DE GRUYTER DOI 10.1515/jisys-2013-0041 — Journal of Intelligent Systems 2013; 22(3): 215–228

Research Article

Veena Karjigi* and Preeti Rao

Knowledge-Based Features for Place Classification of Unvoiced Stops

Abstract: The classification of unvoiced stops in consonant–vowel (CV) syllables, segmented from continuous speech, is investigated by features related to speech production. As burst and vocalic transitions contribute to identification of stops in the CV context, features are computed from both regions. Although formants are the truly discriminating articulatory features, their estimation from the speech signal is a challenge especially in unvoiced regions like the release burst of stops. This may be compensated partially by sub-band energy-based features. In this work, formant features from the vocalic region are combined with features from the burst region comprising sub-band energies, as well as features from a formant tracking method developed for unvoiced regions. The overall combination of features at the classifier level obtains an accuracy of 84.4%, which is significantly better than that obtained with solely sub-band features on unvoiced stops in CV syllables of TIMIT.

Keywords: Knowledge-based features, place of articulation.

***Corresponding author: Veena Karjigi,** Siddaganga Institute of Technology, Department of Electronics and Communication, Tumkur 572103, India, e-mail: veena.karjigi@gmail.com
Preeti Rao: Department of Electrical Engineering, Indian Institute of Technology, Bombay, India

1 Introduction

Phone classification has been attempted in the literature either by a using a uniform set of features across phones, such as mel frequency cepstral coefficients (MFCCs), e.g., ref. [4], or a hierarchical structure with heterogeneous features specific to the broad phone class [1, 7]. Heterogeneous features provide the advantage of embedding speech-specific information in the form of knowledge-based (KB) features derived from the knowledge of speech production. The speech signal can be treated as the output of a sequence of articulatory gestures. The time duration that articulators stay in a particular gesture varies; however, the sequence of gestures by and large remains the same and KB features capture these variations. These features have been shown to be robust compared with the

traditional cepstral/spectral coefficients with speaker age and gender variations in different recognition/classification tasks [6, 10]. However, their performance on train-test matched datasets has generally been poorer than that of spectral/spectro-temporal features, e.g., refs. [3, 7]. Motivated by their potential for robustness, in this work we focus on investigating and evaluating KB features.

Of the broad phone classes, stops are particularly characterized by their dynamics. Stops are classified by the presence or absence of voicing and place of articulation (PoA). Voiced and unvoiced stops in English are {b,d,g} and {p,t,k}, respectively. Three places of articulation for stops in English are labial (/b/ and /p/), alveolar (/d/ and /t/), and velar (/g/ and /k/). 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. These are among the most difficult sounds to handle in automatic speech recognition. We focus on place classification of unvoiced stops on this work, which are short-duration dynamic sounds and thus more challenging for automatic classification. As we investigate KB features, it is important to know the acoustic characteristics of stops under consideration.

In general, stops occur with vowels, and hence production of stops is explained with reference to vowel–consonant–vowel (VCV) syllables. Figure 1 shows the waveform, spectrogram, formants (shown with dots), and important time instants in the production of the VCV syllable /ɑkɑ/. The first phase in the production of a stop is closure, which starts when the first vowel ends. Articulators will be ready for production of the stops even before the closure starts because of co-articulation, and hence formants deviate from their steady states in VC transition as shown in Figure 1. The oral cavity is closed during the closure with the constriction at a particular place in the vocal tract. Because of the complete closing of the oral cavity, formants disappear in the closure duration. Closure is followed by release in the form of noise burst. During the release, the part of the vocal tract downstream from the constriction (front cavity) is excited by the sudden reduction in the intra-oral pressure. Owing to this, formants shift rapidly either upward or downward in frequency as well as amplitude, depending on the place of constriction of the consonant and the following vowel, which is shown as consonant–vowel (CV) transition in the figure. Burst release, voicing onset point, and start of vowel steady state constitute the important temporal landmarks or boundaries for the CV context, while the end of vowel steady state and voicing offset time correspond to the VC context as shown in Figure 1.

Unvoiced stops used in this work are extracted from CV syllables of the standard American English database TIMIT. The production mechanisms of these unvoiced stops along with their relationship to the spectral shape of the release burst and formant transitions from and toward vowels are detailed next.

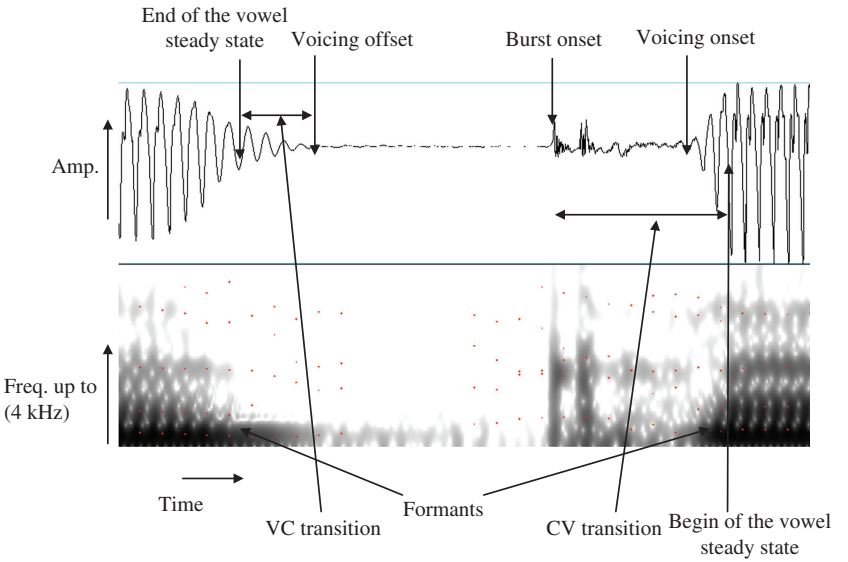

Figure 1. Waveform, Spectrogram, Formants (Shown with Dots), and Important Points in Transitions to and from the Stop in the VCV Syllable /ɑkɑ/.

The labial consonant is produced when both lips constrict and hence there is no front cavity. This results in the absence of strong resonances. Hence, labials do not exhibit well-defined spectra. When a labial stop occurs along with the vowel, the first three formants (F1, F2, and F3) of the vowel transit downward while moving toward the production of labial stop. F2 and F3 transitions are prominent especially with front vowels. The alveolar consonant is produced by making the constriction between the tongue blade and the alveolar ridge, which results in resonances with high frequency. F1 moves down while transiting from a vowel to an alveolar place. Because of the forward movement of the tongue, F2 moves up when the alveolar stop is produced following a back vowel, and it is the opposite for front vowels. F3 rises when a vowel is followed by an alveolar. The velar consonant is produced when the back of the tongue touches the soft palate and thereby excites the long front cavity. The long front cavity results in reduced acoustic losses, and hence velars show compact peaks in the spectrum and compactness reduces as the place of constriction moves toward the lips. F1 comes down and F2 and F3 move toward each other while transiting from a vowel toward a velar; F2 and F3 move toward each other (F2 increases and F3 decreases), known as the velar pinch.

The understanding of the production mechanism of unvoiced stops and their relationship to the spectral shape of the release burst and formant transitions

reveal that the burst release and vocalic transition comprise the phonetic subsegments that carry cues important to the perception of PoA of stops, which often occur with vowels. While voice onset time (VOT) and formant trajectories are directly linked to the PoA, the accurate estimation of formants is difficult especially in the unvoiced region. An alternative way of finding similar information is to find the sub-band energy concentrations in the spectrum.

We consider recently proposed features available for English [16] and evaluate them on the TIMIT database. To the sub-band features of Suchato [16], we add new features related to fixed sub-bands based on our previous work on Marathi stops [10]. Also, we revisit our formant-related features derived from the vocalic region [10], which were not satisfactory because of errors in formant tracking, and improve the formant estimation. Further, we extend the same to the burst region by postprocessing the formant tracks. We evaluate the new features (both sub-band energies and formant features) separately and in classifier combination mode for three-way place classification of stops in CV syllables extracted from TIMIT.

2 Feature Extraction

The accurate localization of segment boundaries is a prerequisite for segment-based feature extraction and classification. Burst release and voicing onset constitute the important temporal landmarks or boundaries for the CV context. These landmarks are detected by a two-stage approach proposed by Patil et al. [14]. When tested for detection of these two landmarks in the CV syllables of TIMIT test set, 64%, 87.2%, and 92.4% of burst onsets and 57.7%, 81.7%, and 89.6% of voicing onsets were located within ±5, ±10, and ±15 ms of actual labels, respectively.

2.1 Sub-band Features in Burst Region

The recent work of Suchato [16] used 12 features derived from the average power spectra to classify stops in English on the basis of PoA. Features were derived from specific frequency bands selected on the basis of the acoustic characteristics of stops in American English. Three frequency ranges include the second and third formant frequency range (1.25–2.5 kHz and 1.5–3 kHz), and the high-frequency range (3.5–8 kHz). The features included (i) attributes related to the spectral shape of the release burst in terms of relative amplitude and energy in the high-frequency range compared with the second and third formant (F2–F3) range; (ii) the largest

amplitude of the burst spectrum in the high-frequency and F2–F3 range with reference to the first formant prominence in the vowel spectrum; (iii) ratios of the biggest peak of the burst spectrum to the voicing spectrum in all the three frequency bands mentioned above; (iv) the overall energy distribution of burst spectrum was obtained by computing the centers of gravity in the frequency scale of the power spectrum obtained from the region of burst duration, 10 and 20 ms from burst onset, to account for dynamic characteristic stops; (v) two temporal cues, VOT and closure duration. To this set of features, we add new features based on our previous work where similar features were defined but with different frequency bands suitable for PoA distinction of Marathi stops [10], the details of which are described next.

1. Sub-band centroids and their log amplitudes (six features): Many features of Suchato [16] involve finding peaks in the specific frequency bands that may not exist always or may not be prominent. Thus, in Suchato's work, it was possible that not all features were derived for all the tokens. We propose that, in the absence of peaks in the particular frequency band, the amplitude of the center of gravity is used as a compensation. In addition to these, we compute the sub-band centroids and their log amplitudes in the three frequency bands for all the tokens.
2. Sub-band spectral tilts (two features): Spectral tilts in the F2–F3 and high-frequency bands are measured in terms of spectral slopes.
3. Burst level (one feature): In addition to these, overall burst energy is known to be an important factor in identifying the PoA [9]. Burst level describes the strength of the burst, which is defined as the rms energy of the segment obtained by centering a Hamming window of 20 ms at the burst onset.
4. Temporal changes in the spectral shape (two features): Also, dynamic features from burst to voicing onset in the F2–F3 band were computed. The temporal trajectory of spectral tilts sampled at 1-ms intervals was computed in the interval between the burst and voicing onset in the F2–F3 band. The value of the largest tilt and the slope of a line fitted to the trajectory of tilts of the same band were derived to represent the dynamic characteristics on the lines of Lahiri et al. [11].

Also, for word initial stops, closure may not be there. Therefore, closure duration also has been omitted in this work. In all, 11 new features were added to Suchato's 11 features and there were 22 features.

2.2 Formant-Related Features

Deviations of the first three formants in CV/VC transitions are the commonly used formant features used for PoA distinction [2, 16]. While Ali et al. [2] used formants

estimated using an auditory front end, Suchato [16] used manually estimated formants. Cosine coefficients obtained by encoding these formant frequency and amplitude trajectories were used for classification of stops in the initial position of CVC syllables [13]. These trajectories are obtained by using the McCandless formant tracking algorithm with roots of the LP polynomial as the possible candidates [12]. The first six formant values obtained (using a linear prediction analysis) over a specified interval starting from the burst onset along with MFCCs were used for classification of stops in TIMIT [18].

We used the 11 features of Suchato [16], (i) absolute values of the first three formants at voicing onset and those of F2 and F3 at burst onset; (ii) variation of F2 and F3 over a 20-ms transition interval starting from burst and voicing onset; (iii) difference between F3 and F2 at burst and voicing onset, and obtain the formant trajectories by using the classic McCandless method [12], as used in our previous work [10], but extend it to extract formants in the burst region also by post-processing the trajectories obtained by the McCandless algorithm.

2.2.1 Original McCandless Algorithm

The classic method of formant extraction [12] estimates the first three formants in vowel-like sounds, which considers peaks of the LP spectrum as possible candidates for formants and uses peak enhancement and knowledge about speech to maintain continuity in the formants, and finally smoothes the formant tracks. Voiced segments are detected by using energy and pitch, and anchor points are marked in each segment at instants where there is highest energy in that segment. Estimation starts at these anchor points and moves in both left and right directions till the end of that segment. Because the formant estimation was only in the voiced regions, formants were extracted at every 5 ms assuming that in such regions articulators do not move much in 5 ms.

2.2.2 McCandless Algorithm as Used in This Work

In this work, the focus is on estimating formants in the CV transitions where formant estimation had to be done at a much higher rate because of the highly varying characteristics of the signal in such regions. Hence, formants were estimated at every 1 ms from the steady state of the vowel where formants are most likely to be correct back to the burst onset. The first anchor points are found as steady-state points in the vowel. Energy in the lower frequency band [0:1000] Hz is computed in an interval of 100 ms beyond the voicing onset, assuming that

the longest vowel in continuous speech can go up to 100 ms. The time instant at which the above-mentioned energy falls below 3 dB from the maximum energy of the vowel toward the end of the vowel is defined as the steady state.

McCandless [12] uses the 14th-order LP spectrum for speech signals sampled at 10 kHz. Speech signals used in this work are sampled at 16 kHz. The 18th-order LP spectrum is computed here because in many cases the 14th-order LP spectrum was unable to resolve the close-by peaks even at the steady state. Also, as the algorithm takes care of the spurious peaks better than the missing peaks, it is assumed that the 18th-order LP spectrum is more suitable. Knowledge about the frequency band of each formant along with peaks of the spectrum was used to find the initial estimates [15].

2.2.3 Postprocessing

McCandless applies continuity constraints for error-free formant estimation only in the voiced region. To ensure correct estimation of formants in the unvoiced region, postprocessing of formant tracks was suggested where formant estimation is done only in the voiced regions and the longest continuous formant segment for each voiced segment was found by applying a fixed threshold to the interframe frequency difference [17]. To estimate formants around this anchor segment, beam search was performed within the available formant candidates obtained by peak picking in the LP spectrum. At any point that this search fails because of the non-availability of valid candidates (available candidates within the specified frequency range), formant estimation was stopped at that point. Therefore, this method also does not ensure estimation of formants in the transitions where formants are not prominent.

We extend the above beam search algorithm to formant tracking in the burst region by changing the frequency range of beam search for each frame based on previous formant estimates and filling in locally with spectral centroid if formant candidates are missing altogether. The steps involved in postprocessing are as follows:

1. Discontinuities in the formant tracks were identified by applying thresholds to interframe frequency difference, which is small for the first formant compared with the second and third formants as deviations of later formants are relatively large in transitions. The longest continuous segment thus obtained is assumed to be estimated correctly.
2. To re-estimate the formants in the remaining frames, beam search was conducted on either side of the longest continuous segment with all possible formant candidates within a chosen frequency range for each formant track.

3. The formant values of four previous or following frames (one that is appropriate based on frames to be beam searched are toward the left or right of the longest continuous segment) are used to set the frequency range for beam searching. A fixed deviation around the average formant value of these four frames was computed and was used for local beam searching, as the formant value of the current frame is expected to lie around this average formant value.
4. Candidates within the frequency range mentioned above are listed, and the one that is near the average formant value mentioned above is selected as the formant for the current frame.
5. If no candidates are found in the specified range, the spectral centroid in frequency is assigned as the formant value for that frame.

Figures 2 and 3 show the first three formants for the syllables "kah" and "pao" respectively. These CV syllables are taken from a TIMIT utterance as obtained by the VTR formant database [5]. Improvements with postprocessing can be observed in both the figures.

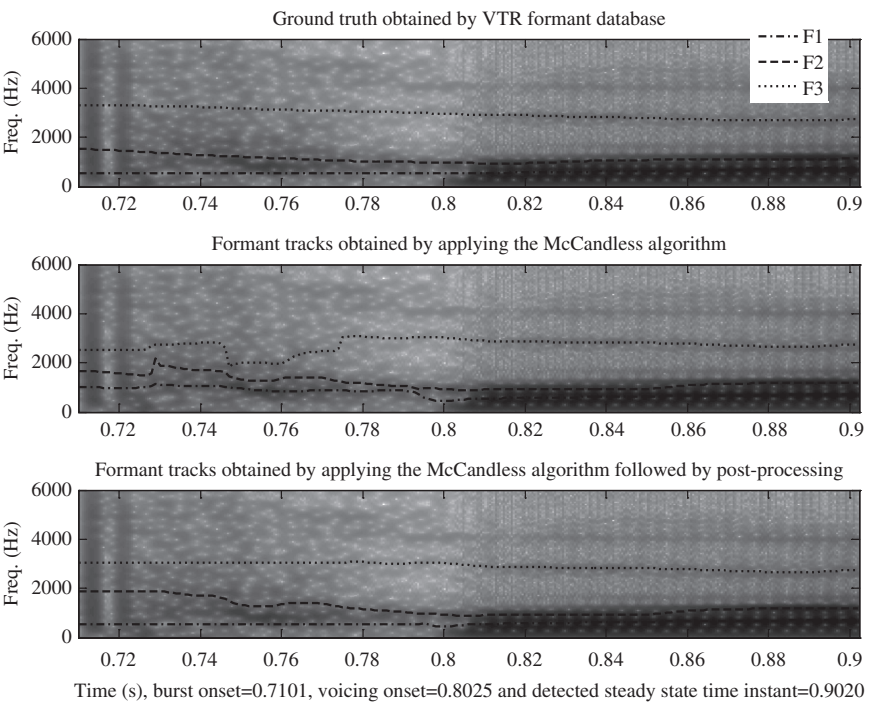

Figure 2. Ground Truth, Formant Tracks Obtained by Applying the McCandless Algorithm Before and After Post-Processing for the CV Syllable "kah".

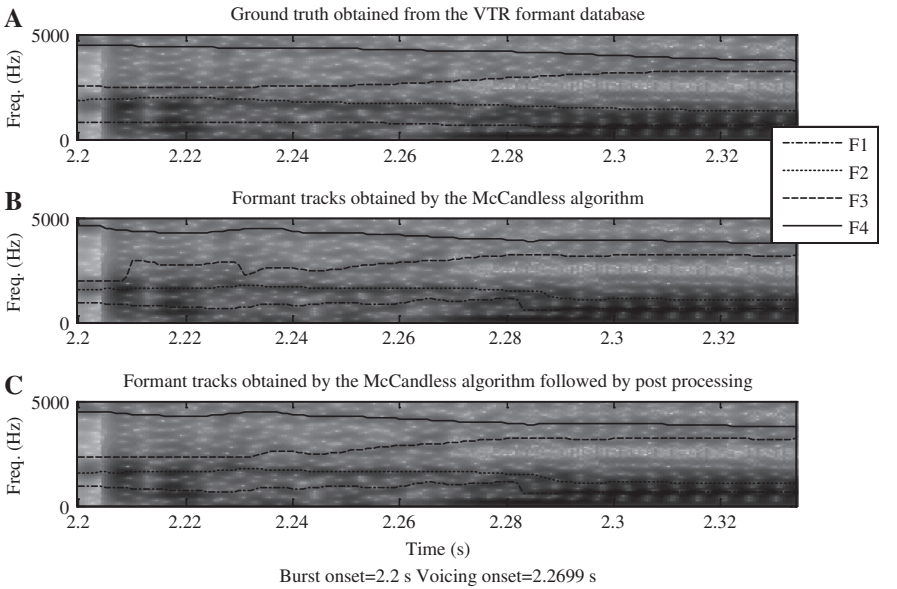

Figure 3. Ground Truth, Formant Tracks Obtained by Applying the McCandless Algorithm Before and After Postprocessing for the CV Syllable "pao".

3 Experiments and Results

The performances of the feature sets discussed in Section 2 are evaluated by classification experiments on unvoiced stops of the TIMIT database. The feature vectors are input to a Gaussian mixture model (GMM) classifier previously trained on labeled training data provided with the database. The training was done with 4-fold model aggregation as in Hazen and Halberstadt [8].

3.1 Database Details

We evaluated the sub-band and formant features on the CV syllable data of TIMIT. The database is divided as a train and test set with utterances from 462 and 168 speakers, respectively. Furthermore, as in other phone classification tasks, e.g., refs. [3, 7], a development set is formed with 50 test speakers. The train set was used for training phone models, the development set for tuning classifier parameters, and the test set for final performance evaluation. Also, the VTR formant database [5] is used for evaluating the formant estimates. The VTR formant database is a subset of the TIMIT database whose train set consists of two sentences, one each in SI and SX, spoken by each of the 162 speakers, thus forming 324

sentences from TIMIT train and test set consists of eight sentences (all SI and SX) uttered by 24 speakers in the core test set to form 192 sentences. We consider the CV syllables with unvoiced stops in these 516 sentences for evaluation. The details of distribution of unvoiced CV syllables in the TIMIT and VTR formant databases are given in Table 1.

3.2 Classifier Details

A GMM classifier with full covariance matrices was trained by the expectation maximization algorithm. Both feature sets were evaluated for a fixed set of reduced dimensions (22, 20, 10, and 5 for the original feature vector of dimension 22; 11, 10, and 5 for the original feature vector of dimension 11) using principal component analysis. The number of mixtures was varied from 1 to 6. Reduced dimension feature vectors were evaluated using the development set to finalize the combination of feature vector dimensionality and number of GMM mixtures based on the best obtained classification accuracy. In each case, only the combination of dimension and mixture that resulted in the highest classification accuracy is reported.

3.3 Evaluation of Formant Estimation

Although formant transitions are known to be useful for PoA distinction, formant estimation itself is a challenge. Thus, as mentioned in Section 3.1, first we evaluated our formant estimation method using the VTR formant database [5]. The database consists of the first four formant estimates labeled at every 10 ms for all sentences. Initial trajectory estimates were obtained on the basis of an automatic format tracking algorithm, which is followed by extensive manual correction based on the known phone characteristics. We linearly interpolate these trajectories to get the formant values at every 1 ms to evaluate formant estimation. The

Table 1. Details of CV Syllables of Unvoiced Stops in the TIMIT and VTR Formant Databases.

	p	t	k	Total
CV syllables with unvoiced stops in TIMIT				
Train	1533	2973	2487	6993
Development	183	291	246	720
Test	424	683	534	1641
CV syllables with unvoiced stops in VTR formant database				
Train	126	246	162	534
Test	83	120	88	291

difference between the ground truth and automatically extracted formant values is computed for a number of frames over the burst duration (as formant estimation is difficult in this region), and the average error per frame was computed. The average error per frame for the first three formants across stops in the database for CV syllables are given in Table 2. Postprocessing resulted in improvements in formant estimation, as seen from Table 2.

Furthermore, to see the effect of formant estimation errors on the classification, we classify the unvoiced stops in these CV syllables by formant features discussed in Section 3.2. For this, GMMs were obtained by features derived from ground truth formant estimates of the train set and evaluated on the features obtained by automatically estimated formants of the test set and the obtained classification accuracies are presented in Table 3. Classification accuracies also reflect the improvements in postprocessing.

3.4 Evaluation of Features on CV Syllables

Both sub-band as well as formant features discussed in Section 3 are evaluated using the CV syllable database described in Table 1, and the percentages of classification accuracy are presented in Table 4. The annotation provided with the TIMIT database was used for feature extraction on train data, whereas automatically detected landmarks mentioned in Section 3 were used for extracting features from the test data. Adding the new sub-band features to Suchato's features [16] and postprocessing of formant tracks improved the classification accuracy in the case of sub-band and formant features, respectively, as seen from Table 4. Also, a combination of the augmented sub-band features and formant features (derived after postprocessing of formant tracks) at the classifier level by product rule resulted in further improvement. McNemar's significance test conducted for performance difference between the best individual (sub-band features) and the classifier combination resulted in a p-value of 0.0012.

Table 2. Average Frame-Level Formant Estimation Error (Hz) for CV Syllables of the VTR Formant Database.

	F1 error	F2 error	F3 error
Before postprocessing	299.2	262.2	342.5
After postprocessing	235.5	201.1	261.1
% Reduction in error	21.3	23.3	23.7

Table 3. Classification Results (%) Obtained for Stops in CV Syllables Extracted from the Test Set of the VTR Formant Database.

	% Accuracy obtained on test set
Ground truth	61.9
Before postprocessing	44.3
After postprocessing	51.9

4 Discussion

As seen from Table 4, unvoiced stops in CV syllables of TIMIT were classified with an accuracy of 84.4% with automatically detected landmarks. To see the effect of errors in landmark detection on the classification accuracy, features were obtained from manually labeled landmarks for the same test set, and this resulted in 86.8% accuracy. Therefore, improvement in the landmark detection stage is expected to improve the overall results.

It is difficult to compare the results obtained in this work with the literature, as most of the authors do not test their features on standard datasets. Suchato [16] obtained an accuracy of 94.9% for place classification of unvoiced stops in CV syllables with his features (both sub-band as well as formant) derived from data of four speakers, manually labeled landmarks, and manually transcribed formants. We, however, used a dataset of 110 speakers and automatically detected landmarks and formants. Ali et al. [2] used the auditory front end to derive the burst spectrum and formant-related features from the stops detected in the continuous sentences of 60 speakers from seven dialect sentences of TIMIT and obtained 90% accuracy on three-way place classification of six stops by using hard decision thresholds for each feature. Even Ali et al. [2] did not report the details about the 60 speakers used in their study; however, as they report their results on TIMIT, roughly we can say that our results are comparable to that of Ali et al. Next, although Zheng et al. [18] gave all the details of the train and test set used in their study, they do not

Table 4. Classification Accuracy (%) Obtained by Different Feature Sets Derived in this Work.

No.	Feature set	Authors	Acc.
1	Sub-band	Suchato (11)	75.0
2	Sub-band	This work (22)	82.1
3	Formant (before postprocessing)	This work (11)	64.2
4	Formant (after postprocessing)	This work (11)	66.0
5	Combination of (2) and (4) by product rule	This work	84.4

report the results obtained with formant features only. However, using MFCCs augmented with formant features in a support vector machine classifier, they report an 89.1% accuracy. With this, we claim that the features used in this work are comparable with the existing literature, and we have set up a platform for other researchers to compare their results on similar tasks.

5 Conclusion

Knowledge-based acoustic features have been explored for place classification of unvoiced stops of the TIMIT database. Augmentation of previously available subband features in the burst region by the new features proposed in this work leads to improvement in classification accuracy. Also, the difficult task of tracking the formants in transition has been attempted, and postprocessing of formant tracks was found to improve the formant estimation especially in the burst region. In the future, addition of well-established spectral features to the KB features used in this study is expected to improve the classification accuracy.

Received May 17, 2013; previously published online June 26, 2013.

Bibliography

[1] A. M. A. Ali, J. V. Spiegel, P. Mueller, G. Haentjens and J. Berman, An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech, in: *Proc. ICASSP*, pp. 118–121, Phoenix, AZ, 1999.

[2] A. M. A. Ali, J. V. Spiegel and P. Mueller, Acoustic–phonetic features for the automatic classification of stop consonants, *IEEE Trans. Speech Audio* **9** (2001), 833–841.

[3] J. Bouvrie, T. Ezzat and T. Poggio, Localized spectro-temporal cepstral analysis of speech, in: *Proc. ICASSP*, pp. 4733–4736, Las Vegas, NV, 2008.

[4] P. Clarkson and P. J. Moreno, On the use of support vector machines for phonetic classification, in: *Proc. ICASSP*, pp. 585–588, Phoenix, AZ, 1999.

[5] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen and A. Alwan, A database of vocal tract resonance trajectories for research in speech processing, in: *Proc. ICASSP*, pp. 60–62, Toulouse, France, 2006.

[6] O. Deshmukh, C. Espy-Wilson and A. Juneja, Acoustic-phonetic speech parameters for speaker-independent speech recognition, in: *Proc. ICASSP*, pp. 593–596, Orlando, FL, 2002.

[7] A. K. Halberstadt, Heterogeneous acoustic measurements and multiple classifiers for speech recognition, Ph.D. Thesis, Massachusetts Institute of Technology, 1998.

[8] T. J. Hazen and A. K. Halberstadt, Using aggregation to improve the performance of mixture Gaussian acoustic models, in: *Proc. ICASSP*, pp. 12–15, Seattle, WA, 1998.

[9] A. Jongman, S. E. Blumstein and A. Lahiri, Acoustic characteristics for dental and alveolar stop consonants: a cross-language study, *J. Phonetics* **13** (1985), 235–251.
[10] V. Karjigi and P. Rao, Landmark based recognition of stops: acoustic attributes versus smoothed spectra, in: *Proc. Interspeech 2008*, pp. 1550–1553, Brisbane, Australia, 2008.
[11] A. Lahiri, L. Gewirth and S. E. Blumstein, A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence form a cross-language study, *J. Acoust. Soc. Am.* **76** (1984), 391–404.
[12] S. S. McCandless, An algorithm for automatic formant extraction using linear prediction spectra, *IEEE Trans. Acoust. Speech Signal Process.* **22** (1974), 135–141.
[13] Z. B. Nossair and S. A. Zahorian, Dynamic spectral shape features as acoustic correlates for initial stop consonants, *J. Acoust. Soc. Am.* **89** (1991), 2978–2991.
[14] V. Patil, S. Joshi and P. Rao, Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach, in: *Proc. Interspeech*, pp. 2543–2546, Brighton, UK, 2009.
[15] S. Seneff, Modifications to formant tracking algorithm of April 1974, *IEEE Trans. Acoust. Speech Signal Process.* **24** (1976), 192–193.
[16] A. Suchato, Classification of stop place articulation, Ph.D. Thesis, Massachusetts Institute of Technology, 2004.
[17] K. Xia and C. Espy-Wilson, A new strategy of formant extraction based on dynamic programming, in: *Proc. Interspeech*, pp. 55–58, Beijing, China, 2000.
[18] Y. Zheng, M. Hasegawa-Johnson and S. Borys, Stop consonant classification by dynamic formant trajectory, in: *Proc. Interspeech*, pp. 2481–2484, Jeju, Korea, 2004.