Automatic Pronunciation Feedback for Phonemic Aspiration

Vaishali Patil, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India.

{vvpatil, prao}@ee.iitb.ac.in

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. Aspiration, an important phonemic attribute in plosives of Indo-Aryan languages such as Hindi, Marathi and Gujarati, is rarely found in the world׳׳s languages. The improper production of the aspiration contrast is thus often the most important cue to non-native accents of spoken Hindi. A system for the detection of phonemic aspiration in unvoiced and voiced stops based on discriminative acoustic features is shown to be effective for rating non-native accents and providing reliable phoneme-level feedback.

Index Terms: computer-assisted language learning, pronunciation scoring, non-native accent, phonemic aspiration

1. Introduction

The computer-assisted learning of spoken language is closely tied to automatic speech recognition (ASR) technology. The automatic assessment of a non-native learner based on carefully designed speaking tests coupled with focused phone-level feedback would go a long way into expanding the reach of the language education industry. While intelligibility is a prime requirement, the absence of non-native accents, as indicated by segmental (phone articulation) and suprasegmental (prosody) differences from native speech, are desirable. A key manifestation of foreign accent is the improper production of the target language (L2) phones. This is especially true when the phones in question do not belong to the phonology of the learner׳׳s native language (L1).

The challenges of automation 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. In an application such as pronunciation assessment, however, language models would obscure genuine pronunciation errors by the non-native learner. Further, for better raw phone recognition accuracy, the acoustic models need to be trained on actual non-native speech. Such a speech database may not be easily available.

In recent past work, widely used direct measures of pronunciation quality from hidden Markov model (HMM) log-likelihoods in state-of-the-art mel-frequency cepstral coefficient (MFCC) feature based systems have not been found completely effective [1]. The MFCC features are a generic representation of the spectral envelope of the signal. More accurate judgement as well as meaningful feedback may be obtained via acoustic features that can be mapped to specific phonological attributes. In the present work, we investigate this approach to the automatic assessment of pronunciation of stop consonants of Hindi, belonging to the Indo-Aryan group, among the few language groups of the world where aspiration is a phonemic attribute. The improper production of the aspiration distinction is an important cue of non-native accent, in addition to vowel quality and intonation [2]. Limiting the scoring to the relevant aspects only would improve the reliability of the system by ignoring other natural variabilities of speech, and facilitate the use of specific discriminatory acoustic features for these aspects.

The 4-way contrast of Hindi stops where voicing and aspiration are independent for each place-of-articulation (PoA) are typically challenging for a learner from a different native language group. In the present study, we consider speakers of Tamil (as L1), a Dravidian language whose phonology is devoid of phonemic aspiration. Hindi is the native tongue of 200 million people in India and Tamil, that of 70 million. Hindi is the national language of India and together with English serves as a link language across the multilingual country. With widespread internal migration, the need for spoken language acquisition of the common languages is high.

The detection of phonetic differences involving aspiration has been previously attempted for Korean unvoiced stops, where a spectral tilt feature was added to vowel onset time (VOT), a traditionally used acoustic measure to distinguish aspirated stops from unaspirated (tense) [3]. While this achieved a good discrimination between aspirated and lax stops, it was less effective for the aspirated-tense case. Based on a phonological observation that aspiration is marked by breathy voice in the following vowel, Clements and Khatiwada [4] investigated the acoustic distinction between aspirated and unaspirated Nepali affricates on a small set of speakers to find that the acoustic measures of breathiness were not reliable across speakers. Voice quality features have been shown to enhance detection accuracies for phonemic aspiration in unvoiced Marathi stops [5], [6]. Voiced aspirated plosives, due to their rare occurrence in the world׳׳s languages, have been studied minimally [7], [8]. Our recent work proposed and evaluated multiple acoustic features, extracted from the consonant and the following vowel regions, for the reliable detection of aspiration in word-initial voiced stops [9].

A goal of the present study is to develop and evaluate a speaker-independent automatic system for the robust detection of aspiration in Hindi voiced and unvoiced stops that can be used in a pronunciation scoring task for non-native speech. An objective measure of intelligibility is proposed based on maximum likelihood classification that is further validated by human listener ratings. The performance of the proposed system is compared with that of a baseline MFCC-HMM ASR system in the context of rating non-native pronunciation and providing corrective feedback.

2. Database and baseline system

Hindi and Tamil belong to distinct language groups that differ prominently in the plosive system. While both languages contain oral stops of 4 places of articulation, voicing and aspiration are used distinctively only in Hindi as depicted in

Table 1. Tamil does not distinguish aspiration or even voicing; stops are voiceless and weakly aspirated in initial position, and voiced after nasals [10]. Since our work is targeted towards a pronunciation assessment task, we collect data from native and non-native Hindi speakers in the form of read-out words containing the target phones in word-initial position across vowel contexts. For training the acoustic models, we use an already available database of Marathi spoken words by 20 native Marathi speakers. Marathi and Hindi are both Indo-Aryan languages and share the stop series of Table 1. In both, the unvoiced, aspirated labial is rarely used and therefore omitted.

Table 1. *IPA chart showing stops of Hindi and Tamil languages.*

Language	PoA of unvoiced and voiced stops			
	Labial	Dental	Retroflex	Velar
Hindi	p	t̪ t̪^{h}	ʈ ʈ^h	k k^h
	b b^h	d̪ d̪^{h}	ɖ ɖ^h	g g^h
Tamil	p (b)	t̪ (d̪)	ʈ (ɖ)	k (g)

2.1 Training and testing datasets

The training database comprises Marathi spoken words sampled at 16 kHz. Two distinct meaningful words with word-initial stops corresponding to each Hindi phone in Table 1 and each of the 8 vowels of the language (/ə/, /a/, /i/, /I/, /u/, /U/, /e/, and /o/) are formed and each word uttered in two carrier sentence contexts by 20 native speakers (equal male and female). The total number of words in each stop category appears in Table 2. The utterances were manually transcribed at phone level to use for acoustic model training.

Table 2. *Count of stop consonant-vowel (CV) pairs from train and test datasets.*

Data sets Stop category		Marathi train (20)	Hindi native (20)	Hindi non-native (10)
Unvoiced	Unaspirated	2560	1280	640
	Aspirated	1920	960	480
Voiced	Unaspirated	2560	1280	640
	Aspirated	2560	1280	640

For the pronunciation assessment evaluation, testing datasets were recorded by 20 native Hindi speakers and 10 speakers of Tamil L1. All were college-going adults. The non-native speakers had been exposed to Hindi reading and writing during their school years but had had limited exposure to the spoken language. They were fluent in Tamil, and used Hindi to varying extents as they currently lived outside their home state. The test dataset involved one meaningful word of Hindi corresponding to each consonant and vowel context embedded in 2 carrier phrases read out by each speaker. The speakers were presented the list of written words in Hindi script along with its English meaning. Table 3 shows examples of the words (that also happen to be minimal pairs) along with their typical pronunciations by native Hindi and Tamil speakers. Each dataset has an equal number of male and female speakers. A native Hindi listener was able to correctly identify every one of the speakers as native or not by listening to a small set of utterances (less than 20 words) by the speaker. It was observed by the listener that phonemic aspiration was the main discriminating attribute. Voicing was always realised correctly even though voicing is allophonic in Tamil stops.

Table 3. *Examples of stops in word initial as articulated by native and non-native speakers.*

Stop	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

2.2 System frame work and baseline

Acoustic features computed from the segment of interest are used in a statistical classifier, previously trained on native speech (the Marathi database in this case), to derive a measure of correctness of pronunciation in terms of the aspiration attribute of oral stops across places of articulation. Since the test speech comprises of known utterances, an alignment with the word‟s phonetic transcription is first achieved using manner broad class models in an available state-of-the-art MFCC-HMM ASR system [11]. Such broad phonetic class based alignment can provide better robustness to speaker and language variability expected in the context of non-native speech since most confusions in a phone recognizer tend to be within the same manner class [12]. The broad classes are: vowels, sonorant consonants, unvoiced fricatives, unvoiced affricates, unvoiced stops, voiced affricates, voiced stops, silence and voice bar. The acoustic models are context independent, 3-state HMM with 8 mixtures, diagonal covariance and flat-start initialization. The standard 39 dim MFCC, delta and acceleration feature vector was computed at 10 ms intervals.

The aligned segments corresponding to voiced and unvoiced stops are processed for the extraction of acoustic-phonetic (AP) features as described in the next section. A Gaussian Mixture Model (GMM) classifier (6 mixtures, full covariance) is trained on the feature vectors of each class: unvoiced unaspirated (UU), unvoiced aspirated (UA), voiced unaspirated (VU), voiced aspirated (VA). Two-way classification is carried out on test unvoiced stop segments, and similarly on test voiced segments. For comparison, we also have a baseline system, implemented by extending the broad class MFCC-HMM system by separating the unvoiced and voiced stops further into 2 classes each to get UU, UA, VU and VA classes.

3. Acoustic-phonetic feature extraction

Aspiration is perceived as a release of breath following the stop burst. The aspiration feature has traditionally been associated with the timing of voicing onset [13]. Aspiration is also accompanied by an increased glottal opening in many languages including Hindi and the presence of aspiration noise during the following vowel [14]. Acoustic correlates of aspiration thus include VOT, aperiodicity of the vowel waveform and spectral shape attributes: H1-H2 (amplitude of the first harmonic relative to the second, reflecting the glottal open quotient) and spectral tilt. The latter two have been extensively studied as acoustic correlates of breathy voice quality in vowels where spectral tilt has been measured in various ways including H1-A3 and A1-A3 where An is the highest amplitude in the region of the nth formant [15], [16]. Thus phonemic aspiration is clearly multidimensional in terms of articulation, and trade-offs can be expected in both the production and perception of a specific realization. Therefore multiple acoustic features have been considered for reliable detection. The implementation of feature extraction is presented next.

3.1. Acoustic landmark detection

The extraction of the acoustic-phonetic features needs the precise temporal locations of landmarks corresponding to burst onset and vowel onset in the CV region of each utterance. The segmentation achieved by the broad-class HMM recognizer of Sec. 2.2 is coarse and must be refined as presented here.

The release burst onset is detected by the largest peak in the rate-of-rise (ROR) of the smoothened energy in 3500-8000 Hz within a 40 ms vicinity of the coarse boundary [17], [18]. This achieves burst localization to acceptable precision. However, cues to vowel onset are dependent on the nature of the consonant and especially difficult for aspirated and voiced stops. We employ different methods for vowel onset detection in the case of unvoiced stops and voiced stops. The rise of periodicity is a prominent cue to vowel onset after an unvoiced stop. Periodicity measured by the autocorrelation function peak computed from sliding 25 ms windows at 1 ms intervals, throughout a region of 40 ms around the initial boundary, 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. In the case of voiced stops, we use the rapid rise in the signal amplitude envelope in the low frequency band (50 Hz – 600 Hz) to detect the precise vowel onset in the vicinity of the initial coarse boundary [19]. While a median localization error of 5 ms is observed with respect to manually detected onsets, experimentally measured acoustic parameters extracted based on the automatically labeled onsets are seen to correspond well with those extracted using manual labels indicating the efficacy of the landmark detection methods.

3.2. Feature implementation

The features used in the pronunciation scoring task are from previously published (or to appear) work on acoustic-phonetic features for aspiration detection [5], [9]. VOT has been widely used to discriminate unvoiced aspirated stops from unaspirated stops in English where the former appear in word-initial context as allophones for voiced stops [13]. It was shown that including spectral tilt (A1-A3) and noise (signal-to-noise ratio – SNR) features improves the classification performance for unvoiced stops further [5]. In the case of voiced stops, performance with VOT alone is barely above chance. Including the A1-A3 and SNR improves it greatly. It was later demonstrated that a performance more comparable to that on unvoiced stops was obtained only after including all the further features listed in Table 4 for voiced stops [9].

Table 4. *Features used for aspiration detection in the AP-GMM system.*

Class of stops	Features in AP-GMM system
Unvoiced	VOT, H1-H2, A1-A3, SNR
Voiced	VOT, H1-H2, A1-A3, SNR, F1F3-sync, Low-band-slope, B3-band energy

The feature implementation available in [5], [9] is briefly reviewed here. VOT is the duration between burst onset and vowel onset. H1-H2 and spectral tilt measurements are obtained in the vowel region from magnitude spectra from 25 ms Hamming-windowed DFTs computed at 1 ms hop and averaged over a selected 5 ms duration. Low-band slope and B3-band energy provide further descriptions of spectral roll-off from the second and third formant regions respectively. The SNR provides the ratio of harmonic energy to aspiration noise energy. It is computed using a 25 ms analysis window placed at a selected time instant beyond the vowel onset. Signal power is obtained from the DFT spectrum but aspiration noise power is estimated using cepstral liftering [20]. Cepstral liftering separates the source from the vocal tract shaping and helps make the SNR less sensitive to formant influences. Since aspiration noise dominates the higher frequency region where formants are weak, an independent method to estimate the noise strength is to measure the uncorrelatedness of the signal components in two different frequency regions. "F1-F3 sync" is such a feature proposed by Ishi [16], computed using F1 and F3 bands of width 600 Hz around the automatically detected formant values corresponding to that token. The index represents 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. The time-instants for spectral shape and noise features have been experimentally shown to be most discriminative at 13 ms and 23 ms respectively after the detected vowel onset [9].

4. Experiments and results

Table 5 shows the performances of the AP and MFCC features on 2-way classification (aspirated, unaspirated classes) of voiced and unvoiced stops. Table 4 lists the specific features used by the AP-GMM system. A 20-fold cross-validation (leave-one-speaker-out) experiment was carried out on the Marathi dataset. We observe that the MFCC features achieve an accuracy comparable to the AP features (the voiced stops performance is a bit lower) on the Marathi dataset. Next, both systems were trained on the full 20 speaker Marathi dataset, and tested on the 20 native speaker Hindi dataset. As seen in Table 5, the AP features show comparable performances on both language datasets whereas the MFCC features" performance decreases significantly. The AP features are clearly more robust to the cross-language transfer, as might be expected from their phonological basis.

Table 5. *Recognition accuracies on stops of Marathi and native Hindi datasets.*

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
Voiced stops	85.1	84.9	80.8	77.8

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. 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. The test database is as described in Sec. 2.1, where each of the 20 native and 10 non-native speakers read out 240 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) rating of non-native accent via ranking of speakers (with respect to correct recognition as the intended target phone by native listeners), and (iii) accuracy of phone-level feedback with respect to human perception.

4.1 Detection of non-native accent

Each test word is automatically segmented and the classifier makes a two-way forced choice between unaspirated and aspirated classes for each test CV segment. For each speaker, 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", separately for the unvoiced stops and voiced stops. Figures 1 and 2 show the obtained %correct for each speaker for the unvoiced and voiced stops respectively for each of the two different classifier systems. We see that the measured intelligibility varies across speakers with the non-native speakers‟ group doing worse overall. 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.

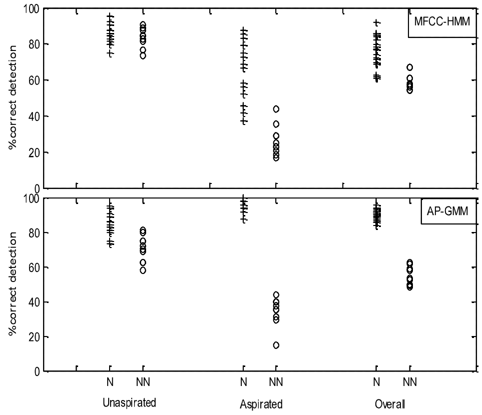

Figure 1: *Percentage correct achieved target unvoiced stops in native (N,+) and non-native (NN,o) datasets.*

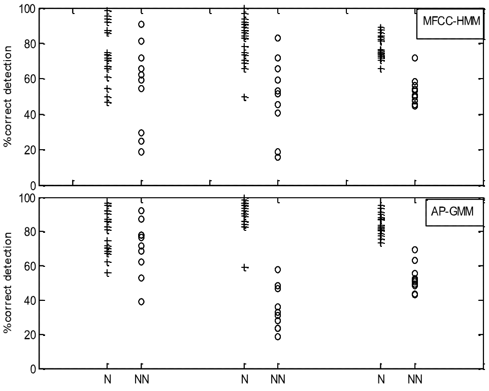

Figure 2: *Percentage correct achieved target voiced stops in native (N,+) and non-native (NN,o) datasets*

We note 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 poor realization of aspirated targets, the AP system also indicates 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 aspiration in the target Hindi word-initial unaspirated stops. For instance, „t̪alI‟ of Table 3 was sometimes pronounced „t̪^{h}alI‟.

4.2. Ranking non-native accent

In Figures 1 and 2 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* than the best ranked non-native speaker. We use this observation to choose a smaller set of speakers for the subjective validation of the ranking trends by human perception tests. The speakers used in the perception test data include the worst-rated native speaker from the baseline system, and 3 non-native speakers with various automatic intelligibility ratings, separately for the unvoiced and voiced stops.

Three judges, all fluent speakers of Marathi and Hindi, only one of whom is a trained speech scientist, labeled every voiced and unvoiced stop segment of each chosen speaker with one of 3 categories: unaspirated, aspirated, unsure. So as to not bias the judges, the *isolated* stop segments extracted from the word were presented for listening in random order. Each listener classified 512 and 448 segments each of the voiced and unvoiced stop CVs respectively, presented in randomized order over 4 sessions of approximate duration15 minutes each. The recognition task was chosen for the perception experiment rather than a "quality" evaluation in order to reduce subjectivity.

Table 5. *Percentage correct production of unvoiced stops as detected by listeners and by the automatic systems.*

Speaker ID	Perceptual results (%)			Classifier results (%)	
	Subject 1	Subject 2	Subject 3	MFCC-HMM	AP-GMM
N-8	94.6	93.8	93.8	60.7	89.3
NN-6	67.9	71.4	69.6	67.0	62.5
NN-1	58.9	58.0	59.8	54.5	58.9
NN-9	43.8	46.4	50.0	56.3	53.6

Table 6. *Percentage correct production of voiced stops as detected by listeners and by the automatic systems.*

Speaker ID	Perceptual results (%)			Classifier results (%)	
	Subject 1	Subject 2	Subject 3	MFCC-HMM	AP-GMM
N-20	92.2	89.8	89.8	65.6	82.0
NN-4	75.0	74.2	68.0	71.9	69.5
NN-9	57.0	57.0	50.8	53.9	51.6
NN-5	53.1	53.1	50.0	45.3	52.3

The results appear in Tables 5 and 6. A target is considered correctly achieved only if its perceived value (aspirated/unaspirated) matches the target. The maximum number of instances rated "unsure" by any judge, were below 3% of the total targets for the native speaker, and less than 4% for the 3 non-native speakers. The unsure cases were ignored in the analysis of this section. Tables 5 and 6 show the %correct target achieved according to each of the judges as well as each automatic system, arranged in decreasing intelligibility as per the 3 judges (whose speaker rankings turned out to be identical). We observe that the rank ordering of the AP system matches the subjective ranking. This is not the case with the baseline system which ranks N-20 lower than NN-4.

4.3 Accuracy of phone-level feedback

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 [21]. We use the log of ratio of likelihoods of the target and the opposite models as an estimate of the "goodness of pronunciation" of an uttered phone.

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

where $L(x| \wedge 1)$ is the likelihood of an arbitrary point x in the feature space for model of class 1 (likewise $L(x| \wedge 2)$ for class 2). Class 1 represents the target class while class 2 the opposite class.

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 wrong articulation. This is illustrated by Fig. 3 which shows the distribution of the log likelihood ratios over the native dataset for voiced stops for each of the AP and MFCC systems. As expected, the native utterances lie mostly to the right of the zero log-likelihood point. We choose a region around log likelihood ratio = 0 of width given by a fixed fraction (0.1) of the standard deviation of the native distribution to indicate "unsure" in the 3-way classification (correct/wrong/unsure) of the non-native utterances.

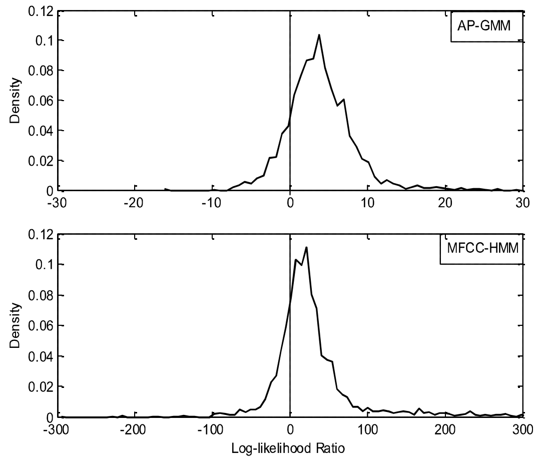

Figure 3: *Distribution of log-likelihood-ratio from AP-GMM and MFCC-HMM systems over native data set for voiced stops.*

The same speakers‟ data (extracted consonant segment only) ratings by the human judges, as already available from Task (ii) as presented in Sec. 4.2, were assigned the same numerical values viz. +1, 0, -1 for target correctly achieved, unsure and opposite achieved respectively. Next the correlation between corresponding ratings was computed between judges, and between each judge and each automatic system output. The inter-judge correlation is found to be 0.70.

Table 7. *Correlation between subjective and objective phone-level ratings for unvoiced and voiced stops.*

Class	Average correlation ratings	
	AP-GMM	MFCC-HMM
Unvoiced stops	0.53	0.39
Voiced stops	0.52	0.13

Table 7 shows the average of the correlation coefficients between each of the subject‟s ratings and the corresponding objective rating. We observe that the AP features provide phone-level feedback that is closer to subjective ratings when compared with that of the MFCC-HMM system which is especially poor for voiced stops.

5. Conclusion

The non-native accent that appears in a language learner‟s speech is, at the segmental level, related to the phonological differences between the speaker‟s L1 and the target language. Exploiting such relevant distinctions with suitable discriminating acoustic features can lead to reliable automatic assessment of degree of nativeness as well as detection of phoneme-level pronunciation errors. In this work, we presented the design and evaluation of a pronunciation scoring system for spoken Hindi where the learners‟ L1 is Tamil. The incorrect production of the aspiration distinction in voiced and unvoiced oral stops of Hindi is a prominent characteristic of non-native Indian speakers whose L1 does not belong to the Indo-Aryan language group. A statistical classifier using acoustic-phonetic features for aspiration detection was proposed based on the acoustic characteristics of voiced and unvoiced stops of Marathi. A number of methods are presented to evaluate the performance of the system in a pronunciation assessment context.

The AP features based system was shown to provide a measure of intelligibility that separates native and non-native speakers well. The acoustic-phonetic features outperformed a standard MFCC-HMM system on overall speaker intelligibility scoring as well as phoneme-level error detection. A discriminative classifier is expected to enhance the performance of the AP system further. Future work will extend the system to include other salient phonological attributes of spoken Hindi and larger scale evaluations on non-native speakers. A drawback of the AP features approach is that specific features are needed for specific phonological distinctions. Finding ways to select features automatically from suitable labeled training data would extend the scope of such work (similar to the suggestion by Strik et al [1]).

6. Acknowledgements

This work was supported in part by Bharti Centre for Communication at IIT Bombay.

7. References

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

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

[3] Cho, T., Jun, S., and Ladefoged, P., “Acoustic and aerodynamic correlates of Korean stops and fricatives”, Journal of Phonetics, 30, pp. 193-228, 2002.

[4] Clements, G., N. and Khatiwada, R., “Phonetic realization of contrastively aspirated affricates in Nepali”, In: Proc. ICPhS XVI 2007, Saarbrucken, Germany, pp. 629-632, Aug. 2007.

[5] Patil, V. and Rao, P., “Acoustic features for detection of aspirated stops”, In: Proc. of National Conf. on Communication 2011, Bangalore, India, pp. 1-5, Jan. 2011.

[6] Patil, V. and Rao P., “Automatic pronunciation assessment for language learners with acoustic-phonetic features”, In: Proc. of SLP-TED Workshop at COLING-2012, Mumbai, India, pp. 17-23, Dec. 2012.

[7] Rami, M. K., Kalinowski, J., Stuart, A. and Rastatter, M. P., “Voice onset times and burst frequencies of four velar stop consonants in Gujarati”, J. Acoust. Soc. Am. 106(6), pp. 3736-3738, Dec. 1999.

[8] Miller A. L., “Guttural vowels and guttural co-articulation in Juhoansi”, Journal of Phonetics, 35, pp. 56-84, 2007.

[9] Patil, V. and Rao P., “Acoustic features for detection of phonemic aspiration in voiced plosives”, to appear In: Proc. of Interspeech 2013, Lyon, France, Aug. 2013.

[10] Balasubramanian, T., “Aspiration of voiceless stops in Tamil and English: an instrumental investigation”, CIEFL Newsletter, pp. 14-18, 1975.

[11] Young S., et al., “The HTK Book v3.4”, Cambridge University, 2006.

[12] Scanlon, P., Ellis, D. P. W. and Reilly, R. B., “Using broad phonetic group experts for improved speech recognition”, IEEE Trans. Audio, Speech and Lang. Process, 15(3), pp. 803–812. Mar. 2007.

[13] Lisker, L. and Abramson, A., “Cross-language study of voicing in initial stops: Acoustical measurements”, Word, 20(3), pp. 384-422, Dec. 1964.

[14] Ridouane, R., Clements, G., N. and Khatiwada, R., “Language-independent bases of distinctive features”, Tones and features: Phonetic and Phonological Perspectives, pp. 264-291, 2011.

[15] Hanson, H. M., “Glottal characteristics of female speakers: Acoustic correlates”, J. Acoust. Soc. Am., 101(1), pp. 466-481, Jan. 1997.

[16] Ishi, C. T., “A new acoustic measure for aspiration noise detection”, In: Proc. ICSLP 2004, Jeju Island, Korea, pp. 629-632, Oct. 2004.

[17] Liu, S. A., “Landmark detection for distinctive feature-based speech recognition,” J. Acoust. Soc. Am., 100(5), pp. 3417-3430, Nov. 1996.

[18] Patil, V., Joshi, S. and Rao, P., “Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach”, In: Proc. of Interspeech 2009, Brighton, U.K., pp. 2543-2546, Sep. 2009.

[19] Prasanna, S. and Yegnanarayana, B., “Detection of vowel onset point events using excitation information”, In: Proc. of Interspeech 2005, Lisbon, Portugal, pp. 1133-1136, Sep. 2005.

[20] Murphy, P. J., and Akande, O. O., “Noise estimation in voice signals using short-term cepstral analysis”, J. Acoust. Soc. Am. 121(3), pp. 1679-1690, March 2007.

[21] Witt, S. and Young, S., “Language learning based on non-native speech recognition”, In Proc. of Eurospeech 1997, Rhodes, Greece, pp. 633-636, Sep. 1997.