Duration Modeling for Hindi Text-to-Speech Synthesis System

N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, A.G. Ramakrishnan[†]

Hewlett-Packard Labs India
Bangalore - 560 030.
{nsridhar,partha.talukdar,kalika}@hp.com

[†]Indian Institute of Science, Bangalore - 560012, India.
ramkiag@ee.iisc.ernet.in

Abstract

This paper reports preliminary results of data-driven modeling of segmental (phoneme) duration for Hindi. Classification and Regression Tree (CART) based data-driven duration modeling for segmental duration prediction is presented. A number of features are considered and their usefulness and relative contribution for segmental duration prediction is assessed. Objective evaluation of the duration model, by root mean squared prediction error (RMSE) and correlation between actual and predicted durations, is performed.

1. Introduction

Accurate estimation of segmental durations is crucial for natural sounding text-to-speech (TTS) synthesis [1]. 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. 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, i.e., as only text is provided for the synthesis, 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 [2], which is implemented in the MITalk system [3]. In this system, 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 [4, 5, 6, 7]. However, rule-based models often over-generalize 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) [8, 9], linear statistical models [10] and Artificial Neural Networks [11] have been increasingly used for duration modeling.

In this paper, duration modeling for Hindi is performed using data-driven approach based on CART. 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 attributes of the instance down the tree until a leaf node is reached [12]. For each node, the decision tree algorithm selects the best attribute, and also the question to be asked about that attribute. The selection is based on what attribute and question about it divide the learning data so that it gives the best predictive value for right 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 are presented. Section 4 describes the features considered for duration modeling and subsequent generation of feature vectors from which the CART based duration model is trained. Section 5 describes the stepwise construction of CART model for analysis on the contribution and relative importance of various features. In Section 6, objective evaluation of the duration model, by root mean squared prediction error and correlation between actual and predicted durations, is presented.

2. Hindi TTS System

Hindi, the official language of India, is spoken as a first language by 33 percent of the Indian population, and by many more as a lingua franca. Only around 5 percent of Indians use English as a means of communication. This fact, coupled with the prevalent low literacy rates, make the use of conventional user interfaces difficult in India. Speech user interfaces using Hindi and other local language Text to Speech systems, provide an ideal means of making information and other ICT based services more accessible to a large proportion of Indians.

The development of a high quality Hindi TTS system [14] is under progress at HP Labs India. This effort is a part of the Local Language Speech Technology Initiative [15], which brings together motivated groups around the world, providing tools, expertise, support and training to enable TTS to be developed in local languages. The aim of LLSTI is to develop a TTS framework around Festival that will allow for a rapid development of TTS in any language.

This paper reports our ongoing work on duration modeling carried out as part of the HP Labs India Hindi TTS System.

3. Speech corpus used

The present study of segmental durations in natural speech is based on a corpus of around 22 minute duration, which consists of 250 sentences taken from three short stories. All the sentences are spoken by a native Hindi male speaker in expressive story reading style. The speaker is also a professional radio artist. The recorded data is manually segmented at phoneme level using Praat [13], thus yielding a total of 12535 segments. The data is divided randomly into training data (11282 segments, 90% of the total segments) and test data (1253 segments, 10% of the total segments). A total of 70 phonemes are analysed for their context-dependent durations.

4. Feature vector generation

Based on the literature [8, 9, 10, 16], a number of features are considered 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 and word boundary strength are not considered even though they are known to affect duration pattern. However, 'stress' in Indian languages is not as clearly studied (both acoustically and perceptually) as in a stress language like English. Each segment in the corpus is annotated with the following features together with the actual segment (phoneme) duration:

- Segment identity; e.g., /a/, /k/, /S/.

- Segment features; e.g., vowel length, vowel height, consonant type, consonant voicing.

- Previous segment (immediate left context) features; e.g., vowel length, vowel height, consonant type, consonant voicing.

- Next segment (immediate right context) features; e.g., vowel length, vowel height, consonant type, consonant voicing.

- Parent syllable structure; e.g., onset, coda, onset size, coda size.

- 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 part of, otherwise 0.

- Parent syllable final; Returns 1 if the segment is the last segment in the syllable it is part of, otherwise 0.

- Parent syllable position type; The type of syllable position in the word it is part of. 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 the parent syllable; The position of the syllable in the word it is part of. The index counts from 0.

- 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.

- Number of phrases in the utterance.

The speech corpus used for modeling and analysis is currently not optimal for duration modeling, since we could not take care of to take care of data sparsity problem or cover feature space. However, to reduce the problem caused due to a small data set, care has been taken to represent the feature space in a generalized manner. For example, the segmental context (immediate left and right context) is represented using various features (front vowel, consonant type etc.) instead of the absolute identities.

5. Generation of CART duration model

Classification and Regression Tree based duration model is trained with feature data described in Section 4. Since there is no previous knowledge about the usefulness of the features and their relative importance, CART's 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 the 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 building process, 'Wagon' classification and regression tree tool [17] is used. Detailed analysis on the usefulness of the proposed features and their relative importance is given in Section 6.

5.1. Prediction of segmental duration

The segmental durations are predicted by traversing the decision tree 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 and position of the syllable in parent word. The leaf node contains the predicted value of segmental duration.

An example partial decision tree 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/.

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 duration model is trained with training data (11282 segments, 90% of the total segments) and evaluated with test data (1253 segments, 10% of the total segments).

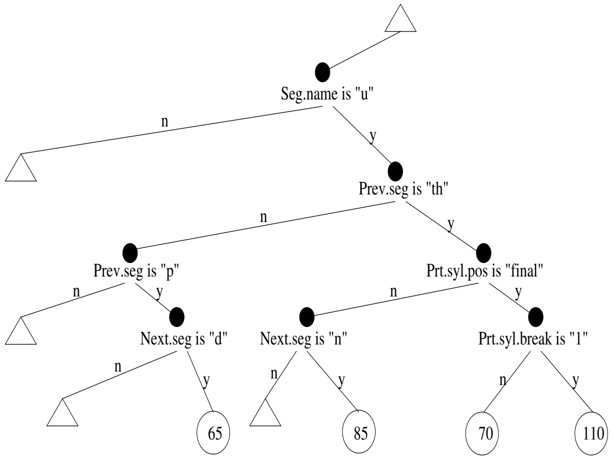

Figure 1: An example partial decision tree (CART) for segmental duration prediction. The triangles depict omitted parts.

Correlation obtained between the actual and predicted durations is 0.7987 and RMSE of prediction is 18.92 ms.

To assess the effectiveness of the features considered, CART's are built in a step-wise fashion (as described in Section 5) and the results are shown in Table 1.

Feature used	Correlation (cumulative)
Segment Identity	0.6234
Next Segment (onset/coda)	0.7112
Next Segment Vowel rounding (1/0)	0.7302
Next Segment Consonant Type	0.7423
Previous Syllable Break	0.7511
Syllable Coda Size	0.7587
Syllable Position (in word)	0.7691
Previous Segment Consonant Type	0.7744
Previous Segment Vowel Height	0.7869
Syllable Break	0.7987

Table 1: Analysis on usefulness of features in predicting segmental duration.

The first column gives the names of the feature, 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. From the results, we observe that the most important feature that contributed to segmental duration prediction is the identity of the segment itself. Other important features in decreasing order of importance are:

- Next segment's syllable structure - whether the next segment is onset or coda of its parent syllable.

- Next segment type - vowel rounding and consonant type of next segment.
- Previous syllable break information.
- Parent syllable coda size.
- Syllable position in word.
- Previous segment type - consonant type and vowel height of previous segment.
- Parent syllable break information.

Though the segmental identity is the most important feature, the prediction of segmental duration is improved greatly by additional features *viz.* syllable structure, immediate context type (right and left) and syllable break information.

7. Conclusions

Preliminary results on data-driven Hindi duration modeling is presented. Classification and Regression Tree based approach for modeling segmental duration is followed. A number of features are considered and their usefulness and relative contribution to segmental duration prediction is assessed. Work is in progress on preparing large annotated speech corpora and better prosody learning.

8. References

[1] Venditti, Jennifer J. and Jan P. H. van Santen., "Modeling Segmental Durations for Japanese Text-To-Speech Synthesis", In SSW3, pages 31–36, 1998.

[2] 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.

[3] Jonathan Allen, M. Sharon Hunnicut, and Dennis H. Klatt, "From Text to Speech: The MITalk system", Cambridge University Press, Cambridge, 1987.

[4] Carison, R. and B. Granstrom, "A search for durational rules in real speech database", Phonetica, vol. 43, pp. 140-154, 1986.

[5] van Santen, J. P. H., "Contextual effects on vowel durations", Speech Communication, vol. 11, pp. 513-546, 1992.

[6] Bartkova, K. and C. Sorin, "A model of segmental duration for speech synthesis in French", Speech Communication, vol. 6, pp. 245-260, 1987.

[7] Simoes, A.R.M., "Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portugese", In Workshop on Speech Synthesis, ESCA, Autrans, pp. 173-176, 1990.

[8] Riley, M.D., "Tree-based modeling for speech synthesis", In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 265-273, 1992.

[9] Hyunsong Chung and Mark A. Huckvale, "Linguistic factors affecting timing in Korean with application to speech synthesis", in Eurospeech, Denmark, 2001.

[10] van Santen, J.P.H., "Assignment of segmental duration in text-to-speech synthesis", Computer Speech and Language, vol. 8, pp. 95-128, 1994.

[11] Campbell, W., "Syllable-based Segmental Durations", In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 43-60, 1992.

[12] Mitchell, T.M., Machine Learning, McGraw-Hill, New York, 1997.

[13] Boersma, P. and D. Weenik., "Praat: A System for Doing Phonetics by Computer", (http://www.praat.org/), 2001.

[14] Ramakrishnan, A.G. et.al., "Tools for the Development of a Hindi Speech Synthesis System", In 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 109–114, 2004.

[15] Roger Tucker, "Local Language Speech Technology Initiative (LLSTI)", (http://www.llsti.org), 2003.

[16] 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.

[17] 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/speechtools.html, 2002.