Classification of place-of-articulation of stop consonants using temporal analysis

A. P. Prathosh[1]*, A. G. Ramakrishnan*[2] *and T. V. Ananthapadmanabha*[3]

[1]Xerox research center India, Banglore, India.
[2]Department of Electrical engineering, Indian Institute of science, Banglore, India.
[3]Voice and speech systems, Banglore, India.
Prathosh.AP@xerox.com, ramkiag@ee.iisc.ernet.in, tva.blr@gmail.com

Abstract

This paper proposes acoustic-phonetic features for classification of place-of-articulation of stop consonants derived from their temporal structures. The speech signal corresponding to a stop is characterized by several temporal features such as sub-band zero-crossings and envelope fits. Classification experiments on the stops from the TIMIT (read speech) and the Buckeye (conversational speech) databases using a support vector machine classifier demonstrate that the performance of the proposed features (84.6 %) is comparable to that obtained by MFCCs (85.1 %) in many aspects. Further, the classification accuracy is boosted (90.1 %) with the combination of temporal and MFCC features, which substantiates their supplementary nature.

Index Terms: stop consonants, place of articulation, temporal features, sub-band crossings, acoustic phonetics

1. Introduction

1.1. Background

Studies on stop consonants have attracted the focus of several researchers due to their challenging nature. Stop sounds are produced by building up pressure behind a complete closure of the vocal tract followed by a rapid release of air-flow through the constriction resulting in a sudden rise in the energy which is termed as the burst [1]. Depending upon the place at which the constriction occurs, stops in English are divided into three categories namely, bilabials (/p/ and /b/ - closure is formed by the lips), alveolars (/t/ and /d/ - closure is formed by tongue blade and alveolar ridge) and velars (/k/ and /g/- closure formed by tongue body and soft palate). In the case of aspirated stops, the burst is followed by an interval during which the glottis is spread letting the air flow, resulting in a noise-like signal referred to as the aspiration noise. If the stop is followed by a voiced phone, the vocal folds start to vibrate a short interval after the burst release, which is called the voicing onset time.

1.2. The problem of classification of stops

Automatic classification of stop consonants based on their the place of articulation (PoA), or equivalently, the automatic identification of PoA of stops, from the acoustic signal is a classical problem in speech analysis. It finds application in many areas such as automatic speech recognition (ASR), speech pathology and phonetic studies. ASR systems can be broadly classified into two categories - statistical modeling based systems and distinctive-feature based systems [2]. It has been shown that the detection of PoA of stops plays an important role in both the kinds of systems: while the accuracy of a statistical ASR can be improved by incorporating the PoA information[3], detection of PoA is an integral part of a distinctive-feature based speech recognizer [4]. Automatic classification of stops can aid computer-based speech therapies and also phonetic and perceptual studies.

1.3. Previous work

The problem of classification of stops has a long history in speech science. Divergent views on acoustic invariance arose amongst speech scientists because of studies on stops. Broadly, the acoustic cues proposed for the classification of stops fall into two categories: (i) features based on the spectral characteristics of stop-burst and (ii) features based on formant transitions from onset to mid-point of the following vowel. Some studies argue that context-independent acoustic cues exist [5] for stops while some contend this view [6, 7]. Studies such as those by Delattre et. al. [8] and Alwan [9] emphasize on the transitions of the second formant as a cue for automatic identification of PoA. Winitz et. al. [10] chose burst-based cues instead of the formant transitions for the classification of unvoiced stops. Studies by Blumstein and Stevens [11, 5] suggested that the gross shape of the burst spectrum considered over the first few milliseconds from the burst serves as a sufficient cue for stop classification. They argue that velars have a 'compact' spectral shape, whereas bilabials and alveolars possess a 'diffuse-falling' and 'diffuse-rising' spectral shapes, respectively. Other than the above described cues, some studies make use of auditory-front end based features along with the spectral cues such as spectral center of gravity [7] and time varying spectral cues [12]. The role of a few temporal cues such as VOT and closure duration are also examined in stop classification [13]. Numerous other works exist on stop classification as may be found in the study by Suchato [4].

1.4. Objectives of this work

It is believed that cues derived from both the burst and the formant transitions contribute to the identification of PoA of stops. However, there is evidence to show that the features derived from the signal spanning the burst interval are sufficient [14]. Burst features are preferred to formant transitions since these facilitate classification in all the contexts i.e., even when stops do not succeed or precede vowels. Further, despite the wealth of literature on stop classification, there have not been many attempts to extract temporal features of the signal around the stop bursts despite references to the usefulness of temporal features [15]. However, on examination of several stop segments, one can visualize distinct differences between the temporal struc-

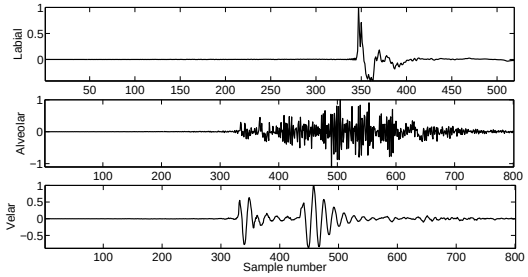

Figure 1: *Differences between the temporal structures of three classes of stops. The bilabial stop /p/ (top trace) resembles an ideal impulse; the alveolar stop /t/ (middle trace) is dense in terms of zero-crossings and the velar stop /k/ (bottom trace) is lesser dense in terms of zero-crossings.*

tures of stops with different PoA. Further, the relative energy of the source component can be expected to be different for different stops due to the differences in the type of release [1]. Nonetheless, source features have seldom been utilized for classification of stops. Motivated by the above facts, in this paper, we propose features derived from the characteristics of the temporal structures and the excitation source component around the burst, for the classification of stops based on their PoA. The features are used in a support vector machine (SVM) classifier to classify the stops from two large speech corpora viz., the TIMIT database [16] and the Buckeye corpus [17] comprising read and conversational speech, respectively. We compare our results with those obtained using the spectral feature Mel Frequency Cepstral Coefficients (MFCC) (in the rest of the paper, we interchangeably use the terms MFCCs and spectral features since MFCCs are derived from smoothing the mel-filtered magnitude spectrum of the speech signal) and also examine the supplementary information in the temporal and spectral features.

2. Proposed method

2.1. Distinct temporal structures of stops - An illustration

Figure1 depicts a typical waveform each for each class of stop, taken from the TIMIT database. On examination of the acoustic waveforms corresponding to several such stops having different PoA, the following empirical observations may be made on their temporal structures.

Alveolar stops are very 'dense' in that the number of zero-crossings per unit time is higher than those of other stops. Velar stops are sparser than alveolars in terms of zero-crossings with a spread burst. Also, it is observed that the pattern of the concentration of energy around the burst-onset contains discriminative information. For a labial, most of the energy is concentrated around the burst which makes it tend to be like an impulse whereas the energy is spread throughout the burst-interval for alveolars. Thus intuitively, we believe that the zero-crossing patterns and the pattern of the concentration of energy around the burst-onset can classify the PoA of stops.

2.2. Sub-band zero-crossings for signal discrimination

The zero-crossing rate (ZCR) of a zero-mean stationary random process is known to correspond to its weighted spectral centroid. However, the complete spectral profile of a given signal is not obtained from the ZCR alone. For instance, the ZCRs for a sinusoid and a square wave of the same fundamental frequency are the same while they have different frequency distributions. Hence, the ZCR alone cannot yield the required discriminability between stops with different PoA. However, as stated in the article by Kedem [18], the sequence of higher-order crossings can uniquely determine the normalized spectral distribution function of a Gaussian process. Here, the term higher-order crossings refers to the ZCR in the linear-filtered versions of a given time series.

As an illustration, consider two signals, $s_1[n]$ and $s_2[n]$ obtained as the superposition of two sinusoids of different frequencies $f_1 = 100\ Hz$ and $f_2 = 2000\ Hz$, i.e., let $s_1[n] = A_1 sin(2\pi f_1 n) + A_2 sin(2\pi f_2 n)$ with $A_1 = 10A_2$ and $s_2[n] = B_1 sin(2\pi f_1 n) + B_2 sin(2\pi f_2 n)$ with $B_2 = 10B_1$, $n = 1, 2, .., 2000$. Now, due to the dominance of f_1 in s_1, ZCR for s_1 corresponds directly to f_1 while it has no information regarding f_2. Similarly ZCR for s_2 can estimate f_2 but not f_1. However, the number of extremum points in s_1 or equivalently the ZCR in its first-differenced (high-pass filtered) version will give the estimate of the higher frequency component (f_2). Similarly, the lower of the frequencies (f_1) can be estimated from s_2, using the ZCR of its integrated (low-pass filtered) version. Thus, it may be ascertained that information regarding the frequency profile of a given signal can be obtained by ZCRs in different frequency bands. In other words, the ordered set of ZCRs in different sub-bands can discriminate signals with different frequency profiles (in this case $s_1[n]$ and $s_2[n]$).

Motivated by the above facts, for the current problem, we consider the set of ZCRs in several sub-bands of the speech signal as one of the feature sets. Specifically, ZCRs in the speech signal filtered using a Mel-filter bank is used in this work which we term the sub-band zero-crossing rate (SZCR). Mel-filter banks are chosen to account for the auditory processing involved in the human perceptual system. According to the *dominant-frequency principle [18],* the ZCR of a given signal admit values in the neighborhood of the frequency which is significantly dominant in the spectral distribution of the signal. Therefore the SZCR in each sub-band corresponds to the spectral centroid of the speech signal within that sub-band. Since the center frequencies of the filters in the Mel-bank progressively increase, the SZCR coefficients will be ordered. Further, it has been shown that for a discrete-time signal, the higher order crossings approach a degenerate state as the number of coefficients increases [18]. Thus we hypothesize that these temporal features (we prefer to call SZCR as temporal features since they are computed in the time domain) provide useful information about the PoA of stops with lesser length of feature vector than conventional MFCCs.

2.3. Burst structure and source features

From Fig. 1, we see differences in the distribution of the energy around the burst, which can possibly distinguish one kind of stop from another. In this section, we define features for quantifying the distribution of the energy around the burst of a stop consonant.

1. Kurtosis and skewness measures: As discussed earlier, the bursts of labial stops are peakier in nature than those of the other stops. The peakedness of a distribution can be quantified by the fourth standardized moment or the kurtosis measure. Further, the asymmetry of the signal around the burst can be quantified using the third standardized moment or the coefficient of skewness. Thus we include the kurtosis and skewness measures of the normalized Hilbert envelope (HE) of the burst in the fea-

ture set. Normalized HE of the burst is used because it ensures that it mimics a probability mass function in that it has all positive values and sums to unity. Fig. 2 illustrates the use of kurtosis and skewness measures in discriminating the burst envelopes of different stops.

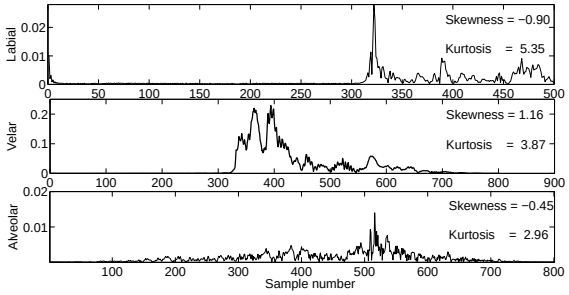

Figure 2: Illustration of the use of kurtosis and skewness measures in discriminating the burst envelopes of different stops. The top, middle and bottom traces, respectively, depict the normalized HE of a labial (/b/), velar (/k/) and an alveolar (/t/) stop. It can be seen that the kurtosis for the labial stop is higher than that for the alveloar stop indicating that the labial stop is more 'peaky' in nature. Also the labial burst has higher absolute skewness than alveolar stop, indicating that the labial burst is more asymmetric in nature compared to the alveloar burst. The skewness of the velar stop is positive indicating that the envelope is more tilted to the right.

2. Source feature: To quantify the differences in the source component, the ratio of the $l_2 - norms$ of the integrated linear prediction residual (ILPR, an estimate of the source signal [19, 20, 21]) and the speech signal corresponding to the burst interval is used as another feature.

2.4. Implementation details of feature extraction

The burst and voicing onsets for stops in a given utterance are automatically detected using the algorithms reported in our earlier works [22, 23]. Since the objective of this study is to analyze the temporal structure for stop classification, we chose the speech signal of 50 milliseconds duration starting from 20 milliseconds prior to the closure-burst transition for analysis. This 50 millisecond interval generally corresponds to the burst-interval for most unvoiced stops without including the aspiration interval, if any. However for some stops, especially voiced ones, this interval may include the following vowel too, in which case, only the interval up to the vowel-onset is considered for analysis. Further, a hanning window is applied to smooth out the edges to facilitate the envelope-based analysis. The signal used for analysis is normalized with respect to its $l_2 - norm$ to make sure that all the tokens have the same energy (unit norm vectors). Although burst-energy is known to be a parameter of significance, it is deliberately not considered in this study since our motive is to examine the usefulness of the temporal structure alone.

The filter-banks used for the computation of SZCR are implemented in the frequency domain, spaced according to the Mel-scale spanning the entire frequency range. To avoid the influence of low-energy noisy components on the computation of SZCR, instead of actual zero crossings in each band, the crossings at the level of 10 % of the maximum value of the unfiltered signal on both positive and negative sides are considered. In summary, given a signal corresponding to a stop, the SZCRs, envelope features and the source feature are computed and concatenated to form the final temporal feature vector (of dimension equal to number of filters used for SZCR + two for kurtosis and skewness + one source measure).

2.5. SVM-RBF for classification

We use a support vector machine (SVM) for classification of the PoA of stops. The radial basis function kernel is used which is implemented using the LibSVM package [24]. All the features are z-scored before training and testing to ensure proper normalization.

3. Experiments and results

3.1. Baseline system

To compare the performance of the proposed features with the spectral features (SF), we build a baseline system with Mel-frequency cepstral coefficients (MFCC) along with the delta and delta-delta coefficients as the feature vector with the SVM classifier. These features are used widely in the state-of-the-art ASR systems. MFCCs quantify the average spectral energy of a signal in different auditory frequency bands thereby characterizing the spectral shape of the signal, which is a distinguishing factor among the stops. MFCCs are computed for the same frame location and duration for which the temporal features are computed.

3.2. Databases and experiments

For all our experiments, we consider two large corpora namely, (i) the TIMIT database [16] containing 6300 utterances spoken by 630 speakers of different dialects of North America and (ii) the Buckeye corpus[17] comprising several hours of spontaneous American English speech of 40 speakers from central Ohio, USA. Both are labeled at the phone level which provides the ground truth for validation. All the stops in the TIMIT database and a large subset from the Buckeye corpus are considered for evaluations, irrespective of their position of occurrence, leaving those occurring in the stop-stop clusters, since it is known that burst may be absent in some such cases [25, 26]. The task is a three-class classification problem by placing bilabials (/p/ and /b/), velars (/k/ and /g/) and alveolars (/t/ and /d/) in a class each. The accuracies reported here are obtained by performing a grid search on the parameters of the SVM kernel.

In our first experiment, we conduct three-fold cross-validation tests on stops (around 25, 000 in number) from both the databases for the cases of voiced, unvoiced and combined cases separately. For all these experiments, accuracies are reported for temporal features alone (TF), spectral features alone (SF) and both the temporal and spectral features concatenated with each other (CF). In this experiment, the number of sub-bands used for the computation of SZCR and MFCCs are fixed at 12 and 13, respectively. However, in our second experiment, we vary the number of sub-bands used for SZCR computation and number of MFCCs, and report the consequent variation in the accuracy on the TIMIT test set. This examines the discrimination capabilities of the SZCR vis-a-vis MFCCs. In our final experiment, we compare the learning abilities of the features by reporting the classification accuracies on a test-set by varying the number of training samples. For this experiment, the training samples are taken from the training set and the entire test set

is used for testing. The second experiment is carried out on the TIMIT test database and the third using TIMIT training and test databases for training and testing, respectively. For the third experiment, a given number of training samples are randomly selected for training every time.

3.3. Results and discussion

Table 1 reports the cross-validation accuracies separately for stops from the TIMIT and Buckeye databases obtained using the proposed temporal features (TF), spectral features (SF) and the combined features (CF). The first and the second entries in each cell correspond to the TIMIT database and the Buckeye corpus, respectively. The following observations may be made from Table 1: (a) In general, the accuracy (for all the features) are better for unvoiced stops than their voiced counterparts. This is due to the fact that the bursts are more pronounced and of longer duration in the case of unvoiced stops and hence the features are better manifested. (b) The accuracies offered by the TF alone are almost equal to those offered by SF alone for all the cases. This suggests that TF possess as much information about the PoA as the SF. (c) When the TF and SF are combined, the accuracy increases by about 4-5 % in all the cases, confirming the presence of complementary information between the temporal and spectral features. (d) The accuracy for the stops in TIMIT (90.1 %) is better than that for stops in Buckeye (73.3 %) by 14-17 %. This is because the TIMIT database contains read-speech, where the bursts are known to manifest better than in free-style conversations which constitute the Buckeye corpus. Further, TIMIT has been carefully hand-labeled whereas most of the labels in Buckeye have been obtained by force alignment. It is interesting to note that the unanimous agreement between six transcribers on PoA of stops in Buckeye is 74 % as well [17]. It is also noted that the accuracies on the TIMIT set using only SZCR, SZCR+source features, SZCR+kurtosis and skewness are respectively, 79.6 %, 82.6 % and 81.8 %. From these observations, it can be inferred that the SZCR contributes the most for the classification accuracy compared to other features.

Table 1: Cross-validation accuracies, of the proposed temporal features (TF, 12 bands), spectral features (SF, 13 MFCCs) and the combined features (CF). The first and second entries in each cell of the table correspond to the result on the TIMIT and Buckeye corpus, respectively.

Feature type	TF (%)	SF (%)	CF (%)
All stops	84.6, 68.6	85.1, 67.2	90.1, 73.3
Unvoiced stops	86.8, 69.9	87.2, 68.8	91.5, 74.1
Voiced stops	81.2, 67.2	81.6, 65.6	87.2, 70.5

Table 2: Confusion matrix for the identification of PoA of stops from the TIMIT (first entry in each cell) and Buckeye (second entry in each cell) databases, using CFs.

	Alveolars	Bilabials	Velars
Alveolars	92.3, 64.2	2.7, 11.2	5.0, 24.6
Bilabials	5.2, 27.3	90.0, 56.6	4.8, 16.1
Velars	6.8, 11.2	4.0, 9.4	89.1, 79.4

Table 2 is the confusion matrix for the different classes of stops from both the databases. Bilabials are classified with least accuracy in the case of Buckeye corpus, probably because /b/ tend to have weak burst in conversational speech.

Figure 3 illustrates the results of the second and third experiments. It is seen that the accuracies with TF are always better than those with SF for all the sizes of the feature vector. SZCRs computed using only six-bands offer an accuracy of around 83 % which saturates after nine sub-bands. This corroborates that higher-order crossings degenerate around 9-10 bands for speech signal as stated in Kedem's study [18]. Also the TF needs lesser training samples (per-class) than SF to offer a given accuracy as shown by Fig. 3.

Our results compare well with those reported in the literature. Halberstadt's perception studies [27] report 6.3% as the average error made by human subjects in a PoA identification task which might be considered to be a rough benchmark. Our study offers 90.1 % accuracy which is about 3 % less than that. Many previous works, including those by Ali [7], Nathan and Silverman [12] and Suchato [4] report accuracies in the range of 82-91 % which compare well with our study. However, our study considers 25, 000 stops for analysis while most of the previous studies analyze a smaller number of stops, ranging from a few hundred to one or two thousand. Given that our study examines all the stops irrespective of their position, does not take into account the formant transition features and considers only temporal features, the results seem significant.

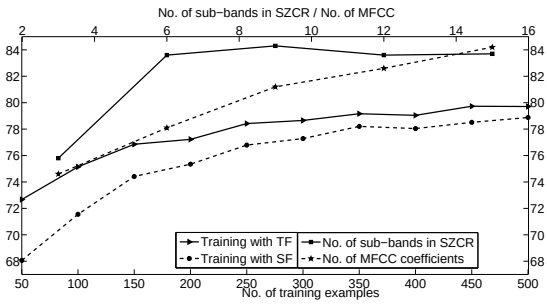

Figure 3: Illustration of classification accuracies for stops on the TIMIT test database for two different experiments. (i) Accuracy as a function of the number of training samples (lower abscissa) for temporal (TF, 12 sub-bands) and spectral features (SF, 13 MFCCs). (ii) Accuracy Vs. feature dimension (upper abcissa) for TF and SF (500 training examples each). Note that both the plots share the same ordinate or y-axis.

4. Conclusion

In this paper, we proposed temporal features for identification of place-of-articulation of stop consonants. Motivated by the differences in the temporal structures and the excitation source signal of the stops around the burst-onset, we employed subband zero-crossings, kurtosis and skewness measures and relative source energy as features. Several classification experiments on the TIMIT database of read speech and the Buckeye corpus of conversational speech confirmed that temporal features are as effective as the spectral features, whereas combinedly they can boost the classification accuracy. Further, it was shown that temporal features perform well with lower number of features and training samples than the spectral features. Future research may aim towards (i) further improving the accuracies on conversational speech, (ii) combining the contextual information such as formant transitions with the proposed features to improve the performance and (iii) studying the inter-corpus variability of the proposed features.

5. References

[1] K. N. Stevens, *Acoustic phonetics*. Cambridge, MA:MIT Press, 1998, ch. 1-8, pp. 1–485.

[2] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, pp. 3417–3430, 1996.

[3] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, pp. 1253–1264, 2011.

[4] A. Suchato, "Classification of stop consonant place of articulation," Ph.D. dissertation, Mass. Inst. of Technology, Cambridge, MA, 2004, phD dissertation.

[5] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am*, vol. 64, no. 5, pp. 1358–1368, 1978.

[6] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Am*, vol. 27, no. 1, pp. 322–335, 1983.

[7] A. M. A. Ali, V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic recognition of stop consonants," *IEEE Trans. Speech and Audio Process.*, vol. 9, pp. 833– 841, 2001.

[8] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am*, vol. 73, no. 1, pp. 769–773, 1955.

[9] A. A. Alwan, "Modeling speech perception in noise: The stop consonants as a case study," Ph.D. dissertation, Mass. Inst. of Technology, Cambridge, MA, 1995.

[10] H. Winitz, M. Scheib, and J. Reeds, "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech," *J. Acoust. Soc. Am*, vol. 54, no. 4, pp. 1309–1317, 1972.

[11] S. E. Blumstein and K. N. Stevens, "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, 1979.

[12] K. Nathan and H. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Trans. Speech and Audio Process.*, vol. 2, pp. 395–405, 1994.

[13] V. W. Zue, "Acoustic characteristics of stop consonants: A controlled study," Ph.D. dissertation, Mass. Inst. of Technology, Cambridge, MA, 1979.

[14] A. Bonneau, L. Djezzar, and Y. Laprie, "Perception of the place of articulation of French stop bursts," *J. Acoust. Soc. Am*, vol. 1, pp. 555–564, 1996.

[15] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English speech: a dynamic approach*. Springer Science & Business Media, 1993.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgrena, *DARPA- TIMIT, Acoustic-phonetic continuous speech corpus.*, US Department of Commerce, Washington, DC, 1993, (NISTIR Publication No.4930).

[17] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, pp. 89–95, 2005.

[18] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, 1986.

[19] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no-12, pp. 2471–2480, Dec. 2013.

[20] T. V. Ananthapadmanabha, "Acoustic factors determining perceived voice quality," in *Vocal fold Physiology - Voice quality control*, O.Fujimura and M. Hirano, Eds. San Diego,Cal.: Singular publishing group, 1995, ch. 7, pp. 113–126.

[21] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *J. Acoust. Soc. Am. EL*, vol. 137, accepted for publication, 2015.

[22] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 460–471, 2014.

[23] A. P. Prathosh, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Estimation of voice-onset time in continuous speech using temporal measures," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. EL122–EL128, 2014.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[25] J. B. Henderson and B. H. Repp, "Is a stop consonant released when followed by another stop consonant?" *Phonetica*, vol. 39, pp. 71–82, 1982.

[26] P. K. Ghosh and S. S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," *J. Acoust. Soc. Am.*, vol. 126, pp. EL1–EL7, 2009.

[27] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Mass. Inst. of Technology, Cambridge, MA, 1998.