Discrimination of Sonorants from Fricatives Using a Scalar Feature Derived from Linear Prediction Coefficients

T. V. Ananthapadmanabha
Voice and speech systems,
Bangalore, India - 560012
tva.blr@gmail.com

A. G. Ramakrishnan
Electrical Engineering,
Indian Institute of Science,
Bangalore, India - 560012
ramkiag@ee.iisc.ernet.in

A. Madhavaraj
Electrical Engineering,
Indian Institute of Science,
Bangalore, India - 560012
madhavaraj@mile.ee.iisc.ernet.in

P. Balachandran
Private Technology Consultant,
Trivandrum, India - 695102
pbn.tvm@gmail.com

***Abstract*—The sum of linear prediction coefficients (LPCs) is proposed as an effective feature in discriminating between sonorants and fricatives in continuous speech. On the closed set of sonorant and fricative frames of the entire TIMIT test database, a classification accuracy of 98.23% is obtained. When this feature is combined with three other features derived from the LPCs, the feature vector achieves an accuracy of 98.27% using a linear support vector machine classifier. The accuracy increases to 98.41% with mel frequency cepstral coefficients also added. The robustness of the feature has been tested on additive white, babble and pink noise.**

***Keywords*—Linear prediction, phonetic feature, manner class, V/U classification, segmentation, sonorant, fricative**

I. Introduction

Recent neuro-physiological experimental studies on speech perception [1], [2] have shown that humans extract phonetic features rather than directly the phones. Also, it is recognized that integration of knowledge of phonetic classes into a statistical based ASR system [3], [4] supplement its performance. Our ultimate goal is to find a suitable robust, speaker independent, acoustic correlate for each of the 'phonetic features' (PFs). Towards this purpose, in this paper, we are addressing one specific task, namely, discriminating between 'sonorants' and 'fricatives' from a continuous speech signal. In this work, we consider the class of sonorants as consisting of vowels and voiced consonants excluding voiced stops (/b/, /d/, /g/) - all of them voiced, with the only exception being /hh/, which is unvoiced and the class of fricatives as the unvoiced phones /s/, /sh/, /f/, /ch/ and the mixed voiced-unvoiced phones /z/, /zh/, /jh/, /th/, /dh/ and /v/.

In the literature, this problem has been studied under the context of manner classification and landmark detection [5]-[6] and also in the context of extraction of distinctive features [7]-[9]. The problem of identifying sonorants vs. unvoiced fricatives may also be looked upon as a V/U classification problem, which has been extensively studied in the literature [10]-[19].

978-1-5386-8235-7/18/$31.00 ©2018 IEEE

Several methods have been proposed for the identification of broad phonetic classes and/or their onsets from a speech signal. Liu [5] has used the change of energy between two frames spaced 50 ms apart, over five sub-band signals, for detecting the onsets or landmarks of four broadly defined classes. Salomon et. al. [6] have used a set of twelve temporal parameters to achieve manner classification. A team of researchers have used landmark based approach [7] for feature extraction and employed SVMs to identify the distinctive features, which in turn may be used for manner classification. King and Taylor have used Mel-frequency cepstral coefficients (MFCCs) and their temporal derivatives to train a neural network to identify distinctive features comprising broad manner classes [8]. Juneja and Wilson combined MFCCs with temporal features and used an SVM classifier for manner classification [9].

Most of the methods on V/U classification use the following temporal features [10]: (i) the relative energy of a frame (low for unvoiced frames), (ii) the ratio of energies in the low to high frequency bands (typically high for voiced frames), (iii) the number of zero-crossings per unit interval (high for unvoiced frames), (iv) the value of normalized autocorrelation at one sample lag, which indirectly relates to the first reflection coefficient in linear prediction (LP) analysis and captures the gross spectral slope (lowpass for voiced and highpass for unvoiced), (v) periodicity detection (voiced sounds are periodic) and (vi) pitch prediction gain. Zekeriya et. al. [15] have proposed measures similar to the above for V/U classification. Deng and O'Shaughnessy [16] have used an unsupervised algorithm for V/U classification.

Other features have also been considered. Alexandru Caruntu et. al. [12] have used zero-crossing density, Teager energy and entropy measures. Dhananjaya and Yegnanarayana [17] have used glottal activity detection, which in turn requires epoch extraction. Molla et. al. [18] have modeled the speech signal as a composite signal of intrinsic mode functions. The trends of these functions are compared with thresholds obtained on a training data for V/U classification. Statistics of LP coefficients (LPCs) have also been used for V/U classification [19]. Speech signal as a composite signal made up of harmonic and noisy structures along with a probabilistic model has been proposed for V/U segmentation of noisy speech [21].

In this paper, we demonstrate that the sum of LPCs, a scalar measure, is highly effective in discriminating the sonorant class from the fricative class in a speech signal [20]. LP is a very successful speech analysis technique [22]. According to the frequency domain interpretation, LP technique estimates an optimal all-pole digital filter, $\frac{1}{A(z)}$, that best approximates the short-time spectrum of a frame of speech signal. The reciprocal (all-zero) filter, A(z), called the digital inverse filter, is given by

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_p z^{-p} \qquad (1)$$

where a_1, a_2,..., a_p are the LPCs and p is the number of LPCs, whose value can be set during the estimation of LPCs. Here, z^{-1} is the unit delay operator given by $z^{-1} = e^{-j2\pi fT}$, where T is the sampling interval. Hence, the gain of the filter $A(z)$ at $z = 1$ (or frequency = 0), i.e., $A(1)$ is given by

$$A(1) = 1 + a_1 + a_2 + ... + a_p \qquad (2)$$

which corresponds to the sum of the LPCs.

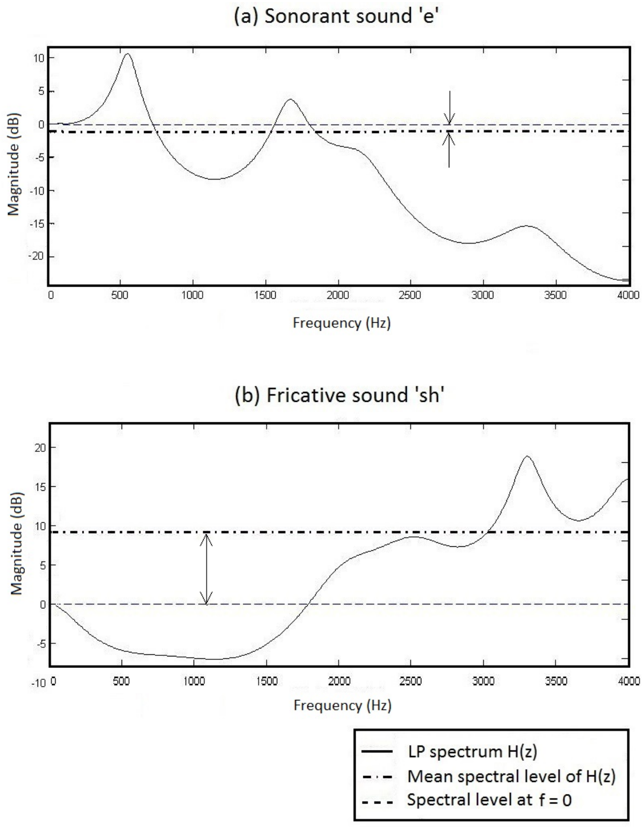

Fig. 1. Normalized spectral plots of the LP filter $H(z)$ [reciprocal of the inverse filter $A(z)$] for a sample frame each of sonorant and fricative sounds.

We report an interesting application of $A(1)$ for discriminating sonorant from the fricative segments of a speech signal and test its effectiveness using the TIMIT test database. The rationale for using $A(1)$ as a feature arises as follows. In the autocorrelation method, the values of LPCs are independent of the input signal level. Hence, the zero-th lag of the autocorrelation, $R(0)$, for a frame of a digital speech signal, may be assumed to be unity. For a given bandwidth, assuming $R(0)$ (which is the same as the signal energy) as unity implies that the mean magnitude squared spectral level is a constant (by Parsevals theorem). The value of $A(1)$ represents the spectral level at $f = 0$ of the inverse filter $A(z)$. We interpret this spectral level at $f = 0$ relative to the mean spectral level. Typically, for unvoiced sounds, the spectral level at $f = 0$ of the LP filter, $H(z) = 1/A(z)$, is much lower than the mean spectral level and vice-versa for voiced sounds. Figure 1 shows the difference between the mean spectral level and spectral level at $f = 0$ for a single frame each of the sonorant phone /a/ and fricative /sh/, respectively. Thus, $A(1)$ acts as a discriminating factor.

II. Characteristics of the contour of sonorant-fricative discrimination index (SFDI)

The speech signal is divided into frames of 20 ms with a frame shift of 5 ms. Hanning window is applied on the mean-removed frames. The stable, autocorrelation method of LP technique [22] is used. The computed LPCs depend only on the spectral shape and not on the signal level. In other words, the value of $A(1)$ is independent of the energy of the frame. The number of LPCs is chosen as $(fs + 2)$, where fs is the sampling frequency in kHz. LPCs are computed on the preemphasized and windowed speech signal. The computed value of $A(1)$ is assigned as a constant value to the mid 5-ms segment of the speech frame. Since the frame shift is also 5 ms, this assignment results in a staircase-like contour of $A(1)$.

An illustrative example: The utterance (sa2.wav) “Don't ask me to carry an oily rag like that” from the TIMIT database is analyzed. The speech wave and the computed $A(1)$ values are shown in Fig. 2(a) for a part of the utterance. The TIMIT labeled boundaries of the phones are also shown in the figure. Here, we have set the value of $A(1)$ to be zero for the silence frames, which are identified by a threshold on the energy. $A(1)$ rises sharply at the onset of the fricative /s/, reaches the maximum value of 23 and falls sharply at the end of the fricative segment. We note that for sonorants, the maximum value of $A(1)$ is low (< 1.1). We make use of this property of $A(1)$ and study its utility for the 2-class problem of sonorants vs. fricatives.

For the TIMIT test set, the maximum value of $A(1)$ observed for fricatives is 119 and the minimum is 0.058 for sonorants. The large value of A(1) for fricatives dominates over the sonorants. The variation in $A(1)$ within a fricative segment

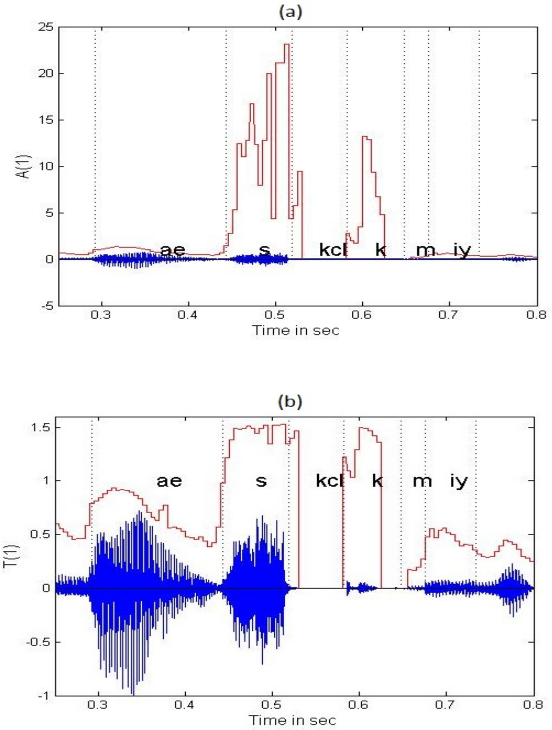

Fig. 2. (color online) Plots of the frame-wise values of (a) the sum of LPCs, A(1) and (b) the SFDI, $T(1) = \tan^{-1}[A(1)]$ for the utterance, "ask me".

is not of interest as long as the value is above a threshold. Hence, we prefer to compress the range of $A(1)$ using

$$T(1) = \tan^{-1}[A(1)] \tag{3}$$

We have used the inverse tan function instead of a logarithmic function just in case A(1) were to take on zero or a negative value, though during our investigation, we did not come across a single instance, where $A(1)$ is zero or negative. Such a compression of $A(1)$ also helps in graphic visualization along with the plot of the normalized signal waveform. The upper bound for $T(1)$ is $\pi/2$. Fig. 2(b) shows the plot of $T(1)$, which compresses the range of $A(1)$ and swamps out the variations when $A(1)$ is large. Henceforth, $T(1)$ is termed as the sonorant-fricative discrimination index (SFDI).

For the stop segment, marked as /k/ in Fig. 2, $T(1)$ reaches the maximum value of 1.4 for a part of the segment. Stop bursts are usually preceded by a silence or a low level voicing, which may be utilized for their detection [23]. Other phones also exhibit mixed characteristics for $T(1)$. However, the detection of stop bursts and other phones is not the topic of this paper.

III. Experiments and results

A. Histogram of SFDI for sonorants and fricatives

For computing the histograms, the labeled boundaries in the TIMIT database [24] are utilized to identify the sonorant and fricative segments. SFDI is computed frame-wise over these segments. For the purpose of computing histograms, /hh/ is excluded from sonorants, since it behaves like an unvoiced sound. Two groups of fricatives are studied, one with the phones /dh/ and /v/ excluded and the other with these phones included, since they often manifest as flaps, glides and stops [7] [25]. The value of SFDI is computed for all the frames within each segment of the two classes, sonorants and fricatives. The frame-wise accuracy is computed for the two classes on the entire training set of the TIMIT database. The normalized histograms of SFDI for the two classes (Fig. 3) show a clear separation between them. The number of sonorant frames falls sharply for SFDI values exceeding 1.18, whereas most fricative frames have SFDI values greater than 1.18.

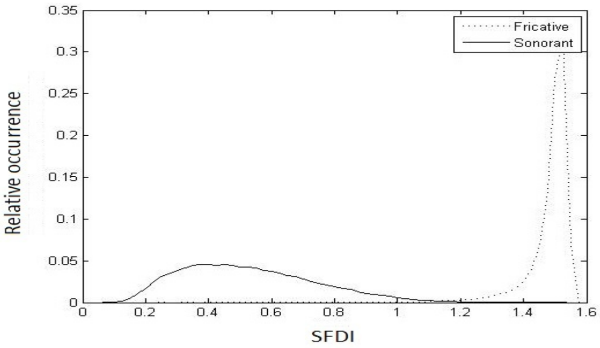

Fig. 3. Normalized histograms of the value of SFDI across the sonorant (solid) and fricative (dashed) segments in the entire TIMIT training set.

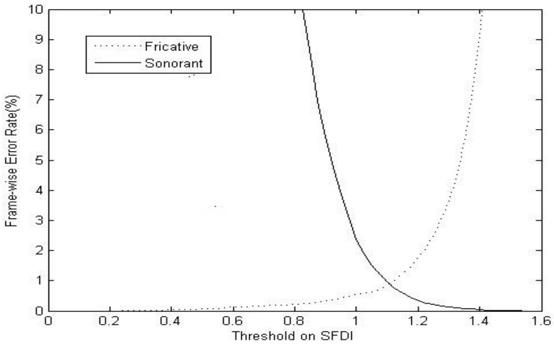

Fig. 4. Frame-wise error rate in the classification of sonorants (solid) and fricatives (dashed) as a function of the value of the threshold applied on SFDI, for the entire TIMIT training set.

B. Arriving at the threshold and the frame-wise performance

Assume that a threshold based logic is used and whenever SFDI is less than a threshold T, the frame is assigned to the sonorant class; else, to the fricative class. If SFDI for a known frame of a sonorant (fricative) is lesser (greater) than the threshold T, then that frame is considered to be correctly classified. The ratio of the number of correctly classified frames to the total number of frames gives the accuracy.

TABLE I
FRAME-LEVEL SONORANT-FRICATIVE CLASSIFICATION ACCURACY USING THE SCALAR SFDI FEATURE FOR DIFFERENT ADDITIVE NOISE TYPES AT DIFFERENT SNR VALUES. THE TEST DATA IS FROM THE ENTIRE TIMIT TEST DATABASE OF 1679 UTTERANCES AND THE TOTAL NUMBER OF TEST FRAMES IS 6,93,899.

SNR level	White noise	Babble noise	Pink noise
Clean speech	98.23		
20 dB	97.46	94.53	93.84
15 dB	97.08	93.60	92.07
10 dB	96.21	92.08	89.02
5 dB	94.15	89.11	83.88
0 dB	89.89	81.01	80.60

TABLE II
SONORANT-FRICATIVE DISCRIMINATION PERFORMANCE (DP) OF SFDI AND ITS COMBINATION WITH VARIOUS OTHER FEATURES ON THE ENTIRE TIMIT CLEAN SPEECH TEST SET.

Feature type	DP in %	DP with /dh/ and /v/ phones in fricative class
SFDI (arctan of sum of LPCs)	98.23	96.64
SFDI, rangeLP, maxLP, std-devLP	98.27	96.69
MFCC	97.04	95.67
MFCC, SFDI	98.17	96.46
MFCC, SFDI, rangeLP, maxLP, std-devLP	98.41	97.21

As the threshold is increased, the error rate falls sharply for the sonorants since less number of sonorant frames have a higher value of SFDI. On the other hand, as the threshold is increased, the fricative area under the normalized histogram below the threshold increases, thereby increasing the error rate for the fricatives.

Figure 4 shows the frame-wise error rate vs. threshold for the entire TIMIT training set comprising 4620 utterances (having a total of 18,78,941 frames). The cross-over point of error rates occurs at a threshold of about 1.18 and the corresponding error rate is about 0.8%. The discriminability of the scalar SFDI feature is also shown in terms of the receiver operating characteristic (ROC) graph in Fig. 5. It can be seen that the area under the ROC curve is almost close to one, which is desirable. For testing the effectiveness of the various sets of features, we have used the entire TIMIT test set containing 1679 utterances (with 6,93,899 frames). The accuracy (98.23%) obtained for clean speech is shown in Table I.

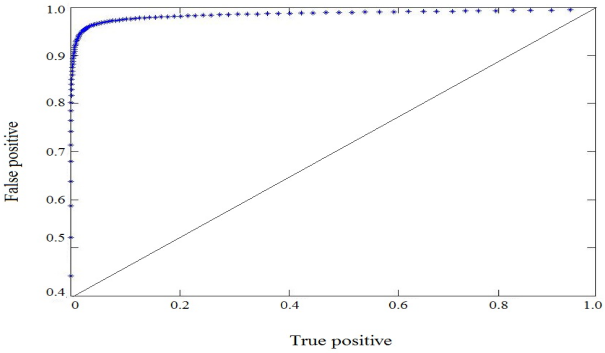

Fig. 5. (color online) ROC curve of the threshold based classifier using SFDI alone as a scalar feature on the sonorant and fricative segments of the entire TIMIT training set (18,79,941 frames from 4620 utterances).

C. Extended experiments

1) Robustness: In order to study the robustness of the SFDI measure, white Gaussian, pink and babble noise are separately added to the speech samples in different experiments, with an appropriate scale factor to achieve the desired global SNR. SNRs of 20 down to 0 dB, in steps of 5 dB, are considered in the experiments. The corresponding frame-wise accuracies are shown in Table I for the entire TIMIT test dataset. Assuming noise and speech signal to be uncorrelated, the spectral level shifts uniformly across all the frequencies and hence SFDI also increases. Thus, as the SNR decreases, the threshold T increases. The frame-wise accuracy even at 0 dB SNR is 89.9%, which is respectable.

We can clearly see that the performance of SFDI depends on the SNR level. To obtain the best possible sonorant-fricative classification accuracy for a test speech signal of unknown SNR, we can use the technique proposed in [26] to estimate the SNR of the test speech and then use the threshold corresponding to the estimated SNR (using look-up method) for classification purposes.

2) Performance of SFDI in tandem with other features: Based on the success of SFDI feature in discriminating sonorants from fricatives, we have tried to combine SFDI with 13-dimensional MFCC features and other LP-derived features (maximum value, range and standard deviation of LPCs). It can be seen from Table II that the performance of the MFCC features improves when concatenated with SFDI, but not sufficiently to match the performance of the standalone SFDI feature, whereas when we use the 3 LP-derived features along with SFDI, they perform marginally better. Best performance is obtained when 13-dimensional MFCC, 3-dimensional LP derived features and SFDI are concatenated together. Since these experiments involve multi-dimensional features, we have used SVM with linear kernel as the classifier. The optimal regularization parameter for SVM is calculated by 5-fold cross-validation on the training set. Using this parameter, the linear-SVM is trained on the training set and the model obtained is used for evaluating the test set. It is rather counter-intuitive to see that the performance of SFDI does not improve when concatenated with MFCC features. To confirm that this is not an artifact of the classifier, we have repeated these experiments on noisy speech and the results are tabulated in Table III. The performance trend of the different proposed feature sets on noisy speech is consistent with that on clean speech.

3) Comparison with previous work: Although a strict comparison with previous studies is not possible since the size of the database used in some cases and the tasks addressed in others are different, we make some broad observations for comparative purposes and these must not be construed as a criticism of the earlier results. Comparative evaluation

TABLE III
SONORANT-FRICATIVE CLASSIFICATION ACCURACIES (IN %) OF MFCC, SFDI, AND COMBINATIONS OF VARIOUS SETS OF FEATURES ON NOISY SPEECH OBTAINED BY ADDING AWGN NOISE TO TIMIT TEST SET AT DIFFERENT SNRS.

Feature type	20 dB	15 dB	10 dB	5 dB	0 dB
MFCC	97.00	96.23	94.38	91.05	83.80
SFDI	97.46	97.08	96.21	94.15	89.89
MFCC, SFDI	97.06	96.25	94.47	91.78	87.37
SFDI, LP derived feats.	97.51	96.96	96.21	94.16	90.11
MFCC, SFDI, LP derived feats.	97.64	97.12	96.35	94.50	90.75

TABLE IV
PERFORMANCE COMPARISON OF SFDI WITH STATE-OF-THE-ART METHODS ON CLASSIFICATION OF SLIGHTLY DIFFERENT BROAD PHONETIC CLASSES ON VARYING SIZED SUBSETS OF TIMIT TEST DATABASE. S: SONORANTS; F: FRICATIVES; V: VOICED; UV: UNVOICED; SIL: SILENCE; G: GLOTTAL; B: BURST. ZC: ZERO CROSSINGS; NR: NOT REPORTED. OUR (SFDI) RESULTS ARE BASED ON THE ENTIRE TEST SET OF 6,93,899 FRAMES, WHEREAS THE RESULTS OF [19] ARE BASED ON ONLY 2010 TEST FRAMES.

Method	Test data size	Classes handled	Feature type	Accuracy on clean speech(%)	Accuracy on noisy speech at 0 dB. (%)
Ours	6300 (utt)	S/F	SFDI (1-D)	98.23	89.89
[19]	2010 (frames)	V/UV	Intrinsic mode functions	99.57	98.96
[18]	380 (utt)	V/UV	Strength of glottal activity	94.40	85.70
[8]	1680 (utt)	V/F	MFCC (39-D)	93.00	NR
[5]	1680 (utt)	G/S/B	Sub-band energy difference	89.00	NR
[17]	1680 (utt)	V/UV/Sil	Signal energy, energy around harmonic, LP/HP, ZC	NR	NR
[6]	120 (utt)	4 classes	Temporal features (4-D)	74.80	NR
[9]	504 (utt)	5 classes	MFCC, temporal features	68.30	NR

of SFDI performance with the reported accuracies of state-of-the-art methods is given in Table IV for those reported for the TIMIT database. The reported accuracies for clean speech in the literature are in the range of 68.3% to 99.57% compared to 98.23% of the proposed scalar SFDI feature. The reported accuracies for 0 dB SNR are 98.96% and 85.7% [17] compared to 88.7% of the proposed feature. However, the highest accuracies of 99.57% for clean speech and 98.96% for 0 dB SNR are based on a set of only 2010 frames, whereas the proposed method is tested on the entire TIMIT test database of 6,93,899 frames.

IV. CONCLUSION

This study has demonstrated that a simple scalar measure SFDI, which is the inverse tan of sum of LPCs, along with a threshold based logic, may be effectively used to distinguish the sonorants from the fricatives. The experiments show that the discrimination given by SFDI is high even at 0 dB SNR. The results obtained are comparable, or in some respects, better than the state-of-the-art methods. Future research would be to utilize this property of SFDI, along with additional features, for automatic segmentation of a speech signal into different phonetic classes.

REFERENCES

[1] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang, Phonetic feature encoding in human superior temporal gyrus, Science, vol. 343(6174), pp. 1006 1010, 2014.

[2] Bahar Khalighinejad, Guilherme Cruzatto da Silva, and Nima Mesgarani, Dynamic encoding of acoustic features in neural responses to continuous speech, The Journal of Neuroscience, vol. 37(8), pp. 2176 2185, 2017.

[3] W. J. Barry and W. A. van Dommelen, The integration of phonetic knowledge in speech technology. Springer, vol. 25, 2005.

[4] P. Niyogi and P. Ramesh, The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets, Speech Communication, vol. 41, no. 2, pp. 349 367, 2003.

[5] S. A. Liu, Landmark detection for distinctive feature-based speech recognition, The Journal of the Acoustical Society of America, vol. 100, no. 5, pp. 3417 3430, 1996.

[6] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, Detection of speech landmarks: use of temporal information, The Journal of the Acoustical Society of America, vol. 115, no. 3, pp. 12961305, 2004.

[7] M. Hasegawa-J, J. Baker, S. Greenberg, K. Kirchhoff, J. Muller, K. Sonmez, S. Borys, K. Chen, A. Juneja, K. Livescu, S. Mohan, E. Coogan, and T. Wang, Landmark-based speech recognition, Report of the 2004 Johns Hopkins summer workshop, 2005.

[8] S. King and P. Taylor, Detection of phonological features in continuous speech using neural networks, Computer Speech & Language, vol. 14, no. 4, pp. 333 353, 2000.

[9] A. Juneja and C. Espy-Wilson, Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning, Proc. IEEE Int. Conf. Neural Information Processing, pp. 726 730, 2002.

[10] J. P. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm, ICASSP, Tokyo, vol. 11, pp. 473 476, April 1986.

[11] A. P. Lobo and P. C. Loizou, Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition, Proc. Int. Conf. Acoustics Speech and Signal Processing, Hong Kong, no. I, pp. 820 823, 2003.

[12] A. Caruntu, A. Nica, G. Toderean, E. Puschita, and O. Buza, An improved method for automatic classification of speech, IEEE Inter. Conf. on Automation, Quality and Testing Robotics, vol. 1, pp. 448 451, 2006.

[13] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, A multifeature voiced/nonvoiced decision algorithm for noisy speech, Proc. Int. Symp. Circuits and Systems, Kos, Greece, pp. 2525 2528, May 2006.

[14] D. Arifianto, Dual parameters for voiced-unvoiced speech signal determination, Proc. Int. Conf. Acoustics Speech and Signal Processing, Honolulu, no. IV, pp. 749 752, April 2007.

[15] S. Zekeriya, O. E. Yetgin, and O. Salor, Voiced-unvoiced classification of speech using autocorrelation matrix, pp. 1802 1805, 2014.

[16] H. Deng and D. O'Shaughnessy, Voiced-unvoiced-silence speech sound classification based on unsupervised learning, IEEE International Conf. on Multimedia and Expo, pp. 176 179, 2007.

[17] N. Dhananjaya and B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs, IEEE Sig. Proc. Letters, vol. 17, no. 3, pp. 273 276., March 2010.

[18] M. K. I. Molla, K. Hirose, S. K. Roy, and S. Ahmad, Adaptive thresholding approach for robust voiced/unvoiced classification, IEEE Inter. Symp. Circuits and Systems (ISCAS), pp. 2409 2412, 2011.

[19] K. Pattanaburi, J. Onshaunjit, and J. Srinonchat, Enhancement pattern analysis technique for voiced/unvoiced classification, Computer, Consumer and Control (IS3C),International Symposium on, pp. 389392, 2012.

[20] Ananthapadmanabha, T. V., A. G. Ramakrishnan, and Pradeep Balachandran, ”An interesting property of LPCs for sonorant vs fricative discrimination,” arXiv preprint arXiv:1411.1267, 2014.

[21] R. Rehr, M. Krawczyk, and T. Gerkmann, A posteriori voiced/unvoiced probability estimation based on a sinusoidal model, ICASSP, 2014.

[22] J. Makhoul, Linear prediction: A tutorial review, Proc. of the IEEE, vol. 63, no. 4, pp. 561 580, 1975.

[23] 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, The Journal of the Acoustical Society of America, vol. 135, no. 1, pp. 460 471, 2014.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, Acoustic-phonetic continuous speech corpus, DARPA TIMIT, U.S. Department of Commerce, Washington, DC, no. 4930, 1993.

[25] S. Zhao, The stop-like modification of /dh/: A case study in the analysis and handling of speech variation, Ph.D. thesis, M.I.T., Cambridge, 2007.

[26] K. V. Vijay Girish, A. G. Ramakrishnan and T. V. Ananthapadmanabha, Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation, Proc. 17th Annual Conf. of the International Speech Communication Association (INTERSPEECH 2016), San Fransico, USA, Sept. 8-12, 2016.