Proc. of Workshop on image and signal processing (WISP), December 2007, Guwahati, India

FOUR-WAY CLASSIFICATION OF PLACE OF ARTICULATION OF MARATHI UNVOICED STOPS FROM BURST SPECTRA

Veena Karjigi, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology, Bombay, India
{veena, prao}@ee.iitb.ac.in

ABSTRACT

Acoustic features computed from the release burst spectrum are evaluated for the classification of Marathi unvoiced and unaspirated stops characterized by four places of articulation. The burst onset spectra are found to provide significant information on place of articulation as determined by feature evaluation measures and classification experiments on a database of Marathi word-initial stops. Classification accuracies are compared with those of cepstral coefficients computed from the same analysis data.

Index Terms— acoustic features, unvoiced stops, four-way place of articulation

1. INTRODUCTION

Close but contrasting phonemes in a language can be distinguished from each other in terms of linguistically motivated distinctive features, characterized by corresponding acoustic correlates in the speech signal. Present day speech recognizers typically employ raw spectral information in the form of cepstral coefficients. Acoustic-phonetic features, being more directly related to underlying distinctive articulatory properties, are expected to be more robust to variations at the phone level due to context, regional accent or vocal tract differences due to gender or age.

In this work, acoustic attributes for unvoiced, unaspirated stops of Marathi which differ in their place of articulation (PoA) are investigated. Marathi is the official language of the Indian state of Maharashtra with roughly 70 million native speakers. It distinguishes four places of articulation for stop consonants in contrast to the three used in English. The four places of articulation are labial [p], dental [t̪], retroflex [ʈ] and velar [k] each of which can occur as unvoiced-unaspirated, unvoiced-aspirated, voiced-unaspirated and voiced-aspirated [1]. The dental consonant is produced by making a constriction of the vocal tract with the tongue blade, immediately behind the upper front teeth. The retroflex consonant is produced by curling the tip of the tongue upwards towards the hard palate to make a constriction behind the alveolar arch. Phonologically, dental and retroflex stops are clubbed in the same class (coronal) and distinguished from labials as well as velars [2]. The retroflex and dental stop consonants would typically both be categorized as alveolar [t] by an English listener. Considering their contrastive role in Marathi (as also in some other Indian languages), it is of interest to investigate the acoustic correlates of these different PoA.

Finding acoustic attributes for the classification of the stops based on PoA has been a subject of active research. Spectral shape of the release burst and formant transitions in adjacent vowels have been widely investigated. Burst shape (and amplitude) attributes are found to be relatively invariant to vowel context [3]. Early work [4] using burst spectra showed that labials and alveolars had diffuse spectra as compared with the peaked spectra of velars. The labials are further distinguished from alveolars by the spectral location of major energy concentration. Zue [5] computed burst spectra by linear prediction (LP) using the first 10 to 15 ms of the waveform after the burst release. These LP spectra were further smoothened and the location of the biggest peak in the resultant spectrum was measured as the burst frequency. The labials showed large variations in the burst frequency and they did not exhibit prominent peaks. Blumstein and Stevens [6] derived a 14th order LP spectrum by placing a modified raised cosine window of length 26 ms at the burst onset. They proposed three templates, diffuse flat or falling for labials, diffuse rising for alveolars and compact for velars, using which a high classification accuracy was obtained on word initial stops. Several experiments have been reported on the importance of the first 10 ms of the waveform after the burst onset [5,6]. Suchato [7] used average power spectra for measuring 15 acoustic attributes of American English stops out of which 3 were extracted purely from the burst spectrum.

There has been very limited amount of work on extending the above acoustic attributes to the classification of stops in Indian languages, several of which have more than three PoA in their inventory. Lahiri et al. [8] investigated the known acoustic attributes for the labial, dental and alveolar stops of Malayalam. It was observed that the dental stops could not be reliably separated from labials based on burst spectra alone. Features based on the change in spectral energy concentration from burst onset to voicing onset were found to perform better in terms of grouping the Malayalam dentals with alveolars.

The present work is restricted to acoustic-phonetic features extracted from the release burst segment for PoA classification of Marathi unvoiced and unaspirated stops. Acoustic features related to the release burst proposed in the literature for English unvoiced plosives are tested on a database of Marathi plosives and improvements are proposed. The acoustic features are compared with more general raw spectrum features (such as the widely used MFCC) computed on the same data via stop classification experiments.

2. DATABASE AND ANALYSIS

2.1. Database

Marathi words with one of the four stops {p, t̪, ʈ, k} in the word-initial position followed by one of the eight vowels and two diphthongs of the language were used for the analysis. Two distinct words for each stop-vowel combination were chosen from the dictionary to obtain 80 words. The words were each embedded in two different carrier phrases (one statement and one question). Five male and five female speakers of standard Marathi [1] were selected for the study. This led to a data set of 80 x 10 x 2 = 1600 tokens (or 400 per stop consonant), recorded at a sampling rate of 16 kHz in quiet condition. The time locations of the release burst and the voicing onset were manually labeled. The burst onset was marked as the time instant after the closure silence at which a rapid change in the waveform amplitude sets in. The first negative to positive going zero crossing in the first cycle of the periodic waveform following the burst was labeled as the voicing onset.

2.2. Acoustic Analysis

Since the goal is to extract acoustic-phonetic features from the release burst spectrum of the unvoiced plosive, the analysis data is restricted to the region between the labeled burst and voicing onsets. This duration is known as the voicing onset time (VOT) of the unvoiced stop. A statistical study of measured VOT across the Marathi word data set is summarized in Table 1. The observations are consistent with articulatory properties [9]. Retroflex exhibit lowest VOT due to the relatively fast movement of the active articulator involved (the tongue tip), which offsets the effect of the more posterior PoA. Further, although the place of constriction for the dental is posterior to that of labial, the VOTs are comparable. A possible explanation is the observed occasional presence of aspiration in the word initial [p]. In Marathi, the aspirated [p] has linguistically evolved to be replaced by [f] due to which there is a spread in the allophonic varieties of [p].

Place of	VOT (ms)	
	Mean	Std. dev.
Labial	17.0	7.7
Dental	15.2	6.1
Retroflex	9.9	4.0
Velar	28.5	10.6

Table 1. *VOT mean and std. dev. for the 4 PoA*

Based on the observations, the data extracted for burst spectrum analysis was limited to either a fixed 10 ms or the VOT, whichever was lower.

2.2.1. Computation of average power spectrum

A smooth power spectrum was obtained, following the method of [10, 7], by averaging the power spectra of a series of windowed data segments each of duration 6.4 ms. The Hanning data window was shifted every 1 ms starting from a center value of 7.5 ms before the burst onset to 7.5 ms after the burst onset. If the VOT was found to be less than 7.5 ms, the last window was centered at 3.2 ms before the voicing onset so as not to encroach on the voiced region. Thus the maximum number of spectra averaged was 16. The time averaging of power spectra serves to compensate for possible errors in the manual labeling of the burst onset.

2.2.2. Burst spectrum characteristics

Stops in Marathi contrast in four PoA as opposed to the three of English. Hence it is important to characterize differences in the spectra obtained from our database with that of the stops in English as noted in the literature. Figure 1 shows typical average power spectra of the four stops from the data of a female speaker. Similar to [4,6], we find Marathi labials showing diffuse falling spectra but with a higher roll-off in the low frequency region (0-750 Hz) compared to the rest of the frequency band. The velars show a compact peak near F2 of the following vowel.

However, burst spectral characteristics of the English alveolar [t] (diffuse, rising) do not completely describe the observed spectra of the two Marathi coronals. While the three coronal stops share the diffuseness characteristic [11], it is seen that the dental [t̪] has a diffuse flat spectrum and retroflex [ʈ] exhibits a slightly more compact and high-energy spectrum up to 4 kHz with an abrupt decrease in energy beyond that.

The observed spectral characteristics are generally consistent with articulatory properties corresponding to the resonances of the vocal tract volume downstream from the place of constriction. Labials do not exhibit clearly defined peaks in the spectrum because of the absence of the anterior cavity while velars exhibit a clear low frequency peak due to

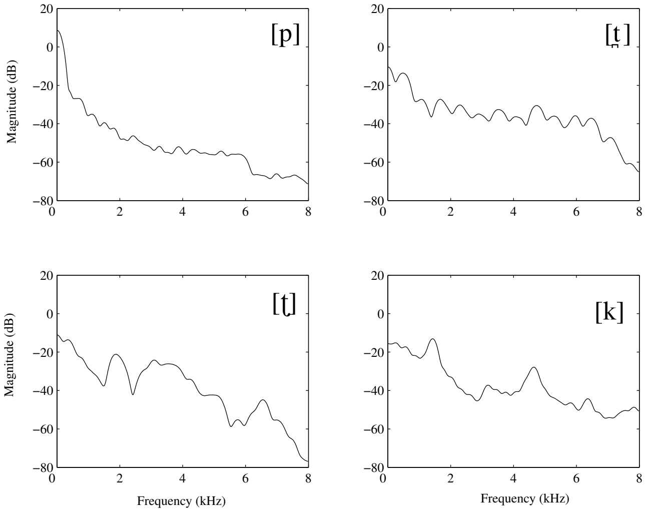

Figure 1: *Average power spectra of the four stop bursts from a female speaker*

the long anterior cavity [12]. The lower frequency concentration of retroflex stops relative to dentals is explained by the longer anterior cavity for retroflex articulation. Increased posterior cavity for labials, dentals and retroflex increases acoustic losses thereby giving rise to diffuse peaks. That an abrupt decrease in energy with frequency distinguishes apicals (such as the retroflex) from laminals (such as the dental) has been noted previously [11].

3. ACOUSTIC FEATURES AND EVALUATION

We see above that the gross shape of the burst spectrum and the locations of spectral prominences distinguish the average power spectra of the four places of articulation. Acoustic features that capture these characteristics could be effective in the automatic recognition of the unvoiced plosives. We measure the effectiveness of the individual features with respect to the classification problem using the information theoretic mutual information (MI) [13]. The MI has been used for feature selection in an HMM-based phone recognizer and shown to correlate well with recognition scores for the ranking of features [14].

We start with acoustic features previously proposed for three-way classification of English stops and evaluate these for the three-way classification of the Marathi stops with the two coronal stops clubbed in one class to be distinguished from labial and velar. Next, we propose modifications to the feature definitions considering the observed burst spectrum characteristics of the Marathi plosives discussed in Sec. 2.2.2, and evaluate their effectiveness by the same measures.

Suchato [7, 3] proposed several acoustic attributes relating to the burst spectrum and formant transitions for stop consonant place of articulation classification. Of these, three attributes relate purely to the burst spectrum shape and are computed from the average power spectrum described in Sec. 2.2.1. These three acoustic features and the associated parameters are detailed below.

1. Energy difference:

$$E_{diff} = 10\log\left(\frac{E_1}{E_2}\right) \tag{1}$$

where, E_1=*E* [3500:8000] (energy of the burst spectrum in the range 3500-8000 Hz), and E_2=*E* [1250:3000]. E_{diff} was shown to achieve reasonable separation of the three PoA of English stops [7].

2. Amplitude difference:

$$A_{diff} = 20\log\left(\frac{A_1}{A_2}\right) \tag{2}$$

where, A_1 is the amplitude of the biggest peak of the burst spectrum in the range 3500-8000 Hz and A_2 is the average peak amplitude of the burst spectrum in the range 1250-3000 Hz. A_{diff} was shown to separate alveolar stops from English labials and velars [7].

3. Center of gravity in frequency ("*cgFa*"): of the average power spectrum obtained from the data between burst and voicing onsets computed over the frequency region 0-8 kHz.

The above features were tested on the Marathi unvoiced stops data for three-way classification (labial, dental+retroflex, velar). Table 2 shows the MI for the above mentioned three features. We see that while the *cgFa* shows reasonable discrimination ability for the three-way classification, the remaining two features, E_{diff} and A_{diff} perform poorly on the Marathi data. (The quantitative measures are supported by the visual inspection of the probability distributions of feature values which overlap significantly across the classes). Based on the study of Sec. 2.2.2 of the Marathi consonant burst spectra, the parameter values of the energy and amplitude ratio attributes of Eq. (1) and (2) were modified as described next. Further, new features were explored for the two-way classification of the coronal stops [ʈ] and [t̪]. Later, few features are added to account for the four-way distinction.

3.1. Features for three-way classification

E_{diff} is modified for the three-way classification to consider the steep spectral roll-off of labials in the low frequency region distinguishing them from the relatively flat spectral shape of retroflex and dentals. The modified feature (E_{ml}) is given by Eq. (1) where, E_1=*E* [750:2500] and E_2=*E* [0:750].

Next, A_{diff} was modified based on the observation that burst spectra become increasingly diffuse proceeding from velar toward labial PoA. Labials exhibit insignificant peaks beyond 500 Hz whereas retroflex and dental show larger spectral peaks. Velars exhibit relatively high and narrow peaks. The amplitude ratio feature was thus modified to A_{hl} given by Eq. (2), where, A_1 is the amplitude of the biggest peak in the region 500-7500 Hz and A_2 is the average amplitude of the burst spectrum in the region 0-

500Hz. Drawing on the same spectral characteristic, a new feature (S_{lf}) is defined as the spectral slope obtained by fitting a straight line to the burst spectrum in the region 0-1.5 kHz using linear regression. All the three above mentioned features are expected to take positive values for velars, negative values for retroflex and dental and large negative value for labials. Further, cgFa was recomputed from the average power spectrum of Sec. 2.2.1, and modified to the frequency range of 0-7 kHz to account for the microphone characteristics.

The feature evaluation measures obtained on the Marathi database for the four new features appear in Table 2. We see that the modified energy and amplitude ratio features improve significantly upon the features of Suchato [7].

Suchato's features	Normalized MI	Modified features	Normalized MI
E_{diff}	0.1111	E_{ml}	0.5095
A_{diff}	0.1553	A_{hl}	0.5805
cgFa	0.3152	*cgFa*	0..4228
		S_{lf}	0.3541

Table 2. *Feature evaluation for the three-way classification*

3.2. Features for two-way classification of coronals

From the discussion of Sec. 2.2.2 comparing the burst spectra of the two coronal stops, we note that [ʈ] is characterized by an abrupt fall in spectral energy, while [t̪] exhibits a more gradual decrease in energy across frequency. To capture this distinction, an energy ratio is defined (E_{hm}) as in Eq. (1) in the two frequency bands, where E_1=E[5000:7000] and E_2=E[1500:3500]. The energy variation with frequency is also captured by the spectral slope in the region 2-6 kHz. A slope feature (S_{mf}) is defined over this frequency region derived in the same way as S_{lf}. Both E_{hm} and S_{mf} are expected to be highly negative for [ʈ] and less so for [t̪].

3.3. Additional features

To increase the separation between labials and dentals, a new feature E_{hl} is defined as in Eq. (1), where E_1=E [5000:7000] and E_2=E[0:750]. Because of the prominent energy in low frequency region for labials and high frequency region for dentals, E_{hl} was expected to show large negative values for labials and larger positive values for dentals. In addition, it showed a relatively high four-way distinction.

Further, spectral prominences in each of the four sub bands (0-750, 750-2500, 2500-5000 Hz and 5000-7000 Hz) were computed (similar to cgFa) and named as cgB1, cgB2, cgB3 and cgB4. Features are ranked using a greedy algorithm [13]. The 11 features in the order of ranking are: A_{hl}, S_{mf}, cgB1, cgB4, E_{hl}, cgB3, cgB2, E_{hm}, S_{lf}, cgFa and E_{ml}.

4. CLASSIFICATION EXPERIMENTS

Based on the feature evaluation of Sec. 3, two feature vectors were obtained: one with the 11 acoustic features and other with only 8 best features. These two feature vectors are tested in a GMM classifier framework for the four-way classification of PoA. A diagonal covariance GMM classifier was trained using EM algorithm with 1, 3, 5 and 8 mixtures per class. Also tested in the same framework were MFCC vectors of two different dimensions {first 8 and first 13 coefficients} with the MFCC obtained from 20 ms Hamming windowed data centered at the burst onset (i.e. extending 10 ms beyond the burst onset). Two different classification tasks are defined.

(a) Task 1: The training set comprised of tokens (800) from three male and two female speakers and testing set comprised of tokens (800) from the remaining two male speakers and three female speakers and vice-versa. Hence there were 2 sets of training-testing in the round-robin (1600 test items). Classification results in % accuracy (percentage of 1600 test tokens identified correctly) are given in Table 3.

Feature set	No. of GMM mixtures			
	1	3	5	8
8 AP features	78.63	77.44	78.88	77.69
11 AP features	78.63	77.50	77.38	77.00
First 8 MFCCs	70.81	74.94	76.75	74.81
First 13 MFCCs	71.50	74.25	75.25	77.81

Table 3. *Classification results: Trained and tested with different speaker sets, each including males and females*

(b) Task 2: The training set comprised of tokens (800) from female speakers and testing set comprised of tokens (800) from male speakers and vice-versa. Hence there were 2 sets of training-testing in the round-robin (1600 test items). Classification results in % accuracy are given in Table 4.

Feature set	No. of GMM mixtures			
	1	3	5	8
8 AP features	77.38	77.94	78.00	77.13
11 AP features	78.13	76.06	76.19	76.00
First 8 MFCCs	69.13	71.94	74.13	74.31
First 13 MFCCs	67.75	71.50	71.13	72.06

Table 4. *Classification results: Trained with male and tested with female speakers and vice-versa*

5. DISCUSSION

Burst onset spectra are found to provide significant information on place of articulation as demonstrated by acoustic feature evaluation and classification results on Marathi unvoiced stops. While previously proposed articulatory-acoustic features for the classification of English unvoiced plosives were found inadequate for the three-way separation of Marathi stops, suitably modified features performed significantly better. The proposed set of 11 as well as the best 8 AP features derived from a study of burst spectrum characteristics across the four PoA of Marathi stops, compare favorably in classification accuracy with the 13-MFCC and 8-MFCC vectors extracted from frames aligned with the burst onset.

In the speaker-independent classification task, both the AP feature sets obtain a maximum classification accuracy similar to that of the 13-MFCC and 8-MFCC vectors. Moving to the cross-gender classification, the performance of the AP features decreases only slightly compared with the steep reduction in accuracy recorded with the MFCCs. The lower dimension MFCC vector fares slightly better than the full 13-MFCC in the cross-gender task indicating that the essential (PoA-specific) shape of the burst spectrum is captured by 8-MFCC and finer spectral detail in the 13-MFCC may reduce robustness to irrelevant variations.

In summary, the results of the present work provide support for the notion that acoustic features have the potential to capture essential phonetic distinctions in a robust manner. Wider testing conditions including variations in dialect, speaking rate and recording conditions would be useful to further validate this. The proposed acoustic feature set has not been systematically optimized for parameter settings (e.g. frequency ranges). A more efficient set of features could result from the fine-tuning of the individual features combined with feature selection to reduce redundancy. Finally, the present work was restricted to the release burst spectral shape of the unvoiced stop. Important acoustic cues to PoA are known to lie in the transition and voicing onset regions. Future work will be directed towards improving classification accuracy by extending the analysis data to include more of the speech waveform for the place detection of Marathi unvoiced stops.

6. ACKNOWLEDGEMENT

The authors wish to acknowledge useful discussions with K. Samudravijaya.

7. REFERENCES

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

[2] Ladefoged P. and Maddieson I., *The Sounds of the World's Languages,* Blackwell, 1996.

[3] Suchato A. and Punyabukkana P., "Factors in classification of stop place of articulation", *Proc. ICSLP*, pp. 2969-2972, Sep. 2005.

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

[5] Zue V.W., "*Acoustic characteristics of stop consonants: A controlled study*", Sc.D. Thesis, MIT, May, 1976.

[6] Blumstein S.E. and Stevens K.N., "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *J. Acoust. Soc. Am.,* vol. 66, pp. 1001- 1017, Oct., 1979.

[7] Suchato A., "*Classification of stop place of articulation*", Ph.D. Thesis, MIT, Jun., 2004.

[8] Lahiri A., Gewirth L. and Blumstein S.E., "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, pp. 391-404, Aug., 1984.

[9] Cho T. and Ladefoged P., "Variation and universals in VOT: evidence from 18 languages", *J. Phonetics.,* vol. 27, pp. 207-229, Jul., 1999.

[10] Stevens K.N., Manuel S.Y. and Matthies M., "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production", *Proc. ICPhS*, pp. 1117-1120, Aug., 1999.

[11] Hamann S.R., "*The phonetics and phonology of retroflexes*", Ph.D. Thesis, Netherlands Graduate School of Linguistics, Jun., 2003.

[12] Stevens K.N., *Acoustic Phonetics,* MIT Press, 2000.

[13] Battiti R., "Using mutual information for selecting features in supervised neural net learning", *IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537-550, Jul., 1994.

[14] Omar M.K., Chen K., Hasegawa-Johnson M. and Brandman Y., "An evaluation of mutual information for selection of acoustic features of phonemes for speech recognition", Proc. *ICSLP*, pp. 2129-2132, Sep., 2002.