Landmark based recognition of stops: acoustic attributes versus smoothed spectra

Veena Karjigi, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India

{veena, prao}@ee.iitb.ac.in

Abstract

Landmark based recognition of unvoiced word-initial stops is investigated. The relative effectiveness of acoustic-phonetic attributes versus more global spectral shape features is experimentally evaluated for four-way place classification of unvoiced, unaspirated stops. Various feature sets derived from the burst and vocalic transition regions of word initial consonants are compared via GMM based classification under speaker, gender, and vowel-context variability. While a set of acoustic attributes derived from the burst shows the best invariance to vowel context, it is found that global spectral shape features provide the most robust representation of the vocalic transition region by overcoming the problem of errors in explicit formant tracking. A combination of features from the burst and vocalic regions was superior to burst-only cues, but still far from the near perfect identification achieved in human perception.

Index Terms: Landmark based recognition, unvoiced stops, acoustic attributes, burst, vocalic transition

1. Introduction

Landmark based recognition of speech has recently re-emerged as an active research area. The approach is anticipated to provide increased robustness to the various sources of variability that challenge present day HMM based systems using standard frame-based spectral representations [1,2]. In a landmark based system, important acoustic events related to speech production (e.g. stop release bursts, vowel onsets) are detected, and a fine analysis of signal segments in the vicinity provides the acoustic cues for recognition. The signal regions around landmarks are considered to be information rich with respect to the linguistic features of the underlying phones and hence expected to better preserve invariances across factors such as context and regional accent, and speaker gender and age. An advantage of landmark based signal representations is the possibility to define acoustic features tailored to the phone class defined by the landmark. The increased flexibility afforded by heterogenous feature sets however gives rise to several research issues. For instance, in the case of stop consonants, the release burst and the adjoining formant transition regions are known to provide the main cues to place of articulation. However the choice of the actual acoustic representation to be used in a statistical recognition framework still remains. Is the information on place of articulation best captured by raw spectral / cepstral representations, or, by acoustic attributes derived by processing the spectra and tailored to the underlying articulatory distinctions? From an information theoretic point of view, information is lost as more processing is done. This assumes, of course, that the signal statistics are completely known or that infinite training data is available. Processing the data to convert the information into a potentially more useful form such as knowledge based acoustic attributes can help to make more effective use of finite training data [3]. The objective of the present work is to consider this question for the class of unvoiced initial stops by automatic classification experiments using different acoustic representations and ways of combining them.

Unvoiced stops in word initial position (prevocalic) are formed by the sequence of release burst, a frication segment and possible aspiration noise followed by a vocalic transition segment. Stop consonants are considered difficult to classify due to their dynamic nature and sensitivity to coarticulatory effects. The frontness of the following vowel influences the precise place of constriction of the stop. Both the perception and the automatic classification of prevocalic stop consonants have been subjects of active research. Perception studies on unvoiced initial stops across languages (English, French, Spanish, Dutch) have found that nearly perfect identification of stops can be achieved only when all the main cues (burst duration, burst spectrum and onset of vocalic formants) are present [4,5,6,7]. Research on automatic classification has focused on discovering the form of the cues that perform best in the place identification task. The acoustic representations investigated have ranged from static spectral cues observed at burst onset [8] to dynamic information (time evolution of static properties) observed throughout the burst and vocalic transition regions [9]. Further, a variety of acoustic attributes have been investigated based on applying knowledge of specific articulatory-phonetic distinctions [1,10]. In contrast to this, have been smoothed spectrum representations that capture only the global spectral shape. While similar to the MFCC based features used in state-of-the-art recognisers, in a landmark based system the spectra are aligned with the detected speech events. A prominent example of landmark based smoothed spectral representations are the cepstral coefficients of Nossair and Zahorian [4]. Dynamic spectral shape features obtained from the smoothed temporal trajectories of the cepstral coefficients beginning with the burst onset were found very effective in the encoding of place information in unvoiced initial stops in CVC isolated utterances [4].

In the present work, we consider the class of unvoiced, unaspirated stops of Marathi, a prominent language of the Indo-Aryan family. Like several related Indian languages, it distinguishes *four* places of articulation for stops in contrast to the three used in English. The four places of articulation are labial [p], dental [t̪], retroflex [ʈ] and velar [k] each of which can occur as unvoiced-unaspirated, unvoiced-aspirated, voiced-unaspirated and voiced-aspirated. Thus the unvoiced, unaspirated stops of Marathi differ from the prevocalic unvoiced stops of English in aspiration. Also, they differ from the corresponding stops of French, Dutch and Spanish by the presence of an additional coronal place of articulation, increasing possibly the complexity of the place identification task. We experimentally evaluate various choices for the

Accepted after peer review of full paper

acoustic representation, and ways to combine cues from burst and vocalic transition regions in a statistical recognition framework. A special focus of this work is the evaluation of the robustness of the features to variations that are related to speaker vocal tract (of which gender difference is one extreme example [1]) and vowel context, which could arise possibly from differences in accent or language.

2. Database description

Utterances containing each of the four stops {p, t̪, ʈ, k} in word-initial position followed by one of the eight vowels of the language were recorded from each of 18 native speakers (10 male and 8 female). The 8 vowels are categorized as follows: {front: I,i,e}, {back: U,u,o}, {center: A,a}. Visually observed formant locations were found consistent with the vowel frontness. Two meaningful mono-/bi-syllabic words were chosen for each stop-vowel combination to obtain 64 distinct words, which were each embedded in two different carrier phrases (one statement and one question). The resulting data set thus comprises of a total of 64 x 18 x 2 = 2304 tokens (or 576 per stop consonant), recorded at a sampling rate of 16 kHz in quiet conditions.

The spoken words were next isolated and normalized so that maximum amplitude of each word utterance is the same to compensate for varying loudness levels of different speakers. The time locations of the release burst and the voicing onset of the initial consonant were manually labeled. The burst onset was marked as the time instant after the closure silence at which a rapid change in the waveform amplitude sets in. The first negative to positive going zero crossing in the first cycle of the periodic waveform was labeled as the voicing onset. Table 1 gives a statistical summary of measured VOT (voicing onset time, or burst duration) across the Marathi words database. We note the relatively low VOTs of the unaspirated stops, with retroflex being the lowest. The latter is explained by the relatively fast movement of the active articulator involved (the tongue tip) which offsets the effect of the more posterior PoA. Labials and dentals have comparable VOTs while velars are the longest.

Place of articulation	VOT (ms)	
	Mean	Std. dev.
Labial	17.1	7.8
Dental	15.0	5.9
Retroflex	9.7	3.8
Velar	27.8	10.5

Table 1. *Statistics of the VOT for the four stops.*

3. Implementation

3.1. Smoothed spectrum coefficients

The global shape of the spectrum can be encoded by cepstral coefficients. We use the implementation of [4] to compute the cepstral coefficients by a cosine transform of a nonlinearly scaled and warped magnitude spectrum. The resulting "discrete cosine transform coefficients" (DCTCs) represent global spectral shape. Temporal trajectories of the first 6 DCTCs sampled at fixed intervals are encoded using 3 cosine basis vector coefficients and that of the 7th DCTC encoded with 2 coefficients to obtain a 20-dim feature vector of "dynamic spectral coefficients" (DSC). The coefficients encode the smoothened trajectory of each DCTC with the first coefficient corresponding to the average value across time. The VOT of our unvoiced stops are much lower than those of the unvoiced aspirated stops of English [4]. As a consequence, it was found more effective to increase the sampling of the temporal trajectory to every 1 ms, with analysis window length = 6.4 ms.

A further design decision involves the duration over which the DCTCs trajectory is observed. Nossair and Zahorian experimentally found a fixed duration of 60 ms from burst onset to be most effective [4]. In view of the fact that the Marathi stops have lower VOTs, we re-examine the choice of the fixed duration. We also investigate the possibility of using variable durations based on the detected voicing onset location. A fixed duration from the voicing onset onwards is appended to the burst region to obtain a continuous segment over which DCTCs trajectories are computed. While the method of [4] is based on the burst onset landmark only, the latter approach requires both landmarks. It also gives us the flexibility to compute the dynamic spectral coefficients separately at each landmark and later combine the feature vectors [5].

3.2. Acoustic attributes

Acoustic attributes are the acoustic correlates of the articulatory distinctions between the phones to be classified, i.e. place of constriction in the present problem. Given that the knowledge of speech production is as yet incomplete, experimental investigations must be relied upon to come up with precise features within the broadly accepted acoustic cues to place identification. Features are extracted from each of the burst and vocalic transition regions. To limit the number of features in each set, feature selection by a greedy algorithm based on Shannon mutual information is applied whenever required [11].

3.2.1. Burst

Important acoustic cues provided by the burst region are the burst duration and level, spectral shape and its dynamics [7]. A compact set of spectral shape cues only, derived from an average power spectrum over the burst region, were shown to provide 78% accuracy in the classification of the 4 Marathi stops [12]. We now investigate the benefits of expanding the set of spectral shape attributes and further incorporating cues related to burst duration, level and dynamics. The burst region of the unaspirated stops is characterized by frication noise without well defined formants. A more global spectral shape description is easily obtained in terms of spectral prominences in fixed broad frequency bands. The bands selected (in Hz) are [0:750], [750:2500], [2500:5000] and [5000:7000]. Across the fixed sub-bands, the frequencies and amplitudes of the respective centers of gravity were found more effective features than were sub-band spectral tilts. To the set of sub-band features, overall burst peak frequency and amplitude are added. The burst peak is described by the ratio of the amplitude of the biggest peak in the frequency band [500:7000] to the average spectral amplitude in the band [0:500], and its frequency by the center of gravity in [0: 7000]. The burst duration (VOT) and r.m.s. level over a fixed window located at the burst onset provided important additional static cues to place.

Preliminary experiments revealed that the static cues, while effective overall, were marked by a prominent dental-

retroflex confusion in the back vowel context. Spectral tilt cues were reconsidered. It was observed that spectral tilt in the band [2000:6000] which is expected to be low for dentals, was not found so at the burst onset but at a later instant in the burst. The set of static burst cues was therefore enhanced with two dynamic features (on the lines of [13]) derived from the temporal trajectory of tilts sampled at 1 ms intervals from burst onset to voicing onset. These were: the value of the largest tilt in [2000:6000] band, and the slope of a line fitted to the trajectory of tilts of the same band.

3.2.2. Vocalic transition

Formant transitions are known to provide important perceptual cues to place of articulation. We use McCandless formant tracking algorithm to estimate the first three formants from the detected steady state of the following vowel back to voicing onset [14]. The formant frequency values and their log amplitudes at the voicing onset and steady state were obtained. Further features were derived from these measurements. Shifts in first and second formant frequencies from one end to the other end of the transition and log amplitudes of the three formants at voicing onset were found to carry significant information about the place of constriction as was determined by feature selection. The encoding of formant trajectories using cosine basis vector coefficients as in [4] was also considered, but was not found to perform as well as the formant parameters sampled at the two ends.

4. Experiments

Classification experiments were designed to test the various feature sets and their combinations for the place detection of the 4 consonants. Variability due to speaker, gender and vowel context were introduced as "cross-speaker", "cross-gender" and "cross-context" tests respectively. Cross-context testing is expected to reveal the extent of invariance of the specific feature set with respect to large changes in the frontness of the following vowel.

For each experiment, a GMM classifier with diagonal covariance matrices was trained by the EM algorithm for the 4-way place classification. The GMM was configured with 3 mixtures per class based on preliminary experiments showing the most consistent performance across subsets of the data for this choice. The different classification tasks are described below.

(a) Cross-speaker (CS): tested for speaker independence by a leave-one-speaker-out cross-validation. The training set was comprised of all tokens from 17 speakers and testing set was comprised of all tokens from the remaining one speaker.

(b) Cross-gender (CG): tested for more extreme speaker variability. The training set was comprised of tokens from male only (or female only) speakers, and testing set, from the remaining female (or male) speakers.

(c) Cross-context (CC): tested for variability in vowel frontness by excluding the test vowel frontness (i.e. one of the front, back, center categories) from the training set. The leave-one-frontness-out cross-validation thus involved 3 complete rounds each round covering train and test tokens from all the 18 speakers.

In each of the above experiments, the consonant classification accuracy is estimated for each of the feature sets under consideration. Table 2 shows the results in terms of percent correct over the 2304 test tokens from the cross-validation testing. The feature sets are organized by the signal region they correspond to and whether they are smoothed spectral coefficients (DSC) or acoustic attributes (AA). Specific combinations of the feature sets are also considered.

No.	Feature set (dim)	CS	CG	CC
Burst region				
1	AA (12)	86.2	84.8	70.4
2	DSC (20)	87.0	83.2	59.6
Vocalic region				
3	AA (8)	49.8	40.8	35.7
4	DSC: 20 ms from voicing onset (20)	61.0	51.8	35.9
Full region- fixed duration with one landmark				
5	DSC: 20 ms from burst onset (20)	87.8	84.9	54.7
Full region- variable duration with two landmarks				
6	DSC: burst+20 ms of vocalic (20)	86.9	77.3	44.5
Burst + vocalic (20 ms from voicing onset) - direct fusion of the 2 feature sets				
7	AA+AA (12+8=20)	87.6	84.8	72.7
8	AA+ DSC (12+8=20)	88.2	86.6	73.3
9	DSC+DSC (20+8=28)	88.6	85.5	60.1

Table 2. *Classification accuracies (%) for various feature sets and tasks.*

5. Results

From Table 2, we can draw the following conclusions regarding the effectiveness of features for automatic place identification. The conclusions were drawn after verifying that the differences noted are statistically significant by applying McNemar's test for two algorithms tested on the same data [15] ($p<0.01$ unless mentioned otherwise).

5.1. Region wise comparison

Burst region: We see that the accuracies of the DSC and AA are comparable in the CS and CG tasks but the AA show a significant superiority in the CC task indicating that the articulatory-acoustic features are far more context-invariant.

Vocalic region: The DSC (20) are significantly superior in the CS and CG tasks. This is not surprising given the known difficulties of formant estimation in the highly dynamic vocalic transition regions. In fact, a vowel-context wise analysis of performance revealed that the AA-vocalic features performed well in center context only, where the first three formants are relatively well separated.

Full region: When fixed duration starting with burst onset was used to obtain DSC, the best performance was found at 20 ms duration with performance degrading with further increases. This duration is likely to include the vocalic transition in the case of non-velar stops. DSC over variable duration based on actual VOT (item 6) showed no advantage over DSC from fixed duration segments.

5.2. Combining feature sets

We consider the combination of feature sets that have been separately obtained from each region (burst, vocalic). From the results of Sec. 5.1, we note that the possibilities worth investigating are: 1) AA from both regions, 2) AA from burst

combined with DSC from the vocalic region, and 3) DSC from both regions. In order to keep the total feature vector dimensionalities more or less comparable, the DSC vector from the vocalic region is pruned to 8 elements (from the original 20 dim) by the feature selection method mentioned in Sec. 3.2.

We observe that DSC over a duration that includes the full burst and a fixed duration of vocalic transition (item 6) is clearly outperformed by the combination of DSC feature sets of the individual regions (item 9) under all three variabilities (although in the CS condition, the p value = 0.02). We find that adding AA-vocalic information to AA-burst changes only CC performance over AA-burst significantly. Similarly, combining the DSC-vocalic feature vector with the AA-burst feature (item 8) leads to a significant improvement in vowel context invariance while retaining the CS and CG performance of the DSC features' combination (item 9). However, adding DSC-vocalic information (without pruning) to burst information of either form, gave poorer results in all the three tasks when compared with adding pruned DSC-vocalic features.

While Table 2 reports overall accuracy over consonant and vowel contexts, some specific observations were made. Adding DSC-vocalic information either to AA-burst or to DSC-burst showed the most marked improvements in the case of back vowels. Dentals in front vowel context were confused with other consonants with vocalic features alone, and this persisted in the feature combinations. Similarly, the retroflex consonant confusion in back vowel context that arises in the AA-burst features persists even when vocalic transition information is added.

6. Conclusions

The problem of landmark based place recognition of unvoiced, unaspirated word initial stops of Marathi was investigated. The database comprised of 64 Marathi words recorded from 18 native speakers. Various feature sets derived from the (manually labeled) burst and vocalic transition regions were compared for the 4-way place identification in a GMM classifier under speaker, gender, and vowel-context variations. A summary of results in point form follows.

1) A maximum classification accuracy of 89% was obtained by combining feature vectors separately extracted from the burst and vocalic transition regions of the consonant. While the acoustic-phonetic feature set for burst representation proposed in this paper shows the best invariance to vowel context, it is found that the global spectral shape features provided the most robust performance in the vocalic transition region by overcoming the problem of errors in explicit formant tracking.

2) The independent encoding of DSC in burst and vocalic regions retains more information about place compared to DSC that are derived from a single continuous region starting at the burst onset. Since the dynamics of the two regions are different, fitting their separate temporal trajectories is likely to be more accurate even if it increases the overall dimensionality. This indicates the utility of detecting both burst and voicing onset landmarks.

3) Both AA and DSC features show relatively high performance in the cross-gender task, unlike the results reported on MFCC in HMM-based systems [1]. A possible interpretation of this is that performance gains are obtained by landmark-aligned analysis irrespective of the precise form of the features extracted. Of course, it may be noted that the present study was restricted to stops, and their acoustic properties are probably less affected by changes in vocal tract length than are those of vowels.

We conclude that landmark based recognition is promising for automatic speech recognition when training data is constrained in some way, or when large feature space dimensionality (needed in order to capture context dependencies) is impractical. Methods for landmark detection in continuous speech are already under active research [2].

A preliminary perception test (not reported here) of 5 listeners' responses to a set of Marathi CV utterances by 2 speakers, and to their isolated burst-only and burst-less segments, indicated that automatic classification was slightly better than the average listener score on burst-only stimuli but significantly worse on burst-less and full stimuli. Future work could address improvements in the representation of vocalic region cues to place identification. Research on novel ways to combine features and/or classifiers may help to further narrow the gap between human and machine identification.

7. References

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

[2] Hasegawa-Johnson M. et al., "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop", *Proc. ICASSP*, vol. 1, pp. 213-216, Mar. 2005.

[3] Meng H.M., "The use of distinctive features for automatic speech recognition", M.S. Thesis, MIT, May 1991.

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

[5] Neagu A. and Bailly G., "Cooperation and competition of burst and formant transitions for the perception and the identification of French stops", *Proc. ICSLP*, pp. 2127-2130, Dec. 1998.

[6] Feijoo S., Fernandez S. and Balsa R., "Acoustic and perceptual study of phonetic integration in Spanish voiceless stops", *Speech Communication,* vol. 27, 1-18, Feb. 1999.

[7] Smits R., Ten Bosch L. and Collier R., "Evaluation of various sets of acoustic cues for the perception of prevocalic stops consonants. II. Modeling and evaluation", *J. Acoust. Soc. Am.,* vol. 100, no.6, pp. 3865-3881, Dec. 1996.

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

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

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

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

[12] Karjigi V. and Rao P., "Four-way classification of place of articulation of Marathi unvoiced stops from burst spectra", *Proc. Workshop on Image and Signal Processing*, IIT, Guwahati, pp. 42-46, Dec. 2007.

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

[14] McCandless S.S., "An algorithm for automatic formant extraction using linear prediction spectra", *IEEE Trans. ASSP.,* vol. 22*,* no.2*,* pp. 135-141, Apr. 1974.

[15] Gillick L. and Cox S.J., "Some statistical issues in the comparison of speech recognition algorithms", *Proc. ICASSP*, vol. 1, pp. 532-535, May 1989.