# Acoustic Analysis of Syllables across Indian Languages

*Anusha Prakash*[1], *Jeena J Prakash*[2], *Hema A Murthy*[2]

[1]Dept of Applied Mechanics, Indian Institute of Technology Madras, India
[2]Dept of Computer Science & Engineering, Indian Institute of Technology Madras, India
am13s002@smail.iitm.ac.in, hema@cse.iitm.ac.in

## Abstract

Indian languages are broadly classified as Indo-Aryan or Dravidian. The basic set of phones is more or less the same, varying mostly in the phonotactics across languages. There has also been borrowing of sounds and words across languages over time due to intermixing of cultures. Since syllables are fundamental units of speech production and Indian languages are characterised by syllable-timed rhythm, acoustic analysis of syllables has been carried out.

In this paper, instances of common and most frequent syllables in continuous speech have been studied across six Indian languages, from both Indo-Aryan and Dravidian language groups. The distributions of acoustic features have been compared across these languages. This kind of analysis is useful for developing speech technologies in a multilingual scenario. Owing to similarities in the languages, text-to-speech (TTS) synthesisers have been developed by segmenting speech data at the phone level using hidden Markov models (HMM) from other languages as initial models. Degradation mean opinion scores and word error rates indicate that the quality of synthesised speech is comparable to that of TTSes developed by segmenting the data using language-specific HMMs.

**Index Terms**: acoustic analysis, Indian languages, syllables, text-to-speech synthesis

## 1. Introduction

Developing speech technologies is a challenging task, particularly when dealing with multiple languages. In India, which has about 1652 languages[1] along with dialectal variations, developing speech recognition, synthesis, translation systems, etc., is quite a herculean task. This is also compounded by the bilingualism (or multilingualism) feature that is quite common in India. Most Indian languages primarily belong to the Indo-Aryan or Dravidian language group. They share a common sound base, consisting of about 11-15 vowels and 33-35 consonants. It is the phonotactics which varies across the languages. A naive analysis is performed to study multilingual characteristics which might lead to enabling easier development of speech technologies.

Most of the literature available focuses on analysing acoustic properties of individual Indian languages [1–3]. Similarities among different languages from a language identification perspective are studied in reference [4]. From a multilingual perspective to better design speech technologies, reference [5] analyses subtle variations in phonetic features across multiple languages. In contrast to this, reference [6] exploits the similarities that exist among Indian languages to design a common label set for building multiple text-to-speech (TTS) synthesisers. Going a step further, [7] performs cross-borrowing of hidden Markov models (HMM) to aid in the easier training of TTSes. This technique is used to validate the studies carried out in the current work.

The work presented in this paper analyses syllables across six Indian languages in the context of continuous speech. The analysis is performed using both textual and speech data. Syllables are analysed as they are the fundamental units of speech production. Moreover, Indian languages are syllable-timed [8]. Syllables are defined as $C^*VC^*$ units, where $V$ represents a vowel and $C$ represents a consonant. The $C^*$ indicates that it can be zero or more consonants. Acoustic properties of most frequent syllables are analysed across six Indian languages - Bengali, Hindi, Marathi, Kannada, Tamil and Telugu. The first three languages belong to the Indo-Aryan language group and the rest to the Dravidian language group. To obtain syllable level segmentation of speech data, a hybrid segmentation algorithm is used [9]. In order to compare syllables across languages, syllables are denoted using standard notations specified in the common label set [6]. Analysis of word types and syllable types across the languages are carried out. Similarities among languages are determined based on the analyses. Such analyses are relevant in a multilingual scenario. Based on the conclusions drawn from the analysis, statistical parametric based text-to-speech (TTS) systems are developed by cross-borrowing hidden Markov models (HMM) for segmenting speech data. Subjective evaluation is conducted to assess the quality of synthesised speech.

The rest of the paper is organised as follows. Section 2 describes the datasets used in the work. Section 3 explains the hybrid segmentation algorithm. The analysis work is detailed in Section 4. Section 5 describes the experiments carried out in the TTS framework. The work is concluded in Section 6.

## 2. Datasets used

Six Indian languages are used in the experiments- three belonging to the Indo-Aryan language group and the others to the Dravidian group. Both female and male data are used. Details of the datasets used are given in Table 1. Additionally, 5 hours of male and female data of Rajasthani, an Indo-Aryan language, was used to substantiate the acoustic analysis performed.

[1]This figure is from the 1961 Census report of India

Table 1: Details of speech data

| Data | Bengali | | Hindi | | Marathi | | Kannada | | Tamil | | Telugu | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Duration (hours) | 4.78 | 4.98 | 5.16 | 5.18 | 3.56 | 4.8 | 3.4 | 3.95 | 4.56 | 5.37 | 4.24 | 5.0 |

## 3. Hybrid segmentation

The first step is to segment the data at the syllable level for analysis. Syllable segments are obtained using the hybrid segmentation algorithm [9]. The hybrid segmentation algorithm is a segmentation algorithm where machine learning is used in tandem with signal processing. Using flat start initialisation, where all the monophone HMMs are assigned with global mean and variance, HMMs are re-estimated. But these HMMs do not consider any boundary information. This boundary information is obtained from signal processing. Group-delay based segmentation method gives a set of syllable boundaries. The closest group-delay boundary in the vicinity of an HMM boundary is considered as the correct syllable boundary. Then embedded re-estimation is performed by restricting segmentation to the syllable boundary. This is done a couple of times iteratively. Further boundary correction is achieved using short term energy and spectral flux as cues. The resulting boundaries are the syllable boundaries.

In order to build a phone based TTS synthesiser, speech data needs to be segmented at the phone level. This is obtained by re-estimating HMMs after splicing the speech waveform at the syllable level. Re-estimating HMMs within the syllable boundary is an improvement over re-estimating over the entire utterance as reported in reference [10].

## 4. Analysis across languages

Reference [11] reports that the properties of a syllable are largely influenced by its position in a word. The acoustic properties of the same syllable vary when the syllable is in the beginning, middle or end of a word. Hence, the syllables are postfixed with BEG, MID and END tags, depending on their position in the word.

The text in each language is syllabified. The number of instances of a syllable occurring varies in the language. The probability of the count of syllables occurring in a language follows a Zipfian (Zipf) distribution. The Zipf distributions of the top 300 syllables for different languages are plotted in Figure 1. The Zipf distributions have a long tail. In the current work, therefore, the analysis is restricted to only the top 300 syllables, referred to as the "top syllables" in this paper.

Since the Indian languages considered in this work have different scripts, syllables in the native script are mapped to labels in the common label set [6]. The common label set is a set of sounds across 13 Indian languages where similar sounds are grouped together and denoted by a label. This makes it easier to compare syllables belonging to different languages.

We analyse the number of phones that constitute a syllable from the top syllables list. The statistics for the same are given in Figure 2. It is seen that syllables consisting of two phones occur most frequently across all languages. It is mainly of form $CV$,

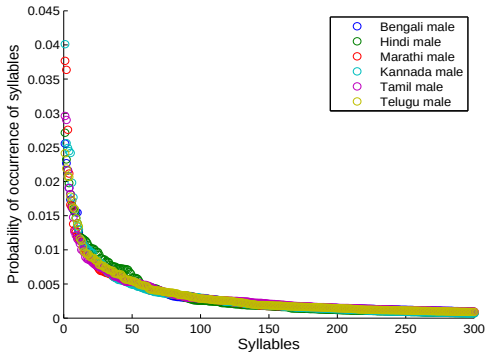


Figure 1: Zipfian distribution (Zipf) of frequency of top syllables for male data across languages

and in very few instances of the form $VC$. This is in keeping with the fact that the writing system of Indian languages is *akshara*-based, which has a $C^*V$ structure. The next frequently occurring syllable structure consists of three phones, mostly $CVC$ structure, and sometimes $CCV$. Syllables with a single phone are vowels ($V$ structure). Syllables composed of four or more phones occur rarely and are due to the presence of English words in the text in most cases.

We also study the probability of mono-syllabic, bi-syllabic, tri-syllabic words in the database. In the Indian languages considered, words are defined by blanks in the script similar to English. Unique words in the database are syllabified. The statistics of the same are given in Figure 3. For Indo-Aryan languages, the occurrence of words is in the following descending order: bi-syllablic, tri-syllabic, mono-syllabic, and further reduces as the number of syllables in the word increases. This pattern is in contrast to that of Dravidian languages. For Dravidian languages, words consisting of three or more syllables are common. Words containing six or more syllables are least probable in Indo-Aryan languages compared to Dravidian languages, where this probability is quite high. This highlights the *agglutinative* nature of Dravidian language scripts [12]. In agglutinative languages, multiple words are concatenated together to form a single word. The meaning of the word before and after concatenation does not change. For example, in Tamil, the words "vandu", "kondu", "irukkiraan" are concatenated together to form a single word "vandukondirukkiraan".

For each instance of a syllable, acoustic features, namely, duration, average energy and average $f0$ are calculated. A frame size of 25ms and a frame shift of 10ms are used for determining these values. The values of average energy and average $f0$ are normalised to zero mean and variance one for every syllable. The durations of each instance of a syllable are normalised with respect to the average syllable rate of the speaker since syllable rate varies with speakers. The average syllable rate is the number of syllables uttered in 1 second. The syllable rates for

Table 2: Syllable rates of different speakers

| Data | Bengali | | Hindi | | Marathi | | Kannada | | Tamil | | Telugu | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Syllable Rate | 4.91 | 5.14 | 5.01 | 4.94 | 4.15 | 4.22 | 5.73 | 4.94 | 5.66 | 5.18 | 6.52 | 5.2 |

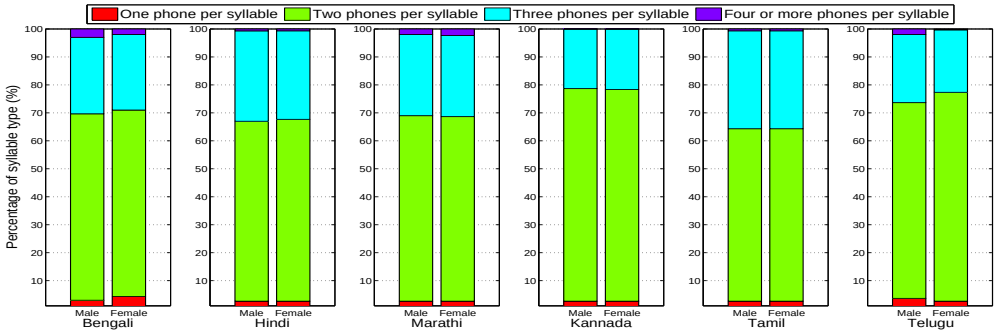


Figure 2: Percentage of types of syllables based on number of constituent phones for different datasets

speakers of different languages are given in Table 2.

The average durations of the top 300 syllables are plotted after normalising the syllable durations with respect to the syllable rate of the speaker for different languages (Figures 4 and 5). It is clearly seen from Figure 4, which is the distribution for male data, that syllables in the Dravidian languages have lower average syllable durations, when compared to those belonging to the Indo-Aryan languages. We hypothesise that the shorter average duration of syllables in Dravidian languages are due to its agglutinative nature. As the word becomes longer in terms of the number of syllables, speakers tend to shorten the word in terms of duration. This is not so in the case of distributions for the female data (Figure 5). Distributions of Bengali and Hindi data overlap with those of Dravidian languages. Additionally, 5 hours of Rajasthani male and female data are used and their distributions conform to the said pattern for both male and female data. We conclude that further studies have to be carried out to analyse the anomaly occurring for the female data in Hindi and Bengali. Similar analyses are performed for average energy and average $f0$ of syllables. They are not presented in the paper because of lack of conclusive results. We also have not done any study across genders, rather concentrating only from a language-specific perspective.

The studies highlight the similarities among Dravidian and among Indo-Aryan language groups. Using this grouping, TTS synthesisers are built with the aid of another language belonging to the same language group.

## 5. Hidden Markov model based speech synthesis (HTS)

To build a phone-based HMM speech synthesis system (HTS) [13] for a new language, speech data should be segmented at the phone level. Segmentation using the conventional flat start method, where HMMs are initialised with global mean and variance, does not result in good synthesis quality [10]. The hybrid segmentation technique has to be performed carefully and is time-consuming. As an alternative, data can be segmented easily with the help of an existing TTS system for a similar language.

HTS systems are developed for four languages- Bengali, Hindi, Tamil and Telugu. HTS version 2.3 is used for the same. Male data is used to build TTSes for these languages. One set of TTSes is built using segmented phone level speech data obtained from the hybrid segmentation algorithm. The next set of TTSes is developed using speech data segmented at the phone level with the help of a similar language in the same language group. Phone level segmentation is obtained by using context-independent monophone HMMs from another language as the initial HMMs and performing embedded training for about 5 iterations. Similar languages from which monophone HMMs are borrowed are termed as source languages. For Indo-Aryan languages, Hindi is considered as the source language and for Dravidian languages, Tamil is considered as the source language. Another experiment is conducted wherein the source language is from a different family group. For Hindi, Tamil is used as the source language, and vice-versa. In case a phone in the given language is not present in the source language, the monophone HMM corresponding to a similar phone is borrowed from the source language. This work differs from the work done in reference [7], where segmentation of speech data in the source language is performed using a bootstrap algorithm.

To evaluate the quality of the output synthesised sentences, degradation mean opinion score (DMOS) [14] and word error rate (WER) tests [15] are conducted. In DMOS test, listeners are asked to rate the quality of the synthesised speech on a scale of 1-5, 5 being the best. The scores are then calculated with respect to the natural or recorded speech. In WER test, listeners are asked to transcribe a set of semantically unpredictable sentences (SUS) or nonsensical sentences and then WER is calculated based on the number of insertions, deletions, and substitution of words.

An average of 10 native listeners evaluated a set of 15 and 10 synthesised sentences for the DMOS and WER tests for each language, respectively. Results of the evaluation are shown in Tables 3 and 4. The second column in both tables indicates the results obtained when languages are segmented using the hybrid segmentation algorithm. The last column indicates the results

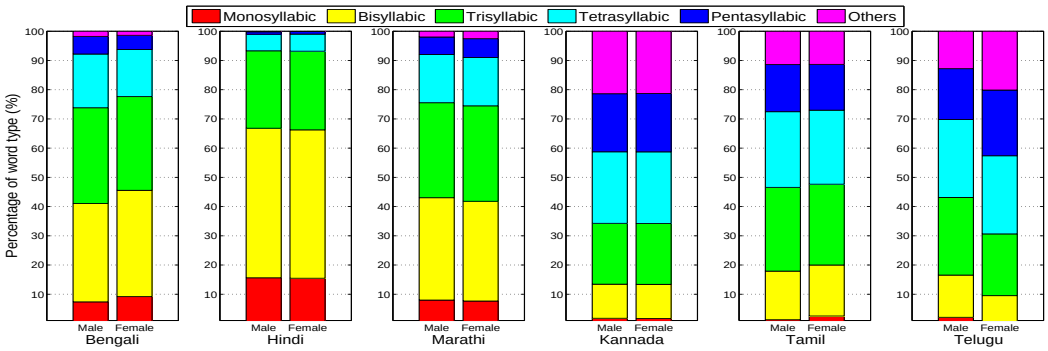


Figure 3: Percentage of types of words based on number of constituent syllables for different datasets

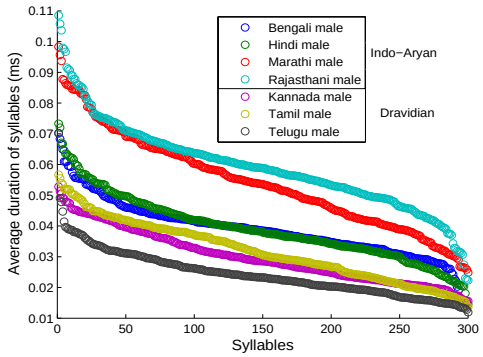


Figure 4: Distribution of average duration of top syllables for male data across languages

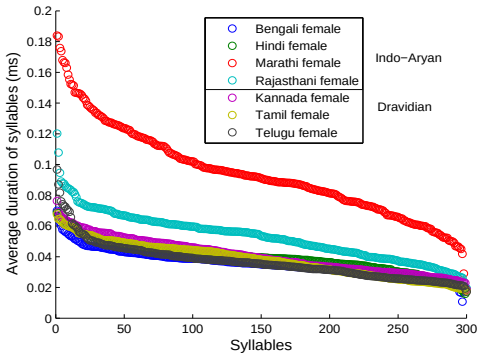


Figure 5: Distribution of average duration of top syllables for female data across languages

Table 3: Degradation mean opinion scores (DMOS)

| Language | Using hybrid segmentation | Source language | Using source language |
|---|---|---|---|
| Bengali | 2.99 | Hindi | 3.24 |
| Hindi | 3.45 | Tamil | 3.43 |
| Tamil | 4.21 | Hindi | 4.22 |
| Telugu | 3.28 | Tamil | 3.46 |

Table 4: Word error rates (WER) (%)

| Language | Using hybrid segmentation | Source language | Using source language |
|---|---|---|---|
| Bengali | 9.82 | Hindi | 3.67 |
| Hindi | 6.06 | Tamil | 5.4 |
| Tamil | 0.4 | Hindi | 3.9 |
| Telugu | 12.5 | Tamil | 9.54 |

when HMMs from a source language are used to segment data at the phone level. From both DMOS and WER scores, it is evident that using HMMs from a source language leads to better synthesis quality in terms of naturalness and intelligibility in most cases. This may be due to robust models in the source language as a result of accurate phoneme segmentation. Hence, phoneme segmentation is crucial in the context of developing good quality TTS systems. It is interesting to note that cross-borrowing of models across language groups also leads to good synthesis quality. This can be attributed to the fact that Indian languages share a common set of sounds. Over time, due to the borrowing of sounds (or words) across languages, the divergence between the language groups has narrowed down [16].

## 6. Conclusions

The work analyses acoustic properties of syllables across six Indian languages. Similarities among Dravidian languages and Indo-Aryan languages are determined. Based on this, monophone HMMs are borrowed from similar languages as initial models to segment speech data at the phone level in a given language and then building a TTS system for the language. This sort of analysis is crucial when developing speech technologies in a multilingual scenario and also for low-resource languages. This is also helpful in the study of prosody across Indian languages.

## 7. Acknowledgements

The datasets used in this work are part of the project "Development of Text-to-Speech Synthesis for Indian Languages Phase II" (Ref. no. 11(7)/2011-HCC(TDIL)). The authors would like to acknowledge the Department of Information Technology, Ministry of Communication and Technology, Government of India for the same.

# 8. References


[1] S. Sinha, S. S. Agrawal, and A. Jain, "Dialectal influences on acoustic duration of Hindi phonemes," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, India, November 2013, pp. 1–5.

[2] S. S. Vel, D. M. N. Mubarak, and S. Aji, "A study on vowel duration in Tamil: Instrumental approach," in *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, India, December 2015, pp. 1–4.

[3] B. Rajapurohit and Central Institute of Indian Languages, *Acoustic Studies in Indian Languages: Research Papers Prepared at the Summer Institute in Advanced Phonetics, 1984*. Central Institute of Indian Languages, 1986.

[4] D. Sengupta and G. Saha, "Study on similarity among Indian languages using language verification framework," *Adv. Artificial Intellegence*, vol. 2015, pp. 1–24, 2015.

[5] P. BHASKARARAO, "Salient phonetic features of Indian languages in speech technology," *Sadhana*, vol. 36, no. 5, pp. 587–599, 2011.

[6] B. Ramani, S. Lilly Christina, G. Anushiya Rachel, V. Sherlin Solomi, M. K. Nandwana, A. Prakash, S. Aswin Shanmugam, R. Krishnan, S. Kishore, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *SSW8*, Barcelona, Spain, 2013, pp. 291–296.

[7] A. Prakash, M. R. Reddy, T. Nagarajan, and H. A. Murthy, "An approach to building language-independent text-to-speech synthesis for Indian languages," in *Twentieth National Conference on Communications (NCC)*, Kanpur, India, February 2014, pp. 1–5.

[8] S. Rupak Vignesh, A. Shanmugam, and H. A. Murthy, "Significance of pseudo-syllables in building better acoustic models for Indian English TTS," in *IEEE International conference on Acoustics, Speech and Signal processing (ICASSP)*, Shanghai, China, 2016.

[9] S. A. Shanmugam and H. Murthy, "A hybrid approach to segmentation of speech using group delay processing and HMM based embedded reestimation," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, Singapore, September 2014, pp. 1648–1652.

[10] S. Aswin Shanmugam and H. A. Murthy, "Group delay based phone segmentation for HTS," in *National Conference on Communication (NCC)*, Kanpur, India, February 2014, pp. 1–6.

[11] Venugopalakrishna.Y.R., Vinodh.M.V., H. A. Murthy, and C. Ramalingam, "Methods for improving the quality of syllable based speech synthesis," in *Proc. of Spoken Language Technology (SLT) 2008 workshop*, Goa, India, December 2008, pp. 29 –32.

[12] Wikipedia, "Agglutinative language," https://en.wikipedia.org/wiki/Agglutinative_language.

[13] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 3, pp. 1039–1064, November 2009.

[14] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," in *Computer, Speech and Language*, vol. 19, 2005, pp. 55–83.

[15] C. Benoit, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-tospeech synthesis intelligibility using semantically unpredictable sentences," in *Speech Communication*, vol. 18, no. 4, 1996, pp. 381–392.

[16] M. B. Emeneau, "India as a lingustic area," *Language, Linguistic Society of America*, vol. 32, no. 1, pp. 3–16, 1956.