A Syllable-Based Framework for Unit Selection Synthesis in 13 Indian Languages

Hemant A Patil[1], Tanvina B Patel[1], Nirmesh J Shah[1], Hardik B Sailor[1], Raghava Krishnan[2], G R Kasthuri[2]
T Nagarajan[3], Lilly Christina[3], Naresh Kumar[4], Veera Raghavendra[4], S P Kishore[4],
S R M Prasanna[5], Nagaraj Adiga[5], Sanasam Ranbir Singh[5], Konjengbam Anand[5],
Pranaw Kumar[6], Bira Chandra Singh[6], S L Binil Kumar[7], T G Bhadran[7], T Sajini[7], Arup Saha[8], Tulika Basu[8]
K Sreenivasa Rao[9], N P Narendra[9], Anil Kumar Sao[10], Rakesh Kumar[10], Pranhari Talukdar[11], Purnendu Acharyaa[11],
Somnath Chandra[12], Swaran Lata[12], Hema A Murthy [2]

[1]Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India,
[2]Department of Computer Science and Engineering, IIT Madras, India,
[3]SSN College of Engineering, India,
[4]International Institute of Information Technology, Hyderabad, India,
[5]Indian Institute of Technology, Guwahati, India,
[6]CDAC, Mumbai, [7]CDAC, Trivandrum, [8]CDAC, Kolkata, India,
[9]Indian Institute of Technology, Kharagpur, India,
[10]Indian Institute of Technology, Mandi, India,
[11]University of Guwahati, India,
[12]Technology Development for Indian Languages, Ministry of Information Technology, Govt. of India, New Delhi

{hemant_patil, tanvina_bpatel, nirmesh_shah, hardik_sailor}@daiict.ac.in, raghav1105@gmail.com, grkasthuri@gmail.com, nagarajant@ssn.edu.in, lilly.christina88@gmail.com, elluru.nareshkumar@gmail.com, raghavendra.veera@gmail.com, kishore@iiit.ac.in, {prasanna, nagaraj, ranbir}@iitg.ac.in, konjengbam.nitdgp@gmail.com, {pranaw, bira, slbinil, bhadran, sajini, arup.saha, tulika.basu}@cdac.in, ksrao@iitkgp.ac.in, narendrasince1987@gmail.com, anil@iitmandi.ac.in, mintu.rakesh@rediffmail.com, phtassam@gmail.com, pbacharyaa@gmail.com, schandra@deity.gov.in, slata@mit.gov.in, hema@cse.iitm.ac.in

***Abstract*—In this paper, we discuss a consortium effort on building text to speech (TTS) systems for *13* Indian languages. There are about *1652* Indian languages. A unified framework is therefore attempted required for building TTSes for Indian languages. As Indian languages are syllable-timed, a syllable-based framework is developed. As quality of speech synthesis is of paramount interest, unit-selection synthesizers are built. Building TTS systems for low-resource languages requires that the data be carefully collected an annotated as the database has to be built from the scratch. Various criteria have to addressed while building the database, namely, speaker selection, pronunciation variation, optimal text selection, handling of out of vocabulary words and so on. The various characteristics of the voice that affect speech synthesis quality are first analysed. Next the design of the corpus of each of the Indian languages is tabulated. The collected data is labeled at the syllable level using a semiautomatic labeling tool. Text to speech synthesizers are built for all the *13* languages, namely, Hindi, Tamil, Marathi, Bengali, Malayalam, Telugu, Kannada, Gujarati, Rajasthani, Assamese, Manipuri, Odia and Bodo using the same common framework. The TTS systems are evaluated using degradation Mean Opinion Score (DMOS) and Word Error Rate (WER). An average DMOS score of $\approx$ *3.0* and an average WER of about *20* % is observed across all the languages.**

Keywords: Indian languages, Text-to-Speech (TTS), text optimization, speaker selection, recording, labeling, pronunciation dictionary.

I. Introduction

The languages of India belong to several language families, the major ones being the Indo-Aryan languages (a sub-branch of Indo-European languages) spoken by *72 %* of Indians and the Dravidian languages spoken by *25 %* of Indians. The *3 %* of the population who speak other languages belong to the Austroasiatic, Tibeto-Burman, and a few minor language families and isolates [1], [2]. Though various Indo-Aryan and Dravidian languages may seem mutually exclusive when first heard, there is a much deeper underlying influence that both language families have had on each other down to a linguistic science [2]. The distribution of the various language families of the South Asian subcontinent is available at [3]. Hindi is the most widely spoken language and primary tongue of *41 %* of the people; there are *21* other official languages, *viz.*, Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Malayalam, Kannada, Odia, Punjabi, Assamese, Bodo, Kashmiri, Sindhi, Punjabi, Dogri, Konkani, Maithili, Nepali, Santhali, and Sanskrit. Most languages in the Indian Republic are written in Brahmi-derived scripts such as Devanagari, Tamil, Telugu, Kannada, Odia, Eastern Nagari, Assamese or Bengali, etc.

Indians account for *700* million speakers of Indo-Aryan languages and *200* million speakers for Dravidian languages [2], which form about one sixth of the population of the world. Speech and language technologies for Indian languages are not

only topical but also imminent. We present here an initiative by *12* institutions across the country for building speech synthesis systems in Indian languages. The languages for which text-to-speech synthesis systems are developed are Hindi, Marathi, Bengali, Gujarati, Rajasthani, Assamese and Odia belonging to the Indo-Aryan languages, Tamil, Telugu, Malayalam and Kannada belonging to the Dravidian languages and Manipuri and Bodo belonging to the Tibeto-Burman language.

Early efforts in building TTS systems for Indian languages were based on synthesis by rule. These were primarily knowledge-based systems and required a large amount of effort and time in deriving *coarticulation* and *prosodic* rules for each Indian language [4]. In fact, the study reported in [4] represents the first Akshara-based TTS system in an Indian language. A formant-based speech synthesis system for Hindi was developed in [5]. This system is based on the Klatt synthesis system for English [6]. In [7], attempts have been made to capitalize on the common phone-base across several Indian languages to build a multilingual speech synthesis system using the festival framework [8]. Extensive prosodic analysis for each of the languages was performed using a broadcast news corpus [9]. The rules were then incorporated in the speech synthesis system. There have also been some efforts in using unit-selection synthesis techniques for building TTS systems [10], [11], [12], [13], [14], [15]. Apart from unit-selection synthesis techniques, statistical parametric speech synthesis synthesis systems using Hidden Markov Model (HMM) is also used for speech synthesis [16]. HMM-based speech synthesis system (HTS) systems require a question set (QS) for every language [17]. On the other hand for the syllable-based USS system, we only require positional context for syllables. Hence, in the present work, we focus here only on syllable-based USS system.

Indian Languages are known to be syllable-based [13]. Syllable as a unit for speech synthesis is confirmed by the efforts of other researchers [11], [12], [13], [18], [19], [20], [21]. In [21] a comparison is made between unit selection synthesisers based on diphones and syllables. It was observed that syllable-based synthesisers performed significantly better than diphone based synthesisers for Indian languages. Most of these efforts are based on a few of hours of speech databases. Therefore, they have not fully exploited the potential of unit-selection techniques. Moreover, the efforts have been limited to Hindi, Telugu and Tamil languages. In this paper, a conscious effort is made to arrive at a common framework for building synthesis systems for all Indian languages.

The rest of the paper is organised as follows: Section II describes the procedure for selecting text collection, optimization and syllable coverage. As online pronunciation dictionaries are seldom available for Indian languages, most words in Indian languages are out of vocabulary. In Section III we discuss some of issues in building letter to sound rules for various Indian languages. Finally in Section IV, we discuss the building of a generic text to speech synthesis system for Indian languages in the unit selection synthesis (USS) framework. Section V discusses the evaluation methodology used. The performance of the system is also evaluated in this Section. Finally, we conclude in Section VI.

II. Building Speech Corpora

Given that Indian languages are low resource languages building speech corpora involves significant effort. This includes the following tasks, namely, text selection, choice of talent and labeling. Text must have sufficient syllable coverage. The footprint of USS systems depends upon the size of the corpus. Therefore, the size of the corpus should be small enough to keep the footprint small, while at the same time large enough to include various contexts.

The text data collected should be from various fields of words and context so that a variety of context are captured in the spoken utterance. The voice talent must be chosen such that the voice is amenable for signal processing. Finally, the data has to be labeled accurately so that the synthesis is accurate. These tasks are discussed in the following sections.

A. Text Collection

The first task that needs to be performed is selection of the text corpus for recordings. It is important to ensure that the chosen text covers the most commonly used words and phrases in a given language to get good coverage of frequently used syllables. To accomplish this task, a large amount of text from the Internet was collected. This included news data, blogs and short stories in Indian languages. To ensure coverage of all domains, and given that Indian languages are low resource languages, efforts were made to generate text to cover domains not available on the web. Naturalness in Indian language speaking arises from colloquialism and the use of English words frequently in a native tongue conversation. To facilitate this, colloquial words and English words were also collected. In Indian languages words tend to be polysyllabic. This is not appropriate for TTS as the speaker tends to swallow some of the syllables. The vocabulary was chosen such that majority of the words were at most trisyllabic.

1) Text Processing and Cleaning: Indian languages have their own scripts. This is different from that of most European languages that use Latin script. As most operating systems, and small form factor mobile devices support UTF8, all text that was collected was first converted to UTF8. Conversion was required since a number of online resources are font-based. Most blogs, websites do not necessarily contain text that is grammatical correct. A manual task was undertaken to carefully clean the text. As websites include numerals and abbreviations, tools were developed to convert them to text as required by the spoken sentence.

2) Text Optimization: The ultimate goal of text optimization is to cover a large number of syllables from less text material. This processing can be done in a number of ways. Each language has performed text optimization in various ways. Hence a common description is ideally not possible. The text optimization procedure from various languages is presented here in Table I. Nevertheless, for all Indian languages it was observed that children's stories are not heavily polysyllabic, thus making them suitable for Indian language synthesisers. As is expected as the coverage increases, the availability of new syllables is also very small. The vocabulary chosen for optimal text selection varies from about *50,000* words to about *3.5* lakh words.

TABLE I. TEXT SELECTION AND OPTIMIZATION METHODOLOGY CARRIED OUT FOR VARIOUS INDIAN LANGUAGES

Sr. no.	Languages	Type of Data collected	Methodology of Text Optimization
1	Hindi	Children stories as they are easier to read. Text selected to ensure maximum syllable coverage. Sentences picked such that there were no words in any any of the sentences with more than *2-3* syllables	Sentences with 5 or more new syllables were chosen first. Thereafter, sentences with *4* or more new syllables were chosen and so on. This ensured that syllable coverage was maximized in a minimum number of sentences
2	Tamil	Children stories as they are easier to read. Text selected to ensure maximum syllable coverage. Sentences picked such that there were no words in any any of the sentences with more than *2-3* syllables	Sentences with 5 or more new syllables were chosen first. Thereafter, sentences with *4* or more new syllables were chosen and so on. This ensured that syllable coverage was maximized in a minimum number of sentences
3	Marathi	Around *50* million word text corpus crawled from web, such as online news sites, blogs, etc.	Sentences of length *5* to *20* words selected from the corpus so that new selected sentence must give at least one new uncovered syllable
4	Bengali	Bengali text corpus is collected from Anand Bazaar newspaper, literature, education and story books.	OTS has been done by assigning weights to several parameters that were defined for each sentence [22].
5	Malayalam	Online news sites and few other blog sites were downloaded automatically using web crawlers. Data from domains like tourism, medicine, story, etc., were prepared manually.	Set a threshold value for the syllable frequency. Syllables below threshold were replaced with its lower syllable, phone combinations if lower syllable exist else retains the syllable. Select sentences to cover *4* apparitions of these syllables
6	Telugu	Collected most frequently used words and phrases in the language and the text from short stories	Given a list of syllables marked with the occur-rence, a minimal number of sentences were selected to get optimum coverage.It was a greedy approach. To take care of the missing syllable coverage, high frequency words were included for recording
7	Kannada	Children stories as they are easier to read. Text selected to ensure maximum syllable coverage. Sentences picked such that there were no words in any any of the sentences with more than *2-3* syllables	Sentences with *5* or more new syllables were chosen first. Thereafter, sentences with *4* or more new syllables were chosen and so on. This ensured that syllable coverage was maximized in a minimum number of sentences
8	Gujarati	Sentences having *200000*+ words from various sources such as newspaper articles, magazine articles, stories, essays from diverse groups to cover different kinds of words	Each line was scored based upon the number of new syllables in the line and the frequency of the syllables. The syllabification script is used for this purpose. The optimized text contains *100000*+ words and to record speech of about *20* hours
9	Rajasthani	Most of the stories are taken from the Rajasthani granthagar. Topic's like, essay, article, finance, sports, agriculture, medical, science and technology, festival's, political issue's are taken from Rajasthani magazine	Text is selected in story by story manner. Then from each story find the total syllable and unique syllable. Then calculate the entire unique syllable including the previous text. Thereafter, estimate the new unique syllables and calculate the percentage of useful script. Finally calculate the percentage of text(story) for recording.
10	Assamese	Collected stories from "buhri air shadhu" and "Isapar shadhu" story books and news bulletins were identified from different newspapers. Covered almost all the words that generally uttered in Assamese news reading	Text is chosen such that it gives new syllables as compared to the already existing syllables. As the database increases, rate of new syllables from a sentence decreases.
11	Manipuri	Manipuri story telling application stories were identified from the children story book Phung-gawaarii Singbul	Each sentence is chosen such that it gives maximum percentage of new syllables. The process is carried out iteratively with gradual decrease of percentage of new syllables until the text corpus contains all the unique syllables.
12	Odia	*8* lakh words text corpus built by manual typing (from novels, Odia Jahnamamun (Chandamama) stories, science, history, geography, etc.) Around *50,000* domain-specific basic words collected from popular Odia Shabdakosha by manual typing.	Approximately *2500* sentences (more than *90 %* of the syllables) of length *5* to *20* words selected from the corpus so that new selected sentence must give at least one new uncovered syllable
13	Bodo	*3.5* lakh word text corpus by manual typing taken from popular novels, science, literature books, etc. Phonetically rich sentences and words	A collection of *5000* Phonetically rich sentences (containing approximately *95%* of the syllables) were selected .Each line was scored based upon the number of new syllables in the line and the frequency of the syllables

3) Syllable Coverage: A syllable is a speech sound unit having a vowel at the nucleus surrounded by none or more consonants. A syllable is typically of the form *C*VC**, where *C* is a consonant, *V* is a vowel and *C** indicates there may be none or more consonant present. Indian languages are syllable-timed, in that the syllables in Indian languages can be thought of as being of approximately the same tempo at the syllable-level. This will change with the speaking rate. Syllables are large speech sound units that can accommodate the acoustic variability of speech and still small enough to be modeled. Table II shows the syllable coverage in the raw text and the syllable coverage in the optimized text.

B. Recording of Speech Database

1) Voice Artist Selection: To record the speech database, a process of selecting a voice talent (speaker) was carried out. Choosing a voice talent is not an easy task. The first task is to choose a professional voice talent who can speak with the same tempo. News readers, radio jockeys are preferred for this tasks. The gender, age group and voice quality have to be considered in selecting a good voice talent. This process then involved selection from a group of *3-5* potential voice talents for both the female and male artists (professional radio jockeys were chosen for this purpose). Thereafter, *10-15* minutes of speech from each voice talent was recorded. The recorded speech was subjected to signal processing algorithms such as *duration* and *intonation* modification using standard signal processing toolkits [23], [24]. It was also taken into consideration that

TABLE II. STATISTICS OF THE TOTAL COLLECTED TEXT CORPUS FOR *13* INDIAN LANGUAGES BEFORE AND AFTER TEXT OPTIMIZATION

Sr. no.	Languages	Raw Text				Optimized Text				
		Number of sentences	Total words	Unique words	Number of unique syllables	Number of sentences	Total words	Unique words	Number of unique syllables	Percentage of syllables covered
1	Hindi	55000	582512	54050	7100	1900	40666	9560	6193	87.23
2	Tamil	75000	786548	68000	8340	2497	41259	22459	7184	86.13
3	Marathi	1519950	16419046	578644	9180	5386	51881	24177	8137	88.44
4	Bengali	50000	538124	78716	4374	7762	84206	22371	4374	100
5	Malayalam	378654	3560283	496682	13683	4232	46008	25291	6591	48.16
6	Telugu	4643	110241	32528	1614	4643	110241	32528	1614	100
7	Kannada	41037	365399	55479	4182	1279	16621	10321	3706	88.62
8	Gujarati	17000	202847	34876	5698	8450	104526	26573	5438	95.43
9	Rajasthani	6926	67758	15455	4215	2851	33983	9978	3352	94.72
10	Assamese	1806	32721	10487	3474	1291	27047	9476	3460	99.59
11	Manipuri	2007	26028	9607	2337	963	13095	6431	2337	100
12	Odia	35404	737654	67678	5803	2973	29948	11151	4065	70.04
13	Bodo	15000	162400	18436	3134	8123	144526	17152	2913	94.98

the consistency in quality rate and amplitude was maintained. Finally, a speaker was selected whose voice seemed pleasant to listen to as well as amenable to signal processing.

2) Speech Data Recording: After the collection of text data and the selection of voice artist (for male and female), the selection of studio, recording setup and the recording of the voices were carried out. A speech corpus of $\approx$ *5-10* hours of data has been recorded for all the languages. The type of recording was mono, sampling rate was *48* kHz and the number of significant bits per sample was *16*. The following issues were taken care during recording of speech data for both male and female artists.

- *Appropriate recording room:* Professional recording studios were used for the task.
- *Transcription errors:* Sometimes voice talents tend to speak different from the text. The text was then corrected to ensure consistency between the spoken sentences and the written sentences.
- *Pronunciation errors:* Users have a tendency to pronounce words in their native fashion. Pronunciations were therefore corrected either by changing the text or asking the speaker to re-record the text.
- To avoid speaker fatigue, breaks were included between recordings. During each recording session the speaker was played some sentences from earlier recordings to ensure that the same tempo was preserved in subsequent recordings.
- Although optimal text selection is good for reducing the size of the corpus, voice talents found it difficult to switch from one topic to another. It was ensured that topic across a paragraph was not changed during the recording.

Since in unit selection synthesisers, the entire waveform is stored, the speech utterances were downsampled to *16* kHz. This is primarily required to reduce the footprint of the synthesis system developed.

Unlike speech recognition, speech synthesisers require that the speech data be labeled accurately. Manual labeling can lead to a lot of inconsistency. To overcome this issue, two different approaches were used, one based on the traditional ergodic-hidden Markov Model (e-HMM) and the other based on a semiautomatic labeling syllable-based labeling tool kit [25].

C. Speech Data Labeling

The accuracy of the labeling depends upon the accuracy of the e-HMM (or the speech recognition system) used. The accuracy of the transcription depends critically on the amount of training data.

Low resource languages (such as Indian languages) hardly have any vocabulary independent continuous speech recognition systems that are readily available to perform transcription. As mentioned earlier, manual labeling is time-consuming and can be inconsistent across labelers. Alternatively, signal processing cues may be used to label speech data. A syllable typically consists of a vowel surrounded by a consonant. The vowels are of high energy surrounded by consonants which are of low energy. In other words, syllable boundary is a junction where coarticulation in speech reduces and again increases. The high energy regions in the short-term energy contour correspond to syllable nuclei, and the valleys at both ends approximately correspond to syllable boundaries [26]. In the context of speech synthesis, the boundaries must be accurate, or more importantly consistent. To enable this, a semi-automatic tool was developed [25] based on group delay functions. In addition to the group delay-based boundary detection, vowel onset point detection [27] and EHMM were also used [8]. The group delay-based algorithm requires a parameter called Window Scale Factor (WSF) that must tuned for different syllable rates. As the syllable-rate for TTS systems has to be more or less constant, this parameter was set once for every langauges.

Figure 1 shows the view of the DONLabel tool developed [25]. It shows different panels which includes the menu bar for playing the loaded speech file and to facilitate the viewing of the speech signal in the wave panel. The text panel shows the syllables that are generated according to the *letter-to-sound* (LTS) rules that belong to a particular language. The panel abbreviated as GD is the group delay panel. It is shown that

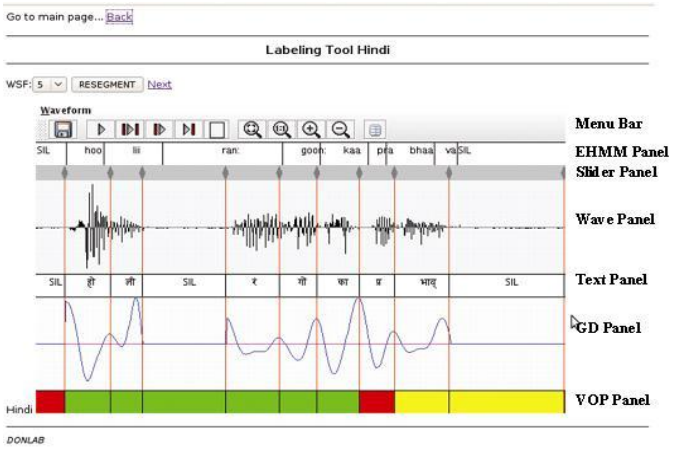

Fig. 1. The DONLabel tool [25] is used for labeling a speech data. The wave panel shows a speech utterance in Hindi Language followed by a text panel of syllabified text and the group delay (GD) panel. The panel on the top of the waveform corresponds to the labels obtained using E-HMM.

the peaks of the group delay nearly correspond to the syllable boundary [26]. However, at times, there might be minor adjustments that are needed to be made to the boundaries. These adjustments are made manually according to the text panel and after hearing the speech waveform.

The tool was found to be quite accurate for most of the Indian languages. The performance of the tool was tested on *100* sentences for the *13* different Indian languages. The percentage of correctness was calculated and the errors are computed for the best WSF. The accuracy is obtained as the percentage of correctness of the tool, i.e.,

$$\%\ of\ Correctness(PoC) = [1 - \frac{(insertions + deletions)}{Total\ no.of\ segments}] * 100. \quad (1)$$

Figure 1 shows the labels for a given sentence in Hindi that uses an appropriate WSF. The labels can be saved and stored as **.lab*. The tool works very well for languages that have word terminal syllables as vowels. The files are generated in a format that is compatible with Festival. Table III shows the percentage of correctness of the tool for *13* Indian languages. Therefore, the speech data is labeled and stored in the form of **.lab* files to built TTS systems using Festival framework [8].

D. Pronunciation Dictionary

The main aim of the database collection was development of TTS systems. Treating all words in the language as (out-of-vocabulary) OOV leads to the TTS being very slow. A pronunciation dictionary is created to facilitate fast synthesis of the speech. The pronunciations are given for every word in terms of a sequence of syllables. After the task of recording the speech data, pronunciation dictionaries were built for all the *13* languages. A large vocabulary covering a minimum of *100,000* words in UTF-8 was first generated. The words in the dictionary are alphabetically ordered. Abbreviations, numbers are made part of the pronunciation dictionary so that the words are read correctly. Table IV shows the details of the number of words that are there in the pronunciation dictionary for each of the Indian languages. It is quite common that most Indian language text will invariably contain English words. To facilitate reading of English words, the words from the CMU dictionary [28] were transliterated in terms of Indian language syllables. Occasionally English words that are not in the dictionary are spelled.

III. Letter to Sound Rules

A study of the various Indian language synthesisers proposed in this paper, suggested the use of a common Indian language parser that could parse about *90 %* of the text. Peculiar pronunciations were incorporated in the pronunciation dictionary. Additionally, to speed up the process of synthesis pronunciation dictionaries were also developed. For some languages like Tamil, the orthography is incomplete in that the same grapheme in different contexts can sound different. Appropriate tagging was employed to differentiate between the various realisations of the same sound. Sometimes the written form includes a sound, while in the spoken form, the sound is deleted. This is especially true for Hindi, Bengali, Marathi, Gujarati, Odia, Assamese, Manipuri and Bodo, where the vowel between two aksharas is deleted. Appropriate language specific rules were employed to fix this. In languages like Bengali, when a consonant cluster contains the same two consecutive consonants and is followed by the vowel /a/, the vowel is transformed to an /o/. In Telugu it is common to find vowel epenthesis for foreign words. Appropriate rules were developed for Telugu to address this. By and large it was observed that less than *20 %* rules were specific to a language. Most rules were common across all the Indian languages.

IV. Syllable based TTS within the festival framework

The text files, wav files and the corresponding generated lab files are used as input to the training phase of the festival text-to-speech synthesis (TTS) system. In order to build UTF-8 voices for Indian languages, some customisation was made to the festival framework. In addition, adaptations were made in order to accommodate syllable as the basic unit. This is primarily to enable building of meaningful classification and regression trees in the context of syllables. This implies

- Extraction of better linguistic and phonetic features which serve as questions while building Classification and Regression Trees (CART) for syllables
- Definition of appropriate acoustic distance measures to capture similarity between syllable units.

The following is a summary of modifications that were made to adapt festival to use the syllable as a basic unit. In Section IV-A, we discuss customisation of the phoneset in festival to accommodate syllable as a unit for synthesis. In Section IV-B appropriate linguistic features are discussed. These are required for building the CART. Parameters that are relevant for building the CART are discussed in Section IV-C. Punctuation is seldom used in any Indian language. Therefore, phrase boundary prediction is required. This is discussed in Section IV-D.

TABLE III. THE PERCENTAGE OF CORRECTNESS (POC) AS OBTAINED FOR THE DONLABEL LABELING TOOL *13* INDIAN LANGUAGES

Languages	Hindi	Tamil	Marathi	Bengali	Malayalam	Telugu	Kannada	Gujarati	Rajasthani	Assamese	Manipuri	Odia	Bodo
PoC	86.83	77.38	80.24	71.71	78.68	85.40	77.27	68.64	54.07	73.61	68.33	75.31	85.56

TABLE IV. THE NUMBER OF WORDS INCLUDED IN THE PRONUNCIATION DICTIONARY FOR *13* INDIAN LANGUAGES

Languages	Hindi	Tamil	Marathi	Bengali	Malayalam	Telugu	Kannada	Gujarati	Rajasthani	Assamese	Manipuri	Odia	Bodo
No. of words	1,16,205	1,12,932	1,10,627	80,000	1,03,105	1,58,493	3,37,556	1,06,739	8,834	90,444	24,000	98,309	1,62,400

A. Customizing the Phoneset

- The phoneset was modified to describe the syllables in the database.
- Phoneset as the name suggests has a default set of tags to describe phones which is generally assumed to be the basic unit within the festival paradigm. Tags were altered in order to characterise syllables
- The onset and coda of a syllable being susceptible to coarticulation with the preceding and following syllables, it is also important to keep track of the context in which syllable units stored in the repository were originally spoken. The syllables in the database were tagged based on the following characteristics:
  - Each syllable had three different identities *begin*, *middle* and *end*, depending upon its location in a word.
  - type of syllable, namely, V, C*V, VC*, C*VC*
  - long or short vowel.
  - type of vowel.
  - place of articulation of the first consonant.
  - place of articulation of the last consonant.
  - manner of articulation of the first consonant.
  - manner of articulation of the last consonant.

B. Linguistic and phonetic features for clustering

- The traditional unit used in the FestVox[8] framework is the phoneme. The default features proposed attempt to capture intra-syllable and inter-syllable relationships. For syllable based speech synthesisers more levels of hierarchy are required, with syllable being the basic unit. The features should be chosen such that the clustered units preserve the gross acoustic properties of the syllable[29] [30]. The following features are used to cluster syllables of the same type: the same type.
  - Word length (number of syllables) of adjacent words.
  - Distance of the syllable from the beginning of the phrase is defined by a two tuple $< number\ of\ words, number\ of\ syllables >$.
  - Distance of the syllable from the end of the phrase is defined by another two tuple $< number\ of\ words, number\ of\ syllables >$.
  - Relative position of the parent phrase in the utterance.
  - Position of the syllable with respect to phrase boundary.
  - Identity of neighbouring syllables.
  - The features of the previous syllable (as defined in the syllable set).
  - The features of the next syllable (as defined in the syllable set).
- Phrase boundaries must be marked in the speech corpora to determine some of these features. In the absence of a text chunker and a tagger, perceived phrase boundaries occurring in utterances in the speech corpora were labeled manually.

C. Optimizing parameters for clustering syllables

In order to better cluster syllables, in addition to redefining linguistic and phonetic features, acoustic measures also have to be optimized. Some of the parameters that were customized are:

- Cluster Size - This parameter controls the number of leaves in the CART. Syllables being larger than phones have fewer instances, the cluster size was reduced to ≈ 5.
- Duration Penalty Weight (duration_pen_weight) - The duration_pen_weight parameter is a measure of the importance given to the duration of the unit when clustering units. A high value of duration penalty ensures that syllables of dissimilar durations are clustered separately.
- Fundamental pitch penalty weight (F_0_pen_weight) - The F_0_pen_weight parameter specifies how much importance is given to F_0 while clustering units. Higher penalty weight was fixed for F_0 also.
- ac_left_context - In speech, owing to coarticulation, the preceding and succeeding units affect the production of a given unit. When diphones are used, a larger left context is used to ensure spectral continuity at the boundary between diphones. This was reduced for the case of syllables as the energy is low at the boundary in comparison with diphones. A large context was found to hurt performance.

D. Phrase boundary prediction and inserting pauses

- Phrase boundary prediction - Punctuation marks are seldom used in Indian languages. At the time of synthesis, it is difficult to comprehend a sentence properly, without pauses at the end of phrase boundaries. Also given that phrase based contextual features are used during training, it is important to predict phrase boundaries and insert pauses at the end of the phrase. For the purpose of building a CART to predict phrase

TABLE V. THE DMOS AND WER EVALUATION OF TTS SYSTEM FOR *12* INDIAN LANGUAGES

Languages	Hindi	Tamil	Marathi	Bengali	Malayalam	Telugu	Kannada	Gujarati	Rajasthani	Assamese	Manipuri	Odia
Listeners	20	20	10	20	20	20	10	20	16	20	7	10
Hours	12	12	10	10	10	10	4.5	7	3.5	4	4.5	5
DMOS	3.597	3.23	2.82	3.67	3.0907	4.26	3.16	3.06	3.51	2.91	2.5715	2.63
WER in %	7.02	7.52	10.6	3.636	8.058	26.48	40.67	13.33	28.68	27.665	23.9583	14.2

boundaries, commas are manually introduced in the text of the train data. It was seen that some languages had explicit positions where phrase boundaries occur [7], [31] (for example, case marker in Hindi). While for other languages a set of appropriate tags have to be identified (for example, Tamil) to tag words. This is to ensure correct phrase boundary prediction. The set of tags are then used as questions in CART to predict phrase boundaries.

- Handling silences - To enable insertion of pauses, only two units of silences were used: SSIL, the silence at the end of a phrase and LSIL, the silence at the end of a sentence. The silence at the end of a phrase will be of a short duration while the silence at the end of a sentence will be of a long duration.
- Geminates - In some Indian languages it is very important to preserve the intra-word pause while speaking, as the word spoken without the intra-word pause would have a completely different meaning. These intra-word pauses can occur because of geminates, or at a syllable boundary of a long word. This is addressed by introducing short silences.

In addition to the above, it was felt that for some languages appropriate suprasegmental costs must be incorporated to reduce the mismatch at the boundary between two syllables. Acoustic cost measures based on intonation, energy, duration and spectra, pause prediction based on word terminal syllables and pauses were incorporated to address these issues.

V. EVALUATION OF TTS SYSTEMS

A variety of subjective testing procedures have been developed for assessing the quality of text to speech synthesis systems. Mean pinion score (MOS) is an important measure for evaluation of speech synthesis systems [32]-[33]. In particular [34] discusses different measures for evaluation of text to speech synthesis systems. The quality of synthesis speech depends on the quality of the original voice. The score obtained for synthesized speech is therefore normalized to that of natural speech. This is referred to as degradation Mean Opinion Score (DMOS). This involves randomly playing a set of semantically unpredictable sentences. Here, DMOS is used for evaluating the *naturalness* of the system. The primary reason for this is that sometimes the synthesised voice may have very good quality. Hence, the objective of DMOS is to determine the performance of our system relative to that of the natural voice. To get a score for the *intelligibility* of the system, word error rate (WER) is used [35]. The methodology for text-to-speech (TTS) listening tests has to be rigorous in voice sample presentation and subject selection. For DMOS evaluation and WER evaluation based on semantically unpredictable sentences (SUS), *15-20* sentences each were chosen and the evaluation was done with about *20* subjects. It is important to mention that these tests were performed by individual institutions. General guides for conducting DMOS and WER were given [36]. These scores are therefore informal. We are currently working on a standardisation of the procedures [1]

A. Degraded Mean Opinion Score

Listening tests involve preparing several samples of synthesized output from multiple TTS systems, randomizing the system sentence combinations and asking listeners to score each output audio. The DMOS obtained must be unbiased. To ensure this the same listeners were not asked to participate in different tests. For DMOS we play randomly the natural sentences (from the database) and synthesized sentences (text from the web, etc). An original file (sentence x) was played followed by a synthesised sentence (sentence y). All these sentences had different text. Headphones of reasonable quality was used for evaluation. Further, each listener was made to listen to a different set of sentences. The DMOS scores for *12* languages and the duration of system built are shown in Table V. Although no significant effort was made to build the CART in festival (Section IV), an average DMOS of $\approx$ *3* is obtained.

B. Word Error Rate Procedure

To get a score for the intelligibility of the system, word error rate is used [35]. This involves randomly playing a set of semantically unpredictable sentences. The details of the Word error rate for *12* languages are also shown in Table V. The WER for some of the languages, especially Kannada is rather high. It was observed that the size of the database for building the voice is rather small, this is primarily the reason in the main for the poor performance of Kannada.

$$WER = (\frac{insertions + deletions + substitutions}{total\ no.\ of\ words}) * 100 \quad (2)$$

VI. SUMMARY AND CONCLUSIONS

In this paper a unified framework for building text to speech synthesis systems in Indian languages is presented. In particular, syllable-based synthesis systems seem to be the key to building USS for Indian languages. The development of syllable-based TTSes across *13* Indian languages confirms the importance of syllable as a fundamental unit for synthesis. As syllable encompass coarticulation, the prosody modification required was significantly less. It was also observed during text parsing that a major subset of the rules were common across

[1]The DMOS and WER is not available for Bodo at the time of this writing as the number of subjects were insufficient to give a meaningful statistic.

all Indian languages. Peculiarities were treated as exceptions. Given the common label set it should possible to build synthesisers for new languages with very little effort.

ACKNOWLEDGEMENT

This work is carried out as a research project funded by Department of Electronics and Information Technology (DeitY), Govt. of India. In this project, different official languages of India, *viz*., Hindi, Gujarati, Tamil, Malyalam, Telugu, Assamese, Manipuri, Odia, Marathi, Bengali, Bodo and Rajasthani are chosen to build TTS systems such that these systems will be an aid for the visually challenged and those with disorder like cerebral palsy. The authors would therefore like to thank DeitY, New Delhi, India for their kind support to carry out the research work.

REFERENCES

[1] "World fact book," https://www.cia.gov/library/publications/the-world-factbook/geos/in.html#People. {Last accessed 18^{th} July 2013}.

[2] "Languages of India," http://http://en.wikipedia.org/wiki/Languages_of_India#cite_note-2. {Last accessed 18^{th} July 2013}.

[3] "South asian language families," http://http://en.wikipedia.org/wiki/File:South_Asian_Language_Families.jpg.

[4] S R Rajesh Kumar, "Significance of durational knowledge for a text-to-speech system in an indian language," M.S. thesis, Indian Institute of Technology, Madras, Department of Computer Science and Engineering, Chennai, India, 1990.

[5] X A Furtado and A Sen, "Synthesis of unlimited speech in indian languages using formant-based rule," *Sadhana*, vol. 21, no. 3, pp. 345–362, June 1996.

[6] D H Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, March 1980.

[7] N. Sridhar Krishna and Hema A. Murthy, "Duration Modeling of Indian Languages Hindi and Telugu," in *5th ISCA SSW*, 2004, pp. 197–202.

[8] A.W. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," `http://festvox.org/festival/`, 1998.

[9] DDNews, *Database for Indian languages*, India, Speech and Vision Lab, IIT Madras, Chennai, 2001.

[10] Samuel Thomas, "Natural sounding text-to-speech synthesis based on syllable-like units," M.S. thesis, IIT Madras, 2007.

[11] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan, and Hema A. Murthy, "Text-to-speech synthesis using syllable-like units," in *National Conference on Communication*, 2005, pp. 227–280.

[12] S. P. Kishore, Rohit Kumar, and Rajeev Sangal, "A data-driven synthesis approach for Indian languages using syllable as basic unit," in *International Conference on Natural Language Processing*, Mumbai, India, 2002.

[13] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Eurospeech*, 2003, pp. 1317–1320.

[14] M Sreekanth and A G. Ramakrishnan, "Festival based maiden TTS system for Tamil language," in *3rd Language and Technology Conference*, Poznan, Poland, October 2007, pp. 187–191.

[15] N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, and A.G. Ramakrishnan, "Duration Modeling for Hindi Text-to-Speech Synthesis System," in *ICSLP*, 2004, pp. 789–792.

[16] Keiichi Tokuda, Heiga Zen, and Alan W. Black, "An hmm-based speech synthesis system applied to english," in *Proc. IEEE Workshop on Speech Synthesis*, 2002.

[17] Ramani B, S Lilly Christina, G Anushiya Rachel, Sherlin Solomi V, Mahesh Kumar Nandwana, Anusha Prakash, Aswin Shanmugam S, Raghava Krishnan, S Kishore Prahalad, K Samudravijaya, P Vijayalakshmi, T Nagarajan, and Hema Murthy, "A common attribute based unified hts framework for speech synthesis in indian languages," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 311–316.

[18] Y. R. Venugopalakrishna, P. Srihari Krishnan, Samuel Thomas, Karthik Bommepally, Karthik Jayanthi, Suket Murarka, Hemant Raghavan, and Hema A Murthy, "Design and development of text-to-speech synthesis systems for indian languages," in *National Conference on Communication*, January 2008, pp. 259–262.

[19] Y R Venugopalakrishna, M V Vinodh, Hema A. Murthy, and C. S.Ramalingam, "Methods for Improving the Quality of Syllable based Speech Synthesis," in *Spoken Language Technology (SLT) 2008 workshop*, December Goa, 2008, pp. 29–32.

[20] E.Veera Raghavendra and Kishore Prahallad, "Database pruning for indian language unit selection synthesisers," in *ICON*, December 2009, pp. 67–74.

[21] Samuel Thomas, M. Nageshwar Rao, Hema A Murthy, and C S Ramalingam, "Natural sounding speech based on syllable-like units," in *EUSIPCO, Florence, Italy*, 2006.

[22] K Sreenivasa Rao, Krishnendu Ghosh, Ramu Reddy Vempada, Sudhamay Maity, and N P Narendra, "Development of text-to-speech synthesis system in bengali," *International Journal of Speech Technology*, vol. 14, pp. 167–182, 2011.

[23] Music TMH Speech and Hearing, "Wavesurfer," http://www.speech.kth.se/wavesurfer/, 2011.

[24] Paul Boersma and David Weenink, "Praat: Doing phonetics by computer," http:http://www.fon.hum.uva.nl/praat/, 2011.

[25] P. G. Deivapalan, Mukund Jha, Rakesh Guttikonda, and Hema A Murthy, "DONLabel: An Automatic Labeling Tool for Indian Languages," in *National Conference on Communication (NCC)*, IIT-Bombay, 2008, pp. 263–266.

[26] Prasad V. K., Nagarajan T., and Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communications*, vol. 42, pp. pp.429–446, 2004.

[27] S.R.M. Prasanna, B.V. Sandeep Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks and modulation spectrum energies," *IEEE Trans. Audio Speech and Language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[28] CMU, "The carnegie mellon university pronunciation dictionary," www.speech.cs.cmu.edu/cgi-bin/cmudict, 2008.

[29] N Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, and A G Ramakrishnan, "Duration modeling for hindi text to speech synthesis system," in *International conference on spoken language processing*, South Korea, 2004.

[30] K Sreenivasa Rao and B Yegnanarayana, "Intonation modelling for indian languages," *Computer Speech and Language*, vol. 23, pp. 240–256, April 2009.

[31] Ashwin Bellur, Badri Narayanan, Raghav Krishnan, and Hema A Murthy, "Prosody modeling for syllable-based concatenative speech synthesis for hindi and tamil," in *NCC*, 2011, pp. 216–220.

[32] R D Johnston, "Beyond intelligibility: the performance of text-to-speech synthesisers," *BT Technological Journal*, pp. 100–111, 1996.

[33] P L Salza, E Foti, L Nebbi, and Oreglia, "Mos and pair comparison combined methods for quality evaluation of text-to-speech systems," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 650–656, 1996.

[34] Mahesh Viswanathan and Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[35] Simon King, "Degradation MOS and word error rate for text to speech synthesis systems," Private Communication.

[36] IIT Madras, "Guidelines for mos evaluation for text-to-speech synthesis systems," http://www.iitm.ac.in/donlab/mosguidelines/, 2000.