Decision Tree Based Segmental Duration Prediction for Amharic TTS system

Tadesse Anberbir*, Hyunjung Shin† and Dong Yoon Kim*

*Department of Computer Engineering, Ajou University, Suwon, Korea
† Department of Industrial and Information Systems Engineering, Ajou University, Suwon, Korea
E-mail: tadanberbir@gmail.com

Abstract—**This paper reports preliminary results of data-driven decision tree based segmental duration prediction for Amharic Text-to-speech synthesis system. Using a manually annotated speech corpus, we extracted a limited number of important features for the construction of training and testing sets. Then, a decision tree based segmental duration prediction was assessed by objective evaluation of the duration model, using root mean squared prediction error (RMSE) and correlation between actual and predicted durations, and we found promising results. The proposed method can be integrated into the TTS duration module for predicting the duration of speech sounds in various textual, prosodic, and segmental contexts.**

I. Introduction

Text-to-Speech (TTS) synthesis is a process which artificially produces synthetic speech for various applications. In TTS synthesis, naturalness is the main goal and it can be achieved mainly by incorporating prosodic features which include duration of segments, intonation patterns and stress. Particularly modeling the segmental duration based on the context is crucial. This paper reports the preliminary results of segmental duration prediction method for Amharic Text-to-speech synthesis system.

In TTS systems, accurate estimation of segmental duration is one of the most important factors that determine the naturalness of synthesized speech. So far many researches have been conducted in the field of duration modeling and interesting results are obtained for various languages [1-4]. However, still now the task of duration modeling is challenging mainly because it is language dependent and the features considered for modeling are limited to those features that can be automatically derived from the input text only. For instance, in Amharic language, although geminates play a key role for naturalness, it can not be predicated from input text. This makes the duration modeling of geminates a bit challenging. In this study we tried to predict only the duration of the basic units based on the context.

Amharic, the official language of Ethiopia, is one of the least supported and least researched languages in the world. Although, recently, the development of different natural language processing (NLP) tools for analyzing Amharic text has begun, it is often very far comparing with other languages [5]. Particularly, researches conducted on the application of machine learning techniques for NLP problems and their applications for language technologies such as speech synthesis are unavailable. To the knowledge of the authors, so far there is only one published work [6] in the area of speech synthesis but no attempts have been made in analysis and modeling of prosody of Amharic, particularly in the area of segmental duration prediction.

This paper presents a description of, as far as the authors know, the first published segmental duration model for Amharic TTS system using a data-driven approach. We proposed a decision tree based modeling techniques for segmental duration prediction based on studies for other languages. The paper reports our ongoing work on prosody modeling (mainly on automatic prediction of geminates duration) for Amharic TTS as part of the PhD study.

II. Amharic TTS System

AmhTTS is parametric and rule-based system that employs a Cepstral method and uses a Log Magnitude Approximation (LMA) filter. The system is designed based on the general speech synthesis system [7]. The input is Amharic text, and the output is synthetic speech. The text analysis sub-system converts Amharic text into a sequence of mapped characters, and then this sequence is used to get information for synthesis. The speech synthesis sub-system generates speech from pre-stored parameters under the control of systems rules. The database contains data for rules and syllable parameters with suitable formats. Fig. 1 shows the design of Amharic speech synthesis system.

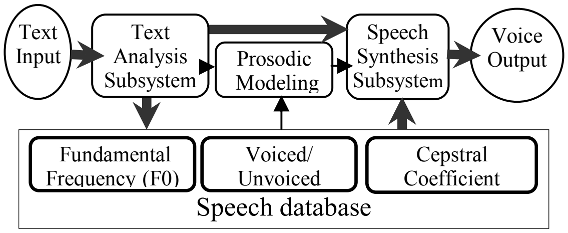

Fig 1. Amharic Speech Synthesis System

III. Speech Corpus Used

As a corpus, a recorded speech consists of about 183 sentences was used from study [6]. Then the recorded data was manually segmented at phoneme level and the duration is determined using Praat [8], thus yielding a total of 100 segments for this particular study. Finally, the important features was extracted based other studies [1, 2, 3, and 4]. In Amharic, except the geminates which is hard to be predicted from input text, the other features can be automatically derived from text.

978-1-4244-4522-6/09/$25.00 ©2009 IEEE

The following are the list of features which we extracted for each segment in the corpus together with the actual segment (phoneme) duration:

- Phonemic identity ;(vowel, consonant).
- Preceding phoneme type; (voiceless fricative, voiceless stop, voiced stop, voiced fricatives, glides, ejectives, nasals)
- Next phoneme type; (voiceless fricative, voiceless stop, voiced stop, voiced fricatives, glides, ejectives, nasals)
- segment position in syllable; levels: onset, nucleus, coda
- Position in a word (beginning, median, end)
- Word length (number of syllables)
- Singleton/Geminate decision (Yes/No)

Since our main interest is speech synthesis of textual data, we focused at the derivation of features that could be extracted from text. But it is important to note that this database was not optimal for the purpose of duration system construction because no attempt was made to cover the greatest number of distinct feature vectors. We used these features as a preliminary experiment. Moreover features like “stress” could not be considered because they are not clearly studied in Amharic.

IV. Decision tree duration model

In data mining, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees.

In our study, a decision tree based duration model was trained with feature data listed in Section III using j48 classifier in Weka software [9]. Waka is a collection of machine learning algorithms for data mining tasks and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is well-suited for developing new machine learning schemes.

Decision tree was chosen in our pilot experiment on Amharic segmental duration modeling. Despite the fact that recently it was shown that other methods like neural networks are superior comparing to Decision tree and CART, the main reasons for choosing this method were:

- Standard tools for their generation are widely available and applied for many langauges and,
- Decision tree and CART modeling is particularly useful in the case of less researched languages like Amharic, for which the most relevant features that affect the duration pattern and the way they are inter-related have not been studied in detail.(for e.g., “geminates” in Amharic language are not well studied).

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.

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

V. Conclusions

In this paper, a preliminary results on decision tree based data-driven duration modeling for Amharic TTS system was presented. Limited number of features was considered and objective evaluation of the duration model, by root mean squared prediction error (RMSE) and correlation between actual and predicted durations was performed and we found promising results.

As a future work, we have a plan to annotate a large corpus and continue studying influential features on Amharic segmental duration and finally integrate the duration model into our TTS system and compare with other studies.

Acknowledgment

H.Shin would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from Korean Government (MOEHRD).

References

[1] N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, A. G. Ramakrishnan, “Duration Modeling for Hindi Text-to-Speech Synthesis System”, in Proc. ICSLP 2004, South Korea, 2004.

[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] Mixdorff, H., Nguyen, D.T. and Wu, N.T. (2005): Duration Modeling in a Vietnamese Text-to-Speech System. Proceedings of Specom2005, Patras, Greece, 2005.

[4] Qing Guo, Nobuyuki Katae, Hao Yu1, Hitoshi Iwamida, “Decision Tree based Duration Prediction in Mandarin TTS System”, Journal of Chinese Language and Computing 97-106.

[5] A. Alemu, L. Asker, and M. Getachew, “Natural language processing for Amharic: Overview and suggestions for a way forward,” in 10th Conf. Traitement Automatique des Langues Naturelles, Batz-sur-Mer, France, 2003.

[6] Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, “Unit Selection Voice for Amharic Using Festvox”, 5th ISCA Speech Synthesis Workshop, Pittsburgh, 2004..

[7] Takara, T.; Kochi, T., 2000. General Speech Synthesis System for Japanese Ryukyu Dialect. Proc. of the 7th WestPRAC, 173-176.

[8] http://www.fon.hum.uva.nl/praat/

[9] http://www.cs.waikato.ac.nz/ml/weka/