Novel VTEO Based Mel Cepstral Features for Classification of Normal and Pathological Voices

Hemant A. Patil, Pallavi N. Baljekar

1 Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India.
2 Department of Electronics and Communication, Manipal Institute of Technology (MIT), Manipal University, Manipal, India.

hemant_patil@daiict.ac.in, pallavi.baljekar@learner.manipal.edu

Abstract

In this paper, novel variable length Teager Energy Operator (VTEO) based Mel cepstral features, viz., VTMFCC are proposed for automatic classification of normal and pathological voices. Experiments have been carried out using this proposed feature set, MFCC and their score-level fusion. Classification was performed using a 2nd order polynomial classifier on a subset of the MEEI database. The equal error rate (EER) on fusion was 3.2% less than EER of MFCC alone which was used as the baseline. Effectiveness of the proposed feature-set was also investigated under degraded conditions using the NOISEX-92 database for babble and high frequency channel noise.

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

1. Introduction

The main motivation to carry out classification of normal vs. pathological voices is to provide an automated, reliable method for non-invasive objective evaluation of patients. Pathologies of the larynx affect the morphology of the vocal folds, resulting in incomplete closure of vocal folds and their asymmetric vibration due to increased mass and loss of elasticity. The incomplete closure of glottal folds causes increased turbulence in airflow at the glottis and their asymmetric vibration causes the vocal folds to vibrate independently at two different frequencies, which may cause transient or permanent diplaphonia or the perception of more than one fundamental frequency component in a voice.

It has been assumed by many researchers that the perceived abnormality in the voice is due to changes at the glottal 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 pitch frequency (i.e., jitter) [1] 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 [2]. The disadvantages of these perturbation and noise measures is that they can be applied reliably to only nearly periodic signals or type 1 voices as described in the classification by Titze [3], since they require accurate determination of pitch period, which is difficult in case of severely pathological voices. There have been other studies which have been based on source-filter theory [4]. However, these methods are sub-optimal in a sense because they do not take into account the nonlinearity in the speech production mechanism. In recent years, there has been a paradigm shift towards using nonlinear features, mainly derived from chaos and information theory [5], based on the assumption that speech is the output of a nonlinear dynamical system. However, these concepts have been borrowed and brought from either nonlinear control theory (e.g., Lyapunov stability criterion) or chaos and fractal theory. Moreover, these features are computationally difficult to implement. Another nonlinear feature that has been used in the speech processing literature is the AM autocorrelation envelope of the first formant of speech signal. To the best of authors’ knowledge, this is the only work employing the Teager Energy operator (TEO) to capture the nonlinearity in the properties of glottal airflow [6]. Very recently, a method has been developed to fuse features (i.e., feature-level fusion) which capture the amplitude envelope fluctuations, i.e., the 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 [7]. Apart from this, there have been very few studies which have used fusion of features, especially fusion of both source and system features to characterize the effects due to pathology in the perceived speech. In this paper, we propose fusion of novel nonlinear Variable length Teager Energy Operator (VTEO) based features with MFCC. VTEO characterizes the glottal source properties, by modeling the effects due to the nonlinear sources of speech production, while MFCC perceptually models the vocal tract configuration characterizing the system features. As a result we study how the source (VTEO profile) and system (MFCC) together provide complementary information for a very accurate and reliable detection of pathological voices. In addition, the robustness of this proposed feature set under degraded conditions is also studied.

This paper is organized as follows: In section 2 we briefly review the VTEO and its ability to capture the source related information and describe the proposed Variable length Teager Energy Mel Frequency Cepstral Coefficients (VTMFFC) feature. In section 3 we describe the classification experiments and their results. In section 4 we summarize our findings and discuss the further possibilities.

2. Variable Length Teager Energy Operator (VTEO)

In [8], the authors hypothesize that the airflow in the vocal tract is not planar as assumed in the linear source-filter theory, but rather is a nonlinear turbulent flow. They found that even slow, nonpulsatile fluids give rise to turbulent and nonlaminar
flow due to difficulty of the fluid in conforming to the changes in cross-section and direction in the vocal tract. According to them, the flow is not laminar but separates into various paths leading to the generation of vortices. It is these vortices in the vicinity of the false vocal folds that provide the nonlinear sources of excitation to the vocal folds during the closed phase. Thus, TEO is an operator which captures the energy of these nonlinear sources and is found to be proportional to the square of amplitude and frequency. It is defined in both the discrete and continuous domain. For the discrete case it is defined as:

$$\psi \{x(n)\} = x^2(n) - x(n-1)x(n+1) \approx A^2 \omega^2$$  \hspace{1cm} (1)

for small values of $\omega$ i.e., $\sin(\omega) \approx \omega$.

This concept of TEO was further generalized in [9] to use the $i^{th}$ past and $i^{th}$ future samples instead of just the 2 adjacent samples. The location of these 2 samples from the current sample is called the dependency index (DI). This generalized version of the TEO operator is called the VTEO and is defined as:

$$\xi \{x(n)\} = x^2(n) - x(n-i)x(n+i) \approx i^2 A^2 \omega^2$$  \hspace{1cm} (2)

DI is termed so, because the VTEO operator brings out the hidden dependencies in the sequence of the speech signal. Thus, the VTEO, maps the discrete-time samples of a signal onto a running estimate of the signal’s energy. Even though the expressions (1) and (2) above are valid only for bandpass signals, if we bandpass filter the signal above, we will get a damped sinusoid and we will not get dominant bumps in the TEO or VTEO profile. As a result, the nonlinearities in the interaction of the source with system are neglected.

### 2.1. VTEO as speech source information

Since VTEO is a generalization of TEO, which is known to capture the nonlinear sources of sound production namely the vortices, so also VTEO has been shown to capture this source information. Figure 1(a) and (b) depict the speech signal and corresponding differenced EGG waveform taken from the CMU-ARCTIC database [12]. Figure 1(c) shows the corresponding VTEO profile for DI=4. It is interesting to see in this figure that the peaks in the VTEO profile are in close proximity of locations corresponding to the differenced EGG waveform which corresponds to the glottal closure instants (GCl), indicating that the VTEO successfully captures the airflow properties at the glottis (in particular glottal activity).

Figure 2, depicts the VTEO (DI=4) for a normal and pathological speaker taken from MEEI database. As can be seen, for a normal speaker, due to complete glottal closure, there is not much turbulence at the source which is reflected in the regularity of the VTEO profile peaks. On the other hand, in the case of the pathological voice, due to incomplete closure, there is increased turbulence at the source, which is reflected in the irregular structure in the running estimate of signal energy of the VTEO profile. This evidence again reiterates the fact that VTEO for DI=4 is very good at capturing the airflow properties at the glottal source.

### 2.2. Variable length Teager Energy Mel frequency cepstral coefficients (VTMFCC)

VTMFCC as a feature was first proposed in [10]. It’s implementation is shown as a block diagram in Figure 3. It is computed by first preprocessing (framing, Hamming Windowing, Pre-emphasis) the speech signal $x(n)$ to give $x_p(n)$. Next, the VTEO as defined in Eq.(2) is taken, followed by Mel scale warping of the VTEO magnitude waveform and then the usual log and DCT computation is carried out, as done for MFCC. In [12], it was used for biometric application using the hum and was shown to give a better classification accuracy than MFCC parameters taken alone and was shown to provide complementary information to the MFCC features.

From the analysis presented in Section 2.1, it is evident that the proposed VTMFCC feature set represents perceptually meaningful information (due to Mel scale warping) of speech source characteristics.

### 3. Experiments

In this section, an automatic method of classification of normal and pathological voices is presented. The robustness of the proposed features is also tested under degraded conditions using the NOISEX-92 database [13].
3.1. Data and Methods

The corpus used for the experiments is the commercially available MEEI database [11]. For this work, a subset of 173 pathological and 53 normal speakers was used according to [2]. The MEEI database consists of samples either at 50 kHz or 25 kHz, so all samples were downsampled to 25 kHz sampling frequency. Since the number of pathological samples is approximately 3 times the normal samples, 3s of pathological data for patients and 1s of normal data of sustained phonation /ah/ for control people 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 accuracy, as plotted in Figure 4(b), was calculated as an average for all these 48 trials.

For each case, the speech signal was first framed into blocks of 256 samples or 10.2 ms duration with 50% overlap corresponding to 128 samples overlap. Each Frame was multiplied by a Hamming window and 12 MFCC coefficients and 12 VTMFCC coefficients for DI equal to 1 to 20 were extracted per frame. These features were then fed to a 2nd order polynomial classifier to generate the true and false scores [14]. These scores were then used to plot the Detection Error Tradeoff (DET) curves and get Equal Error Rate (EER) as the performance measure [15].

For data fusion, a score-level fusion with $\alpha=0.5$ was carried out, i.e., VTMFCC and MFCC features were fused with equal weights to give the fused scores, as follows:

$$C_f = \alpha C_{MFCC} + (1 - \alpha) C_{VTMFCC}$$

(3)

3.2. Results

Experiments were carried out to test the discrimination capability of VTMFCC feature alone for DI varying from 1 to 20. MFCC which was used as the baseline for comparison gave an accuracy of 95.65% and the corresponding EER was 4.35%. For the VTMFCC feature alone the classification using DI=4, gave the best relative performance with an accuracy of 94.31% and EER of 5.69%. Score-level fusion with $\alpha=0.5$ was done according to Eq. (3). The EER and classification accuracy as a function of DI is plotted in Figure 4(a) and 4(b), respectively.

As can be seen from the DET plot shown in Figure 5, with this fusion EER decreased to 1.15%, i.e., a significant drop in EER by 3.2% and there was an increase in classification accuracy by almost 2.2%. Thus, it is clearly seen that the proposed features do provide significant complementary information to that provided by MFCC alone. It was shown in Section 2, the VTEO operator is successful in capturing the source airflow properties, showing that the VTMFCC features characterize the source-like perceptually relevant information from the speech signal, whereas the MFCC features, account for the vocal tract (system) characteristics. Hence, on fusing these source and system features, we obtain a significant decrease in the EER, giving a very good classification accuracy of 98.85%.

The significant increase in accuracy for DI=4, may be attributed to the fact that around DI=4, it is capturing the sharp transitions in the speech signal. Thus, for pathological voices, which show a much larger irregularity in their signal structure as compared to the normal samples, VTMFCC for DI=4 captures these irregularities in an optimal manner and hence gives a high classification accuracy and minimum EER.

3.2.1 Evaluation under degraded conditions:

To verify the robustness of the features under degraded conditions, two noise types, viz., babble and high frequency channel noise for the case of additive noise from the NOISEX-92 database were used. These noises were used for 3 different SNR values of the input signal, 35 dB (severely degraded conditions), 5 dB (slightly degraded conditions) and 10 dB (good conditions). The results are shown in Table 1 for these different SNRs.

As can be seen from the table, the fusion performs significantly better in all 3 cases, viz., clean, babble and HF channel noise for different SNR values. Figure 6, illustrates...
the DET plot for the severely degraded condition, i.e., -5 dB. As is evident from the plot, in all 3 cases, viz., clean (red), babble noise (blue) and HF channel noise (green), the fusion (solid line) performed significantly better than the MFCC (dashed line) and VTMFCC (DI=4)(dotted line) features taken alone.

Table 1. EER values of VTMFCC (DI=4), MFCC, and their score-level fusion for the two noise types, babble and HF channel noise for 3 different SNR values.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Fusion babble</th>
<th>MFCC babble</th>
<th>VTMFCC babble</th>
<th>Fusion HF channel</th>
<th>MFCC HF channel</th>
<th>VTMFCC HF channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>1.34</td>
<td>5.10</td>
<td>6.21</td>
<td>4.84</td>
<td>5.39</td>
<td>7.96</td>
</tr>
<tr>
<td>5</td>
<td>1.19</td>
<td>4.69</td>
<td>5.99</td>
<td>4.09</td>
<td>4.65</td>
<td>5.69</td>
</tr>
<tr>
<td>10</td>
<td>1.15</td>
<td>4.54</td>
<td>6.21</td>
<td>4.13</td>
<td>4.50</td>
<td>5.99</td>
</tr>
</tbody>
</table>

Figure 6: DET plots of VTMFCC (DI=4), MFCC and their score level fusion for clean speech (red), babble noise (blue) and HF channel noise (green) for SNR=-5dB.

The reason for the significant increase in EER in the case of HF Channel noise as compared to the babble noise can be attributed to the fact that one of the main discriminating characteristics of pathological voices is their increased high frequency components and so by introducing high frequency characteristics of pathological voices it increases the EER. In spite of this fact, we see that the proposed feature set on score-level fusion does give significantly better results and it can be inferred that it is robust under degraded conditions.

4. Conclusions

In this paper, novel VTEO-based Mel cepstral features have been proposed to take into account the nonlinearity in the speech production mechanism caused due to nonlinear sources and capture perceptually meaningful source information. Even though the MFCC features give better classification accuracy as compared to VTMFCC when taken alone, the inherent problem with MFCC is that it is based on source-filter theory and does not account for the coupling of vocal tract with source. It has been shown that these VTMFCC features provide complementary information to state-of-the-art MFCC features and show a significant decrease in EER value on score-level fusion of MFCC and VTMFCC with equal weights. The robustness of these features has also been investigated under degraded conditions for two types of noise. It has been observed that, the proposed feature set works well in presence of babble noise, but loses its discriminating power in presence of high frequency noise.

The VTEO-based features could open promising new directions to characterize the nonlinearity in speech production mechanism by providing a more realistic way of characterizing the airflow properties (in particular, glottal activity) at the glottal source and thus may be used as an alternative to features derived from chaos theory and information theory, which are computationally intensive approaches.

5. Acknowledgements

The authors would like to thank the DA-IICT, Gandhinagar, India authorities for their kind co-operation in carrying out this research work and ECE Dept. MIT, Manipal, India for providing the MEEI database.

6. References