Proc. of the National Conference on Communications (NCC), Jan 2011, Bangalore, India

Acoustic Features for Detection of Aspirated Stops

Vaishali Patil, Preeti Rao
Department of Electrical Engineering,
Indian Institute of Technology Bombay,
Mumbai, India.
{vvpatil, prao}@ee.iitb.ac.in

***Abstract*—Aspiration is an important phonemic feature in several Indian languages. Unlike English, languages such as Marathi have lexicons in which words with different meanings differ only in the aspiration feature of the initial voiced or unvoiced stop. Thus the reliable discrimination of aspirated stops from their unaspirated counterparts is important in automatic speech recognition for such languages. The important acoustic distinctions include durational features as well as fine spectral structure features. Traditional frame-based spectral representations such as MFCCs used in HMM-based recognizers do not explicitly encode these cues. In this work, we explore various acoustic features for aspiration detection in voiced and unvoiced stops of Marathi. Enhancements to available methods of aspiration detection borrowed from voice quality measures are found to provide improved detection of phonemic aspiration in stops. The performance of a landmark-based acoustic feature classifier is compared with MFCC-HMM baseline system for the recognition of aspirated and unaspirated stops.**

Keywords- aspiration, acoustic features, MFCC

I. INTRODUCTION

The phonological structure of stops in many Indian languages (Hindi, Marathi, Gujarati etc.) is such that they exist as aspirated-unaspirated pairs corresponding to a particular place of articulation (PoA) for both the voicing manners i.e. voiced and voiceless. In these languages, aspiration is used along with voicing element to form four categories of stop consonants for each PoA. Thus the aspiration attribute of a stop is necessary for its complete description. The current investigation is based on voiced and voiceless stops in Marathi, the language of Maharashtra state. Several examples of words can be cited where the word initial stop differs only in aspiration but the resulting two words have different meaning. For e.g. /ʈa ʈ/ and /ʈʰa ʈ/, is a pair of words from Marathi with unvoiced stops in word initial which differ only in aspiration feature but form two words which mean 'plate' and 'magnificence', respectively. /b a g/ and /bʰ a g/ is a similar such pair with voiced stops in word initial. Thus distinction of aspirated-unaspirated stops is important in phone recognition application of these languages. Such a facility could also be useful for the detection of pronunciation errors for non-native learners of spoken Marathi who may not have been exposed to the aspiration-based distinction in their native tongue.

In standard statistical approach of HMM recognizer conventional MFCC based features are computed frame wise indicating uniform mode of processing over the whole speech signal irrespective of the signal variations over different phone utterances which are not taken into account. The current investigation is based on the acoustic-phonetic approach of knowledge-based systems, using explicit knowledge about speech production in the recognition process. We focus on extraction of acoustic-phonetic features for detection of aspiration in stops. Though not adequate in isolation, a durational parameter measured from burst release to onset of voicing has been suggested as prominent cue for aspiration manner distinction of stops from Gujarati and Hindi languages which have similar phonemic structure as Marathi [1][2].

As the aspirated-unaspirated pairs of stops are characterized by same PoA they have similar acoustic properties of the transient phase and the frication burst phase that occur at the release. However in case of the aspirated stops, before the onset of glottal vibration associated with following vowel, the glottis is abducted initially and then the vocal folds are brought together to produce modal voicing. Hence the turbulence noise is seen to be followed by an interval of breathy voicing, which is further followed by modal vowel. This portion of following vowel with simultaneous aspiration is described as "superimposed aspiration" and is characterized by periodic voicing with noise excitation overlaid, imparting breathy or muffled voice quality to the corresponding voiced segment[3][4]. While the above mentioned attributes due to superimposed aspiration in case of aspirated stops are observed in most tokens of these consonants, it should be noted that there may be considerable variability in individual tokens depending on PoA of that stop along with vowel context and speaker, thus posing a challenge.

Presently we investigate the task of detecting aspirated stops as an aspirated-unaspirated distinction task, a 2-class problem where the performance of conventional MFCCs is compared with that of the acoustic feature set. We also check the contribution of features extracted on the basis of superimposed imposed aspiration for this task when added to the previously known durational parameter. The next section describes the baseline and proposed systems including the implementation of features used in the classification. Following this experiments for the evaluation of the system are presented. Finally conclusions are drawn about the relative performances of baseline and acoustic feature based recognition.

II. ASPIRATED/UNASPIRATED STOP IDENTIFICATION

The task undertaken is the distinction of aspirated-unaspirated stops in Indian languages focusing on Marathi. The features explored are the traditional MFCCs and the acoustic features which comprises of the durational measure as well as aspiration detecting parameters. We compare the performance of MFCC feature vector in an HMM based recognizer applied to the task of detecting aspirated stops with that of the acoustic feature set in a suitable classifier framework.

A. Baseline system

The HTK toolkit is used where HMM based recognizer is operated in force alignment mode [5]. The standard 39 dim MFCC, delta and acceleration feature vector is computed for the 16 kHz sampled signals at 10ms intervals. Mel scaled 26 filter banks spanning the 8 kHz frequency range are used for computation of MFCCs. Acoustic models are generated for 11 broad classes listed in Table I below. All broad class models were context independent monophones with 3-state HMMs. 8 Gaussian mixtures (diagonal covariance) per state were trained with flat-start initialization.

As the current focus is on distinguishing aspirated and unaspirated stops, log-likelihood values are obtained for both types of phones in forced alignment mode. The aspiration manner resulting in higher value of log-likelihood is assigned to the respective token. Further difference of the two log-likelihoods gives a distance metric which is an indicator of the separability between aspirated-unaspirated stops achieved using the MFCC feature sequence.

TABLE I. BROAD CLASS GROUPING FOR GENERATING ACOUSTIC MODELS USING HTK

Stops categories	**Others**
Unvoiced-unaspirated	Affricates
Unvoiced-aspirated	Fricatives
Voiced-unaspirated	Vowels
Voiced-aspirated	Semivowels
	Nasals
	Voice bar
	Silence

B. Acoustic features based classification

Acoustic features are computed on the signal corresponding to a stop as determined by the acoustic landmarks of burst release and voicing onset. The acoustic features investigated can be divided into two categories viz. the one which includes durational measure while the other which differentiates these two stop classes on the basis of presence or absence of aspiration noise. The perceptual correlate of aspiration noise is breathy or muffled voice quality in the voicing region of the following vowel. Hence several voice quality measures are potential candidates for the detection of phonemic aspiration.

1) Durational measures: A durational measure of "voicing onset time", (VOT), corresponding to the duration between the burst release and onset of vibration of vocal folds is known to be a prominent cue to the distinction of aspirated-unaspirated stops with the same PoA. In a study of Gujarati velar stops, a significant distinction was observed in VOT across the aspirated-unaspirated pair of unvoiced stops [1]. However it was not seen in case of voiced stops [1]. The study of durational characteristics of stops in Hindi indicated that the duration of release segment (measured from burst onset to onset of the following vowel) is systematically affected due to aspiration. The "release duration", of aspirated stops is shown to be longer than that of unaspirated stops for a particular PoA in case of both unvoiced and voiced manners. The high degree of variability in this parameter value even for a particular PoA is seen through the large values of their standard deviations. The speaking rate is further expected to influence it in continuous speech [2]. Also, in the study of classification of stops, "release duration" serves as one of the cues to the identification of the PoA of stops [6]. Thus the variations in this timing parameter across different PoA will tend to reduce the differentiating pattern for aspirated-unaspirated distinction when observed across PoA. This indicates that though the durational parameter may differentiate the stops on the basis of aspiration manner, there is scope for improvement with additional features that may capture the presence of aspiration.

2) Features based on aspiration noise: As the aspirated-unaspirated pair of stops is characterized by same PoA they have similar acoustic attributes of the transient and frication burst phase that occur at the release. However in case of the aspirated stops turbulence noise is followed by an interval of breathy voicing, which is further followed by modal vowel. This portion of following vowel with simultaneous aspiration has been suggested to characterize the affricates as aspirated-unaspirated by presence or absence of "superimposed aspiration" respectively in the study of Nepali language affricates [3]. This concept of "superimposed aspiration" is also suggested by K. Stevens [4] where the aspiration in "/ha/" is considered. A study of Korean stops also points out that voice quality of the following vowel is capable of differentiating the three Korean stops viz. lenis, fortis and aspirated [7]. Thus, analysis of the initial part of following vowel is suggested to give an index or measure of aspiration.

The extraction of aspiration related spectral features, as presented below, involves the frequency region around the third formant. The formant estimate is taken from the output of an automatic formant tracker developed in our lab that provides formant values for each frame from vowel onset to vowel steady state. The automatic formant tracker uses linear prediction based root solving for generation of formant candidates. Dynamic programming followed by post processing is further used to obtain a reliable formant track.

a) Spectral features (H1-H2, H1-A3 and A1-A3): The effect of superimposed aspiration in a vowel region is perceived as breathy voice. Hence features identifying breathy voice quality imparted due to aspiration noise have been investigated in the past literature [3][7]. Difference

between the first and second harmonic (H1-H2) and spectral tilt (H1-A3) are the two prominent features considered to be indicators of breathiness [8]. These acoustic correlates of breathiness can be employed for analysis of voice quality of the following vowel to work indirectly as indicators for presence or absence of aspiration [3][7]. In the present work these features are computed for the 5 frames with window centre shifted over 6 to 10 ms from the manually labeled instant of vowel onset point (VOP) and then averaged. Computation of H1-A3 uses third formant values for these 5 frames as estimated by the automatic formant tracking algorithm. Another spectral tilt indicator (A1-A3) is computed as difference between strongest component in the range of 100 to 1000Hz and the one in the range of 1800 to 4000Hz thus avoiding the need of precise formant value extraction [9].

b) Synchronization index (f1f3sync): Presence of aspiration noise in the high frequency region around third formant is an acoustic correlate of breathiness due to the tendency of the weak higher harmonics to be dominated by aspiration noise. This was evident through visual observation of band pass filtered waveform around the third formant [10]. To get a corresponding automatic measure in the work on voice quality Ishi [9] has proposed to compute a synchronization index by obtaining correlation of signal components in the fixed frequency regions of 100-1500 Hz and 1800-4000 Hz. However in our work the feature is computed by using the estimated formant regions, F1 and F3 bands (600 Hz around the formant values) automatically obtained for the vowel of that CV token.

c) Sub-band spectral power and sub-band slope: G. Krom [11] has investigated some spectral parameters to serve as relevant predictors of breathiness and roughness in voice. These parameters are spectral levels, spectral slopes and harmonic to noise ratios in four frequency bands. These four bands are specified as band0 (60 to 400 Hz), band1 (400 to 2000 Hz), band2 (2000 to 5000 Hz) and band3 (5000 to 8000 Hz). Spectral slopes are calculated as level differences between these frequency bands [11]. An additional feature cumulative slope is computed as level difference between the spectral level of band3 and band0.

d) Signal to noise ratio (SNR): Cepstral based Harmonic-to-Noise (HNR) computation, proposed by Murphy and Akande [12] for voice quality studies, avoids the need for precise computation of harmonic strength from the spectrum. It uses the cepstrum with window duration required to be pitch dependent and greater than than 60 ms for lower values of male pitch (around 80 Hz) [12]. However for the current investigation of noise component in a vowel due to the preceding aspirated stop, the impact persists for a smaller duration which would affect the accuracy of HNR estimation. Hence we choose to compute "signal-to-noise ratio" (SNR), using 25 ms analysis window where noise power estimation is done using cepstral liftering and signal power is obtained from the regular FFT spectrum.

The potential of each acoustic feature to differentiate the aspirated-unaspirated stops was judged initially through the ANOVA distribution plots separately for voiced and voiceless manners. It was seen that different features exhibit varying differentiating ability across these two classes of stops. Those features are chosen with higher values of F-ratio obtained from the ANOVA distribution plot as they are the indicators of better discriminating ability of that feature. For the unvoiced stops aspiration detecting parameters used are H1-H2, SNR and B1-band energy while H1-H2, A1-A3 and f1f3_sync are the ones used in case of voiced stops. Variation in the speakers forming the data sets of voiced and unvoiced stops in addition to the high degree of inconsistency in the discriminating ability of these features can be accounted for this. The task of detection of aspirated stops is dealt with as two separate classification tasks, i.e. distinction in case of stops for each of the voicing manners.

A GMM classifier is used where the acoustic models are obtained by training using EM algorithm with number of mixtures varying from 1 to 7. The best resulting accuracies are reported. In order to evaluate the distinguishing ability of the acoustic features, classification of aspirated-unaspirated stops was done using two sets of acoustic features. In set 1 only the durational parameter of "release duration" is included while set 2 comprises of the breathiness (or aspiration) detecting measures in addition to the durational parameter of set 1. The same classification experiments are carried using features of set 1 and set 2 for comparison.

III. EXPERIMENTS

A. Database details

In the current investigation the experiments are conducted on words data of voiced and unvoiced stops in Marathi, one of the prominent Indian languages. It has very similar pattern of stops as that of Hindi for both voicing manners as indicated in Table II [13].

TABLE II. IPA CHART SHOWING MANNER AND POA OF STOPS IN MARATHI [13].

PoA / MoA	Labial	Dental	Retroflex	Velar
Voiced stops	b bʰ	d̪ d̪ʰ	ɖ ɖʰ	g gʰ
Unvoiced stops	p pʰ	t̪ t̪ʰ	ʈ ʈʰ	k kʰ

For the database of voiced and unvoiced stops, words with CVs in initial position are formed in combination with 8 vowel contexts. The eight vowels used are /ə/, /a/, /i/, /I/, /u/, /U/, /e/, and /o/. The words are uttered in two carrier phase sentences, one sentence and other a question. The utterance of unvoiced aspirated stop 'pʰ', failed to exhibit the expected closure by all the speakers and hence is not included in the current analysis. The data of unvoiced stops comprises of total 2240 stops from recordings of 10 speakers (5M, 5F). The dataset of voiced stops includes 1024 stop tokens

recorded from 4 speakers (2M, 2F). It was ensured that Marathi is the mother tongue of all speakers so as to avoid any misarticulation. The word initial CV of interest is manually labeled to mark the burst onset of 'C' and onset of the following vowel 'V', with the help of PRAAT waveform editor [14]. Manual labeling is preferred so as to extract the durational features precisely and also chose appropriate region of analysis. This data set was created so as to have equal count of tokens from each class of aspirated-unaspirated stops with all vowel contexts to facilitate appropriate evaluation of the distinguishing ability of the features extracted for aspiration detection task through classification experiments.

B. Evaluation

To compare the separability achieved by the two feature sets (MFCCs and acoustic features) they should be represented on the same dimension. Likelihood ratio discriminability measure is hence computed using the following equation 1 [15].

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

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). The distribution of $d(x)$ for class1 and class 2 can characterize the separability. Likelihood values using MFCC feature vector are extracted from the HMM based recognizer while those using acoustic feature set are obtained from Gaussian probability models of the GMM classifier. The following plot in Fig. 1, gives distribution of the discriminability measure for unvoiced stops.

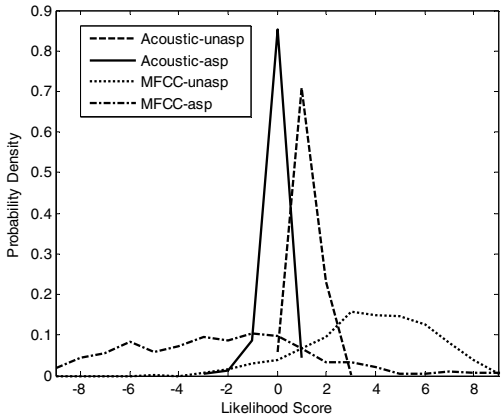

Figure 1. Separability of aspirated-unaspirated unvoiced stops using probability models constructed from MFCCs and acoustic features

As seen the distribution plots corresponding to MFCC models show greater overlap as compared to that between the distributions constructed from the acoustic parameters. This suggests that acoustic measures will be better able to differentiate the aspirated-unaspirated classes of unvoiced stops. Similar pattern is also seen in case of the two classes of voiced stops as seen in plots of Fig. 2.

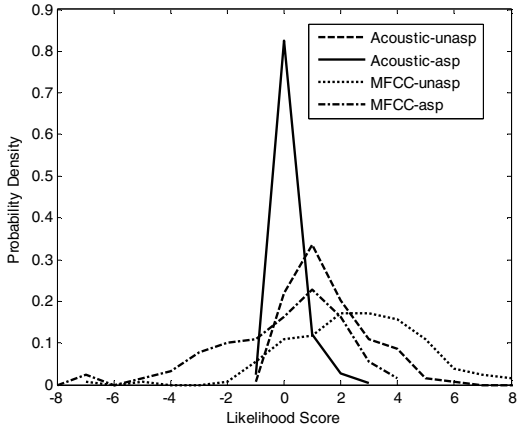

Figure 2. Separability of aspirated-unaspirated voiced stops using probability models constructed from MFCCs and acoustic features

IV. RESULTS AND DISCUSSION

The detection experiments are performed using partial data as train set and remaining data as test set. For unvoiced stops, data of 7 speakers (speaker 1-7) forms the train set while speakers 8-10 are included in the test set. In case of voiced stops data of speakers 2-4 is the train set while testing is done on data of speaker 1. The acoustic features form the feature vector for a GMM based classifier, while the MFCC based feature vector is plugged in HTK operating in force alignment mode. The role of acoustic features is tested using the two feature sets where only the durational measure is the feature in set 1 while in set 2 the breathiness detecting features are added to it. Significance test is carried out to check the nature of improvement by comparing the classifier outputs available using features of set 1 and set 2 for a particular classification task. The corresponding classification results are indicated in Table III given below.

TABLE III. COMPARISON OF CLASSIFICATION ACCURACIES OBTAINED USING MFCC AND ACOUSTIC MEASURES

Features / **Task of classification**	**MFCC**	**Acoustic features**		
		Set 1	**Set 2**	**p-value**
Unvoiced stops	86.6 %	88.4 %	92.3 %	p<0.001
Voiced stops	67.6 %	56.3 %	73.4 %	p<0.001

The performance of the acoustic features is seen to be consistently higher than that of the conventional MFCCs for both voicing manners by around 6%. Also percentage classification accuracy is seen to improve on using the features of set 2. For unvoiced stops the improvement in case of set 2 over that for set 1 is of 4% while for voiced data it is around 17 %. The classification accuracies as seen in Table III indicate that the performance of detecting aspirated stops using acoustic features is higher in case of unvoiced stops as compared to that in case of voiced ones. Also the p-values available from the respective significance tests comparing the performance for set 1 and set 2 are very small (p<0.001)

indicating that the inclusion of aspiration noise detection improves performance significantly.

Superior separation between the aspirated-unaspirated stops is seen to be achieved using the acoustic measures as compared to those possible using MFCCs, with significantly less computational complexity. Also, the dimension of MFCC feature vector is 39 while only 4 acoustic features chosen in the current task are capable of better detecting the aspirated stops. The HMM models do encode the duration feature but not explicitly. Lower accuracy for acoustic feature set 1 comprising of only timing parameter and improvement achieved by including the aspiration detecting features in set 2 indicates the importance of the latter features in this task.

Though not insufficient, the data of voiced stops needs to be increased to validate the current results for speaker variations across more number of speakers. Similar such investigation can be extended to the whole class of plosives which includes affricates. The acoustic cues to aspiration investigated in this work evolve from the characteristics of voice quality which may be also influenced by additional effects like jitter and shimmer. Glottal-to-noise excitation (GNE) ratio suggested to be a measure for amount of noise in voicing region which is independent of these influences should be tested for this task [16]. Further, the detection of aspiration noise, that has been restricted to the vowel region so far, can be extended to the burst release in order to explore additional cues for aspiration.

As the acoustic features rely on parameters derived from typical regions defined by the burst onset and vowel onset landmarks, future work needs to be directed to the reliable automatic detection of these landmarks. HMM-based phonetic segmentation using MFCC feature vectors can give initial estimates of the required landmarks which can be relocated using the 2-stage approach proposed in our earlier work [17]. Thus an appropriate blend of HMM frame work with acoustic phonetic approach can help in the tasks of phone recognition and improve accuracy in related tasks such as pronunciation error detection.

REFERENCES

[1] M. K. Rami, J. Kalinowski, A. Stuart and M. P. Rastatter, "Voice onset times and burst frequencies of four velar stop consonants in Gujarati", *J. Acoust. Soc. Am.,* vol. 106, no. 6, pp. 3736-3738, Dec. 1999.

[2] K. Samudravijaya, "Durational characteristics of Hindi stop consonants", Eurospeech 2003, pp. 81-84, Sep 2003.

[3] G. N. Clements and R. Khatiwada, "Phonetic realization of contrastively aspirated affricates in Nepali", *Proc. ICPhS XVI*, pp. 629-632, Aug. 20 07.

[4] K. N. Stevens, "Acoustic phonetics", Cambridge, MIT Press, 1999.

[5] Young S. et al., "The HTK Book v3.4", Cambridge University, 2006.

[6] V. Karjigi and P. Rao, "Landmark based recognition of stops: acoustic attributes versus smoothed spectra", *Proc. ICSLP*, pp. 1550-1553, Sep. 2008.

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

[8] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", *J. Acoust. Soc. Am.,* vol. 101, no. 1, pp. 466-481, Jan. 1997.

[9] C. T. Ishi, "Anew acoustic measure for aspiration noise detection", *Proc. ICSLP 2004*, pp. 629-632, Aug. 2004.

[10] D. H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am.,* vol. 87, no. 2, pp. 920-857, Feb. 1990.

[11] G. Krom, "Spectral correlates of breathiness and roughness for different types of vowel fragments", *Proc. ICSLP 1994*, pp. 1471-1474, Sep. 1994.

[12] P. J. Murphy and O. O. Akande, "Noise estimation in voice signals using short-term cepstral analysis," *J. Acoust. Soc. Am.,* vol. 121, no. 3, pp. 1679-1690, Mar 2007.

[13] "Marathi language", http://en.wikipedia.org/wiki/Marathi

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.3.01) [Computer program]", Retrieved from http://www.praat.org/ , 2005.

[15] P. Niyogi and P. Ramesh, "The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets", *Speech Communication*, vol. 41, pp. 349-367, 2003.

[16] D. Michaelis, M. Frohlic, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices", *J. Acoust. Soc. Am*., vol. 103, no. 3, pp. 1628-1639, Mar. 1998.

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