

DURATION MODELING OF INDIAN LANGUAGES HINDI AND TELUGU

N. Sridhar Krishna, Hema A. Murthy

Department of Computer Science and Engineering,
Indian Institute of Technology, Madras, Chennai - 600036
Email: {sridhar,hema}@lantana.iitm.ernet.in

ABSTRACT

This paper reports a preliminary attempt on data-driven modeling of segmental (phoneme) duration for two Indian languages Hindi and Telugu. Classification and Regression Tree (CART) based data-driven duration modeling for segmental duration prediction is presented. A number of features are proposed and their usefulness and relative contribution in segmental duration prediction is assessed. Objective evaluation of the duration models, by root mean squared prediction error (RMSE) and correlation between actual and predicted durations, is performed. The duration models developed have been implemented in an Indian language Text-to-Speech synthesis system [1] being developed within Festival framework [2].

1. INTRODUCTION

Duration is one of the most important prosodic features that contributes to the perceived naturalness of synthetic speech. Variation in segmental duration serves as a cue to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases thereby increasing the naturalness and intelligibility. In natural speech, segmental durations are highly context dependent. For example, in our study on DD news bulletins [3], we observed instances of vowel /e/ that are as short as 35 ms in the word “pehla” and as long as 150 ms in the word “rahega”. So the primary goal in duration modeling is to model the duration pattern of natural speech, considering various features that affect the pattern. An important restriction being that, due to the nature of the Text-to-Speech synthesis problem, only those features that can be automatically derived from text can be considered.

The approaches to segmental duration modeling can be divided into two categories: rule-based and corpus-based. The most prevalent rule-based duration model is a sequential rule based system proposed by Klatt [4] which is implemented in the MITalk system [5]. In this, starting from some intrinsic rule, the duration of a segment is modified by rules that are applied sequentially. Models of this type have been developed for several languages [6, 7, 8, 9]. However,

rule-based models often generalize too much and cannot handle exceptions well without getting exceedingly complicated. When large speech corpora and the computational means for analysing these corpora became available, new data-driven approaches based on Classification and Regression Trees (CART) [10, 11], linear statistical models [12] and Artificial Neural Networks [13] are proposed.

In this paper, CART based data-driven duration modeling for two Indian languages, Hindi and Telugu, is presented. Classification and Regression Trees are models based on self learning procedures that sort the instances in the learning data by binary questions about the attributes that the instances have. It starts at the root node and continues to ask questions about the attribute of the instance down the tree until a leaf node is reached [14]. The decision tree algorithm selects the best attribute and question to be asked about that attribute at each node. The selection is based on what attribute and question about it divide the learning data so that it gives the best predictive value for the classification. CART modeling is particularly useful in the case of less researched languages like Indian languages, for which the most relevant features that affect the duration pattern and the way they are inter-related have not been studied in detail.

This paper is organised as follows. Section 2 gives the background for the work presented in this paper. In Section 3, details about the speech corpus that is used for the duration analysis is presented. In Section 4, various features considered to derive from the text, and generation of feature vectors from which the CART is trained, is described. Section 5 describes the stepwise construction of CART model for the analysis on contribution and relative importance of various features. In Section 6, objective evaluation of the duration models, by root mean squared prediction error and correlation between actual and predicted durations, is presented.

2. BACKGROUND

Research on Text-to-Speech conversion for Indian languages is a much younger enterprise in comparison with the Text-

to-Speech research for English and other European languages. The major obstacle for speech synthesis research in Indian languages is, we neither have the databases annotated with prosodic and linguistic information nor the tools required to generate the appropriate linguistic information (for example, the syntactic, morphological and lexical information) that is essential to predict various prosody events from the text. Further, a Text-to-Speech system, in general, is targeted for one particular language. In India, there are 18 officially recognised languages each with its own set of dialects. It is very difficult to have one speech synthesizer for each language (and for each dialect!).

In our work, a multilingual Text-to-Speech system [1] for Indian languages is being built within the Festival framework [2]. As a starting point, a common multilingual diphone database is prepared and linguistic/prosodic processing modules are being developed for two Indian languages Hindi and Telugu. The two languages are chosen so as to represent one from each of the Aryan and Dravidian family of languages. The languages Hindi and Telugu are also the first and second largest spoken languages within the India respectively.

This paper reports our work on duration modeling of the two Indian languages Hindi and Telugu. The duration models developed in this paper have been implemented in the Text-to-Speech synthesis system described above.

3. SPEECH CORPUS

The work presented here concentrates on duration modeling within a news-reading style. For language Telugu, the corpus used for study includes one single speaker *Doordarshan* news bulletin [3]. The corpus is of around 14 minute duration and consists of 156 read sentences. The corpus is segmented at phoneme level, thus yielding a total of 6846 segments. The corpus is divided into train data (5477 segments, 80% of the total segments) and test data (1369 segments, 20% of the total segments).

For language Hindi, the corpus used for study include one single speaker *Doordarshan* news bulletin [3] of duration around 12 minute and consists of 121 read sentences. The corpus is segmented at phoneme level, thus yielding a total of 5014 segments. The corpus is divided into train data (4083 segments, 80% of the total segments) and test data (1021 segments, 20% of the total segments).

For segmentation, initially both the corpora are segmented automatically using the technique based on aligning the pre-generated, labeled synthesized prompts [15]. For this purpose, synthesized prompts are generated using the Indian language Text-to-Speech synthesis system [1] with a basic duration model that uses average phoneme durations written by hand. After automatic labeling, the segmentation results are visually inspected and corrected using a labeling

tool, Emulabeller [16]. Both the waveform and the spectrogram are used to determine the segment boundaries, and the boundaries identified are confirmed by listening to the speech. In Table 1, the segments(phonemes) in both the corpora are listed, and the average and standard deviation of the segment durations is given.

4. FEATURE VECTOR GENERATION

Based on the literature [10, 11, 12, 17], number of features are proposed for segmental duration prediction. Only those features that can be automatically derived from text are considered. For example, information about the focus or stress, accent assignment, word boundary strength etc. are not considered even though they are known to affect the duration pattern considerably. The reason for not considering features like 'accent assignment' is, the features in turn need to be predicted from the text. The feature 'stress' in Indian languages is not as clearly defined (both acoustically and perceptually) as in a stress language like English.

Each segment in both the corpora (Telugu and Hindi) is annotated with the following features together with the actual segment(phoneme) duration:

- Identity of the current phoneme; This feature is categorical and it has 42 possible values.
- Identity of the preceding phoneme; This feature is categorical and it has 42 possible values.
- Identity of the following phoneme; This feature is categorical and it has 42 possible values.
- Position in the parent syllable; Position of the segment in the syllable it is related to. The index counts from 0.
- Parent syllable initial; Returns 1 if the segment is the first segment in the syllable it is related to, otherwise 0.
- Parent syllable final; Returns 1 if the segment is the last segment in the syllable it is related to, otherwise 0.
- Parent syllable position type; The type of syllable position in the word it is related to. This may be any of: 'single' for single syllable words, 'initial' for word initial syllables in a poly-syllabic word, 'final' for word final syllables in poly-syllabic words, and 'mid' for syllables within poly-syllabic words.
- Number of syllables in the parent word
- Position of parent syllable in the word; The position of the syllable in the word it is related to. The index counts from 0.

Phoneme name	Telugu		Hindi	
	Mean (in ms)	Std.Dev (in ms)	Mean (in ms)	Std.Dev (in ms)
a	51.0	20.1	52.2	14.3
A	97.4	28.0	89.1	33.5
@	79.3	14.6	-	-
i	56.5	29.0	54.3	21.9
I	88.8	26.6	66.4	25.1
u	50.4	36.3	63.0	19.9
U	79.4	21.3	60.8	21.8
e	58.1	19.7	72.2	24.9
E	90.6	23.6	-	-
ai	111.0	24.5	108.7	27.8
o	58.0	16.7	73.8	31.3
O	90.4	26.3	-	-
au	117.9	28.1	87.2	25.6
k	64.0	18.8	70.6	24.9
kh	87.2	26.2	92.1	39.5
g	46.1	18.9	53.3	18.1
gh	48.0	11.4	42.2	12.6
ch	69.5	17.9	72.0	19.4
j	63.1	26.0	65.1	27.7
T	55.5	15.6	58.8	26.0
td	55.8	5.8	60.1	13.2
D	40.7	22.7	42.1	15.0
N	41.6	12.7	57.7	18.7
th	66.4	17.2	66.8	29.4
tth	65.3	24.1	66.7	28.2
dh	43.9	18.8	58.3	22.9
ddh	62.4	25.1	76.8	31.8
n	51.0	17.9	54.2	20.0
p	65.8	20.3	70.6	21.7
f	72.7	24.6	79.9	20.2
b	51.3	16.8	70.7	26.5
bh	78.6	29.4	132.1	34.8
m	57.8	18.0	60.6	24.7
y	32.8	18.0	39.6	15.7
r	27.9	11.7	40.6	33.7
v	37.5	15.2	52.6	26.3
l	45.0	14.7	46.3	23.7
L	42.2	12.8	-	-
s	77.2	21.2	79.1	25.2
sh	84.9	24.9	87.0	28.5
h	51.3	21.3	55.8	21.4

Table 1. Mean and standard deviation of segment(phoneme) durations in the corpora.

- Parent syllables break information; Break level after the parent syllable. This feature is categorical and it has 4 possible values: 0 for word internal syllables, 1 for syllables occurring in word boundary, 3 for syllables occurring in phrase boundary, 4 for syllables occurring in sentence boundary.
- Phrase length (in number of words).
- Position of phrase in the utterance.

5. GENERATION OF CART DURATION MODELS

Two classification and regression trees, one for language Hindi and one for language Telugu, are generated using the feature data described in the Section 4. Since there is no previous knowledge about the usefulness of the features and their relative importance, classification and regression trees are built in a step-wise fashion to establish the usefulness and relative importance of the features. In this approach each single feature is taken in turn and a tree consisting of nodes containing only the conditions imposed by that feature is built. The single best tree is then kept and each remaining feature is taken in turn and added to the tree to find the best tree possible with just two features. The procedure is then repeated for a third, fourth, fifth feature and so on. This process continues until no significant gain in accuracy is obtained by adding more features. For running the CART tree building process, 'Wagon' classification and regression tree tool [18] is used. Detailed analysis on the usefulness of the proposed features and their relative importance is given in Section 6.

5.1. Segmental duration prediction

The segmental durations are predicted by traversing the decision tree (CART) starting from the root node, taking various paths satisfying the conditions at intermediate nodes, till the leaf node is reached. The path taken depends on various features like, the segment identity, preceding and following segment identities, position of the segment in parent syllable, position of the syllable in parent word etc. The leaf node contains the value of segmental duration prediction.

An example partial decision tree (CART) for segmental duration prediction is shown in Figure 1. The tree assigns different durations for segment /u/ when it occurs in different contexts. A duration value of 110 ms is assigned when it satisfies the following criteria: the preceding segment is /th/, parent syllable is the final syllable in the parent word, and there is a break (or pause) after the parent syllable. A duration value of 70 ms is assigned when it satisfies the following criteria: the preceding segment is /th/, parent syllable is the final syllable in the parent word, and the parent syllable is not at the end of a phrase break. A duration value

of 85 ms is assigned when the preceding segment is /th/ and the following segment is /n/. A duration value of 65 ms is assigned when the preceding segment is /p/ and the following segment is /d/.

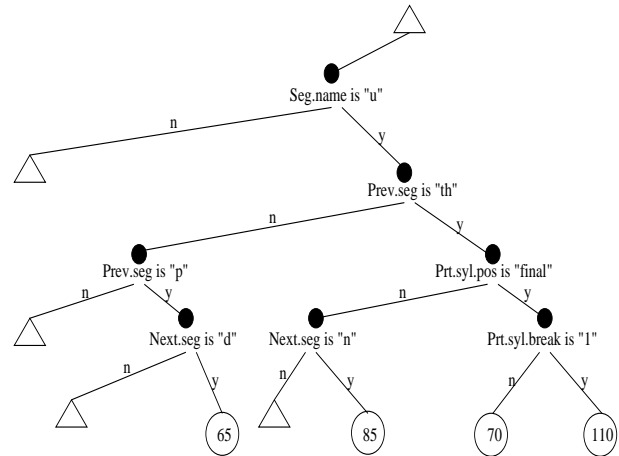


Fig. 1. An example partial decision tree (CART) for segmental duration prediction. The triangles depict omitted parts.

6. OBJECTIVE EVALUATION AND DISCUSSION

Objective evaluation of the duration models, by root mean squared prediction error (RMSE) and correlation between actual and predicted durations, is performed.

The following two subsections gives the evaluation results of the CART-based duration models for languages Telugu and Hindi.

6.1. Telugu

The duration model is trained with train data (5477 segments, 80% of the total segments) and evaluated with test data (1369 segments, 20% of the total segments). Correlation obtained between actual and predicted durations is 0.8014 and the root mean squared error(RMSE) of prediction is 22.86 ms.

To assess the effectiveness of the features considered, CART trees are built in a step-wise fashion as described in Section 5. and the results are shown in Table 2. The first column gives the feature names, and the second column gives the correlation obtained between actual and predicted durations by the addition of the successive features in the CART modeling process. The most important feature that contributed to the segmental duration prediction is the identity of the segment being predicted. The next two most important features are the identity of preceding and following segments. The other important features for the

segmental duration prediction are, the parent syllable final position and the parent syllable break information followed by number of syllables in the parent word.

Feature used	Correlation between actual and predicted durations (cumulative)
Segment.name	0.6417
Next.Segment	0.7163
Previous.Segment	0.7658
Syllable.final	0.7726
Syllable.break	0.7906
Word.numsyls	0.8011

Table 2. Analysis on usefulness of features in segmental duration prediction, for language Telugu.

6.2. Hindi

The duration model is trained with train data (4083 segments, 80% of the total segments) and evaluated with test data (1021 segments, 20% of the total segments). Correlation obtained between actual and predicted durations is 0.7526 and the root mean squared error (RMSE) of prediction is 27.14 ms.

To assess the effectiveness of the features considered CART trees are built in a step-wise fashion as described in Section 5. and the results are shown in Table 3. The First column gives the feature names, and the second column gives the correlation obtained between actual and predicted durations by the addition of the successive features in the CART modeling process. The most important feature that contributed to the segmental duration prediction is the identity of the segment being predicted. The next most important feature is the identity of preceding segment. The other important features are the parent syllable position type and parent syllable break information followed by parent syllable position in the word. In contrast to Telugu, identity of the following segment has not contributed to the segmental duration prediction in Hindi.

7. CONCLUSIONS

A preliminary attempt on data-driven duration modeling of two Indian languages Hindi and Telugu is described. Classification and Regression Tree (CART) based duration modeling for segmental duration prediction is presented. A number of features are proposed and their usefulness and relative contribution in segmental duration prediction is assessed. An important observation being that, in contrast to Telugu, identity of the following segment has not contributed to the

Feature used	Correlation between actual and predicted durations (cumulative)
Segment.name	0.5692
Previous.Segment	0.6981
Syllable.position.type	0.7116
Syllable.break	0.7406
Syllable.pos.in.word	0.7514
Syllable.final	0.7515

Table 3. Analysis on usefulness of features in segmental duration prediction, for language Hindi.

segmental duration prediction in Hindi. Objective evaluation on unseen data has given inferior but comparable results with the duration models for much researched languages like English.

8. CRITICISM

- The corpora used for modeling and analysis is not optimal for duration modeling, as no attempt is made to take care of data sparsity problem or to cover the feature space.
- Modeling and analysis is done on smaller data sets.

9. REFERENCES

- [1] Sridhar Krishna, N., Hema A. Murthy, Timothy A. Gonsalves, "Text-to-Speech in Indian Languages.," in *International Conference on Natural Language Processing*, Mumbai, India, 2002, pp. 317–326.
- [2] Black, A.W., Paul Taylor, and Richard Caley, *The Festival Speech Synthesis System: Manual and source code available at*, <http://www.cstr.ed.ac.uk/projects/festival.html>, CSTR web page.
- [3] *Database for Indian languages*, India, Speech and Vision Lab, IIT Madras, Chennai, 2001.
- [4] Dennis H. Klatt, *Synthesis by rule of segmental durations in english sentences.*, In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication Research*, pages 287-300., Academic Press, New York., 1979.
- [5] Jonathan Allen, M. Sharon Hunnicut, and Dennis H. Klatt, *From Text to Speech: The MITalk system.*, Cambridge University Press, Cambridge., 1987.

- [6] Carison, R. and B. Granstrom, "A search for durational rules in real speech database.," *Phonetica*, vol. 43, pp. 140–154, 1986.
- [7] van Santen, J. P. H., "Contextual effects on vowel durations.," *Speech Communication*, vol. 11, pp. 513–546, 1992.
- [8] Bartkova, K. and C. Sorin, "A model of segmental duration for speech synthesis in french.," *Speech Communication*, vol. 6, pp. 245–260, 1987.
- [9] Simoes, A.R.M., "Predicting sound segment duration in connected speech: An acoustical study of brazilian portugese.," *In Workshop on Speech Synthesis, ESCA, Atrans.*, pp. 173–176, 1990.
- [10] Riley, M.D., "Tree-based modeling for speech synthesis.," *In: G. Bailly, C. Benoit, and T. Sawallis (Eds.), Talking machines: Theories, models and designs.*, pp. 265–273, 1992.
- [11] Hyunsong Chung and Mark A. Huckvale, "Linguistic factors affecting timing in korean with application to speech synthesis," in *Eurospeech*, Denmark, 2001.
- [12] van Santen, J.P.H., "Assignment of segmental duration in text-to-speech synthesis.," *Computer Speech and Language*, vol. 8, pp. 95–128, 1994.
- [13] Campbell, W., "Syllable-based segmental durations.," *In: G. Bailly, C. Benoit, and T. Sawallis (Eds.), Talking machines: Theories, models and designs.*, pp. 43–60, 1992.
- [14] Mitchell, T.M., *Machine Learning*, McGraw-Hill, New York, 1997.
- [15] Malfrere, F. and T. Dutoit, "High quality speech synthesis for phonetic speech segmentation.," in *Eurospeech*, Rhodes, Greece, 1997, pp. 2631–2634.
- [16] Cassidy, S., *The EMU Speech Database System*, <http://www.shlrc.mq.edu.au/emu/>, 2002.
- [17] Lee, S. and Y.H. Oh, "Tree-based modeling of prosodic phrasing and segmental duration for korean tts systems.," *Speech Communication*, vol. 28, pp. 283–300, 1999.
- [18] Taylor, P., R. Caley, and A.W. Black, *The Edinburgh Speech Tools Library, 1.2.1 edition*, University of Edinburgh, <http://www.cstr.ed.ac.uk/projects/spechtools.html>, 2002.