Exploring Text To Speech Synthesis in Non-Standard Languages

Jesin James[1], Catherine I. Watson[1], Deepa P. Gopinath[2]

[1] Dept. of Electrical and Computer Engineering, University of Auckland, NZ
[2] College of Engineering, Trivandrum, India

Abstract

The development of Text to Speech synthesis systems as part of the Human Language Technology is a dominating research field in the present age. Languages are the backbone of every culture and developing technologies in every language is a basic requirement for users. Technology development focusing on the standard languages is a discouraging trend for many users who are not well-versed in these languages, but are in need of assistive technology. This paper focusses on the challenges in developing speech systems for non-standard languages, in light of a TTS development work in Malayalam language (non-standard language used in South India).
Index Terms: Under-resourced, Non-standard languages, Text To Speech synthesis, Indian languages, Malayalam

1. Introduction

Speech is the most basic form of human interaction and it has dominated all human ways of communication. This primary communication method was structured and categorized depending on the location where people lived, the differences in the words, tones, accents spoken etc. This gave birth to languages as we know them today. Language is the capability to convey ones emotions and ideas to a larger public. History has proved that the understanding of the scripts in ancient languages have paved way to envisage the development of the world as we see it today. In [1], S. Houston emphasizes this by stating Humankind is defined by Language; Civilization is defined by Writing. UNESCO has highlighted language diversity as a crucial element of the cultural diversity of the world [2][3].

Citing from the statistics depicted in Ethnologue website, presently there are a total of 7097 living languages in the world. Out of this, only 572 (8%) languages come under the category of "institutional", meaning they are used and sustained by institutions beyond the home country. Further analyzing the status of languages across the globe, a very discouraging fact emerges that only less than 100 (1.4%) languages have the required resources for high level language technology like sufficient speech corpus, parsers, POS taggers, morphological analyzers etc. [4] This is the era of advanced research studies in Human Language Technology including Text To Speech synthesis, Speech Recognition, Machine Translation, Natural Language Understanding etc. But it is a fact that a large majority of these works are happening only in a few privileged languages (1.4%) and the fruits of these research works do not reach people of all linguistic backgrounds. Such languages that are not having the necessary technical support to develop various speech processing technologies are termed as under-resourced languages or non-standard languages. There have been many distributed attempts[8] across the globe to analyse various non-standard languages and the predominant handicap reported by almost all researchers has been the non-availability of standard resources. The Basic Language Resource Kit (BLARK), a concept defined by Krauwer in [5] lists out the basic requirements for Text To Speech development in any language. BLARK comprises of written language corpora, spoken language corpora, mono and bilingual dictionaries, terminology collections, grammars, modules (e.g. taggers, morphological analysers, parsers, speech recognizers, text-to-speech), annotation standards and tools etc. Also there are many languages in the world where communication happens by speech only, as they do not have an acceptable written script. This makes speech processing complex [6].

The paper is organized as follows: Section 2 deals with the motivation in pursuing research in TTS systems in under-resourced languages. This is followed by Section 3 which explains various works that happened in other non-standard languages and the challenges faced by the researchers. Section 4 and 5 describes the background and methodology of the TTS development attempt in Malayalam language. Section 6 deals with the results that were obtained and Section 7 concludes the paper.

2. Motivation

Advancements in Human Language Technology is happening on a daily basis. It definitely makes our life easier and faster. But the real beneficiaries of these developments should be people who are vocally or visually challenged [8]. Statistics show that there are about 285 million visually impaired worldwide [9] out of which 90% are from developing countries. Also about 7% of the children in the age group of 13-17 face speech or language disorders [10]. These people are well distributed among all linguistic backgrounds, but the problems are more pronounced in developing and under-developing countries due to the lack of modern facilities. The TTS systems can be used by such people for book, newspaper reading and even for disabled children as an aid in studies. There is a growing awareness among disabled people on how these systems can make their life more simplified, but language becomes a barrier here as well.

Many studies show that cultural and linguistic backgrounds are dominating factors in acceptance of speech technology by people. [11] A sense of familiarity with a voice always encourages the people to use a particular technological system. A majority of the well-developed TTS technologies are in English, and there is a trend for non-English speakers to not accept them [12]. Especially for people with disabilities, learning a new language other than their mother tongue, to use

a TTS system will be burdensome. Hence a necessity arises to develop high quality TTS systems in every non-standard language, for the ease of use of people and preventing the language from being extinct at a later stage. When such an attempt is made in minor languages, the researchers face the problem of inadequate resources as cited in the next section.

3. Related Work

One of the pioneering works in developing prosodic models in English is reported by Hirshberg et.al [32] in developing intonational phrasing using CART. In the experiment design for Australian aboriginal language Pitjantjatjara [13], the author has stated that developing a proper TTS for the language is not realizable presently, due to the non-availability of any resources for the same. So, the work reports using a major-language TTS to design a TTS for Pitjantjatjara. In [8] the author states that for a low-resourced language like Vietnamese, speech processing services or commercial products (ASR, TTS) do not exist yet. For developing such services, large amount of resources listed previously [16] are required. Due to this, cross lingual acoustic modelling has been used to develop a TTS for the language.

The authors in [14],[15] have reported TTS development and prosody incorporation research work in Yoruba and Ibibio which are languages spoken in Nigeria, Africa. They are both tone languages and hence prosody modelling is of utmost importance. Yoruba is spoken by about 28 million people, and Ibibio is being spoken by about 2 million people. But even then not much has been done to build computational resources for them. For the 171 living Philippine languages, there is no standard database available [17]. Due to this the researchers working in the language have to travel to the locality where the language is spoken, collect data from native speakers and then continue the research. There is an ongoing project to collect the corpora (Philippine Languages Online Corpora-PLOC), but it has been completed only for 8 languages. Even though the internet contains a plethora of text from many languages, a majority of world languages are not well represented in it. The Amharic language spoken in Ethiopia is one such language. An effort to build relevant corpus for Amharic has been carried out by Pellegrini et.al [18]. Celtic languages are Irish, Welsh, Breton, Scottish Gaelic, Cornish and Manx which are used in some parts of Europe. In [19], Delyth describes the difficulty in developing speech technology in these languages due to the non-availability of standard speech corpus. There are no newspapers and broadcast radio availability is also very limited. Only source of written speech data is from the internet, which are mainly translations from English. All this prevents any significant research for a TTS system in Celtic language

India has 22 official languages which are widely used in different parts of the country, but they are all low resource languages. An effort was taken by Hemant et.al. [21] to develop speech corpus for 13 of these languages. In the work reported in [22], a Kannada language TTS system is developed in Festival framework by first converting Kannada text to English and then using the TTS system. The work in Marathi language also reports the requirement of a larger and more standardized database to develop TTS system with good naturalness in the output speech [20]. There is a lack of standard databases in Indian languages. Efforts are being taken by various researchers, but they all have to be synchronized for working towards a common goal.

By investigating the status of Text To Synthesis in minor languages, it is very evident that there is a predominant "language divide" [7] in terms of the significant technological advancement and reachability to all the people in need. To bridge this divide, the only solution is to encourage research in these languages and attempt to develop standard resources for all languages. This has motivated the authors of this work to conduct prosody studies in Malayalam, a low-resourced Indian language.

4. TTS in Malayalam Language

Malayalam is one of the 22 official languages in India with 38 million native speakers particularly in the state of Kerala and the Laccadive Islands. There have been some attempts to develop TTS in Malayalam language. In [23] attempt is made to model the duration pattern of Malayalam speech in Festival speech engine. [24] reports using cluster analysis of duration patterns to improve the quality of Malayalam TTS. Arun et.al. explains an attempt to implement a concatenative synthesis method using ENSOLA technique [25]. But these developed systems and analysis have not produced TTS systems with the required speech quality and naturalness. This is because there is paucity in prosodic modeling, non-availability of standard databases, lack in enhancement of database to handle bilingual text for speech generation etc.

Prosody is defined as patterns of intonation and stress in a language and the various prosodic features are pauses, syllable prolongations, overall timing structure etc. The human brain introduces these features into natural speech to make it more understandable to the listeners; so, the inclusion of these prosodic features into synthesized speech will be similar to mimicking the human brain. The prosodic feature that is investigated and implemented in this work is "Pause". Pause is the temporary stop introduced between words, phrases, sentences or paragraphs in speech. The inclusion of pauses in speech generally has two motives: one is to take sufficient breath to speak further and second is to provide the listeners' sufficient time to decipher what was spoken. Pause is comprised of two parameters: pause position (where the pause has to be inserted in a sentence) and pause duration (for how long the pause has to be inserted). Modeling these two parameters is called Pause Modeling in speech and incorporating these pause patterns correctly will improve the overall duration model of synthesized speech.

Pause analysis has been conducted for other languages [26][27][28] like Mandarin Chinese, Japanese and Bangla, but pause patterns vary with language and speaking styles. Therefore to develop pause model for Malayalam language, speech corpus in Malayalam has to be analyzed in detail to derive the pause duration model. There has been no previous reported work in pause modeling for Malayalam. Since Malayalam is an under-resourced language, the availability of resources and standard speech corpus to suit the requirements of the study are limited. Hence such resources have to be collected and some have to be even recorded to complete the research work.

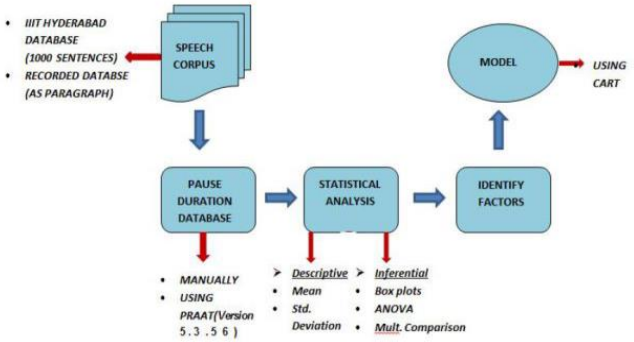

Figure 1: Methodology for pause analysis

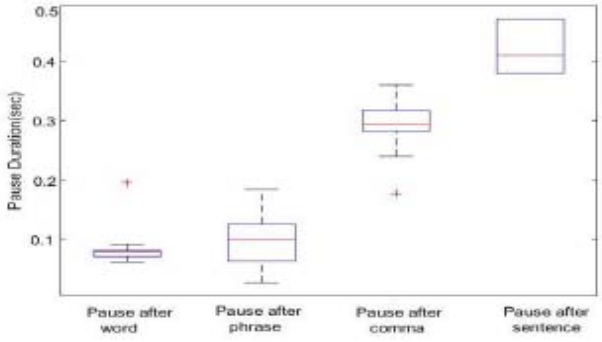

Figure 2: Box plot based on pause position (Database I)

5. Methodology

The basic methodology adopted for Pause Modelling of Malayalam TTS has been depicted in Figure 1. Getting an adequate speech corpus for analysis is the primary task. Since Malayalam is a non-standard language, the availability of a variety of databases that suit the needs of this particular study is limited. A database of 1000 Malayalam sentences compiled by IIIT-Hyderabad is used as Database I [29]. This database has 1000 .wav file recordings of individual sentences. Since the database has only individual sentences, it is inadequate for the analysis of pauses after sentence and after paragraph. So a new database (Database II) is recorded for the purpose which consists of 15 minutes (760 words, 7 paragraphs) of Malayalam speech read by a female Malayalam speaker. To analyse pauses, a pause duration database is required, and it is not existing for the language. So, a pause duration database is manually developed using the Praat speech analysis tool. Each of the sentences from the two databases are read in Praat and the pauses are manually marked and their durations measured to form the database. Along with each pause, its various parameters like location of the pause, the words/syllables before and after pauses and breath groups are also marked to form a consolidated database.

Once the pause duration database is developed, statistical analysis tools like mean, standard deviation (Descriptive analysis), box plots, ANOVA and Multiple comparison (inferential analysis) tools are employed. This analysis helps identify the various factors that affect pauses in Malayalam language. Once these factors are identified, a pause duration model is developed and an implementation of the model on Malayalam sentences is also conducted.

6. Results and Discussion

An attempt is made to identify the various factors that affect pauses in sentences. From the discussion with a linguist and reference to previous research work in other languages, the first factor studied is the Position of pause. The pause duration databases I and II are categorized in terms of the pause position and analyzed. The Database I has only individual sentences, so analysis of pause after word, phrase and comma are only possible. Database II consists of speech in the form of paragraphs, so along with the pause after word, phrase and comma; the analysis of pause after sentence and paragraph can also be conducted. The analysis tools used are box plots and ANOVA. In the box plots obtained, the medians of the boxes pertaining to each position did not fall on the same line and the overlap between the boxes is also minimal (Seen in Figure

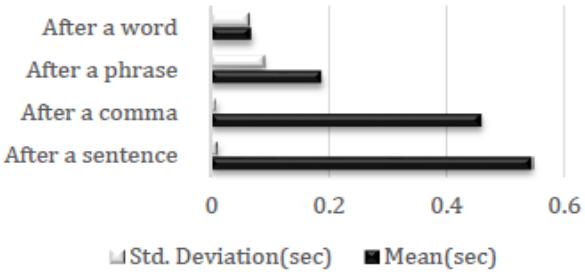

Figure 3: Descriptive analysis of pauses

2). The significance of Position of Pause as a factor for modelling can be further substantiated by the results from ANOVA analysis ($F_{3,280} = 283.2$, $p < 0.05$). This clearly establishes that the pause at different positions have values in different ranges, and hence can be used as a factor for modelling pauses.

Based on the position of pause as the primary factor, a general descriptive analysis of the pauses that occur at various positions in a sentence is conducted and depicted in Figure 3.The preliminary analysis suggests that the pause after word and phrase have high deviation values, as a result, further factors have to be identified to model them. But, pause after comma and sentence have low deviation values. So, they can be directly modelled by their mean values. [30][31]

The pause after phrase and word are investigated further by studying the databases. From inspection and discussions, more factors that affect these pauses are recognized (listed in Table I). To finalize these factors, statistical analysis tools like box plots, ANOVA were used.

Table I: Factors identifies for pause modelling

No.	Factor
1	Position of pause
2	Number of syllables before pause
3	Number of syllables in the word preceding the pause
4	Number of words before pause
5	Number of syllables after pause
6	Number of syllables in the word succeeding the pause
7	Number of words after pause

Pause Duration Model and Statistical Testing:

The identified factors are then used to model duration of pause after word and phrase using CART (Classification And Regression Tree) in MATLAB. Since the standard deviation values are low, the pause after comma and sentence are modelled directly by their mean values: 0.546s and 0.46s respectively.

Thus a complete model for pause is developed for Malayalam language.

In order to understand the effectiveness and accuracy of the model, RMSE (Root Mean Square Error) and Correlation tests are conducted. 60% of the database is used for training and 40% for testing. The developed model is implemented on the output of an existing Malayalam TTS system (in eSpeak), and the RMSE obtained is 0.025s with 90.85% Correlation in comparison with the testing corpus. This proves that the developed model predicts pause duration effectively.

7. Conclusion

An attempt is made to develop a Pause duration model for an under-resourced language. This is the first work reported in this regard for Malayalam language. Parts of the work have been reported in [30][31], but this paper focusses on the challenges in doing such a work for the first time in a non-standard language. Conducting research studies to develop TTS in under-resourced languages like Malayalam is a challenge, due to the many reasons emphasized in the Section 2 and 3 of the paper. If a user-friendly and high quality system is developed in a language, there will definitely be many users for it, especially physically challenged and aged people who prefer to communicate in their local language. A plethora of research work is happening in standard languages like English, but this is useful for only a minority of people who are well-versed in these languages. This leaves a very wide population still challenged because of their physical disabilities, even though there is such a huge advancement in technology. Also, researchers who attempt to work in the field of non-standard languages are often faced with the encumbrances of rarity in standard resources, difficulty in gathering the required resources, non-availability of written material in the language, inconvenience of traveling to the particular region to learn about the language etc. Researchers should be enthusiastic to work in their own languages and a Text To Speech technology accessible to all without any language divide should be the focus of future research.

8. References

[1] The First Writing, Script Invention as History and Process,S.Houston, ed. Cambridge Press, 2008.

[2] Jean-Marie Favre, Dragan Gasevic,Ralf La¨mmel,AndreasWinter, “Guest Editors’ Introduction to the Special Section on Software Language Engineering”,IEEE Transactions OnSoftware Engineering, December 2009

[3] UNESCO Ad Hoc Expert Group on EndangeredLanguages, “Language Vitality and Endangerment “

[4] Kevin P. Scannell “The Crúbadán Project: Corpus buildingfor under-resourced languages” in the Proceedings of the 3rdWeb as Corpus Workshop

[5] Steven Krauwer “The Basic Language Resource Kit(BLARK) as the First Milestone for the Language ResourcesRoadmap" in the Proceedings of SPECOM 2003

[6] L. Besacier, B. Zhou,Y. Gao, “Towards speech translationof non written languages,” in Proc. SLT Workshop, Aruba,2006, pp. 222–225.

[7] Laurent Besacier , Etienne Barnard , Alexey Karpov , TanjaSchultz, ”Automatic speech recognition for under-resourcedlanguages: A survey “, Speech communication, 2014

[8] T Dutoit, “An Introduction to Text-to-Speech Synthesis”,Kluwer Academic Publishers.

[9] Statitics from IAPB(The International Agency forthe Prevention of Blindness)

[10] CDC/NCHS, National Health Interview Survey, 2012

[11] Rie Tamagawa ,Catherine I. Watson, I. Han Kuo ,Bruce MacDonald,Elizabeth Broadbent,”TheEffects of Synthesized Voice Accents on User Perceptions of Robots”,International Journal of Social Robotics , 2011

[12] Jane M. Carey, Charles J. Kacmar “Cultural andLanguage Affects on Technology Acceptance and Attitude: Chinese Perspectives” International Journal of Information Technology,2010

[13] Harold Somers,” Faking it: Synthetic text-to-speechsynthesis for under-resourced languages – Experimental design“ACL Anthology”

[14] Odéjobí, O.A., Wong, S.H.S., Beaumont, A.J.,”A modularholistic approach to prosody modelling for Standard Yorubaspeech synthesis” , Computer Speech and Language, 2008

[15] Ekpenyong, M., Udoh, E., Udosen, E., Urua, “Improvedsyllable-based text to speech synthesis for tone languagesystems”, Lecture Notes in Computer Science, 2014

[16] Viet-Bac Le and Laurent Besacier, “Automatic SpeechRecognition for Under-Resourced Languages: Application toVietnamese Language” IEEE Transactions On Audio, Speech,And Language Processing, 2009

[17] Shirley Dita , Rachel Edita O. Roxas,” PhilippineLanguages Online Corpora: Status, issues, and prospects”, ACLAnthology

[18] T. Pellegrini and L. Lamel, “Investigating automaticdecomposition for ASR in less represented languages,” in Proc.ICSLP’06, Pittsburgh, PA,2006

[19] D. Prys, “The BLARK Matrix and its relation to thelanguage resources situation for the Celtic languages,” in Proc.LREC’06, Genova, Italy,2006.

[20] Mr.S.D.Shirbahadurkar,Dr.D.S.Bormane,”Marathi Langu- ge Speech Synthesizer Using Concatenative Synthesis Strategy(Spoken in Maharashtra, India)”, Second InternationalConference on Machine Vision , 2009

[21] Hemant A Patil, “A Syllable-Based Framework for UnitSelection Synthesis in 13 Indian Languages”, OrientalCOCOSDA ,2013

[22] Anusha Joshi, Deepa Chabbi , Suman M and SupritaKulkarni, “Text To Speech System For Kannada Language”IEEE ICCSP, 2015

[23] Bindhu;V. Rijoy;Deepa;Nimmy“Duration modeling for text to speech synthesis system using festival speech engine developed for Malayalam language”, (ICCPCT), 2015

[24] K. S. Sreelekshmi ; Deepa P. Gopinath “Clustering ofduration patterns in speech for Text-to-Speech Synthesis” ,IEEE INDICON, 2012

[25] Arun Gopi,Shobana Devi Sajini,Bhadran, “Implementationof malayalam text to speech using concatenative based TTS forandroid platform” , IEEE ICCC, 2013

[26] Hiroya Fujisaki, Sumio Ohno, Seiji Yamada, “Analysis Of Occurrence Of Pauses And Their Durations In Japanese Text Reading”, The 5th International Conference on Spoken Language Processing December, 1995

[27] Sudipta Acharya, Shyamal Kr.Das Mandal, “Occurrence And Duration Modeling Of Sentence Medial Pause For Bangla Text Reading At Different Speech Rate”, in proc. of (Oriental COCOSDA), IEEE,2012

[28] Jian Yu , Jianhua Tao , “The Pause Duration Prediction for Mandarin Text to- Speech System”, in proc. of IEEE NLPKE,2005

[29] Kishore Prahallad, E.Naresh Kumar, Venkatesh Keri,S.Rajendran, Alan W Black, “The IIIT-H Indic Speech Databases,” INTERSPEECH, 2012

[30] James, J., Gopinath, D.P.,“Modeling pause duration for Malayalam language TTS”,IEEE ICALIP, 2014

[31] Jesin James, Deepa P. Gopinath, “Pause Duration Model for Malayalam language TTS”IEEE ICACCI, 2015.

[32] Hirshberg,Wang, "Predicting Intonational Phrasing fromText",Proceedings of the 29th annual meeting on Associationfor Computational Linguistics,1991