PSO Based Optimized Reliability for Robust Multimodal Speaker Identification

Md. Tariquzzaman, Jin Young Kim, Seung You Na
School of Electronics and Computer Engineering
Chonnam National University
300 Yongbong-Dong, Buk-gu, Gwangju 500-757, South Korea
tareq@moiza.chonnam.ac.kr; beyondi@jnu.ac.kr; syna@jnu.ac.kr

Abstract: - Speaker recognition in real environment with reliable mode is a key challenge for ubiquitous service in human computer interface. In this paper, we present a robust multimodal speaker identification system with optimized reliability of different modalities. We propose an extension of modified convection function's optimizing factors to account optimum reliability simultaneously in audio, face and lip information. The proposed reliability measure is applied to a multimodal speaker identification framework for robust speaker identification. Particle swarm optimization (PSO) algorithm has been employed to optimize the modified convection function's optimizing factors. In the face-based expert, the image quality has been degraded with jpeg compression technique in enrollment and test session. Similarly, Lip-based expert's image quality also degraded to create mismatch in enrollment and test image. Finally, an artificial illumination in opposite direction has been added to test face and lip image with different intensities, respectively. The VidTimit audio DB was collected in office environment has a high level of signal distortion. We have applied local principal component analysis (Local PCA) to both face and lip modalities for reducing the dimension of features vector. The overall speaker identification experiments are performed using VidTimit DB. Experimental results show that our proposed optimum reliability measures effectively enhanced the identification rate (IR) of 8.67% in comparison with the best classifier system i.e., audio classifier and most notably retained the consistency of multimodal integration framework.

Key-Words: - Speaker identification, Human Computer Interface, Particle Swarm Optimization, local PCA, Optimum Reliability Measures

1 Introduction

In this information era, communication and dealings between people and different organizations are performed through human computer interface; where identification of the user is a vital concern to the organizations. Multimodal Speaker identification, a goal of biometrics technology, are deploying increasingly in the secured organizations. The development of the speaker recognition technology particularly multimodal speaker recognition is still an active area of research which has a range of applications from national security, communication system security, computer security, computer network security, e-commerce to forensic [2][4]. Besides, increasing demands of intelligence interface with system like humanoid robot create a need for automatic speaker recognition. With the current-state-of-the-art multimodal speaker identification system may perform well under controlled environment condition. However, in real environment voice information can easily be exposed of different levels of noises due to the channel distortion of the various transmission medium, its surrounding environments and codec distortion. Meanwhile, the visual signal also can be degraded due to image quality, the distribution of illumination over image, illumination direction change and occlusion. As a result, degradation in the performance of single modality based speaker identification as well as the multimodal speaker identification system is occurred. Nonetheless, multimodal biometrics system with a high degree of careful design structure might lead a better performance in comparison with single modality based speaker identification system. For designing a multimodal biometrics system, a number of issues are focused which are mainly categorized into three; which modality to be fused, where to be fused and how the different modalities information to be fused; the later issue includes the reliability measure. Beside the single modality based speaker identification or verification system [1][4-5][7-8][12][15][21][23], there are different approach existing in the literature[6][9][14] [19][22] for multimodal fusion. Specifically in multimodal fusion, reliability measure
is one of the key issues in fusion process. The ultimate goal of measuring the reliability is to reduce classification error rate and thus to enhance the performance and robustness of a multimodal speaker identification system through maintaining the integration structure of multimodal biometrics system.

In this paper, we will propose an extension of the modified reliability convection function \[20\] to account optimum reliability simultaneously in audio face and lip information through introducing optimizing factors. The proposed reliability measure is applied to a multimodal speaker identification framework for robust speaker identification. Particle swarm optimization (PSO) algorithm has been employed to optimize the modified convection function's optimizing factors. In the face-based expert, the image quality has been degraded with jpeg compression technique in enrollment and test session. Similarly, Lip-based expert's image quality also degraded to create mismatch in enrollment and test image. Finally, an artificial illumination in opposite direction has been added to test face and lip image with different intensities, respectively. The VidTimit audio DB was collected in office environment has a high degree of signal distortion. For individual experts’ speaker model generation, we have implemented the classical GMM \[16\]. The entire experiments are performed using VidTimit DB \[19\].

2. System Architecture

In this section, we will describe the individual classifiers and the overall baseline system.

2.1 Baseline system

A schematic diagram of a bimodal speaker identification system is shown in Figure 1.

Certainly, in a multimodal speaker identification system there are different individual classifiers which are termed as audio-based, face-based and lip-based classifier/expert system. Briefly the overall process can be described as follows:

The speech signal, face/lip image features are fed to its respective expert which is termed as classifier. These classifiers calculate the set of likelihood score. From each set of observation likelihood score, the expert reliability parameters are estimated. The fusion weighting parameters are calculated using basic mapping rule using the reliability parameters of different classifiers. The popular late integration approach is applied for integration. Using the weighting factors and the set of normalized score, finally the set of combined score is obtained through weighted sum rule. In the decision stage, the highest score of likelihood is decided as the identified speaker from the combined score. Typically, The N-score integration rule is applied to integrate all the three experts.

In experiments presented in this paper, MFCC features are extracted from speech signal and CMS \[7\] is performed on these features vector. Before feature extraction and energy based VAD is applied on speech signal. Similarly, DCT features of visual signal are extracted from the individual image and local PCA is applied to reduce the dimension of the features vector in both face and lip based classifier system. We have created the the face DB from the VidTimit DB using AdaBoost algorithm. Manually we have created lip ROI DB from the original VidTimit DB considering the lip center as the base point for ROI determination and thus created the lip database.

In individual classification process, each modality expert generates the log-likelihood score generally termed as \text{S}(O_{mn} | \lambda_n) \text{ where O is either voice or visual information. For a bimodal case, when m= 1, it indicates that it is audio information while at m=2 it is visual information and n of \lambda_n is the n-th speaker of the GMM model. For the multimodal case, as in our experimental system, m=2 and m=3 represent face expert and lip expert, respectively.}

2.2 Reliability measures and audio-visual information integration

Broadly, the reliability measure can be categorized into two i.e., at the signal level or at the expert score level. The score level reliability measure uses the statistical rank of each of the experts. Even there are some methods in literature considering the signal level confidence measure; however, the higher level
confident confidence measure is more desirable. In the literature, there are different methods for score based reliability measure which are mainly score entropy, dispersion, variance, cross classifier coherence and score difference. The score difference [14][22] is our special interest which has been taken as the baseline method in this paper. The baseline method takes the statistical nature of observation probability based on the rank of information and min-max normalized method is performed on all observation probabilities. The entire process for reliability parameters calculation is followed by the following sequences:

In first, the individual modality generates the likelihood scores which are normalized using min-max normalization method differently. Mathematically we can express

\[ S(O_m | \lambda_j) - \text{MinP}_m \]

\[ \text{MaxP}_m - \text{MinP}_m \]  

where \( m \) represent the modality either audio or video expert weighting factor,

Secondly, the different modalities reliability parameters are calculated using the highest rank of the normalized score, i.e.,

\[ \rho_m = \text{MaxP}_m - \text{Max2P}_m \]

Here, \( \text{MaxP}_m \) is the highest rank of the expert score where \( m \) represent the modality either audio or video and \( \text{Max2P}_m \) is the second highest rank of the individual expert score. After normalization \( \text{Max2P}_m \) becomes 1 if there is no preferred range and the above equation can be written as

\[ \rho_m = 1 - \text{Max2P}_m \]

Thirdly, for mapping in between the reliability and the expert weighting factor, the weighting factor is calculated from the reliability function as follows:

\[ a_1 = \frac{\rho_1}{\rho_1 + \rho_2}, \quad a_2 = 1 - a_1 \]

Fourthly, each modality score is integrated using the weighting factor as

\[ S(O_1, O_2 | \lambda_n) = a_1 S(O_1 | \lambda_n) + a_2 S(O_2 | \lambda_n) \]

Then N-score integration rule is applied to integrate all the three experts.

3. Proposed Reliability Measure

From the above described section, we see that integrated score of observation are calculated using equations (2) to (5). Specifically, the reliability value which is expressed by equation (3) is determined from the audio and visual modality information individually and it has an important role to enhance the speaker recognition performance. For each expert, reliability value expressed by the equation (3) which was derived from the normalized observation probabilities. Explicitly, even though the fusion final goal is to maximize the recognition rate; explicitly, even though the convection function’s i.e., \( \rho_m = f(S(O_m | \lambda_1), S(O_m | \lambda_2)......S(O_m | \lambda_n)) \) final goal is recognition rate; Even though optimum reliability measure is an open issue in literature [18][24],indeed, there is no optimization parameter in the convection function that make an optimum reliability and thus maximizes the recognition rate in the speaker identification system. We can think two different possible conditions in the reliability measure regarding reliability function expressed by equation (3).

a. Overestimation: Reliability value is estimated in higher order rather than its optimum level; In this case, we should regulate the reliability function so that the reliability value reaches at the optimum point.

b. Underestimation: Reliability value is estimated poorly from its ground truth; In this case, also we should regulate the reliability function to raise the reliability value. Introducing an optimization factor in reliability function, we can control the reliability function so that an improvement in audio-visual speaker identification could be achieved thus for the ultimate goal of integration to be fulfilled.

Thus, we have modified the convection function expressed by equation (3) through introducing optimization factors on different modalities i.e., introducing the optimization factors let \( f_m \) on \( \text{Max2P}_m \), so that the above two case can be controlled. Mathematically, we can express the modified reliability function as follows:

\[ \rho_m = 1 - \left( \text{Max2P}_m \right)^{f_m} \]

In equation (7) \( f_m \) i.e., \( f_1 \) and \( f_2 \) are the optimization variables. Considering \( \text{Max2P} \) we can conclude the following condition for the modified reliability values for different limit value of \( f_m \).

- \( 0 \leq f_m \leq 1 \): \( \rho_m \leq \rho_m \)
- \( f_m = 1 \): \( \rho_m = \rho_m \)
- \( f_m > 1 \): \( \rho_m' > \rho_m \)

Figure 2 shows a typical example of relationship and physical meaning between \( \text{Max2P} \) and reliability.
values for different values of the optimization factor $f$.

Moreover, we have modified the equation (4) and (5) as in the following to account the integration process in parallel rather than the N-score method:

$$q_i = \frac{\rho_1}{\rho_1 + \rho_2 + \rho_3}; \quad q_j = \frac{\rho_2}{\rho_1 + \rho_2 + \rho_3}; \quad q_k = \frac{\rho_3}{\rho_1 + \rho_2 + \rho_3}$$  \hspace{1cm} (7)

Finally, each modality score is integrated using the weighting factor as

$$S(O_1, O_2, O_3 | \lambda_a) = \alpha_1 S(O_1 | \lambda_a) + \alpha_2 S(O_2 | \lambda_a) + \alpha_3 S(O_3 | \lambda_a)$$  \hspace{1cm} (8)

It is expected that the optimum values of $f_m$ in equation (6), could play an important role in the overall integration system. However, we cannot find the optimum values $f_m$ in linear searching way and thus we need optimization. For optimization we need a target function which is in detail in the following section.

4. Optimization Target Function

The optimization object function can be defined with the help of speaker identification rate. Mathematically, the optimization object function is defined in the following:

$$x(f_1, f_2, f_3) = \frac{\sum_{k=1}^{K} \sum_{l=1}^{L_k} \delta(\arg \max_j (P_j (X_{kl}))), k)}{K \sum_{k=1}^{K} L_k}$$  \hspace{1cm} (8)

In the above function, $\delta(i, j)$ is the delta function, $X_{kl}$ is the $l$-th speech feature vectors of the $k$-th speaker, and $P_j$ is the observation probability of given feature sequence for $m$-th speaker. The parameter $\arg \max P_j$ is the index of the speaker having maximum probability and $P_j$ is defined by equation (6). The optimization object function having optimization variable $f_1, f_2, f_3$ and $f_3$ expressed by equation (8) is nonlinear function and it is impossible to find the closed solution of the function. Thus the PSO approach is applied for optimizing the object function.

4. Experiments

4.1 Experimental database and specifications

For the experiments of this research work, the VidTIMIT audio-visual database was employed. The VidTIMIT database contains 43 speaker (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus.

4.2 Experiments and results

For our proposed method validation, we have used VidTimit database contains 10 sentences for each speaker in audio and visual signal level. For Audio case, we have grouped the 10 utterances into three; Group I (1-4 utterances), Group II (5-7 utterances) and Group III (8-10 utterances). However, we have taken only 9 utterances of each speaker and divided into two: Group I (1-3 utterances), Group II (5-7 utterances), Group III (8-10 utterances). Group I is used for speaker model generation while Group II and III are used for PSO validation involved in test stage for speaker identification. We have added an artificial illumination to the respective testing visual image as follows:

$$1(y, x) = w(y, x) + |\varphi|d + \delta$$  \hspace{1cm} (9)

where, $y = 1, 2, ..., M_y$, $x = 1, 2, ..., N_x$, and $d$ is either $y$ or $x$ depending on the illumination direction and $\varphi = -\delta / M_y or N_x$. In our experiment we have added the artificial illumination from up to down (UD) direction for face image while for lip case it was in the down to up (DU). Examples of the different face and lip images with artificial illumination in different directions are shown in Fig. 3 and 4. For features dimension reduction, we have applied local PCA [10] to both the visual features vector. Before features extraction we have transformed the color image to gray image. Table 1. shows the overall specifications
for the experiment.

![Image 55x680 to 145x730]
![Image 146x682 to 193x730]
![Image 194x679 to 238x730]
![Image 238x683 to 282x731]

Table 1: GMM Based Experts Specifications

<table>
<thead>
<tr>
<th>Modality</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>Frame Level Features: MFCC Features Dimension: 17 No. of Mixtures in GMM: 3</td>
</tr>
<tr>
<td>Lip Image ROI</td>
<td>Input Image: 64×64 pixel Features: DCT Dimension reduction: local PCA No. of PC’s: 10 No. of Mixtures in GMM: 3 Train Sentences image QF: 25 Test Sentences image QF: 5</td>
</tr>
</tbody>
</table>

Table 2: Only Audio Based Expert Performances

<table>
<thead>
<tr>
<th>Average SNR (dB)</th>
<th>Train DS Group</th>
<th>Test DS Group</th>
<th>IR (%)</th>
<th>Avg. IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.58</td>
<td>I (5-7)</td>
<td>II</td>
<td>91.47</td>
<td>86.04</td>
</tr>
<tr>
<td></td>
<td>I (8-10)</td>
<td>III</td>
<td>80.62</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Only Face Based Expert Performances

<table>
<thead>
<tr>
<th>δ</th>
<th>light Direc.</th>
<th>m</th>
<th>Train DS Gr.</th>
<th>Test DS Group</th>
<th>IR (%)</th>
<th>Avg. IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>UD</td>
<td>-δ/δ</td>
<td>I</td>
<td>III(8-10)</td>
<td>79.06</td>
<td>79.06</td>
</tr>
<tr>
<td>180</td>
<td>UD</td>
<td>-δ/δ</td>
<td>I</td>
<td>II (5-7)</td>
<td>79.06</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Only Lip Based Expert Performances

<table>
<thead>
<tr>
<th>δ</th>
<th>light Direc.</th>
<th>m</th>
<th>Train DS Gr.</th>
<th>Test DS Gr.</th>
<th>IR (%)</th>
<th>Avg. IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>DU</td>
<td>-δ/δ</td>
<td>I</td>
<td>III</td>
<td>49.61</td>
<td>49.22</td>
</tr>
<tr>
<td>160</td>
<td>DU</td>
<td>-δ/δ</td>
<td>I</td>
<td>II</td>
<td>48.83</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Baseline Audio-visual Integrated SI System Performance

<table>
<thead>
<tr>
<th>Modality</th>
<th>Test DS Group</th>
<th>IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio + Face+ Lip ROI II (5-7)</td>
<td>60.46</td>
<td></td>
</tr>
<tr>
<td>Audio + Face+ Lip ROI II (8-10)</td>
<td>61.24</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Proposed Optimum Reliability Based Multimodal SI System Performances

<table>
<thead>
<tr>
<th>PSO</th>
<th>Optim. ( f_1 )</th>
<th>Optim. ( f_2 )</th>
<th>Optim. ( f_3 )</th>
<th>Test DS Gr. using optim. ( f_m )</th>
<th>IR (%)</th>
<th>Avg. IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>2.6653</td>
<td>0.1764</td>
<td>0.0062</td>
<td>III</td>
<td>95.64</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>2.6999</td>
<td>0.1717</td>
<td>0.0051</td>
<td>II</td>
<td>93.79</td>
<td></td>
</tr>
</tbody>
</table>

The different experimental results are presented through Table 2 to 5. Table 2 shows the performance of audio based classifier with a high level of signal distortion due to office noises. Only face and lip expert performances are depicted in Table 3 and Table 4. Due to the uncertainty of lighting condition on the comparatively poor quality face and lip images, the face based classifier system performances are degraded in comparison with its counterpart audio based classifier whereas the lip based performance is very poor. Table 4 shows the experimental results of our adopted baseline system [14][22] and its performance is very far from the integration/fusion phenomena as it fails to reach higher or equal to the performance of best classifier system i.e., here the audio based classifier system.

We have presented our proposed method based experimental results in Table 5. It is seen from the Table 5 that the proposed method based system fulfilled the desire of fusion strategy through an improvement of speaker identification rate of about 8.67% in average using the optimum reliability.

5. Conclusion

In this paper we have presented a robust multimodal speaker identification system considering the optimal reliability in the integration process of audio and visual information. Introducing our proposed optimization factors in the existing convection function improved the performance of the multimodal robust speaker identification system significantly. The optimization factors were optimized by PSO algorithm. The entire experiments were performed using the VidTimit database. With the multimodal speaker identification system, we confirm that the proposed modified convection function could be a promising solution in biometrics.
technology. For further study, we will focus on audio-visual speaker verification and particularly the audio-visual speech recognition in real environment.

Acknowledgements
This work is supported by the MKE, Korea, under the ITRC support program supervised by the IITA (IITA-2009-(C1090-0903-0008)) and National Research Foundation of Korea Grant funded by the Korean Government (2009-0077345).

References