Efficient Speaker Verification System Based on Heart Sound and Speech


Abstract—This paper proposed an integration of heart sound and speech for biometrics application. The method the method selects the best fusion and normalization techniques for biometric system. The framework is developed and test the verification task. The approach in this paper is biometrics recognition, for example, providing features that can’t be easily copied, such as the Mel-Frequency Cepstral Coefficient (MFCC) as a feature vector and vector quantization (VQ) as the matching model algorithm. A simple yet highly reliable method is introduced for biometric applications. Experimental results show that the recognition rate of the Heart sound-speaker verification (HS-SV) provides an average EER of 17.8% while the average EER for the speech speaker verification model (S-SV) is 3.39%. In order to reach a higher security level an alternative to the above approach, which is based on multimodal and a fusion technique, is implemented into the system. The best performance of the work is based on simple-sum score fusion with a pricewise-linear normalization technique which provides an EER of 0.69% while the fusion type Main Rule provide an EER of 1.1%.

Keywords—Fusion, Speaker Verification, Vector Quantization.

I. INTRODUCTION

SIMPLE fusion of two leads to a wider separation between the user clusters in the combined feature space. However, improper fusion can lead to designs which require extra training and extra storage hence increased complexity. In addition, recognition based on various biometrics such as fingerprints, speech, face, signature, ECG and heart sound used feature extraction (LPC, MFCC, CEPSTRUM) and classification (VQ, NN, GMM and HMM) in the design of the biometrics system. In this paper, we extend this framework to the application of speech and heart sound biometrics. The objective of this work is to design a speaker recognition (SR) system using VQ trained with possibly limited training data. Researchers have used neural network [1]-[2]. An speaker verification (SV) task using large amounts of labeled speech data is not only time consuming but also requires a huge amount of training data for better generalization of neural network. Generalization will depend on the empirical optimization of the Neural Networks (NN) structure and the training methods. On the other hand, HMM, [3]-[4] has been shown to perform significant better than the VQ based or the NN approach for SV tasks. The benefit however must be weighed against the high increased computational requirements using HMM during training and recognition as well as increased computation as more features are added into the processing stage. Thus, the design of a Speaker Recognition (SR) system using VQ with this specific objective in mind should follow the criteria mentioned below to achieve the end results.

- Use of simple algorithm for ease of implementation and without increases in computational effort.
- Capable of achieving high performance with limited training data.
- Reasonable handling of the temporal information.
- To maximize the probability of making correct fusion and normalization decisions.

This paper is thus structured as follows: Section 2 provides the speaker recognition framework, and type of database used in the experiments. In Section 3, the feature extraction and vector quantization method is discussed. Section 4 presents the experimental results and discussions. This section also introduces the decision level scheme for multimodal fusion, and finally section 5 concludes the paper.

II. DESIGN OF SPEAKER VERIFICATION BASED ON HEART SOUND AND SPEECH

Fig.1 Basic Structure of Automatic Speaker Verification based on Heart/Speech.
\[ \text{Mel}(f) = 2595 \log \left( 1 + \frac{f}{700} \right) \]  

\( f \) is the frequency in Hz.

**A. Verification System**

The work carried out here makes use of signal processing theory, computational technologies, electronic recording stethoscopes, and sensors to revitalize the traditional approach of biometrics authentication. Fig. 1 shows the speakers Verification (SV) process to determine the identity of the test speaker speakers using specific information retained in the speech or the heart signal. Establishing whether a speakers claimed identity is correct by automatic means based on the acoustic of his/her voice/heart sound is formed as automatic speaker verification (SV).

**B. Data Base**

**Heart Sound:** The database consists of cycles of heart sounds from a large number of subjects. There are 20 clients and 40 imposters. The database is divided into training and test data. A group of 20 subjects are modeled by the system the data are all end point detected with ECG segmentation. The frame size was 0.025 ms with 0.010 ms overlap, and about half of the heart sound cycles are used as training data and the remaining half for testing with the client subjects. Roughly 4700 cycles of heart sounds of imposter subjects were used for testing.

**Speech:** The same set of clients and imposters are used in speaker recognition. The performance of VQ varies strongly with the amount of tokens available. The data base used to evaluate the system is trained with 5 Training tokens of each digit for the construction of the codebook with 27 clients. The system was tested with 10 true client tokens and 70 imposter tokens for each digit for each speaker. The database consists of isolated digits from a large number of speakers. Ten isolated digits (digits ‘Sefer-0’ to ‘tesah-9’) are used in the experiment. The data are all end point detected to remove excess silence and to minimize storage. The frame sizes are 20 ms with 15 ms overlap.

**III. Feature Extraction and Vector Quantization**

A representation of heart sound or speech would be provided by a set of cestrum coefficients. Mel-frequency cepstrum coefficients (MFCC) are used in this paper as the feature representation of the heart signal. Previously the MFCC was mainly used in the field of speech processing i.e. for speech/speaker recognition application and has delivered excellent results due to its robustness under various conditions [5]-[6]. As heart sound and speech are both acoustic signals, it is a reasonable to use MFCC in heart sound recognition or identification tasks. The heart signals are pre-emphasized to spectrally flatten the signals. After frame blocking, the signals were hamming windowed to minimize the spectral distortion. MFCC coefficients were then calculated by taking a discrete cosine transform (DCT) of the logarithmic spectrum scale after it was warped to the Mel scale [7].

**IV. Evaluation of Speaker Verification Based on Speech and Heart Sounds**

**A. Speech Speaker Verification (S-SV)**

Assessing the effects of codebook size begins with codebook sizes of 16, 32, 64, 128 and 256. All experiments showed that no improvement could be seen as the codebook size increased, and this is probably because of the limited size of trainings data used to develop the codebook. If more training tokens were used, a better EER performance would be expected over the entire range of codebook sizes before settling on a fixed point.

Here 20 client speakers were chosen as a set of client speakers and were enrolled in the speaker verification (SV) system. The SV system is evaluated on single isolated digits. The performance of each client is listed in TABLE I with EER varying from 0.2% to 10.9% which is the EER for all the digits. The overall average EER for the client is 3.39%.

**B. Heart Sound Speaker Verification (HS-SV)**

In this section the performance of the HS-SV model with 10 clients is compared with the S-SV model which is shown in Fig. 4. The verification scores based on the output of this model is used by applying the EER threshold to make a decision to accept or reject a client. As with the previous experiment, the threshold is speaker specific.
The performance of each 20 client is shown in Fig.4 and Fig.5. From the figures it can be seen that there are considerable variations in performance across the clients. The HS-SV model has an average EER of 17.8% with a range of 0.6% to 32%. The S-SV model on the other hand Fig.3 shows an average EER of 3.9% with a range of 0.2% to 10.9%. It was expected that the S-SV model would perform better than the HS-SV model. The heart sound is based on the lub and dub sounds only. Variations of this lub and dub sound would depend on whether the person is fat or thin, young or old, sick or healthy and the auscultation sites may vary according to the patient’s anatomy. While each digit displays different performance in isolation, each digit emphasizes a different aspect of time. Varying speech signal and rankings of the digits may vary from client to client.

C. Multimodal Score Fusion of Speech Output Score and Heart Sound

One important property of VQ output scores, which is useful to ASV task, is the ability to represent the statistical properties of data. In this work, different biometrics of the same person are acquired and combined to complete and improve the recognition process. According to [9], there are several levels at which fusion can take place such as,

- Sensor level
- Match score level

In this paper, the match score level as seen in TABLE I is implemented for each individual biometrics process. The fusion process fuses the speech and heart sound Euclidean distance values into a single score, which is then compared to the system acceptance threshold. The next step is the score normalization stage, as it is critical in the design for matching score -fusion level. The output score from the speech and the heart are of different numerical ranges. The work applied the simple sum ($\sum$) method to normalized scores of heart and speech biometrics. The normalized scores were obtained by using the following techniques: min-max, Z-score, median-MAD, Double sigmoid, Tanh and piecewise-linear. The results from the analysis are shown in TABLE I.

It can be seen from the table that certain combinations, for example fusion type Simple Sum combine with piecewise -linear, provide the best result with an average equal error rate of less than 0.69%. The pice-wise linear normalization and simple sum fusion gave less than 1% EER for 7 of its client subjects followed by max-min, Z-score and Tanh which had 4 of the subjects with the same EER. The worst performance came from a combination of double segment and median -MAD in the double sigmoid, where there were 3 subjects which did not show resilience to the errors in the estimates of the densities.

### TABLE I

<table>
<thead>
<tr>
<th>Fusion Type</th>
<th>EER before Fusion</th>
<th>Normalization Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client 1</td>
<td>4.8</td>
<td>2.5</td>
</tr>
<tr>
<td>Client 2</td>
<td>3.4</td>
<td>2.7</td>
</tr>
<tr>
<td>Client 3</td>
<td>10.0</td>
<td>2.9</td>
</tr>
<tr>
<td>Client 4</td>
<td>0.5</td>
<td>1.1</td>
</tr>
<tr>
<td>Client 5</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Client 6</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Client 7</td>
<td>2.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Client 8</td>
<td>1.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Client 9</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Average</td>
<td>2.6</td>
<td>2.3</td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th>Fusion Type</th>
<th>Normalization Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client 1</td>
<td>4.8</td>
</tr>
<tr>
<td>Client 2</td>
<td>3.4</td>
</tr>
<tr>
<td>Client 3</td>
<td>10.0</td>
</tr>
<tr>
<td>Client 4</td>
<td>0.5</td>
</tr>
<tr>
<td>Client 5</td>
<td>0.4</td>
</tr>
<tr>
<td>Client 6</td>
<td>0.2</td>
</tr>
<tr>
<td>Client 7</td>
<td>2.1</td>
</tr>
<tr>
<td>Client 8</td>
<td>1.1</td>
</tr>
<tr>
<td>Client 9</td>
<td>1.0</td>
</tr>
<tr>
<td>Average</td>
<td>2.5</td>
</tr>
</tbody>
</table>

- Rank level.
- Decision level.
- Feature level.
median -MAD, on the other hand, shows that all of the subjects had an EER of more than 3%.

![Fig. 5 Shows comparison of fusion type of Simple Sum and Main rule.](image)

The speaker acceptance or rejection decision in Fig. 5 is carried out by comparing the results with fusion type based on the Main Rule. Speaker specific (SS) threshold were used to evaluate the SV system. The threshold are determined by the EER criteria. The use of EER in both type provide a standard set of measurement that details the performance of SV system. The figure shows the performance improvement fusion type Simple Sum over the Main Rule. It can be seen that there are large differences in the distribution of error between this two fusion types.

V. CONCLUSION

Traditional unimodal biometrics (Speech, Fingerprint, Face, Signature, Iris) indicate the high potential of the proposed methodology. However the performance of the reliability estimation in a multimodal has not been widely studied or evaluated. This paper presents a methodology of reliable estimation in the multimodal biometrics verification scenario. A reliable fusion strategy has shown to be an efficient and reliable way of predicting and correcting erroneous classification decisions in multimodal (speech and heart) systems. The use of Main Rule with Pricewise linear provide an EER of 1.1% In this work the advantages of implementing unimodal in order to perform an efficient biometric fusion score of two modalities is clearly shown to be superior with the use of simple -sum score fusion and pricewise-linear normalization technique. There is an improvement of 96.1% compared to the HS-SV model and 79.9% compared to the S-SV model.

ACKNOWLEDGEMENT

This research project is supported by Center of bio-medical Engineering (CBE) University Technology Malaysia university research grant (GUP-flagship Q.J130000.2436.00G31) and MOHE

REFERENCES


