Sādhanā Vol. 36, Part 5, October 2011, pp. 587–599. © Indian Academy of Sciences

Salient phonetic features of Indian languages in speech technology

PERI BHASKARARAO

Tokyo University of Foreign Studies, Tokyo, Japan
e-mail: bhaperi@gmail.com

Abstract. Speech signal is the basic study and analysis material in speech technology as well phonetics. To form meaningful chunks of language, the speech signal should have dynamically varying spectral characteristics, sometimes varying within a stretch of a few milliseconds. Phonetics groups these temporally varying spectral chunks into abstract classes roughly called as allophones. Distribution of these allophones into higher level classes called phonemes takes us closer to their function in a language. Phonemes and letters in the scripts of literate languages – languages which use writing have varying degrees of correspondence. As such a relationship exists, a major part of speech technology deals with the correlation of script letters with chunks of time-varying spectral stretches in that language. Indian languages are said to have a more direct correlation between their sounds and letters. Such similarity gives a false impression of similarity of text-to-sound rule sets across these languages. A given letter which has parallels across various languages may have different degrees of divergence in its phonetic realization in these languages. We illustrate such differences and point out the problem areas where speech scientists need to pay greater attention in building their systems, especially multilingual systems for Indian languages.

Keywords. Acoustic-phonetic segment; allophone; code-point; multiple source; phone; phoneme.

1. Introduction

Speech signal of an utterance in a language is the only physical event that can be recorded and reproduced. The signal can be further processed in two directions – signal and linguistic processing. During linguistic processing, signals are cut into chunks of varying degrees of abstraction such as acoustic-phonetic segments, allophones, phonemes, morphophonemes, etc, which will be ultimately correlated with the letters in the script of a language. We will examine some of the important aspects of linguistic processing of speech signals of Indian languages.

1.1 *Basic sound units from a linguistic perspective*

The spoken output of a natural language is a dynamically changing acoustic signal. From an acoustic-phonetic point of view, based upon distinct changes in time and frequency domains, cer-

tain 'acoustic-phonetic segments' (APS) can be demarcated. As every change in these domains cannot be interpreted as a segment boundary, decisions of segment demarcation are based on phonetic criteria. From linguistic point of view, the same signal that is demarcated into APSs can be correlated with linguistic abstractions such as allophones, phonemes, morphophonemes, etc. However, the correlation of strings of one or more APSs with 'allophones' can never be of one-to-one type. This inherent lack of correlation is further compounded by the fact that the generator of speech signal, viz., the human speaker cannot and does not produce invariable and exact copies of the signal for different instances of the 'same' allophone. However, there are some 'landmark' events in a speech signal that have a good amount of correlation to some classes of speech sounds. Correlating such acoustic landmarks with the abstract speech sounds is a possible task. Although the number of possible speech sounds that a human can produce is extremely large, out of those 'articulatory possibilities of man', only a few groups of them are utilized by any particular natural language. For instance, voiced aspirates sounds such as [b^h, d^h, d^h, j^h, g^h] are common among Indo-Aryan languages such as Hindi and Marathi, but they are absent in Kashmiri and English.

The relevance of phonetics for speech technology, specifically text-to-speech synthesis is well known (van Santen 2005). As one of the major purposes of speech technology is building systems that convert speech to text or generate speech from text, it may depend heavily upon finding out the correlates of groups of APSs with allophonic variants. After the APSs are identified, further processing of the segment stream should be performed by the linguistic processing module. In this module, a string of APSs (consisting of one or more such segments) are associated with the 'allophones' of the language. Each of these allophones are recognized as members of classes called 'phonemes'. This results in a string of phonemes. Each string of phonemes (consisting of one or more phonemes) is then associated with another underlying unit called 'morphophoneme'. Then certain chunks of morphophonemes in the string are matched with an underlying string that correlates with the mental lexicon or mental grammatical elements that are stored in the memory. These strings are either words in the dictionary or various grammatical items such as suffixes etc. Once this process is completed, the resulting string of distinct words is then transformed into the corresponding string of code-points which gets transformed onto way the string is written in the concerned script.

1.1a *Script letters to phonemes*: As letters in a script are more easily perceivable than to the abstract allophones, we will take the route of script→morpho-phoneme→phoneme→ allophone→APS. A majority of the Indian scripts are derived from an ancient script called Brahmi. The Brahmic scripts are: Assamese; Bengali; Devanagari (mainly used in Hindi, Marathi, Sanskrit, Nepali); Gujarati; Gurumukhi (for Panjabi); Kannada; Oriya; Tamil and Telugu. In addition, Perso-Arabic script is used for writing Urdu, Kashmiri and Sindhi – however, these languages are also written in Devanagari, a Brahmic script.

The script system of each of these languages consists of a table of characters. Each of these characters correspond to one or more phonemes. Although most of the characters and phonemes have a one-to-one equation, some of them have one-to-many or many-to-one equation. Some examples follow (characters are marked by { } and phonemes by//):

one character = one phoneme: Telugu, Hindi:{k} ↔ /k/

one character = two phonemes: Telugu: {eː} = /eː/, /æː/

two characters = one phoneme: Telugu: {t^h}, {d^h} ↔ /d^h/

two characters = two phonemes of different value: Panjabi: {d^hV} = /tV́/ (where V́ represents a vowel with a high tone).

1.1b *Words to sentences and sandhi process*: Except for 'single word' utterances such as 'come!', 'yes', 'one', most of the utterances in a natural language are composed of strings of words and grammatical forms. Every word in a sentence should be recoverable either from its lexicon or in conjunction with its grammar. For instance, the following lexical words and suffixal additions (and the grammatical rules that control them) are recoverable from the lexicon and grammar of Hindi: /laɽaki:/ 'girl'; /laɽaki-yã:/ 'girls'. /laɽaki:/ can be usually recovered directly from the lexicon and /laɽaki-yã:/ will be recovered by applying an approriate plural inflectional rule on the lexical form /laɽaki:/. A speaker of Hindi language will put together the necessary lexical words and their inflected words to form the uttered sentences. However, in natural conversation, one does not utter them as separate words with pauses between each word. They are uttered as a continuous string which is reflected as a continuous waveform of the output. When words are uttered in a continuous way, in most of the languages, the phonemes at junctions between the words may undergo changes – sometimes the changes in phonemes even jump over the junctions and affect phonemes that are further away from the junctions. Such changes are called '*sandhi* changes'. For a natural sounding machine speech output, *sandhi* changes constitute a very important component. Most of the *sandhi* changes can be captured by context-sensitive rules. *Sandhi* rule sets are specific to a specific language. Although two languages may share some *sandhi* rules, they will not have the same set of *sandhi* rules. After a string of text is converted into a string of phonemes, before allophonic rules are applied, the phoneme string undergoes *sandhi* process. If the phoneme string does not undergo *sandhi* process, then the output will sound very mechanical (uttered as separate words instead of a natural sounding continuous sentence).

1.1c *Phonemes to allophones*: Phonemes are a class of allophones. Allophones are context-dependent realizations of a phoneme. Most of the words in a language are made of strings of phonemes and the surrounding phonemes have an influence on the selection of an allophone of a given phoneme. Sometimes the context of influence may extend beyond left or right adjacency.
Examples:
An example from Telugu showing the realization of letter {m} into its phonemic equivalent /m/ and the allophonic realizations of /m/:

1. {m} → /m/
2a. /m/ → [ɱ]/ V___V, #
2b. /m/→ [m]/ elsewhere

where:
{m} is the script letter
/m/ is a phoneme
[ɱ] is a labiodental nasal allophone
[m] a bilabial nasal allophone.
Allophone assignment rules will locate the phoneme /m/ and check its left and right contexts. If they find a vowel (V=any vowel) on the left, and on the right if a V or word boundary (#) is found, then the /m/ is replaced by its allophone [ɱ]. If any part of the above context is not satisfied, /m/ is replaced by [m].
An example from Bengali:
Text-to-phoneme rule affecting the three fricative letters:

(1) {ʃ}, {ʂ}, {s} → /ʃ/

Phoneme-to-allophone rules effecting the output of the above rule:

(2a) /ʃ/ → [s]/ __DS
(2b) /ʃ/ → [ʃ]/ elsewhere

In Bengali script there are three 'fricative' letters. They are maintained in the spelling of the language. However, their pronunciation collapses into one phoneme. Rule 1 above, shows the merger of the three script letters {ʃ}, {ʂ}, {s} into one phoneme: /ʃ/. However, this phoneme /ʃ/ is phonetically realized as two allophones: Rule (2a) /ʃ/ → [s]/ __DS converts the input /ʃ/ into the output allophone [s] if it is followed by a dental sound such as /t, n, l/; (2b) takes the remaining instances of /ʃ/ as input and gives out the output allophone [ʃ].

1.1d *Underlying forms to final phonetic output*: Here we will trace the derivation of the final phonetic output from the deepest underlying form of a string of script letters constituting an utterance. We will explain this through four examples from two languages.
Examples:
Hindi

(1) Utterance meaning: 'He made profit'

(a) Script input: {laːbʰa paːyaː}
(b) Pre-*sandhi* phoneme string output: //laːbʰapaːyaː//
(c) Post-*sandhi* phoneme string output after applying *sandhi* rules: /laːbpaːyaː/
(d) Allophonic string output from the above: [laːbpaːyaː]

Explanation: The pre-*sandhi* sequence //bʰap// in (b) is converted into /bʰp/ in (c) by the application of Schwa-deletion rule; and //bʰ + p// is then converted into /bp/ in (c) by the application of 'deaspiration rule'. Both these rules (Schwa-deletion and deaspiration) are *sandhi* rules. The post-*sandhi* phoneme string /bp/ is realized as the allophone string [bp].

(2) Utterance meaning 'If there is happiness…'

(a) Script input: {yadi saṁtoṣa hai}
(b) Pre-*sandhi* phoneme string output: //yadisantoṣhai//
(c) Post-*sandhi* phoneme string output after applying *sandhi* rules: /yadisantoṣhai/
(d) Allophonic string output: [yədisəntoʂɦiɛ)

Explanation: {saṁtoṣa} of (a) becomes //saṁtoṣ// in (b) by the application of Schwa-deletion rule. The string /yadisantoṣhai/ of (c) is converted into the allophonic string of [yədisəntoʂɦiɛ] in (d) by applying the appropriate 'phoneme-to-allophone' conversion rules.

Telugu

(3) Utterance meaning: 'Bring it!'

(a) Script input: {adi teː}
(b) Pre-*sandhi* phoneme string output: //aditeː//
(c) Post-*sandhi* phoneme string output after applying *sandhi* rules: /atteː/
(d) Allophonic string output: [ɐtteː]

Explanation: //adi teː// is converted into /atteː/ by the application of two *sandhi* rules: (i) deletion of /i/ between two homorganic consonants (/d/ and /t/ in this case) followed by assimilation of the

two homorganic consonants (/d/ and /t/ here becoming /tt/). /a/ is converted into the appropriate [ɐ] allophone because of the presence of a low vowel (/eː/) in the next syllable.

(4) Utterance meaning: 'That is enough'

(a) Script input: {adi sari}
(b) Pre-*sandhi* phoneme string output: //adisari//
(c) Post-*sandhi* phoneme string output after applying *sandhi* rules: /assari/
(d) Allophonic string output: [ɜssaɾi]

Explanation: //adi sari// is converted into /assari/ by the application of two *sandhi* rules: (i) deletion of /i/ between two homorganic consonants (/d/ and /s/ in this case) followed by assimilation of two homorganic consonants (/d/ and /s/ here becoming /ss/). /a/ is converted into the appropriate [ɜ] allophone because of the presence of a high vowel (/i/) in the next syllable; /r/ is converted into the appropriate tap allophone [ɾ] because it occurs between two vowels (/a/ and /i/).

Comparison. In the Hindi example (1) the final phonemic output of sequence //b^{h}p// is /bp/; whereas in the Telugu example (3), a similar sequence //dit// results first in //dt// and then in the final phonemic output /tt/. Note that Telugu assimilates /d/ to /t/ by removing the voicing from /d/; whereas Hindi does not remove the voicing from //b// in //bp//. In the Hindi example (2), the final phonemic output of sequence //dis// is still /dis/; whereas in Telugu example (4), a similar sequence //dis// is converted into the final phonemic string /ss/. Thus, Telugu employs a stronger assimilation process.

These examples altogether show that the *sandhi* rule sets differ from language to language. They also show what looks similar to the same phoneme across two languages may (and will) have different allophonic outputs.

1.1e *Allophones to acoustic-phonetic segments*: An allophone is a group of one or more APSs. Phonetics distinguishes two types of sounds – segmental and suprasegmental. [p], [b], [s], [ʃ], [v], [r] are some examples of segmental sounds. Vowel length, consonant length, stress, pitch of a vowel or other voiced continuants (such as [m], [l]) are suprasegmental sounds. All segmental sounds are composed of one or more APSs arranged sequentially in time. Among suprasegmentals, vowel and consonant length can also be viewed as a time-wise sequential arrangement. Pitch variations that contribute to the intonation or tone of a segment has to be viewed as an APS that is superimposed on another sound.

Derivation of the string of allophones from the underlying string of letters is mostly in the realm of linguistic processing. Matching the allophone string with a string of APSs is at the interface between linguistic processing and the actual speech signal. In the following sentence, an example from Telugu, we will trace the route of allophone-to-APSs. The route is as follows: {script string} → //pre-*sandhi* phoneme string// → /post-*sandhi* phoneme string/ → [allophone string] → APS string → signal

In the following example, !! demarcate the sets of changes that the preceding string undergoes to produce the following string.
Meaning of sentence: *'I showed him to you'*

(a) {vaːḍi-ni niːku cuːpiṁceːnu}
! ṁc → nc !
(b) //vaːḍinini:kucuːpinceːnu//
! ḍinin → ṇṇ !

(c) /vaːn̪n̪iːkucuːpinceːnu/
! aː→ ɐː u → ů c → ʦ I → ɪ c → ʧ u → ů !
(d) [ʊɜːn̪n̪iːkůtsuːpɪnʧeːnů]
(e) Acoustic-phonetic segment string:

The APS string will consist of values of one or more of the items listed under A to G in table 1. (In column Z, the phonetic name of the allophone is given). This table lists the various segments that are necessary to define the quality of each of the concerned sounds. The terms used in the table are: phonetic description, pitch pulse, transition from, transition to, steady-state, plosion spike and VOT (Voice Onset Time) lag. 'Phonetic description' is a taxonomic term used in articulatory phonetics. 'Pitch pulse' indicates whether the sound is voiced (with vocal cord vibration) or not. 'Transition from' and 'transition to' show whether the spectral peaks of the pitch pulses change dynamically while moving from one sound to another sound – these transitions are clear indicators of change of place of articulation (a phonetic term). 'Steady state' shows that the particular speech sound is produced without noticeable changes in the place of articulation. 'Closure' shows whether the airstream is completely stopped for a certain amount of duration during the concerned segment. 'Plosion spike' refers to the burst after the release of a closed set of articulators (phonetic term). 'VOT lag' stands for the duration after 'plosion spike' and the onset of the first pitch pulse.

1.2 *Phonetic events*

A better understanding of the terms given here can be achieved by quickly going through the various gestures of the organs of speech in the production of speech sounds and their gross acoustic correlates. Fine tuning of the acoustic correlates is achievable on the basis of the analysis of a single model speaker of a given language.

Pitch pulse. Each vocal cord vibration in the production of voicing in all voiced sounds produces a pitch pulse. The background 'buzzing' heard during the closure of a voiced stop is also composed of pitch pulses. Pitch pulses are the major source of acoustic energy in a majority of speech sounds in the languages of the world.

Stricture. The degree of opening between the two articulators in defining a place of articulation is a stricture. Strictures roughly correspond to 'manners of articulation'. Zero stricture stands for total absence of an opening as in the case of stop sounds. Fricatives have a stricture that is slightly wider than zero stricture. Approximants have wider stricture than fricatives and vowels have the widest stricture. During the production of a zero stricture, if there are no concurrent pitch pulses, acoustic energy will be totally absent. However, in the case of voiced stop sounds, the pitch pulses present in the background laryngeal vibration will give out muffled energy which is termed as 'voice bar'.

Strictures are prolongable or momentary. Prolongable strictures are those that can be maintained for a longer time as in the case of stops, fricatives, laterals, etc. Momentary strictures are non-prolongable as in the case of flaps. Trills have intermittent strictures – they are alternant instances of momentary and opener strictures. However, trills are prolongable by producing the alternant strictures as long as necessary. Sounds produced with prolongable strictures can differ in duration, wheras sounds with momentary gestures cannot have durational difference. The nature of a stricture has an impact on the type of acoustic event that one may expect to find in the production of the concerned sound.

Table 1. Sample of allophone-APS strings.

	Z Phonetic description	A Pitch pulses	B Transition from	C Transition to	D Steady-state	E Closure	F Plosion spike	G VOT lag
[ʊ]	labio-dental approximant	+	silence	> [ɜ]	short	−	−	−
[ɜː]	open-mid central unrounded vowel	+	< [ʊ]	> retroflex	long	−	−	−
[ɳɳ]	retroflex nasal consonant	+	< [ɜ]	> front vowel	long	oral closure + nasal murmur	−	−
[iː]	close front unrounded vowel	+	< retroflex	> velar	long	−	−	−
[k]	voiceless velar plosive	−	< front vowel	> back vowel	short	voiceless closure	Velar plosion	+
[ů]	extra short close back rounded vowel	+	< velar	> dental	extra short	−	−	−
[ʦ]	dental affricate	−	< back vowel	> back vowel	short	voiceless closure	Dental plosion	heavily fricativized
[uː]	long back rounded vowel	+	< dental	> bilabial	long	−	−	−
[p]	bilabial plosive	−	< back vowel	> front vowel	short	voiceless closure	Bilabial plosion	+
[ɪ]	close-midfront unrounded vowel	+	< bilabial	> dental	short	−	−	−
[n]	dental nasal	+	< front vowel	> palatal	short	oral closure + nasal murmur	−	−
[ʧ]	palatal affricate	−	< dental	> front vowel		voiceless closure	Palatal plosion	heavily fricativized
[eː]	long close-midfront unrounded vowel	+	< dental	> dental	long	−	−	−
[n]	dental nasal	+	< front vowel	> back vowel	short	oral closure + nasal murmur	−	−
[ů]	extra short close back rounded vowel	+	< dental	> silence	extra short	−	−	−

Source of acoustic energy. The major source of acoustic energy is the pitch pulse of all the voiced sounds. Strictures are a major source of energy for all voiceless plosives, voiceless fricatives, voiceless affricates, flaps and voiceless trills. In the case of voiced plosives, voiced fricatives, voiced affricates, voiced trills, pitch pulses as well as strictures contribute to the acoustic energy. In the case of voiced plosives, the two sources are used consecutively; first the pitch pulses during the closure duration followed by the stricture source from the plosion and the tapering off fricative noise during the VOT.

Nasal hum. In the case of a nasal sound, the airstream is divided into two channels – the oral and the nasal channels. This division gives rise to 'nasal hum' with its distinct acoustic characteristics.

Formants. The major spectral peaks in the pitch pulses. The spectral peaks of consecutive pitch pulses have present roughly in the same region giving rise to an impression of continuity of peaks. This 'continuity' of peaks is known as a formant.

Transitions. Transition is a gradual change of the region of the formants over two or more contiguous pitch pulses. They are the major indicators of the places of articulation of the consonant sounds they are pointing to.

VOT lag. The time gap between the plosion spike of a plosive and the onset of the first glottal pulse of a following voiced continuant sound is a VOT.

1.3 *Specific characteristics of Indian speech sounds*

All Indian languages having natural languages share several features and sounds with the other languages of the world as one cannot expect a language or a group of languages entirely composed of speech sounds that are not found anywhere else. However, some amount of confusion may arise because of misleadingly similar or same terms to denote different features. For instance, the phonetic realization of 'alveolar' plosives of English is somewhat different from the phonetic realization of 'alveolar' plosives of Malayalam. In the following section, we will discuss the essential nature of Indian speech sounds that set them apart from the other languages of the world. A clear understanding of this is necessary for successful implementation of speech technology.

1.3a *Plosives*: Plosive is a manner of articulation. It is defined by the zero stricture which is its main component, followed by burst spike and then by a short stretch of friction noise which is considered as a part of the VOT lag. If a voiced continuant sound (such as a vowel, nasal, lateral) precedes the plosive, that sound may contain a final transition that indicates the place of articulation of the plosive. However, the burst and the friction spectra of the plosive and the transition out of the plosive into the following voiced continuant sound, if any, are strong indicators of its place of articulation.

There are some places of articulation that are typical of a good number of Indian languages. We will examine the nature of such places of articulation.

1.3b *Places of articulation: Dental-alveolar-retroflex*: Retroflex sounds are typical of a majority of Indian languages. Transitions leading into the stricture, burst and friction spectra as well as the transitions leading out of them differentiate retroflexes from alveolars and dentals. Retroflex sounds are present in all the Dravidian languages, Indo-Aryan languages (except Assamese), in some Tibeto-Burman languages, Austro-Asiatic languages. Alveolar plosives are characteristic of Malayalam and Assamese languages. Malayalam alveolar plosives have a good amount of frication at the time of release which sets them apart from their retroflex counterparts. This APS need to be carefully considered for technology applications.

1.3c *Laryngeal activity*: Here, we will examine the usage of both quantity and quality of pitch pulses in Indian languages.

Presence-absence of voicing. Presence or absence of voicing in a speech sound gives rise to voiced-voiceless distinction. This basic distinction that is found in all the languages of the world is employed in Indian languages to a great extent. All Indo-Aryan languages, and some Dravidian languages and Austro-asiatic languages have clear-cut contrast between voiced and voiceless plosives. Among Dravidian languages, Tamil and Malayalam treat voicing partially at allophonic level. In Tamil, among the words from native vocabulary (vocabulary that is not borrowed from other sources such as Indo-Aryan languages or English), voiced and voiceless plosives behave as allophones, i.e., sounds that are decided by the context of their occurrence in a word. A phoneme such as /p/ is realized as its voiced allophone [b] when it occurs between two voiced sounds (such as vowels, nasals, etc.). In other contexts (such as in word initial position, or when it is doubled), /p/ is realized as its voiceless allophone [p]. In borrowed words such as /bassu/ for English 'bus', the initial /b/ is realized as [b] – this behaviour does not follow the context-sensitive rule applied to native words. However, in the writing system of Tamil, /bassu/ will be written as {passu} since there is no separate letter to write /b/ in Tamil. Such a situation of rewriting {p} $\rightarrow$ /b/ $\rightarrow$ [b] has to be handled by dictionary lookup by the system.

Besides plosives, very few other manners of articulation show a contrast for voicing. Urdu has contrastive sets of voiceless and voiced fricatives. Tibeto-Burman languages such as Mizo show voiced-voiceless contrast among their continuants (such as nasals, laterals and trills). Voicelessness among plosives in Indian languages is generally characterized by the total absence of voicing during the stricture period. However, the acoustic correlates of voiceless continuants of Tibeto-Burman languages are more a matter of quantity of voicing and is discussed in the later sections.

Aspiration. Aspiration in the case of voiceless stops is quantified as the amount of VOT lag (the time delay between the burst of a plosive and the onset of the first pitch pulse of the following voiced continuant (such as a vowel), if any. It can be deduced that if no voiced continuant follows an aspirated plosive (in final utterance), then it is difficult to maintain clear aspiration. Many Indian languages which have contrastive aspiration among their plosives, do not have much of aspiration in the plosion that occurs in utterance final position. In careful speech, they are released with an extra 'puff of air' (which has the acoustic correlate of voiceless glottal friction). Thus, the quantity of VOT lag has to be finely tuned to get the appropriate result depending upon the position of the voiceless aspirated plosive in an utterance.

Quantity of voicing. Languages differ considerably in the 'amount' of voicing that is present during their closure stricture. English voiced plosives are supposed to be 'partially' voiced

as compared to 'fully' voiced plosives of some other European languages such as Spanish or French. Voiced plosives of Indian languages are typically 'fully' voiced. 'Partial' or 'full' voicing of a voiced plosive is a timing phenomenon. If pitch pulses are present throughout the stricture duration (i.e., if the voice bar is really voiced throughout), then that voicing is 'full' voicing. If pitch pulses are present during part of the stricture duration, then it is a case of partial voicing. In English, the contrast between voiced and voiceless plosives is conveyed mostly by the presence of aspiration at the time of release of voiceless plosives (versus absence of any release-aspiration in the case of voiced plosives). On the other hand, in an Indian language such as Urdu or Hindi, release aspiration does not play a key role in separating voiceless and voiced plosives. The reason is that these languages maintain a contrast between voiceless aspirated and unaspirated plosives, whereas English does not have such a contrast. Hindi and Urdu utilize the feature of aspiration to separate their voiceless aspirates from their voiceless unaspirates, whereas English uses the same feature of aspiration to separate its voiced from voiceless plosives.

Another usage of quantitative differences in voicing is found in the case of voiceless continuants in languages such as Mizo. It was found (Bhaskararao & Ladefoged 1992) that a voiced continuant such as a voiced nasal has modal voicing throughout its stricture. On the other hand, a corresponding voiceless nasal is characterized by the absence of pitch pulses during the initial three-thirds of the stricture period but with the rest of the period containing pitch pulses indicating an off-glide of voicing. This voiced offglide is essential in identifying the place of articulation of different voiceless nasals (Ladefoged 1971; Ohala 1975). Speech synthesis, when attempted for such languages need to take into consideration this feature of voiced offset among voiceless continuants.

Quality of voicing. Voiced sounds in all the Indian languages are of 'modal' variety. 'Modal voice' is produced by 'regular vibrations of the vocal folds at any frequency within the speaker's normal range' (Ladefoged & Maddieson 1996). In addition to voiceless and voiced states of the larynx, Gujarati uses the 'breathy voiced' type of laryngeal vibration which will have a direct impact on the nature of the pitch pulses that one obtains for these sounds. In the production of breathy voice 'vocal folds vibrating but without appreciable contact…' (Ladefoged & Maddieson 1996). All the vowels in Gujarati can be pronounced either with normal voicing or with 'breathy voice' giving rise to contrastive sets. Both type of vowels are phonemic and can separate one word from another. APSs for such breathy voiced pitch pulses have to be worked out.

Pitch variation. Pitch is the perceptual correlate of fundamental frequency. All natural languages use relative variations in pitch to bring out intonational differences. Intonational differences can signal syntactic differences (such as differences between interrogative and declarative sentences) or to bring out paralinguistic features such as emotional and attitudinal differences on the part of the speaker. Although it may be too early to model paralinguistic features of intonation at the current state of speech synthesis, intonational patterns have to be modelled for quality speech synthesis. In additional to the use of relative pitch variation for syntactic purpose, some languages use relative pitch variations to signal lexical differences. Such languages are called tone languages. Several of the languages of the Tibeto-Burman family within India are tone languages. The number of contrastive tones that these languages use may vary from a minimum of two tones to a maximum of five tones. For instance, Manipuri has a two-way contrast of tones whereas Mizo has a three-way contrast. While synthesizing tones for a tonal language, it should be remembered that it is the relative variation of pitch that defines the tonal distinction but not

any absolute values of fundamental frequency. As tones in a tonal language carry lexical load, the built-in lexicon of the language should carry tone markings and the synthesizer should be able to generate the appropriate tonal pattern for the given word. It should be remembered that all tonal languages have their intonation systems too. The lexically decided tonal pattern will interact with the syntactically controlled intonation pattern. This gives rise to 'allotonal' variations within the tones. Hence, the tonal patterns have to be modelled after examining their interaction with the intonation patterns.

Another unique case of contrastive pitch variation occurs in Panjabi. An underlying rewrite process of script letter sequences into phoneme sequences occurs here. In Panjabi the written sequences {b^hV, d^hV, $ḍ^hV$, j^hV, g^hV, hV} are realized phonemically as: /pV́, tV́, ṭV́, cV́, kV́, V́/. Here V́ stands for a vowel with a high tone. Hence, speech synthesis for Panjabi will have to handle the transfer of each of the five members of voiced aspirated plosive set to the corresponding member from the voiceless unaspirated plosive set. In addition, the corresponding vowel has to be assigned the appropriate fundamental frequency. Notice that the letter sequence {hV} is transformed into the phoneme /V́/. Dogri also has a similar set of rules but its writing has evolved using a separate 'accent mark' (svara chihna) to take care of this change.

Airflow direction and quality of voicing. A majority of the languages of the world use pulmonic egressive airstream (air pushed out of the lungs) for producing speech sounds. However, there are several languages that use other types of airstreams. Among Indian languages, Sindhi is the only language that uses voiced glottalic ingressive plosives (implosives) in a systematic way. For instance, the velar series of plosives in Sindhi consist of: /k/, /k^h/, /g/, /g^h/, and /ɠ/, the last being a voiced implosive. Although some rough estimates of the acoustics of implosive sounds are found for other languages such as Degema of Nigeria (Lindau 1984), we do not have clear-cut acoustic parameters for synthesizing the implosives of Sindhi. Although good aerodynamic readings of Sindhi implosives are available (Nihalani 1974), concrete acoustic parameters of these sounds are a desideratum for successful system implementation.

1.3d *Manners of articulation*: Among manners of articulation, trills, flaps and approximants of Indian languages need special consideration as sets of sounds produced with the manners in Indian languages have some inherent differences compared to those outside the subcontinent.

Trills. Most of the Indian languages have at least one distinct trill phoneme. The trills of Indian languages are 'typical trills' in the sense there is a periodic vibration (roughly around 30 Hz) at the stricture usually by the tip of the tongue. This vibration is different from the concurrent laryngeal vibration (that gives rise to the pitch pulses). The usual symbolization of a trill as [r] causes some amount of ambiguity as compared to the value of the rhotic [r] of English. In standard English, the sound represented by the letter {r} is never trilled. It is mostly realized as an approximant or a rhotic colour over the adjacent vowel. On the other hand, the sound represented by the letter {r} in all the Indian languages is a typical trill. Hence, trying to synthesize an Indian [r] using the parameters of English {r} will not produce satisfactory results at all. The Italian or Spanish trills have more similarity with the Indian trills. Another complexity has to be observed – in Indian languages, a trill phoneme /r/ generally has a tap allophone [ɾ] in word initial position and a trill allophone [r] with two or three tappings when it occurs between two vowels. When the trill phoneme is geminated as /rr/, then it is always trilled (not tapped). Malayalam has two contrasting trills, an advanced trill [r̟] as in /kar̟i/ 'soot' and a retracted trill [r̠] as in /kar̠i/ 'curry'. The advanced trill [r̟] has a 'higher locus for the second formant' and the retracted trill [r̠] has a

lower third formant (Ladefoged & Maddieson 1996). However, more precise APSs of these two sounds have to be worked out and tested through synthesized samples.

Secondary articulations. In sound produced with secondary articulation, there are two articulations simultaneously occurring at two places of articulation. 'Palatalization' is the main variety of secondary articulation that needs to be considered in the Indian scenario. Ladefoged & Maddieson (1996) describe palatalization as the 'superimposition of a raising of the front of the tongue toward a position similar to that for **i** on a primary gesture'. The only language in India that has systematic secondary articulation in its phonology is Kashmiri. Kashmiri is well-known for a large set of palatalized consonants that contrast with their non-palatalized (i.e., regular) counterparts. Although palatalization is 'more apparent at the release than at the formation of a primary constriction' (Ladefoged & Maddieson (1996), Kashmiri palatalization consistently begins before the formation of the primary stricture (Bhaskararao *et al* 2009). Russian is another language which is often used as a text-book example for palatalized consonants. However, we should not synthesize the Kashmiri palatalized sounds on the Russian model but evolve a more precise model that is based upon actual analysis of the sounds in the language. Modelling the palatalized consonants of Kashmiri taking into consideration this particular feature will be essential to produce a better synthesized output.

1.3e *Vowels*: As compared to consonants, vowels of Indian languages do not have that many significantly different features. Two major elements have to be considered for vowels of Indian languages.

Vowel quantity. A majority of Indian languages maintain a contrast between short and long vowels. As with tone and intonation, durational differences (in milliseconds) between a short vowel and its long counterpart is a relative matter. The overall tempo of an utterance determines the mean durations that a short or a long vowel will have in an utterance. On the other hand, a short vowel phoneme (say represented by /V/) and a its phonemically longer counterpart /Vː/ will have a few allophones each which vary in their relative durations. A short vowel might become extremely short – sometimes just having two pitch pulses – say, at 100 Hz fundamental, its length will be 20 msec. In a language such as Telugu, a long vowel phoneme may have an extra long allophone when it occurs in the grammatical position of a conjoiner: e.g., Telugu: /vaːḍ-uːad-iːvacceːru/ 'He and she came'. Here, /uː/ and /iː/ function as conjoiners (denoting 'and'). In this function, these two vowels have extra-long allophones (which may be tentatively represented as [uːː] and [iːː], respectively). This case also shows that parsed grammatical information should also be invoked during speech synthesis.

Vowel quality. Indian languages are supposed to have syllable-timed rhythm whereas English is a typical example for a language with stress-timed rhythm (Abercrombie 1967). As stress does not have any phonemic value in Indian languages, it does not control the quality or quantity of vowels in a word. Thus Indian languages do not exhibit drastic changes in the quantity or quality of a vowel which usually depends upon the syllabic stress. On the other hand, languages such as Telugu exhibit elaborate vowel harmony phenomenon. In Telugu, the quality of a vowel in a syllable is contextually decided by the quality of the vowel that occurs in the next syllable (table 2).

In each of the words (in both the columns), we have the same vowel phoneme in the first syllable in each of the five rows. They are /i/ /eː/ /a/ /u/ /o/, respectively, in rows 1 to 5. In

Table 2. Telugu vowel harmony.

	A			B		
	Meaning	Phonemic transcription	Phonetic transcription	Meaning	Phonemic transcription	Phonetic transcription
1	‘cat’	/pilli/	[pilli]	‘girl’	/pilla/	[pɪlla]
2	‘nail’	/meːku/	[meːku]	‘goat’	/meːka/	[mɛːka]
3	‘tie’	/kaṭṭu/	[kaṭṭu]	‘bundle’	/kaṭṭa/	[kɜʈʈa]
4	‘rust’	/tuppu/	[tuppu]	‘bush’	/tuppa/	[tʊppa]
5	‘a core’	/koːṭi/	[koːṭi]	‘fort’	/koːṭa/	[kɔːʈa]

Column A, the words have a close vowel in the second syllable (/i/ or /u/), whereas in Column B, the corresponding words have an opener vowel in the second syllable (/a/). The ‘closeness’ or ‘openness’ of a the vowel in the second syllable controls the ‘closeness’ or ‘openness’ of the allophone of the vowel in the first syllable. Thus in Column A, we get the ‘close’ allophones [i], [eː], [a], [u], [oː], whereas in Column B we get their ‘open’ counterparts [ɪ], [ɛː], [ɜ], [ʊ], [ɔː]. Selection of the appropriate allophone of the vowel is highly important for a clean synthesis of these vowels in Telugu.

2. Conclusions

We have demonstrated that converting an underlying text string of an Indian language to an appropriate speech output is succesfully achieved by paying sufficient attention to the intermediary stages out of which the *sandhi* process is a crucial step as these languages have rich and complex morphophonological processes. Some of the salient phonetic features of Indian languages are discussed systematically. Based upon these features, evolving precise acoustic-phonetic segment sets for different languages of the subcontinent requires detailed analysis-by-synthesis approach to these elements.

References

Abercrombie D 1967 *Elements of general phonetics* (Chicago: Aldine)

Bhaskararao P, Ladefoged P 1992 Two types of voiceless nasals. *J. Int. Phonetic Association* 21(2): 80–88

Bhaskararao P, Hassan Sheeba, Naikoo I A, Ganai P A, Wani N H, Ahmad T 2009 A phonetic study of Kashmiri palatalization. In: Minegishi M, *et al* (ed) *Field Research, Corpus Linguistics and Linguistic Informatics – Working Papers in Corpus-based Linguistics and Language Education - No 3*, Tokyo: Tokyo University of Foreign Studies, 1–17

Ladefoged P 1971 *Preliminaries to linguistic phonetics* (Chicago: Univ. of Chicago Press)

Ladefoged P, Maddieson I 1996 *The sounds of the world’s languages* (Oxford: Blackwell)

Lindau M 1984 Phonetic differences in glottalic consonants. *J. Phonetics* 12: 47–55

Nihalani P 1974 An aerodynamic study of stops in Sindhi. *Phonetica* 29: 193–224

Ohala, J 1975 Phonetic explanations for nasal sound patterns. In: Ferguson C A, Hyman L, Ohala J (eds) Nasalfest: Papers from a Symposium on Nasals and Nasalization. Stanford: Language Universals Project, 289–316

van Santen J P H 2005 Phonetic knowledge in text-to-speech synthesis. In: Barry W J, van Dommelen W A (eds) *The integration of phonetic knowledge in speech technology*, Dordrecht: Springer, 149–166