Detection of Phonemic Aspiration for Spoken Hindi Pronunciation Evaluation

Vaishali V. Patil[a]*, Preeti Rao[a],

*Corresponding author: prao@ee.iitb.ac.in, Tel: +91 9757154083
[a]Department of Electrical Engineering, Indian Institute of Technology Bombay, India.

Abstract

The computer-assisted learning of spoken language is closely tied to automatic speech recognition (ASR) technology which, as is well known, is challenging with non-native speech. By focusing on specific phonological differences between the target and source languages of non-native speakers, pronunciation assessment can be made more reliable. The four-way contrast of Hindi stops, where voicing and aspiration are phonemic for each of 5 distinct places-of-articulation, are typically challenging for a learner from a different native language group. The improper production of the aspiration contrast is thus often the salient cue to non-native accents of spoken Hindi. In this work, acoustic-phonetic features, motivated by an understanding of the production of the aspirated plosives, are evaluated for the classification of plosives along the aspiration dimension. Several new acoustic measures are proposed for the reliable detection of the aspiration contrast in unvoiced and voiced plosives. The acoustic–phonetic features are shown to perform well in the two-way classification task, and also appear robust to cross-language transfer where statistical models trained on Marathi speech were tested on native Hindi utterances. In experiments on native and non-native utterances of Hindi words by Tamil-L1 speakers, the acoustic-phonetic features clearly separate the non-native speakers from native on pronunciation quality of aspirated plosives. The acoustic-phonetic features also outperformed an ASR system based on more generic spectral features in terms of phone-level feedback that was consistent with human judgement.

Keywords

Computer-assisted language learning, pronunciation scoring, non-native accent, phonemic aspiration, acoustic-phonetic features

Detection of Phonemic Aspiration for Spoken Hindi Pronunciation Evaluation

1.0 Introduction

The computer-assisted learning of spoken language is closely tied to automatic speech recognition technology. The automatic assessment of a non-native learner based on carefully designed speaking tests coupled with focused phone-level feedback can potentially go a long way into expanding the reach of the language education industry. Fluency in spoken language by a language learner must be judged based on the achieved articulation and prosody in comparison with that of native speakers of the language (Celce Murcia and Goodwin, 1991). While acquiring intelligible speech is the prime requirement for a language learner, the absence of non-native accent, as indicated by segmental (phone articulation) and suprasegmental (prosody) differences from native speech, is desirable given the possible consequences of reduced processing ease by native listeners (Lev-Ari and Keysar, 2010). A key manifestation of foreign accent is the improper production of the target language (L2) phones.

Automatic speech recognition (ASR) systems would seem to provide the solution to automatic pronunciation error detection by the ability to decode speech into word and phone sequences and provide acoustic likelihood scores indicating the match with previously trained native speech models. However the challenges here are linked to the known deficiencies of state-of-the-art ASR systems where phone recognition accuracies are relatively low and an acceptable performance in practical tasks is achieved only through the constraints of a powerful language model that represents the vocabulary, grammar and typical usage of the language (Chelba et al., 2012). In an application such as pronunciation assessment, however, the use of lexical and higher-order context would actually obscure genuine pronunciation errors by the non-native learner by ignoring minor phone-level differences in favour of vocabulary and syntax based predictions of the text. Further, for better phone-level recognition accuracy, the acoustic models should ideally be trained on actual non-native speech so that the achievable phonetic realizations of target phones by the learner are taken into account. However, large enough non-native speech databases are not easily available (Carranza et al., 2014).

A powerful way to deal with the problem of poor phone recognition accuracies from ASR is to exploit any available knowledge about the type of pronunciation errors (Franco et al., 2012). It is observed, for instance, that the errors made by a non-native speaker learning a second language (L2) tend to be heavily influenced by her own native tongue (L1) in terms of both phonetic and phonological influences (Flege and Port, 1981; Bhela, 1999; Bhada, 2001; Best et al., 2001; Wiltshire and Harnsberger, 2006). The segmental errors arise from (1) the absence of certain L2 phones in the L1 leading to phone substitutions by available similar phones (e.g. “vet” instead of “wet”), and (2) phonotactic constraints of L1 leading to improper usage of phones in certain word-level contexts (e.g. “s-a-low” instead of “slow” as when L1 does not support consonant clusters). Knowledge of the common phone-level errors in the non-native speech can help to refine the search space, by restricting it to the expected native and non-native forms, in automatic speech decoding thus improving accuracy in phone mispronunciation detection. However, since the phone errors typically involve phone substitution by closely matching phones borrowed from the speaker’s L1 (Wiltshire and Harnsberger, 2006), the required phone discrimination from the acoustic signal is all the more challenging.

A widely used distance measure in pronunciation assessment involving phone articulation quality is a normalized “acoustic likelihood” score obtained within a statistical speech recognition framework (Strik et al., 2007; Franco et al., 2012; van Doremalen et al., 2013). An effective method to measure phone pronunciation quality has been the ratio of log-likelihoods computed for the correct and mispronounced version of the phone (Strik et al., 2009; Franco et al., 2012). The reliability of this measure depends on the acoustic features used to represent the achieved phonetic realization. Standard statistical model-based systems use Mel-frequency Cepstral Coefficients (MFCC), a compact representation of the short-time spectral envelope, across classes of speech sounds (Jurafsky and Martin, 2008). On the other hand, research by speech scientists over the years has suggested that acoustic-phonetic features obtained through an understanding of the specifics of the speech production can be usefully applied to discriminate phones that differ in a single articulatory attribute (Niyogi and Ramesh, 2003; Truong et al., 2004; Stouten and Martens, 2006, Strik et al., 2009). In the present work, we investigate this latter approach for the automatic assessment of pronunciation of a specific class of sounds, namely the plosives of Hindi. Hindi belongs to the Indo-Aryan language group, which is among the few language groups of the world where aspiration is a phonemic attribute in both unvoiced and voiced plosives. The improper production of the aspiration distinction is an important cue to non-native accents, in addition to vowel quality and intonation (Wiltshire and Harnsberger, 2006). Limiting the scoring to the distinctive acoustic aspects is expected to contribute to the robustness of the system by ignoring other natural variabilities of speech that are irrelevant to pronunciation accuracy. Further, this approach facilitates the exploitation of specific discriminatory acoustic features.

A four-way contrast of plosives is a phonological feature of many languages of the Indo-Aryan group including Hindi and Marathi (Masica, 1993; Bhaskararao, 2011). The four-way contrast of Hindi plosives where voicing and aspiration are distinctive for each place-of-articulation (PoA) are typically challenging for a learner from a different native language group where aspiration is not used to signal a phonological contrast. This may be expected based on the “feature hypothesis” (Flege and Port, 1981; McAllister et al., 2002) according to which L2 features not used to signal phonological contrast in L1 are difficult to perceive and produce for the L2 learner. Further, while aspiration is allophonic in unvoiced plosives in several languages, appearing, for example, in the word-initial context, voicing occurring concurrently with aspiration, giving rise to what are known as “breathy voiced” sounds, is relatively rare (Lisker and Abramson, 1964, Gordon and Ladefoged, 2001). In the present study, we consider Tamil-L1 learners of spoken Hindi. Tamil is a Dravidian group language whose phonology is devoid of phonemic aspiration. Hindi is the native tongue of 400 million people in India and Tamil, that of 70 million (Census of India, 2011). Hindi is the national language of India and together with English serves as a connecting language across the multilingual country. With widespread internal migration, the need for spoken language acquisition of the common languages is high. Knowledge of typical segmental errors can make for a more robust pronunciation assessment system that delivers focused feedback by (i) weighting specific segmental scores higher relative to other segments which may be influenced by more speaker-dependent characteristics, and (ii) exploiting acoustic features that are motivated by the specific dimension of the phonological contrast to reliably detect pronunciation errors.

A goal of the present study is to develop and evaluate an automatic method for the speaker-independent detection of aspiration in voiced and unvoiced plosives that can be used in a pronunciation scoring task for spoken Hindi. Studies on phonemic aspiration in the world’s languages have been largely confined to unvoiced consonants and voice onset time has been the chief distinctive feature explored (Lisker and Abramson, 1964; Cho and Ladefoged, 1999).In the present work, acoustic characteristics of both unvoiced and voiced aspirates are studied in order to identify features and feature extraction methods for a pronunciation assessment system based on the statistical modeling of distinctive acoustic features. The performance of the proposed system is eventually compared with that of a baseline system that uses the MFCC features available in a traditional ASR system using Hidden Markov Models (HMM) for phone recognition (Franco et al., 2012). In the next section, we discuss the phonology, production and acoustic characteristics of Hindi plosives forming the basis for a discussion of acoustic features to detect phonemic aspiration in speech. This is followed by a description of our native and non-native speech datasets, and the system framework used for pronunciation assessment. We next present acoustic feature extraction methods that are designed to maximize discrimination between aspirated and unaspirated phones in native speech. Descriptive statistics computed on our native Hindi word-initial plosives database are provided to obtain insights on the discrimination of the measures across the aspiration contrast and their possible dependence on factors such as place of articulation and speaker gender. Finally, experiments that serve toexamine the correlation between system predicted pronunciation errors and subjective judgments of pronunciation quality are presented in the context of rating non-native pronunciation and providing corrective feedback with respect to aspirated plosives.

2.0 Production and acoustic characteristics of the aspiration contrast

The plosives of Indo-Aryan languages such as Hindi share the production characteristics of plosive sounds of other languages including English in terms of a complete closure at the place of articulation giving way to a transient burst at release, followed by frication and aspiration. The stops and affricates evidence frication and aspiration to different extents. As shown in Table 1, for each of the five places of articulation (including one affricate place), the plosives are distinguished by four contrasting combinations of voicing and aspiration attributes, namely unvoiced aspirated, unvoiced unaspirated, voiced aspirated and voiced unaspirated. The aspirated and unaspirated plosives are distinguished primarily by the aspiration phase following the burst release. The aspiration phase occurs when there is no constriction of the vocal tract and the vocal folds are still being adducted to produce modal voicing at the onset of the following vowel in the consonant-vowel (CV) context. Aspiration is perceived as a release of breath accompanying the plosive.

Table 1 ***Plosives of Hindi and Tamil languages***

Language	**PoA of unvoiced and voiced plosives**				
	Labial	**Dental**	**Retroflex**	**Palatal**	**Velar**
Hindi	p p^{h}	t̪ t̪^{h}	ʈ ʈ^{h}	ʧ tʃ^{h}	k k^{h}
	b b^{h}	d̪ d̪^{h}	ɖ ɖ^{h}	ʤ ʤ^{h}	g g^{h}
Tamil	p (b)	t̪ (d̪)	ʈ (ɖ)	ʧ (ʤ)	k (g)

Ohala and Ohala (1972) studied Hindi stops to find that following the release of the aspirated stops, the absence of constriction in the oral cavity lowered the glottal resistance resulting in air rushing out in great volume. It has been observed that aspiration is accompanied by an increased glottal opening in many languages, including Hindi, as well as the presence of aspiration noise during the following vowel (Ridouane et al., 2011). It is well known that the varying extent of glottal adduction is associated with voice quality, where it is considered to be greatest for pressed voice, moderate for modal voice and least for breathy voice quality (Klatt and Klatt, 1990; Hanson, 1997). Thus there is seen to be a similarity between the movement of the articulatory organs associated with aspiration release and breathy voice. In the case of voiced aspirated plosives, the abduction starts halfway through the closure and reaches its maximum at the burst release (Ladefoged and Maddieson, 2005). Due to the overlap of the glottal gesture with the vowel, a strong presence of non-modal (breathy) voice quality extends into the vowel region (Dutta, 2007).

The aspiration feature in unvoiced stops has traditionally been associated with the timing of voicing onset relative to the burst release (Lisker and Abramson, 1964) with longer duration voicing onset time associated with aspirated stops. Measurements of the voicing onset time similar to those reported by Lisker and Abramson (1964) were noted by Bengueral and Bhatia (1980) but are indicated to be clearly insufficient in distinguishing Hindi stops across the four categories (two of voicing and two of aspiration).The voice onset time, or more accurately, the vowel onset time in the context of voiced stops, has a high variance within a voicing-manner class since it is influenced by the place of articulation, and also possibly the speaking rate (Samudravijaya, 2003). From the preceding discussion on the production of aspirated plosives, it would seem that possibly more robust acoustic cues to aspiration may arise from some of the other articulatory distinctions, viz. increased glottal opening, more gradual closure and breathy voice quality of the subsequent vowel. These aspects have been extensively studied in terms of the corresponding acoustic correlates of breathy voice quality in vowels where H1-H2 (amplitude of the first harmonic relative to the second) represents the glottal open quotient, and spectral tilt representing glottal closure rate (Hanson, 1997; Ishi, 2004).

The acoustics of phonetic distinctions involving aspiration have been previously studied for several languages with most focusing on the unvoiced consonants. Spectral tilt, in addition to voicing onset time, the traditionally used acoustic measure, has been suggested to distinguish Korean aspirated stops from unaspirated (tense) (Cho et al., 2002). While this showed good discrimination between aspirated and lax stops, it was less effective for the aspirated-tense case. Based on the phonological observation that aspiration is marked by breathy voice in the following vowel, Clements and Khatiwada (2007) investigated the acoustic distinction between aspirated and unaspirated Nepali affricates on a small set of speakers to find that the acoustic measures of breathiness or “superimposed aspiration” (the part of the aspiration that coexists with glottal pulsing, as defined by Mikuteit and Reetz (2007)), showed considerable variation across speakers. The Khoisan language family has a rich set of phonation types that includes aspirated consonants as well as breathy vowels (Traill, 1980). Observations of H1-H2 and Harmonics-to-Noise ratio (HNR) in vowels following aspirated click consonants are reported to be similar to those in breathy vowels (Miller, 2007).Voiced aspirated plosives, due to their rare occurrence in the world’s languages, have been the topic of fewer studies (Benguerал and Bhatia, 1980; Dixit, 1989; Scheifer, 1986; Davis, 1994; Dutta, 2007; Mikuteit and Reetz, 2007). Lisker and Abramson (1964) in their classic study found that while the word-initial stops of most languages are effectively separated by the voice onset time (the interval between the burst release and the onset of glottal vibration), this is insufficient for the voiced aspirates and voiced inaspirates of the two four-category languages, Hindi and Marathi.

Rami et al. (1999) investigated the four velar stop consonants in Gujarati (with the same phonology of stops as Hindi) and observed the voice onset time as a function of voicing and aspiration. While the voice onset time of unvoiced stop /k^h/ is reported to be significantly longer that of /k/, the voiced stops /g/ and /g^h/ show no statistical differences (Rami et al., 1999). In a previous study, Davis (1994) found that the lag-VOT (that is, the interval between burst onset and the onset of the following vowel) discriminated all four velar stops in Hindi word-initial utterances. Dutta (2007) conducted an acoustic-phonetic study of the four-way contrast of Hindi stops, with spectral analysis in the vowel region following the voiced aspirated stops indicating breathy characteristics over a substantial portion of the vowel. Mean values of H1-H2 in the initial part of the vowel were reported to be significantly higher in voiced aspirated stops over that of voiced unaspirated stops. Further, the spectral tilt as captured by H1-A2 (where A2 is amplitude of the highest harmonic in the second formant range) indicated more gradual closing of vocal folds in voiced aspirated stops compared to the unaspirated stops.

Word-initial unvoiced stops of Marathi were shown to better separate into aspirated and unaspirated classes when breathy vowel quality features were combined with the voicing onset time (Patil and Rao, 2011). Marathi is an Indo-Aryan language that shares the phonology of plosives with Hindi, with the added presence of an allophone for each of the unaspirated affricates in non-front vowel context. The vowel quality features used were the glottal open quotient (via H1-H2), spectral tilt and aperiodicity in the region following the manually labeled vowel onset. The same set of features was found to be inadequate for the classification of the voiced plosives. However augmenting the spectral tilt and noise features with different measurements of essentially the same articulatory attributes was shown to improve two-way classification accuracy on Marathi voiced plosives (Patil and Rao, 2013a), and is considered more extensively in the coming sections.

In summary, the literature reviewed suggests that phonemic aspiration is multidimensional in terms of articulation. Trade-offs in the extents of various attributes can be expected in the realization of the phonemic contrast in natural speech. Therefore multiple acoustic features must be considered for the reliable detection of aspiration.

3.0 Database and baseline system

For the development and evaluation of the pronunciation assessment system for aspirated plosives, suitable datasets of utterances by native and non-native speakers were created. Both languages, Hindi and Tamil, contain oral plosives (stops and affricates) of five places of articulation. However, voicing and aspiration are used contrastingly only in Hindi as depicted in Table 1.Tamil does not distinguish aspiration or even voicing; the plosives are voiceless and weakly aspirated in initial position (much like English), and voiced after nasals (Balasubramanian, 1975). We obtain speech recorded by native and Tamil-L1speakers of Hindi in the form of read-out words containing the target phones. Several meaningful Hindi words are available across vowel contexts for all the plosives in word-initial position.

We also need data for the training of acoustic models used in the statistical classifier. We explore the use of already available Marathi training data (from previous studies on Marathi plosives) to train acoustic models for Hindi plosives. Marathi, like Hindi, is an Indo-Aryan language and shares its plosive phone series listed in Table 1. Although specified in Table 3, the unvoiced, aspirated labial is rarely used (substituted instead by the fricative phone /f/) in both languages and hence is omitted in the database. With native Marathi speakers more easily available to us given our geography, it was more practical to use a Marathi speech dataset for training. This also facilitated the incorporation of distinct word lists in the train and test datasets, which is a more realistic scenario in practice. Of course, we need to be aware of the underlying assumption on the phonetic similarity of the plosives across the two languages. In a comparative study of voice onset time for Hindi and Marathi stops of same manner and place of articulation, Lisker and Abramson (1964) comment that "in general Marathi stops are phonetically similar to Hindi stops". Motivated by the phonological similarity of the two languages, recent work in statistical modeling for automatic speech recognition has attempted to overcome a shortage of Hindi training data by augmenting it with a larger Marathi speech corpus (Mohan et al. (2014)). They obtained mixed results where it helped to augment the corpus provided the Hindi acoustic models were weighted more relative to the Marathi models in the eventual system for Hindi limited vocabulary recognition. We present our own acoustic measurements on Hindi plosives in Sec. 4.4 together with observations of statistically significant differences with respect to measurements on the corresponding Marathi plosives in our dataset.

3.1 Training and testing datasets

The training database comprises Marathi spoken words sampled at 16 kHz where two distinct meaningful words with word-initial plosives and each of the 8 vowels of the language (/ə/, /a/, /i/, /I/, /u/, /U/, /e/, and /o/) (common to Hindi and Marathi) are formed and each word is uttered in two carrier sentence contexts (one sentence and one question) by 20 native speakers of Marathi (same numbers of male and female).

The testing datasets comprising Hindi words were recorded by 20 native Hindi speakers and 10 speakers of Tamil L1, all engineering graduate students at IIT Bombay in Mumbai (equally distributed across the two genders). The Tamil-L1 speakers had been exposed formally to Hindi reading and writing during their school and undergraduate years in their home state where they had very limited exposure to the spoken language. They were fluent in Tamil and, currently living in Mumbai, used some Hindi for day-to-day communication with the locals. They were all familiar with English and fluent in reading and writing, with English being the sole medium of instruction in higher education. The non-native speakers were not specifically assessed for Hindi proficiency. However, as described in Sec. 3.2, the non-native speaker utterances involving aspirated plosives were correctly labelled as "non-native" by each of two native Hindi listeners.

Table 2 ***Minimal pair words with word-initial plosive as articulated by native and non-native speakers***

Plosive	Word	Meaning	Native pronunciation	Non-native pronunciation
t̪	ताली	Clap	t̪alI	t̪alI
t̪	थाली	Plate	t̪^{h}alI	t̪alI
b	बाग	Orchard	bag	bag
b^{h}	भाग	Section	b^{h}ag	bag
tʃ	चोटी	Plait	tʃoʈI	tʃoʈI
tʃ^{h}	छोटी	Short/little	tʃ^{h}oʈI	tʃoʈI
ʤ	जुठा	Lipped food	ʤUʈ^{h}a	ʤUʈa
ʤ^{h}	झुठा	Liar	ʤ^{h}Uʈ^{h}a	ʤUʈa

The test dataset involved one meaningful word of Hindi corresponding to each plosive consonant and the above mentioned 8 vowel contexts embedded in 2 Hindi carrier phrases (one sentence and one question) read out by each speaker. The speakers were presented the list of written words in Hindi script along with its meaning in English to further rule out potential ambiguities between minimal pair words. Table 2 shows a few example words (that also happen to be minimal pairs) along with their typical pronunciations by native Hindi and Tamil speakers. We observe the absence of aspiration in the pronunciation of Tamil speakers. Voicing however is always realised correctly even though voicing is allophonic in Tamil plosives. This may be explained possibly by the speakers' strong familiarity with spoken English. Table 3 summarises the training and test datasets in terms of the number of test tokens of each category. The complete Hindi word list is provided in the Appendix.

Table 3 ***Description of word-initial plosives in the database. Number of distinct speakers in each linguistic category in parentheses. The Marathi native dataset was used as training data.***

Data sets **Plosive category**		**Marathi native (20)**	**Hindi native (20)**	**Hindi non-native (10)**
Unvoiced	Stops	4480	2240	1120
	Affricates	1920	640	320
Voiced	Stops	5120	2560	1280
	Affricates	1920	640	320

3.2 Subjective validation of test datasets

To confirm the assumed competence levels of the native and non-native speakers of our test dataset, a perception test was carried out with two native listeners who were asked to label a test speaker as native or non-native based on a set of 5 words uttered by the speaker. In the interest of limiting the testing duration, the test phones were restricted to the two nearest PoA corresponding to the two different manners (dental and palatal corresponding to a stop and an affricate respectively). Thus we had 8 unique plosive phones (4 voicing-manner x 2 place) each represented by 5 words (drawn from across vowel contexts) giving rise to a total of 8 sets per speaker. The native listeners were asked to identify whether a given plosive's 5-word set was uttered by a native speaker of Hindi or a non-native given the text transcription corresponding to each word set.

A total of 20 speakers' data (10 native and 10 non-native) was used in the test giving rise to 20x8 five-word sets (i.e. 800 stimuli but grouped into 160 sets for set-level labeling) to be labeled across the two listeners. The sets were presented in random order. The results indicated that the word sets of all 8 phones for all the 10 native speakers were correctly labeled as "native". On the other hand, on the 10 non-native speakers' data, all 4 unaspirated plosives' word sets were misclassified as "native" with just 3 exceptions out of the 40 sets. Of the 3 exceptions, 2 corresponded to the unvoiced affricate (/tʃ/) and one to the unvoiced stop (/t̪/). The non-natives' aspirated plosives were correctly labeled "non-native" as expected, except for 6 sets out of the 40, where it was found that the non-native speakers had indeed correctly articulated the voiced and unvoiced affricates. In summary, the difference in the articulation of aspirated plosives between native Hindi and Tamil speakers is clearly perceived by native listeners. The occasional confusion observed was restricted mainly to the articulation of the affricates. Further, the unaspirated plosives of native and non-native speech seem practically perceptually equivalent.

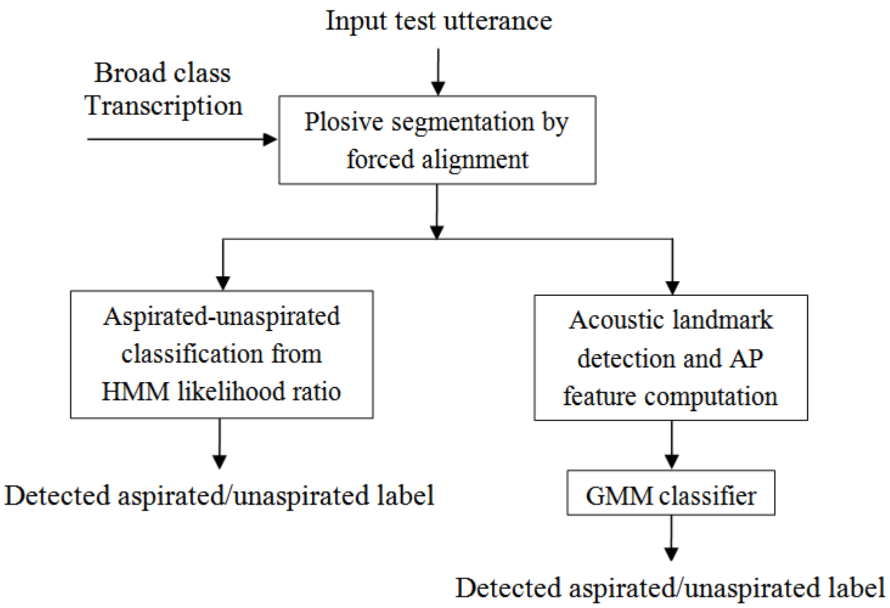
An interesting related question is whether non-native speakers who do not produce the aspiration contrast, perceive it. A listening test involving the minimal pair words of Table 2 by 4 different native speakers were presented in random order to three non-native listeners. The listeners were asked to write the words they perceived in Hindi script. It was observed that the listeners identified the unaspirated plosives correctly but had a recognition accuracy of only 50% on the aspirated plosives (confusing these with the corresponding unaspirated plosives).This could indicate that the production and perception of phonemic aspiration are related in adult language learners (Raphael et. al., 2007).

3.3 System framework

In the context of pronunciation assessment and feedback, it is necessary to evaluate each phone of the test utterance for correctness in articulation. The overall approach is to use the knowledge of the phonetic contrast to design acoustic features that capture the desired contrast (aspiration in this case) while ignoring most other variability in the speech signal across other phonetic attributes and across speakers. A set of such “distinctive” features is then represented further by statistical distributions computed from a “training” dataset that represents as much as possible of the expected signal variability for a given phone class. The resulting “acoustic model” can be used to achieve phone decoding in automatic speech recognition based on the match between the test utterance acoustics and the acoustic models of the hypothesized phone classes. Better features help to obtain more accurate acoustic models, in terms of better generalizability, from limited training data.

In the present work, we first implement a phone-level segmentation of the utterance. Next, the segment is scored for phonetic quality with respect to the acoustic model of the intended phone. A common framework that facilitates comparison of the proposed system with a conventional ASR based system is presented next. Figure 1 gives a flow chart of the overall processing steps involved in detecting the plosive segment within an input word utterance and classifying it as aspirated or unaspirated. First, a broad phone class segmentation is achieved by the forced alignment of the amplitude-normalized acoustic waveform of each utterance with the underlying broad phone-class transcription using an available state-of-the-art ASR system (Young et al., 2006) that employs MFCCs as the features. The standard 39 dim MFCC, delta and acceleration feature vector was computed at 10 ms intervals using a 25 ms Hamming analysis window. The broad classes are: vowels, sonorant consonants, unvoiced fricatives, unvoiced affricates, unvoiced stops, voiced affricates, voiced stops, silence and voice bar. Broad class acoustic models are more robust to training-testing mismatches as would be expected to occur in the pronunciation assessment scenario (Patil et al., 2009).The acoustic models are context independent, 3-state HMM with 8 Gaussian mixtures, diagonal covariance and are initialized assuming equal duration phones (Jurafsky and Martin, 2008; Young et al., 2006). We thus obtain the aligned segment corresponding to the plosive (known voicing category only and whether stop or affricate).This segment can next be processed for aspiration detection using each of the two methods being compared viz. the baseline MFCC system and the acoustic-phonetic (AP) features presented in the next section. In the case of the proposed AP feature based classification, a Gaussian Mixture Model (GMM) classifier (with 6 fullcovariance mixtures) is trained on the feature vectors of each class using the 20 speaker Marathi data as training data. This makes the systems comparable in terms of the underlying distributions assumed in the statistical modeling thus allowing us to focus purely on the relative performances of the acoustic features.

Figure 1 ***Automatic classification system framework***


Input test utterance
Broad class Transcription
Plosive segmentation by forced alignment
Aspirated-unaspirated classification from HMM likelihood ratio
Acoustic landmark detection and AP feature computation
Detected aspirated/unaspirated label
GMM classifier
Detected aspirated/unaspirated label

4.0 Feature extraction

As we have seen in the previous section, native listeners clearly can detect Tamil-L1 pronunciations of aspirated plosives. This motivates an investigation of acoustic-phonetic features that capture the aspiration contrast in unvoiced and voiced plosives. From previous work in different languages, reviewed in Sec. 2, distinctive acoustic cues to aspiration are found in the burst release and in the region immediately following the vowel onset point.

The spectrograms in Figures 2(a) and 2(b) show contrasting pairs of unvoiced dental stop CVs (/t̪u/ and /t̪ʰe/) and voiced velar stop CVs (/ɡa/ and /g^{h}e/) respectively extracted from words from native Marathi speech with the plosive in the word-initial position. The acoustic landmarks of onset of the plosive (burst onset time or BOT) and onset of the vowel (vowel onset point or VOP) are marked for reference. The onset of the burst is easily detected by the abrupt energy change that separates the closure from the burst release. The VOP is marked by the onset of periodicity (as also seen in the waveforms) in the unvoiced plosive CVs. The voiced plosives however require the detection of the vowel formant structure to indicate the VOP. The bottom panels of the figures show the aspirated plosive CVs. Superimposed aspiration (seen as the noise obscuring the vowel formant structure) is clearly observed over the early vowel region in aspirated plosives. Aspirated affricates show the same acoustic characteristic, as seen in Figures 3(a) and (b) of unvoiced affricate CVs (/tʃI/ and /tʃ^{h}u/) and voiced affricate CVs (/ʤa/ and /ʤ^{h}o/) respectively. Further, aspirated affricates show two phases in the region preceding the vowel onset, viz. frication and aspiration. The unaspirated affricates are characterized by a single frication phase. We hypothesize that, acoustic features sensitive to the aspiration contrast can be extracted from the burst release segments and post-vowel onset regions of Marathi plosives in CV context.

Figure 2(a) ***Word-initial CVs of unvoiced dental stop, /t̪ɯ/ (top panel) and /t̪ʰe/ (bottom panel)***

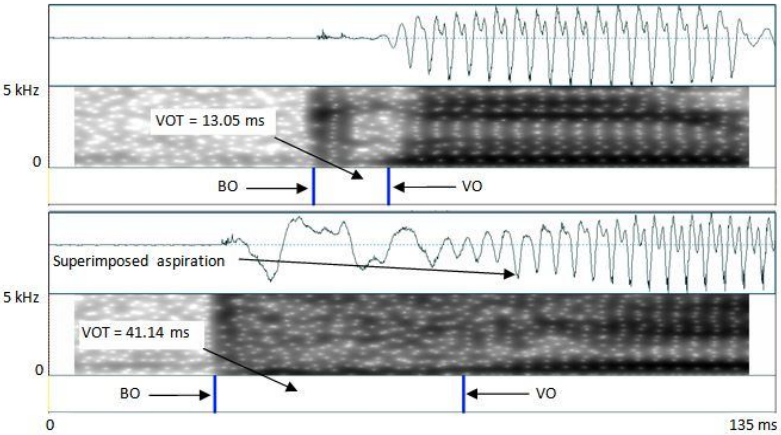


Figure 2(b) ***Word-initial CVs of voiced velar stop with acoustic landmarks, /ɡa/ (top panel) and /ɡ^{h}e/ (bottom panel)***

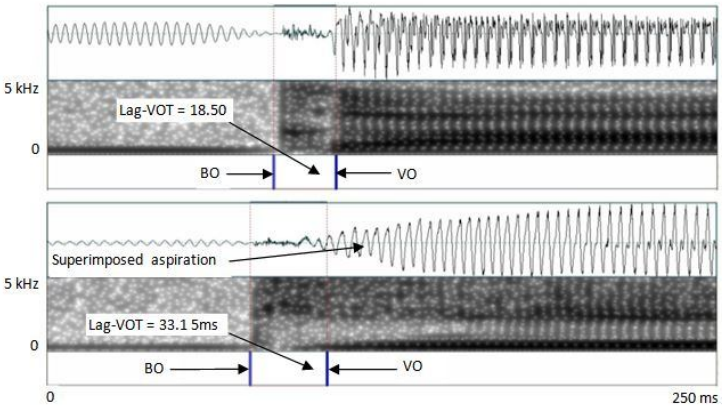


Figure 3(a) ***Word-initial CVs of unvoiced affricates with acoustic landmarks / tʃI/ (top panel) and / tʃ^{h}u/ (bottom panel)***

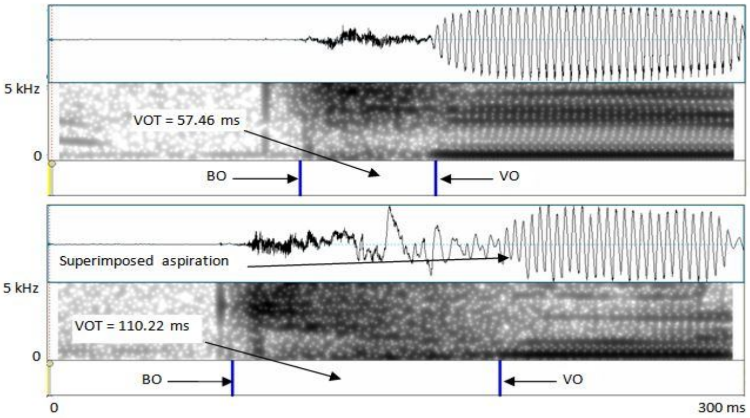
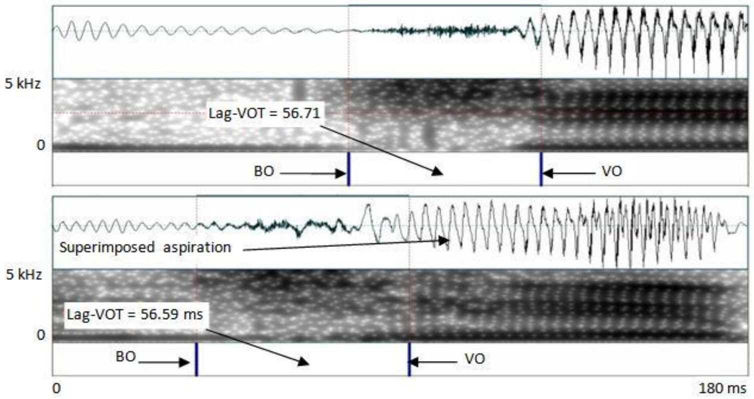


Figure 3(b) ***Word-initial CVs of voiced affricates with acoustic landmarks, /ʤa/ (top panel) and /ʤ^{h}o/ (bottom panel)***



We note that the detection and accurate localization of the acoustic landmarks of burst and vowel onset are required for reliable feature extraction. We therefore present a brief review of methods for landmark detection followed by a discussion of feature extraction.

4.1 Acoustic landmark detection

The extraction of AP features needs precise temporal locations of the landmarks corresponding to BOT and VOP in the CV region of each utterance. The segmentation achieved by the broad-class HMM recognition system of Section 3.3 is coarse and must be refined by a further localized search for the specific acoustic events in the vicinity of the detected coarse boundaries of the plosive. The release burst onset is detected by the largest peak in the rate-of-rise (ROR) contour of the smoothened energy in 3500-8000 Hz within a 40 ms vicinity of the coarse boundary with 1 ms resolution (Liu, 1996; Patil et al., 2009). This achieves burst localization to acceptable precision, i.e. a median localization error of 5 ms with respect to manually detected onsets. However, cues to vowel onset are dependent on the nature of the consonant and especially difficult for aspirated and voiced plosives. As the voicing dimension of the plosives is known from the broad class segmentation we employ different methods for vowel onset detection in the case of unvoiced and voiced plosives. The onset of periodicity is a prominent cue to vowel onset after an unvoiced plosive. Periodicity, as represented by the height of the autocorrelation function peak, computed from a sliding 25 ms window, is measured at 1 ms intervals throughout a region of 40 ms around the initial boundary. This vector of 40 periodicity values is input to a previously trained decision tree to detect the vowel onset. The decision tree is trained on the manually labeled vowel onsets of the Marathi database.

Periodicity is not a suitable cue to the vowel in the case of voiced plosives since the voice bar is likely to persist throughout the closure and burst regions. We exploit instead the rapid rise in the signal amplitude envelope in the low frequency band (50 Hz – 600 Hz) to detect the vowel onset in the vicinity of the initial coarse boundary. The Hilbert envelope of the filtered signal provides an accurate instantaneous estimate of the vowel amplitude (Prasanna and Yegnanarayana, 2005). Table 4 presents the VOP localization error with reference to manually marked vowel onsets for the different phone classes as computed over a total 13440 tokens across the 4 classes of the native Marathi dataset. We see that the localization is relatively good in the unvoiced plosives and less accurate in the voiced plosives. The feature extraction methods should ideally take into account possible landmark detection errors. Eventually these are expected to affect the phone classification accuracies.

Table 4 ***Percentage of vowel onset point (VOP) landmarks within a given tolerance duration with respect to manual annotation***

Deviation in ms	Obstruent class			
	Unvoiced stop	**Unvoiced affricates**	**Voiced stop**	**Voiced affricates**
± 5 ms	66.3	78.0	47.8	38.0
± 10 ms	82.3	92.1	61.9	57.3
± 15 ms	88.4	95.8	69.4	72.4
± 20 ms	93.1	97.5	74.4	81.9

4.2 Feature implementation

The voice onset time has been widely used to discriminate unvoiced aspirated plosives from unvoiced unaspirated plosives in English where the former appear in word-initial context as allophones for voiced plosives (Lisker and Abramson, 1964). It was shown that including the breathy voice quality features of spectral tilt (A1-A3, where A1-A3 correspond to the highest harmonic amplitudes in the first and third formant regions respectively) and noise (in terms of the signal to noise ratio or SNR) computed in the region immediately following the manually labeled vowel onset, additionally, improves the classification performance for unvoiced Marathi stops (Patil and Rao, 2011).In the case of voiced stops, Lisker and Abramson (1964) observed prominent overlap in the distributions of voice onset time for the aspirated and unaspirated classes for each place of articulation. A related duration feature, the vowel onset time or lag-VOT, defined for voiced stops as the interval between the burst onset and vowel onset, showed an aspiration detection performance barely above chance (Patil and Rao, 2011). Including the A1-A3 and SNR features improved this. It was later demonstrated that a performance more comparable to that on unvoiced stops was obtained only after including further supplementary features (as presented later in Table 5) that capture essentially the same underlying attributes viz., spectral tilt and noise (Patil and Rao, 2013a). Additionally, a pre-onset energy distribution feature computed over the burst region before the VOP was demonstrated to improve the otherwise relatively low performance of voiced affricates. The features developed for the present pronunciation scoring task build upon this previously published work on acoustic-phonetic features for aspiration detection in plosives. Further, automatic landmark detection is used across the phone classes.

Table 5 ***Acoustic-phonetic features used in aspiration classification for the different phone classes***

Class of plosives	AP features
Unvoiced stops	lag-VOT, H1-H2, A1-A3, SNR
Unvoiced affricates	lag-VOT, H1-H2, A1-A3, SNR, pre-onset energy ratio
Voiced stops	lag-VOT, H1-H2, A1-A3, SNR, F1F3-sync, Low-band-slope, B3-band energy
Voiced affricates	lag-VOT, H1-H2, A1-A3, SNR, F1F3-sync, Low-band-slope, B3-band energy, pre-onset energy ratio

From the discussion in Sec. 2 we note that the potential distinguishing properties of aspiration in voiced plosives are the vowel onset time, glottal OQ (captured by H1-H2), spectral tilt and aspiration noise. The vowel onset time is easily obtained as the time interval between burst onset and vowel onset, both landmarks detected automatically by the methods presented earlier. All the other signal features involve spectrum measurements on the speech signal post the vowel onset and are made on 25 ms Hamming windowed signal segments. To increase reliability, the signal measurements are averaged over 5 windows spaced at 1 ms hop intervals in the region of interest. The H1-H2 is the ratio of the amplitudes of first two harmonic peaks detected via local maxima in the DFT magnitude spectrum. (Preliminary experiments with spectrally corrected H1-H2, as prescribed by Hanson (1997), provided no advantage likely due to sensitivity to errors associated with automatic formant estimation.) The spectral tilt and aspiration noise too can be measured from the signal in different ways giving rise to the different acoustic features as presented next.

Spectral tilt has been estimated by a number of different acoustic parameters in the context of voice quality measurements. These include H1-A3 (Klatt and Klatt, 1990; Hanson, 1997), A1-A3 by Ishi (2004) where fixed bands around average first and third formant regions are used, and H1-A2 by Cho et. al. (2002). All these measurements capture the rolling off of the spectrum from the low frequency band to the higher formant region. We consider fixed bands around average formants like Ishi(2004) to obtain A1-A3 as the difference between the strongest harmonic components, one each in the range of 100 to 1000 Hz (F1 band) and 1800 to 4000 Hz (F3 band). The F2 region energy, used as an indicator of aspiration in unvoiced stops by Cho et. al. (2002), is also captured by de Krom’s (1994) breathiness feature in the form of the spectral slope computed as the difference in band energies of F2 band (400-2000 Hz) and the first harmonic region (60-400 Hz). We term this “low band slope”. Additionally, a normalized “B3 band energy” is included where B3 is the band (2000-5000 Hz) which takes on low values at increased tilt. The multiple distinct measures of the spectral tilt attribute help improve the representation of voiced aspirated plosives over that obtained by A1-A3 alone (Patil and Rao, 2013a; Patil, 2014).

Aspiration noise is the component of the vowel signal corresponding to the vocal tract filtered noise that accompanies the glottal source signal. A cepstrally liftered noise floor, corresponding to between-harmonics spectral power, is obtained using the method of Murphy and Akande (2007). The SNR feature is the ratio of speech signal power to this estimated aspiration noise power. Another acoustic property of breathy voice quality is the lowered correlation between signal components in low and high frequency bands arising from the dominance of aspiration noise in the high frequency region. This is captured by “F1-F3 sync”, a feature proposed by Ishi (2004) who employed the fixed frequency bands mentioned earlier in the context of A1-A3. Based on observations by Klatt and Klatt (1990) regarding the appearance of aspiration noise in vowel specific formant regions, we found it more useful however to restrict the bands to a width of 600 Hz around automatically detected first and third formants regions from LP analysis of the windowed data. The measure represents the correlation of the amplitude envelopes of the two band-pass filtered signals over a 25 ms region centered at a specific time instant beyond the vowel onset.

Further, in the case of affricates, aspiration noise is also conspicuous in the burst region as seen in Figure 3. Relative to frication noise, the presence of aspiration noise is detected by an increase in low frequency energy. We define a feature, the “pre-onset energy ratio”, to capture this. It is the ratio of the energy in frequency band 3000-7000 Hz to that in the 60-3000 Hz band. It is computed at 1 ms intervals with a 6 ms data window over the region 10 to 20 ms prior to the vowel onset. When aspiration noise appears in the otherwise purely fricated region, indicating an aspirated affricate, the pre-onset energy ratio is expected to drop. Table 5 shows the selected set of AP features for the different phone classes. The reader is reminded that the vowel onset time is identical to the voice onset time (VOT) in unvoiced plosives. In the case of voiced plosives, the former is positive valued (unlike the voice onset time) and is also known as the after-closure time or (positive) VOT (Mikuteit and Reetz, 2007). We will refer to the vowel onset time as "lag-VOT" across stop categories, noting that it is identical to VOT in the case of unvoiced plosives.

4.3 Selection of region of analysis

Since the breathiness of the vowel arises from the co-articulation with the preceding aspirated stop, it is important to select the analysis region suitably for the breathy voice features. Also different languages have been observed to have different durations of breathiness extending into the vowel (Ladefoged and Maddieson, 2005). Therefore we test the discriminability of the various features evaluated for different analysis interval locations ranging from vowel onset to about 30 ms into the vowel. Variation in the extent of aspiration noise beyond vowel onset can be observed through the waveforms of aspirated unvoiced plosives and voiced plosives as is seen in the waveforms of the bottom panels in Figure 2(a) and 2(b) which correspond to the aspirated category of unvoiced and voiced stops respectively. The aspiration noise is seen to be restricted over a smaller time interval beyond the VOP in the case of the unvoiced stop compared to that of the voiced stop where it extends further into the vowel.

The ability of a particular feature to differentiate aspirated-unaspirated classes can be captured by histogram based separability measures. One of these is the reciprocal of the average classification error estimate obtained from the feature distributions for each class (Theodoros, 2008).We compute the average separability across the features of a given category measured from the overlap of the distributions of the aspirated and unaspirated classes for each voicing category separately. The two feature sets tested for variation in separability with analysis region are: spectral shape (H1-H2, spectral tilt), and aspiration noise (SNR, and F1-F3 sync added in the case of voiced plosives).

Figure 4 shows the variation of the average separability of each feature set as a function of distance from vowel onset in case of unvoiced stops. It is seen that the computed average separability starts to decrease in value beyond 8 ms from vowel onset. (The slightly higher or equivalent separability observed at 3ms from the vowel onset is not completely reliable as the analysis region in this case extends into the burst region precedingthe vowel onset.) Accordingly in case of unvoiced plosives the analysis interval for both the feature sets is selected to be around the instant of the global maximum in separability, i.e. 8 ms from vowel onset.

Figure 5 shows the corresponding variation of the average separability as a function of distance from vowel onset for voiced stops. It is observed that the spectral shape based differences are strongest near vowel onset. On the other hand, Figure 5 indicates that noise measures are most discriminative further removed from the vowel onset. Accordingly the analysis interval for the spectral shape features is selected to be around the instant 13 ms from vowel onset, and that for the noise features to be at 23 ms in case of voiced plosives.

Figure 4 ***Variation in the average separability of aspirated-unaspirated classes of unvoiced stops as a function of distance from vowel onset***

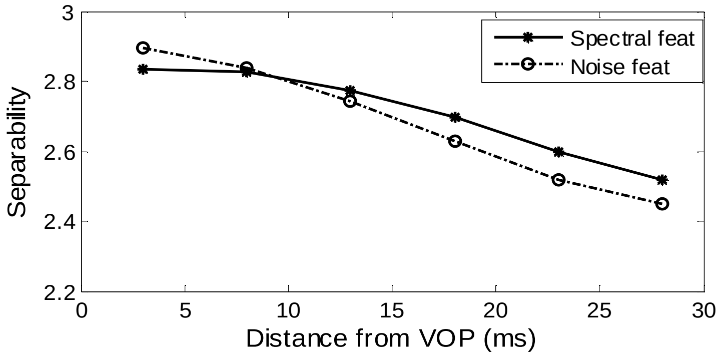
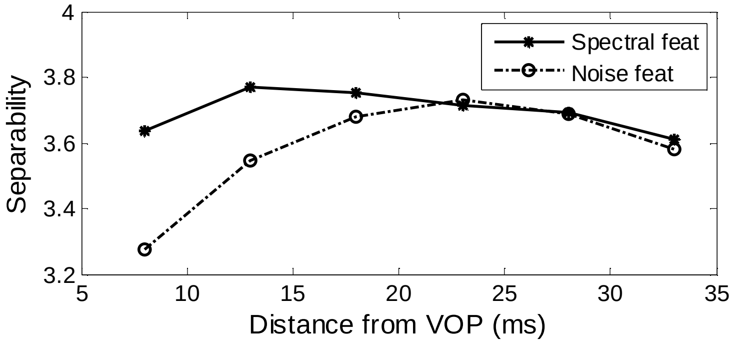


Figure 5 ***Variation in the average separability of aspirated-unaspirated classes of voiced stops as a function of distance from vowel onset***



4.4 Acoustic measurements

We present descriptive statistics (mean and standard deviation) for some of the acoustic measures used in this work for each of the Hindi plosives computed from the 20 native speakers' data in Table 3. We pick the four features (lag-VOT, H1-H2, A1-A3, SNR) that are common across the plosive classes. Mean and standard deviation were obtained separately for each phone class specified by its voicing, aspiration and place of articulation. The feature values, computed using the automatically detected landmarks, are presented in Table 6 separately for unvoiced and voiced plosives for the different PoA (after averaging across vowel contexts). We observe that the measures do indeed show numerical differences between unaspirated and aspirated categories for every plosive class. The lag-VOT means capture the aspiration distinction in voiced plosives for each place of articulation, in line with the observations of Davis (1994) on Hindi velars. While the lag-VOT means are also dependent on place of articulation, the remaining three (measures related to breathy voice quality in the following vowel) appear to be prominently influenced by the aspiration attribute only.

Table 6 (a) ***Mean and standard deviation (in parentheses) of various acoustic measures for Hindi unvoiced plosives averaged across speakers and vowel contexts***

Features	Place of articulation							
	Dental		Retroflex		Palatal (Affricate)		Velar	
	Unasp	Asp	Unasp	Asp	Unasp	Asp	Unasp	Asp
VOT	25	74	19	65	55	90	40	90
(ms)	(10)	(28)	(8)	(30)	(27)	(49)	(16)	(29)
H1-H2	4.9	12.3	4.2	13.8	6.9	11.6	7.0	11.8
(dB)	(7.8)	(6.0)	(7.8)	(6.3)	(8.6)	(6.7)	(9.2)	(6.2)
A1-A3	24.5	34.0	23.5	33.9	23.4	28.8	33.8	34.8
(dB)	(9.67)	(12.4)	(11.0)	(12.8)	(8.0)	(17.3)	(17.7)	(13.2)
SNR	-15.2	-21.5	-14.5	-22.2	-15.8	-20.56	-18.9	-21.1
(dB)	(5.2)	(5.3)	(6.1)	(5.0)	(4.3)	(5.6)	(6.1)	(4.4)

Table 6 (b) ***Mean and standard deviation (in parentheses) of various acoustic measures for Hindi voiced plosives averaged across speakers and vowel contexts***

Features	Place of articulation									
	Labial		Dental		Retroflex		Palatal (Affricate)		Velar	
	Unasp	Asp	Unasp	Asp	Unasp	Asp	Unasp	Asp	Unasp	Asp
Lag-VOT	12	26	20	34	12	24	64	99	38	59
(ms)	(10)	(30)	(8)	(26)	(5)	(22)	(23)	(39)	(20)	(30)
H1-H2	2.3	12.2	2.3	12.4	1.6	12.1	4.2	11.8	3.1	11.4
(dB)	(5.6)	(6.7)	(5.8)	(5.8)	(5.4)	(6.5)	(7.2)	(5.7)	(7.4)	(6.0)
A1-A3	31.0	36.7	25.9	34.0	25.9	33.1	24.4	31.6	30.1	36.2
(dB)	(14.2)	(11.5)	(10.6)	(10.3)	(11.1)	(10.1)	(8.2)	(10.8)	(15.5)	(13.2)
SNR	-15.0	-21.9	-15.0	-20.1	-14.3	-20.2	-14.8	-18.4	-16.4	-20.7
(dB)	(4.9)	(4.8)	(4.5)	(4.1)	(4.2)	(4.4)	(4.0)	(4.9)	(5.2)	(4.9)

Table 7 (a) ***Mean and standard deviation (in parentheses) of various acoustic measures for Marathi unvoiced plosives averaged across speakers and vowel contexts. (with * indicating statistically significant differences from Hindi)***

Features	Place of articulation							
	Dental		Retroflex		Palatal (Affricate)		Velar	
	Unasp	Asp	Unasp	Asp	Unasp	Asp	Unasp	Asp
VOT	17 *	56 *	12 *	47 *	62	89	30 *	77 *
(ms)	(8)	(21)	(7)	(21)	(21)	(26)	(21)	(24)
H1-H2	7.4 *	12.2	6.1	11.5 *	9.0	12.2	9.1	11.7
(dB)	(9.4)	(6.6)	(9.0)	(6.5)	(9.2)	(7.1)	(9.4)	(6.6)
A1-A3	29.5 *	36.9 *	28.4 *	36.0	30.0 *	34.6 *	34.7 *	36.6
(dB)	(10.1)	(9.4)	(10.6)	(8.9)	(7.7)	(9.6)	(12.5)	(10.6)
SNR	10.5 *	17.0 *	10.4 *	17.2 *	11.4 *	14.1 *	15.1 *	15.4 *
(dB)	(5.8)	(5.4)	(5.6)	(5.5)	(4.5)	(5.0)	(5.5)	(5.0)

Table 7 (b) ***Mean and standard deviation (in parentheses) of various acoustic measures for Marathi voiced plosives averaged across speakers and vowel contexts. (with * indicating statistically significant differences from Hindi)***

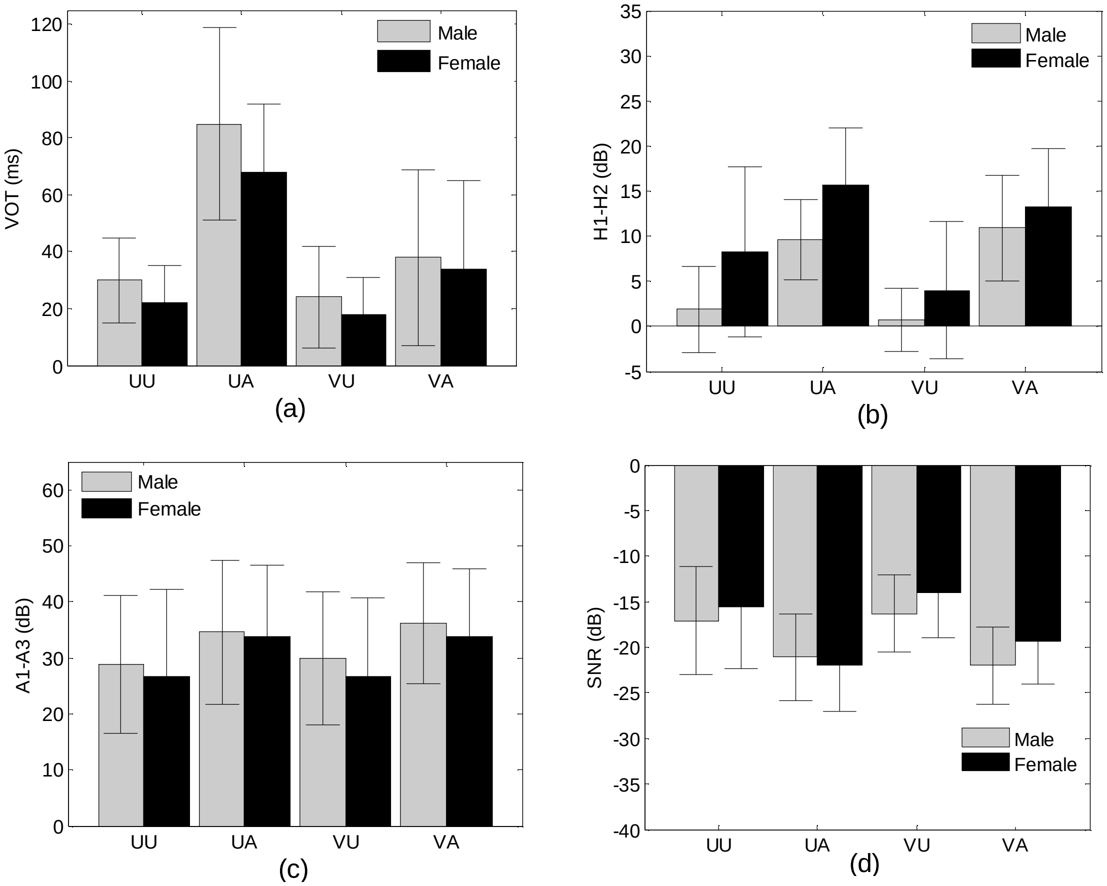
Features	**Place of articulation**									
	Labial		**Dental**		**Retroflex**		**Palatal (Affricate)**		**Velar**	
	Unasp	**Asp**	**Unasp**	**Asp**	**Unasp**	**Asp**	**Unasp**	**Asp**	**Unasp**	**Asp**
Lag-VOT (ms)	11 (9)	15 * (18)	14 * (8)	18 * (13)	9 * (4)	16 * (17)	63 (28)	65 * (24)	36 (19)	44 * (25)
H1-H2 (dB)	3.6 (7.3)	11.9 (5.8)	4.2 * (7.3)	11.9 (5.1)	3.1 (7.2)	12.2 (5.2)	7.5 * (7.9)	11.6 (5.4)	6.8 * (8.9)	12.1 (5.5)
A1-A3 (dB)	27.2 * (11.6)	36.9 (9.8)	24.8 (9.2)	35.9 (8.9)	22.9 * (9.3)	34.7 (8.8)	25.7 (7.3)	32.7 (6.7)	32.4 (14.3)	38.4 (10.9)
SNR (dB)	12.5 * (5.1)	19.3 * (4.4)	11.8 * (4.9)	17.6 * (3.7)	10.1 * (4.6)	17.6 * (4.3)	13.6 * (4.7)	16.3 * (4.0)	17.9 * (5.6)	19.7 (4.4)

In order to study possible phonetic differences between the corresponding plosives of Hindi and Marathi in our dataset, Table 7 provides the means and standard deviations for the Marathi plosives computed on the 20 Marathi speakers' dataset (using one instance of each word to keep the total number of utterances equal to that of Hindi).

We note that the standard deviations reflect speaker dependencies but also the errors in feature implementation that might arise from the automatic landmark detection errors. Statistically significant differences (p<0.01) in the same acoustic measure for the corresponding Hindi plosive as determined from a 2-sample t-test with equal variances are indicated with an asterisk. We note that the H1-H2 and A1-A3 distributions are reported similar for most plosive pairs while the VOT (lag-VOT) and SNR show statistically significant differences for the dataset at hand. It may be noted that there are no previous available studies comparing the phonetic aspects of Hindi and Marathi plosives.

While both genders are represented equally in our datasets, it is of interest to determine whether our acoustic measures differ systematically with gender. Previous literature has studied the variation of VOT with gender (Morris et al., 2008; Oh, 2011) and reported opposing trends in the different languages considered. On the other hand, the dependence of voice quality measures on gender is well documented (Hanson, 1997; Hanson et al., 2001). To facilitate comparisons, Figure 6 presents box plots (mean and standard deviation indicated) of the distributions for male and female speakers in our native Hindi dataset for each stop class and each of four measures. Overall, we note that the lag-VOT serves the aspiration contrast best for the unvoiced stops, while the H1-H2 and SNR measures are most effective for the voiced stops where the lag-VOT is not so discriminative across aspiration states. We observe that the lag-VOT is consistently lower for females compared to males across voicing and aspiration states. H1-H2 is higher in females as expected (Hanson, 1997). The A1-A3 distribution is more comparable across genders. The SNR distribution indicates that the vowel region following the plosive is most breathy for the voiced aspirates, as expected, but that the breathiness in higher for males compared to females. The affricates were observed to show the same trends, not reported due to the possible unreliability arising from the low number of affricates available for the averages per gender-class.

Figure 6 ***Mean and standard deviation values of (a) lag-VOT (ms), (b) H1-H2 (dB), (c) A1-A3 (dB) and (d) SNR (dB), across gender for the 4 stop classes. (Unvoiced-Unaspirated (UU), Unvoiced-Aspirated (UA), Voiced-Unaspirated (VU), Voiced-Aspirated (VA))***



In summary, the acoustic features proposed for the aspiration contrast in plosives show the potential to discriminate aspirated plosives from unaspirated. The observed dependence of the measures on place of articulation and gender is expected to be compensated for to an extent by the statistical modeling used in automatic speech recognition where the underlying distributions are typically assumed to be multi-modal, e.g. Gaussian mixture models (Jurafsky and Martin, 2008).

5.0 Experiments and results

Keeping in mind the end application for this work, it is of interest to evaluate the accuracy and robustness of the acoustic-phonetic features for aspiration detection. In order to focus on the aspiration distinction, a purely two-way (aspirated-unaspirated) classification framework is employed, as presented in Section 3.3, where the classifier makes a choice between options that differ only in the aspiration dimension of the word-initial plosive. The AP features' performance is compared with that of the MFCC features, both systems as depicted in Figure 1, in the two-way classification of native speech plosives in two distinct contexts: same language training-testing and cross-language training-testing. The evaluation is next extended to the pronunciation assessment context by employing the classifiers to detect phonemic aspiration in Hindi utterances by native and non-native speakers.

5.1 Classification of native speech

We present experiments involving the two-way (aspirated-unaspirated) classification of word-initial plosives recorded by native speakers of Marathi. The Marathi-trained acoustic models are also evaluated for the classification of native Hindi utterances. This is the context of the pronunciation assessment task, where we train the acoustic models for the aspirated and unaspirated plosives on an already available Marathi dataset. Given the observed dependence of the acoustics of aspiration on the voicing and manner (stop, affricate), the acoustic-phonetic features of Table 5 are evaluated for classification accuracy separately for each of the 4 plosive classes, with all stops for each voicing and aspiration combined across the places of articulation.

A 20-fold cross-validation (leave-one-speaker-out) classification experiment was carried out on the Marathi dataset to obtain the results shown in Table 8. Also shown in Table 8 are the performances with systems that were trained on the full Marathi dataset of 20 speakers, and tested on the Hindi dataset of 20 native speakers. The native-Hindi test data provides a more realistic evaluation of the acoustic models since the test set is different from the train set not only in the speakers but also in the words. In the more constrained Marathi train-test experiment, we observe that the MFCC features achieve accuracy comparable to the AP features for unvoiced stops and a numerically lower performance with voiced stops. However for unvoiced as well as voiced affricates the MFCC-HMM system provides the higher accuracy. A closer analysis revealed that the inferior performance of the AP features on affricates was due to the greater dependence of the features (lag-VOT and pre-onset energy) on the accurate detection of the vowel onset point. A more prominent difference between the MFCC and AP features appears in the cross-language accuracies. The AP features show a similar performance on Marathi and Hindi where there is a train-test difference in the uttered words due to the lexical differences. On the other hand, the MFCC features' performance drops markedly. The AP features are designed to target the aspiration distinction and clearly generalize better. The MFCC models seem to over fit to language-dependent phenomena, a possible explanation for the steep drop in performance across test datasets. This strength of acoustic-phonetic feature design has been observed in previous work related to the detection of different phonetic distinctions (Niyogi and Ramesh, 2003). The particularly large drop in MFCC performance on unvoiced plosives could be due to the allophonic utterance of the Marathi dental/alveolar unvoiced affricate '/ts/'which is not present in Hindi. This unaspirated allophone is acoustically similar to the aspirated affricate leading to poorly trained models for the unvoiced affricates.

Table 8 ***Classification accuracies with Marathi-trained acoustic models for aspiration state of plosives of the Marathi and native Hindi datasets in Table 3***

Class	**% accuracies in AP-GMM**		**% accuracies in MFCC-HMM**	
	Marathi	**Hindi native**	**Marathi**	**Hindi native**
Unvoiced stops	90.5	90.2	90.3	76.4
Unvoiced affricates	83.2	80.9	89.3	48.8
Voiced stops	85.1	85.0	80.8	77.8
Voiced affricates	79.2	79.2	83.0	81.7

5.2 Evaluation of AP-GMM and MFCC-HMM system for pronunciation assessment

We next present an evaluation of the acoustic-phonetic system for pronunciation assessment and compare it with the baseline MFCC-HMM system on the same tasks. Both systems are trained on the full Marathi dataset of 20 speakers. The tasks are designed to demonstrate the suitability of the systems for overall rating of the pronunciation quality of phonemic aspiration of a non-native learner and the accuracy of phone level feedback (Patil and Rao, 2013(b)).The test database is as described in Section 2.1, where each of the 20 native and 10 non-native speakers read out 304 words each embedded in a carrier phrase. The automatic systems are evaluated on this dataset for the (i) detection of non-native accent with respect to ground-truth about the speaker's L1, (ii) correlation with native listeners' judgments of phone realization using the methodology presented next.

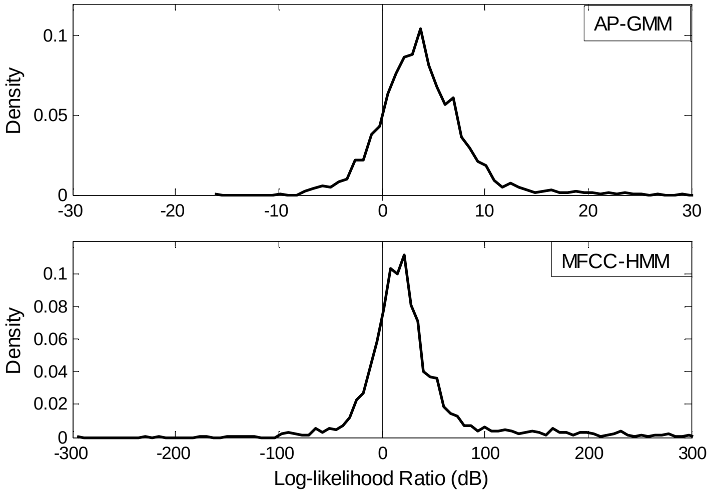
A pronunciation assessment system that provides focused feedback in terms of flagging poorly articulated phones can be very useful in computer-aided language learning. In the classifier framework, the normalized likelihood of the target model, given the observation, provides a measure of the match between the test utterance and the native-trained model (Witt and Young, 1997). We use the log of ratio of likelihoods of the target over that of the opposite models as an estimate of the "goodness of pronunciation" of an uttered phone (Niyogi and Ramesh, 2003).

$$d(x) = \log\left(\frac{L(x|\wedge 1)}{L(x|\wedge 2)}\right) \dots\dots\dots \quad (1)$$

where$L(x|\wedge 1)$ is the likelihood of an arbitrary point x in the feature space for the model of class 1 (likewise $L(x|\wedge 2)$ for class 2). Class 1 represents the target class while class 2 the opposite class. That is, if the target is the aspirated phone, Class 1 would correspond to "aspirated" and Class 2 to "unaspirated".

A ratio much greater than 1.0 would indicate native-like articulation of the target while a ratio much less than 1.0 would indicate non-native-like articulation. This is illustrated by Figure 7 which shows the distribution of the log likelihood ratios (in dB) over the Hindi native dataset (20 speakers) for voiced stops for each of the AP and MFCC systems. As expected, the native utterances lie mostly to the right of the 0 dB log-likelihood point. Similar plots were observed for the remaining plosive classes.

Figure 7 ***Distribution of log-likelihood-ratio from AP-GMM and MFCC-HMM systems over native data set for voiced stops.***



5.2.1 Detection of non-native accent

Each test word is automatically segmented and the classifier makes a two-way forced choice between unaspirated and aspirated plosive classes for each test CV segment. For each speaker and plosive class, we compute the percentage of instances that the target is correctly achieved (i.e. the classifier output matches the intended target phone) as an objective measure of speaker "intelligibility". Figure 8 and 9 show the obtained % correct for each speaker for the unvoiced and voiced stops respectively for each of the two classification systems.

Figure 8 ***Scatter plot of percentage correct achieved target for unvoiced stops in native (N,+) and non-native (NN,o) datasets***

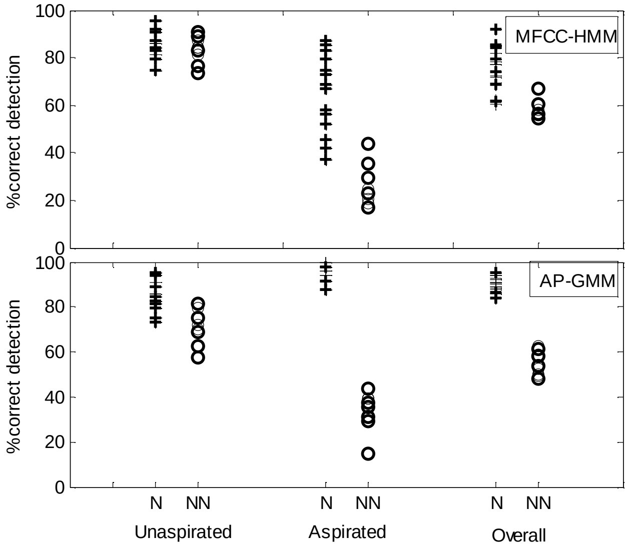
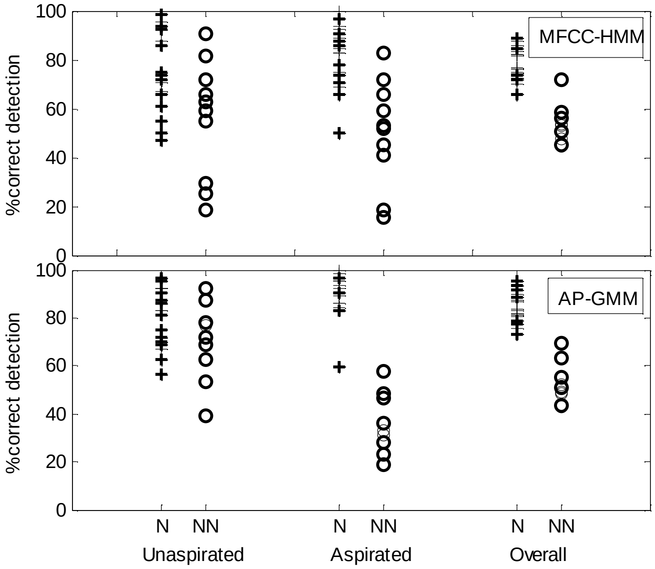


Figure 9 ***Scatter plot of percentage correct achieved target for voiced stops in native (N,+) and non-native (NN,o) datasets***



We see that the measured intelligibility varies across speakers, with the non-native speakers' group doing worse overall as should be expected. Given that native listeners were able to accurately detect the non-native speaker utterances of words containing aspirated plosives in the listening test reported in Sec. 3.2, this observation on the separation between native and non-native speakers can be viewed as a validation of the objective measure. The results suggest, for instance, that a lower than 40% correct realization of aspiration according to the objective measure is a strong indicator of non-native like pronunciation.

Observations of the individual scores of the 10 non-native speakers showed that their relative positions matched across the voiced and unvoiced stops, indicating that the phonemic aspiration contrast is acquired by Tamil-L1 learners similarly across both voicing classes. We observe that the overall intelligibilities of the native (N) and non-native (NN) speakers are better separated by the AP system relative to separation achieved by the baseline MFCC system. While the non-native speakers show the expected poor realization of aspirated targets, the AP system also indicates a few compromised unaspirated targets by the non-native speakers. This is not surprising in view of the allophonic usage of aspiration in Tamil word-initial stops, leading to the incorrect introduction of some aspiration in the target Hindi word-initial unaspirated stops. Thus it is possible that the phonetic realization of unaspirated plosives differs from that of natives even if native listeners find them perceptually equivalent as indicated in Sec. 3.2.

Figure 10 ***Scatter plot of percentage correct achieved target for unvoiced affricates in native (N,+) and non-native (NN,o) datasets.***

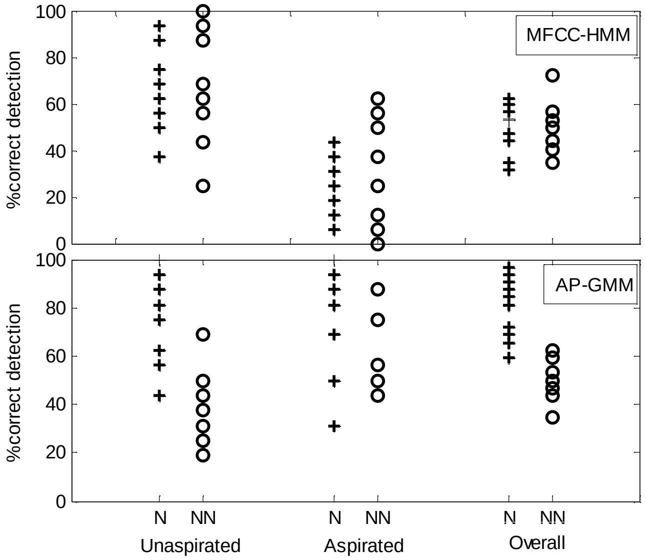
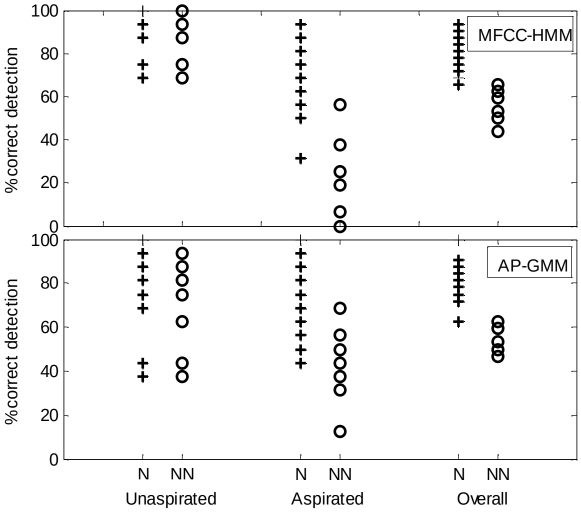


Figure 10 and 11 show the obtained %correct for each speaker for the unvoiced and voiced affricates respectively. Similar to the case of stops, the AP system obtains a better separation of native and non-native speakers compared with the MFCC features. The AP features show a reduced separation with affricates compared with that for stops in Figures 8 and 9. This is consistent with the relatively poor classification accuracy of affricates captured in Table 8 and attributed to the difficulty of VOP landmark detection.   The MFCC system shows a complete overlap in the predicted intelligibility of unvoiced affricates, in line with its particularly poor classification accuracy on this phone classes seen in Table 8.

Figure 11 ***Scatter plot of percentage correct achieved target for voiced affricates in native (N,+) and non-native (NN,o) datasets.***



5.2.2 Correlation with subjective segmental judgments

In Figure 8 to Figure 11 we observe an overlap in the overall intelligibility scores especially in the case of the MFCC system. That is, some native speakers are rated lower by the system than the best ranked non-native speaker. We use this observation to choose a smaller set of speakers for the subjective validation of the objective predictions at phone level via human perception tests. The speakers used in the perception test data include 6 good-to-poor rated native speakers from the baseline system, and 3 non-native speakers with various automatic intelligibility ratings (spanning the full range), separately for each of the 4 broad classes: the unvoiced and voiced stops and affricates. We would like to evaluate the automatic systems in terms of predicting the judgment of a native listener with the eventual goal of providing reliable feedback at the segment level in computer-aided language learning.

Two judges, fluent speakers of Marathi and Hindi with Masters degrees in engineering, labeled every voiced and unvoiced plosive segment of each chosen speaker with one of 3 categories: unaspirated, aspirated, unsure. So as to not bias the judges, the isolated stop-vowel and affricate-vowel (CV) segments extracted from the word were presented for listening in random order. The presented segments extended from the closure to until 40 ms beyond the vowel onset making them comparable with the data used in the automatic classification. Each listener classified 1152 and 1008 segments each of the voiced and unvoiced stop CVs respectively, presented in randomized order over 6 sessions of approximate duration 30 minutes each. Similarly 288 CVs each of unvoiced and voiced affricates were also rated. The judges labeled each CV with one of 3 tags: unaspirated, aspirated, unsure. It was observed that the maximum number of instances rated “unsure” by any judge was below 2% of the total for the native speakers, and less than 5% for the non-native speakers. The recognition task was chosen, over a quality rating task, keeping in mind the expected categorical perception of plosive aspiration by native listeners.

In order to make the objective ratings comparable with the subjective, we choose a region in Figure 7 around the 0 dB threshold value of the log likelihood ratio and of width given by a fixed fraction (0.1) of the standard deviation of the native distribution to indicate “unsure” in the system classification. We thus obtain a 3-category rating for each token by each of the automatic systems. A measure of the match between corresponding ratings was computed between judges, and between the consensus judgement and corresponding rating from each of the automatic systems. Table 9 presents the percentage agreement for each plosive class between the two judges. We further compute the Cohen's kappa coefficient, a statistical measure of inter-rater reliability valid for two raters (Cardillo, 2007). We note that the judges substantially agree across tokens of all plosive classes with higher agreement on the stops relative to the affricates. The use of segmented utterances in the listening test and the availability of an "unsure" label could explain the deviation from perfect agreement. It was observed that of the fraction of the utterances where the judges disagreed, less than a third corresponded to native speech. While the judges tend to differ more on the affricates compared to the stops, a higher proportion of disagreements was observed to occur on the velars and labials in both unvoiced and voiced stop categories.

Table 9. ***Inter-subject and system-subject correlations in terms of percentage agreement of ratings for the different plosive classes based on the data of 9 speakers (6 native and 3 non-native). Cohen's kappa coefficient is in parentheses***

Class	**Total count**	**Inter-subject % agreement (K)**	**System-subject correlation ratings**	
			AP-GMM	**MFCC-HMM**
			% agreement (K)	**% agreement (K)**
Unvoiced stops	1008	92.6 (0.72)	86.0 (0.52)	75.6 (0.32)
Unvoiced affricates	288	88.2 (0.60)	78.4 (0.35)	38.6 (-0.07)
Voiced stops	1152	89.1 (0.67)	82.1 (0.47)	70.1 (0.20)
Voiced affricates	288	87.9 (0.63)	69.2 (0.32)	79.5 (0.44)

We next consider the subset of tokens on which both human judges agree, and evaluate the match between each system and the human rating. Table 9 shows the percentage agreement computed between the corresponding ratings from human judgement and each of the automatic systems together with the estimated Cohen's kappa coefficients. The AP-GMM system shows good agreement for both unvoiced and voiced stops. As pointed out in Sec. 2.1, the relatively low performance of the AP system on affricates may be explained by the critical dependence of the features on the accurate detection of the vowel onset point. From Table 9, we note that the AP features provide phone-level feedback that is better matched with subjective ratings when compared with that of the MFCC system in all plosive classes except the voiced affricates.

From a closer observation of the AP system classification errors with respect to human judgement, it was seen that the velars contributed most in both the unvoiced and voiced stop categories. While some errors come from poor landmark detection, the rest can be attributed to speaker variability with respect to the chosen features. We speculate that worse performance on velars may be linked to the relatively low "ease of articulation" of velars, at least in the voiced case (Shariatmadari, 2006), leading to more ambiguity in the phonetic realizations across speakers. The phone-level feedback from the AP system for the class of unvoiced affricates is considerably lower than that of the other plosives which is also the case with the MFCC system. As noted earlier, the inter-judge correlation too is lower in the case of affricates reflecting the ambiguity associated with the acoustics of aspiration in affricates.

6.0 Conclusion

Motivated by a prominent phonological characteristic of Indo-Aryan languages, we propose a method to identify non-native accents of spoken Hindi. The incorrect production of the aspiration contrast in voiced and unvoiced oral stops and affricates of Hindi is a characteristic of non-native Hindi speakers whose L1 does not belong to the Indo-Aryan language group. We discuss several acoustic attributes that are the phonetic correlates of the phonological contrast involving aspiration for unvoiced and voiced plosives across places of articulation and vowel contexts. Exploiting such relevant distinctions via discriminating acoustic features facilitates the automatic assessment of a language learner's accent and can provide reliable segmental feedback. We consider the detection of the aspiration contrast in Hindi plosives as realised by native speakers and by non-native speakers with Tamil L1. The other Dravidian group languages, Telugu and Malayalam, share the phonology of Tamil plosives, potentially widening the applicability of the present work.

Several acoustic-phonetic features motivated by the understanding of the production of aspirated plosives are presented and evaluated for the classification of word-initial plosives in CV context in native speech. The features capture the release characteristics, and the distinctive glottal pulse shape and aspiration noise in the vowel region following the voiced and unvoiced aspirated plosives. The lag-VOT (interval between burst onset and following vowel onset) is the chosen duration feature, common across voicing categories. Statistical descriptions indicate that the proposed acoustic measures differentiate the aspirated plosives from their unaspirated counterparts in the case of unvoiced as well as voiced categories.

The acoustic-phonetic features are compared with more standard automatic speech recognition features in a baseline MFCC-HMM system for classification based on aspiration. The AP features are observed to be more robust to cross-lingual training and testing as demonstrated by a classification experiment using Marathi speech-trained acoustic models on native Hindi utterances. Their phonological basis makes them comparatively less sensitive to irrelevant spectral variations arising from language dependent phonetic realizations. On a test set of native and non-native Hindi speech comprising word-initial plosive utterances, the AP features based system separates the native and non-native speakers on the basis of correct detections of aspirated plosives. The acoustic-phonetic features also outperformed the MFCC-HMM system in terms of phone-level feedback that was more consistent with human judgment. The aspiration contrast in unvoiced affricates showed particularly low inter-judge agreement and correspondingly proved more difficult to characterize acoustically. Future improvements in acoustic landmark detection along with better trained models of the aspiration contrast are promising directions for improvement. The observed gender dependence of the AP features suggests that the use of larger gender specific datasets could help improve the acoustic model training and lead to more accurate aspiration detection. In the present work, we have clubbed together training samples of stop consonants across PoA and vowel context to obtain a single acoustic model for specified voicing and aspiration. With a considerably larger native Hindi speech database, it should be possible to train PoA and vowel context specific acoustic models as well in order to further improve aspiration detection in the language learning context. Finally, this work can be viewed as a validation of the power of carefully designed acoustic-phonetic features in speech recognition tasks such as pronunciation training where phonetic errors must be detected when the acoustic similarity among phonetic realizations is high.

Reference List

Bada, E. (2001). Native language influence on the production of English sounds by Japanese learners.*The Reading Matrix*, 1,

Balasubramanian, T. (1975). Aspiration of voiceless stops in Tamil and English: an instrumental investigation. *CIEFL Newsletter*, 14-18.

Benguerel, A. and Bhatia, T. (1980). Hindi stop consonants: An acoustic and fibroscopic study. *Phonetica*, *37*, 134-148.

Best, C. T., McRoberts, G. W., &Goodell, E. (2001).Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*,*109*(2), 775-794.

Bhaskararao, P. (2011). Salient phonetic features of Indian languages.*Sadhana*, 36, 587-599.

Bhela, B. (1999). Native language interference in learning a second language: Exploratory case studies of native language interference with target language usage. *International Education Journal*, 1, 22-31.

Cardillo, G. (2007) Cohens kappa: compute the Cohen's kappa ratio on a 2x2 matrix. http://www.mathworks.com/matlabcentral/fileexchange/15365, retrieved on November 8, 2015.

Carranza, M., Cucchiarini, C., Burgos, P., &Strik, H. (2014). Non-native speech corpora for the development of computer assisted pronunciation training systems, Conference Paper, Research Gate.

Celce Murcia, M., & Goodwin, J. (1991).Teaching pronunciation.In M. Celce Murcia (Ed.), *Teaching English as a second language*. Boston: Heinle&Heinle.

Census of India (2011), website of the Govt. Of India http://www.censusindia.gov.in/, retrieved on July 15, 2015.

Chelba, C., Bikel, D., Shugrina, M., Nguyen, P., and Kumar, S., "Large scale language modeling in automatic speech recognition," Tech. Rep., Google, 2012.

Cho, T., &Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics*, *27*(2), 207-229.

Cho, T., Jun, S., and Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics, 30*, 193-228.

Clements, G., N. and Khatiwada, R. (2007). Phonetic realization of contrastively aspirated affricates in Nepali. *In Proceedings International Congress of Phonetic Sciences,ICPhS XVI*, Saarbrucken, Germany, August 2007, pp. 629-632.

Davis, K.(1994), Stop voicing in Hindi, *Journal of Phonetics*,22, 177-193.

Dixit, R. P. (1989). Glottal gestures in Hindi plosives. *Journal of Phonetics*,17(3), 213-237.

Dutta, I. (2007). Four-way contrast in Hindi: An acoustic study of voicing, fundamental frequency and spectral tilt. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.

Flege, J., E. and Port, R. (1981). Cross language phonetic interference: Arabic to English. *Language and Speech*, 24, 125-146.

Franco, H., Ferrer, L. and Bratt, H. (2012). Adaptive and discriminative modeling for improved mispronunciation detection. In *Proceedings of SRI Publications*, Stockholm, Sweden, June 2012, pp. 53-58.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383-406.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America, 101*, 466-481.

Hanson, H. M., Stevens, K. N., Kuo, H. K. J., Chen, M. Y., & Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics*, *29*(4), 451-480.

Ishi, C. T. (2004). A new acoustic measure for aspiration noise detection. *In ProceedingsInternational Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, August 2004, pp. 629-632.

Jurafsky, D. and Martin, J. H. (2008). Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd edition, 2008, Prentice Hall.

Klatt, D. H. and Klatt, L.C. (1990).Analysis, synthesis, and perception of voice quality variations among female and male talkers.*Journal of the Acoustical Society of America, 87*, 920-857.

Ladefoged, P. and Maddieson, I. (2005).The Sounds of World's Languages.Blackwell Publishing, 2005.

Lev-Ari,S.andKeysar, B.(2010), Why don't we believe non-native speakers? The influence of accent on credibility, Journal of Experimental Social Psychology, 46(6), 1093-1096.

Lisker, L. and Abramson, A. (1964). Cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384-422.

Liu, S. A. (1996). Landmark detection for distinctive feature-based speech recognition.*Journal of the Acoustical Society of America, 100*, 3417-3430.

Masica, Colin P. *The Indo-Aryan Languages*.Cambridge University Press, 1993.

McAllister, R., Flege, J. E. and Piske, T. (2002).The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian.*Journal of Phonetics, 30*, 229-258.

Mikuteit, S., andReetz, H. (2007). Caught in the ACT: The timing of aspiration and voicing in East Bengali. *Language and speech*, *50*(2), 247-277.

Miller A. L. (2007). Guttural vowels and guttural co-articulation in Juhoansi. *Journal of Phonetics, 35*, 56-84.

Mohan, A., Rose, R., Ghalehjegh, S. H., & Umesh, S. (2014). Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. *Speech Communication*, *56*, 167-180.

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, *36*(2), 308-317.

Murphy, P. J. and Akande, O. O. (2007). Noise estimation in voice signals using short-term cepstral analysis. *Journal of the Acoustical Society of America, 121*, 1679-1690.

Niyogi, P. and Ramesh, P. (2003). The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets. *Speech Communication*, *41*, 349-367.

Oh, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, *39*(1), 59-67.

Ohala, M. and Ohala, J. (1972).The problem of aspiration in Hindi phonetics.*Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, No. 6*, 39-46.

Patil, V., Joshi, S. and Rao, P. (2009). Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach. *In Proceedings of Interspeech 2009*, Brighton, U.K., September 2005, pp. 2543-2546.

Patil, V. and Rao, P. (2011). Acoustic features for detection of aspirated stops. *InProceedings of National Conference on Communication*, Bangalore, India,January 2011, pp. 1-5.

Patil, V. and Rao P. (2013a). Acoustic features for detection of phonemic aspiration in voiced plosives. *In: Proceedings of Interspeech,* Lyon, France, August 2013, pp. 1761-1765.

Patil, V. and Rao P. (2013b). Automatic pronunciation feedback for phonemic aspiration.*In: Proceedings of SLaTE,* Grenoble, France,September 2013, pp. 116-121.

Patil, V. (2014).Automatic classification of obstruents of Marathi and Hindi.Ph.D.Dissertation, Indian Institute of Technology, Bombay, India.

Prasanna, S. R. M., and Yegnanarayana, B. (2005).Detection of vowel onset point events using excitation information.*InProceedings of Interspeech,* Lisbon, Portugal, September 2005, pp. 1133-1136.

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.

Raphael, L. J., Borden, G. J. and Harris, K. S. (2007).*Speech Science Primer: Physiology, Acoustics, and Perception*. (5thed.). Lippincott Williams & Wilkins.

Ridouane, R., Clements, G., N. and Khatiwada, R. (2011).Language-independent bases of distinctive features.*Tonesand features: Phonetic and Phonological Perspectives*, 260-287.

Samudravijaya, K. (2003). Durational characteristics of Hindi stop consonants. *In Proceedings of Eurospeech*, Geneva, Switzerland, September 2003, pp. 81-84.

Scanlon, P., Ellis, D. P. W. and Reilly, R. B. (2007). Using broad phonetic group experts for improved speech recognition. *IEEE Trans. on Speech And Audio Processing, 15(3),* 803–812.

Schiefer, L. (1986). F0 in the production and perception of breathy stops: Evidence from Hindi. *Phonetica*, *43*(1-3), 43-69.

Shariatmadari, D. (2006). Sounds difficult? Why phonological theory needs' ease of articulation'. *School of Oriental and African Studies Working Papers in Linguistics*, *14*, 207-226.

Stouten, F., & Martens, J. P. (2006). On the use of phonological features for pronunciation scoring. In *Proceedings International Conference on Acoustics, Speech and Signal Processing ICASSP May 2006, pp.329-332.*

Strik, H., Troung, K., Wet F. and Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection.*InProceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 1837-1840.

Strik, H., Truong, K., Wet, F. and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51(10)*, 845-852.

Theodoros, G. (2008). Histogram based class separability measure. Retrieved from http://www.mathworks.in/matlabcentral/fileexchange/18791-histogram-based-class-separability-measure, Retrieved on November 8, 2015.

Traill, A. (1980). Phonetic diversity in the Khoisan languages. In (J. W. Snyman, ed.) Bushman and Hottentot Linguistic Studies, pp. 167-189, University of South Africa, Pretoria.

Truong, K., Neri, A., Cuchiarini, C. and Strik, H. (2004). Automatic pronunciation error detection: an acoustic-phonetic approach. *In Proceedings of InSTIL/ICALL Symposium*, Venice, Italy, June 2004, pp. 135–138.

vanDoremalen, J., Cucchiarini, C., &Strik, H. (2013). Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch. *The Journal of the Acoustical Society of America*, *134*(2), 1336-1347.

Wiltshire, C. R. and Harnsberger, J. D. (2006). The influence of Gujarati and Tamil L1s on Indian English: a preliminary study. *World Englishes, 25(1),* 91-104.

Witt, S. and Young, S. (1997). Language learning based on non-native speech recognition. *In Proceedings of Eurospeech*, Rhodes, Greece, September 1997, pp. 633-636.

Young S. et al., (2006). The HTK Book v3.4, Cambridge University, 2006.

Appendix : **Hindi word list arranged by voicing and place of articulation**

Unvoiced plosives

Sr.	Words	Transcription	Sr.	Words	Transcription
1	कलाप्रेमी	kəlapɾemi	1	खडा	kʰəɖ a
2	काजू	kɑdʒu	2	खामोशी	kʰɑmoʃi
3	किमया	kɪməja	3	खिचडी	kʰɪtʃəɖi
4	कीमती	kimət̪i	4	खीसकाना	kʰisəkɑna
5	कुटी	kʊʈi	5	खुदा	kʰʊd̪a
6	कुदना	kud̪əna	6	खूबी	kʰubi
7	केतकी	ket̪əki	7	खेती	kʰet̪i
8	कोयला	kojəla	8	खोदाई	kʰod̪ai

Sr.	Words	Transcription	Sr.	Words	Transcription
1	चखाना	tʃəkʰɑnɑ	1	छतरी	tʃʰət̪əɾi
2	चालाकी	tʃɑlɑki	2	छापना	tʃʰɑpənɑ
3	चिडियाघर	tʃɪɖɪjɑgʰəɾə	3	छिपाना	tʃʰɪpɑnɑ
4	चीखना	tʃikʰənɑ	4	छीडकना	tʃʰiɖəkənɑ
5	चुनावी	tʃʊnɑʋi	5	छुडाना	tʃʰʊɖɑnɑ
6	चूसना	tʃusənɑ	6	छूटना	tʃʰuʈənɑ
7	चेतावनी	tʃet̪ɑʋəni	7	छेदना	tʃʰed̪əna
8	चोटी	tʃoʈi	8	छोटा	tʃʰoʈɑ

Sr.	Words	Transcription	Sr.	Words	Transcription
1	टमाटर	ʈəmɑʈəɾə	1	ठगाना	ʈʰəgɑnɑ
2	टालना	ʈɑlənɑ	2	ठानलेना	ʈʰanəlena
3	टिकाऊ	ʈɪkɑu	3	ठिकाना	ʈʰɪkɑnɑ
4	टीकाकार	ʈikɑkɑɾə	4	ठीकाना	ʈʰikɑnɑ
5	टुकडा	ʈʊkəɖɑ	5	ठुकराना	ʈʰʊkəɾɑnɑ
6	टूटाफूटा	ʈuʈɑpʰuʈɑ	6	ठूसना	ʈʰusənɑ
7	टेलिफोन	ʈelɪpʰonə	7	ठेकेदारा	ʈʰeked̪ɑɾɑ
8	टोपी	ʈopi	8	ठोकर	ʈʰokəɾə

Sr.	Words	Transcription	Sr.	Words	Transcription
1	तथापि	t̪ət̪ʰɑpɪ	1	थकान	t̪ʰəkɑnə
2	तारिका	t̪ɑɾɪkɑ	2	थाली	t̪ʰɑli
3	तिमाही	t̪ɪmɑɦi	3	थिरकाना	t̪ʰɪɾəkɑnɑ
4	तीसरा	t̪isəɾɑ	4	थीरकाना	t̪ʰiɾəkɑnɑ
5	तुलसी	t̪ʊləsi	5	थुलथुल	t̪ʰʊlət̪ʰʊlə
6	तूफानी	t̪upʰɑni	6	थूकदान	t̪ʰukəd̪ɑnə
7	तेजोमय	t̪edʒoməjə	7	थेगली	t̪ʰegəli
8	तोता	t̪ot̪ɑ	8	थोडा	t̪ʰoɖɑ

Sr.	Words	Transcription
1	पचास	pətʃɑsə
2	पालना	pɑlənɑ
3	पिताजी	pɪt̪ɑdʒi
4	पीछडा	pi tʃʰəɖɑ
5	पुछना	pʊtʃʰənɑ
6	पूजारी	pudʒɑɾi
7	पेशा	peʃɑ
8	पोशीदा	poʃid̪ɑ

Voiced plosives

Sr.	Words	Transcription	Sr.	Words	Transcription
1	गति	gət̪ɪ	1	घटना	gʰəʈənɑ
2	गाडी	gɑɖi	2	घाटिका	gʰɑʈɪkɑ
3	गिटार	gɪʈɑɾə	3	घिसापीटा	gʰɪsapiʈɑ
4	गीता	git̪ɑ	4	घीसापीटा	gʰisapiʈɑ
5	गुजारिश	gʊdʒɑɾɪʃə	5	घुसाना	gʰʊsanɑ
6	गूढप्रश्न	guɖʰəpɾəʃnə	6	घूसखोरी	gʰusəkʰoɾi
7	गेहूँ	geɦun	7	घेरना	gʰeɾənɑ
8	गोपाल	gopɑlə	8	घोडागाडी	gʰoɖɑgɑɖi

Sr.	Words	Transcription	Sr.	Words	Transcription
1	जगाना	dʒəgɑnɑ	1	झरोका	dʒʰəɾokɑ
2	जामीन	dʒɑminə	2	झाडी	dʒʰɑɖi
3	जिम्मेदारी	dʒɪmmed̪ɑɾi	3	झिलमिली	dʒʰɪləmɪlɪ
4	जीना	dʒinɑ	4	झील	dʒʰilə
5	जुदाई	dʒʊd̪ai	5	झुकाना	dʒʰʊkɑnɑ
6	जूता	dʒut̪ɑ	6	झूठा	dʒʰuʈʰa
7	जेलखाना	dʒeləkʰɑnɑ	7	झेलना	dʒʰelənɑ
8	जोशीला	dʒoʃilɑ	8	झोपडी	dʒʰopəɖi

Sr.	Words	Transcription	Sr.	Words	Transcription
1	डरपोक	ɖəɾəpokə	1	ढकना	ɖʰəkəna
2	डाकखाना	ɖɑkəkʰana	2	ढाचा	ɖʰɑtʃɑ
3	डिब्बा	ɖɪbba	3	ढिलाई	ɖʰɪlɑi
4	डीका	ɖika	4	ढीला	ɖʰilɑ
5	डुबोना	ɖʊbona	5	ढुलमुल	ɖʰʊləmʊlə
6	डूबाना	ɖubana	6	ढूढ	ɖʰuɖʰə
7	डेरा	ɖeɾa	7	ढेला	ɖʰelɑ
8	डोलना	ɖolənɑ	8	ढोलवादन	ɖʰoləʋɑd̪ənə

Sr.	Words	Transcription	Sr.	Words	Transcription
1	दगाबाजी	d̪əgabadʒi	1	धनादेश	d̪ʰənnɑd̪eʃə
2	दादागिरी	d̪ɑd̪agɪɾi	2	धातु	d̪ʰɑt̪ʊ
3	दिखाऊ	d̪ɪkʰɑu	3	धिक्कारना	d̪ʰɪkkɑɾəna
4	दीपावली	d̪ipɑʋəli	4	धीरज	d̪ʰiɾədʒə
5	दुनिया	d̪ʊnɪja	5	धुलाई	d̪ʰʊlai
6	दूरभाष	d̪uɾəbʰɑʃə	6	धुसरता	d̪ʰusəɾət̪a
7	देशवासी	d̪eʃəʋasi	7	धेनु	d̪ʰenʊ
8	दोषित	d̪oʃɪt̪ə	8	धोबी	d̪ʰobi

Sr.	Words	Transcription	Sr.	Words	Transcription
1	बचाना	bətʃɑnɑ	1	भगवान	bʰəgəʋɑnə
2	बालवाडी	bɑləʋɑɖi	2	भावना	bʰɑʋənɑ
3	बिजली	bɪdʒəli	3	भिगाना	bʰɪgɑnɑ
4	बीताना	bit̪ɑnɑ	4	भीषण	bʰiʃəɳə
5	बुढापा	bʊɖʰapɑ	5	भुलना	bʰʊlənɑ
6	बूढिया	bu ɖʰɪja	6	भूगोल	bʰugolə
7	बेहोशी	beɦoʃi	7	भेजना	bʰedʒənɑ
8	बोली	boli	8	भोलापन	bʰolapənə