Proc. of Frontiers of research on Speech and Music (FRSM), January 2007, Mysore, India

INVESTIGATION OF ACOUSTIC ATTRIBUTES OF MARATHI UNVOICED STOPS FOR CLASSIFICATION

Veena Karjigi, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, INDIA 400076
Email: {veena, prao}@ee.iitb.ac.in

Samudravijaya K.
Tata Institute of Fundamental Research
Homi Bhabha Road
Mumbai, INDIA 400005
Email: chief@tifr.res.in

Abstract

Close but contrasting phonemes in a language can be distinguished from each other in terms of the linguistically motivated distinctive features. Distinctive features are characterized by corresponding acoustic correlates in speech signal. Studies of the acoustic correlates of distinctive features of phones of a language are valuable from both, the scientific and technological points of view. In this work, acoustic attributes for classification of unvoiced, unaspirated stops of Marathi which differ in their place of articulation are investigated. Three features were derived from the burst spectrum computed from a set of VCV syllables with three vowel contexts to distinguish these stops and were evaluated for their discriminating ability.

1. Introduction

One reason for the unsatisfactory performance of current generation statistical speech recognition systems is the inferior modeling of phonetic level linguistic information by the standard frame-based spectral features. More accurate phonetic characterization of the speech signal is possible by obtaining acousto-phonetic features directly related to the underlying distinctive articulatory features of the phone such as voicing, manner and place of articulation. Phonemes are thus treated as complexes of distinctive features to be extracted by speech signal processing, and then used of recognition.

Stop consonants represent one of the broad categories of phones in all major languages. The production of a stop involves a complete closure of the oral cavity followed by release in the form of noise burst. The stop consonants are differentiated from each other in terms of the manner of articulation (whether voiced and/or aspirated) and the place of articulation. During the release of the closure, the part of the vocal tract downstream from the constriction is excited by the sudden reduction in the intra-oral pressure. Due to this, formants shift rapidly either upwards or downwards in frequency as well as amplitude depending on the place of constriction of the consonant and the following vowel. Hence stops can be classified based on the place of articulation by studying the spectral distribution of the burst and formant transitions from burst onset to the steady state of following vowel. Because of their short duration, classification of stops is a challenging problem. Finding acoustic attributes for the classification of the stops based on place of articulation has been the subject of active research. Past research results, reviewed next, have found that acoustic attributes extracted from the burst spectrum are more useful in the classification of unvoiced stops as compared to formant transitions.

In this work, we focus on the class of unvoiced, unaspirated stops of Marathi. Marathi contains four stops in this category corresponding to four distinct places of articulation (PoA) which provide contrasts in similar phonetic environments. These are labial, dental, retroflex and velar. English, on the other hand, contains only labial, alveolar and velar stop consonants. Acoustic features for PoA classification for the Marathi unvoiced stops are investigated in this work. The evaluation of the acoustic attributes is carried out by statistical measures over a data set spanning utterances of several speakers representing typical phonetic contexts.

The next section gives a brief summary of the previous research results on unvoiced stop classification in English followed by a review of work on extending these to other languages. The proposed features and evaluation experiment are discussed in Section 3. Results are reported in Section 4 and in Section 5, future work is discussed.

2. Previous Work

Early work [1] using burst spectra showed that labials and alveolars had diffuse spectra as compared with the peaked spectra of velars. The labials are distinguished from alveolars by the location of major energy concentration: low frequencies (500-1500 Hz) for labials and high frequencies (greater than 4000 Hz) for alveolars. In addition to this, alveolars showed peaks around 500 Hz. Winitz et al.[2] conducted perceptual tests with only burst portion as stimuli and the performance was improved when stimuli had both burst and 100 ms of the adjacent vowel. Blumstein and Stevens [3] studied the spectrum of the burst, and proposed three templates, diffuse flat or falling for labials, diffuse rising for alveolars and compact for velars. Kewley-Port [4] proposed time varying features derived from running spectra based on LP and 1/3rd octave filters in lieu of static spectra of Blumstein and Stevens [3]. This was done for voiced stops but further Lahiri et al. [5] used the concept of running spectra for both voiced and unvoiced stops. Suchato [6] used the average power spectrum for measuring acoustic attributes related to the burst and onset spectrum of American English stops. There has been very limited amount of work on extending the above acoustic attributes to the classification of stops of Indian languages. Many Indian languages have more than three places of articulation for unvoiced stops in their phonetic inventory. Considering their contrastive role in language, it is of interest to investigate the acoustic correlates of these different places of articulation.

Based on phonological theory, which groups dental and alveolar stops in the same category to be distinguished from labial stops, Lahiri et al [5] investigated the known acoustic attributes [3] for the labial, dental and alveolar stops of Malayalam. However, it was found that the dental stops could not be reliably separated from labials based on burst spectra alone. A new measure based on the change in the distribution of spectral energy concentration from burst onset to voicing onset was found to perform better in terms of grouping the Malayalam dentals with alveolars. Also mentioned is the observation that dentals and alveolars may further be distinguished from each other based on the total energy of the burst compared to that at the onset of voicing. A similar observation is made by Verma and Chawla [7] in the context of Hindi retroflex and dental stop consonants.

3. Experiment

Three features were investigated for the classification of stops /p/, /t/, /th/ and /k/ based on the works cited earlier. Features were derived from the static burst spectrum with a small set of database.

3.1 Database

VCV syllables were recorded with three vowels /a/, /i/ and /u/ and four unvoiced, unaspirated consonants /p/, /t/, /th/ and /k/ from three male and two female speakers by embedding them in the carrier sentence "Please say VCV again" at a sampling rate of 16000 Hz. Even though the effect of preceding vowel is not expected to be significant, 9 VCV syllables (with all possible combinations of three vowels in two positions) were collected from every speaker for each stop consonant to increase the number of VCV syllables. This led to 45 tokens for each stop consonant.

3.2 Analysis procedure

The time locations of the burst and voicing onset were detected manually by looking at the waveform, spectrogram and listening to the utterance. To compute the burst spectrum, an interval of 26 ms from the burst onset or the complete duration of the burst, whichever was smaller, along with 2 ms of the waveform each on either side was Hamming windowed. That is, either a 30 ms window was centered at 13 ms from the burst onset or a window of size equal to the burst duration plus 4 ms was centered at the middle of the burst. The voicing onset spectrum was computed from an interval starting 2 ms before the voicing onset and extending to 28 ms after using a 30 ms Hamming window. LP spectra of LP orders 14 and 18 were obtained for burst and voicing onset respectively by using autocorrelation method.

The burst and voicing onset spectra obtained from four VCV syllables, uttered by a single speaker, corresponding to four places of articulation with the vowel /a/ at both the initial and final positions are shown in Fig.1.

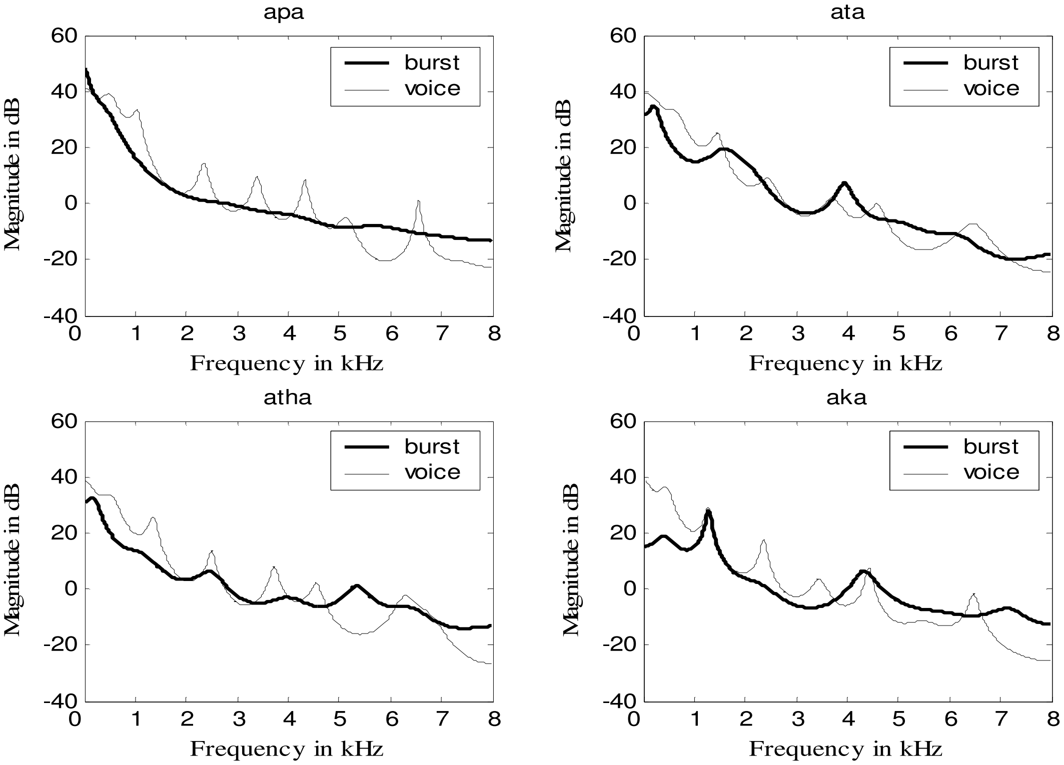

Fig. 1: Examples of burst and voicing onset spectra for VCV syllables corresponding to four places of articulation with the vowel /a/ at both the initial and final positions. The thick line represents the burst spectrum and the thin line represents the spectrum at the voicing onset

We see that the burst spectrum for /p/ decreases monotonically with a higher rate in the low frequency region (0-750 Hz) compared to the rest of the frequency band. Spectra for /t/ and /th/ exhibit diffuse flat or falling characteristics until around 4 kHz and /t/ burst shows falling spectra while /th/ burst shows flat spectra after 4 kHz. Burst spectrum for /k/ shows compact peak near F2 of the following vowel. /k/ preceding the back vowels, showed one more prominent peak in the high frequency region.

It may be mentioned here that spectra for retroflex and dental consonants differ from the spectrum of alveolar consonant /t/ of English which exhibits diffuse rising spectrum. Spectra of the /t/ burst extracted from English sentences spoken by an Indian speaker did not show diffuse rising spectra indicating that for the Indian speaker the articulation of /t/ of English is similar to the retroflex /t/ of Indian languages [8]. In the same vein, the labio-dental fricative of English is realized as the dental plosive /th/ in Indic languages [8].

3.3 Attributes selected

Since the place of articulation for retroflex and dental consonants are close to each other, first we tried to perform a three way classification by putting /t/ and /th/ in the same group which was followed by an attempt to separate /t/ from /th/.

A. Three way distinction:

Two features were derived from the burst spectrum to classify the four stops into three groups: labials in the first, retroflex and dentals in the second and the velars in the third group.

Feature 1: Because of the low energy of the labial spectra in the frequency band 750-7500 Hz compared to retroflex and dental stops which in turn was less than that of velars, the energy difference in the two frequency bands 0-750 Hz and 750-7500 Hz specified in Eq. 1 was tested for the three way distinction.

$$E_{lh}(dB) = 10\log(\frac{1}{hi1-lo1}\sum_{lo1}^{hi1}A^2(f)) - 10\log(\frac{1}{hi2-lo2}\sum_{lo2}^{hi2}A^2(f)) \text{ ---- (1)}$$

where A (f) is the magnitude of the burst spectrum as a function of frequency (f) in Hz.
lo1 and hi1 are 750 and 7500 Hz respectively
and lo2 and hi2 are 0 and 750 Hz respectively
This was expected to be large for velars compared to retroflex and dental stops which was large compared to labials

This feature which measures the tilt of the burst spectrum is similar, but not identical, to a feature used by Suchato [6] for discriminating stops based on the three places of articulation: labial, alveolar and velar.

Feature 2: From the monotonic decreasing behavior of the labial spectra, diffuse falling or flat spectra of retroflex and dental consonants and peaked spectra of velars in the frequency band greater than 500 Hz, the difference in the spectral magnitude at two frequencies as given in Eq. (2) was used for the three way classification.
A_{0peak} (dB) = 20log (A (0)/A (peak_lcn)) ---- (2)
where peak_lcn is the location (in frequency) of the most prominent peak in the region 500-7500 Hz. This was expected to be small for velars compared to retroflex and dental stops which were smaller than labials.

B. Distinction of /t/ and /th/:

Feature 3: From the burst spectra of /t/ and /th/ it was observed that, /t/ showed falling spectrum while /th/ showed flat spectrum in the range 5-7 kHz. Hence the energy difference in the two frequency bands, 1.5-3.5 kHz and 5-7 kHz, (E_{mh}) was used to distinguish /t/ from /th/ using Eq. (1), where, lo1, hi1, lo2 and hi2 are 1.5, 3.5, 5 and 7 kHz respectively. This was expected to be large for retroflex compared to dentals.

4. Results

The discriminating capacity of the features was evaluated by using ANOVA [9]. Box-and-whiskers plots, F-ratio and P-value were used to evaluate the features. Confidence level was set to 99%. Hence if P-value between the two groups is less than 0.01, then it was assumed that means of two groups differ significantly. Fig. 2 shows the box-and-whiskers plot for the feature, $E_{lh.}$ As expected, /p/ showed the largest negative value and for /k/, it was the smallest and the values for /t/ and /th/ are overlapping. Table 1 gives the details of ANOVA results for pair wise distinctions for three groups.

Fig. 3 shows the box-and-whiskers plot for the feature, A_{0peak}. In general spectra for /k/ showed rising characteristics in the initial frequency range in addition to compact peaks which are the main features. Accordingly, it is clear from Fig. 3 that the feature value A_{0peak} is either negative or a small positive value for /k/. /p/ is found to have the largest value compared to other classes because of the high rate of fall in the initial frequency region. In this case also the values for /t/ and /th/ are overlapping. Details of the ANOVA results for pair wise distinctions for the three groups are illustrated in Table 2.

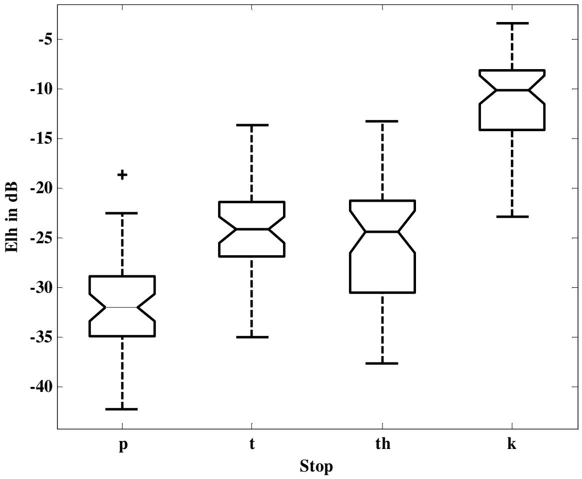

Fig. 2: Box-and-whiskers plot for the feature E_{lh}.

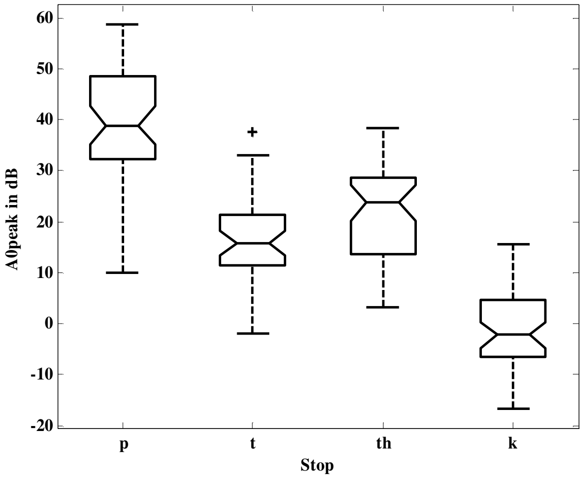

Fig. 3: Box-and-whiskers plot for the feature A_{0peak}.

	F-Ratio	P-value
/p/ and /t/, /th/	51.75	4.14×10^{-11}
/p/ and /k/	405.8	0
/t/, /th/ and /k/	211.15	0

Table 1: ANOVA results for the feature E_{lh}.

	F-Ratio	P-value
/p/ and /t/, /th/	136.94	0
/p/ and /k/	457.42	0
/t/, /th/ and /k/	176.69	0

Table 2: ANOVA results for the feature A_{0peak}.

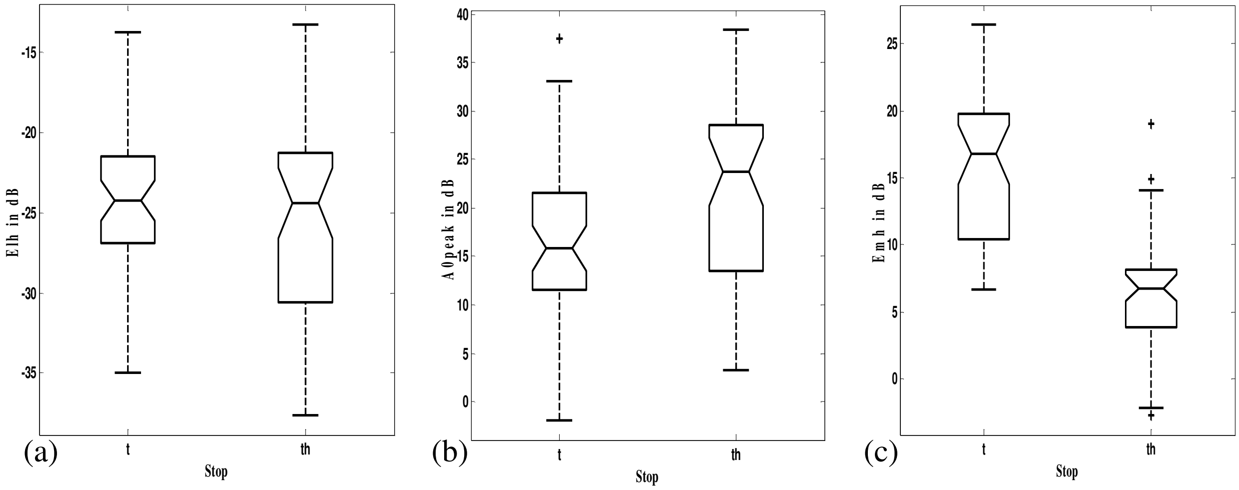

Fig. 4: Box-and-whiskers plots for features E_{lh}, A_{0peak} and E_{mh} to distinguish /t/ form /th/ (a) Feature E_{lh} (b) Feature A_{0peak} and (c) Feature E_{mh}.

Fig. 4 shows the box-and-whiskers plots for the /t/ and /th/ distinction using the above mentioned three features. ANOVA results for the same are given in Table 3.Fig. 4 and Table 3 reveal that the feature, E_{mh} separates the two phonemes better than either of E_{lh} and A_{0peak}.

Feature	/t/ and /th/	
	F-Ratio	P-value
E_{lh}	0.39	0.5351
A_{0peak}	7.02	0.0096
E_{mh}	82.83	2.54×10^{-14}

Table 3: ANOVA results for /t/ and /th/ distinction

5. Discussion

In this work, features were derived only from the burst spectrum. The feature, A_{0peak} that characterizes the difference between the spectral magnitude at 0 Hz and that of the most prominent peak in the frequency region 500-7500 Hz appears to perform three way distinction between /p/, /t/ and /th/ and /k/ better than E_{lh}, a measure of difference between the energy in the two frequency bands 0-750 Hz and 750-7500 Hz. The feature, A_{0peak} for /k/ showed comparatively high values when preceded by front vowels because of the less compact peaks. Also the values were found to be slightly higher for one speaker compared to others. Similar observations were made for the feature E_{mh} that defines the difference between the energy in the two frequency bands 1.5-3.5 kHz and 5-7 kHz for /t/ and /th/ distinction. It was found that /th/ preceded by the vowel /u/ showed decreasing spectra in the higher frequency region compared to other two vowels and hence in such cases the values for the feature E_{mh} for /th/ were overlapping with those of /t/. This indicates that results could be improved if vowel spectrum is also considered along with the burst spectrum in which case the features extracted are expected to be more speaker independent as well as context independent. Our future work will be focused on extracting features of the burst spectrum considering also the voicing onset spectrum. Further work will extend the acoustic analysis to aspirated and voiced stops.

References

[1] M. Halle, G.W. Hughes and J.P.A.Radley, "Acoustic properties of stop consonants", *J. Acoust. Soc. Am.,* vol. 29, no. 1, pp.107-116, Jan. 1957.

[2] H. Winitz., M.E. Scheib and J.A. Reeds, "Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech", *J. Acoust. Soc. Am.,* vol. 51, no. 4 (part 2) pp.1309-1317, Apr., 1972.

[3] S.E. Blumstein and K.N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *J. Acoust. Soc. Am.,* vol. 66, no. 4, pp. 1001-1017, Oct., 1979.

[4] D. Kewley-Port, "Time varying features as correlates of place of articulation in stop consonants", *J. Acoust. Soc. Am.,* vol. 73, no. 1, pp. 322-335, Jan., 1983.

[5] A. Lahiri, L. Gewirth and S.E. Blumstein, "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study", *J. Acoust. Soc. Am.,* vol. 76, no. 2, pp. 391-404, Aug., 1984.

[6] A. Suchato, Classification of stop place of articulation, PhD Thesis, MIT, June, 2004.

[7] R. Verma and P. Chawla, "Comparative analysis of Hindi retroflex and dental CV syllables and their synthesis", Workshop on Spoken Language Processing, TIFR, Mumbai, Jan., 2003.

[8] "Dental consonant", http://en.wikipedia.org/wiki/Dental_consonant .

[9] J. Miller, *Statistics for Advanced level*, 2nd Edition, Cambridge University Press, 1989.