.93	99.55	99.23
Hyb-CRF	34.52	6.47	10.8	93.53	93.53	93.53
Hyb-CART	24.52	4.62	7.77	95.00	95.38	95.18

Table 2: Recall, Precision & F-measure for PAUSE & NON-PAUSE

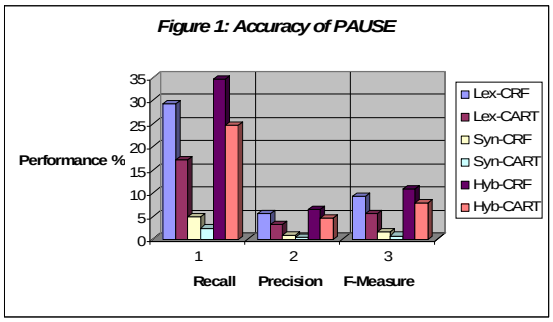
of 0.66 and similarly 84.22 % of CRF corresponds to average perception improvement of 1.04, which indeed convey that 0.35 % improvement from CART to CRF improved the average perception by 0.38.

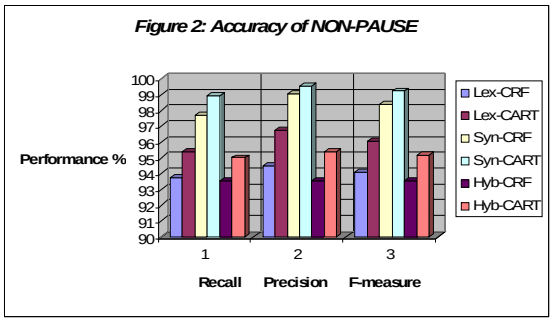
	CART %	CRF %
Lexical	83.01	83.57
Syntactic	83.68	83.05
Hybrid	83.87	84.22

Table 3: Overall Performance

7 Conclusion

In this paper, we have described the Lexical, Syntactical and their hybrid model of pause prediction. Using both syllable sequence and Parts of speech information we got good results for Telugu which is a syllabic language. In our future work, we will be extending to prosody modeling in this frame work.





References

[1] Alan W Black and Paul Taylor, "Assigning Phrase Breaks from Part-of-Speech Sequences", *Proceedings of Eurospeech, 1997.*

[2] Tina Burrows, Peter Jackson, Katherine Knill, Dmitry Sityaev, "Combining Models of Prosodic Phrasing and Pausing", *Proceedings of INTERSPEECH, 2005.*

[3] Chiu-yu Tseng and Bau-Ling Fu, "Duration, Intensity and Pause Predictions in Relation to Prosody Organization", *Proceedings of INTERSPEECH, 2005.*

[4] Jesse Davis, Mark Goadrich, "The Relationship Between Precision-Recall and ROC Curves", *Proceedings of* 23rd international conference on Machine learning, *2006.*

[5] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location", *Proceedings of Computational Linguistics, 1994*

[6] S.P. Kishore, Rajeev Sangal and M. Srinivas, "Building Hindi and Telugu Voices using Festvox", *In Proceedings of ICON, 2002.*

[7] Thorsten Brants, TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.