SPECTRAL TRANSITION MEASURE FOR DETECTION OF OBSTRUENTS

Maulik C. Madhavi[1], *Hemant A. Patil*[1] *and Bhavik B. Vachhani*[2].

[1] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India,
[2] TCS Innovation Labs, Mumbai, India.

{maulik_madhavi, hemant_patil}@daiict.ac.in, bhavik.vachhani@tcs.com

ABSTRACT

Obstruents are very important acoustical events (i.e., abrupt-consonantal landmarks) in the speech signal. This paper presents the use of novel *Spectral Transition Measure* (STM) to locate the obstruents in the continuous speech signal. The problem of obstruent detection involves detection of phonetic boundaries associated with obstruent sounds. In this paper, we propose use of STM information derived from state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) feature set and newly developed feature set, *viz.*, MFCC-TMP (which uses Teager Energy Operator (TEO) to exploit *implicitly Magnitude* and *Phase* information in the MFCC framework) for obstruent detection. The key idea here is to exploit capabilities of STM to capture high *dynamic* transitional characteristics associated with obstruent sounds. The experimental setup is developed on entire TIMIT database. For *20 ms* agreement (tolerance) duration, obstruent detection rate is found to be *97.59 %* with *17.65 %* false acceptance using state-of-the-art MFCC-STM and *96.42 %* with *12.88 %* false acceptance using MFCC-TMP-STM. Finally, STM-based features along with static representation (i.e., MFCC-STM and MFCC-TMP-STM) are evaluated for phone recognition task.

Index Terms— Mel frequency cepstral coefficients, obstruents, spectral transition measure, Teager Energy Operator.

1. INTRODUCTION

Speech signal can be classified into three categories, *viz.*, obstruent, sonorant and silence, based on the characteristics of speech production mechanism. All the sonorants are voiced by nature whereas obstruents are *voiced* as well as *unvoiced* [1]. Obstruents are produced by a constriction of articulators while air escapes from the lungs. They are typically spanned over different place of articulation from bilabial to glottal area and classified based on *manner* of articulation, *viz.*, stop (or plosive), affricates and fricatives [1].

Authors would like to thank Department of Electronics and Information Technology (DeitY), Government of India, New Delhi, India and DA-IICT, Gandhinagar for supporting this research work and providing the necessary resources.

Spectral features such as *Spectral Centre of Gravity* (SCG), energy in different frequency bands, formant transitions are also being used for recognition of *place* and *manner* of articulation for production of obstruents [2]. Earlier studies in the analysis of obstruents involved use of *temporal* and *spectral* (formant)-based information to detect *place* and *manner* of articulation. Fricatives-affricates are distinguished by features such as silence duration, frication duration, rise time, amplitude rise time [3]- [4]. For voiced plosive and fricative discrimination, zero-crossing rate and *Root Means Square* (RMS) energy were exploited [5]. In this paper, STM information is exploited to locate the obstruents. STM was *originally* proposed for syllable perception [6], then it was modified in the context of the phone boundary detection [7]. In this work, obsturent detection problem is posed as to detect the phone boundaries associated with obstruents, which is an extension of our earlier work [8]. As obstruents are produced with obstruction in the vocal tract cavity, rapid variations of spectral transition measure (STM) are expected in the vicinity of obstruents. This capability of STM motivated the authors to exploit it for obstruent detection task. Thus, novelty of the present work is to exploit spectral transition information (in order to capture the dynamics of the speech sound) for detection of obstruents. Results are shown on state-of-the-art MFCC (i.e., Mel Frequency Cepstral Coefficients) and newly proposed feature set, *viz.*, MFCC-TMP (i.e., Mel Frequency Cepstral Coefficients to capture Magnitude and Phase spectrum information via TEO).

2. MOTIVATION

Fig. 1 shows speech signal and corresponding phone-label along with corresponding (narrowband) spectrogram for a segment of an utterance taken from TIMIT database [9] (using *wavesurfer* software [10]). In Fig. 1, the spectrogram of different obstruent sound exhibits the energy distribution in different frequency regions. The arrows in Fig. 1, represents the boundaries associated with obstruents sounds. Since obstruents sounds are having *noise-like* characteristics and being *impulsive* in nature, it can be observed that there is *sudden* change in spectral energy distribution which is evident via spectrogram. From a speech production point of view, all the

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

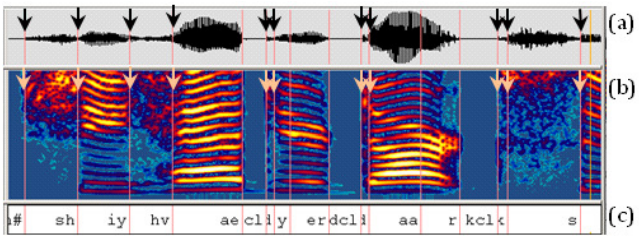


Fig. 1. (a) Time-domain speech signal, arrows in figure represents the phonetic boundaries associated with obstruent sounds, (b) corresponding spectrogram and (c) phone-labels for a segment of an utterance, *viz.*,"She had your dark " taken from TIMIT test database [9].

sonorant sounds are voiced in nature and few obstruents are unvoiced, vocal source information transits at few obstruent-sonorant boundaries [11]. In addition, high constriction (for plosive and affricates) and noise information (for fricative and affricates) play dominant role for the obstruent sound production. Furthermore, from speech perception viewpoint, *neural firing* patterns are directly related to the speech spectrum as suggested by *place theory of hearing* [11]. Hence, such *transitional* stimuli (i.e., at spectral change) indicate key motivation for using *STM* to detect phone boundaries [6].

The *STM* is derived from the *linear* regression coefficient information. Linear regression coefficients take *large* value due to the rapidly varying cepstral information in the vicinity of spectral transition and hence resulting STM is large in the vicinity of spectral transition. Such high transitions are the hypothesized phone boundaries [7]. From Fig. 1, the phone boundaries associated with obstruent are subject to higher spectral variation, resulting in higher *Mean Square Error* (MSE) in linear regression and which is reflected in STM contour. STM, at frame, i, can be computed as a MSE for linear regression [6], i.e.,

$$C_i = \frac{1}{D}\sum_{l=1}^{D} a_l^2(i), \tag{1}$$

where C_i is STM at given frame *i*, D is the dimension of the *spectral* feature vector (*13* in this case) and $a_l(i)$ is the *regression coefficient* or the rate of change of the spectral information and defined as [6].

$$a_l(i) = \frac{\sum_{k=-F}^{F} f_l(k+i)k}{\sum_{k=-I}^{I} k^2}, \tag{2}$$

where i represents the current *frame index*, l represents the coefficient index and F represents the *number of frames* (on each side of the current frame) used to compute these *regression* coefficients, which is similar to delta-coefficients [12]. One of the motivation behind development of delta representation of cepstral parameter was to compensate the spectral undershoot effect [13]. Since STM is derived using delta cepstrum, the information represented in terms of STM may capture perceptually-related linguistic information . We used $F = 25$ for a *1 ms* frame step corresponding to an interval of *50 ms* centered around the current frame at which $C(i)$ is computed. The rationale behind using *50 ms* interval is because of the fact that the perceptually essential interval for the perception of syllable (/CV/) unit is about *50 ms* [6]. Fig. 2 exhibits that how STM can be used to detect obstruent in the speech signal.

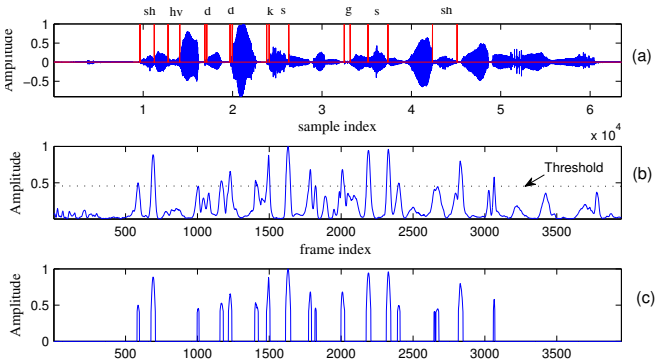


Fig. 2. (a) Speech signal with *manually* marked obstruents boundaries, (b) corresponding STM contour and a constant threshold *0.4*, (c) hypothetical obstruent boundaries which crosses amplitude threshold (*0.4*), for the TIMIT sentence, "She had your dark suit in greasy wash water all year".

STM for each frame is computed using eq. (1) and thus STM contour for an utterance is obtained. The peaks in this contour indicate probable transition of phones. It is evident from Fig. 2 shows that STM contour amplitude varies between *0* and *1*. In addition, generally it is below *30 %* to *40 %* of its maximum value. As spectral variation is higher around obstruent boundary (due to highly *dynamic* nature of obstruent sounds), large STM peaks may correspond to the obstruent boundaries. For a fixed threshold *0.4*, hypothesized obstruent boundary segments are shown in Fig. 2. The choice of threshold plays an important role in the detection task. In particular, a smaller threshold may result into many spurious boundaries and increases the *false alarm*, whereas higher value of threshold may *miss* obstruent boundaries. Hence, it would be worth to consider adjacent STM information, which exploits the rapidness behavior of STM.

3. EXPERIMENTAL SETUP

3.1. Feature Extraction

The details of the feature extraction computation and feature vector formation scheme is as follows. ***Step 1.*** *Computation of static features:* First, spectral features (f_i) is computed from speech signal. The state-of-the-art feature set, *viz.*, Mel Frequency Cepstral Coefficients (MFCC) was used as one of the speech representation. In addition, MFCC-TMP is used as another feature representation. The key difference between MFCC and MFCC-TMP is in the method for energy computation. In MFCC feature extraction, conventional l^2 is computed for subband energy computation. Hence, the energy of

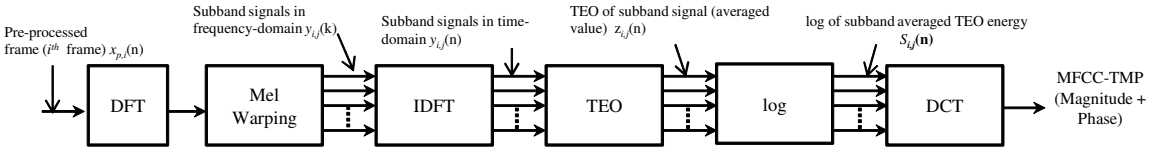


Fig. 3. Schematic block diagram of MFCC-TMP feature extraction. (After [14])

subband signal is equal to the sum of squared values of magnitude spectrum [15], whereas in MFCC-TMP, Teager energy is used instead of l^2 energy. Time-domain representation is used to compute TEO of subband signal [14]. Frame duration and shifting is *20 ms* and *1 ms*, respectively. *13-D* feature vector is computed by *30* Mel spaced subband filters. The feature extraction scheme for MFCC-TMP is shown in Fig. 3. Finally, normalized subband energy is computed followed by logarithm and Discrete Cosine Transform (DCT) operations to get proposed feature set, *viz.*, MFCC-TMP, i.e.,

$$MFCC-TMP_i(k) = \sum_{j=1}^{N_F} Sl_{i,j} cos\left(\frac{k(j-0.5)\pi}{N_F}\right), \quad (3)$$

where $k = 1, 2 \cdots N_c$, N_c = dimensions of feature vector (*13* in this work), N_F = number of filters used in the Mel filterbank (*30* in this work).

Step 2: After feature computation, STM is computed from features using eq. (1) and eq. (2).

Step 3: In order to incorporate the dynamics around the present STM value, adjacent values are also considered. The i^{th} frame of concatenated feature vector is defined as:

$$\mathbf{x}_i = [C_{i-N} \cdots C_{i-1} C_i C_{i+1} \cdots C_{i+N}]^T, \quad (4)$$

where N corresponds to number of adjacent STM values into consideration on either side in the preparation of concatenated feature vector. The practical framework for feature concatenation is very similar to the one used in [16] (where *9* adjacent frames are concatenated, i.e., *N=9*). The next subsection discusses about experimental setup.

3.2. Speech Corpus

The proposed approach is applied on the training and testing part of the TIMIT American English acoustic-phonetic corpus [9]. We have used the total *18* obstruents sounds (which are shown in Table 1) from TIMIT database along with training and testing examples. Total number of obstruent boundaries is *31128* in entire TIMIT database.

3.3. Setup for Performance Evaluation

In this paper, problem is formulated as a *binary* classification problem, where two classes correspond to obtruent boundaries and region of speech other than obstruent boundaries. Since STM is used to capture the phonetic boundaries, we may think of using proposed features in the vicinity of current STM values. This would incorporate the information about STM values and how fluctuations in STM value. Let us now define few nomenclatures. In particular, let B = anchor point corresponding to phonetic boundaries of obstruent (i.e., the ground truth), F_C = frame center point, ξ_C = tolerance within frame, ξ_A = frame agreement duration, $\mathbf{x}_t$ = t^{th} observation vector. We assign a two distinct labels to each frame based on the location of F_C and B. If $|F_C - B| < \xi_C$, then $\mathbf{x}_t$ = is considered to be belonging to *class 1* otherwise it belongs to *class 2*. It is as if a basket of $2\xi_C$ duration. For testing the algorithm, we may introduce agreement duration along with tolerance, i.e., if $|F_C - B| < \xi_C + \xi_A$, then $\mathbf{x}_t$ is considered to be belonging to *class 1* else it is assigned to *class 2*. Here, the basket size, $2(\xi_C + \xi_A)$ is flexible based on the value of ξ_A. In the experiments, ξ_C was kept as constant (i.e., *1 ms*) that models the frame where in obstruent boundaries are within $2\xi_C$ duration around F_C (tight central tendency for *1 ms*). Under the assumption that features of *class 1* are associated with phone boundaries of obstruent sounds and features of *class 2* are associated with other speech information than obstruent phone boundaries. The description of obstruent detection task is illustrated in Fig. 4. Gaussian Mixture Model (GMM) of *64* components to model the obstruent boundaries, is used for classification. Earlier GMM has been used extensively for speaker verification task in [17] and phoneme recognition task [18].

Phonetic Class	**Phonetic Symbols(# Train : # Test)**	
	Voiced	**Unvoiced**
Strong Fricative	/z/(3773:1273), /zh/(151:74)	/s/(7475:2639), /sh/(2238:796)
Affricates	-	/ch/(822:259), /jh/(1209:372)
Weak Fricative	/v/(1994:710), /dh/(2826:1053), /hv/(1154:369)	/f/(2216:912), /th/ (751:267), /hh/(957:256)
Stop	/b/(2181:886), /d/(3548:1245), /g/(2017:755)	/p/(2588:957), /t/(4364:1535), /k/(4874:1614)

Table 1. Obstruents Phonetic Symbols and Statistics of Training and Testing Samples.

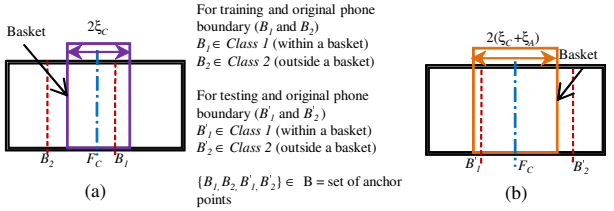


Fig. 4. Schematic diagram for performance evaluation for (a) training samples and (b) testing samples.

3.4. Performance Measures

The performance of proposed obstruent detection system is evaluated under two different evaluation metrics, *viz.*, % Detection Rate (DR) and % False Alarm Rate (FAR). % DR is defined as follows,

$$\%\ Detection\ Rate = \frac{N_{C1}}{N_{TC1}} \times 100, \quad (5)$$

where N_{C1} is the total number events when average frames within a basket considered are correctly classified as class *1*. NT_{C1} total number of baskets considered in a test experiment. % FAR is defined as follows.

$$\%\ False\ Alarm\ Rate = \left(1 - \frac{N_{C2}}{NT_{C2}}\right) \times 100, \quad (6)$$

where N_{C2} is the total number frames in *class 2* are correctly classified as *class 2*. NT_{C1} is total number of frames in *class 2* in a test experiment. For better performance, % DR should be higher and % FAR should be lower. To quote *statistical* significance of our experimental results, *95 %* confidence interval is also mentioned. For N trials, if the probability of success is p, then confidence interval is $[p-B, p+B]$, where B is the *band of confidence* and is defined as [19].

$$B = z_c \sqrt{\frac{p(1-p)}{N}} \times 100, \quad (7)$$

where z_c is called the level of *coefficient of confidence*. It is *1.96* for *95 %* confidence interval [19].

4. EXPERIMENTAL RESULTS

The experimental results are performed by two different representations. Fig. 5 shows the % DR and % FAR performance for different agreement duration in the step of *1 ms* along with *95 %* confidence interval. It can be observed from Fig. 5, for agreement duration of *2 ms*, % DR crosses to *90 %*, indicating that the very slight change in the agreement duration efficiently captures the STM information. Table 2 indicates the performance of obstruent detection system in terms of % DR and % FAR for different agreement duration using different feature sets, *viz.*, MFCC and MFCC-TMP along with *95 %* confidence interval.

From Table 2, it can be observed that as agreement duration ξ_A increases, % DR increases and % FAR decreases. It indicates that *tight* central tendency may miss out few acoustical events for obstruent boundary detection. In addition, it can be observed that as agreement duration increases, % DR increases, since we are incorporating the evidences from larger vicinity around the actual ground truth. Newly proposed feature set, *viz.*, MFCC-TMP, performs better for less agreement duration (i.e., tight agreement condition) whereas MFCC performs better than MFCC-TMP for large agreement duration. However, MFCC-TMP shows promising resulting in terms of False Alarm rejection. From Fig. 5 (b), it can be observed that MFCC-TMP consistently outperforms MFCC for *all* the agreement duration. This may be due to effectiveness of energy computation via TEO (which exploits instantaneous frequency). As shown in Fig. 3, the architecture of MFCC-TMP,

ξ_A (ms)	MFCC		MFCC-TMP	
	% DA	% FAR	% DA	% FAR
0	80.7 (80.26, 81.14)	23.41 (23.02, 23.81)	85.03 (85.00, 85.08)	19.68 (19.65, 19.72)
5	94.14 (93.88, 94.40)	21.21 (20.96, 21.48)	94.21 (94.17, 94.25)	16.79 (16.76, 16.83)
10	96.55 (96.54, 96.93)	19.63 (19.41, 19.85)	96.05 (96.02, 96.10)	14.61 (14.58, 14.64)
15	97.07 (96.88, 97.26)	18.26 (18.06 ,18.46)	96.64 (96.61, 96.68)	13.4 (13.37, 13.43)
20	97.59 (97.43, 97.77)	17.65 (17.44, 17.86)	96.42 (96.38, 96.46)	12.88 (12.85, 12.92)

Table 2. % DR and % FAR various agreement duration (along with *95 %* confidence interval).

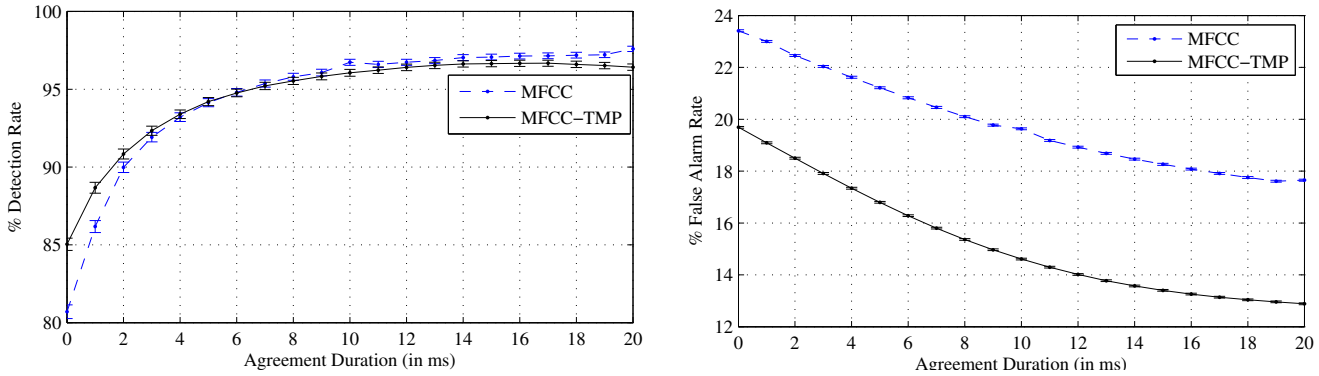
Feature sets	% Correct Detection
MFCC	*54.39*
MFCC-STM	***57.06***
MFCC-TMP	*54.24*
MFCC-TMP-STM	***56.35***

Table 3. Phone recognition performance on TIMIT dataset.

suggests that it has a capability to compute subband energy. TEO has a capability to capture instantaneous frequency for different subband [20]. The instantaneous frequency might be different across the obstruent boundaries and hence may play an important role in obstruent detection task. In addition, the effectiveness of STM has been tested for phone recognition task. TIMIT training data and testing data (excluding */sa/* files) are considered. It is well known fact that MFCC-delta representation improves speech recognition performance. STM is one-dimensional representation of delta features [6]. A group of *39* phones are considered as per details suggested in [21]. *13-D* static features (MFCC and MFCC-TMP) and *14-D* static features with their STM representation (MFCC-STM and MFCC-TMP-STM) are considered. The performance of % correct detection is shown in Table 3. From Table 3, it is found that STM-based feature adds information which results in improved the performance than that of static representation.

5. SUMMARY AND CONCLUSIONS

In this paper, the problem of obstruent detection is viewed as binary classification of two acoustical events which exploits boundary information around STM and information around its neighbourhood to detect the obstruents. Here, two different feature sets, *viz.*, MFCC and MFCC-TMP are used to compute STM. Proposed algorithm for obstruent detection can be used to classify the obstruents and refine newly developed features to improve the obstruent detection performance. In addition, STM-based features along with static representa-



(a) (b)

Fig. 5. (a) % Detection Rate (DR) (with *95 %* confidence interval) (b) % False Alarm Rate (FAR) (with *95 %* confidence interval) for various agreement duration ξ_A .

tion are tested for phone recognition task. It should be noted that, our claim here is not regarding relative performance of MFCC and MFCC-TMP. We are using MFCC-like representation via MFCC-TMP in order to capture spectral transition information to efficiently detect obstruent boundary. Our future work will be directed towards analysis of MFCC-TMP for phone boundaries of other phonemes as well.

REFERENCES

[1] Peter Ladefoged, *A Course in Phonetics*, Wadsworth Publishers, Belmont, California, 2005.

[2] J. Hoelterhoff and H. Reetz, "Acoustic cues discriminating german obstruents in place and manner of articulation," *J. Acous. Soc. Am.*, vol. 121, no. 2, pp. 1142, 2007.

[3] Z. Mahmoodzade and M. Bijankhan, "Acoustic analysis of the persian fricative-affricate contrast," in *16^{th} Int. Congress of Phonetic Sciences, ICPHS XIV, Saarbrucken*, 2007, pp. 6–10.

[4] L. J. Gerstman, "Noise duration as a cue for distinguishing among fricative, affricate, and stop consonants," *J. Acous. Soc. Am.*, vol. 28, no. 1, pp. 160, 1956.

[5] P. Howell and S. Rosen, "Production and perception of rise time in the voiceless affricate/fricative distinction," *J. Acous. Soc. Am.*, vol. 73, no. 3, pp. 976–984, 1983.

[6] S. Furui, "On the role of spectral transition for speech perception," *J. Acous. Soc. Am.*, vol. 80, no. 4, pp. 1016–1025, 1986.

[7] S. Dusan and L. R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries.," in *INTERSPEECH*, Pittsburgh, PA, USA, 2006, pp. 17–21.

[8] B. B Vachhani, K. D. Malde, M. C. Madhavi, and H. A. Patil, "A spectral transition measure based melcepstral features for obstruent detection," in *Int. Conf. on Asian Lang. Process. (IALP), 2014*. IEEE, 2014, pp. 50–53.

[9] J. S. Garofolo et al., "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[10] K. Sjőlander and J. Beskow, *Wavesurfer [Computer program] (Version 1.8.5)*, 2009.

[11] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Pearson Education India, 2002.

[12] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech & Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[13] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *IEEE Int. Conf. Acoust., Speech, and Signal Process., ICASSP'86*, Tokyo, Japan, 1986, IEEE, vol. 11, pp. 1991–1994.

[14] H. A. Patil and M. C. Madhavi, "Significance of magnitude and phase information via vteo for humming based biometrics," in *5^{th} IAPR Int. Conf. on Biometrics (ICB), 2012*. IEEE, 2012, pp. 372–377.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[16] V. Keri and K. Prahallad, "A comparative study of constrained and unconstrained approaches for segmentation of speech signal.," in *INTERSPEECH*, Makuhari, Japan, 2010, pp. 2238–2241.

[17] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[18] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process., ICASSP*. IEEE, 2003, vol. 1, pp. I–68.

[19] N. A. Weiss and M. J. Hasset, *Introductory Statistics*, Addison Wesley, Reading, MA, 1993.

[20] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Int. Conf. on Acoustics, Speech and Signal Process., ICASSP*. IEEE, 1990, pp. 381–384.

[21] K-F. Lee and H-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 37, no. 11, pp. 1641–1648, 1989.