Proc. of Wireless Personal Multimedia Communications (WPMC), September 2007, Jaipur, India

Identification of stop consonants for acoustic keyword spotting in continuous speech

Veena Karjigi, Bhavik Patel, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology, Bombay
e-mail: {veena,bhavik,prao}@ee.iitb.ac.in

***Abstract*— Acoustic segmentation of speech based on landmark detection is an important stage in keyword spotting based on acoustic matching. In the present work, the class of plosive sounds is considered for detection and classification in continuous speech. Acoustic-phonetic features extracted in the vicinity of landmarks or speech events are shown to be reliable for the detection of unvoiced stops with high temporal accuracy.**

***Keywords*— Keyword spotting, landmark detection, acoustic-phonetic features, unvoiced stops**

I. INTRODUCTION

There are several voice communication applications where the detection of specific target words or phrases in an incoming stream of unconstrained speech is important. These include command interfaces to machines, possibly operated by untrained users, and searching large spoken audio databases by keywords. Available large vocabulary continuous speech recognition (LVCSR) systems can be used for the purpose by applying text keyword spotting to the output recognized speech. Current day LVCSR systems achieve reasonable recognition accuracy owing largely to the use of sophisticated language modeling. As such they have a limited applicability in the context of unconstrained speech with possibly out-of-vocabulary target words. A conceptually (and computationally) simpler alternative to speech recognition followed by text mining, is keyword spotting based on acoustic matching between the keyword (represented by its phone sequence) and the speech signal.

In the present work, we consider acoustic keyword spotting based on the detection of target phonemes in the continuous speech signal. Spectral representations such as the MFCC, widely used in LVCSR systems, can be used as the basis for acoustic matching between the input signal and keyword templates. However acoustic-phonetic representations for phoneme detection are expected to provide greater robustness in the context of speaker, language and environment variabilities [1]. Further, a knowledge-based approach based on acoustic-phonetic features enables the use of separate detectors, each best suited to a phonetic class.

Acoustic-phonetic (AP) representations for speech are typically "landmark-based". That is, the features used for recognition are extracted in a two-step process involving first detecting acoustic landmarks [1]. Landmarks are the time instants in the speech signal where characterstics of the signal change abruptly. Hence they are considered as information rich areas and features extracted from the signal region around the landmarks are the most salient. The actual AP features extracted will depend on the nature of the landmark. For example, the place of articulation of a stop consonant is best captured by the spectral characteristics of the signal near the burst and voicing onsets. Similarly, the begin and end of frication are important landmarks for the identification of fricatives, and landmarks corresponding to voicing onset and offset for sonorants which include vowels, semivowels and nasals.

Labeling a speech segment with a sequence of phonemes thus involves the reliable detection of landmarks followed by signal analysis in the vicinity of the landmarks to extract features suitable to the broad phonetic class corresponding to the specific landmark. In the present work, we consider the broad phonetic class of plosives for landmark detection. Plosives comprise the stops and affricates, and are considered particularly challenging to recognise due to their highly dynamic characteristics. In the present work, we develop acoustic-phonetic features for the recognition of unvoiced stop consonants. Acoustic-phonetic (AP) features have been designed and evaluated extensively on the English unvoiced stop consonants (p,t,k). Several Indian languages, on the other hand, have a significantly larger set of linguistically distinct stop consonants and therefore require separate studies on AP features for recognition. Our work is currently focussed on the identification of unvoiced, unaspirated (UVUA) stop consonants of Marathi (which also happen to be common to several other Indian languages). The UVUA stops correspond to four distinct places of articulation (PoA) viz. Labial, dental, retroflex and velar [2].

In the next section, we review previous literature pertinent to the detection of landmarks and the extraction of AP features for recognition of unvoiced stops. This is followed by a description of the Marathi speech database that was developed for the training and evaluation of the classification system. The algorithms developed for landmark detection and classification are discussed next, followed by a presentation of the classification results. In addition to AP features, we also evaluate MFCC-based features extracted at the landmarks as more general spectral features.

II. PREVIOUS WORK

Landmarks are time instants at which the signal properties change abruptly. They have been detected by observing energy changes in specific bands of the speech spectrum. An automatic method for detection of the burst release of aspirated stops, as proposed by Liu [3], is based on the time derivative of the energy computed in 4 frequency bands in the range of 800 Hz to 5000 Hz. The energy derivatives in these 4 bands show simultaneous peaks, indicating the burst onset. Further, the onset and offset of periodicity before and after the burst are detected from the energy derivative in the low frequency band (0-400 Hz). The burst landmark was recognized with error rate of 12%. Salomon et al. [4] improved the performance of the Liu's landmark detector for noisy speech by refining the

definition of the energy derivative to suit the underlying local characteristics of the signal such as whether it belongs to silence, periodic or aperiodic regions of speech. An overall detection rate of 80.2% was obtained for event detection. Prasanna and Yegnanarayana [5] used excitation information contained in the LP residual for detecting the presence of pitch periodicity and thus the voicing onset. 90.2% of vowel onset points were detected correctly within a resolution of ±30 ms. Das and Hansen [6] used the signal band pass filtered with 300-1200 Hz band and computed Teager energy of the signal and average Teager energy of forward and backward differences of signal. By this, the amplitude modulation component (AMC) of the signal in the band is obtained. Next, the time instant at which there occurs a rapid increase in AMC is marked as the voicing onset.

Stops can be classified in terms of place of articulation by suitable features extracted from the vicinity of the burst and onset of voicing landmarks. Burst and voicing spectral shapes, formant transitions and temporal features have been used in the previous literature. Features extracted from the burst spectrum are relatively insensitive to the vowel context [7]. Early work [8] 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.

Stops are short duration sounds and are, further, prone to co-articulation effects from the following vowel. The vowel region also provides some useful information for the classification of stops. Stevens et al. [9] and Suchato [7] used average power spectra derived from both burst and voicing regions and used features relating amplitudes of burst and voicing spectra in specified frequency regions and formant related parameters to classify English stops. Hoelterhoff and Reetz [10] distinguished German obstruents based on manner and place of articulation by considering only two PoAs: labial and alveolar. They computed relative amplitude (energy in the burst spectrum relative to energy in the voicing spectrum) in fixed frequency bands. These bands were of 1 kHz bandwidth, equally spaced in the region 0-8 kHz, and showed varying abilities to distinguish the obstruents.

Representing the use of more generic spectral features is the work of Nossair and Zahorian [11]. They used DCTCs (discrete cosine transform coefficients) computed from amplitude and frequency scaled versions of the Fourier transformed spectrum for classification of English word initial stops. By obtaining 7 DCTCs from a static window placed at the burst onset, an accuracy of around 82% was obtained. Further, by using a larger portion of the signal (60 ms from the burst onset) including the following vowel, they obtained DCTC trajectories over time. Each of these was represented by 3 low cosine basis vector coefficients which were then used for classification. An improvement of 12% was observed over features extracted only from the static window. The similar broad approach of using spectral vectors extracted from an extended region aligned with specific landmarks was used by Gangashetty et al. [12] to classify CV syllables in an MFCC based SVM classifier.

We may conclude that, in addition to using static features extracted from burst only, dynamic features derived over a larger segment of the signal which includes the vowel and formant transitions improves the classification accuracy for stops.

III. DATABASE

Two databases were constructed, one each for training and testing the classification system. Training data comprised of 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. 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 [2] 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 testing database comprised of 16 Marathi sentences spoken by each of one male and one female speaker recorded at a sampling rate of 16 kHz. The sentences contained a high concentration of plosives especially UVUA stops. In all, test data had 219 plosives.

The time locations of the release burst and the voicing onset for both training and testing data are labeled as follows. 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.

Fig.1 shows the landmarks for plosives the Marathi utterance /kithi kaam/. Labels 'b' indicate the burst onsets corresponding to the three burst /k, th and k/ and labels 'v' indicate the corresponding voicing onsets.

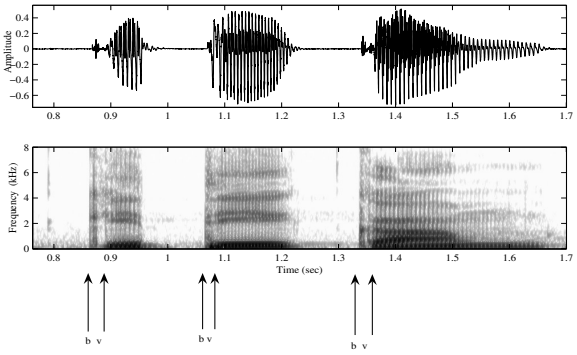

Fig.1. Manually marked landmarks with the waveform and spectrogram of Marathi utterance /kithi kaam/.

IV. LANDMARK DETECTION FOR PLOSIVES

The two landmarks for plosive phones are burst onset and voicing onset. The time derivative (or the rate of rise) of energy in different frequency bands is used to detect the burst onset The voicing onset is detected by the Teager energy method proposed in [6].

A. Burst onset detection

Similar to landmark detection by Liu [3], spectrogram is computed every 1 ms using 6 ms Hamming window and 512 point FFT. Further, energies (E_1 and E_2) in two frequency bands (2-3.5 kHz and 3.5-5 kHz respectively, which are the fourth and fifth bands in [3]) are computed followed by the computation of rate of rise (ROR) in both the bands. ROR at time t is defined as,

$$ROR_i(t) = E_i(t) - E_i(t - t_{step}) \quad (1)$$

where, i is the band number and t_{step} is 20 ms. The closure silence is detected by using a threshold on energy in the

frequency region: 0-8 kHz. An interval of 10 ms after the end of silence is searched for the burst onset. Instants where RORs in both the bands cross a threshold of 9 dB are noted and the later of the two (indicating the time where both RORs have crossed 9 dB) is labelled as burst onset. Fig. 2 (b) and (c) show RORs in bands 1 and 2 respectively for the word /kaathyakoot/ which has four bursts; the ellipses indicate the regions where RORs in both the bands have crossed the 9 dB threshold.

B. Voicing onset point (VOP) detection

Following the computationally simple method of [6], voicing onset detection is performed using Teager Energy Operator (*TEO*) defined in Eq (2).

$$TEO[x(n)] = x(n)^2 - x(n-1)x(n+1) \quad (2)$$

We note that the *TEO* at instant n, depends only on three samples of speech input which helps in providing high temporal resolution in marking energy fluctuations.

Modelling speech resonances by amplitude modulation (AM) and frequency modulation (FM) using Teager energy, we can separate the low frequency (300-900 Hz) AM component of given speech signal [13]. This AM component is enhanced in the vowel region relative to the burst region in stop-vowel clusters [6]. AM component ($x_a(n)$) and FM component ($x_f(n)$) are given by following equations:

$$x_a(n) = \sqrt{\frac{TEO(n)}{1 - x_f{}^2(n)}} \quad (3)$$

$$x_f(n) = \arccos(1 - 0.5\frac{TEO(n)}{ATEO(n)}) \quad (4)$$

where *ATEO* is the average of the Teager energies of the forward and backward differences of the bandpass filtered signal.

An 85-ms segment (considering the maximum VOT of stops) following the detected burst onset point is processed to obtain AM and FM components. A rapid increase in AM component can be seen at the start of vowel region as shown in Fig.2 (d). To obtain reliable detection and localisation of voicing onset, $x_a(n)$ is convolved with the negative of the modulated Gaussian window g(n), given by Eq. (5).

$$g(n) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{n^2}{2\sigma^2}} \cos(\omega n) \quad (5)$$

where, σ = 0.2, ω=0.0114 and n=800.

The "detection function" so obtained shows a prominent peak at the voicing onset. But sometimes, when the burst is strong, the output goes high even in the burst region as shown by the label in Fig.2 (e). But in such cases, the detection function falls below a threshold after some time. Considering these situations, a simple method is devised to reliably detect the voicing onset. Local peaks exceeding a threshold are marked in the detection function. Whenever the detection function dips below some threshold between two valid peaks, the valid peak after the dip is marked as a voicing onset.

C. Results

The burst and voicing onset detector is evaluated on the testing data described in Sec III. A correct detection is said to occur when the automatically detected location is within ±30 ms of the manually labelled landmark; else, a deletion is recorded. Landmarks not actually present but detected by the algorithm are considered as insertions. The detection accuracy is computed as the ratio of detected to total landmarks. Results of landmark detection are shown in Tab. 1.

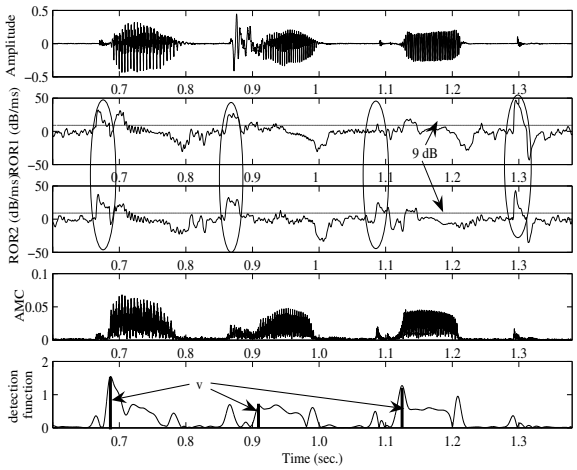

Fig.2. (a) Speech signal, (b) and (c) RORs in band 1 and 2, (d) AM component, (e)detection function for the Marathi word /kaathyakoot/

	Total	Detec-ted	Deleti-ons	Insert-ions	% acc.
Burst	219	195	24	16	89.04
VOP	205	186	19	3	90.24

Tab. *1. Evaluation of burst and voicing onset detection*

Tab. 2 shows % accuracy of detected landmarks for various tolerances. It can be noted that detection rate reaches above 89% for both the landmarks for ±15 ms tolerance. Burst landmark shows better temporal resolution compared to VOP landmark.

	±5 ms	±10 ms	±15 ms	±20 ms	±30 ms
Burst	85.84	87.67	89.04	89.04	89.04
VOP	64.88	84.88	89.27	89.76	90.24

Tab. *2. Landmark detection rates (%) for different tolerances*

V. CLASSIFICATION OF UNVOICED STOPS

Since knowledge of the following vowel context improves the identification of stops as discussed in Sec II, both static and dynamic features have been investigated in the present work. The effect of the preceding vowel is not significant, and only singleton stops followed by vowels or semivowels were used in the analysis.

A. Acoustic analysis

The average power spectrum [9, 7] is used to obtain the burst and voicing spectra. To compute the burst spectrum, the analysis data is restricted to the region between the labeled burst and voicing onsets. This duration is known as voicing onset time (VOT) of the unvoiced stop. VOT measured for Marathi database showed a lower mean of 10 ms for retroflex and the largest mean of 28 ms for velars [14]. Based on these observations, the data used for burst spectrum analysis was limited to either a fixed 10 ms or the VOT, whichever was lower.

A smooth power spectrum was obtained, following the method of [9, 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. Similarly, average voicing spectrum is computed by averaging a series of spectra obtained by moving a 6.4 ms Hanning window starting from 12.5 ms after the voicing onset untill 27.5 ms after the voicing onset with a hop size of 1ms, thus averaging 16 segments. The time averaging of power spectra serves to compensate for possible errors in the manual labeling of the burst onset.

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. Fig. 3 shows typical average power spectra of the four stops from the data of a female speaker. Similar to [8], 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 F_2 of the following vowel.

However, burst spectral characteristics of the English alveolar [t] (diffuse, rising) do not completely describe the observed spectra of dentals and retroflex. It can be 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 consistent with articulatory movements [14].

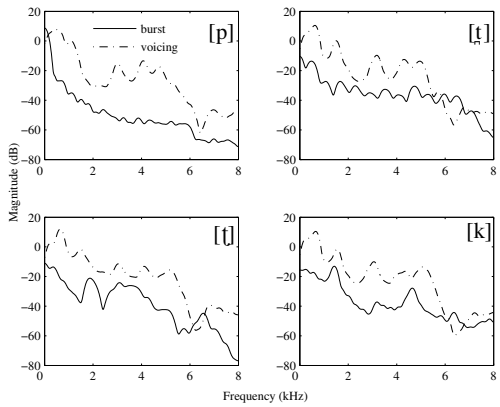

Fig. 3. *Average power spectra of the four stop bursts followed by the vowel /a/ from a female speaker*

B. Acoustic feature extraction and evaluation

Spectral features derived by Suchato [7] for English stops are evaluated for the Marathi database and modifications are suggested considering the observed spectral characteristics of the Marathi stops. First static features derived from only the burst spectrum are discussed followed by the dynamic features related to both burst and voicing spectra.

Static features:

Three attributes of Suchato [7] are related purely to the burst spectral shape. The three features were tested on the Marathi unvoiced stops data and improvements are suggested along with the addition of new features in [14]. A brief review is here. Two of these features give the burst spectral tilt in terms of energy and amplitude differences. Energy difference E_{diff} is defined as the ratio of total energy in the high frequency band, 3.5-8 kHz to that in the mid frequency band, 1.25-3 kHz. By considering the steep spectral roll-off of labials, this feature was modified as E_{ml} where, mid frequency band was replaced by low frequency band of 0-0.75 kHz and high frequency band was changed to 3-7 kHz. Amplitude difference was defined as,

$$A_{diff} = 20\log\left(\frac{A_1}{A_2}\right) \quad (6)$$

where, A_1 is the amplitude of the biggest peak of the burst spectrum in the range 3.5-8 kHz and A_2 is the average peak amplitude of the burst spectrum in the range 1.25-3 kHz. Owing to the insignificant peaks in the labial spectrum beyond 500 Hz, A_{diff} was modified to A_{hl} given by Eq. (6), where, A_1 is the amplitude of the biggest peak in the region 0.5-7.5 kHz and A_2 is the average amplitude of the burst spectrum in the region 0-0.5 kHz. In addition to these, additional features like spectral center of gravity and spectral slope in different sub bands are also added.

Dynamic features:

Suchato [7] used features relating amplitude of the peaks in burst and voicing spectra in four frequency bands. Two of these frequency bands are the same as those used to derive static features. In addition to these, the mid frequncy band, 1.25-3 kHz was split into F_2 (1.25-2.5 kHz) and F_3 (1.5-3 kHz) regions thus making four subbands.

Two of these features measure the amplitudes of mid and high frequency prominences of the burst normalized with respect to the vowel first formant prominence. Normalization with the first formant prominence reduces the speaker dependency. These two features are given by, Eq. (6), where, A_1 is the amplitude of the first formant of the vowel for both the features and A_2 is the biggest peak in the region 1.25-3 kHz of the burst spectrum for one and 3-8 kHz for the other.

Owing to the modified static features explained earlier, where frequency bands used for E_{ml} and A_{hl} were different from that of E_{diff} and A_{diff}, different frequency regions are used in deriving dynamic features too. Also because of the absence of clearly defined peaks in the spectra of labials, the definition of these features is slightly modified. Instead of the amplitude peak, spectral amplitude at the center of gravity in that particular region is used in deriving the features. Modified features $A_{vf1,bmid}$ and $A_{vf1,bhi}$ are given by Eq. (6), where, A_1 is the vowel first formant prominence and A_2 are the spectral amplitudes at center of gravities in regions 1-2.5 kHz and 5-7 kHz.

Other dynamic features used by Suchato [7] were amplitudes of the biggest peak in particular frequency band of the burst spectrum relating to the amplitudes of the biggest peak in the same frequency band in the voicing spectrum. These three features are given by Eq. (6), where, A_1 are the amplitude of the biggest peaks in the regions 1.25-2.5 kHz, 1.5-3 kHz and 3-8 kHz of the burst spectrum and A_2 are the biggest peaks in same frequency regions of the voicing spectrum. Again considering the observed spectral characteristics, modified features $A_{vf2,bf2}$, $A_{vf3,bf3}$ and $A_{vhi,bhi}$ are given by Eq. (6), where, A_1 and A_2 are the spectral amplitudes at center of gravities in regions 0.75-2.5 kHz, 1.5-3 kHz and 5-7 kHz, of burst and voicing spectrum respectively. Finally, similar to [10], relative amplitudes in 8 frequency bands are computed by using the burst and voicing average power spectra obtained in Sec. V A. resulting in 8 additional features, RAB1 to RAB8.

VOT is further included as a dynamic feature. In all, there are 11 static and 14 dynamic features resulting in 25 AP features.

C. Results of classification

Based on the discussion in Sec. V B, a feature vector is formed comprising the 25 AP features. A diagonal covariance GMM classifier was trained on the Marathi words database (with manually marked landmarks) using EM algorithm with 1, 3, 5 and 8 mixtures per class for the four-way classification of PoA. Further, to obtain comparisons with general spectral vector approaches, MFCC and their first and second derivatives were computed over an interval starting from the burst onset to 100 ms (assuming that the VOT of unvoiced stops does not exceed 60 ms) after that. The speech signal in this interval is divided into 11 frames by centering 20 ms Hamming window at the burst onset and sliding the window by 10 ms each time. A feature vector of MFCC (only the lower coefficients: 2 to 6 were considered due to the superior performance reported in [14]) and their and first and second derivatives were obtained for each frame, resulting in a 135-dim feature vector. This feature vector was reduced to 25 elements (to make it comparable with the dimensionality of AP feature vector) by using a mutual information based greedy algorithm [15]. A GMM classifier was trained on the training data MFCC vectors. On the testing data, both the feature vectors (AP features and MFCC-based of length 25) were computed for both manual and automatically detected landmarks which resulted in four sets of feature vectors. Thus computation of AP features involved both the landmarks while MFCC used only the burst onset. Classification results in % accuracy (% of 162 and 161 test tokens identified correctly for manual labeling and automatically labeled data respectively) are given in Tab. 3.

Feature set	No. Of GMM mixtures			
	1	3	5	8
Manual labeling				
AP features	83.33	88.27	85.19	88.89
MFCC+δ+δ^2	72.84	80.25	86.42	87.04
Automatic labeling				
AP features	81.37	85.71	80.75	88.82
MFCC+δ+δ^2	69.56	83.85	85.71	86.34

Tab. 3. *Classification results for both manual and automatically detected landmarks*

VI. CONCLUSION

The automatic landmark detection methods have shown good localisation and accuracy for burst and voicing onsets. Most of the deletions involved voiced stops due to the presence of the voice bar at low frequencies during closure causing the closure to go undetected. Some of the deletions were due to weak release burst in the unaspirated stops for which ROR was not prominent. Insertions are found for vowels following silence, where energy increases sharply as also noted in [16].

A comparison of the classification performances of the AP features with MFCC-based features is reported. It must be noted that, unlike the traditional HMM speech recognition systems which employ regularly spaced frame-based MFCC vectors, the present work obtained MFCC feature vectors aligned with acoustic landmarks. The classification performance of the proposed set of 25 AP features derived from a study of burst and voicing spectrum characteristics across the four PoA of Marathi stops is found to be comparable to that of the 25-MFCC based features (obtained by reducing the 135-MFCC, comprising of lower order MFCC and their derivatives). Whether this continues to hold outside the clean recording and matched training-testing conditions remains to be investigated. It may also be noted that the AP features were derived from a restricted region of the signal (at the most 15 ms interval of the speech signal around the burst onset and 15 ms interval in the voicing region) while the MFCC based features were derived from a larger part of the signal (100 ms of the signal after the burst onset) potentially capturing formant transition information better. Classification results obtained for manual and automatically detected landmarks are comparable indicating the efficacy of the automatic landmark detection.

In summary, the results presented encourage the further exploration of AP features for segmentation of speech into phonemes, an important step towards keyword spotting based on acoustic matching. The set of AP features can be extended by adding formant transitions or features extracted from a series of spectra from burst onset to the voicing region which capture fine variations to further improve the classification accuracy. Future work will be directed towards the detection and classification of other broad phone classes.

REFERENCES

[1] C.Y. Espy-Wilson, T. Pruthi, A. Juneja and O. Deshmukh, "Landmark-based approach to speech recognition: An alternative to HMMs", *Proc. ICSLP*, pp. 886-889, Aug. 2007.

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

[3] S. A. Liu, "Landmark detection for distinctive feature based speech recognition", Ph.D. Thesis, MIT, May 1995.

[4] A. Salomon, C.Y. Espy-Wilson and O. Deshmukh, "Detection of speech landmarks: Use of temporal information", *J. Acoust. Soc Am.,* vol. 115, no. 3, pp.1296-1305, Mar. 2004.

[5] S.R.M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information", *Proc. ICSLP*, pp. 1133-1136, Sep. 2005.

[6] S. Das and J.H.L. Hansen, "Detection of voicing onset time (VOT) for unvoiced stops (/p/,/t/,/k/) using the Teager energy operator (TEO) for automatic detection of accented English", *Proc. NORSIG*, pp. 344-347, Jun. 2004.

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

[8] 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.

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

[10] J. Hoelterhoff and H. Reetz, "Acoustic cues discriminating German obstruents in place and manner of articulation", *J. Acoust. Soc. Am.,* vol. 121, no. 2, pp. 1142-1156, Feb. 2007.

[11] Z.B. Nossair and S.A. Zahorian, "Dynamic spectral shape features as acoustic correlates for initial stop consonants", J. *Acoust. Soc. Am.,* vol. 89, no. 6, pp. 2978-2991, Jun. 1991.

[12] S.V. Gangashetty, C. Chandra Sekhar and B. Yegnanarayana, "Spotting consonant-vowel units in continuous speech using autoassociative neural networks and support vector machines", *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, pp. 401-410, Oct. 2004.

[13] P. Maragos, J.F.Kaiser and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. on Signal Processing*, vol. 41(10), pp. 3024-3051, Oct. 1993.

[14] V. Karjigi and P. Rao, "Four-way classification of place of articulation of Marathi unvoiced stops from burst spectra", *Workshop on image and signal processing*, Dec. 2007, submitted for publication.

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

[16] P. Niyogi and M.M. Sondhi, "Detecting stop consonants in continuous speech", *J. Acoust. Soc Am.,* vol. 111, no. 2, pp.1063-1076, Feb. 2002.