Acoustic features for detection of phonemic aspiration in voiced plosives

Vaishali Patil, Preeti Rao

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

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

Abstract

Plosives in Indo-Aryan languages such as Hindi and Marathi display a 4-way contrast involving the two dimensions of voicing and aspiration. While many studies are available on the acoustics of aspiration in unvoiced stops due to their more universal presence in the world's languages, voiced aspirated plosives have been less studied. Rather than the release duration cue of aspiration in unvoiced stops, the acoustic realization of aspiration in voiced plosives is marked by the coarticulatory breathiness of the following vowel. We consider the automatic detection of aspiration in Marathi word-initial voiced stops and affricates via several features relating to extent and timing of breathiness of the following vowel. The effectiveness of the features is evaluated by classification performance on a database of Marathi words. A practical application of this work to the detection of non-native pronunciation of voiced obstruents is presented.

Index Terms: voiced obstruents, phonemic aspiration, breathy vowels, manner classification, acoustic-phonetic features

1. Introduction

In state-of-the-art ASR systems, phone recognition accuracies are relatively low, and an acceptable performance in practical tasks is achieved only through the constraints of a powerful language model. Inherently better phone recognition, however, would help in many situations such as cross-language contexts and the assessment of non-native pronunciation where the dependence on a language model is not possible. Acoustic features tailored to exploit phonetically relevant distinctions in the language have been a promising direction in speech recognition research [1, 2]. Such phone class-specific features can potentially be incorporated in a hierarchical classifier for superior phone recognition [3]. With this perspective, we consider a class of speech sounds common to Indo-Aryan languages but otherwise rare in the world's languages. Languages such as Hindi, Marathi and Gujarati, among others, display a four-way contrast, in plosives for each place of articulation, involving the dimensions of voicing and aspiration. We investigate acoustic features that can be mapped to the specific phonological feature based on available previous studies as well as our own observations from the analysis of a Marathi speech dataset.

Contrastively aspirated voiced plosives in Marathi occur with 5 places of articulation (PoA), including one affricate, as shown in Table 1. Studies on the acoustic realization of phonemic aspiration in consonants are available for the class of phones comprising voiceless-unaspirated, voiceless-aspirated and voiced-unaspirated stops (3-way contrast) that occurs in several languages. In English, the latter two represent the voicing distinction in word-initial stops, and are acoustically distinguished by release duration [4]. However many studies have indicated that release duration or vowel onset time (VOT) are weak cues to aspiration in *voiced* stops [4, 5]. An exception is a study of Bengali voiced stops which showed phonemic aspiration being distinguished by release duration in inter-vocalic geminates [6]. Certain Khosean (African click) languages employ contrastive aspiration in voiced guttural consonants. This phonological contrast has been found to be acoustically linked to the breathy voice quality of the vowel following the consonant suggesting the usefulness of measures such as spectral tilt and harmonic-to-noise ratio for the discrimination of the voiced consonants [7]. The use of voice quality measures for phonemic aspiration has been previously attempted for Korean unvoiced stops [8]. A spectral tilt measure was added to vowel onset time (VOT) to distinguish three levels of aspiration. The acoustic distinction between aspirated and unaspirated Nepali affricates (voiced and unvoiced) on a small set of speakers in two vowel contexts found that the acoustic measures of breathiness were not reliable across speakers [9]. Voice quality features, computed near manually labeled vowel onsets, were shown to enhance detection accuracies for phonemic aspiration in unvoiced Marathi stops [10].

In summary, previous studies suggest the uncertain role of release duration and the prominent role of voice quality of the following vowel in cueing phonemic aspiration in stops. Most studies have been confined to unvoiced stops, selected vowel contexts and very few speakers. Acoustic measurements have been manually implemented in most cases. Questions about which acoustic features best represent the phonemic aspiration contrast, including breathy voice quality, in the context of voiced aspirated plosives across a large range of PoA, and how these measures may be automatically computed from recorded speech, have not been addressed so far. Thus a goal of the present study is to identify speaker-independent acoustic features for the robust detection of aspiration in Marathi and Hindi voiced plosives. Based on the production and acoustics of voiced aspirated plosives, acoustic features are proposed for aspiration detection. The feature computation and choice of analysis parameters are experimentally validated via two-way classification performance on a Marathi speech database. Finally, the features are shown to be useful in the pronunciation assessment of voiced plosives of Hindi by non-native speakers.

2. Dataset and acoustic characteristics

Table 1 gives the voiced plosives in Marathi at all PoA and example word pairs where the plosive in word initial position differs only in aspiration manner. The palatal PoA is an affricate (rather than a stop) that appears allophonically as palate-alveolar unaspirated in certain vowel contexts. For the training and development of features and acoustic models for

classification, we use an available Marathi speech database. Cross-validation classification experiments on this dataset are used to arrive at the best set of speaker-independent features for aspiration detection. The trained models are further applied to a Hindi speech dataset in a pronunciation assessment task across native and non-native speakers.

Table 1. *PoA of voiced plosives and word examples with only word initial 'C' differing in aspiration.*

PoA	IPA	Word	Meaning
Velar	ɡ	ɡoʈa	Round stone/pebble
	ɡʰ	ɡʰoʈa	The ankle bone
Palatal	ʤ	ʤaɖI	Thickness
	ʤʰ	ʤʰaɖI	Thick wood
Retroflex	ɖ	ɖəg	Fear/danger
	ɖʰ	ɖʰəg	Cloud
Dental	d̪	d̪ar	Door/entrance
	d̪ʰ	d̪ʰar	Edge of a tool/ flow
Labial	b	bag	Garden
	bʰ	bʰag	A part

2.1 Training and testing datasets

The training database comprises Marathi spoken words sampled at 16 kHz. Two distinct meaningful words with word-initial plosives corresponding to each 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 plosive category appears in Table 2. The unaspirated affricates number more than the aspirated due to the allophonic variations recorded for each word.

Table 2. *Count of voiced plosives in train and test data sets*.

Plosives \ Data sets		Marathi train	Hindi native	Hindi non-native
Stops	Unaspirated	2560	1280	640
	Aspirated	2560	1280	640
	Total	5120	2560	1280
Affricates	Unaspirated	1280	320	160
	Aspirated	640	320	160
	Total	1920	640	320

For the pronunciation assessment evaluation, test datasets were recorded by 20 native Hindi speakers and 10 speakers of Tamil L1. Hindi and Tamil belong to distinct language groups that differ prominently in the plosive system. While both languages contain oral plosives of 5 places of articulation, voicing and aspiration are used distinctively only in Hindi. Tamil does not distinguish aspiration or even voicing; stops are voiceless and weakly aspirated in initial position, and voiced after nasals [11]. 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. We recorded speech from the native and non-native Hindi speakers in the form of read-out words containing the target phones in word-initial position across vowel contexts. Hindi has the same plosives and vowels as Marathi except that there is no allophone of the unaspirated affricate. The test dataset involved one meaningful word of Hindi corresponding to each consonant and vowel context embedded in 2 carrier phrases. Each dataset has an equal number of male and female speakers. A native Hindi judge 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 that phonemic aspiration was the main discriminating attribute. Voicing was always realised correctly even though voicing is allophonic in Tamil stops.

2.2 Speech production and acoustic characteristics

Voiced plosives are produced with glottal vibration in the closure region until the release burst. This is followed by the onset of the vowel (in non utterance-final positions). While vowel onset time (VOT) prominently differentiates unvoiced stops in aspiration, it is not so in voiced stops [4, 5]. However, similar to unvoiced aspirated stops, voiced aspirated stops are characterized by glottal abduction. In the latter, the abduction starts halfway through the closure and reaches its maximum at the burst release [12, 13]. 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. The increased glottal open quotient (OQ) in breathy voice in reflected in the amplitude of the first harmonic relative to that of the second (H1-H2) [14, 15]. The more gradual glottal closure is reflected in higher spectral tilt, and aspiration noise leads to increased aperiodicity following the vowel onset. These characteristics are evident in the spectrograms of Fig. 1 and 2 where relatively weak higher formants and aperiodicity are observed following the vowel onset in the case of the aspirated plosives. Different languages have been observed to have different durations of breathiness extending into the vowel [12].

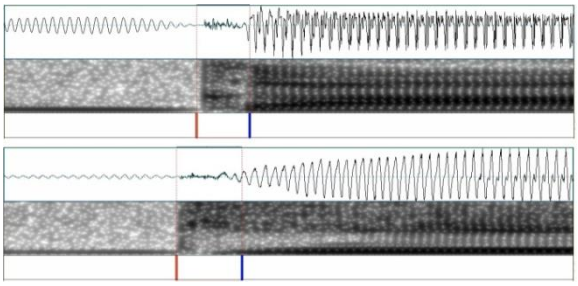

Figure 1: *Word-initial CVs over 250 ms of velar stops, /ga/ (top) and /gʰE/ (bottom).*

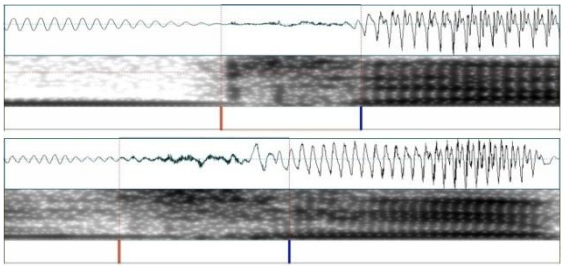

Figure 2: *Word-initial CVs over 180 ms of affricates /ʤa/ (top) and /ʤʰo/ (bottom).*

We note that VOT durations are comparable across the aspirated-unaspirated classes. In the case of the voiced affricates, an additional cue appears in the form of aspiration noise just before the vowel onset with a clear alteration of spectral structure from the more frequency-localised frication.

We see that 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 (e.g. by a particular speaker or vowel context). It is therefore best to consider multiple acoustic attributes for reliable detection across speakers.

3. Feature extraction

In the previous section, several acoustic parameters suggested themselves based on the physiological and acoustic properties of voiced plosives. We consider the computation of these parameters from recorded speech. As seen from Fig. 1, the parameters must be estimated in the vicinity of the burst and vowel onset landmarks. Thus the implementation involves suitable segmentation of the signal followed by the required spectral or temporal analyses. Acoustic landmark detection is discussed next followed by feature computation and selection.

3.1. Acoustic landmark detection

The temporal landmarks are located in the utterance by a two-staged procedure [16]. A coarse first stage of segmentation involves alignment of the broad class phonetic transcription using an available state-of-the-art MFCC-HMM ASR system. Since our focus is on discrimination within the class of voiced plosives, we assume the availability of a broad class phone transcription. In the recognition task, this may be available from a previous step in the hierarchical system, and in the pronunciation task, known from the intended words in the list. The broad classes are: vowels, sonorants, unvoiced fricatives, unvoiced affricates, unvoiced stops, voiced affricates, voiced stops, silence and voice bar. The acoustic models are context independent, 3-state HMM with diagonal covariance and flat-start initialization. The standard 39 dim MFCC, delta and acceleration feature vector was computed at 10 ms intervals. Broad class based recognition is robust since most confusions in a phone recognizer tend to be within the same manner class [17]. However the decoder segmentation is not accurate enough for the location of landmarks with the required precision, and is further refined follows.

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 [16, 18]. Cues to vowel onset are dependent on the nature of the consonant and especially difficult for aspirated and voiced stops. In the case of voiced plosives, 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].

3.2. Feature implementation

From Sec. 2.2, we see that the potential distinguishing properties of aspiration in voiced plosives are the duration (VOT), glottal OQ (H1-H2), spectral tilt and aspiration noise. The latter two can be measured from the signal in different ways giving rise to different acoustic features as presented next.

VOT is the duration between burst onset and vowel onset.

H1-H2 is the ratio of the first two harmonic amplitudes computed in the vowel region. H1-H2 and tilt measurements are obtained from magnitude spectra from 25 ms Hamming-windowed DFTs computed at 1 ms hop and averaged over a selected 5 ms duration.

Spectral tilt has been estimated by a number of different acoustic parameters in the context of voice quality detection. These include H1-A3 (A3=highest spectral amplitude in 3rd formant region) [14, 15], A1-A3 by Ishi [20] where fixed bands around average first and third formant regions are used, and H1-A2 by Cho [8]. All these measurements capture the rolling off of the spectrum from the low frequency band to the higher formant regions. We consider fixed bands around average formants like Ishi [20] to compute A1-A3 as the difference of the strongest spectral component in the range of 100 to 1000 Hz (F1 band) and the one in the range of 1800 to 4000 Hz (F3 band). The F2 region energy used as an indicator of aspiration in unvoiced stops by Cho [8] is also captured by de Krom's [21] 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 [2000-5000 Hz] takes on low values at increased tilt. The multiple measurements of spectral tilt were motivated by preliminary classification experiments which showed that each provided some additional information to the discrimination task.

Aspiration noise is measured by the ratio of harmonic energy to noise energy. An "SNR" feature 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 [22]. 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 [20], 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.

Since the breathiness of the vowel arises from the coarticulation with the preceding aspirated stop, it is important to select the analysis region suitably for the breathy voice features. We tested the discriminability of the various features as a function of the analysis interval location from vowel onset to midpoint, chosen based on average vowel duration. Fig. 3 shows the variation of the average separability as measured from the overlap of the distributions across aspirated/unaspirated classes for each of the two feature sets: Spectral shape (H1-H2, spectral tilt), and aspiration noise (SNR, F1-F3 sync). We observe that the spectral shape based differences are strongest near vowel onset even though non-modal voice quality is known to extend over most of the vowel duration [13]. On the other hand, noise measures are most discriminative nearer to the vowel midpoint. The reduced effect of the preceding unaspirated plosive (especially affricate) may well be a contributing factor to this. Accordingly we select the analysis interval for the spectral shape features to be around the instant 13 ms from vowel onset, and that for the noise features to be at 23 ms.

Further, in the case of affricates, aspiration noise is also conspicuous in the burst region as seen in Fig 2. This can be estimated by E1-E2 where E1 = E[3000:7000] i.e. energy in the range of 3000-7000 Hz and E2 = E[60:3000] i.e. energy in

the range of 60-3000 Hz. When aspiration noise appears in the otherwise fricated region, E2 increases causing E1-E2 to decrease. The energies are obtained by averaging 6 ms windowed magnitude spectra at 1 ms hop over the 10-20 ms duration *prior* to the vowel onset.

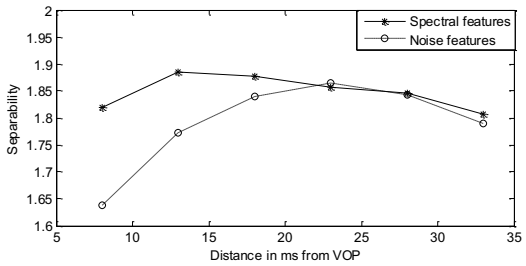

Figure 3: *Variation in separability with analysis time instant measured from vowel onset*

4. Classification experiments

Two-way classification experiments are carried out using the speech datasets of Sec. 2.1 to evaluate the performance of aspiration detection. Features as described in the previous section are computed on the automatically segmented speech. Landmark detection errors with respect to manual labeling were noted to be within 10% for a 15 ms tolerance for the burst onset, and within 20% for the vowel onset for the same tolerance, with affricates performing worse than the stops on average. A GMM (6 mixtures, full covariance) is trained on the feature vectors of each class: voiced unaspirated and voiced aspirated, separately for stops (7 features) and affricates (8 features).

4.1. Two-way classification of Marathi plosives

Table 3 shows the accuracies obtained in a 20-fold (leave-one-speaker-out) cross-validation classification experiment on the Marathi speech database. Various feature sets are tested as explained in the caption. We note that including the additional measurements of spectral tilt (over the basic A1-A3 feature) leads to significant increase in accuracy. A further accuracy increase is obtained upon including both the noise features. Finally, we add the pre-onset noise feature (E1-E2) to the affricate feature set to obtain the accuracies shown in parentheses. This feature did not enhance the performance on stops but, as expected, provided a useful cue to aspiration in voiced affricates.

Table 3. *Accuracy (%) achieved using various feature sets with automatic and manual labeling.* ***S1****= VOT, H1-H2, A1-A3;* ***S2*** *= S1,low-band-slope, B3 band energy;* ***S3*** *= S2, SNR, F1-F3-sync. Parenthesized: E1-E2 added. All differences are significant (p<0.01)*

Class of plosive	**Automatic labels**			**Manual labels**
	S1	**S2**	**S3**	**S3**
Stops	83.5	84.5	85.1	86.2
Affricates	74.5	76.9	77.6 (79.2)	78.0 (79.0)

From Table 3, we also note that while manual landmark detection provides better accuracies over automatic labeling, this is not true for the affricate class. This stems from the inherent difficulty of precise manual annotation of vowel onset instant in voiced aspirated affricates. Next, with the best feature set of each class, we trained acoustic models for the aspirated-unaspirated distinction of stops, and of affricates, over the full 20-speaker Marathi data. These models were used to classify the Hindi voiced plosives of the 20 native speaker set of Table 2 using automatic segmentation. The best feature set classification accuracy for stops was 84.9, and that for affricates 79.2. This shows that the acoustic features generalize well for cross-language transfer, at least between Marathi and Hindi with their shared set of phones.

4.2 Non-native pronunciation detection

Using the Hindi test speech, we compute the percentage of instances that the target is correctly achieved for each speaker as a measure of speaker's "nativeness". Fig. 4 shows the obtained %correct for each speaker for the proposed system and, for comparison, with an MFCC-HMM system trained similarly for the same 2-class task. We see that the %correct varies across speakers with the non-native speakers' group doing worse overall. Compared to the MFCC features, the proposed features show higher accuracy on the native speakers' data as well as better separation between native (N) and non-native (NN) speakers.

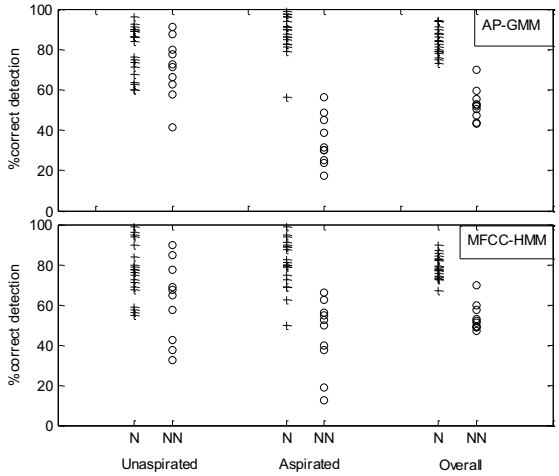

Figure 4: *Percentage correct achieved target of voiced plosives in native (N,+) and non-native (NN,o) datasets.*

5. Conclusion

Phonologically motivated features are investigated for the detection of phonemic aspiration in Marathi and Hindi voiced plosives. Multiple acoustic attributes and features contribute towards robust detection, in line with the research goals of ASAT for speech recognition [23]. Experiments on non-native speech confirm previously reported findings that features based on phonetic properties are superior to the standard ASR system features for pronunciation error detection [2]. Future work will be focused on larger datasets and other Indo-Aryan language including Gujarati, marked also by occurrence of phonemic breathiness in vowels [24].

6. Acknowledgement

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

7. References

[1] Niyogi, P., Burges, C. and Ramesh, P., "Distinctive feature detection using support vector machines", In: Proc. ICASSP-1999, Pheonix, Arizona, pp 425–428. March 1999.

[2] Strik, H., Truong, K., Wet, F. and Cucchiarini C., "Comparing classifiers for pronunciation error detection", In: Proc. of Interspeech 2007, Antwerp, Belgium, pp. 1837-1840, Aug. 2007.

[3] Halberstadt, A. and Glass, J., "Heterogeneous measurements and multiple classifiers for speech recognition", In: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP-1998), Sydney, Australia, pp. 995–998, December 1998.

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

[5] 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

[6] Mikuteit, S., "Voice and aspiration in German and East Bengali stops: A cross-language study", In: Proc. of Interspeech 2005, Lisbon, Portugal, pp. 2873-2876, Sep. 2005.

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

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

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

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

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

[12] Ladefoged P. and Maddieson I., "The Sounds of World's Languages", Blackwell, 2005.

[13] 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.

[14] Klatt, D. H. and Klatt, L.C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am. 87(2), pp. 920-857, Feb. 1990.

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

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

[17] 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.

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

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

[21] Krom, G., "Spectral correlates of breathiness and roughness for different types of vowel fragments", In: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP'94), Yokohama, Japan, pp. 1471-1474, Sep. 1994.

[22] 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.

[23] Lee C.-H., Clements M. A., Dusan S., Fosler-Lussier E., Johnson K., Juang B.-H., and Rabiner L.R., "An overview on automatic speech attribute transcription (ASAT)", In: Proc. of Interspeech 2007, Antwerp, Belgium, pp. 1825-1828, Aug. 2007.

[24] Bhaskararao P., "Salient phonetic features of Indian languages", Sadhana, 36(5), pp. 587-599, Oct. 2011.