Combining Evidences from Variable Teager Energy Source and Mel Cepstral Features for Classification of Normal *vs.* Pathological Voices

Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar-382007, Gujarat, India. Email: hemant_patil@daiict.ac.in

Abstract

In this paper, novel Variable length Teager Energy Operator (VTEO) based Mel frequency cepstral coefficients, *namely,* VTMFCC are proposed for automatic classification of normal *vs.* pathological voices. Experiments have been carried out using this proposed feature set, Mel Frequency Cepstral Coefficients (MFCC), and their score-level fusion. Classification was primarily performed using a discriminatively-trained *2nd* order polynomial classifier on a subset of the MEEI database for a feature dimension of 12. The equal error rate (EER) on fusion was reduced by 3.2% than MFCC alone which was used as the baseline. The classification accuracy was analyzed for different dimensions of feature vector. Furthermore, results obtained for the *2nd* order classifier were compared with the results obtained from the *3rd* order polynomial classifier for different feature dimensions. In addition, the effectiveness of dynamic features, in particular, delta, delta-delta, and shifted delta cepstral features have been investigated for this particular problem. It has been observed that the score-level fusion (with equal weights) of proposed feature set and state-of-the-art MFCC gave better classification performance than MFCC alone for various evaluation factors considered in this paper.

Index Terms:-Pathological voice, nonlinearity, VTEO, glottal closure instant (GCI), VTMFCC, polynomial classifier.

1. Introduction

Exploiting excitation source and vocal tract system information to develop automated and non-invasive method to classify normal vs. pathological voice is the objective of this study. It has been assumed by many studies that the perceived abnormality in the voice is primarily due to changes at the glottal excitation source and is not as much due to abnormalities in the vocal tract configuration. Hence, many of the earlier studies have concentrated on quantifying the perturbations in fundamental or pitch frequency (Fo) called as jitter[1], [2] and pitch amplitude (i.e., shimmer) [1] together called *perturbation* factors and quantifying the noise at the glottal source using *noise* measures such as harmonics-to- noise ratio (HNR), glottal-to-noise excitation ratio (GNE), normalized noise energy (NNE), etc. [3]-[5]. However, the issue with these features is that they can be applied reliably to only nearly periodic signals or Type 1 voices as described in [6], since they require accurate determination of pitch period (i.e. To=1/Fo), which is difficult in case of severely pathological voices. There have been other studies which have been based on source-filter theory [7], [8] and nonlinear dynamics [9], [10]. However, these nonlinear dynamical methods are based on chaos and fractals, which are computationally complex.

To alleviate these issues, computationally simple nonlinear feature based on Teager energy operator (TEO) is proposed [11], [12], [39]. To the best of author's knowledge, these are some of the earliest studies in this area employing the TEO to capture the nonlinearity in the properties of glottal airflow. For example, study reported in [12] used the AM autocorrelation envelope of the first formant (F_1) of speech signal. Apart from these techniques there have been other studies which have attempted to characterize the non-stationary characteristics of speech using time-frequency analysis, wavelets, and wavelet packets [13]-[15]. Another linear and perceptually relevant feature that has been used is MFCC [16], [17]. Fusion of Modulation Spectral features (mRMS) with MFCC which are hypothesized to capture the mucosal wave variation due to increased mass and the noise introduced due to incomplete closure of vocal folds [17], [18]. Other recent approaches in this problem are reported in [40]-[42]. Apart from this study, there have been very few studies which have used fusion of features, especially fusion of both source and system-related features to characterize the effects due to pathology in the perceived speech.

In this paper, we exploit excitation source-related nonlinear variable length Teager energy operator (VTEO)-based Mel cepstral feature set, i.e., VTMFCC. The key idea behind VTMFCC is to represent source information in perceptually meaningful way via Mel warping. This work is an extension of author's work reported in [19], where preliminary results were obtained on fusion of MFCC and VTMFCC which showed a robust performance even under degraded conditions. In this paper, a more detailed analysis has been performed where the effect of feature dimension, order of polynomial classifier as well as the efficacy of dynamic features has been investigated.

2. VTMFCC Feature Set

For the discrete-time signal, TEO is defined as [20]-[23]:

$$\psi\{x(n)\} = x^2(n) - x(n-1).x(n+1) \approx A^2\omega^2, \quad (1)$$

for small values of ω, i.e., $\sin(\omega) \approx \omega$, where A and ω are the amplitude and frequency of a sinusoidal signal. It can be observed from eq. (1) that TEO represents the running estimate of signal's energy by considering the joint contribution of signal's amplitude and frequency. This concept of TEO was further generalized in [24] to use the i^{th} past and i^{th} future samples instead of just the two adjacent samples. The location, *i* of these two samples from the current sample is called the *dependency index* (DI). This generalized version of the TEO operator is called the VTEO, $\xi_i\{.\}$ and is defined as:

$$\xi_i\{x(n)\} = x^2(n) - x(n-i).x(n+i) \approx i^2A^2\omega^2. \quad (2)$$

DI is termed so, because the VTEO operator brings out the hidden *dependencies* in the sequence of the samples of speech signal.

2.1. Evidence of VTEO as Excitation Source Information

Fig. 1(a) and Fig. 1(b) depict the speech signal and

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

corresponding differenced electroglottograph (EGG) waveform taken from the CMU-ARTIC database [25], [38]. Fig. 1(c) and Fig. 1(d) are the VTEO profiles of speech signal shown in Fig. 1(a) for DI=1 and DI=4, respectively. It can be observed that the peaks in the VTEO profile are in close proximity of locations corresponding to the differenced EGG waveform for both DI=1 and DI=4, which corresponds to the glottal closure instants (GCI), indicating that the VTEO successfully captures the airflow properties at the glottis (in particular, *glottal activity* [26]). Moreover, it can be seen that for DI=4 the peaks are significantly higher or dominant than the peaks in the VTEO profile for DI=1.

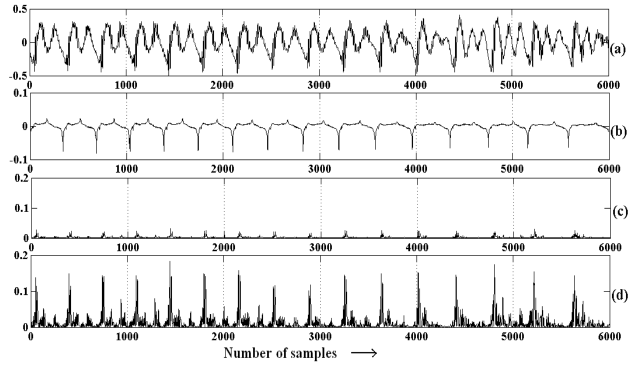

Fig. 1. (a) Speech for a male speaker from CMU-ARCTIC Database [32], (b) Differenced EGG, (c) VTEO profile for DI=1, and (d) VTEO profile for DI=4.

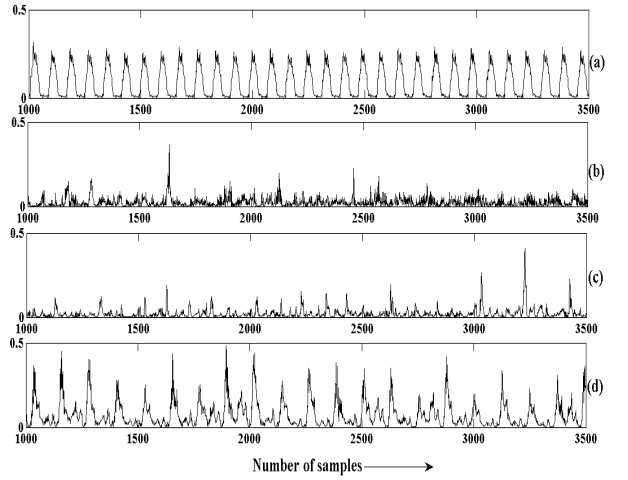

Fig. 2. VTEO profile (with DI=4) of sustained phonation of */ah/* taken from MEEI database for (a) normal speech, (b) paralysis, (c) vocal fold polyp, and (d) Reinke's edema.

Fig. 2 depicts the VTEO profile for DI=4 for a normal speaker and three pathological speakers suffering from paralysis, vocal fold polyp and Reinke's edema, respectively. All the samples were taken from the Massachusetts Eye and Ear Infirmary (MEEI) database [31]. Details of various pathological diseases can be found in [27]. These particular voice disorders have been considered since they represent the three broad categories of the myriad diseases affecting the vocal folds. As can be seen in Fig 2(a), for a normal speaker, due to complete glottal closure, there is not much *turbulence* at the source which is reflected in the regularity of occurrence of peaks of the VTEO profile. On the other hand, in the case of all the pathological voice examples shown in Fig. 2(b)-2(d), due to *incomplete* closure and/or *asymmetrical* vibration of the glottis, there is an increased turbulence and aperiodicity at the source, which is reflected in the irregular structure in the running estimate of signal energy via the VTEO profile. Thus, it is clear that the VTEO profile successfully characterizes the airflow properties through the glottis and captures the changes in the air flow properties by exploiting the dependencies in the sequence of samples brought about by the various pathologies.

2.2. VTMFCC Feature Extraction

VTMFCC was first proposed in [28] and then applied for voice pathology detection [19]. Its implementation is shown via a block diagram in Fig. 3.It is computed by first preprocessing (framing, Hamming windowing, pre-emphasizing) the speech signal $x(n)$ to give $x_p(n)$, the pre-processed speech signal. Next, the VTEO as defined in (2) is taken, followed by Mel scale warping of the VTEO profile and then the usual log and DCT computation is carried out to get VTMFCC. In [28]-[30], this feature set was used for biometric application using the hum and was shown to give a better classification accuracy using score-level fusion of VTMFCC and MFCC than MFCC alone and was shown to provide complementary information to the MFCC feature set.

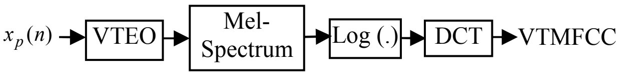

Fig. 3. Block diagram for proposed VTMFCC. After [26], [35].

From the analysis presented in Section II-B and II-C, it is evident that the proposed VTMFCC feature set represents *perceptually* meaningful information (due to Mel scale warping to mimic human perception process for hearing) of excitation source characteristics. In our earlier study, we have observed that DI=4 gives relatively better results for VTMFCC [19] and hence, this choice of DI is used for the experiments reported in this paper.

3. Experimental Results

3.1. Experimental Setup

The corpus used for the experiments is the commercially available MEEI database [31]. For this work, a sustained phonation */ah/* of a subset of 173 pathological and 53 normal speakers was used according to [3]. The MEEI database consists of samples either at 50 kHz or 25 kHz, so all the samples were down-sampled to 25 kHz sampling frequency. Since the number of pathological samples is approximately three times the normal samples, 1 s of pathological data and 3 s of normal data per person was used for training and testing. A *4*-fold cross-validation scheme repeated 12 times giving a total of 48 trials was carried out, using 75% samples for training and 25% for testing, with the training and testing subsets kept independent to each other. The classification accuracy (% number of correctly classified samples out of total samples from normal *vs.* pathological samples) was calculated as an average for all these 48 trials. This ensures that the results reported in this study are statistically significance and independent of train and test set.

3.2. The Polynomial Classifier

In this paper, discriminatively-trained polynomial classifier is used as basis for all the experiments [35]. In this classifier, the input feature vector x is converted to a polynomial vector $\mathrm{p(x)}$, which contains terms till degree *d* for a *d*th order polynomial classifier. The score is produced by taking the *inner product* of this class model with the input test feature vector which is averaged over time to give the final score for that speaker, i.e.,

$$s = \frac{1}{N}\sum_{i=1}^{N} w_{class}^{T} p(x_i), \qquad (9)$$

where N is the number of frames, w_{class} is the model for either the normal or pathological class and $p(x_i)$ is the polynomial expansion for i^{th} testing feature vector. Thus, whichever model gives the higher score, the test sample is said to belong to that class. The details of training algorithm are given in [32],[33].For data fusion, a score-level fusion with β=0.5 was carried out, i.e., VTMFCC and MFCC features were fused with equal weights to give the fused scores as,

$$S_f = \beta . S_{MFCC} + (1-\beta) . S_{VTMFCC} \;, \qquad (6)$$

where S_f, S_{MFCC}, and S_{VTMFCC} represent scores for fusion, MFCC, and VTMFCC, respectively. These scores were used to plot the Detection Error Trade-off (DET) curves and get equal error rate (EER) (a point where false alarm and miss probability is same) as the performance measure [34].

TABLE 1 EER (%) nd Classification Accuracy (%) for Different Feature Dimension (D) for VTMFCC (DI=4)

FD	4	8	12	16	20	24	28	32
Ac	79.50	89.99	94.31	95.54	96.95	**97.62**	73.29	59.63
EER	20.50	10.01	5.28	4.46	3.05	**2.38**	33.30	44.05

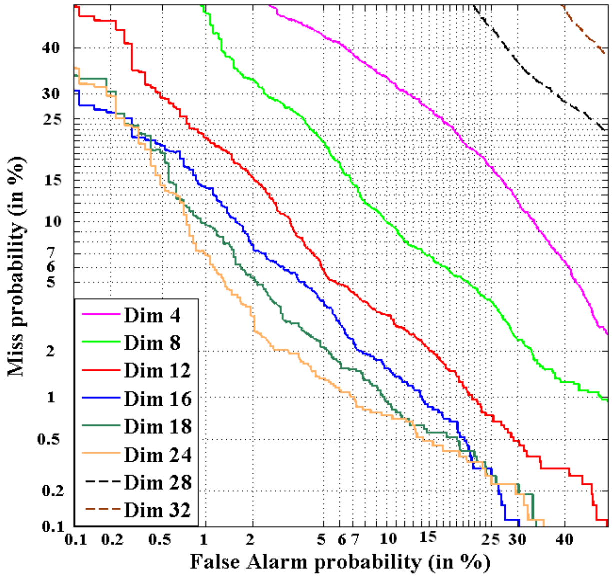

Fig. 4. DET plot for varying feature dimension (FD) for VTMFCC (DI=4) and a polynomial classifier of order 2.

3.3. Effect of Feature Dimension (FD)

The objective of this experiment was to find the relatively optimal feature dimension in the proposed feature set keeping other specifications (DI and degree of classifier) constant. The feature dimension was varied from 4 to 32 in steps of 4. Thus in each case, FD number of VTMFCC (with DI=4) coefficients were extracted per frame. Each frame consisted of 256 samples corresponding to 10.2 ms with a 50 % overlap.The EER values and the corresponding classification accuracy (average of 48 trials) (Ac) has been shown in Table 1.

As can be seen from Table 1, the accuracy is gradually increasing while the EER gradually decreasing till feature dimension of 24 (shown in bold). The accuracy increases from 79.5 % (corresponding to an EER of 20.50 %) to 97.62 %, (corresponding to an EER of 2.38 %), for a feature dimension of 4 and 24, respectively. Thereafter, there is a sudden fall in accuracy to 73.29 % for a feature dimension of 28 which further decreases to 59.63 % and an EER of 44.05% for a dimension of 32.Fig. 6 depicts the DET plots for the different feature dimensions (i.e., 4, 8, 12, 16, 20, 24, 28, and 32). As can be seen that the feature dimension of 24, gives the lowest EER and is thus the *optimal* feature dimension for the proposed feature set.

3.4. Effect of Order of Classifier

The experiments were conducted to investigate effect of polynomial order in the classifier for VTMFCC, MFCC and their score-level fusion for a given feature dimension (as shown in Table 2 and Fig. 5). Some of the observations from Table 2and Fig. 5 are as follows:

1) It can be seen that in all the cases for both MFCC and VTMFCC, *3rd* order classifier performs much better than the *2nd*order classifier.
2) It was found that for the lower feature dimensions, the difference in accuracies was much greater than for higher dimensions, showing small changes from 12 to 16 as compared to the changes from 8 to 12. As seen in the Fig. 7, the gap between the accuracies obtained for VTMFCC feature set gets closer and closer as D increases.

TABLE 2 EER (in %) for Different Feature Dimension for Degree 2 and 3 for VTMFCC (DI=4), MFCC, and their Score-Level Fusion

Feature Set	Degree of Classifier	Feature Dimension (FD)				
		8	10	12	14	16
VTMFCC	Order 2	10.00	8.41	5.28	4.76	4.46
(DI=4)	Order 3	6.96	5.28	3.64	3.53	4.24
MFCC	Order 2	5.88	4.61	4.50	3.79	3.01
	Order 3	4.87	4.05	3.98	3.64	3.53
Fusion	Order 2	7.03	**1.41**	**0.82**	1.15	0.44
	Order 3	2.08	**1.38**	**1.19**	3.53	4.24

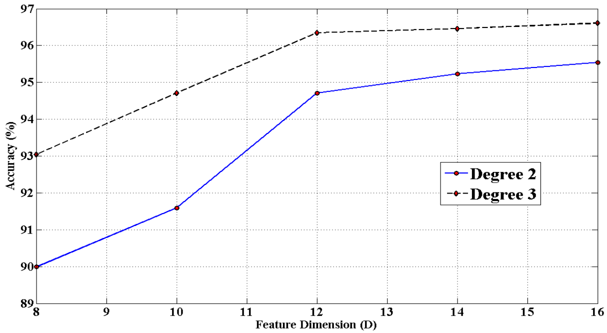

Fig. 5. Comparison of the accuracy for *2nd*and *3rd*order polynomial classifier for varying feature dimension for VTMFCC alone.

3.5. Effect of Dynamic Features

Dynamic features have shown to improve the performance of speech and speaker recognition systems when combined with the static features [36], [37]. Thus, in this set of experiments, we have explored the effectiveness of few dynamic features, i.e., delta (Δ) cepstrum, delta-delta (ΔΔ) cepstrum and shifted delta cepstrum (SDC). The delta cepstrum is the first order derivative of the *cepstral trajectory* of static features (such as MFCC and VTMFCC). Similarly, the delta-delta cepstrum is its second order derivative [36]. Thus, the delta cepstrum and delta-delta cepstrum takes into account the velocity and acceleration of cepstral trajectories, respectively and hence give information about changes in their temporal (dynamic) behavior. On the other hand, the shifted delta cepstrum (SDC) is basically a variation of the delta cepstrum over a longer period of time. It consists of four parameters, denoted as N-d-*p*-*k*. Here, *N* stands for the number of cepstral coefficients extracted per frame, *d* stands for the samples over which delta computation is spread, *p* stands for the gaps between the successive delta computations, and finally *k* stands for the number of delta computations forming one SDC.

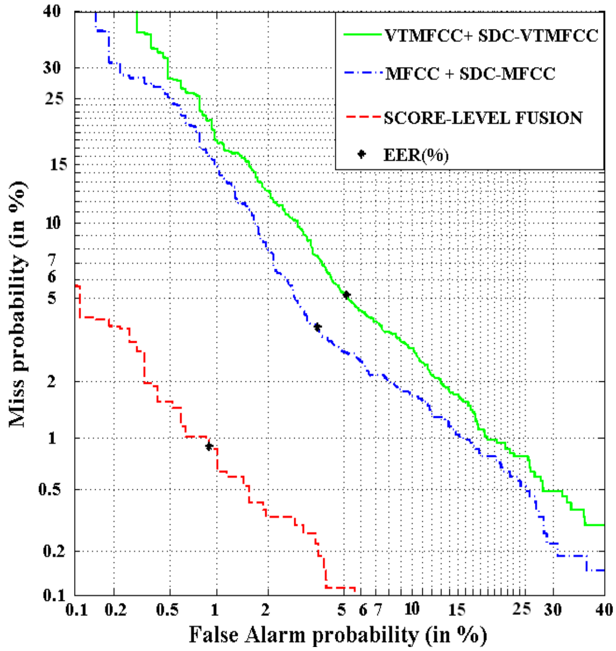

Fig. 6. DET plot of VTMFCC + SDC-VTMFCC, and MFCC + SDC-MFCC dynamic features, and their score-level fusion.

TABLE 3Classification Accuracy (%) and EER (%) Values of Dynamic Features and their Score-Level Fusion

System	Feature Set	Acc(%)	EER(%)
1	MFCC	95.65	4.53
2	VTMFCC	94.31	5.69
3	System 1 + System 2	98.85	1.13
4	MFCC + Δ MFCC	95.95	4.06
5	VTMFCC+ Δ VTMFCC	95.09	4.91
6	System 4 + System 5	**99.03**	**0.97**
7	MFCC + ΔΔ MFCC	96.06	3.94
8	VTMFCC + ΔΔ VTMFCC	94.05	5.95
9	System 7 +System 8	**99.07**	**0.93**
10	MFCC + SDC-MFCC	96.32	3.68
11	VTMFCC +SDC-VTMFCC	94.87	5.13
12	System 10 + System 11	**99.11**	**0.89**

+ indicates score-level fusion

In the following experiments, the speech data was blocked into frames of length 10.2 ms or 256 samples with 50% overlap. *12*-dimensional MFCC and VTMFCC were extracted per frame. The delta and delta-delta parameters were extracted by taking the first and second order derivative per frame. For the SDC parameters, a 12-2-2-5 SDC was taken. Instead of concatenating the SDC frames to get a *kN*-dimensional feature vector, in these set of experiments, we use a weighted sum of the frames, where highest priority is given to the first frame and lowest to the *(k-1)th* frame, so as to get an *N*-dimensional feature vector. In all cases the dynamic features extracted was concatenated with the static features (i.e., MFCC or VTMFCC), forming a feature vector with a dimension of 24 (i.e., 12 MFCC + 12 Δ-MFCC, 12 MFCC + 12 ΔΔ-MFCC, 12 MFCC + 12 SDC-MFCC), and similar concatenation was done for VTMFCC.

The results have been shown in Table 3. It is evident from Table 3 that feature-level fusion of MFCC and Δ MFCC does not perform significantly better than only static MFCC features. This is the case for VTMFCC as well. However, it is interesting to note that score-level fusion of static and dynamic features (i.e., Δ, ΔΔ, SDC) derived from MFCC and VTMFCC (highlighted in Table 3) gave a relatively better performance. On the whole, SDC gave relatively the best performance. This may be due to the fact that SDC captures better temporal variation over longer time duration. However, on comparing the best results obtained among the dynamic features, namely, system 12 in Table 3, with its static counterpart, namely, system 3 in Table 3, we see that there is an increase in accuracy by just 0.26% and a decrease in EER by 0.23%, which is not a very significant improvement in classification performance. Thus, it can be seen that even though the dynamic features do provide some extra information, the relative increase in performance (with respect to its static counterpart) is not as significant to compensate for the two-fold increase in feature dimension which in turn increases the computational cost.

The DET plot for the best performing dynamic feature set, SDC has been plotted in Fig. 9. It can be seen that just as in the case of static features there is a significant decrease in the EER value on fusion of VTMFCC+SDC-MFCC and MFCC+SDC-MFCC (system 12) by almost 2.79% as compared to the EER value of MFCC+SDC-MFCC (system 10). This was relatively the smallest EER value obtained amongst all the features (static and dynamic) investigated for a feature dimension of 12. The accuracy also in this case was the highest equal to 99.11%, which indicates it is a very accurate classification system.

4. Summary and Conclusions

In this paper, novel VTEO-based Mel frequency cepstral features have been proposed to take into account the *nonlinearity* in the speech production mechanism caused due to nonlinear sources and capture excitation source-related information. Even though the MFCC feature set gave better classification accuracy as compared to the VTMFCC when taken alone, the inherent problem with MFCC is that it is based primarily on source-filter theory and does not account for the *coupling* of vocal tract with source. It has been shown that the proposed feature set provides complementary information to state-of-the-art MFCC features and shows a significant decrease in EER value when a score-level fusion of MFCC and VTMFCC with equal weights is carried out. The effect of feature dimension, order of classifier, and the effectiveness of dynamic features for proposed feature set have also been investigated. Further research could be conducted in investigating other data fusion strategies such as feature-level or classifier-level fusion.

5. Acknowledgements

The author would like to thank Dr. Pallavi Baljekar for her kind help and suggestions during this research work,

6. References

[1] S.B. Davis, “Computer evaluation of laryngeal pathology based on inverse filtering of speech,” *SCRL Monograph*, no. 13,1976.

[2] A.M. Smith and D. G. Childers, “Laryngeal evaluation using features from speech and electroglottograph,” *IEEE Trans. on Biomedical Engg.,* vol. 30, no.11, pp.755-759, Nov.1983.

[3] V. Parsa and D.G. Jamieson, “Identification of pathological voices using glottal noise measures,” *J. Speech, Language,HearingRes*., vol.43, no.2, pp. 469–485,2000.

[4] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco and Fernando Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders,"*J. Voice*, vol. 24, no. 1, pp. 47-56,2010.

[5] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, “Normalized noiseenergy as an acoustic measure to evaluate pathologic voice,” *J. Acoust.*
Soc. Am., vol. 80, no. 5, pp. 1329–1334, 1986.

[6] I. R Titze, R. J. Baken, and H. Herzel, “Evidence of chaos in vocal fold vibration,” in *Frontiers in Basic Science*, I. R Titze, Ed., San Diego, CA: Singular Publishing Group,1993, pp.143–188.

[7] M. O. Rosa, J. C. Pereira and M. Grellet, “Adaptive estimation of residue signal for voice pathology diagnosis,” *IEEE Trans. Biomed. Eng,* vol. 47, no.1, pp.96-104, Jan 2000.

[8] L.G. Ceballos and J.H.L. Hansen, “Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection”, *IEEE Trans. on Biomedical Engg,* vol. 43, no. 4, pp. 373-383, 1996.

[9] P. Henrıquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Dıas-de-Marıa, “Characterization of healthy and pathological voice through measures based on nonlinear dynamics,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1186–1195, Aug.2009.

[10] G. Vaziri, F. Almasganj, and R. Behroozmand, “Pathological assessmentof patients’ speech signals using nonlinear dynamical analysis,” *Comput. Biol. Med.*, vol. 40, no. 1, pp. 54–63, 2010.

[11] D. A. Cairns and J. H. L. Hansen, “A noninvasive technique for detecting hypernasal speech using a nonlinear operator,” *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 35–44, Jan. 1996.

[12] J. H. L. Hansen, L.Gavidia-Ceballos, and J. F. Kaiser, “A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment,” *IEEE Trans. on Biomedical Engg*., vol. 45, no.3,pp. 300-313, March 1998.

[13] P. Scalassara, C. Maciel, J. Pereira, “Predictability analysis of voice signals,” *IEEEEngineering in Medicine and Biology Magazine,* vol.28, no.5, pp.30-34, Sept-Oct 2009.

[14] K. Umapathy, S. Krishnan, V. Parsa, D.G. Jamieson, “ Discrimination of pathological voices using a time-frequency approach,” *IEEE Trans. on Biomedical Engg.,* vol. 52,no.3,pp. 421-430, March 2005.

[15] C.D.P. Crovato and A. Schuck, “The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices,”*IEEE Trans. on Biomedical Engg.,* vol. 54, no. 10, pp.1898-1900, Oct. 2007.

[16] J.I Godino-Llorente, P. Gomez-Vilda, “Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. on Biomedical Engg.*, vol. 51, no.2, pp.380-384, Feb. 2004.

[17] M. Markaki, Y. Stylianou, “Voice pathology detection and discrimination based on modulation spectral features,”*IEEE Trans. Audio, Speech, Lang. Process,*vol. 19, no. 7, pp. 1938-1948, 2011.

[18] M. Markaki, Y. Stylianou, J.D. Arias-Londoño, J.I., Godino-Llorente, “Dysphonia detection based on modulation spectral features and cepstral coefficients,”*IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing(ICASSP),* March 2010, pp. 5162-5165.

[19] H. A. Patil, and P.N. Baljekar, “Novel VTEO based Mel cepstral features for classification of normal and pathological voices,” accepted in *INTERSPEECH, Florence, Italy*, 27-31 Aug. 2011.

[20] H. M. Teager, and S. M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,”*In Speech Production and Speech Modelling* , W.J. Hardcastle, and A. Marchal (Eds.) Netherlands: Kluwer, 1990, pp. 241-261.

[21] H.M.Teager, “Some observations on oral air flow during phonation,”*IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 28, pp.599-601, 1980.

[22] H. M. Teager and S. M. Teager, “A phenomenological model for vowel production in the vocal tract,” *Proc.Speech Sciences: Recent Advances,* San Diego, CA: College Hill Press, 1983.

[23] J.F. Kaiser, “On a simple algorithm to calculate energyof a signal,” *in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP* Apr 1990, vol. 1, pp. 381-384.

[24] V. Tomar and H. A. Patil, “On the development of variable length Teager energy operator (VTEO),” *Proc. INTERSPEECH, Brisbane, Australia*, Sept. 2008, pp. 1056-1059.

[25] J. Kominek , A.W. Black, “CMU-ARCTIC speech databases,” *in 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA*, 2004, pp. 223-224.

[26] K.S.R. Murty, B. Yegnanarayana, M.A. Joseph , “Characterization of glottal activity from speech signals,”*IEEE Signal Processing Letters* , vol.16, no.6, pp. 469-472,June 2009.

[27] P.L. Dhingra, *Diseases of the Ear, Nose and Throat*, 3rd ed., New Delhi: Elsevier, 2004.

[28] H. A. Patil and K. K. Parhi, “Novel variable length Teager energy based features for person recognition from their hum,” *Int. Conf. Acoust., Speech, Signal Processing (ICASSP) 2010, Dallas, USA*, pp. 4526-4529.

[29] H. A. Patil, M. C. Madhavi, K. K. Parhi, “Combining evidence from spectral and source-like features for person recognition from humming,”in *INTERSPEECH, Florence, Italy*, 27-31 Aug. 2011.

[30] H. A. Patil, and M. C. Madhavi, Combining evidences from magnitude and phase information using VTEO for person recognition using humming," in special issue of Recent advances in speaker and language recognition and characterization Computer Speech and Language, Elsevier, In Press, Sept. 2017**.**

[31] Kay Elemetrics Corp, Disordered Voice Database Model 4337, Version 1.03, Massachusetts Eye and Ear Infirmary Voice and Speech Lab, 2002.

[32] W. M. Campbell, K. T. Assaleh and C. C. Broun, “Speaker recognition with polynomial classifiers,” *IEEE Trans. on Speech and Audio Processing*, vol.10, no.4, pp. 205-212, May 2002.

[33] H. A. Patil and T. K. Basu, “A novel approach to language identification using modified polynomial networks,” *in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, Studies in Computational Intelligence,* vol.83, B. Prasad and S.R.M. Prasanna (Eds.), Berlin Heidelberg, Germany: Springer-Verlag, March 2008, pp.117-144.

[34] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,”*In Proc. EUROSPEECH’97*, pp. 1895-1898.

[35] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., New York: Wiley-Interscience, 2000.

[36] S.Furui, “Cepstral analysis for automatic speaker verification,” *IEEE Trans. on Audio Speech and Signal Proc*., vol. 29, pp. 254–272, 1981.

[37] R. J. Calvo, R. Fernandez, G. Hernandez, “Application of shifted delta cepstral features in speaker verification,” *Proc. INTERSPEECH, Antwerp, Belgium, 2007*, pp. 734-737.

[38] “CMU-ARCTIC speech synthesis databases,” [Online]. Available: http://festvox.org/cmu_arctic/index.html {Last Accessed June 26, 2019}.

[39] L. Salhi, A. Cherrif, “ Robustness of auditory Teager energy cepstrum coefficients for classification of pathological and normal voices,” The Scientific World Journal, 2013.

[40] K. Doudi and B. Bertrac, “On classification between normal and pathological voices using the MEEI –Kay PENTAX database: issues and consequences,” INTERSPEECH 2014, Singapore.

[41] H. Wu, J. Soraghan, A. Lowit, G. Di-Caterina, “ A deep learning method for pathological voice detection using convolutional deep belief networks,” INTRSPEECH 2018, Hyderabad, pp. 446-350.

[42] A. H. Poorjam, M. A. Little, J. R. Jensen, M. G. Christensen, “A parametric approach for classification of distortions in pathological voices,” ICASSP 2018, pp. 286-290.