Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling

V. Karjigi* and P. Rao

Dept. of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai-400076, India. Tel:(91)22 2576 4670/7695
E-mail:{veena, prao}@ ee.iitb.ac.in

Abstract

Unvoiced stops are rapidly varying sounds with acoustic cues to place identity linked to the temporal dynamics. Neurophysiological studies have indicated the importance of joint spectro-temporal processing in the human perception of stops. In this study, two distinct approaches to modeling the spectro-temporal envelope of unvoiced stop phone segments are investigated with a view to obtaining a low-dimensional feature vector for automatic place classification. Classification accuracies on the TIMIT database and a Marathi words dataset show the overall superiority of classifier combination of polynomial surface coefficients and 2D-DCT. A comparison of performance with published results on the place classification of stops revealed that the proposed spectro-temporal feature systems improve upon the best previous systems' performances. The results indicate that joint spectro-temporal features may be usefully incorporated in hierarchical phone classifiers based on diverse class-specific features.

Keywords: Place of articulation, Unvoiced stops, Spectro-temporal features.

1. Introduction

Modern automatic speech recognition (ASR) systems are typically based on a framework of short-term spectra for acoustic features and first-order Markov chains for acoustic modeling. This framework has been mentioned by several prominent speech scientists as a primary limitation, with sophisticated language models serving to compensate the errors in phone-level recognition. It is widely accepted that one promising direction for improving raw phone recognition scores is to embed longer-term characterisics explicitly within the acoustic features (Morgan et al., 2005; Chen et al., 2004; Hermansky et al., 1999). The speech signal dynamics observed over a phone duration are linked to articulatory gestures arising in speech production and thus are expected to represent important phone characteristics. While time-differencing of coefficients representing frame-level spectra are traditionally used to capture the signal dynamics, this serves to exploit only very short-term dependencies relative to overall duration of a phone. An alternate approach to phone-level recognition is "segment based" where the entire segment, and therefore the dynamics over the phone duration, are represented by a single feature vector (Halberstadt, 1998; Glass, 2003). Such an approach requires a prior step of segmentation of speech into the variable duration phone-sized units. For instance, a segmentation algorithm is applied to the extracted frame-based acoustic feature sequence to obtain segment boundaries of broad phonetic categories such as silence, sonorant, stop and fricative. Features specific to each broad class, extracted over the segment duration, are used next for further manner and place classification of phones (Niyogi and Ramesh, 2003).

While segment-level recognition offers a framework for a variety of phone-class specific features including acoustic-phonetic features to be exploited, widely studied features are the more easily computed spectral representations that capture the gross spectral shape in each analysis frame across the duration of the phone. With frame-level features extracted in relation to specific temporal landmarks in the speech signal, acoustic cues important to phonetic contrasts are better represented (Glass, 2003). Next, the temporal dynamics within a segment can be captured implicitly by simply concatenating the feature vectors of each frame of the segment (Forrest et al., 1988; Bunnell et al., 2004; Hou et al., 2007). Frame-wise spectral coefficients were averaged over selected short sub-segments based on observed stationarity before concatenation (Halberstadt, 1998). However, the encoding of temporal dynamics can be achieved efficiently by a more explicit representation. Assuming that features change slowly with time, each feature's temporal trajectory was encoded using a small number of discrete cosine transform (DCT) coefficients (Nossair and Zahorian, 1991; Wang et al., 1996). Further, motivated by findings on the spectro-temporal processing of the auditory system, joint spectro-temporal features have been used for phone classification via Gabor filter functions or as localized two-dimensional DCT (2D-DCT) applied over patches in a time-frequency representation of the signal (Kleinschmidt, 2002; Rifkin et al., 2007; Ezzat et al., 2007; Bouvrie et al., 2008).

In the present work our focus is on classification of stop consonants using features that capture the dynamic characteristic. In particular, we consider the class of unvoiced stops with a goal to achieve accurate classification of place of articulation (PoA). Unvoiced stops are weak energy, short duration phones characterized by the movement of articulators from the release closure until the adjoining vowel steady state is attained. Cues to place of articulation are hence expected to be linked to the spectro-temporal dynamics throughout the phone segment which includes the burst release, when present, and transitions to or from the adjacent vowel in the typical syllable context. Dynamic characteristics include formant transitions (Fant, 1973) and temporal change in relative enegry in specific frequency bands (Kewley-Port, 1983). In this work, we investigate features that capture the spectro-temporal variation including 2D-DCT and bivariate-surface model coefficients. The latter, obtained by polynomial modeling of the time-frequency surface, have not been previously applied to phone classification. Such a representation potentially provides a different and non-separable approximation of the joint time-frequency surface from that available from periodic transforms such as the DCT. Particular attention is paid to the choice of analysis intervals over which features are extracted from across the stop phone segment. While some automatic classification methods use a fixed duration analysis interval around a detected landmark (Halberstadt, 1998; Nossair and Zahorian, 1991), a few use the actual variable duration sub-segment (Halberstadt, 1998; Bouvrie et al., 2008). We test both approaches to segment selection in our experiments.

There is a lot of work on classification of stops in English which differ in three PoAs. Unlike English, most Indian languages have more than three PoAs, and unvoiced stops exhibit presence or absence of aspiration, a phonemic feature. The automatic classification of unvoiced stops of Indian languages has been investigated rarely (Datta et al., 1980; Sekhar and Yegnanarayana, 2002).

In the present work, we have evaluated our stop PoA classification methods on the unvoiced stops of American English and Marathi datasets. Stops in Marathi differ in four PoAs, namely labial, dental, retroflex and velar. 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. The retroflex and dental stop consonants, which are used to achieve meaningful contrasts in Marathi, would typically be both categorized as alveolar [t] by a native English listener. Each of these four PoAs occur in the four manner classes viz. unvoiced-unaspirated (UVUA), unvoiced-aspirated (UVA), voiced-unaspirated (VUA) or voiced-aspirated (VA). We evaluate the proposed features for classification of UVUA stops in Marathi in addition to standard American English speech database (TIMIT) with three PoAs for UV stops.

To sum up the objectives of this work, we investigate segment-based classification of unvoiced stops by (1) joint spectro-temporal features, including new polynomial features not previously applied to this task, and (2) evaluation of the spectro-temporal features on Marathi datasets with four PoAs as well as on TIMIT datasets. In the next section, we present a review of methods that specifically target joint spectro-temporal characteristics for phone classification. The spectro-temporal representations investigated in the present study are then described. Finally, the results of automatic classification experiments on English and Marathi speech databases are discussed.

2. Spectro-temporal features for phone classification

The critical-band based spectral processing of the auditory system is well understood and has been incorporated into ASR front-ends in the form of mel-scale filtering. However there is growing evidence that the auditory system exhibits spectro-temporal processing of speech signals with auditory neurons tuned to modulations of an extent on the order of 50-200 ms (Chi et al., 2005). Further, it is found that neurons are sensitive to the direction of spectro-temporal patterns, indicating that a combined spectro-temporal processing is a more appropriate description of perception rather than that of sequential stages of spectral and temporal filtering (Depireux et al., 2001; Obleser et al., 2010). Motivated by such findings, several recent works have investigated features that represent joint spectro-temporal variation for phone classification. In this section, we review past work and make a case for the specific new work proposed in this paper.

2.1. Review of related past work

Statistical dependencies of mel frequency cepstral coefficients (MFCCs) across time frames were investigated by examining the spatio-temporal correlation structure of the error obtained after trajectory modeling of the individual time varying coefficients over phone segments (Goldenthal, 1994). It was found that the residual error exhibited clear correlations between different quefrency coefficients at different time instants, an aspect not taken into account by the modeling of individual temporal trajectories. That cepstral coefficients are correlated within and across frames, over durations corresponding to a phone segment, was also shown by Ellis (1997).

More recently, spectro-temporal correlations have been examined in the spectral rather than the cepstral domain. This has the advantage that frequency-localized background noise remains localized rather than spreading over all coefficients as in the case of the cepstral transformation. Mel-filter bank outputs have been shown to compare favorably with MFCCs in syllable recognition tasks while proving to be far more robust under a variety of background noise (Sroka and Braida, 2005). Joint spectro-temporal features were derived by applying principal component analysis to blocks of spectrogram obtained from 15 critical-band filters over a frequency range of 4 kHz and time interval of 1 sec to recognize continuously spoken digits (Kajarekar et al., 2001). It was observed that the principal components so obtained were largely aligned horizontally or vertically, and hence a sequence of spectral and temporal analyses steps suffices in place of more sophisticated joint spectro-temporal analysis, at leat at such large time scales.

Another important research trend has been the use of spectro-temporal filters that emulate auditory processing, to extract directional features from spectrogram-like representations of speech (Kleinschmidt, 2002; Rifkin et al., 2007; Ezzat et al., 2007; Meyer and Kollmeier, 2008). Localized 2-dimensional Gabor functions are used to simulate human spectro-temporal perception. Gabor filters with parameters tuned to empirically determined spectro-temporal modulations, including diagonal modulations, are applied to the speech log mel-spectrogram (treated as the primary feature set) to obtain the Gabor features (termed secondary features). The Gabor features facilitate the capture of modulations in specific directions including purely spectral and purely temporal modulations. For instance, a specific formant transition may be captured best by the corresponding diagonal orientation Gabor filter. To limit the dimensionality of the feature vector however, it is necessary to carefully select the filter parameters for the finite set of Gabor filters from a large set of possibilities. Gabor features obtained from an extensively optimized set of filters showed a performance comparable to a standard mel frequency cepstral coefficients-hidden markov model (MFCC-HMM) recognition system in an isolated word recognition task with clean speech, and superior robustness in noise (Kleinschmidt, 2002). Further, it was found that spectro-temporal features are not suited to short duration phoneme labels but more suitable for long duration diphone or word level classification. Gabor features derived from the wideband spectrogram at different levels with different spectro-temporal positions, orientations and scales were used to train a regularized least squares classifier for a TIMIT 20-class vowel classification task (Rifkin et al., 2007). It was shown, by resynthesizing the spectrogram using selected spectro-temporal components, that these Gabor coefficients capture harmonic and formant structure as well as temporal structure of speech (Ezzat et al., 2007). Very recently, discriminant spectro-temporal filters obtained by maximizing phoneme separability in their projected space have been applied to phone recognition (Mesgarani et al., 2009). Each filter was tuned to discriminate one phoneme from all other phonemes resulting in 39 such filters corresponding to the phoneme classes defined for the TIMIT database.

The complexity associated with the data-dependent optimization of the Gabor filters approach was avoided by using the separable 2D-DCT over small, overlapping patches of spectrogram (Bouvrie et al., 2008). The final feature vector comprised of the low-order DCT coefficients of each of a sequence of localized time-frequency patches corresponds to a smoothed spectro-temporal envelope of the associated patch. With the overall phone segment represented by the concatenation of a large number of patches, the use of a separable transform (providing no diagonal structures as such), is expected to be compensated for to an extent. The choice of the patch size in time and frequency is a critical parameter and is selected empirically to 107 ms by 623 Hz for wideband and 56.75 ms by 780 Hz for narrowband spectrograms respectively with step size of 4 ms by 390 Hz in both cases. To obtain fixed-length feature vectors, the features are averaged in 3:4:3 proportion within a phone segment and in the interval of ±30 ms at the phoneme boundary thus averaging over each of 5 sub-segments similar to the work of Halberstadt (1998). Six lower cepstral coefficients obtained for each block are concatenated to form the final feature vector. These features were used for the classification of vowels and stops. It was observed that in the case of stops, in particular, the localized 2D-DCT features provided substantial improvement over traditional MFCCs. This is attributed to the explicit encoding of vertical as well as horizontal spectro-temporal gratings by the 2D-DCT basis functions.

The studies reported above indicate the value of joint time-frequency analysis, at least over intervals on the range of phone durations. That stop phones should be represented by spectro-temporal features that capture directional structures is also reasonable from the point of view of speech production where articulatory movements are known to influence formant transitions. The patch-based 2D-DCT analysis is non-parametric and therefore simpler to effect than the 2D Gabor filters approach with its data-dependent parameter tuning. However, choosing small fixed patch sizes for the separable transform analyses may not be the most efficient way to encode the longer term trends that are neither horizontal nor vertical in the underlying spectro-temporal envelope. The goal of the present work is to investigate a non-parametric approach that more directly models joint spectro-temporal correlations. Before this, we seek to establish in the next section, the presence of joint spectro-temporal correlations for the phone class of interest in this work, namely unvoiced stops.

2.2. Spectro-temporal characteristics of unvoiced stops

To examine the utility of a spectro-temporal representation for unvoiced stops, we present the analysis of a sample set of VC (vowel followed by unvoiced stop) syllables segmented from a database of Marathi words. 320 vocalic transitions were extracted from VC syllables corresponding to unvoiced, unaspirated Marathi stops of equal numbers for each of the four PoA. The variable duration interval from voicing offset (i.e. onset of the stop closure) back until the steady state of the vowel was used for the log mel-spectrogram computed with 40 mel filters with 6.4 ms frames and 1 ms frame shift. Time-normalization of the variable length segments was achieved by uniform time-scaling and interpolation to obtain 8 equi-spaced points on each trajectory. The resulting set of matrices (each of size 40x8) was processed to obtain the non-negative matrix factorization (NMF) basis (Lee and Seung, 1999) for each PoA. Fig. 1 depicts the corresponding first 20 basis images.

We note from Fig. 1 that, contrary to the observations on general fixed duration continuous speech segments reported in the literature (Kajarekar et al., 2001), isolated vocalic transition segments do indeed exhibit joint spectro-temporal features as evident from the diagonal structures in several of the basis images. In particular, Fig. 1 (a) shows a clear falling diagonal band in basis images 12, 14 and 15, which can be explained by the downward F2 transition in this region, characteristic of labials with front vowels. Fig. 1(b) shows slight upward transition of F2 which is the characteristic of velars in center vowel context(basis image 6). Fig. 1 (c) shows upward transition of F2 (basis images 7 and 10) especially when the F2 of vowels is inferior to the dental locus region (roughly around 1600 Hz) which happens in center vowel context. Similar F2 transition but with a higher slope is observed in retroflexes which is the characteristic of vocalic transitions of retroflexes in center vowels (basis images 2, 10 and 18). F2 and F3 transitions in the form of joining of F2 and F3 (F2 moving upwards and F3 moving downwards) while approaching the closure in retroflexes with center vowel context is seen in basis image 10. Also, downward transition of F1 in dental, retroflex and velar stops can be seen (basis images 10, 20 and 7 of dental, retroflex and velar stops respectively) which is the characteristic of these stops in center vowel context. Phonetic analyses presented here for each PoA with

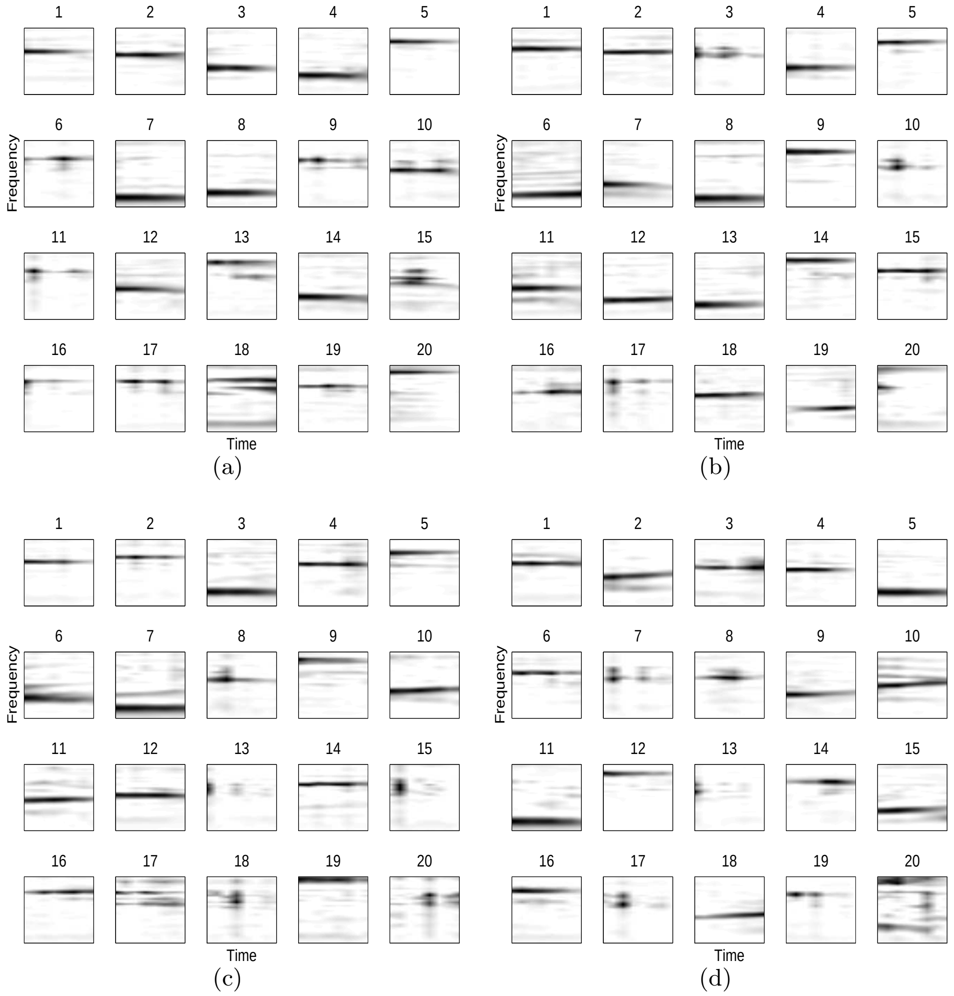


Figure 1: First 20 basis images obtained by applying NMF to datasets of VC transitions from (a) labial, (b) velar, (c) dental and (d) retroflex unvoiced stops

particular vowel context are consistent with Ohala (1995) where acoustic characteristics of VC transitions in Hindi (a language with similar PoAs as that of Marathi) are studied. The presence of bands that are neither horizontally nor vertically aligned in the time-frequency plane demonstrates the existence of joint spectro-temporal correlations that may be usefully exploited for the recognition of stops.

2.3. Need for spectro-temporal features

We have seen that structures of various orientations are a prominent aspect of the spectro-temporal envelope of unvoiced stops and can potentially contribute to discriminatory information for PoA classification. In view of this, it is worthwhile to investigate non-parametric encodings that do not require the extensive parameter tuning of the Gabor filter bank but still are not constrained by the strictly horizontal and vertical gratings of the 2D-DCT. In this work, we explore the classification potential of low-dimensional features that are sensitive to joint spectro-temporal modulations.

As in several of the studies reviewed in Section 2.1, the log-mel spectrogram is a good choice for the primary level feature from which the joint spectro-temporal features are estimated. The mel spectrogram is motivated by the auditory mel scale and can be viewed as a time-frequency resolution optimized spectrogram for speech signals. The log transformation serves to reduce the dynamic range to provide a smooth spectro-temporal envelope representation. In the next section, we review the two sets of secondary features derived from the smooth spectro-temporal envelope that are investigated here for the classification of unvoiced stops namely, the 2D-DCT coefficients and the newly proposed bivariate-polynomial surface coefficients.

3. Joint spectro-temporal representations

The log-mel spectrogram can be viewed as a surface on the time-frequency grid specified by frame number and mel filter index. Fitting the surface with a low-order model could provide a smooth representation that potentially captures the dominant spectro-temporal modulations. The 2D-DCT analysis is one approach to surface modeling. Polynomial basis functions rather than the cosine basis can help to overcome the constraints imposed by the separability of the 2D-DCT representation. Recently second- and third-degree bivariate polynomials fitted to vocal tract area and to line spectral frequencies in consonant-vowel and vowel-consonant tranistions were found to be useful in PoA distinction (Pandey and Shah, 2009). The purpose was to interpolate the unknown area values during the closure of stops from measurements in the two regions on either side within the vowel-consonant-vowel (VCV) utterance in order to provide visual feedback on achieved articulation to speech learners. The system was tested on clearly articulated isolated VCV utterances. We next describe the implementation of the two methods of spectro-temporal envelope modeling investigated in this work.

Let $((x_0, y_0), z_0), ((x_1, y_1), z_1), ..., ((x_{N-1}, y_{N-1}), z_{N-1})$ be the set of N sample points of the time-frequency surface to be fitted. Here x is the temporal index that corresponds to frame number, y is the frequency index that corresponds to the mel filter number and z is the feature value (log amplitude) at (x, y). The spectro-temporal envelope is the surface represented by $z = f(x, y)$.

3.1. 2D-DCT modeling

Let $z = f(x,y)$ be a matrix of N points ($N = PQ$) where P and Q are equal to number of mel filters and number of frames in the segment respectively. Then $F[k, l]$, its 2D-DCT is given by,

$$F\left[k,l\right]=\frac{2w[k]w[l]}{\sqrt{PQ}}\sum_{x=0}^{Q-1}\sum_{y=0}^{P-1}f\left(x,y\right)\cos\frac{\pi l\left(2x+1\right)}{2Q}\cos\frac{\pi k\left(2y+1\right)}{2P} \tag{1}$$

where $k = 0, 1, ...P - 1, l = 0, 1, ...Q - 1$ and $w\left[k\right] = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0 \\ 1 & \text{if k} \neq 0 \end{cases}$

The number of frames in the segment varies with segment length. However the cosine basis used above inherently effects the time-normalization of the basis vector thus making the coefficients of corresponding variable length segments directly comparable. To obtain a low dimensional feature vector for classification, only a fixed number of low order coefficients of the 2D-DCT are retained. This provides a smooth approximation,$\hat{f}(x, y)$, to the underlying spectro-temporal envelope $f(x, y)$.

3.2. Polynomial surface modeling

The spectro-temporal surface is approximated by a bivariate polynomial function with the approximation improving the fit as the polynomial order increases. A second degree bivariate polynomial is given by

$$\hat{f}\left(x,y\right)=c_0+c_1x+c_2x^2+c_3y+c_4y^2+c_5xy \tag{2}$$

A third degree bivariate polynomial is given by

$$\hat{f}\left(x,y\right)=c_0+c_1x+c_2x^2+c_3x^3+c_4y+c_5y^2+c_6y^3+c_7xy+c_8x^2y+c_9xy^2 \tag{3}$$

The number of coefficients in 2^{nd} and 3^{rd} degree bivariate polynomial is 6 and 10 respectively. The task here is to compute the vector of coefficients such that mean squared error between the actual and the reconstructed surface as given by (2) or (3) is minimized (Dyer and He, 2001). For N points on the actual surface $f(x, y)$ there will be N such equations with

$$f\left(x_0, y_0\right) = z_0, f\left(x_1, y_1\right) = z_1, ...f\left(x_{N-1}, y_{N-1}\right) = z_{N-1} \tag{4}$$

The system of simultaneous equations can be represented in matrix form as follows.

$$VC = F \tag{5}$$

where $F_{_{NX1}}$ is the vector with

$$F^T = [z_0 \; z_1 \; ... \; z_{N-1}] \tag{6}$$

$C_{_{DX1}}$ is the vector of coefficients to be estimated and $V_{_{NXD}}$ is the Vandermonde matrix that is formed by evaluating D terms each at N points. To use a bivariate polynomial with D coefficients we require $N \geq D$. If $N = D$ then there exists only one solution. But if N>D, the system is overdetermined and a least squares error approach is applied.

Since the work reported later in this paper showed that polynomial orders beyond 3 were linked to the problem of matrix ill-conditioning, we restrict the description below to 3^{rd} degree bivariate polynomial fitting. The Vandermonde matrix and coefficient vector for 3^{rd} degree bivariate polynomial are given by,

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & y_0 & y_0^2 & y_0^3 & x_0 y_0 & x_0^2 y_0 & x_0 y_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 & y_1 & y_1^2 & y_1^3 & x_1 y_1 & x_1^2 y_1 & x_1 y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N-1} & x_{N-1}^2 & x_{N-1}^3 & y_{N-1} & y_{N-1}^2 & y_{N-1}^3 & x_{N-1} y_{N-1} & x_{N-1}^2 y_{N-1} & x_{N-1} y_{N-1}^2 \end{bmatrix} \tag{7}$$

$$C^T = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 & c_8 & c_9 \end{bmatrix} \tag{8}$$

The mean squared reconstruction error for the segment with N surface values is

$$E = \sum_{n=0}^{N-1} \left[f(x_n, y_n) - \hat{f}(x_n, y_n) \right]^2 \tag{9}$$

where $f(x, y)$ is the actual value and $\hat{f}(x, y)$ is that obtained by surface fitting. By equating the derivative of error wrt c_i to 0 for $i = 0$ to $D - 1$, we solve for the coefficients by

$$C = (V^T V)^{-1} V^T F \tag{10}$$

$(V^TV)^{-1}V^T$ is the pseudo inverse of V. To compare the coefficients across similar segments of different durations, it is necessary to normalize for segment duration (Gish and Ng, 1996). This normalization is done by modifying the Vandermonde matrix. Accordingly, the third degree bivariate polynomial function given in (3) is rewritten as

$$\begin{aligned}\hat{f}(x,y) = c_0 + c_1\left(\frac{x-1}{Q-1}\right) + c_2\left(\frac{x-1}{Q-1}\right)^2 + c_3\left(\frac{x-1}{Q-1}\right)^3 + c_4\left(\frac{y-1}{P-1}\right) + c_5\left(\frac{y-1}{P-1}\right)^2 \quad (11)\\ + c_6\left(\frac{y-1}{P-1}\right)^3 + c_7\left(\frac{x-1}{Q-1}\right)\left(\frac{y-1}{P-1}\right) + c_8\left(\frac{x-1}{Q-1}\right)^2\left(\frac{y-1}{P-1}\right) + c_9\left(\frac{x-1}{Q-1}\right)\left(\frac{y-1}{P-1}\right)^2\end{aligned}$$

An important consideration in polynomial surface modeling is the problem of large approximation errors due to poor fitting at the boundaries of the specified surface (Harder et al., 2010). This can be overcome by separately fitting overlapping sub-segments of the spectro-temporal envelope to obtain a more uniform approximation of the entire region of interest. The choice of the size and number of sub-segments governs the trade-off between the surface approximation accuracy and the total number of coefficients. It may be noted that polynomial fitting is computationally more expensive than 2D-DCT.

In the next section, we explore the potential of the two distinct methods of spectro-temporal surface modeling, 2D-DCT and bivariate polynomial, for the representation of phones characterized by prominent formant transitions over the segment duration. Feature extraction and classification methods are presented together with a description of the classification experiments.

4. Automatic classification experiments

The performances of the joint spectro-temporal feature sets discussed in Section 3 are evaluated by classification experiments on unvoiced stops of English and Marathi datasets. In this section, the test speech datasets are described followed by a discussion of speech segmentation and landmark-based feature extraction. The feature vectors are input to a Gaussian mixture model (GMM) classifier previously trained on labeled training data.

Perception experiments for PoA identification of Marathi unvoiced stops conducted by playing different sub-segments of CV (unvoiced stop followed by vowel) and VC syllables to 5 listeners (not reported here) and those reported in the literature for French, Spanish, English and Hindi (Neagu and Bailly, 1998; Feijoo et al., 1999; Bonneau et al., 1996; Nossair and Zahorian, 1991; Smits et al., 1996; Ahmed and Agarwal, 1969; Ohala and Ohala, 1998) indicate that both the burst and vocalic transition regions individually contain useful acoustic cues to place of articulation. In the case of unreleased stops as in VC context, the vocalic transition provides strong acoustic cues to stop place of articulation. The variable duration segments corresponding to these distinctive regions can thus serve as the ideal input for automatic phone classification. However many automatic phone classification systems operate in the landmark based mode where a fixed duration segment is extracted in the vicinity of a chosen acoustic landmark e.g. (Nossair and Zahorian, 1991). In view of this, we consider both variable and fixed duration segments for PoA classification in this work.

4.1. Databases

The features investigated in this work are evaluated on two distinct datasets viz. American English continuous speech as provided in the TIMIT database, and a Marathi words database specially created for the purpose. Additionally, to facilitate comparison with performance results reported in the literature, specific subsets of the TIMIT database are tested as well.

4.1.1. TIMIT CV and VC dataset

TIMIT is a database of continuous speech comprising 10 sentences uttered by each of 630 American speakers of 8 dialect regions (Lamel et al., 1986). The database is partitioned into a training set of 462 speakers and test set of 168 speakers. The TIMIT database provides manually obtained phonetic transcriptions for all the speech. CV and VC segments extracted from a specific set of 50 test speakers was used as development set with the remaining forming the test set (Halberstadt, 1998). The train set was used for training phone models, the development set for tuning classifier parameters, and the test set for final performance evaluation. Details of the distribution of unvoiced stops appear in Table 1. CVs include word initial and word medial stops. VCs include word medial and word final stops.

4.1.2. Marathi words CV and VC dataset

The CV tokens were obtained from utterances containing the four unvoiced unaspirated stops in word initial position appearing with one of the eight vowels of the language recorded at 16 kHz in quiet conditions in the lab by 20 native speakers (10 male and 10 female). The 8 vowels are categorized as follows: {front: ɪ,i,e}, {back: ʊ,u,o}, {center: ə, ɑ}. Two distinct meaningful mono-/bi-syllabic words were chosen for each CV combination. The words were recorded by each speaker by embedding in two Marathi carrier phrases / ɦɑ ʃəbd̪ ----- ɑɦe/ and /mi ----- mənu kɑ?/ which mean "This word

Table 1: Details of the TIMIT CV and VC data used in classification experiments

	p	t	k	Total
CV syllables				
Train	1533	2973	2487	6993
Development	183	291	246	720
Test	424	683	534	1641
VC syllables				
Train	1541	3603	3880	9024
Development	152	383	384	919
Test	378	898	925	2201

is ———" and "Shall I say ——?". Thus there were 16 words x 2 instances x 20 speakers=640 tokens per stop.

The VC tokens were obtained from utterances containing the four unvoiced unaspirated stops in the word final position following one of the eight vowels of the language recorded at 16 kHz in quiet conditions in the lab by 20 native speakers (10 male and 10 female). Two distinct meaningful mono-/bi-syllabic words were chosen for each VC combination with UVUA stops in the final position. The words were uttered continuously once by each speaker. Thus there were 16 words x 1 instance x 20 speakers=320 tokens per stop.

The spoken words in both CV and VC datasets are isolated and normalized so that maximum amplitude of each word utterance is the same to compensate for varying loudness levels of different speakers. Marathi data was not divided into training/development/test because of the small size. Instead we use the same data for training and testing for landmark detection experiments and 20-fold cross validation (leave one speaker out) in the classification experiments.

4.1.3. Acoustic differences between English and Marathi stops

Apart from the spectral differences due to the differing places of articulation of English stops and Marathi stops, it is observed that prominent durational differences are present between English and Marathi unvoiced stops even with same PoA. A statistical summary of voice onset time (VOT) across the unvoiced stops followed by a vowel from Marathi words and from the TIMIT train datasets appears in Table 2. We note that the Marathi unaspirated stops have considerably lower VOTs. This may be attributed to the fact that the English UV stops are often aspirated and therefore have longer VOT. Marathi on the other hand contains a linguistically contrasting set of unvoiced aspirated stops corresponding to the same 4 PoAs. Longer VOT is an important cue for aspiration. The durational differences will influence our choice of analysis parameters for feature extraction.

Table 2: VOT mean and standard deviation for the Marathi UVUA stops and English UV stops

Place of articulation (Marathi)	VOT(ms)		Place of articulation (English)	VOT(ms)	
	Mean	Standard deviation		Mean	Standard deviation
Labial	17.2	7.9	Labial	41.5	19.8
Dental	14.9	5.8	Alveolar	48.3	21.5
Retroflex	9.5	3.7			
Velar	27.6	9	Velar	59.7	22.1

4.2. Landmark detection for segment-based feature extraction

The accurate localization of segment boundaries is a pre-requisite for segment-based feature extraction and classification. Burst release, voicing onset point and start of vowel steady state constitute the important temporal landmarks or boundaries for the CV context, while end of vowel steady state and voicing offset time correspond to the VC context as shown in Fig. 2.

In this section we discuss the automatic detection of landmarks in the speech signal based on previously reported work (Patil et al., 2009). The initial segmentation of the utterance into broad phone classes is carried out via standard MFCC-HMM recognition (similar to the approach of Niyogi and Ramesh, 2003 and Glass, 2003) based on the observation that phone recognition errors in the HMM based recognition system are largely restricted to within-manner class confusions. Next a refinement of the hypothesized phone boundaries of unvoiced stops is carried out to locate the above mentioned landmarks.

4.2.1. Broad phone segmentation

The initial segmentation into broad classes is carried out by an HMM recogniser with the broad class phone models trained on the TIMIT training set for English and on the complete Marathi words data for Marathi phone models. The distinct broad classes are vowels, semivowels, nasals, unvoiced stops, voiced stops (stops are further separated based on presence or absence of aspiration in case of Marathi), fricatives, affricates and closures of plosives

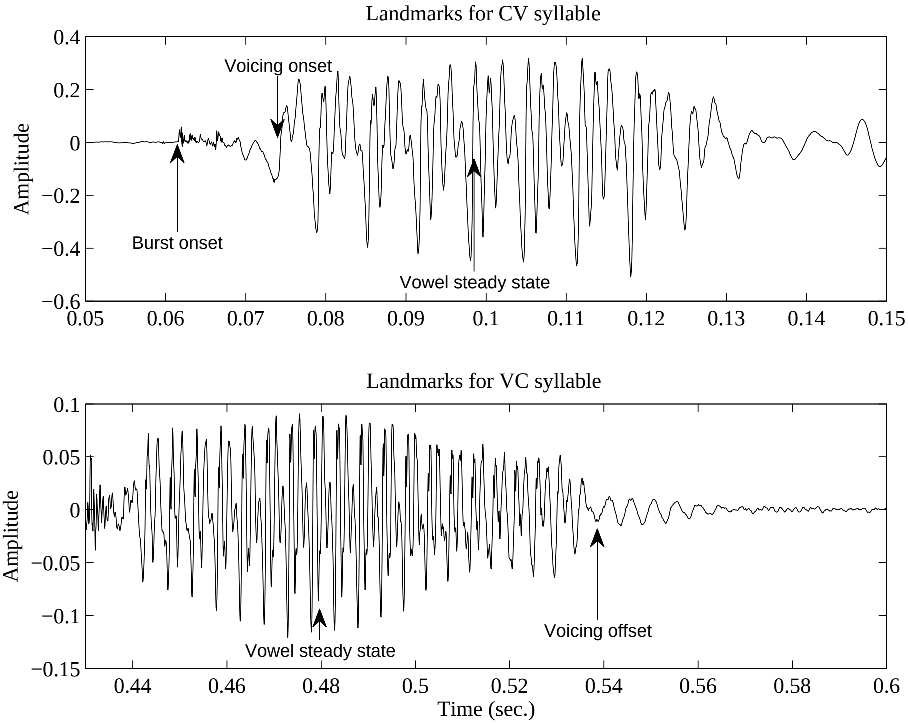


Figure 2: Temporal landmarks used in segmentation (with Marathi UVUA stop t̪ and vowel ɑ)

(separated as belonging to voiced and unvoiced plosives because of relatively large low-frequency energy in the closure of voiced plosives compared to unvoiced in Marathi). The broad class models were context independent, 3-state HMM with diagonal covariance trained with flat-start initialization. The standard 39 dim MFCC, delta and acceleration feature vector was computed using 20 ms frames at 10 ms intervals.

The system was observed to provide a recognition accuracy of 75% on TIMIT unvoiced stops and 88% on Marathi unvoiced unaspirated stops. The phone boundaries of the correctly recognized stops correspond to the landmarks of burst release, voicing onset and voicing offset. However, the HMM decoder segmentation is not accurate enough for use in PoA feature extraction, and is further refined as described in this section. In order to maximize the data available to us for the evaluation of the PoA classification (the main focus of the present study), we included the entire set of unvoiced stops in the PoA classification tests irrespective of whether or not the HMM recognizer made errors. In the case of unvoiced stops missed by the HMM recognizer, the phone boundaries were obtained by using forced alignment with ground truth broad class labels. Future work will address improvements to the accuracy of the HMM broad class recognizer by augmenting the MFCC feature set with suitable acoustic-phonetic features.

4.2.2. Landmark location refinement

The first-stage detected phone boundaries are corrected for average bias as measured on the training set, and are then refined by locating abrupt changes, in a suitable vicinity, of acoustic features appropriate for each landmark type as described next (Patil et al., 2009).

Burst onset: The rate-of-rise (ROR) of energy in the high frequency band of 3.5-8 kHz is computed every 1 ms over the chosen search interval using windows of 6 ms duration. The ROR is computed as the smoothened energy difference over the interval of 10 ms and 2 ms in case of TIMIT and Marathi respectively. This difference in the parameters for TIMIT and Marathi is motivated by the linguistic differences in average burst duration as presented in Table 2. The instant corresponding to the global maximum in ROR is identified as the burst onset.

Voicing onset: The ratio of short time signal energy to LP residual energy in the voiced region is high compared to that of unvoiced regions. A 16th order linear predictor is used here. As the voicing onset coincides with a glottal cycle epoch instant, the epoch instant at which this ratio exceeds an empirically decided threshold is identified as the refined voicing onset point (Prasanna et al., 2001).

Voicing offset: for VC syllables is determined as the instant at which short time energy computed every 1 ms over the chosen analysis interval falls below a fixed threshold and does not rise again in an interval of a 10 ms. This threshold was 4% and 10% of maximum energy in case of TIMIT and Marathi stops respectively.

Vowel steady state onset/offset: The instant at which steady state starts or ends in the case of vocalic transitions was detected automatically by the time instant at which the energy in the low frequency band (<1 kHz) reaches its maximum value over the duration of 20 to 50 ms from the onset/offset.

4.2.3. Landmark detection performance

Accuracy of automatic landmark detection in terms of the percentage correct detection within specific tolerance intervals about the manually marked boundaries appear in Table 3 for TIMIT UV stops and Marathi UVUA stops respectively. The refinement stage provides considerable improvement in localization consistent with previous observations (Patil et al., 2009). The lower performance for TIMIT compared to Marathi can be attributed to the great variation in speakers, sentences of TIMIT compared to Marathi words databases and analysis of only word initial and final stops of Marathi compared to the inclusion of word medial stops also in TIMIT.

Table 3: Automatic detection of required landmarks: % correct detection

Deviation about the manual label	Burst onset		Voicing onset		Voicing offset	
	I stage	II stage	I stage	II stage	I stage	II stage
TIMIT data						
±5 ms	26.5	64.0	46.6	57.7	50.8	62.8
±10 ms	50.9	87.2	78.0	81.7	82.5	87.4
±15 ms	68.4	92.4	93.4	89.6	93.8	94.7
±20 ms	82.0	93.9	97.6	93.4	97.2	97.2
Marathi UVUA data						
±5 ms	75.9	89.9	36.1	60.5	39.9	70.3
±10 ms	94.8	95.7	69.8	84.8	70.6	89.8
±15 ms	98.1	97.5	89.8	92.9	88.8	94.0

4.3. Feature extraction

As discussed in Section 3, the surface formed from mel-filterbank log energies computed at uniform time intervals throughout the selected segment duration, is represented by surface coefficients. The latter are called secondary features that are derived from the primary features i.e. the mel-filterbank energy vectors. The analysis segment corresponds to a selected region of the CV or VC token. In this section, feature extraction including choice of analysis parameters is discussed.

4.3.1. Primary features

Due to the differences in the average duration of TIMIT and Marathi stop segments as presented in Table 2, the parameters used for primary feature extraction are differently chosen, as detailed in Table 4, to span the region of interest with sufficient sampling of features. The analysis window duration is chosen empirically. Windowed data frames are extracted from the specified interval and pre-emphasized followed by a 256-point DFT. Forty Mel filter bank (MFB) log energies are computed with filter center frequencies logarithmically spaced in the range of 0-8 kHz using equal maximum gain triangular filters.

Table 4: Analysis parameters for primary feature extraction

Analysis	English	Marathi
Window size (ms)	15	6.4
Hop size (ms)	5	1
Frequency range (kHz)	0-8	0-8
FFT size	256	256
No. of MFB channels (bins)	40	40

For CV syllables, both burst and vocalic regions are available for analysis. For VC syllables, the burst may not be released and hence we restrict the analysis to transitions only. Considering this, we define both fixed and varying duration intervals as detailed in Table 5. The two distinct intervals considered for full CV are the following.
(i) Fixed duration from burst onset as in Nossair and Zahorian (1991). Classification experiments on the development data with interval durations ranging from 20 ms to 70 ms indicated that a duration of 60 ms was best for the TIMIT stops while 40 ms worked best for the Marathi UVUA stops.
(ii) Full burst interval followed by a fixed 20 ms after voicing onset.
Further, transition-only intervals with variable duration, were considered for both CV and VC.

Table 5: Phone intervals used in primary feature extraction

Type of segment	Notation used	Analysis interval
Full CV	cv-full-bd	fixed interval of 'd' ms from burst onset
	cv-full-bt$d$	burst onset to a fixed duration of 'd' ms after voicing onset
CV Transition	cv-tv	CV vocalic transition
VC Transition	vc-tv	VC vocalic transition

4.3.2. Secondary features

The log-mel spectrogram of the phone sub-segment of interest, obtained as described above, is a block of size $38XQ$. The first MFB log energy is not used because of very low frequency values (<90 Hz) involved and the last band (7-8 kHz) is not considered to account for microphone characteristics.

Low order coefficients obtained by implementing a separable transform on a full time-frequency segment was found to capture the dynamic characteristics of stops (Nossair and Zahorian, 1991). Bouvrie et al. (2008) tried to improve on it by segmentation into local patches in time and frequency. To get a better fit of the low order models to the spectro-temporal envelope it is desirable to segment the log-mel scale spectrogram along the frequency axis and separately model each block. Frequency subdivision is acceptable because correlation is expected to be restricted to neighboring frequencies. Time subdivision, on the other hand, is avoided in order to ensure efficient coding of the temporal dynamics of stop articulation. Frequency subdivision has the added advantage of restricting the spread of any frequency localized noise to a limited number of coefficients.

Block sizes and amount of overlap (required to overcome the effect of large approximation errors at the surface boundaries) were decided experimentally by running classification experiments with different sizes and amount of overlap and by choosing the set of parameters that gave the best results on the development set. Different configurations with varying overlaps were tried which included 4 blocks corresponding approximately to the first four formant ranges, and uniform distribution of mel bands across 6, 9 or 12 blocks. Of these, the configuration with 6 blocks was found to work best. The 6 block configuration is shown in Table 6.

Table 6: Parameters used for secondary feature extraction

Block number	Frequency range used for each block (Hz)
1	92 to 517
2	445 to 1059
3	955 to 1845
4	1693 to 2979
5	2760 to 4621
6	4303 to 6994

The choice of degree of the bivariate polynomial for surface modeling is driven by the assumption that the movement of articulators follows low order dynamics. Use of 2^{nd}, 3^{rd} and 4^{th} degree bivariate polynomials was investigated by computing the polynomial coefficients over a 60 ms duration interval from burst onset of CV syllables so as to approximately cover both burst and vocalic transitions as in Nossair and Zahorian (1991). It was found that using the 4^{th} degree polynomial often resulted in ill conditioning, and that the 3^{rd} degree with 10 coefficients provided for higher classification accuracy over 2^{nd} degree with 6 coefficients. To retain the same dimension feature vector from 2D-DCTs 10 lower order coefficients (upper 3x3 sub matrix and fourth coefficient in the first row) are considered. Hence each block was modeled using 10 2D-DCT coefficients, and separately, with 10 BVP (bivariate polynomial) surface coefficients obtaining feature vector dimensions of 60 corresponding to the 6 block configuration.

4.4. Reconstruction based comparison

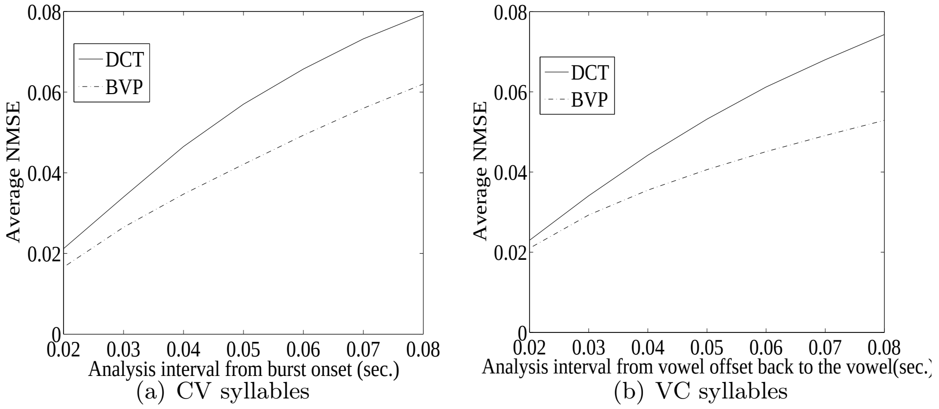
The two different approaches to surface modeling of the spectro-temporal envelope were compared via the normalized mean squared error (NMSE) between the original mel-spectrogram $f(x, y)$ and the reconstructed mel-spectrogram $\hat{f}(x, y)$. The MFB spectrogram over the desired analysis interval was segmented into 6 blocks as in Table 6. The low order 10 DCTs and 10 BVP surface coefficients extracted from each of the six blocks were used to reconstruct each block separately before overlap adding to obtain the full reconstructed log-mel spectrogram. Two types of analysis intervals were considered. These are 1) fixed intervals ranging from 20 ms to 80 ms from the burst onset of CV syllables in steps of 10 ms, and 2) fixed vocalic transition segments of 20 to 80 ms backwards from closure onset for VC syllables in steps of 10 ms. Average NMSE was computed for 6993 CV syllables and 9024 VC syllables extracted from the TIMIT train set.

Fig. 3 shows average NMSE versus segment duration for both burst of CV syllables and transitions of VC syllables. We see that the BVP surface reconstruction has the lower NMSE, and that the difference in performance between the two representations widens as the duration of the segment increases. The higher approximation error may be attributed to the inability of the separable 2D-DCT to capture variations in time-frequency that are oriented away from the strictly horizontal or vertical, such as could arise from joint spectro-temporal modulations.

4.5. PoA classification experiments

4.5.1. Classifier parameter selection

A GMM classifier with full covariance matrices was trained by expectation maximization (EM) algorithm for each task. Because of the difference in the sizes of training data available, the number of mixtures was varied from 1 to 6 in case of TIMIT and 1 to 4 in case of Marathi. All feature sets are evaluated for a fixed set of reduced dimensions (in steps of 10 starting from the full feature vector of dimension equal to 60). The feature vectors are



(a) CV syllables (b) VC syllables

Figure 3: Average NMSE for surface reconstruction from 2D-DCT and BVP coefficients for TIMIT unvoiced stops versus interval duration from (a) burst onset for CV syllables (b) voicing offset back to the vowel for VC syllables

processed by an orthogonal transform to obtain reduced dimension vectors corresponding to the principal components. This transform is determined from a covariance matrix computed from the training dataset feature vectors. Various reduced dimension feature vectors are evaluated using the TIMIT development set to finalize the combination of feature vector dimensionality and number of GMM mixtures based on the best obtained classification accuracy. This exercise is separately repeated for the Marathi data by using cross validation on the entire dataset. In each case, only the combination of dimension and number of mixtures that resulted in the highest classification accuracy is reported.

4.5.2. Classification tasks

Two classification tasks are defined as follows.

Task1: 3-way classification of unvoiced stops extracted from CV and VC syllables of TIMIT with 4-fold model aggregation as in Hazen and Halberstadt (1998).

Task2: 4-way classification of the unvoiced, unaspirated stops of Marathi (both CV and VC syllable) with 20-fold cross validation (leave one speaker out).

5. Results and discussion

As mentioned in Sec. 4.3.2, our implementations of spectro-temporal features, including 2D-DCT, do not use time-segmentation of the surface and thus serve to capture potentially more explicitly the temporal variation across the segment duration. Hence we start with comparing the performance of 2D-DCT computed in this work with related work reported in the literature that use different approaches to surface segmentation before modeling (Nossair and Zahorian, 1991; Bouvrie et al., 2008). Next, we present the results of an exploratory experiment to validate our hypothesis about the superior joint spectro-temporal representation of the BVP features compared with the 2D-DCT on speech segments characterized by prominent formant transitions. We use diphthongs and CV (corresponding to unvoiced stops) of various segment lengths starting from burst release. Finally, the performance of spectro-temporal features in the classification tasks on unvoiced stops as described in Section 4.5.2 are presented. We also compare the classification accuracies achieved by the same features with the accuracies of previous systems as available in the published literature on the specified datasets.

5.1. Comparison of 2D-DCT in this work with related literature

As mentioned in Sec. 4.3.2 secondary features in this work are computed by dividing the time-frequency surface across frequency only, unlike Nossair and Zahorian (1991) where the 2D-DCT was computed over a single unsegmented region and Bouvrie et al. (2008) where the time-frequency surface was divided across both time and frequency. The features of Nossair and Zahorian (1991) were recomputed in this work for comparison because of the non-availability of the published results for the performance of the corresponding features on a standard dataset. Comparison of these features is done in the Table 7 with the help of different classification tasks. With the same data and analysis interval it can be seen that dividing the time-frequency segment across frequency has improved the classification accuracy from 91.0% to 93.5%. Also, comparing the published results of Bouvrie et al. (2008) on the TIMIT core set of all (voiced and unvoiced) stops with the corresponding performance of the 2D-DCT of the present work, we see an improvement with the latter accompanied by a reduction in feature vector dimensionality.

5.2. Preliminary comparison of 2D-DCT and BVP features

Diphtongs: An example of phones with longer duration and large formant deviations is the class of diphthongs. BVP coefficients are expected to

Table 7: Classification accuracies (%) of 2D-DCT in this work and the related work

Feature set	Dimension	Analysis interval	%accuracy
3-way classification of UV stops from CV syllables of TIMIT test set			
Nossair and Zahorian (1991) 2D-DCT implemented in this work	20	60 ms from burst onset	91.0
This work (fixed duration)	60	60 ms from burst onset	93.5
6-way classification of all occurances of stops in the core set of TIMIT			
Bouvrie et al., (2008) published	511	Burst and 30 ms on either side	76.5
This work (variable duration)	61	Burst and 20 ms beyond that	77.6

capture these deviations better than 2D-DCT. Polynomial coefficients were found to better represent the formant tracks in diphthongs compared to a representation using cosine coefficients in a speaker recognition application (Enzinger, 2010). It was also found that spectro-temporal features are more suitable for long duration diphone classification (Kleinschmidt, 2002). Motivated by these reports, an experiment is conducted to evaluate the features 2D-DCT and BVP separately for the classification of diphthongs in American English.

Diphthongs extracted from the TIMIT database are considered for this purpose. The three diphthongs are 'aI', 'ɔI' and 'aʊ' as in 'bike', 'boy' and 'loud' respectively. Details of these in TIMIT train and test sets are given in Table 8. Analysis parameters used for TIMIT UV stops (according to Table 4) are used for primary feature extraction. Secondary features were extracted with the 6 block configuration (given in Table 6). BVP coefficients perform significantly better compared to 2D-DCT on both development and core set in the task of 3-way classification of diphthongs in TIMIT as seen from Table 9. Because of the differences in the size of the train data for the three classes prior probabilities of the classes were considered in classification.

CV segments of varying length The superiority of BVP surface modeling over the separable 2D-DCT for longer time segments for their use in classification of stops is illustrated systematically via the graph of Fig. 4 depicting classification accuracy versus segment duration of CV, starting from burst onset, for the DCT and BVP feature sets (dimension after reduction and

Table 8: Details of the diphthongs in TIMIT

	aI	ɔI	aʊ	Total
Train	1934	304	728	2966
Development	214	30	61	305
Core set	89	16	30	135

Table 9: % Accuracies obtained with classification of diphthongs in TIMIT

	2D-DCT	BVP	p-value
Development	89.8	92.5	0.0433
Core set	89.6	94.8	0.0233

number of mixtures chosen to be the same as that used for the analysis interval cv-full-bd) for both classification tasks.

5.3. Place classification of TIMIT and Marathi unvoiced stops

On the classification tasks defined in Section 4.5.2, we report the performance of the 2D-DCT features and BVP features as extracted from the analysis intervals mentioned in Table 5. Further, a classifier level combination is investigated motivated by the observed complementary behavior of the two sets of spectro-temporal features for segments of different durations. The combination is achieved by the product rule where, for a given class, likelihood values corresponding to each feature set are multiplied, and the class with the largest combined likelihood is assigned to the test token.

Accuracies on the two classification tasks along with significance (p values obtained from McNemar significance test) of the performance difference between 2D-DCT and those after augmenting it with BVP at classifier level appear in Table 10. Classification results are shown for 4 different analysis intervals which include two full-cv regions and two transitions (CV and VC). The class models were trained on features extracted by manual landmark detection. Automatically detected landmarks were applied to feature extraction on the test set. In order to better understand the influence of errors in landmark detection on classification accuracy, classification was also performed with the features computed from manually obtained landmarks and the results are tabulated in Table 11.

From Tables 10 and 11, we see that the classifier combination of 2D-DCT and BVP features is significantly superior to the 2D-DCT classifier

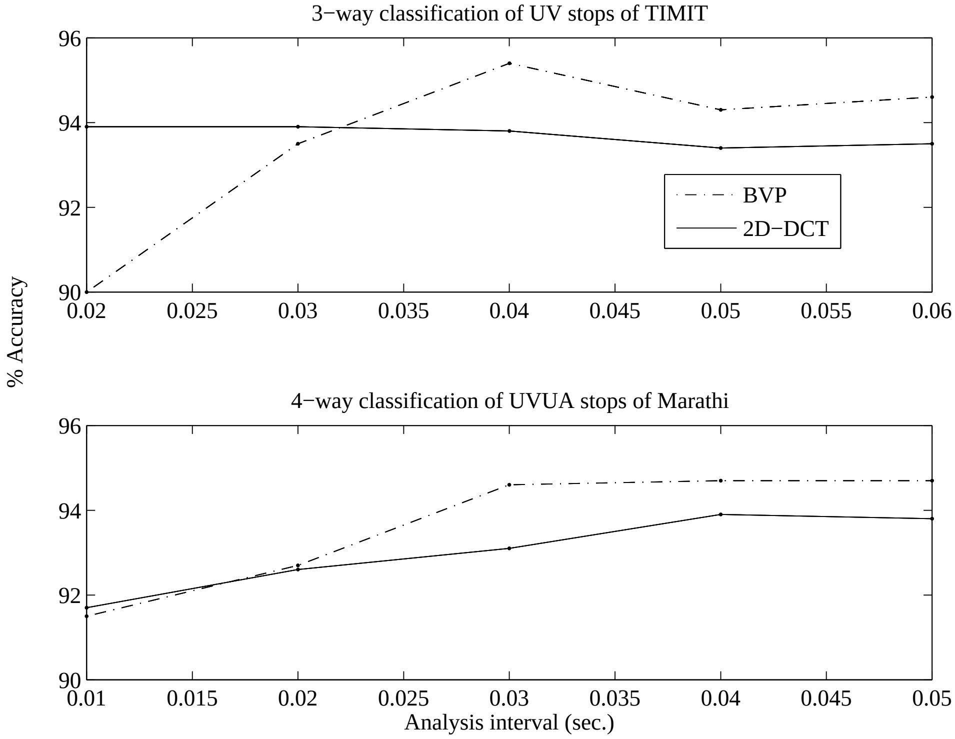


Figure 4: Classification accuracy (%) versus segment duration of CV, starting from burst onset, for the 2D-DCT and BVP feature sets

alone for both datasets and all segment categories. This may be attributed to the complementary behavior of the BVP with its better representation of spectral transitions over longer time intervals, thus bringing in value to the combination with the more standard 2D-DCT representation.

We note that there is an improvement in performance for the CV context when automatic landmark detection is replaced by manual indicating that better landmark detection would help improve system accuracy further. The reduction in performance in CV context based classification due to automatic landmark detection errors is more marked in the case of Marathi data compared with TIMIT data, although landmark detection accuracy itself is superior for Marathi as seen in Table 3. This can be attributed to the fact that Marathi bursts are of smaller duration where a deviation of ±5 ms becomes more significant.

Full CV intervals are considerably more valuable for PoA classification compared to burst-less transitions for both the databases. Also VC transi-

Table 10: Performance of 2D-DCT, BVP and classifier combination (landmarks automatically detected)

Type of segment	2D-DCT	BVP	Product rule	p-value (b/w the 2D-DCT and product rule)
3-way classification of TIMIT UV stops				
cv-full-b60	91.9	92.6	93.5	0.0011
cv-full-bt20	91.2	92.0	93.3	$4.2561\text{x}10^{-5}$
cv-tv	59.9	59.5	62.2	0.0070
vc-tv	82.1	82.9	84.1	0.0013
4-way classification of Marathi UVUA stops				
cv-full-b40	90.5	91.7	91.9	$6.9386\text{x}10^{-4}$
cv-full-bt20	90.3	91.4	92.4	$6.909\text{x}10^{-7}$
cv-tv	62.3	65.5	67.7	$1.3244\text{x}10^{-11}$
vc-tv	80.8	81.3	83.5	0.0032

tions identify the PoA better than CV transitions as seen from Table 10. CV syllables begin with the burst. Therefore an important part of the CV transition is missing when only its vocalic part is considered (Fant, 1973). For both CV and VC syllables of Marathi, the larger confusions were observed to be between dentals and retroflexes across all analysis intervals. The observed trends are consistent with the inferences drawn from the perceptual tests reviewed in Section 4.

5.4. Peformance comparison with available systems

Recent work by Hou et al. (2007) involved a task similar to one considered in the present work, namely, 3-way place classification of unvoiced stops from CV syllables of a subset of the TIMIT database described in Table 12. The best results to date for classification of stops have been obtained by Halberstadt (1998) by the use of phone-specific heterogeneous features in hierarchical classifier combinations using bigram and context dependent models on the TIMIT subset detailed in Table 12. The tasks considered were 3-way place classification of the 6 stops (voiced and unvoiced) and 6-way classification of stops across contexts. Gaussian mixture models were trained on the training set, classifier parameters were tuned on the development set and tested on the core test set. The small size of the core set would make the significance testing somewhat coarse. Features were tested on a larger data set (full test set) in cases where significance tests were conducted.

Table 11: Performance of 2D-DCT, BVP and classifier combination (landmarks manually detected)

Type of segment	2D-DCT	BVP	Product rule	p-value (b/w the 2D-DCT and product rule)
3-way classification of TIMIT UV stops				
cv-full-b60	93.5	94.6	95.4	$1.5705\text{x}10^{-4}$
cv-full-bt20	93.6	95.4	95.5	$3.7753\text{x}10^{-5}$
cv-tv	61.5	61.9	63.4	0.0276
vc-tv	80.1	82.6	82.9	$4.0573\text{x}10^{-6}$
4-way classification of Marathi UVUA stops				
cv-full-b40	93.8	94.7	95.1	$2.1985\text{x}10^{-4}$
cv-full-bt20	93.9	95.3	95.5	$7.6911\text{x}10^{-6}$
cv-tv	70.7	71.2	74.0	$6.7537\text{x}10^{-6}$
vc-tv	79.1	79.4	81.4	0.0061

The full test set involved all data not in the training or development sets. Further, perceptual tests were conducted in that work with the data set comprising 490 tokens (from the development set) that were misclassified by one or other of three different single-classifier systems. Six-way restricted classification accuracies on the same TIMIT core test set have also been reported by Bouvrie et al. (2008) who compared their 2D-DCT patch based features system with Halberstadt's best single-classifier system.

Table 12: Details of subsets of TIMIT database used in related past work

	p	t	k	b	d	g	Total
From (Hou et al, 2007)							
Train	1311	2176	1900				5387
Test	510	711	587				1808
From (Halberstadt,1998)							
Train	2588	3948	3794	2181	2432	1191	16134
Development (DV)	281	413	376	249	239	127	1685
Core test (CR)	136	187	165	132	113	66	799
Full test (FT)	676	954	828	637	602	325	4022
Perceptual test set (PT)	105	92	56	70	99	68	490

First we compare the performance of classifier level combination of 2D-DCT and BVP with 2D-DCT alone for different datasets detailed in Table 12. Then we compare the accuracies obtained by the spectro-temporal feature systems of the present work with the same-dataset accuracies of the above systems as published in the references cited. In each case, only the best results as obtained by the feature dimensionality and classifier parameter selection are presented. The above mentioned literature (Hou et al., 2007; Halberstadt, 1998; Bouvrie et al., 2008) used an interval of complete burst and region around that (±30 ms) for feature extraction. The nearest analysis segment for this in our work is "burst and 20 ms of voicing" (cv-full-bt20) although the 20 ms following the burst need not correspond to vocalic transition in non-CV contexts. Burst and voicing onset landmarks were obtained from the available TIMIT annotation for fair comparison with the experiments of the other authors reported here. To be consistent with Halberstadt (1998) and Bouvrie et al. (2008), burst duration was an additional feature, appended to the dimension-reduced spectro-temporal feature vector.

Halberstadt (1998) used Gaussian mixture models obtained with N-fold model aggregation to get robust models for 6-way stop and 3-way place classification on voiced and unvoiced stops. Full covariance GMMs with 4-fold model aggregation are used for classification in this section. Results reported by Hou et al. (2007); Halberstadt (1998) and those obtained in this work for different classification tasks for different datasets appear in Table 13.

5.4.1. Comparison with unvoiced stops classification accuracies of Hou et al. (2007)

Recent work specifically on the task of classification of unvoiced stops reported 91.9% accuracy for place classification on a TIMIT subset from frame-based features across the CV segment with a time-delay neural network (TDNN) classifier (Hou et al., 2007). The features used were the set of 13 MFCCs with first- and second-derivatives extracted from an interval of 90 ms before and 60 ms after the voicing onset. The TDNN was used due to its known ability to exploit temporal structure and its inherent discrimination ability. Evaluation of the joint spectro-temporal feature based classifiers of this work on the identical dataset, detailed in Table 12, resulted in the higher accuracies of 94.8% and 95.1% using 2D-DCT and BVP surface coefficients respectively and 96.2% when combined at classifier level as seen from Table 13 (a). The combination resulted in significantly improved performance compared to 2D-DCT features alone.

5.4.2. Comparison with performances reported in Halberstadt (1998) and Bouvrie et al. (2008)

We discuss the results in two stages: first with the 6-way stop classification and next with 3-way place classification of stops. Results of evaluation of features on 6-way stop classification using different data sets as published by Halberstadt (1998) and those obtained in this work are tabulated in Table 13 (b). We considered a relevant feature set ("S2") proposed by Halberstadt (1998) obtained by averaging MFCCs across frames over the five non-overlapping segments with addition of burst duration, reported to give the best results of all his individual systems in a 6-way stop classification task. With diagonal covariance GMM, maximum of 96 mixtures per class and 4-fold model aggregation, stops in the development set were correctly classified with 76.3% accuracy in phone classification task. Bouvrie et al. (2008) used the same feature set on 6-way restricted stop classification with a linear regularized least-squares (RLS) classifier to obtain an accuracy of 72.1% on the core test set. As presented in Table 6 also reported are the results (76.5%) obtained on the same task and data by their localized spectro-temporal cepstral features from the narrowband spectrogram (Bouvrie et al., 2008). Evaluation of the spectro-temporal features of the present work on the same data and 6-way classification task with full covariance GMM and 4-fold model aggregation in this work resulted in 78.0% and 77.2% accuracies with 2D-DCT and BVP features respectively. Also the combination resulted in significant performance compared to 2D-DCT only on the full test set with a p-value of 3.0060x10^{-4}. The combined performance is also superior to the results reported by Bouvrie et al. (2008).

Results of perceptual tests and evaluation of features on 3-way place classification using different data sets as published by Halberstadt (1998) for his various single-classifier systems are tabulated in Table 13 (c). Scoring of perceptual tests was done by averaging the 7 listeners' responses as well as by majority voting. The latter resulted in the best human performance of 97.8%. The automatic classification systems were based on either averaging MFCCs over different sub-segments of the phone with two MFCC derivatives obtained at the phone segment boundaries (Systems A, B) or by encoding the trajectory of low order MFCCs using tapered cosine basis vectors over a longer interval (System C). In all systems, burst duration was added to the feature set. System B was considered to be specially tailored to stops with shorter analysis window durations and temporal averaging restricted to the two halves of the phone segment. All the systems used full covariance context independent GMMs with phone priors.

Results of evaluation of the new spectro-temporal features on the same 3-way place classification task (using context-independent models) on the same development and perceptual test data of Halberstadt (1998) are also presented in Table 13 (c). Comparing the performances of the combination of 2D-DCT and BVP with 2D-DCT it can be seen that the combination performs significantly better than 2D-DCT only (p-value of 1.0307x10^{-6}). We note that the combined performance of 2D-DCT and BVP is comparable with that of systems A and C. The BVP features, in particular, show superior classification of the perceptual data (i.e. tokens that are misclassified by one or more of Halberstadt's systems) indicating that such joint spectro-temporal features can usefully complement any of the already available feature sets. It may be noted that the Systems A and C were evaluated with 20-fold and 10-fold model aggregation respectively while we used 4-fold model aggregation. Further the System C uses a longer analysis interval (300 ms) for feature extraction.

6. Conclusion

An investigation of features representing joint spectro-temporal correlation within phonetic segments has been attempted for unvoiced stop classification. The larger goal is to develop a hierarchical classifier where an initial classification into phone classes is followed by phone-class specific classifiers. The use of such a framework provides the opportunity to develop diverse measurements which focus on the phonetically relevant information for discriminating among sounds in a particular phone class. In this work, two approaches to spectro-temporal surface modeling, viz. the 2D-DCT and bivariate polynomial fitting are evaluated for place classification of unvoiced stops. The proposed implementation of 2D-DCT features improves upon previous implementations reported in the literature by the use of frequency-localised but temporally undivided patches in the spectro-temporal domain. The bivariate-polynomial surface modeling was expected to provide for a truly joint spectro-temporal representation. For given phonetic segment type, it is observed that 2D-DCT coefficients represent the discriminative acoustic cues better than polynomial coefficients for shorter analysis durations but polynomial coefficients do better on longer analysis durations when large formant transitions are expected. A classifier combination of BVP and 2D-DCT features performs better than either in isolation on a database of American English (TIMIT) and Marathi unvoiced stops. A comparison of performance with published results on the same task revealed that the spectro-temporal feature systems tested in this work improve upon the best

Table 13: Performances (%accuracy) of proposed features and features in the related work

(a)

3-way UV stop classification on Hou's data (Hou et al., 2007)

Feature set (Authors)	Test set
MFCC(Hou et al.,2007)	91.9
2D-DCT(this work)	94.8
BVP(this work)	95.1
Combination(this work; p-value b/w 2D-DCT and the combination)	96.2(5.2654x10^{-4})

(b)

6-way stop classification on Halberstadt's data (Halberstadt, 1998)

Feature set (Authors)	DV	CR	FT
MFCC based features, S2 (Bouvrie et al., 2008)		72.1	
Localized 2D-DCT (Bouvrie et al., 2008)		76.5	
2D-DCT(this work)	78.1	78.0	78.4
BVP(this work)	79.2	77.2	78.0
Combination(this work; p-value b/w 2D-DCT and the combination)	79.6	79.0	79.6(3.0060x10^{-4})

(c)

3-way place classification on Halberstadt's data (Halberstadt, 1998)

Feature set (Authors)	DV	PT	FT
Perceptual test, Avg. (Halberstadt, 1998)		93.7	
Perceptual test, Majority voting (Halberstadt, 1998)		97.8	
System A (Halberstadt, 1998)	92.9	75.3	
System B (Halberstadt, 1998)		77.1	
System C (Halberstadt, 1998)	93.5	77.8	
2D-DCT (this work)	92.3	82.9	91.5
BVP (this work)	93.2	84.5	92.2
Combination(this work; p-value b/w 2D-DCT and the combination)	93.4	83.9	92.9 (1.0307x10^{-6})

previous systems' performances in terms of classification accuracies on the specified datasets. The results indicate that BVP features may be usefully incorporated in hierarchical phone classifiers based on diverse class-specific features.

Acknowledgement

The authors are grateful to the anonymous reviewers for their valuable comments on a previous version of the submission. The research is partly supported by a project grant to IIT Bombay under the National Programme on Perception Engineering, sponsored by the Department of Information Technology, MCIT, Government of India.

References

Ahmed, R., Agarwal, S.S., 1969. Significant features in the perception of Hindi consonants. J. Acoust. Soc. Amer. 45, 758–763.

Bonneau, A., Djezzar, L., Laprie, Y., 1996. Perception of the place of articulation of French stop bursts. J. Acoust. Soc. Amer. 100, 555–564.

Bouvrie, J., Ezzat, T., Poggio, T., 2008. Localized spectro-temporal cepstral analysis of speech, in: Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., Las Vegas, NV. pp. 4733–4736.

Bunnell, H.T., Polikoff, J., McNicholas, J., 2004. Spectral moment vs. Bark cepstral analysis of children's word-initial voiceless stops, in: Proc. Interspeech'04, Jeju Island, Korea. pp. 1313–1316.

Chen, B., Zhu, Q., Morgan, N., 2004. Learning long-term temporal features in LVCSR using neural networks, in: Proc. Interspeech'04, Jeju Island, Korea. pp. 612–615.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Amer. 118, 887–906.

Datta, A.K., Ganguli, N.R., Ray, S., 1980. Recognition of unaspirated plosives - a statistical approach. IEEE Trans. Acoust., Speech, and Signal Process. 28, 85–91.

Depireux, D.A., Simon, J.Z., Klein, D.J., Shamma, S.A., 2001. Spectro-temporal response filed characterization with dynamic ripples in ferret primary auditory cortex. J.Neurophysiol. 85, 1220–1234.

Dyer, S.A., He, X., 2001. Least-square fitting of data by polynomials. IEEE Instrum. Meas. Mag 4, 46–51.

Ellis, D.P.W., 1997. Correlation of feature dimensions: some numbers.

Enzinger, E., 2010. Characterising formant tracks in Viennese diphthongs for forensic speaker comparison, in: Proc. Audio Engg. Society Int. Conf. - Audio Forensics, Hillerod, Denmark. pp. 47–52.

Ezzat, T., Bouvrie, J., Poggio, T., 2007. Spectro-temporal analysis of speech using 2-d Gabor filters, in: Proc. Interspeech'07, Antwerp, Belgium. pp. 506–509.

Fant, G., 1973. Stops in CV-syllables, Speech Sounds and Features. Mouton, Cambridge MA, MIT Press.

Feijoo, S., Fernandez, S., Balsa, R., 1999. Acoustic and perceptual study of phonetic integration in Spanish voiceless stops. Speech Commun. 27, 1–18.

Forrest, K., Weismer, G., Milenkovic, P., Dougall, R.N., 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. J. Acoust. Soc. Amer. 84, 115–123.

Gish, H., Ng, K., 1996. Parametric trajectory models for speech recognition, in: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP'96), Philadelphia, PA. pp. 466–469.

Glass, J.R., 2003. A probabilistic framework for segment-based speech recognition. Comp. Speech and Lang. 17, 137–152.

Goldenthal, W.D., 1994. Statistical trajectory models for phonetic recognition. Ph.D. thesis. Mass. Inst. of Tech.. Cambridge, MA.

Halberstadt, A.K., 1998. Heterogeneous acoustic measurements and multiple classifiers for speech recognition. Ph.D. thesis. Mass. Inst. of Tech.. Cambridge, MA.

Harder, D.W., Math, M., Khoury, R., 2010. Numerical Analysis for Engineering. University of Waterloo. Free online textbook for numerical analysis.

Hazen, T.J., Halberstadt, A.K., 1998. Using aggregation to improve the performance of mixture Gaussian acoustic models, in: Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., Seattle, WA. pp. 653–656.

Hermansky, H., Sharma, S., Jain, P., 1999. Data-derived nonlinear mapping for feature extraction in hmm, in: Proc. Automatic Speech Recognition and Understanding '99, Keystone, CO. pp. 63–66.

Hou, J., Rabiner, L., Dusan, S., 2007. On the use of time-delay neural networks for highly accurate classification of stop consonants, in: Proc. Interspeech'07, Antwerp, Belgium. pp. 1929–1932.

Kajarekar, S.S., Yegnanarayana, B., Hermansky, H., 2001. A study of two dimensional linear discriminants for ASR, in: Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., Salt Lake City, UT. pp. 137–140.

Kewley-Port, D., 1983. Time-varying features as correlates of place of articulation in stop consonants. J. Acoust. Soc. Amer. 73, 322–335.

Kleinschmidt, M., 2002. Methods for capturing spectro-temporal modulations in automatic speech recognition. Acta Acustica united with Acustica 88, 416–422.

Lamel, L.F., Kassel, R.H., Seneff, S., 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus, in: DARPA Speech Recognition Workshop, Palo, Alto. pp. 61–70.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Mesgarani, N., et al., 2009. Discriminant spectrotemporal features for phoneme recognition, in: Proc. Interspeech'09, Brighton, U.K.. pp. 2983–2986.

Meyer, B.T., Kollmeier, B., 2008. Optimization and evaluation of Gabor feature sets for ASR, in: Proc. Interspeech'08, Brisbane, Australia. pp. 906–909.

Morgan, N., et al., 2005. Pushing the envelope-aside: Beyond the spectral envelope as the fundamental representation for speech recognition. IEEE Signal Process, Mag. 22, 81–88.

Neagu, A., Bailly, G., 1998. Coopeartion and competition of burst and formant transitions for the perception and the identification of French stops, in: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP'98), Sydney, Australia. pp. 2127–2130.

Niyogi, P., Ramesh, P., 2003. The voicing features for stop consonants: recognition experiment with continuously spoken alphabets. Speech Commun. 41, 349–367.

Nossair, Z.B., Zahorian, S.A., 1991. Dynamic spectral shape features as acoustic correlates for initial stop consonants. J. Acoust. Soc. Amer. 89, 2978–2991.

Obleser, J., Leaver, A.M., VanMeter, J., Rauschecker, J.P., 2010. Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Frontiers in Psychology 1, paper 232.

Ohala, M., 1995. Acoustic study of VC transitions for Hindi stops, in: Int. Congress of Phonetic sciences, Stockholm, Sweden. pp. 22–25.

Ohala, M., Ohala, J.J., 1998. Correlation between consonantal VC transitions and degree of perceptual confusion of place contrast in Hindi, in: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP'98), Sydney, Australia. pp. 2795–2798.

Pandey, P.C., Shah, M.S., 2009. Estimation of place of articulation during stop closures of Vowel-Consonant-Vowel utterances. IEEE Trans. Audio, Speech, and Lang. Process. 17, 277–286.

Patil, V., Joshi, S., Rao, P., 2009. Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach, in: Proc. Interspeech'09, Brighton, U.K.. pp. 2543–2546.

Prasanna, S.R.M., Gangashetty, S.V., Yegnanarayana, B., 2001. Significance of vowel onset point for speech analysis, in: Proc. Signal Process. and Commn. Conf., Banagalore, India. pp. 81–88.

Rifkin, R., Bouvrie, J., Schuttle, K., Chikkerur, S., Kouh, M., Ezzat, T., Poggio, T., 2007. Phonetic classification using hierarchical, feed-forward, spectro-temporal patch-based architectures. MIT-CSAIL-TR-2007-007 CBCL-266. Mass. Inst. of Tech.. Cambridge, MA.

Sekhar, C.C., Yegnanarayana, B., 2002. A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances. IEEE Trans. Speech, and Audio Process. 10, 472–480.

Smits, R., Bosch, L.T., Collier, R., 1996. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. J. Acoust. Soc. Amer. 100, 3852–3864.

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Commun. 45, 401–423.

Wang, X., Zahorian, S.A., Auberg, S., 1996. Analysis of speech segments using variable spectral/temporal resolution, in: Proc. Int. Conf. on Spoken Lang. Process. (ICSLP'96), Philadelphia, PA. pp. 1221–1224.