Sādhanā (2018) 43:53
https://doi.org/10.1007/s12046-018-0841-y

Spoken Indian language identification: a review of features and databases

BAKSHI AARTI[1,*] and SUNIL KUMAR KOPPARAPU[2]

[1] Department of Electronics and Communication, UMIT, SNDT University, Mumbai 400020, India
[2] TCS Innovation Labs - Mumbai, TATA Consultancy Services, Yantra Park, Thane 400606, India
e-mail: artigauri@yahoo.com; SunilKumar.Kopparapu@TCS.Com

MS received 14 October 2015; revised 10 April 2017; accepted 19 February 2018; published online 12 April 2018

Abstract. Spoken language is one of the distinctive characteristics of the human race. Spoken language processing is a branch of computer science that plays an important role in human–computer interaction (HCI), which has made remarkable advancement in the last two decades. This paper reviews and summarizes the acoustic, phonetic and prosody features that have been used for spoken language identification specifically for Indian languages. In addition, we also review the speech databases, which are already available for Indian languages and can be used for the purposes of spoken language identification.

Keywords. SLID; phonetic; characteristics; features.

1. Introduction

Spoken language is the most natural mode of communication in today's world, especially given the advances that have happened in the area of automatic speech recognition (ASR). However, to achieve good ASR performance in terms of recognition accuracies, two things are crucial, namely (a) the correct identification of the spoken language, which in turn depends on (b) the availability of a good speech corpus. From the Indian context, there has not been much work done in either of these two crucial areas, especially given that India boasts of many languages.

India has a rich language diversity with 22 languages recognized officially under the eighth schedule of the Indian Constitution. In addition to the officially recognized languages, there are several, running into hundreds, Indian languages spoken in India. A majority of them belong to either the Indo-Aryan or the Dravidian language families; a few languages belong to the Tibeto-Burman and Austroasiatic language families. The last decade or so has seen an increased interest in Indian language research and in particular the identification of language and development of Indian language speech corpus. Each language has several dialects and hundreds of accents, making the task of spoken language identification (SLID) even more complex. However different languages have their own set of grammar rules and other differentiating properties, which could help in identifying the language from spoken speech. While humans use a rich variety of information to distinguish one language from another, features like acoustics, phonetics and phonotactics associated with speech are widely used in automatic SLID task [1].

Speech features are a compact representation and/or parameterization of the raw acoustic speech signal. Acoustic features extracted from the speech signal have been widely used for SLID. Some of the widely used features in speech signal processing are (a) Mel Frequency Cepstral Coefficients (MFCC), (b) Perceptual Linear Predication (PLP) and (c) Linear Predication Cepstral Coefficients (LPCC). The basic aim of parameterization is to extract salient information from the speech data and ignore any unwanted information in speech. Once the basic acoustic features are extracted from speech signal, additional features, depending on the need, can be appended with the objective of modelling the temporal aspects of the speech signal. Some commonly utilized additional features are the delta and acceleration Cepstrum and the shifted delta Cepstrum (SDC) [1]. More recently, with the increasing access to large amounts of data, deep learning has been used in the area of speech research. One of the aspects of deep learning is to allow a deep neural network (DNN) to identify speech feature that can help in representing an acoustic signal rather than having to decide which speech features to use for a particular task. Phonetic features are features that are associated with the sounds of human speech and their production. It could also be associated with the extraction of a sequence of sound units from speech signal. Phonetics can be classified as follows [2]:

- Articulatory phonetics is a study of the physiological processes involved in producing sounds in human speech.

[*] For correspondence

- Acoustic phonetics is the study of acoustic characteristics of speech, including an analysis and description of speech in terms of its physical properties, such as frequency, intensity and duration.
- Auditory phonetics is the study of how listeners perceive speech sounds or briefly it is the phonetic aspects of perception.

Acoustic-phonetic features characterize all speech sounds and play a very important role in forming the linguistic structure. They are a useful cue for SLID (spoken language identification), especially because a specific language might have its own set of valid combination or sequence of sounds. While phonemes can be shared considerably across languages, and possibly could be combined in an unconstrained way, the valid sequential patterns may vary from one language to another [3].

In this paper we first review specific properties of Indian languages in section 2, which describes work done on different Indian languages and their phonological characteristics. Section 3 describes the development of speech corpus for a few Indian languages and we conclude in section 4. The main contribution of this paper is to review the state of the art in spoken Indian language identification and also bring together all the available speech corpus resources that are available for other speech scientists to work on.

2. Indian language identification

2.1 *Automatic Indian language identification*

All Indian languages are phonetic in nature. Phonetics is a branch of linguistics to study the sound structure of human language. Phonetics was studied as early as 500 BC in the Indian subcontinent, with Pāṇini's account of the place and manner of articulation of consonants (C) in his $5^{\{\mathrm{th}\}}$ century BC treatise on Sanskrit. The major Indo-Aryan alphabets today order their consonants according to Pāṇini's classification [4]. Speech sounds or phonemes of all languages are classified into vowels (V) and consonants (C). Most popularly they are represented with some specific symbols generally called as *aksharas*, which are an audible speech sound of that language.

These language-specific properties can be exploited to identify a spoken language reliably. Automatic language identification has emerged as a prominent research area in Indian languages processing. People from different regions of India speak around 800 different languages. Unfortunately, in our opinion, in comparison with other developed countries, work in the area of Indian language identification has not reached a critical level. Some salient phonetic features of Indian languages with focus on specific characteristics of Indian languages such as vowels, plosive, aspiration, affricates, nasals, quantity and quality of voicing and pitch variation have been discussed in [5].

2.2 *Segmental features of Indian languages*

Phonetic behaviour of Indian languages and categorization of phonotactics have been discussed in [6]. While studying phonetic nature of Indian languages, it is necessary to analyse the language features with respect to speaking style. Distinctive features of any language can be coded as being absent or present and they can be also categorized based on the manner they are uttered along with the place of utterance. The manner feature known as *sonorant* in speech sound measures the resonance in the vocal tract and is generally referred to as the *vowels*, *glides* and *nasals* [7]. Sounds produced with more constricted vocal tracts are referred to as *consonants*. Fricatives, stops and affricates are produced with an obstruction in the vocal tract and are referred to as *obstruents* [7].

Acoustic features distinguish structure of speech. The sounds that are produced with open vocal tract are called *vowels* and sounds produced with constriction in vocal chord are called *consonants*. Individual sounds of *vowels*, *glides* and *nasals* can be identified by their *formant* frequencies. It is measured in terms of amplitude peaks in the frequency spectrum of the sound, using a spectrogram. In general, /a/ has relatively high first formant frequency F_1 and it is relatively low for /i/, /u/. For bilabial stops (/p/, /b/) the F_2 and F_3 formants frequencies are relatively low, for alveolar stops (/t/, /d/) the F_2 and F_3 formants frequencies are high while for velar stops (/k/, /g/), the F_2 formant frequency is relatively high and the F_3 formant frequency is close to the F_2 formant frequency [7] as shown in table 1.

Plosives sounds in speech are determined by voice onset time (VOT) and burst release. VOT is the duration measure between release of stop consonant and onset vibration of the vocal chord. It is widely used to distinguish between voiced and voiceless stops.

Spectral features identify breathy voice quality imparted due to aspiration noise. Difference between first and second harmonics ($H_1 - H_2$) and spectral tilt - difference between second harmonic and amplitude of the strong harmonic in the region of third formant frequency ($H_1 - A_3$) are the features used to identify breathiness or murmured characteristics of phonemes [8].

Closed quotient measures closed portion of the glottal cycle using an electrocardiogram (ECG). Its value is low

Table 1. Phonemes and their formants.

Phone	F_1	F_2	F_3
/a/	High	–	–
/i/, /u/	Low	–	–
Bilabial stops (/p/, /b/)	–	Low	Low
Alveolar stops (/t/, /d/)	–	High	High
Velar stops (/k/, /g/)	–	High	$\approx F_2$

for breathy phonation while it is high for modal phonation [9].

Keeping generic structure of sound distribution as shown in table 2, modern Indo-Aryan and Dravidian language families have their specific acoustic, phonetic and prosody characteristics that discriminate them from others and we have tried to focus on the these characteristics of 9 major languages of India, namely Marathi, Gujarati, Hindi, Bengali, Assamese, Tamil, Telugu, Malayalam and Kannada. The language selection was based on (a) wide literature availability, (b) being widely spoken and (c) availability of speech corpus for experiments. Also, a conscious effort was made so that some languages are easy to distinguish from one another (example Hindi and Telugu) and some very hard to distinguish from one another (example Telugu and Kannada). Table 3 shows standardized international phonetic notation for sounds of spoken language.

Indo-Aryan language is a sub-group of Indo-Iranian branch of the Indo-European language family. There are three major divisions of Indo-Aryan language family such as Old (Old and Classical Sanskrit), Middle (Prakrit and Pali) and Modern (Hindustani). Tentative phonological structure of Indo-Aryan family has vowels and diphthongs system, 5 positions of articulations for stops and 3 nasal sounds. In modern Indo-Aryan languages, utterance of mid central vowel /ə/ is replaced by open front unrounded vowel /a/. Utterance of diphthongs /au/ and /ai/ is retained at the same area while some vowel sequences are constricted such as /aũ/ to /ũ/ and /iu/ to /Ī/ [10]. In modern Indo-Aryan language family, phone system has been reduced such that there is no retroflex sound in Assamese, while Marathi and Bengali languages have a tendency to avoid aspiration sounds, and specifically in Marathi there are changes of sound /s/ to /š/ before palatal sounds [10].

Dravidian language family has four major divisions such as South, South-Central, Central and North groups. Tentative phonology structure of Dravidian family has 5 short and 5 long vowels system, 6 positions of articulation for stops and 5 nasal sounds. but there is gradual elimination of contrast sound between short /e/ and /o/ and long /e/ and / o/ [11]. No retroflex or alveolar series of consonant occurs at the beginning of words, and at the end of utterance of stop sounds they were followed by the vowel /u/ [11]. In modern Kannada, phonology has extensive use of secondary phoneme /h/ [11].

Marathi belongs to an Indo-Aryan language family. It is the official language of Maharashtra State. Its grammar and syntax is derived from *Pali* and *Prakrit*. In Marathi language, there are 55 phonemic letters or characters, which are divided as (a) *swara* (vowels 12 letters), (b) *vyanjana* (consonant 45 letters) and (c) *swaradi* (diphthongs 2 letters). The consonants have vowel at the end and when vowels are mixed with consonants it forms *akashara*. According to this classification (see figure 1) vowels are sub-classified as long, short and mixed vowels, and consonants are sub-classified as nasal consonant (3), semivowels, fricatives (3), voiced fricatives and unvoiced fricatives, affricates (7) whisper and independent consonant [12]. It has six vowels, including short and long vowels, but they are not contrastive. There are two additional vowels /6/ and /æ/ in Marathi [13]. Retaining certain characteristics of Indo-Aryan family, Marathi language has its unique phonological characteristics.

- Aspiration is one of the phonetic property of several Indian languages. In Marathi language, word initial stop differs in only the aspiration characteristic (e.g., /b/, /b^{h}/, /t̪/, /t̪^{h}/, /p/, /p^{h}/) but the resulting two words have different meanings. Acoustic characteristics for aspiration detection of voice and voiceless stops in Marathi language were explored in [8].
- Plosive (manner of articulation) is an important characteristic of Indian speech sound. Like other Indo-Aryan languages, Marathi shows 5-way contrast in plosives involving voicing, aspiration and murmured stops with five places of articulation such as velar, palatal, retroflex, dental and labial (e.g., /g/, /g^{h}/, /ʥ/, /ʥ^{h}/, / d/, / d^{h}/, /d̪/, /d̪^{h}/, /b/, /b^{h}/) [14].

Table 2. Distribution of sounds in different Indian languages.

Language	Vow	Dipt	Liq	Glid	Cons	Nasl	Stp	Fric	Affric
Telugu	11	2	3	2	36	3	15	5	4
Kannada	13	2	3	2	34	5	16	5	4
Tamil	10	2	3	2	18	5	4	4	2
Malayalam	10	5	4	2	37	6	16	4	3
Hindi	10	2	4	2	37	5	20	4	4
Marathi	12	2	3	2	45	3	16	3	7
Gujarati	13	2	3	2	35	3	20	3	4
Bengali	14	15	3	3	34	4	20	2	4
Assamese	11	10	2	2	23	3	12	4	–
Oriya	11	9	3	2	31	4	20	2	4
Punjabi	10	6	4	2	28	5	15	2	3
Kashmiri	15	–	2	2	27	2	12	4	5
Sindhi	10	2	4	2	46	4	20	2	4

Table 3. Notation of phonemes.

Notation	Phoneme	Phonetic transcription
/i:/	Feet	/fi:t/
/i/	Fit	/fit/
/e/	Let	/let/
/ae/	Man	/mæn/
/a:/	Car	/ka:/
/ɔ/	Pot	/pɔ t/
/ɔ:/	Caught	/kɔ:t/
/u/	Put	/put/
/u:/	Boot	/bu:t/
/ʌ/	Cup	/kʌp/
/ə:/	Pearl	/pə:l/
/ə/	Attend	/əˋ tend/
/ai/	File	/fail/
/ei/	Sail	/seil/
/ɔi/	Boy	/bɔi/
/ɛə/	Hair	/hɛə/
/uə/	Poor	/puə/
/iə/	Ear	/iə/
/au/	Now	/nau/
/əu/	Know	/nəu/
/p/	Peak	/pi:k/
/b/	Bought	/bɔt/
/t/	Ton	/tʌn/
/d/	Die	/dai/
/k/	Cave	/keiv/
/g/	Gap	/gæp/
/tʃ/	Chin	/tʃin/
/ʤ/	Jeer	/ʤi ə/
/m/	Limp	/limp/
/n/	Send	/send/
/ŋ/	Sing	/siŋ/
/f/	Fast	/fa:st/
/v/	View	/vju:/
/θ/	Thin	/θin/
/ð/	Then	/ð en/
/s/	Sink	/si ŋ/k
/z/	Zoo	/zu:/
/ʃ/	Shoe	/ʃu:/
/ʒ/	Beige	/beiʒ/
/h/	Hit	/hit/
/l/	Tell	/tel/
/r/	Cry	/krai/
/j/	Tune	/tju:n/
/w/	Quick	/kwik/

- In Marathi, performance of detecting aspirated stops /k^{h}/, /t^{h}/, /t̪^{h}/ using VOT, spectral features and synchronization index is higher in case of voiceless stops as compared with that in case of voiced stops /b^{h}/, /d̪^{h}/, /d^{h}/,/g^{h}/ [8].
- This language has phonetic property of modal nasal sounds (e.g., /m/, /n̪/, /n/) as well as contrast breathy nasal sounds (e.g., /m^{h}/, /n̪^{h}/, /n^{h}/) [13].
- The breathy sound /N̤/ has mostly nasal flow with small amount of oral flow at the end of segment and it has lower closed quotient than modal voice /N/ [9].

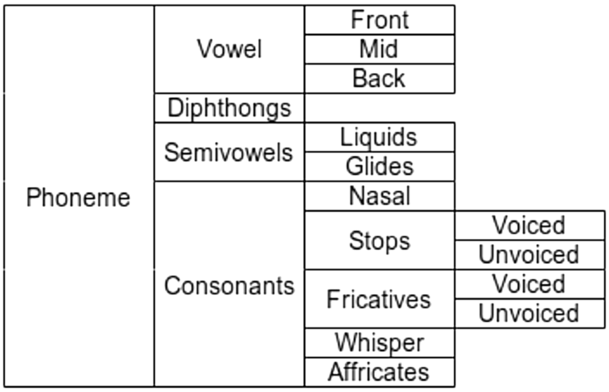

Figure 1. Organization of sounds.

- In Marathi language, affricates are split into alveolar (e.g., /c/, /j/, /z/) and palatal (e.g., /č/, /ǰ/, /ž/) phonemes, which are contrasting in their sound system [15].
- It also shows contrast in approximants involving plain (e.g., /ʋ/, /l/) and murmured (e.g., /ʋ^{h}/, /l^{h}/) sounds while it also shows contrast in tap/flap (e.g., /ɾ/, /ɾ^{h}/) [13].
- It does not use voiceless aspirated /ʦ^{h}/ sound.
- Marathi language uses retroflex nasal sound /ɳ/ most frequently.
- A retroflex lateral approximant /ɭ/ is a prominent feature of Marathi [16].
- Vowel /i/ has low formant frequency F_1 while vowel /a/ has high formant frequency F_1.

Gujarati is a member of an Indo-Aryan language family. It is derived from *Sanskrit* through *Prakrit* and *Apabhramsha*. It is the official language of Gujarat state. Like other Indo-Aryan languages, it has 60 phonemic letters/characters, which are divided as (a) *svara* (vowels 13 letters), (b) *vyanjana* (consonants 35 letters) and 2 diphthongs. Vowels are classified as *hrasva* (short vowel) or *dirgha* (long vowel). All vowels, except /e/ and /o/, occur nasalized, and in murmured and non-murmured forms. Being part of Indo-Aryan family, Gujarati has short and long vowels but they are not contrastive. Vowels are long when nasalized or in a final syllable [17]. Consonants in Gujarati are sub-classified as nasal(3), fricatives (3), affricates (4) and plosives (20). Retaining generic characteristics of Indo-Aryan family, Gujarati has its specific phonological characteristics.

- Gujarati shows 4-way contrast in plosives, which are similar to other Indo-Aryan languages involving voiceless unaspirated, voiceless aspirated, modally voiced unaspirated and breathy-voiced aspirated with five places of articulation, which are velar, palatal, retroflex, dental and labial (e.g., /g/, /g^{h}/, /ʣ/, /ʣ^{h}/, /d/, /d^{h}/, /d̪/, /d̪^{h}/, /b/, /b^{h}/) [18].

- In Gujarati, two voiced velar stops /gʰ/ /g/ have shorter VOT than voiceless velar stops /kʰ/, /k/ while voiced velar stops have more burst frequencies than voiceless velar stops. Aspirated /kʰ/ has longer VOT than unaspirated /k/ [19].
- It has unique phonological characteristic of differentiating modal vowels (e.g., /i/, /e/, /ɛ/, /a/, /ə/, /ɔ/, /o/, /u/) and breathy vowels (e.g., /a̤/, /ɛ̤/, /ə̤/, /ɔ̤/) while the difference in the production of breathy voice and modal voice is dependent on the vocal tract excitation [18].
- In Gujarati, breathy voice is less loud, its average fundamental frequency less and strength of excitation is more while spectral title is higher as compared with modal voice [20].
- Most of the Gujarati speakers use alveolar fricative /s/ sound instead of retroflex /ʂ/ and post-alveolar /ʃ/ fricative sounds [21].
- Retroflex nasal sound /ɳ/ is most frequently used in Gujarati language.
- Gujarati speakers mostly pronounced alveolar plosives /ɗ/ and /t/ as retroflex plosives /ɖ/and /ʈ/ [21].
- A retroflex lateral approximant /ɭ/ is a prominent feature of Gujarati [16].

Hindi is one of the official languages of India. It is the fourth most spoken language in the world. It has the status of official and co-official language for Bihar, Chhattisgarh, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand. It also belongs to an Indo-Aryan language family and much of its vocabulary is derived from *Sanskrit*. It has 58 phonemic letters/characters, which are divided as (a) *svara* (vowels, 10 letters) and (b) *vyanjana* (consonants, 37 letters); in addition 5 *nukta* consonants are adapted from *Farsi/Arabic* language sounds [22]. Hindi alphabets are classified as short, long and semivowels, fricatives (4), stop sounds (20) and nasal sounds (5) and it has (8) aspirated plosives and (2) aspirated fricatives [23]. It has basic 10 vowels (3 lax and 7 tense vowels) and all of them have nasalized form where oral and nasal vowels are contrastive. Like other Indo-Aryan languages, in Hindi, length distinction of long and short vowels has been neutralized. Retaining generic characteristic of Indo-Aryan family, Hindi has its unique phonological features as follows:

- In Hindi vowels (/ə/,/I/,/U/) are always considered as short in length while vowels (/a/,/i/,/e/,/o/,/u/, /ɛ/,/ɔ/) are considered long in length [24].
- Vowels /I/, /U/ have narrow, while /ə/ has narrow, openness of vocal tract. Vowels /i/, /u/ have narrow and /e/,/o/ have medium while /a/ has wide openness of vocal tract [25].
- In case of vowels of Hindi language, pitch is high (about 180 Hz) for high vowels and low (about 160 Hz) for low vowels [25].
- It shows 4-way contrast in plosives involving voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated with five places of articulation such as velar, palatal, retroflex, dental and labial (e.g., /g/, /gʰ/, /ʤ/, /ʤʰ/, /d/, /dʰ/, /d̪/, /d̪ʰ/, /b/, /bʰ/). Hindi has an additional voiceless unaspirated uvular plosive /q/ [24].
- Voiced stops have shorter closure durations than voiceless stop and breathy voiced stops have the shortest closure duration [26].
- In Hindi, vowels after plain have less vowel duration than those following aspirated and breathy voiced stops [26].
- Fundamental frequency F_0 is lower after voiced than after voiceless stops, and even lower after breathy voiced stops than after their plain voiced counterpart [26].
- In Hindi, voiced (breathy) oral stop /Nɦ/ has mostly oral flow in addition to nasal flow and it acts as two distinct segments, one nasal /N/ and /ɦ/. It has higher closed quotient, indicating less breathiness than modal /N/ [9].
- It also has a phonemic difference between the dental plosives and the retroflex plosives. The dental plosives are laminal-denti alveolar and the retroflex series is not purely retroflex; it actually has an apico-postalveolar articulation.
- In Hindi, the durations of unvoiced aspirated stop sounds are twice the durations of unvoiced unaspirated stop sounds [27].
- Its dialect does not show any difference in pronunciations of sound /ʂ/ and /s/.
- This language has a prominent feature of voiced palatal fricative /ʝ/ sound.
- It also has three flaps: a simple retroflex flap /ɽ/, a murmured retroflex flap /ɽ̤/ and a retroflex nasal flap ɽ̃ [24].
- Retroflex nasal sound /ɳ/ is also present in Hindi language but mostly not utilized in Hindi dialect [15].
- A retroflex lateral approximant /ɭ/ is absent in Hindi pronunciation and uses lateral approximant /l/ [15].
- VOT feature is used to distinguish dental stop from retroflex stop. The retroflex has very short VOT lag as compared with dental [28].

Bengali, also called *Bangala* or *Bangla-Bhasa*, belongs to an Indo-Aryan language family. It is the official language of the state of West Bengal. It is derived from *Magadhi Prakrit* and *Pali*. It is also bound stress language and there are 48 phonemic letters/characters, which are divided as (a) *sboroborno* (vowels 14 letters) and (b) *benjonborno* (consonants 34 letters). The vowels are further classified as short and long and (7) nasal vowels. It also has mid-central vowel sound (*swa*) [6]. All oral vowels have nasal counterpart and nasalization is phonemic. They are coalesced and their distinction is no longer phonemic like other Indo-

Aryan languages [29]. It has around (15) diphthongs. The consonants include (20) stops, (2) fricatives, (4) nasals and (3) liquids. Retaining generic characteristics of Indo-Aryan family, Bengali language has its unique phonological characteristics as follows.

- Vowel length is not phonemic in the Bengali language, so their occurrence may be short or long as per the duration of vowel. /u/ has the shortest duration and /i/ and /æ/ have the longest duration [29].
- Voiced stops have shorter closure durations than voiceless stops and breathy voiced stops have the shortest closure duration [26].
- In Bengali, voiced (breathy) oral stop /Nɦ/ has mostly nasal flow and has lower closed quotient only during the /ɦ/ portion while the modal portion of the/Nɦ/has a closed quotient value similar to that of the modal /N/ [9].
- Stop consonants can be divided into velar, palato-alveolar, alveolar, dental and labial. Every series of stops include voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated (e.g., /g/, /gʰ/, /ʣ/, /ʣʰ/, /d̠/, /d̠ʰ/, /d̪/, /d̪ʰ/, /b/, /bʰ/) [29, 30].
- Aspiration is an important phonetic feature of Bengali Language and it is found in voice and voiceless plosives. Bengali consonants are classified on the basis of manner and place of articulation as well as aspiration and voicing. In Bengali, only plosives and stops are aspirated (e.g., /t/, /tʰ/, /t̠/, /t̠ʰ/, /ʧ/, /ʧʰ/) while fricatives and nasals are not [30].
- Bengali language allows aspirated stops both in *onset* (syllable initial) and *coda* (syllable final) position and two consecutive stops can never occur [31].
- It does not have palatal nasal /ɲ/, dental nasal /n̪/ and retroflex nasal /ɳ/ sounds and it extensively uses velar nasal /ŋ/ (*English* ’*ng*’ *sing*) sound.
- It also has a whole series of three sibilants (/s/, /ʂ/, /ʃ/) but Bengali speakers pronounce all of them as /ʃ/ [28].
- In Bengali dialect, there is a tendency to pronounce /pʰ/ as a voiceless bilabial fricative /ɸ/ or /f/, also /bʰ/ or /kʰ/ equivalent as voiced bilabial fricative /ʙ/ or /x/ [32].
- A retroflex lateral approximant /ɭ/ is absent in Bengali pronunciation.

Assamese, also called *Axamia*, belongs to an Eastern Indo-Aryan language family. It is the official language of the state of Assam. It is derived from the *Magadhi Prakrit* and dialects are developed from *Vedic* dialects. It has got many phonological characteristics retaining original attributes from Indo-European family, which makes Assamese language speech unique [33]. It has 44 phonemic letters/ characters, which are divided as (a) *swara* (vowels 11 letters) and (b) consonants 33 letters.

Each consonant letter represents a single sound with an inherent vowel sound /a/. The first (25) consonants letters are called *sparsha barna*. The consonants are broadly categorized as *stops* and the *continuants*. A stop consonant may be voiced or voiceless and aspirated or unaspirated, while the continuants are categorized as (2) frictionless, (4) aspirants, (1) lateral and (3) nasals [34]. It has basic (8) vowels and around (10) diphthongs. Its long and short vowels are coalesced and length distinction has been neutralized. Retaining generic characteristics of Indo-Aryan language family, Assamese has its specific characteristic as follows:

- Assamese vowel has high back rounded vowel /ʊ/, which is unique among Eastern Indo-Aryan languages; it is slightly lower and more centralized than /u/.
- Assamese language inventory has lack of dental-retroflex distinction among the stops. These consonants can be divided into velar, alveolar and labial. Every series of stops includes voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated (e.g., /g/, /gʰ/, /d/, /dʰ/, /b/, /bʰ/) [35].
- It has the unique voiceless velar fricative /x/ sound, which is not present in any other Indian language. It is pronounced somewhat between the sounds /s/, /kʰ/ and /h/ [35].
- Assamese lacks affricates; for e.g. voiceless palatal /ʧ/,/ʧʰ/ are merged into alveolar /s/ while the voiced palatal affricates /ʤ/, /ʤʰ/ are merged into /z/ [36].
- It does not have palatal nasal /ɲ/, dental nasal /n̪/ and retroflex nasal /ɳ/ sounds and it extensively uses velar nasal /ŋ/ (*English* ’*ng*’ *sing*) sound [35].
- All nasal consonants except /ŋ/ occur initially, medially and finally while /ŋ/ does not occur initially. Using a spectrogram plot it is observed that all nasal consonants have a very low value energy ratio as presence of energy at frequency above 3 kHz. Nasal consonant /ŋ/ has more energy at 4 kHz as compared with other nasal consonants [37].
- A retroflex lateral approximant /ɭ/ is absent in Assamese pronunciation.
- Assamese speakers mostly pronounce alveolar plosives /ɗ/ and /t/ as retroflex plosives /ɖ/and /ʈ/ [21].

Tamil is one of the oldest and the longest surviving classical language from the Dravidian language family. It is the official language of the state of Tamil Nadu. It has 30 phonemic letters/characters, which are divided as (a) *uyireluttu* (uyir – life, eluttu – letter, vowels 10 letters), (b) *asmeyyeluttu* (mey – body, eluttu – letters, consonants 18 letters) and (c) (2) diphthong (/ai/ and /au/). The vowels are classified as *kuril* (short), *nedil* (long) and *kuuriyl* (shortened) vowels [38]. The consonants are classified as *vallinam* (hard), *mellinam* (soft or nasal) and *itayinam* (medium). It also has a special character (also referred to as secondary character) called *aytham* in classical Tamil but now it is rare in modern Tamil [38]. These consonants and vowels are combined to form 216 compound characters. It has (10) vowels, which are classified as long and short vowels. It has (2) diphthongs. Consonants are classified as nasal (5), stops (4), affricates (2) and lack of fricatives. Retaining generic features of Dravidian family, Tamil has its unique features that discriminate it from other languages.

- In Tamil language the long vowels are twice as long as short vowels, and diphthongs are pronounced about 1.5 times as long as a short vowel [38].
- It has an epenthetic vowel /ɨ/ sound.
- This language has neither aspirated nor voiced stop like other Indian languages and also it does not have aspirated consonant. Its stop consonants can be divided into velar, retroflex, alveolar, dental and labial (e.g., /k/, /ʈ/, /t/, /t̪/, /p/) [39–41].
- Voiced and unvoiced stops both are present in spoken Tamil language as allophones (e.g., /p/ and /b/, /t̪/ and /d̪/, /t/ and /d/, /ʈ/ and /ɖ/, /k/ and /g/) [39].
- This language has a voiced libodental approximant /ʋ/ and a voiced bilabial approximant /β/ sounds.
- It has six distinction in nasal sounds such as dental n̪/n̺/, alveolar /n/, retroflex /ɳ/, alveolo-palatal /ɳ̥/ and velar /ŋ/ [13].
- It has three /R/ sounds or otherwise fricatives including flapped /ɾ/, alveolar trill /r/ and a unique retroflex approximant /zha/ sound [42].
- Tamil phonology has some consonant clusters that are never word initiated. In Tamil language, there is no stress or accent at word level and all syllables are pronounced with the same emphasis [41].

Telugu belongs to the south central group of Dravidian language family. It is the official language of the state of Andhra Pradesh and Telangana. Telugu is influenced by *Prakrit* and *Sanskrit*. It has 51 phonemic letters/characters, which are divided as (a) *achchu or* swar (vowels 10 letters) and (b) *hallu or* vyanjan (consonants 35 letters). Telugu is syllabic in nature and syllables are formed with vowels and consonants. Telugu words are generally ended with vowels and its grammatical rules are derived from Pāṇini's concepts [43]. To form a sentence *hallulu* (*C*) combines with a *acchulu* (*V*) sound [44]. Telugu has a full *C V* unit with pitch higher than those in other languages [6]. Telugu letters are categorized as vowels, vowel signs, consonants, semivowels, sibilants and aspirates. Telugu has (11) vowels, which are read in two groups as short and long vowel and (2) diphthongs. In addition to basic 10 vowels it has a long vowel /æ/. Consonants are classified as nasal (3), plosives (15), fricatives (5) and affricates (7). Retaining Dravidian family features, Telugu has its unique features that discriminate it from other languages.

- The duration of long vowel is almost twice that of the short vowel and hence the ratio of the short to long vowel duration is 1 : 2.1 [45, 46].
- In Telugu, the longest among short vowels is /o/ and the shortest is / u/. The longest among long vowels is / a:/ and the shortest is /e:/ [45, 46].
- In Telugu, long vowel (/i :/, /u :/) phonemes may have an extra allophone (/i : :/, /u : :/) when they grammatically occur in the position of a conjoiner [5].
- Among short vowels in Telugu (/i/, /e/, /o/, /a/, /u/), front mid vowel /e/ has the longest mean vowel duration while short front high vowel /i/ has the shortest mean vowel duration. Among long vowels in Telugu (/i :/, /e :/, /o :/, /a :/, /u :/), low mid vowel /a :/ has the longest mean vowel duration while long vowel /i :/ has the shortest vowel [47].
- It exhibits the vowel harmony phenomenon, which is not characteristic of any other Dravidian language. In vowel harmony, quality of a vowel in a syllable is decided by vowels of the preceding [5].
- Stop consonants can be divided into velar, retroflex, dental and labial. Every series of stops include voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated (e.g., /k/, /kʰ/, /ʈ/, /ʈʰ/, /d̪/, /d̪ʰ/ /p/, /pʰ/) [43].
- In this language, dental voiceless plosive /tʰ/ is rarely found and is replaced by dental voice plosive /dʰ/ [43].
- In this language, affricates have split into alveolar and palatal (e.g., /c/, /cʰ/, /j/, /jʰ/). These /c/ and /j/ are pronounced as palatal affricates (/ʧ/, /ʤ/) before front vowels and as alveolar affricates (/ʦ/, /ʣ/) before back vowels [48].

Malayalam belongs to the southern group of Dravidian language family. According to the researchers, Malayalam is a branch of *classical Tamil* but has a large number of *Sanskrit* vocabulary [49]. It is the official language of the state of Kerala. It has 53 phonemic letters/characters, which are divided as (a) *svaram* (vowels 10 letters) and (b) *vyanjanam* (consonants 37 letters). It is a syllable-based language in which all consonants have in-built vowel /a/. It has uniform literature dialect throughout the state of Kerala. The vowels are sub-classified as long and short vowels, which occur at all positions in a word, except /o/, which will not occur at the end of a word [50]. It has (11) monothongs and (5) diphthongs [51]. These diphthongs are completely under separate categories; that is, they are phonologically distinct from monophthongs. Consonants are classified as nasal (6), plosives (16), fricatives (4) and affricates (3). Nasal, laterals and voiceless unaspirated stops sounds can be geminated while distinction between single and geminated consonants is phonemic [50]. Retaining generic features of Dravidian language family, Malayalam has unique features that discriminate it from other languages.

- Malayalam has an epenthetic vowel /ɨ/ and /ə/ vowel sound [49].

- In this language, all vowels except /ɨ/ and /ə/ can be short as well as long (e.g., /a/, /aː/, /i/, /iː/). These vowels have significant durational difference resulting use of these vowels in the word may have different meaning [52].
- The consonants of Malayalam have 9 places and 8 manners of articulation. In this regard, alveolars and plosives are the most complicated and plosives are further classified as velar, palatal, retroflex, palato-alveolar, alveolar, dental and bilabial (e.g., /k/, /k^{h}/, /ḱ/, /ḱ^{h}/, /ʈ/, /ʈ^{h}/, /ʧ/, /ʧ^{h}/, /t/, /t̪/, /t̪^{h}/, /p/, /p^{h}/) [51].
- Unaspirated voiceless velar, retroflex, alveolo-palatal, dental and bilabial plosives (e.g., /k/, /k :/, /ʈ/, /ʈː/ /ʧ/, /ʧː/, /t̪/, /t̪ː/, /p/, /p :/) are either single tone (i.e., short) or geminate (i.e., long) in Malayalam [51].
- Malayalam labial, dental and alveolar-palatal plosives are classified by distribution of high-frequency energy compared with low-frequency energy between consonant onset followed by onset vowel using LPC analysis [53].
- To identify voiced percept for bilabial plosives, Malayalam speakers require shorter lead VOTs and longer lead VOTs to identify a voiced percept of velar plosives[54].
- The author has analysed acoustic characteristics of stop consonants during normal and large rate of speech using acoustic parameters, including mean pitch, jitter, shimmer, SNR and HNR. He found no significant difference in stops at the two rates [55].
- Malayalam language shows six contrasting places of articulation for nasal sounds, which are also either single tone (i.e., short) or geminate (i.e., long) such as bilabial /m/ and /mː/, dental /n̪/ and /n̪ː/, alveolar /n/ and /nː/, retroflex /ɳ/ and /ɳː/, palatal /ɲ/ and /ɲː/, except velar /ŋ/ nasal sound [13, 51].
- It has two contrast trills, an advanced trill /ṛ/ and a retracted trill /r̠/.
- A post-alveolar approximant /ʐ/ is a prominent feature of Malayalam language.
- Malayalam has either single tone (i.e., short) or geminate (i.e., long) retroflex lateral /ɭ/ and /ɭː/ and alveolar lateral /l/ and /lː/ approximants [51].
- In Malayalam, nouns with geminate and those with non-geminate sonorants differ systematically in terms of tense vs lax articulation and phonation. The geminate consonants are longer than non-geminates and their nouns differ in their rhythmic relationship between first and second syllables, while non-geminate consonants have short–long relationship between their syllables [56].

Kannada, also called as *Kanarese*, or *Canarese*, belongs to the southern group of the Dravidian language family. Kannada is influenced by Sanskrit, *Prakrit* and *Pali* languages [57]. It is the official language of the state of Karnataka. It has 49 phonemic sounds and the corresponding characters; different characters can be combined to form a compound character [58]. They are divided into three groups, namely, (a) *swaragalu* (vowels – 13 letters), (b) *vyanjanas* (consonants – 34 letters) and (c) *yogavaahakas* (part vowel, part consonants 2 letters: *anusvara* and *visarga*). It has (2) diphthongs, (5) short vowels, (5) long vowels and (2) vowel glides [58]. All Kannada words end with /a/ vowel. Consonants are classified as nasal (5), plosives (16), fricatives (5) and affricates (4). All consonant phonemes occur in all position of the word. Retaining generic features of Dravidian language family, Kannada has unique features that discriminate it from other languages.

- Kannada has short as well as long vowels (e.g., /i/ and /ī/, /e/ and /ē/, /o/ and /ō/), these vowels lengths make a difference in the meaning of words [57].
- Voiceless plosives /p/, /t/, /k/ have long positive VOT while voiced plosives /b/, /d/, /g/ have negative VOT. VOT also helps for gender differentiation when speaking rate is controlled [59].
- To identify voiced percept for bilabial plosives, Kannada speakers require longer lead VOTs and shorter VOTs to identify a voiced percept of velar plosives [54].
- Stop consonants in Kannada can be divided into velar, retroflex, dental and labial. Every series of stops includes voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated (e.g., /k/, /k^{h}/, ʈ,/ʈ^{h}/, d̪, d̪^{h} /p/, /p^{h}/) [57].
- In Kannada, the bilabial voiceless plosive /p/ at the beginning of many words has disappeared to produce a velar fricative /h/ or has disappeared completely [57].
- It has retroflex lateral approximant /ɭ/ in its sound inventory [16].
- It has contrast in singleton and geminate alveolar (/l/ and /lː/ and retroflex (/ɭ/ and /ɭː/) lateral approximants [16].

2.3 *Suprasegmental features of Indian language*

Suprasegmental features are features that accompany phonemes and are seen in this section as related to the language. Some of the suprasegmental features are sound pressure (intonation), stress (accent), tone and duration (consonant and vowel length). These features are not limited to a particular phone or sound like segmental but extended over syllables, phrases or words. They are also known as prosody features.

Stress (accent) is the relative prominence given to certain syllables in a word and one syllable usually stands out more prominently than the other syllables. Stress makes some sounds longer than unstressed syllables; it may introduce aspiration in initial stops [60] as seen in Hindi.

- Hindi is a syllable-timed language, meaning words are not distinguished based on stress alone. Default stress in Hindi is given on the last syllable [61].

- Hindi does have lexical stress and it is expressed in term of syllable lengthening [60].
- In Bengali, stress is at the word initial; the first syllable of the word carries the maximum stress, the third syllable carries somewhat weaker stress and all units with odd number of syllables carry very weaker stress [62].
- In Marathi, stress is at the word initial and is weight sensitive. Words with open syllables with a short vowel have light stress while closed syllables and open syllables with a long vowel have heavy stress. Intensity and duration are the most important clues for describing stress in Marathi [63]. An initial phoneme /a/ always carries stress, and a final phoneme /e/ carries stress if it is not preceded by phoneme/a/, and an initial phoneme /e/ carries stress if it is followed by phoneme /ʌ/ [26].
- In Gujarati, normally stress is on the first syllable when it does not have phoneme /a/. Stress mainly falls on the penultimate syllable of a word but it is attracted away from the final syllable if vowel in that syllable is more prominent than the one in the penultimate [64].
- In Assamese, stress is contrastive; as a result, the location of stress is unpredictable. Therefore, in case of Assamese, the whole sentence expresses prominent stress level [65]. In most of the cases, primary stress falls on the final syllable in the word.
- Tamil is syllable-timed; it has lexical stress with a complex vowel quantity [64]. Stress in Tamil shifts to the second vowel if the second vowel is long and the first is short and stress remains on the first vowel though the second syllable is closed [66].
- Telugu is a mora-timed language. Default stress is on the first syllable in Telugu [61]. Words containing long vowels have stress on the rightmost long vowel.
- For words in Telugu with two short syllables or the first syllable is long and second is short, stress falls on the first syllable. The stress falls on second syllable if the second one is long or both syllables are long. For a trisyllable word, if the first syllable is long then it carries stress; otherwise the penultimate syllable carries stress [67].
- In Malayalam, if the first syllable is short vowel then stress falls on it; it falls on the second syllable if it follows the first short one and if it has a long vowel [67]; otherwise, the primary stress falls on the first syllable; secondary stress falls on the remaining long vowels in the word.
- Kannada is spoken normally with no wide variation in stress; in multi-syllable words, the strongest stress is on the first syllable of the word while the word final syllable has normal stress [68].

Intonation is the relative variation in pitch over a word or a sentence and can be used to distinguish words. While pitch is related to fundamental frequency, the variation in pitch is influenced by language. Intonation difference can address attitudes and emotions of the speaker. In some languages, pitch variation is used to distinguish words either grammatically or lexically [69]. If languages use relative pitch variations to signal lexical differences, such languages are known as tone languages [5].

- In Hindi each content word except the final one has rising contour. Hindi is an accentual phrase (AP) language and it has three types of phrasal tones, namely, AP, intermediate phrase (ip) and intonational phrase (Ip) [70]. In Hindi, alignment of low tone (low pitch accent) is available if prominence is non-initial, so it has optional right shift [71].
- In Bengali, both intonation tunes and underlying tones can exist; however, it not necessarily uses tone as a contrastive feature. Bengali has uniform pattern of pitch contour for focus intonation and it has identical phonological form $H * L_1$ [71], while it does not have lexically specific pitch contour. There are three APs in Bengali and it has the starred tone low on stressed syllable with a sharp rise after it (*L**+*H*) [71].
- Assamese has four gliding pitches: falling (F), rising–falling (RF), rising (R) and falling–rising (FR). In Assamese, alignment of low tone (low pitch accent) is on word initial syllable as *L** can shift rightward. The low pitch accent (*L**) followed by a high boundary tone (Ha) (smooth rise) is the default tonal pattern in Assamese [71].
- Tamil intonation is rising contour that occurs on each lexical word except the last in a phrase. There is double rise (*L***H*..*LHa*) in longer AP [71].
- Telugu intonation is classified into falling (F), rising–falling (RF), rising (R) and falling–Rising (FR). It has 5 tone groups such as period pattern, the mid-level or slightly rising pitch pattern, the steeply rising pitch pattern, the falling or abruptly terminated pitch pattern and the comma pattern [72]. Vowel Length creates high tone plateaus in Telugu [71].
- Malayalam intonation is classified into rising (R), falling (F) and level (L). Question words in Malayalam have common intonation pattern such as MH-LML% (mid, high-(Ip), low, mid and low % (intonation boundary)).

As can be seen, there are several suprasegmental properties that are specifically dependent on the language. These properties can be exploited for the purposes of language identification, especially for natural spoken utterances. These properties of individual languages must carry some prelexical cues that enable development of language identification models.

- Different languages have different set phonemes and they maintain their own specific features.
- All languages have phonotactic constraints on their structural distribution of phonemes. In Gujarati and

Marathi languages, retroflex lateral approximant /ɭ/ and retroflex nasal /ɳ/ sounds never occur at word initial.

- Intonation and stress play important roles in discriminating the languages, such as Telugu is mora-timed language while Hindi is syllable-timed language.

Table 4 shows summary of language specific references used in the paper. However, any language-specific analysis needs a rich speech corpus. In the next section we review the list of available speech corpora.

3. Indian language speech corpus

For spoken language systems, development of speech corpora is essential for any research. In a multilingual country like India, systematic efforts have been taken in developing speech corpora in major languages. Speech corpora for Indian language are classified on the basis of features and purpose of development such as general purpose corpora, specific task corpora, acoustic-phonetic database, lexical, morphological, syntactical and semantic corpora [73].

Speech corpora that have been described in the later part of the paper were collected from peoples of different age groups, accents and sexes. Variability in speech corpora arises due to variation in speaking style, education status, recording environment, sampling rate and transmission channels. These corpora are recorded in different modes such as continuous and spontaneous reading, conversational mode, lecture mode, sentences and phrases. For recording of speech corpora, different software tools have been used and some of the speech corpora include conversation of TV News, TV talk shows, interviews, telephonic conversations and All India Radio.

Here, we have tried to compile information about speech corpora in Indian languages by different Government organizations, Indian academic institutes, research organizations and commercial companies. Table 5 lists the various speech corpora available for Indian languages.

3.1 *General purpose speech corpora*

The first Indian language corpus probably was the *Kolhapur Corpus for Indian English* developed at *Shivaji University Kolhapur*. This corpus consisted approximately of one million words of Indian English drawn from materials published in the year 1978 [74]. *CDAC, Kolkata*, sponsored by *TDIL, DeitY*, has developed an annotated speech corpora in three East Indian languages, namely, Bangla, Assamese and Manipuri. The corpora was developed with help of professional artists. The speech is recorded in a speech studio environment and digitized at a sampling rate of 22, 050 Hz, 16 bits/sample in PCM wave format. In case of speech, phonemes, syllables and breath pause have been annotated. The total size of the speech corpora is about 8.5 GB. Majority of this corpus is for Bangla language (5.12 GB) [75].

EMILLE-CIIL Corpus (Enabling Minority Language Engineering) consists of three components: monolingual, parallel and annotated corpora. It has 14 monolingual corpora, including both written and spoken data. Spoken data consist of 14 South Asian languages. These monolingual corpora consist of total 96 millions of words, including more than 2.6 million words of spoken corpora in Bengali, Urdu, Gujarati, Hindi and Punjabi. It consists of recording of everyday conversation among families and friends. It also includes recording of news, telephonic interviews and interviews. It is collaborative venture between *Lancaster University,* UK, and the *Central Institute of Indian Languages (CIIL)*, Mysore, India [76].

The spoken language group of *TIFR* developed a large multilingual spoken corpora for Indian languages. The speech database has been developed for four different languages such as Hindi, Marathi, Malayalam and Indian English and speech database has been collected over telephone channels. Phonetically rich corpora consisting conversational speech as well as read speech were collected by the '*Wizard of Oz*' speech data collection method [77].

3.2 *Application-based speech corpora*

DA-IICT prosody research team has collected speech data in two Indian languages: Marathi and Gujarati. These speech data were mostly collected from remote villages of Maharashtra and Gujarat states. The data are collected in three different modes, viz. read, spontaneous and lecture mode with age variations of speakers. The data have been collected using a portable handy recorder with 44.1 kHz sampling frequency with 16 bit/sample resolution [78].

Shruti is a read speech corpus designed by *IIT Kharagpur* in association with *Media Lab Asia*. It contains a total of 7383 unique sentences spoken by 34 speakers with different age categories. The speakers are from a region of West Bengal and text is collected from Anandbazar patrika, story book. The phonetic variations of Bengali language are designed from sports, political, geographical and general news [79].

Garhwali language speech database was developed for automatic speech recognition system at *Government PG college, Rishikesh*. Thousands of tokens/words that consist of spontaneous speech as well as phonetically rich sentences have been collected from different sources such as Garhwali newspapers, magazines and story books. A total number of 100 speakers consisting of 50 male and 50 female from different regions of Uttarakhand where Garhwali Hindi is spoken frequently were selected for speaking words/tokens. A *PRAAT* software tool is used for recording of words with two microphones for developing high quality speech as well as a noisy environment speech [80].

Marathi language speech database for ASR was developed by Samudravijaya. It consists of spontaneous as well as phonetically rich sentences of Marathi language. A

Table 4. Language-specific reference.

Language	References	Type
Marathi	[8, 10, 12, 13, 15, 16, 18, 20, 26, 63, 78]	Indo-Aryan
Gujarati	[8, 16–21, 64, 78]	Indo-Aryan
Hindi	[9, 15, 22–24, 26, 27, 27, 60, 61, 70, 71]	Indo-Aryan
Bengali	[6, 9, 26, 28–30, 32, 62, 71]	Indo-Aryan
Assamese	[21, 33–37, 64, 71]	Indo-Aryan
Tamil	[13, 38–42, 66, 71]	Dravidian
Telugu	[6, 43–48, 67, 71, 72]	Dravidian
Malayalam	[13, 49–56, 67]	Dravidian
Kannada	[16, 54, 57–59, 68]	Dravidian

Table 5. List of Indian language corpora.

Sl. no.	Name of corpora and developed by	Language	Duration and size	Usage
1	Kolhapur Corpus	English	1 million words	Not known
2	Annotated Speech Copora	Bangla, Assamese, Manipuri	8.5 GB	Speech and speaker recognition, synthesis
3	EMILLE-CIIL Corpus	Bengali, Urdu, Gujarati, Hindi, Punjabi	6 million words	Language and speech recognition
4	DA-IICT	Marathi, Gujarati	Varies from 4.5 to 11.5 h	Speech and speaker recognition
5	Shruti Speech Corpus	Bengali	7383 unique sentences	Speech recognition
6	Garhwali Speech Database	Garhwali Hindi	Thousands of tokens	Speech recognition
7	TIFR Speech Corpus	Hindi, Marathi, Malayalam	Not known	Spoken language recognition
8	Marathi Speech Corpus-BAMU	Marathi	28420 isolated words	Speech recognition
9	Marathi Speech Database	Marathi	10 sentences	Speech and speaker recognition
10	Assamese Speech Corpus	Assamese, Hindi	Not known	Language and speaker recognition
11	LVCSR, Anna University	Tamil, Telugu	17 h	Speech recognition
12	LDC-IL, Mysore	22 Indian Languages	Different hours duration	Language, speech recognition, synthesis
13	IITKGP-MLILSC	27 Indian Languages	Minimum 1 h	Language identification
14	KIIT Bhubaneswar	Hindi, Indian English	630 sentences	Mobile-based speech recognition
15	TIFR Mumbai, CDAC Noida	Hindi	10 sentences	General purpose
16	CDAC Noida (Travel Domain)	Hindi	8567 sentences	Speech recognition for travel domain
17	Speech Database, IIT Kharagpur	Hindi, Telugu, Kannada, Tamil	17.5 h	General purpose

sequence of phonemes for Marathi language was taken from online text and that was broken into phrases, which was manually validated. *CorpusCrt* was used to derive the sets of Marathi sentences. This database collected speech data from many speakers of different age groups who speak fluently in various dialects of Marathi language [81].

Another *Marathi language speech database* was created at *Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad,* for speech recognition system. It consists of 28420 isolated words and 17470 sentences of Marathi language. Speech corpus was collected from speakers of different age groups, accents, genders and education statuses without noisy environment and echo effect. *CMU lexicon tool* and the *ARPAbet symbol* have been used for speech transcription [82].

Assamese language speech database was collected by authors of *IIT Guwahati* for development of an Assamese phonetic engine [83]. Speech database for this work was collected from 25 native Assamese speakers and 3 non-native speakers in three different modes: lecture, reading and conversation mode. Data were collected by using two channels, namely a telephone channel, where data were recorded at 8 kHz sampling frequency and 16 bits/sample resolution, and the other channel was a head set microphone, where data were recorded at 48 kHz sampling frequency and 16 bits/sample resolution. It was transcribed using the *IPA symbols*.

Tamil language speech database was created at *Department of Computer Technology, Anna university,* for *Large Vocabulary Continuous Speech Recognition (LVCSR)*. All speech data were collected in two stages: in the first stage, 68 speakers read lines from Tamil literary classics, which is around 17-hour speech data, and in the second stage, 29 speakers read newspapers, which is 1-hour speech data. All data were recorded using a microphone and in a quite environment with 16 kHz sampling frequency and 16 bits/sample.

The *Linguistic Data Consortium for Indian Languages (LDC-IL)* helps researchers and developers worldwide in the field of Indian language. It is governed by *Central Institute of Indian Languages, Mysore*. It consists of 45 million corpora in 15 Indian languages. It has speech corpora as raw data as well as segmented data in 22 Indian languages and annotated speech corpora in 16 Indian languages [84].

IITKGP-MLILSC Speech Database consists of 27 Indian languages. For collection of speech database, speech was recorded from TV news bulletins, talk shows, interviews and live shows, while remaining speech data were collected from Pasar Bharati All India Radio. Each language contains minimum 1-hour speech data with at least 10 male and female speakers. *Audacity software* is used to record the speech from TV channels. The speech signal is recorded at 16 kHz sampling frequency and each sample is stored in 16 bits [85].

The activity of development of speech corpora for Indian Languages is being pursued in many Indian academic institutes but still it does not have a common procedure, standard, environmental and recording condition.

4. Conclusion

With recent advances in speech signal processing, there has been a renewed interest in SLID and recognition. While there is a lot of literature that exists for Western and European languages, there is not much information available for Indian languages. Speech as a mode of communication with machines is very apt and important in the context of India, especially because of the low literacy levels, which enables people to communicate only in their (spoken) native language. This requires that there are speech recognition systems for Indian languages. With several languages officially recognized in India the task of building speech applications is complex. One of the first steps in building a reliable speech-based application is the recognition of the language being spoken. In this review paper, keeping the SLID task in mind we have done an extensive literature survey to gather information about the specific language-dependent properties of Indian languages. The idea is to enable researchers to exploit these properties to enable them to build algorithms and systems that can identify the spoken language. We have also listed the speech databases that exist for Indian languages as part of this review paper.

References

[1] Ambikairajah E, Li H, Wang L, Yin B and Sethu V 2013. Language identification: a tutorial. *IEEE Circuits and Systems Magazine* 11: 82–108

[2] Phonetics 2015 http://en.wikipedia.org/wiki/Phonetic

[3] Li H, Ma B and Lee K A 2013 Spoken language recognition: from fundamentals to practice. *Proceedings of the IEEE* 5: 1136–1159

[4] Reddy M V, Hanumanthappa M and Jyothi N M 2014 Phonetic dictionary for natural language processing: Kannada. *International Journal of Engineering Research and Applications* 4(7): 01–04

[5] Bhaskararao P 2011 Salient phonetic features of Indian languages in speech technology. *Sadhana* 36(5): 587–599

[6] Mohanty S 2011 Phonotactic model for spoken language identification in Indian language perspective. *International Journal of Computer Applications* 19(9): 18–24

[7] Koch D B, McGee T J, Bradlow A R and Kraus N 1999 Acoustic-phonetic approach toward understanding neural processes and speech perception. *Journal of the American Academy of Audiology* 10: 304–318

[8] Patil V and Rao P 2011 Acoustic features for detection of aspirated stops. In: *Proceedings of the IEEE 2011 National Conference on Communications (NCC)*, pp. 1–5

[9] Esposito C, Hurst A, *et al* 2005 *Breathy nasals and / Nh/clusters in Bengali, Hindi, and Marathi*. http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.499.8360

[10] Indo-Aryan languages 2017 http://www.languagesgulper.com/eng/Malayalam.html

[11] Dravidian languages 2017 http://www.languagesgulper.com/eng/ Malayalam.html

[12] Shinde R B and Pawar V P 2012 A review on acoustic phonetic approach for Marathi speech recognition. *International Journal of Computer Applications* 59(2): 40–44

[13] Index of sound 2015 https://en.wikipedia.org/wiki/Marathi-language

[14] Marathi Language 2015 http://www.phonetics.ucla.edu/index/sounds.html

[15] M R Mhaiskar 2014 Change in progress: phonology of Marathi-Hindi contact in Eastern Vidarbha. *International Journal of English Language, Literature and Humanities* 2(7)

[16] Retroflex Lateral Approximant 2015 http://self.gutenberg.org/articles/retroflex-lateral-approximant

[17] Gujarati Language Gulper 2017 http://www.languagesgulper.com/eng/Gujarati.html

[18] Esposito C M, *et al* 2012 Contrastive breathiness across consonants and vowels: a comparative study of Gujarati and White Hmong. *Journal of the International Phonetic Association* 42(2): 123–143

[19] Rami M K, Kalinowski J, Stuart A and Rastatter M P 1999 Voice onset times and burst frequencies of four velar stop consonants in Gujarati. *Journal of the Acoustical Society of America* 106(6): 3736–3738

[20] Thati, and Bollepalli B, Bhaskararao P and Yegnanarayana B 2012 Analysis of breathy voice based on excitation characteristics of speech production. In: *Proceedings of the IEEE International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5

[21] Agrawal S S 2008 Analysis of breathy voice based on excitation characteristics of speech production. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/agrawal2008.pdf

[22] Gaurav, Deiv D S, Sharma G K and Bhattacharya M 2012 Development of application specific continuous speech recognition system in hindi. *Journal of Signal and Information Processing* 3: 394

[23] Rajput P and Lehana P 2014 *Effect of the proportion of harmonic and noise part on the quality of synthesized speech using HNM in Hindi language*. MAGNT Research Report, http://brisjast.com/wp-content/uploads/2014/11/Dec-10-2014.pdf, vol. 2, no.7, pp. 116–133

[24] *Hindustani phonology* 2015 https://en.wikipedia.org/wiki/Hindustani-phonology

[25] *Shodhganga* 2017 http://shodhganga.inflibnet.ac.in/bitstream/10603/25106/7/07-chapter%202.pdf

[26] Berkson K H 2012 *Phonation types in Marathi: an acoustic investigation*. PhD Dissertation

[27] Kanth B L, Keri V and Prahallad K S 2011 Durational characteristics of Indian phonemes for language discrimination. In: *Information systems for Indian languages*. Springer, pp. 130–135

[28] *Index of sound* 2017 https://books.google.co.in

[29] *The language gulper* 2015 http://www.languagesgulper.com/eng/Bengali.html

[30] Barman B 2008 *Distinctiveness of aspiration in Bangla*. Daffodil International University

[31] Das A 2009 The distribution of aspirated stops and/h/in Bangla: an optimality theoretic approach. *Linguistics Journal* 4(2): 51–76

[32] *The Indo-Aryan languages* 2015 https://books.google.co.in/books/about/The-Indo-Aryan-Languages.html?id=Itp2twGR6tsC&redir-esc=y

[33] Das A 2009 The distribution of aspirated stops and/h/in Bangla: an optimality theoretic approach. *International Journal of Electrical and Electronics Engineering* 3(5): 2281–2291

[34] Sarma M, Dutta K and Sarma K K 1982 *Assamese numeral corpus for speech recognition using cooperative ANN architecture*. Department of Publication, Gauhati University, http://www.amazon.co.uk/Structure-Assamese-Golockchandra-Goswami/dp/B005VYXSDG

[35] Sarma M and Sarma K K 2013 An ANN based approach to recognize initial phonemes of spoken words of Assamese language. *Applied Soft Computing* 13(7): 116–133

[36] The Language Gulper 2015 http://www.languagesgulper.com/eng/Assamese.html

[37] Devi M, Thakuria L K, Purnendu B A and Talukdar P 2014 A study on acoustic behavior of Assamese nasal phoneme. *Journal of Harmonized Research in Engineering* 2(1): 235–238

[38] Srinivasan A, Rao K S, Kannan K and Narasimhan D 2010 *Speech recognition of the letter'zha'in Tamil language using HMM*. arXiv preprint arXiv:1001.4190

[39] Tamil Phonology 2015 https://en.wikipedia.org/wiki/Tamil-phonology

[40] Pushpa N, Revathi R, Ramya C and Hameed S S 2014 Speech processing of Tamil language with back propagation neural network and semi-supervised training. *International Journal of Innovative Research in Computer and Communication Engineering* 2(1): 2718–2723

[41] Thangarajan R, Natarajan A M and Selvam M 2008 Word and triphone based approaches in continuous speech recognition for Tamil language. *WSEAS Transactions on Signal Processing* 4(3): 76–86

[42] South Asia Language Resource Center 2015 http://www.southasia.sas.upenn.edu/tamil/consonants.html

[43] Telugu Language 2015 https://en.wikipedia.org/wiki/Telugu-language

[44] Nagamani M and Girija P N 2015 *Pronunciation variant and substitutional error analysis for improving Telugu language lexical performance in ASR system accuracy*. http://www.ijser.org/paper/Pronunciation-Variant-and-Substitutional-error-analysis-for-Improving-Telugu-Language-Lexical-performance-in-ASR-system-Accuracy.html.

[45] Girija P N and Sridevi A 1995 *Duration rules for vowels in Telugu*. Department of Computer & Information sciences, AI Lab, University of Hyderabad, file:///C:/Users/Dell/Downloads/Duration-Rules-For-Vowels-In-Telugu

[46] Datta A K, Ganguli N R and Ray S 1980 Recognition of unaspirated plosives—a statistical approach. *IEEE Transactions on Acoustics, Speech and Signal Processing* 28(1): 85–91

[47] Rajashekhar B, *et al* 2013 Vowel duration across age and dialects of Telugu language. *Language in India* 13(2)

[48] Bhat D N S 1973 *Retroflexion: an areal feature*. Working Papers on Language Universals, No. 13, ERIC

[49] Malayalam Phonology 2015 https://en.wikipedia.org/wiki/Malayalam

[50] Malayalam Language Gulper 2017 http://www.languagesgulper.com/eng/Malayalam.html

[51] Jiang H 2010 *Malayalam – a grammatical sketch and a text*. Department of Linguistics, Rice University

[52] Mohanan K P 1986 *The theory of lexical phonology*. In: Studies in Natural Language and Linguistic Theory, Springer, pp. 63–108

[53] Acoustic Characteristics of Speech Sound 2017 http://isites.harvard.edu/fs/docs/icb.topic482062.files/Reetz-Jongman

[54] Cross-Language Study of Stop Perception 2017 http://www.drsavithri.com/articles/Cross%20langauge%20study%20of%20stop%20perception.pdf

[55] George J, Abraham A S, Arya G S and Kumaraswami S 2015 Acoustic characteristics of stop consonants during fast and normal speaking rate in typically developing Malayalam speaking children. *Language in India* 15: 47

[56] Local J and Simpson A P 1999 Phonetic implementation of geminates in Malayalam nouns. *Work* 4(92): 46

[57] Kannada Language 2015 http://aboutworldlanguages.com/kannada

[58] Hemakumar G 2011 Acoustic phonetic characteristics of Kannada language. *International Journal of Computer Science Issues* 8(2): 332–339

[59] Manjunath N, Varghese S M and Narasimhan S V 2010 Variation of voice onset time (VOT) in Kannada language. *Language in India* 10(5): 170–181

[60] Dyrud L O 2001 *Hindi–Urdu: stress accent or non-stress accent*. University of North Dakota

[61] Sirsa H and Redford M A 2013 The effects of native language on Indian English sounds and timing patterns. *Journal of Phonetics* 41(6): 393–406

[62] Bengali Phonology 2017 https://en.wikipedia.org/wiki/Bengali-phonology

[63] Le Grézause E 2015 *Investigating weight-sensitive stress in disyllabic words in Marathi and its acoustic correlates*

[64] The Handbook of Phonological Theory 2017 https://books.google.co.in/books

[65] Sarma P and Sarma S K 2016 A study on detection of intonation events of Assamese speech required for tilt model. *Int. J. Comput. Appl.* 154: 34–38

[66] Vijayakrishnan K G 2015 The path from the prosodic phonology of loans to second language phonology: two case studies. http://www.iitg.ernet.in/wti3/img/abstract/%20Vijayakrishnan.pdf

[67] The Dravidian Languages 2017 https://books.google.co.in/books

[68] Leonard A P 1964 *Partial analysis of the phonology of formal Kannada*. The University of Montana, Missoula

[69] Intonation 2017 https://en.wikipedia.org/wiki/Intonation-(linguistics)

[70] Sengar A, Mannell R, *et al* 2012 A preliminary study of Hindi intonation. *Proc. SST*

[71] The Intonation of South Asian Languages Towards a Comparative Analysis 2017 https://www.reed.edu/linguistics/khan/assets/Khan2016-FASAL.pdf.

[72] Bhuvaneswar C 2017 *Intonation in English and Telugu proverbs: evidence for Karmik linguistic theory*

[73] Agrawal S, Samudravijaya K and Arora K 2006 Recent advances of speech databases development activity for Indian languages. In: *Proceedings of ISCSLP*

[74] Shastri S V 1988 The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal* 12: 15–26

[75] CDAC Corpus 2015 http://cdac.in/index.aspx?=mc-i/f-speech-corpora

[76] EMILLE Corpora 2015 http://catalog.elra.info/search-result.php?keywords=W0037&language=en

[77] Samudravijaya K 2006 Development of multi-lingual spoken corpora of Indian languages. In: *Proceedings of the Chinese Spoken Language Processing Symposium*, pp. 792–801

[78] Malde K D, Vachhani B B, Madhavi M C, Chhayani N H and Patil H A 2013 Development of speech corpora in Gujarati and Marathi for phonetic transcription. In: *Proceedings of Oriental COCOSDA held jointly with 2013 IEEE Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp. 1–6

[79] Shruti Bengali Continuous ASR Speech Corpus 2015 http://cse.iitkgp.ac.in/pabitra/shruti-corpus.html

[80] Upadhyay R K and Riyal M K 2010. Garhwali speech database. In: *Proceedings of O-COCOSDA*.

[81] Samudravijaya K and Gogate M R A 2006 Marathi speech database. In: *Proceedings of the International Symposium on Speech Technology and Processing Systems and Oriental COCOSDA-2006*, Penang, Malaysia, http://speech.tifr.res.in/chief/publ/06ococosdaMarathiDatabase.pdf, pp. 21–24

[82] Gaikwad S, Gawali B and Mehrotra S 2013 Creation of Marathi speech corpus for automatic speech recognition. In: *Proceedings of Oriental COCOSDA held jointly with 2013 IEEE International Conference Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp. 1–5

[83] Sarma B D, Sarma M, Sarma M and Prasanna S R M 2013 Development of Assamese phonetic engine: some issues. In: *Proceedings of the Annual IEEE India Conference (INDICON)*, pp. 1–6

[84] Size of Speech Corpora 2015 http://www.ldcil.org/resourcesSpeechCorp.aspx.

[85] Maity S, Vuppala A K, Rao K S and Nandi D 2012 IITKGP-MLILSC speech database for language identification. In: *Proceedings of the IEEE National Conference on Communications (NCC)*, pp. 1–5