International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 28935-28944
© Research India Publications
http://www.ripublication.com

An Improved Tamil Text To Speech Synthesis System With Annotated Dual Corpora

Kiruthika S
Department of CSE
King College of Technology, Namakkal
Tamil Nadu, India
Krishnamoorthy K
Department of CSE
Sudharsan Engineering College, Pudukottai
Tamil Nadu, India

Abstract

This paper describes the effective implementation of a Tamil Text to Speech Synthesis system by using an annotated dual corpora model. One of the corpora is a repository of annotated syllable units; while the other one is the repository of annotated diphone units. A syllable-like unit is chosen primarily because Tamil language is syllable centered. But polysyllables, as the stand-alone units, failed to bring the TTS betterment in following areas like sentence termination, scientific notations, website link and email addresses. This is because concatenation at syllable boundaries may lead to smaller errors. The lagging fields can be effectively processed by units of diphones. The idea of two different speech corpora forms an effective implementation of Tamil Text to Speech Synthesis System, wherein the naturalness and intelligibility of synthesized speech is attained. All the functional aspects of prosody are dealt to perform annotations. Effective models are adapted for phrasing the given text into chunks of appropriate syntactic categories at various levels. ToBI is kept as the standard for modeling the Intonational aspects of Prosody. CART model is followed for addressing the Duration aspects. All these aspects are met, only if high level linguistic information are properly extracted from the text data. This research brings utmost effective synthesis of speech by addressing all aspects of prosody that attain the naturalness.

Keywords: Prosody; Speech Synthesis; Dual corpora; Annotation

Introduction

The efficiency of any Text to Speech Synthesis System depends on the naturalness of the speech which on the other way depends on the intelligibility of the synthesizer. Designing a synthesizer, which is intelligent enough to handle the minute linguistic properties of a language, is somewhat hard. Dravidian languages like Tamil, Turkish, Japanese and Estonian languages show linguistic richness, thus the development of an intelligent synthesizer is really a difficult task. In Tamil language, for example, all plosives of a given place of articulation are represented by a single grapheme. The pronunciation of such graphemes depends on the context[11]. Therefore an annotated speech corpus which may contain optimal amount of prosodic information will help to deal with Text to Speech Synthesis system of such languages. So we follow a dual corpora based synthesizer.

Limited Prosody Indicators In Tamil Language

Tamil is a phonetically rich language. It is a well-known language for some of its salient accents that comprise the phonetic richness of the language. But it has a very limited amount of prosody indicators. Tamil language has a very little or almost no punctuation aspects. This sometimes leads to a very complex structuring of a sentence, which again leads to a heavy prosody tagging. This needs a virtual punctuating scenario preceded by a complete structural phrasing of the given sentence. This feature of the language makes prosodic phrasing somewhat complicated.

When the punctuation aspect is like this, the syntactic aspect of the language is quiet good. Every word of the language viz., a noun, verb or adjunct, has a morpheme affixed to the root of that individual word. Thus Tamil language is said to be agglutinative in nature.

Meeting The Functional Components of Prosody

The three main functional components of Prosody are:

1. Phrasing
2. Duration
3. Intonation

Of course there are some more prosodic features like Tempo, Melody, Pitch and Rhythm. The features of pitch are addressed by intonation itself. Tempo, melody and rhythm do not play significant role in our Text to Speech Synthesis System. Therefore, the components that are listed above are to be met by our Text to Speech Synthesis System.

- Phrasing is carried out various levels such as sentence, syntactic phrase, semantic phrase and polysyllables. Diphones are separately extracted at special places of articulations viz., sentence boundaries, website addresses, email ids, numerals, etc.

- Duration modeling is carried out using [12] CART (Classification And Regression Trees) model.
- Intonation is carried out using original English ToBI (Tones and Break Indices) conventions.

A. Phrasing

Phrasing needs parsing the text into various levels of articulation, viz., sentences, syntactic and semantic patterns. The text is parsed into sentences at the first level. Each sentence is parsed into syntactic patterns. The purpose of POS tagging is to find out the syntactic category of a word in a sentence. Table 1 gives the list of POS tags [13] and their usage.

Table 1: POS Tags and their descriptions

S.No.	**Tag**	**Description**
1	N	Noun
2	NP	Noun Phrase
3	NN	Noun + noun
4	NNP	Noun + Noun Phrase
5	IN	Interrogative noun
6	INP	Interrogative noun Phrase
7	PN	Pronominal Noun
8	PNP	Pronominal Noun Phrase
9	VN	Verbal Noun
10	VNP	Verbal Noun Phrase
11	Pn	Pronoun
12	PnP	Pronoun Phrase
13	Nn	Nominal noun
14	NnP	Nominal noun Phrase
15	V	Verb
16	VP	Verbal phrase
17	Vinf	Verb Infinitive
18	Vvp	Verb verbal participle
19	Vrp	Verbal relative participle
20	AV	Auxiliary verb
21	FV	Finite verb
22	NFV	Negative Finite verb
23	Adv	Adverb
24	SP	Sub-ordinate clause conjunction Phrase
25	SCC	Sub-ordinate clause conjunction
26	Par	Particle
27	Adj	Adjective
28	Iadj	Interrogative Adjective

29	Dadj	Demonstrative Adjective
30	Inter	Intersection
31	Int	Intensifier
32	CNum	Character number
33	Num	Number
34	DT	Date Time
35	PO	Post Position

For example, the following sentence can be tagged as below:

inRu vIttil niRaiya vElaikal uLLana.

N NP Adj N VP

Sub tags are also there describing other grammatical details of the words, such as tense, number, person and case of the words. The Sub-tags are:

1. Tense: Prs - present tense, Pst – past tense, Fut – future tense
2. 1,2,3 – first, second, and third person respectively,
3. S – singular Pl – plural,
4. Neg – negative, acc – accusative case, dat – dative case

Therefore, the above said tagged sentence is tagged as below:

inRu <N><Prs>
vIttil <NP>
niRaiya <Adj>
vElaikal <N><Pl>
uLLana. <VP><Pl>

Followed by POS tagging, syllable extraction has to be done. Prediction of syllable or diphone units is based on the phrase frequency occurrence in corpus table. It is a higher risk to find phrase boundaries in Dravidian Tamil language because of non occurrence of case markers. However it can be done manually with the help of Morpheme tags. Morpheme tags in Tamil used to tag the phrase correctly and to predict whether it could be processed by syllable or diphone. The Morpheme tags help us to predict the phrase boundaries based on the occurrence of phrases and also it helps us to predict the synthesis mode of each and every phrase.

Table 2: Predicting occurrences of morphemes in Tamil

S.No	MORPHEME	WORDS	OCCURRENCES	PERCENTAGE
1	Adhu	korpAdhu	65	0.65%
2	Kal	VaazthukKal	248	2.48%
3	L	PoruL	1648	16.4%
4	Da	AnbuDan	945	9.4%
5	Um	SudUm	644	6.44%
6	Illai	paarkavIllai	453	4.5%
7	Ven	varuVen	125	1.2%

Here table 1 gives the correlation between predicted morphemes in actual words and it occurrences by CART model.

B. Duration

Duration modeling is carried out by using CART (Classification And Regression Trees) Technique. 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 [12]. CART modeling is an optimal technique for Duration modeling in Indian Languages, where there are multiple duration patterns for a single morpheme.

A Classification And Regression Tree is a self learning Tree Structure. Like any other binary tree, a CART also have a parent node, which can generate only two child nodes; recursively each child node may act as a parent node to create two different child nodes. Those nodes which do not have children are called leaf nodes. Learning in a CART is done using binary questions that start at the root node. The answer for the binary question will select a child node; another binary question and its answer will lead to the next level till the leaf node is reached. Fig1 represents a CART used for Duration modeling of the proposed Text to Speech Synthesis System. The triangles in the given tree represent the leaf nodes.

Clustering is carried out for capturing the gross acoustic properties of the syllables. The phrase boundaries have a major role in fluent connected speech. The energy contours within a phrase vary depending on the relative position of the phrase in the utterance. Therefore, before entering into the CART, syllables are clustered using the following features [14]:

- Word length of the neighboring words in terms of number of syllables constituting the word.
- Distance of the syllable from the beginning of the phrase in terms of number of words and number of syllables.
- Distance of the syllable from the end of the phrase in terms of number of words and number of syllables.
- Relative position of the parent phrase in the utterance.
- Position of the syllable with respect to phrase boundary.
- Identity of neighboring syllables.
- Previous syllables’ features as defined in the syllable set.

C. Intonation

Intonation modeling is carried out by following the ToBI (Tones and Break Indices) standard prescribed for original English. There are no specific Tones and Break Indices have been tabulated for Tamil Language. So English ToBI guidelines are followed. There are two simple tone indices H for High tone; and L for low tone. An asterisk (*) indicates which tone is aligned with the stressed syllable of the word. Rising tone is indicated by L+H; and a falling tone is indicated by H+L. With these basic tone indices, the following pitch patterns are used for modeling. Table 3gives the standard pitch accent patterns that can be used for our system.

Table 3: Tobi Transcription For Pitch Accent Patterns

S.No.	Pitch Pattern	Description
1	H*	Pitch peak
2	L*	Pitch trough
3	L+H*	Rising peak accent
4	L*+H	Scooped accent
5	H+!H*	High pitch unaccented

H and L indicate relative high pitch and low pitch in the intonation contour. Their actual phonetic realization is conditioned by a number of factors, such as pitch range and preceding pitch accents in the phrase. Since Tamil is a kind of post-lexical language, pitch accents occur at stressed syllables. They form their own characteristic patterns in the pitch contour. Basic syntactic information of POS of words in a sentence is considered for forming rules for pause insertion. The prosodic structure of a sentence can be represented by different levels of break markers in the model.

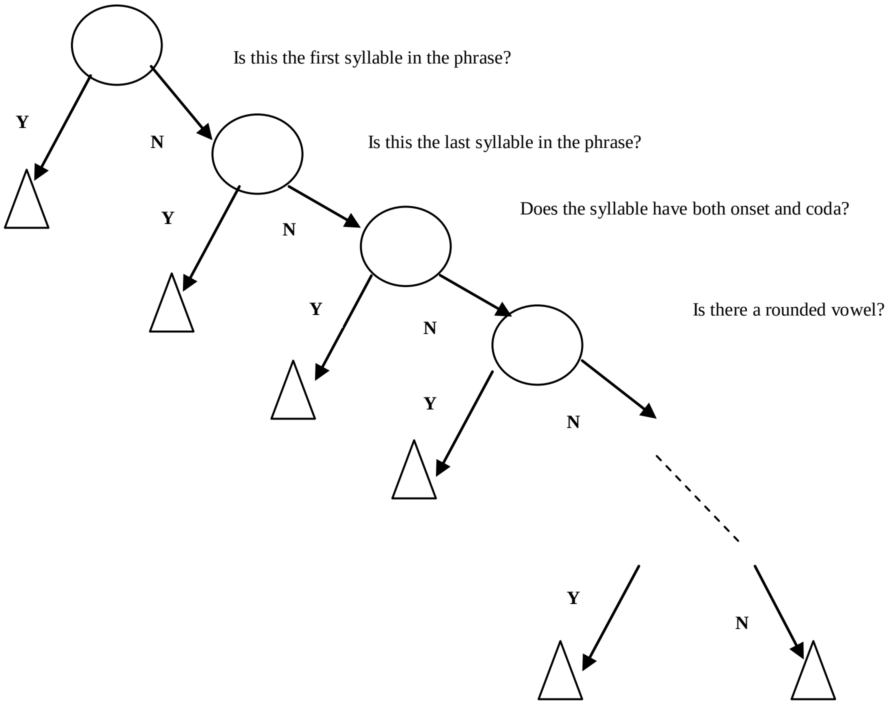
Is this the first syllable in the phrase?
Y
N
Is this the last syllable in the phrase?
Y
N
Does the syllable have both onset and coda?
Y
N
Is there a rounded vowel?
Y
N
Y
N

Figure 1: A Classification And Regression Tree used for Duration Modeling

System Architecture

The system architecture deals majorly with the formation of speech database. The classical method of construction of formal speech corpora needs the following necessary steps:

- Selection of appropriate textual content
- Recording of Textual content
- Parsing the recorded speech content at various levels
- Annotating features to the speech units
- Storing those annotated units into the corpus

The proposed synthesizer is designed to use a dual corpus. One of the corpora is a repository of annotated syllable units; while the other one is the repository of annotated diphone units. Therefore, we need to distinguish normal text and special words viz., website addresses, email ids, numerals, etc. The selected textual content converted into spoken content. The morpheme analysis phase takes care of distinguishing normal text and special words.

The parsing phase divides the content into various levels such as paragraphs, sentences, phrases, syllables and of course diphones at required instances. POS Tagging is done for each phrase. Syllables are segmented from the phrases for further processing. The general format of an Indian language syllable is C*VC*, where C is a consonant, V is a vowel and C* indicates the presence of 0 or more consonants. There are about 35 consonants and 18 vowels in Indian languages. There are defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. A rule based morpheme to syllable converter is used for syllabification [10]. Scientific notations, website link, email address, end of sentences, stress notes are to be processed by diphone units and need to be concatenated with the already processed syllable unit.

The Prosody Annotation phase deals with modeling the various components of prosody such as Phrasing, Duration and Intonation. Phrasing starts with POS tagging, where there are 35 main tags and about 5 sub tags. Furthermore, phrase boundary detection may also be carried out for the complete phrasing scenario. CART technique is used for duration modeling of the phrased units, which is associated with an appropriate clustering algorithm. Tones and Break Indices guidelines for original English are followed for Intonation modeling, since there are no ToBI model exists specially for Tamil. Fig 2 explains the Architecture of the system.

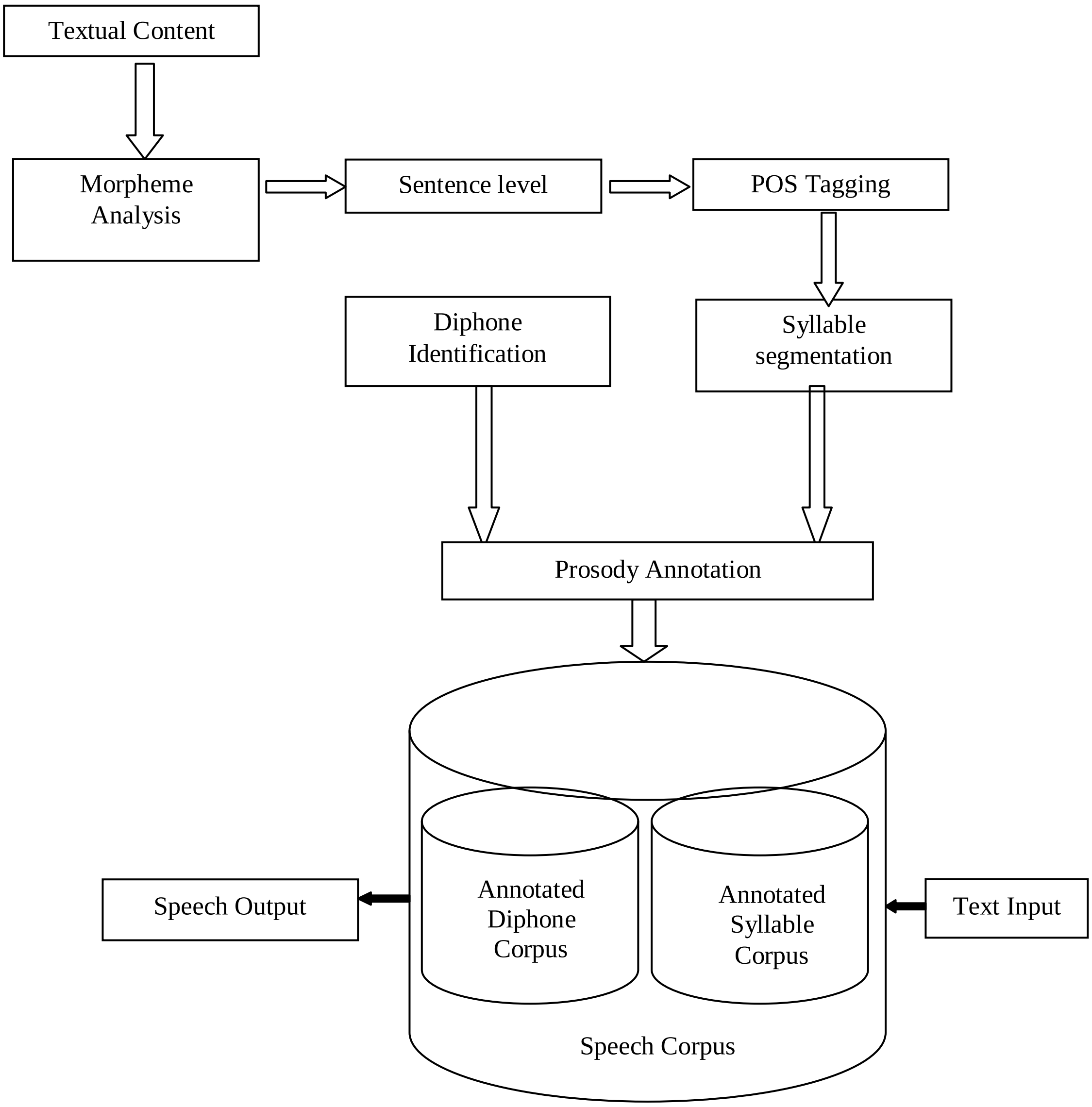

Figure 2: The System Architecture

Conclusions

The proposed system architecture, which implicitly follows a hybrid Prosody model which incorporates the best features of varoius techniques followed for Indian language Text to Speech Synthesis systems. The annotated dual corpus improves the efficiency of the system to an appreciatable extent.

References

[1] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, ―Text-to-speech synthesis using syllablelike units, in National Conference on Communication, Kharagpur, India, Jan 2005, pp 277-280.

[2] R. Thangarajan, A.M. Natarajan and M. Selvam, ―Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language, in WSEAS Transactions on Signal Processing, Issue 3, Volume 4, March 2008.

[3] Alan W. Black and P. Taylor, ―Automatically clustering similar units for unit selection in speech synthesis, Proc. EUROSPEECH 97, Rhodes, Greece, 1997, Vol. 2.

[4] Youngim Jung, Hyuk-Chul Kwon, ―Consistency Maintenance in Prosodic Labeling for Reliable Prediction of Prosodic Breaks, in the Proceedings of the Fifth Law Workshop (LAW V), Portland, Oregon, 23-24 June 2011.

[5] Vinodh M Vishwanath, Ashwin Bellur, Badri Narayan K, Deepali M Thakare, Anila Susan, Suthakar N M and Hema A Murthy,―Using Polysyllabic units for Text to Speech Synthesis in Indian languages, Proceedings of National Conference on Communication (NCC),pp.1-5, 29-31 Jan. 2010

[6] Kishore Prahallad, Arthur R Toth, Alan W Black, ―Automatic Building of Synthetic Voices from Large Multi- Paragraph Speech Databases, in Proceedings of Interspeech, Antwerp, Belgium 2007.

[7] Kiruthika S, Krishnamoorthy K, "Combining Syllable and Diphone units for a Corpus based Tamil Text to Speech Synthesis System", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.34 (2015)

[8] T.Jayasankar, R.Thangarajan, J.Arputha Vijaya Selvi, ―Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis, in International Journal of Computer Applications (0975 – 8887), Volume 25– No.1, July 2011.

[9] S. Saraswathi and T.V. Geetha, ―Design of language models at various phases of Tamil speech recognition system‖, International Journal of Engineering, Science and Technology Vol. 2, No. 5, 2010, pp. 244-257.

[10] Kiruthika S, Krishnamoorthy K, "Annotating Speech Corpus for Prosody Modeling in Indian Language Text to Speech Systems", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012, ISSN (Online): 1694-0814

[11] Shweta Vikram, “Morphology: Indian Languages and European Languages”, International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013, ISSN 2250-3153.

[12] N. Sridhar Krishna, Hema A. Murthy, “Duration Modeling of Indian Languages Hindi and Telugu”, Proceedings of 5 th ISCA SSW.

[13] S. Lakshmana Pandian and T. V. Geetha, “Morpheme based Language Model for Tamil Part-of-Speech Tagging”, Polibits, 38, pp. 19-26, 2008.

[14] Ashwin Bellur, K Badri Narayan, Raghava Krishnan K, Hema A Murthy, “Prosody Modeling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil”, DOI: 10.1109/NCC.2011.5734737, 1-2011.