RESEARCH ARTICLE OPEN ACCESS

Duration for Classification and Regression Treefor Marathi Text-to-Speech Synthesis System

Sangramsing N. Kayte[1], Monica Mundada[1],Dr. Charansing N. Kayte[2], Dr. Bharti Gawali*
[1,3]Department of Computer Science and Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad
[2]Department of Digital and Cyber Forensic, Aurangabad, Maharashtra

Abstract
This research paper reports preliminary results of data-driven modeling of segmentalphoneme duration for Marathi. Classification and Regression Tree 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 and correlation between actual and predicted durations, is performed.

I. Introduction

Accurate estimation of segmental durations is crucial for natural sounding text-to-speech synthesis [1-2]. 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 corpusbased. The most prevalent rule-based duration model is a sequential rule based system proposed by Klatt [2], which is implemented in the MI-Talk classification [3]. In this system, starting fromsome intrinsic rule, the durationof a segment is modified by rules that are applied sequentially. Models of this type have been developed for several languages [4, 5]. 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 analyzing these corpora became available, new data-driven approaches based on Classification and Regression Trees [6], linear statistical models and Artificial Neural Networks [7] have been increasingly used for duration modeling.

In this paper, duration modeling for Marathi is performed using data-drivenapproachbased 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 [1,2,3,8]. 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 organized 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.

II. Marathi TTS System

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighboring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtra, one of the Prakrit languages which developed from Sanskrit. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. From the 13th century until the mid-20th century, it was written with the Modi alphabet. Since 1950 it has been written with the Devanāgarī alphabet[9,10].

The development of a high quality Marathi TTS system [11] is under progress at HP Labs India. This effort is a part of the Local Language Speech Technology Initiative [7,11], 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[8,10,12].

2.1 Devanāgarī alphabet for Marathi

Vowels and vowel diacritics [9]

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	आँ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṅ	aḥ		
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]	[e]	[əi]	[o]	[əu]	[$ə^{n}$]	[əh]	[æ]	[ɔ]

प	पा	पि	पी	पु	पू	पृ	पे	पै	पो	पौ	पं	पः
pa	pā	pi	pī	pu	pū	pṛ	pe	pai	po	pau	paṅ	paḥ

Consonants [9]

क	ka	[kə]	ख	kha	[$k^{h}ə$]	ग	ga	[gə]	घ	gha	[$g^{ɦ}ə$]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[$ts^{h}ə$]	ज	ja	[ʄə/zə]	झ	jha	[$ʄ^{ɦ}ə/z^{ɦ}ə$]	ञ	ña	[ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[$ʈ^{h}ə$]	ड	ḍa	[ɖə]	ढ	ḍha	[$ɖ^{ɦ}ə$]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[$t̪^{h}ə$]	द	da	[d̪ə]	ध	dha	[$d̪^{ɦ}ə$]	न	na	[n̪ə]
प	pa	[pə]	फ	pha	[$p^{h}ə$/fə]	ब	ba	[bə]	भ	bha	[$b^{ɦ}ə$]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṟa	[ɾə]	ल	la	[lə]	व	va	[və/wə]
श	śa	[ʃə]	ष	ṣa	[ʂə]	स	sa	[sə]						
ह	ha	[ɦə]	ळ	ḷa	[ɭə]	क्ष	kṣa	[kʃə]	ज्ञ	jña	[ʄɲə]	श्र	śra	[ʃrə]

III. Speech corpus used

The present study of segmental durations in natural speech is based on a corpus of around 90 minute duration, which consists of 1000 sentences taken from three short stories. All the sentences are spoken by a native Marathi 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 Pratt[13,14], thus yielding a total of 10,000 segments. The data is divided randomly into training data (10,000 segments, 85% of the total segments) and test data (1000 segments, 15% of the total segments). A total of 68 phonemes are analysed for their context-dependent durations.

IV. Feature vector generation

Based on the literature [9,10,11,12,13], 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 Maharashtra 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., /ka/, /ka/, /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. instead of the absolute identities.

V. 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 [15,16] 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/.

VI. 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).

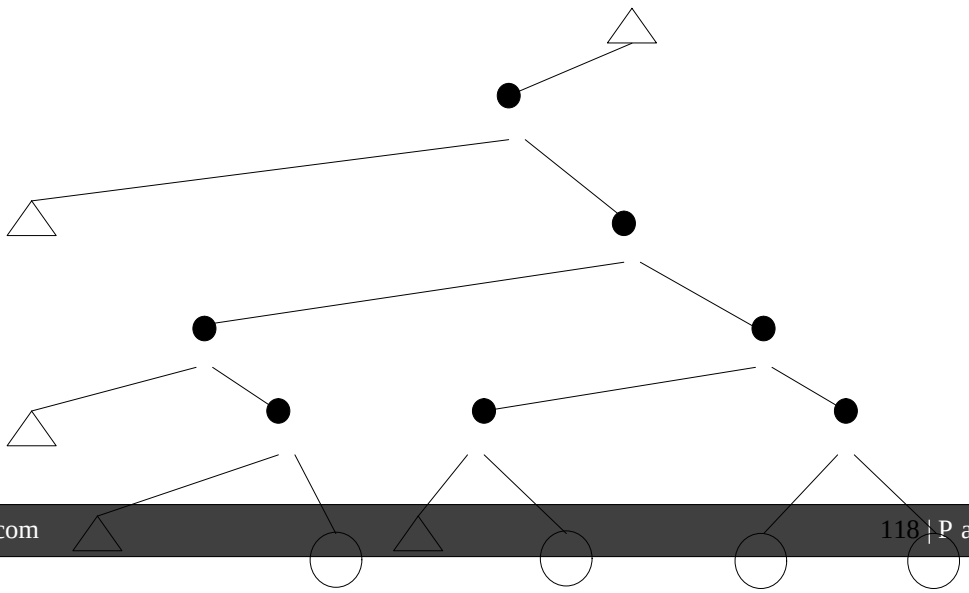

Figure 1: An example partial decision tree (CART) for segmental duration prediction. The triangles depict omitted parts.

Correlation obtained between the actual and predicted durations is 0.8997 and RMSE of prediction is 28.82 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.

Table 1: Analysis on effectiveness of features in calculating segmental time.

Feature used	Correlation
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

The first column gives the names of the feature, and the secondcolumn gives the correlationobtainedbetween 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.

VII. Conclusions

Earliest results on data-driven Marathi 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.

References

[1.] Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014

[2.] Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT)

[3.] Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014

[4.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[5.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep –Oct. 2015), PP 76-81e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197 www.iosrjournals.org

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

[7.] 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.

[8.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Implementation of Marathi Language Speech Databases for Large Dictionary" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 40-45e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[9.] Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015.

[10.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[11.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte " Performance Calculation of Speech Synthesis Methods for Hindi language IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6,

Ver. I (Nov -Dec. 2015), PP 13-19e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[12.] Sangramsing N.kayte "Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach" 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.

[13.] Sangramsing Kayte, Dr. Bharti Gawali "Marathi Speech Synthesis: A review" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

[14.] Sangramsing Kayte, Monica Mundada, Santosh Gaikwad, Bharti Gawali "PERFORMANCE EVALUATION OF SPEECH SYNTHESIS TECHNIQUES FOR ENGLISH LANGUAGE " International Congress on Information and Communication Technology 9-10 October, 2015

[15.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Review of Unit Selection Speech Synthesis International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[16.] 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.