Combining Evidence from Spectral and Source-like Features for Person Recognition from Humming

Hemant A. Patil1, Madhavi C. Madhavi1 and Keshab K. Parhi2

1Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, INDIA
2Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA

{hemant_patil, madhavi_maulik}@daiict.ac.in, parhi@umn.edu

Abstract

In this paper, hum of a person is used in voice biometric system. In addition, recently proposed feature set, i.e., Variable length Teager Energy Based Mel Frequency Cepstral Coefficients (VTMFCC), is used to capture perceptually meaningful source-like information from hum signal. For person recognition, MFCC gives EER of 13.14% and %ID of 64.96%. A reduction in equal error rate (EER) by 0.2% and improvement in identification rate by 7.3 % is achieved when a score-level fusion system is employed by combining evidence from MFCC (system) and VTMFCC (source-like features) than MFCC alone. Results are reported for various feature dimensions and population sizes.

Index Terms— Humming, VTEO, VTMFCC, fusion of Source-System features, polynomial classifier

1. INTRODUCTION

Speaker recognition refers to the task of identification of a person from his or her voice with the help of machines. In this paper, an attempt is made to recognize a person using his or her hum. A hum is produced by closing mouth cavity, so that the vocal tract is coupled with nasal cavity only. This kind of voice biometric system may be applicable to a person with speech disorder and an infant, who is not able to speak [2],[3]. Such kind of biometric pattern (i.e., hum as well) is universally available [4],[5] and it may be useful for speaker forensics [6].

The motivation to use hum produced by the speakers for voice biometrics problem is the following. While producing hum oral cavity is closed and thus the vocal tract is coupled with the nasal cavity. Hence, the hum produced by the speakers is nasalized sounds. The nasal cavity remains steady during hum production and is known to be speaker-specific [6]. The experiments done on nasal sound by Amino et. al. shows that nasal sounds have more inter-speaker variability and less intra-speaker variability. Moreover, humming sounds are vowels-like quasi-periodic sounds. Hence, speaker-specific features derived from hum can be useful for design of speaker-dependent query-by-humming (QBH) system which can be an important utility or a part of music information retrieval (MIR) applications [7]. In this context, authors have reported use of the spectral features for humming-based speaker recognition where performance of MFCC was found to be better than LPC and LPCC [8]. 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 [5]. Recently, Variable length Teager Energy Based Mel Frequency Cepstral Coefficients (VTMFCC) was proposed for this problem [1]. The novelty of this approach lies in exploiting variable length Teager Energy Operator (VTEO) to capture airflow properties in vocal tract which is key to the excitation for both oral tract and nasal cavity. This work was reported on database of 51 subjects. In this paper, this work is extended on large database of 170 subjects. In addition, in this paper, VTMFCC is shown to capture perceptually meaningful speech (or hum) source characteristics which could be useful for person recognition. Furthermore, performance of VTMFCC is evaluated for different feature dimensions and population sizes.

2. VARIABLE LENGTH TEAGER ENERGY BASED MFCC (VTMFCC)

A non-linear energy-tracking operator referred to as Teager Energy Operator (TEO) for discrete-time signal \( x(n) \), is defined as

\[
\text{TEO}_i x(n) = x^2(n) - x(n+1)x(n-1) \tag{1}
\]

TEO algorithm gives good running estimate of the signal energy when signal has sharp transitions in the time-domain. However, in situations where the amplitude difference between two consecutive samples of the signal is very small, 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 [10]. VTEO of discrete-time signal \( x(n) \) for dependency index (DI) \( i \) is defined as

\[
\text{VTEO}_i x(n) = x^2(n) - x(n+i)x(n-i) \tag{2}
\]

where \( \text{VTEO}_i x(n) \) 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/hum signal. Next, computation details of VTMFCC are explained in brief.

Traditional MFCC-based feature extraction involves pre-processing; Mel-spectrum of pre-processed hum, followed by log-compression of subband energies, and finally DCT to get MFCCs per frame [12]. In our approach, we employ VTEO for calculating the energy of hum signal. In VTMFCC the energy is calculated in time-domain instead of subband energy in frequency-domain. Hum signal \( x(n) \) is first passed through pre-processing stage (which includes frame blocking means to break sequence of samples in small frames of 20-30 ms, Hamming windowing and pre-emphasis) to give pre-processed hum signal \( x_p(n) \). Next, we calculate the VTEO of \( x_p(n) \)

\[
\text{VTEO}_i x_p(n) = x_p^2(n) - x_p(n+i)x_p(n-i) \tag{3}
\]
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:

$$\text{VTMFCC}(k) = \sum_{l=1}^{L} \frac{\log[\Psi(i)]}{l} \cos \left( \frac{k(l-0.5)\pi}{L} \right), \quad k = 1, 2, \ldots, N_c,$$

where $N_c$ is the number of VTMFCC per frame, $\psi(i)$ is the filterbank output of $\text{DFT}(\psi(i))$ and $\log[\psi(i)]$ is the log of filterbank output and $\text{VTMFCC}(k)$ is the $k$th VTMFCC. The proposed feature set, viz., VTMFCC, differs from the traditional MFCC in the definition of energy measure, i.e., MFCC employs $L^2$ energy in frequency domain (due to Parseval’s energy equivalence) at each sub-band whereas VTMFCC employs variable length Teager energy in time domain (here term variable is referred for DI across different person recognition experiments). Figure 1 shows the functional block diagram of VTMFCC [1].

![Figure 1: Block diagram for proposed VTMFCC implementation. After[1].](image)

2.1. VTMFCC as perceptually meaningful source information

Let us consider speech vowel /a/ and its VTEO for DI=4. This dependency index is taken due to the fact that better person recognition performance is achieved for this value which will be discussed in section 4. The Figure 2 shows speech signal and hum signal produce by the same speaker and their corresponding VTEO profiles for DI=4.

![Figure 2: (a) Speech signal for vowel /a/ (b) its VTEO profile (c) hum waveform produced (by the same speaker as in Figure 2(a)), (d) its VTEO profile.](image)

As vowels are quasi-periodic in nature, it is evident from Figure 2(b) and 2(d) that VTEO profile gives sharp pulses at regular interval which corresponds (in close agreement) to location of Glottal Closure Instance (GCI) (evident from quasi-periodic nature of vowel /a/ in Figure 2 (a)). It is interesting to note that peaks in Figure 2(d) are more dominant and clearer than the peaks in Figure 2(b). In particular, there is a presence of dominant bumps around the peaks in Figure 2(b) than Figure 2(d). This may be due to the fact that hum is produced by nasal cavities. During production of a humming signal nasal cavities are stationary. Whereas for the production of normal speech, dynamic use pattern of articulators may introduce nonlinearities (via vortices along vocal tract) which may result in bumps in VTEO profile (Figure 2(b)).

3. EXPERIMENTAL SETUP

The database is prepared from 170 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 1 shows the details of corpus.

<table>
<thead>
<tr>
<th>Item</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of speakers</td>
<td>170 (137 male and 33 female)</td>
</tr>
<tr>
<td>Data type</td>
<td>Hum for a Hindi song</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>22050 Hz</td>
</tr>
<tr>
<td>Sampling format</td>
<td>1-channel, 16-bit resolution</td>
</tr>
<tr>
<td>Dimension of feature vector</td>
<td>12</td>
</tr>
<tr>
<td>Training segments</td>
<td>30 s, 60 s, 90 s</td>
</tr>
<tr>
<td>Test segments</td>
<td>1 s, 2 s, 3 s, 4 s, …30 s (30 testing segments)</td>
</tr>
<tr>
<td>Genuine Trials</td>
<td>30 (per testing segment) x3 (per training segment) x 170(no. of speakers) = 15300</td>
</tr>
<tr>
<td>Impostor Trials</td>
<td>(170x170x30x3) = 2585700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 1. Details of Humming Database for Person Recognition</th>
</tr>
</thead>
</table>

4. EXPERIMENTS

Feature extraction was carried out on frames of 23.2 ms duration with an overlap of 50%. Each frame was pre-emphasized with the filter $1-0.97z^{-1}$, followed by Hamming windowing. Polynomial classifiers of 2nd 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). The feature vectors are processed by the polynomial discriminant function. During recognition, the score for each test segment is computed as inner product between the polynomial expansion of feature vectors of test segment and speaker model for each hypothesized person. For person identification, the test segment is assigned to the speaker whose score is the maximum whereas for person verification, if a score for speaker is higher than a threshold, the claimant is accepted; otherwise it is rejected [13].

4.1. Person Identification

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

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

where $N_c$ the number of correctly identified person and $N_t$ is the total number of persons used for machine learning.

<table>
<thead>
<tr>
<th>Table 2. Average %ID Rates for 2nd Order Polynomial Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>95 % CNFINT</td>
</tr>
</tbody>
</table>
The results are shown in Table 2 (for different feature sets) as average % ID rate (computed using $15300 + 2585700 = 2601000$ 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 2 that proposed feature set VTMFCC alone is not performing well compared to MFCC. However, the score-level fusion of VTMFCC with MFCC outperforms MFCC by 7.3%. In particular, scores from MFCC and VTMFCC are fused as 

$$s = \alpha s_M + (1 - \alpha) s_F$$

where $\alpha = 0.75$ and $s_M$, $s_F$ and $s$ are the matching scores for MFCC, VTMFCC and fused system, respectively. This may be due to the fact that VTMFCC captures complementary information (in particular, perceptually meaningful source-like information in some sense). In this experiment, $\alpha = 0.75$ is chosen because it was observed that by running computer simulation by varying $\alpha$ from 0 to 1. Figure 3 shows plot of % person identification (success rate) vs. dependency index (DI). From Figure 3, it is evident that DI=4 gives relatively best person recognition performance. Thus, selection of DI in this work is optimized with respect to the recognition performance.

### 4.2. 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]. Figure 4 shows plot of EER vs. dependency index (DI).

From Figure 4 it is evident that DI=3 and 4 give relatively best person recognition performance. In this work, DI=4 is selected as it gives good results for person identification as well. Figure 5 shows DET curve for MFCC, VTMFCC (for DI=4) and score-level fusion of MFCC and VTMFCC. It is evident from DET curves that the score-level fusion of MFCC and VTMFCC performs better than baseline MFCC at all operating points of the DET curve.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>EER (%)</th>
<th>Opt. DCFx10^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td>MFCC</td>
<td>VTMFCC(DI=4)</td>
</tr>
<tr>
<td>EER</td>
<td>13.14</td>
<td>14.48</td>
</tr>
<tr>
<td>Opt. DCFx10^2</td>
<td>13.05</td>
<td>14.35</td>
</tr>
</tbody>
</table>

### 4.3. Effect of Population Size

The experiment was conducted for different population sizes, viz., 10, 20 … to 170 subjects. A selection of person is made in sequence. It means that a result for 10 subjects is computed from the first 10 subject and so on. The results for person verification and identification for various population size are shown in Figure 6(a) and 6(b), respectively. From Figure 6, it is evident that score-level fusion of VTMFCC gives better result as compared to baseline MFCC. Similar performance improvement is obtained for optimal detection cost function (DCF) function as well. Here, for calculation of DCF, costs associated with FA and FR are taken as 1 and prior probabilities of genuine and impostor trials are assumed to be equal, i.e., 0.5. The combined feature performs better than MFCC alone. Again, this indicates that proposed feature captures complementary information hidden in the sequence of hum signal.

### 4.4. Effect of Feature Dimension

The experiments were conducted to investigate effect of feature dimension on the performance of person recognition system. The results for person verification and identification for various feature dimension are shown in Figure 7(a) and 7(b), respectively.

![Figure 3: Success rate (% ID) vs. dependency index.](image-url)

![Figure 4: Equal error rate (EER) vs. dependency index (DI)](image-url)

![Figure 5: Speaker Detection Performance of MFCC, VTMFCC and their score-level fusion of them](image-url)

![Figure 6(a): Success rate (% ID) vs. population size.](image-url)

![Figure 6(b): Equal error rate (EER) vs. population size.](image-url)

![Figure 7(a): DET curve for MFCC, VTMFCC and their score-level fusion of them](image-url)

![Figure 7(b): DET curve for MFCC, VTMFCC and their score-level fusion of them](image-url)
From Figure 7, it is again evident that score-level fusion of MFCC and VTMFCC gives better performance for almost all the cases of feature dimensions. In addition, results are better for higher feature dimension. This may be due to the fact that as the feature dimension increases the feature occupancy in higher-dimensional feature space decreases and thereby separating different patterns more efficiently. However, one has to consider the computational cost associated with extracting high-dimensional feature vector and preparation of class-specific model. Higher dimension of feature vector leads to more time and space complexity.

5. SUMMARY AND CONCLUSION

In this paper, recently proposed feature set, viz., VTMFCC is used for different person recognition experiments. In addition, VTMFCC is analyzed to indicate perceptually meaningful source information which is complementary to the system features like MFCC. It is observed that the score-level fusion of MFCC and VTMFCC performs well for person recognition task. This observation was also validated on various population sizes and various feature dimensions. One of limitations of the VTMFCC feature could be to find the optimal dependency index (DI) for a particular database, i.e. DI is optimized with respect to its performance. Future work will be directed towards evaluation of proposed feature set under degraded (noisy) conditions.

REFERENCES