NOVEL VARIABLE LENGTH TEAGER ENERGY BASED FEATURES FOR PERSON RECOGNITION FROM THEIR HUM

Hemant A. Patil\textsuperscript{1} and Keshab K. Parhi\textsuperscript{2}

\textsuperscript{1}Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India-382 007.

\textsuperscript{2}Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA.

ABSTRACT

Most of the state-of-the-art voice biometrics systems use the natural speech signal (either read speech or spontaneous or contextual speech) from the subjects. In this paper, an attempt is made to identify speakers from their hum. A new feature set, \textit{viz.}, Variable length Teager Energy Based Mel Frequency Cepstral Coefficients (VTMFCC) is proposed for this problem. Experiments have been carried out for person identification and verification task using Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) with polynomial classifier of 2\textsuperscript{nd} order approximation. It is shown that the speaker identification rate for proposed feature set outperforms LPCC by 13.6\% and is competitive over baseline MFCC. For speaker verification, a reduction in equal error rate (EER) by 1.73\% is achieved when a score-level fusion system is employed by combining evidence from MFCC and VTMFCC.

Index Terms— Voice biometrics, Humming, VTEO

1. INTRODUCTION

In this paper, we propose a voice biometrics system for identification of speakers based on their hum using variable length Teager energy-based acoustic features. A hum is a sound made by singing a wordless tone with the mouth completely closed, forcing the sound to emerge from the nose. To hum is to produce such sound, most often with a melody. As humming contains no linguistic information, voice biometrics based on humming is a challenging research issue. However, a humming-based speaker recognition system may be applicable to a person with speech disorder and an infant, who is not able to speak \cite{1}, \cite{2}. In terms of universality, which is an essential criterion to be considered while designing any biometric systems \cite{3}, humming is more universally available on everyone than speech \cite{4} and has relevance in forensic conditions \cite{5}.

Fig. 1 and Fig. 2 show the hum sampled at 22050 Hz (and spectrograms) produced for a Hindi song, \textit{viz.}, ‘Ye Sham Mastani Madhhosh Kiye Jaay (This beautiful evening charges me),’ by two male speakers of age 21 years. It is evident from the plots (both time-domain and pitch striations in spectrograms) that the hum signal is mostly periodic in nature. In addition to this, pattern of hum signal, pitch contour, formant contour and spectrogram for each speaker are distinct. This motivates us to investigate whether we can
use hum produced by the speakers for voice biometrics problem. This may be due to the fact that hums are nasalized sounds and nasal cavity remains steady during hum production and is known to be speaker-specific [5].

Recently, there has been a growing interest in designing Query by Humming (QBH) system for music retrieval applications [6]. In addition, we may design a QBH system which responds to hum query from a particular speaker. In this context, authors have reported use of spectral features for humming-based speaker recognition where performance of MFCC was found to be better than LPC and LPCC [7]. Jin et al. found that pitch, which is conducive to humming-based music retrieval, is not conducive to human verification and identification (as the pitch in humming is highly dependent on the melody and not on the target speaker). They also reported better performance of LP-based features [4]. In this paper, we propose to use a new feature set based on variable length Teager energy operator (VTEO) and compare it with baseline MFCC. The novelty of the approach lies in exploiting VTEO to capture airflow properties in vocal tract which is key to the excitation for oral and nasal cavity. To the best of the authors’ knowledge this is first work that combines VTEO and MFCC for humming-based biometric problem.

2. VARIABLE LENGTH TEAGER ENERGY BASED MFCC (VTMFCF)

According to Teager [8], the airflow is not propagated in the vocal tract as a linear planar wave. Instead, separate and concomitant vortices are distributed throughout the vocal tract during phonation and the true source of sound production is actually the vortex-flow interactions, for which non-linear model has been suggested based on the energy of airflow. Modeling the time-varying vortex flow is a formidable task and Teager devised a simple algorithm which uses a non-linear energy-tracking operator referred to as Teager Energy Operator (TEO) in discrete-time) for signal analysis (and to characterize airflow properties) with the supporting observation that hearing is the process of detecting energy. The concept was further extended to continuous-domain by Kaiser [9]. According to Kaiser, energy in a speech frame is a function of both amplitude and frequency. By using the dynamics of Simple Harmonic Motion (S.H.M.), he developed TEO for discrete-time signal, as

$$TEO\{x(n)\} = x^2(n) - x(n+1)x(n-1)$$

(1)

From (1) it is evident that TEO of a signal involves non-linear operations (e.g., squaring) on the signal. TEO algorithm gives good running estimate of the signal energy when signal has sharp transitions in the time-domain. However, when the amplitude difference between two consecutive samples of the signal is minute, then the TEO will give zero energy output which indicates that energy required to generate such sequence of samples is zero but that may not be the case in actual physical signal (e.g., speech or hum). To alleviate this problem VTEO was proposed recently and is very briefly discussed here [10]. The dynamics and solution (which is an S.H.M.) of mass-spring system are described by

$$\frac{d^2 x}{dt^2} + \frac{k}{m} x = 0 \Rightarrow x(t) = A \cos(\omega t + \phi)$$

and the energy is given by

$$E = \frac{1}{2} m \omega^2 A^2 \Rightarrow E = (\omega A)^2$$

(2)

From (2), it is clear that the energy of the S.H.M. of displacement signal $x(t)$ is directly proportional not only to the square of the amplitude of the signal but also to the square of the frequency of the signal. Kaiser and Teager proposed an algorithm to calculate the running estimate of the energy content in the signal. Now $x(t)$ can be expressed in discrete-time domain as

$$x(n) = A \cos(\omega n + \phi)$$

where $A$, $\omega$ and $\phi$ are the amplitude, digital frequency (rad/sec) and phase of sinusoidal signal of S.H.M. We have

$$x(n \pm i) = A \cos(\omega(n \pm i) + \phi); \quad i > n$$

$$\Rightarrow TEO\{x(n)\} = x^2(n) - x(n + i)x(n - i) = A^2 \sin^2(\omega i) = A^2 \omega^2 = E_n; \quad (3)$$

$$\Rightarrow TEO \rightarrow TEO \quad as \quad i \rightarrow 1$$

where $TEO\{x(n)\}$ gives the running estimate of signal’s energy and we refer it as for variable length TEO for the dependency index (DI) $i$ which is expected to give running estimate of signal’s energy after considering past $i^{th}$ and future $i^{th}$ sample to track the dependency in the sequence of samples of speech signal.

Traditional MFCC-based feature extraction involves pre-processing: Mel-spectrum of pre-processed speech, followed by log-compression of subband energies, and finally DCT to get MFCC per frame [11]. In our approach, we employ VTEO for calculating the energy of speech signal. Now, one may apply VTEO in frequency domain, i.e., VTEO of each subband at the output of Mel-filterbank, but there is difficulty from implementation point of view. Let us discuss this point in detail. In frequency-domain, (3) for pre-processed speech $x_p(n)$ implies,

$$F\{ TEO(x_p(n))\} = F\{ x_p^2(n)\} - F\{ x_{p(n+i)}x_{p(n-i)}\} \quad (4)$$

Using shifting and multiplication property of Fourier transform, we have

$$F\{ x_{p(n+i)}x_{p(n-i)}\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} X_{ip}(\theta)X_{ip}(\omega - \theta) d\theta$$

where $X_{ip}(\omega) = e^{-j\omega t}X_{p}(\omega)$ and $X_{2p}(\omega) = e^{j\omega t}X_{p}(\omega)$. Hence (4) becomes

$$F\{ TEO(x_p(n))\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[1 - e^{-j\omega t}e^{j\omega t} \right] X_{p}(\theta)X_{p}(\omega - \theta) d\theta \quad (5)$$

Thus (5) is difficult to implement in discrete-time and is also time-consuming. So we have applied VTEO in the time-
domain. Let us now see the computational details of VTMFCC. Speech signal $x(n)$ is first passed through pre-processing stage (which includes frame blocking, Hamming windowing and pre-emphasis) to give pre-processed speech signal $x_p(n)$. Next we calculate the VTEO of $x_p(n)$.

$$VTEO[x_p(n)] = x_p^2(n) - x_p(n+i)x_p(n-i) = \psi(n)$$

The magnitude spectrum of the VTEO output is computed and warped to Mel frequency scale followed by usual log and DCT computation (of MFCC) to obtain VTMFCC as:

$$VTMFCC = \sum_{i=1}^{L} \log[\psi(i)] \cos \left( \frac{kL-0.5}{L} \pi \right), k = 1, 2, ..., N_c.$$

where $\psi(i)$ is the filterbank output of $DFT[\psi(n)]$ and $\log[\psi(i)]$ is the log of filterbank output and $VTMFCC(k)$ is the $k^{th}$ VTMFCC similar to T-MFCC for DI=1 [12]. The proposed feature set, viz., VTMFCC, differs from the traditional MFCC in the definition of energy measure, i.e., MFCC employs $\ell^2$ energy in frequency domain (due to Parseval’s equivalence) at each sub-band whereas VTMFCC employs variable length Teager energy in time domain (here term variable is referred for DI across different recognition experiments). Fig. 3 shows the functional block diagram of VTMFCC. It is clear that output of Mel filterbank in VTMFCC is not squared further in frequency-domain.

$$x_p(n) \rightarrow VTEO \rightarrow \text{Mel-Spectrum} \rightarrow \log(.) \rightarrow \text{DCT} \rightarrow \text{VTMFCC}$$

![Fig. 3. Block diagram for proposed VTMFCC Implementation](image)

### 3. EXPERIMENTAL SETUP

The database is prepared from 51 subjects in the radio room of DA-IICT Gandhinagar (India). Subjects were asked to hum for 20 most popular songs of the legendary singer late Kishore Kumar and Lata Mangeshkar (famous singers in Hindi cinema) out of which 4 hums from each subject were kept for testing and the remaining hums were kept for machine training. Table I shows the details of corpus.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>DETAILS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of speakers</td>
<td>51 (35 male and 13 female)</td>
</tr>
<tr>
<td>No. of hums per speaker</td>
<td>20</td>
</tr>
<tr>
<td>No. of sessions</td>
<td>1</td>
</tr>
<tr>
<td>Data type</td>
<td>Hum for a Hindi song</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>44100 Hz down sampled to 22050 Hz</td>
</tr>
<tr>
<td>Sampling format</td>
<td>1-channel, 16-bit resolution</td>
</tr>
<tr>
<td>Type of hum</td>
<td>Humming of 20 songs (11 Kishore Kumar songs and 9 Lata Mangeshkar Songs)</td>
</tr>
<tr>
<td>Training segments</td>
<td>30 s, 60 s</td>
</tr>
<tr>
<td>Test segments</td>
<td>1 s, 1.25s, 1.5s, 1.75s, ..., 15s (57 test segments)</td>
</tr>
<tr>
<td>Microphone</td>
<td>SCHURE microphone</td>
</tr>
<tr>
<td>Genuine Trails</td>
<td>51 (per test segment) x57=2907</td>
</tr>
<tr>
<td>Impostor Trials</td>
<td>51x51x57-2907=145350</td>
</tr>
</tbody>
</table>

### 4. EXPERIMENTS

Feature analysis was performed using $12^{th}$ order LPC on a 23.2 ms frame with an overlap of 50%. Each frame was pre-emphasized with the filter $1-0.97z^{-1}$, followed by Hamming window and then the mean value is subtracted from each speech frame. Similar pre-processing steps were performed for MFCC [11]. Polynomial classifiers of $2^{nd}$ order approximation are used as basis for all the experiments and have advantage of using out-of-class data to optimize the performance (as opposed to other statistical methods such as HMM or GMM) [13]. Results are shown for person identification and verification experiments.

#### A. Person Identification

In this work, % identification (ID) rate is defined as

$$SR = \frac{N_c}{N_t} \times 100,$$

where $N_t$ the number of correctly identified speakers and $N_c$ is the total number of speakers used for machine learning. The results are shown for different feature sets (FS) in Tables II as average % ID rate (computed using 2907+145350=151164 trials) with $2^{nd}$ order polynomial approximation. To state the statistical significance of our results, we have also included confidence intervals denoted as CNFINT (with a confidence of 95%) in brackets [14]. It is evident from Table II that proposed feature set VTMFCC outperforms LPCC and competitive over MFCC. Optimal dependency index DI=9 is chosen after running the computer simulations for DI=1 to 12 and hence presently selected based on best recognition performance.

<table>
<thead>
<tr>
<th>FS</th>
<th>LPCC</th>
<th>MFCC</th>
<th>VTMFCC(DI=9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% ID</td>
<td>66.40</td>
<td>79.36</td>
<td>80.01</td>
</tr>
<tr>
<td>95%</td>
<td>(65.18, 67.61)</td>
<td>(78.31, 80.40)</td>
<td>(81.07, 83.04)</td>
</tr>
</tbody>
</table>

#### B. Person Verification

For person verification, we consider equal error rate (EER), which is the point at which the false acceptance (FA) and false rejection (FR) rates are equal on Detection Error Tradeoff (DET) curve [15]. Fig. 4 shows DET curve for MFCC, VTMFCC (DI=9) and combined (MFCC+VTMFCC). It is evident from DET curves that the proposed feature set performs better than baseline MFCC at many operating points of DET curve. In addition, when scores from MFCC and VTMFCC are fused (with equal weight factor, i.e., $s = \alpha s_M + (1-\alpha) s_V$), where $\alpha = 0.5$ and $s_M$, $s_V$ and $s$ are the matching scores for MFCC,
VTMFCC and fused system, respectively), then the combined system outperforms baseline MFCC at all the points of the DET curves. This indicates that proposed feature set captures the complementary information hidden in the hum signal. For speaker detection task, the performance measure used is the optimal detection cost function (DCF) which is a weighted sum of FR and FA probabilities [15]. Results are reported in Table III. It is evident that EER for proposed VTMFCC is reduced by 0.35% as compared to baseline MFCC. The reduction is significant (by 1.73 %) when scores from the proposed feature set are fused with that from baseline MFCC. Similar performance improvement is obtained for optimal detection cost function (DCF) function as well. The combined system performs better than MFCC alone. This may be due to the fact that VTEO profile of hum contains perceptual information about the signal. This was verified by sound playback experiment by creating a sound file from VTEO information about the signal. This was verified by sound playback experiment by creating a sound file from VTEO information about the signal. Significant perceptual correlation in hearing was observed between the original signal and its corresponding VTEO profile.

<table>
<thead>
<tr>
<th></th>
<th>FS</th>
<th>MFCC</th>
<th>VTMFCC(DI=9)</th>
<th>FUSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER (%)</td>
<td>14.25</td>
<td>13.89</td>
<td>12.52</td>
<td></td>
</tr>
<tr>
<td>Opt. DCFx10^-2</td>
<td>14.08</td>
<td>13.82</td>
<td>12.35</td>
<td></td>
</tr>
</tbody>
</table>

### 5. SUMMARY AND CONCLUSION

A novel approach using VTEO (to capture airflow properties in vocal tract) and MFCC is proposed for humming-based biometric problem. Score-level fusion of VTMFCC and MFCC gave better performance. One limitation of the proposed method could be to find the optimal dependency index (DI) for a particular database. Future work will be directed towards evaluation of proposed method under noisy conditions.

### 6. REFERENCES


