Hearing versus Seeing Identical Twins

Li Zhang[1], Jing Li[2], Shenggao Zhu[2], Wee Kheng Leow[2], Terence Sim[2]

School of Computing, National University of Singapore, Singapore

Abstract

Identical twins pose a great challenge to face recognition systems due to their similar appearance. Nevertheless, even they look alike, they speak differently. Voice is a natural signal to produce, and it is a combination of physiological and behavioral biometrics without appearance, therefore it is suitable for twin recognition. In this paper, we collect an audio-visual database from 39 pairs of identical twins, and conduct a comprehensive study of audio features on twin recognition. For each audio feature, we use Gaussian Mixture Model (GMM) to model the audio spectral distribution of each subject, and then employ the likelihood ratio of the probe belonging to different classes for verification. Our experimental results demonstrate that spectral features, such as Mel Frequency Cepstral Coefficients (MFCC), are more discriminative than other audio features including Linear Prediction Coding (LPC) and pitch. We further verify that when twin subjects are speaking same text, they can be better distinguished via audio. These results indicate that audio is a good biometric to distinguish between identical twin. Finally, we utilize the fusion of audio and appearance feature at feature level and significantly reduce the error rate for twin recognition.

Keywords: `MFCC, Identical Twins, Gaussian Mixture Model, Fusion`

[1]Corresponding author. Tel.:+65 8613 6907. E-mail addresses: lizhang@comp.nus.edu.sg. Postal address: AS6-502, Computer vision lab, School of Computing, National University of Singapore, Singapore, 117417

[2]E-mail addresses: {lijing,shenggao,leowwk,tsim}@comp.nus.edu.sg

1. Introduction

According to the statistics in [1], twins birth rate has risen from 17.8 to 32.2 per 1000 birth with an average 3% growth per year since 1990. This increase is associated with the increasing usage of fertility therapies and the change of birth concept. Nowadays women tend to bear children at older age and older women are more likely than younger women to conceive multiples spontaneously especially in developed countries [2]. Although currently identical twins still only represent a minority (0.2% of the world's population), it is worth noting that the total number of identical twins is equal to the whole population of countries like Portugal or Greece. This, in turn, has created an urgent demand for biometric systems that can accurately distinguish between identical twins.

Identical twins share the same genetic code, therefore they look very alike. This poses a great challenge to current biometric systems, especially face recognition system. This challenge due to the small appearance variation between identical twins has been verified by Sun *et al.* [2] on 93 pairs of twins using a commercial face matcher. Nevertheless, some biometrics depend not only on the genetic signature but also on the individual development in the womb. Some researchers explored the possibility of using behavior difference to distinguish between identical twins. For instances, Zhang *et al.* [3] utilized the Right-Cauchy tensor to compute the deformation along the fixed facial expression and used the difference of those deformation to differentiate twins. They verified the possibility of using expressive difference for twins recognition. Besides, they also tried to use external motion other than facial expression, such as head motion, as biometric in [4] to recognize twins. Their model was based on the abnormality of the external motion where more abnormal motion got higher weights to enlarge the difference and they reported a very good recognition accuracy. Generally, these behavior-based biometrics provide a good way for twin recognition, but they are intrusive and not very natural, and they are very sensitive to subject behavior consistency and relied on accurate tracking algorithm. Therefore, they are not suitable for twin recognition in the wild.

Besides the behavior-based biometric, several researchers showed encouraging results using fingerprint [5, 2], palmprint [6], ear [7] and iris [8, 2] to distinguish between identical twins. For example, equal error rate for 4-finger fusion reported by Sun *et al.* [2] was 0.49, and equal error rate for 2-iris fusion was also 0.49. Despite the discrimination ability of these biometrics, these biometrics require the cooperation of the subject. Therefore, it is still desirable for the biometric community to identify twins in a natural way.

Close to face biometric, audio biometric has been studied for decades and widely applied in authentication systems. Audio biometric makes use of the distinctive characteristics of information, present in the speech samples of the speaker to verify the authenticity. It is unrelated with appearance and reflects both anatomy (*e.g.*, size and shape of the throat and mouth) and learned behavioral patterns (textite.g., voice pitch, speaking style), therefore it is suitable for twin verification. Unfortunately, most works in this field, such as [9, 10, 11], only focus on general population, while neglecting the challenges caused by identical twins.

In this work, we are going to fill this gap by conducting a comprehensive study of audio biometric on twins. To the best of our knowledge, we are the first to investigate a comprehensive study of audio features for twin recognition. In this paper, we are trying to answer these questions:

1. **Can audio be used to distinguish between identical twins? Is it better than appearance-based approach?**
2. **If audio is effective for twin recognition, which audio feature is the best?**
3. **Under which setting can the audio feature be more discriminating ?**
4. **Even appearance is not effective for twin recognition, can we combine it with audio?**

In the first question, we ask for the feasibility of audio biometric on twins;

in the second and third questions we concern the methodology of how audio can be used for twins recognition. To answer those questions, a twin audio-visual database with 39 pairs of identical twins from a twins festival was firstly collected. For each twin subject, several audio clips speaking different contents were recorded. Each content was repeated at least twice. We utilized traditional Gaussian Mixture Model to compute the spectral distribution of each twin subject and then employed the likelihood ratio for classification. In total six different types of audio features were extracted for comparison including MFCC and LPC. We cannot take recent Deep Neural Network [12] which has been proven to be effective in speaker recognition in our experiments due to the small scale of training data. Besides audio, this database provides many mugshots for each twin subject. Those mugshots were employed to test the discrimination ability of conventional appearance features on identical twins including Eigenface [13], Local Binary Pattern [14] and Linear Discriminating Analysis on Gabor wavelet features (Gabor) [15]. Further, comparison bewteen text-independent setting and text-dependent setting were conducted. Finally, different fusion strategies were proposed to combine the appearance features with audio features. Through these studies, we concluded that audio is a good biometric for twins recognition and MFCC performs the best in both setting; under text-dependent setting, audio feature is more discriminating; by fusing appearance and audio at feature level, the best recognition performance can be achieved (EER equal to 0.05).

The contribution of this works is four fold:

1. **We have collected the largest audio-visual twin database, to the best of our knowledge. We are making this twin audio dataset publicly available.**
2. **We have verified the effectiveness of using audio biometric to distinguish identical twins.**
3. **We show that text-dependent setting can improve the discriminating ability of audio features.**

4. **We show twins can be better differentiated via the fusion of audio and appearance features at feature level.**

2. Related Works

2.1. Biometrics for Identical Twins

Studies on twins are limited, especially for 2D face biometric [2, 16, 17]. Sun et al. [2] are the first to evaluate the performance of using conventional appearance based face recognition for twins. They collected a twin database in 2007 at the fourth Annual Festival of Beijing Twins Day and compared face biometric with iris, fingerprint and a fusion of them. Their experiments were conducted using the FaceVACS commercial matcher and showed that identical twins were a challenge to current face recognition systems. Phillips *et.al* [16] further thoroughly extended the analysis of the performance of face recognition systems in distinguishing between identical twins on another database collected at the Twins Days festival in Ohio in 2009 and 2010. It consisted of images of 126 pairs of identical twins collected on the same day and 24 pairs with images collected one year apart. Facial recognition performance was tested using three of the top submissions to the Still Face Track at Multiple Biometric Evaluation 2010. Based on their experimental results, they claimed that under more realistic conditions, distinguishing between identical twins was very challenging. Klare *et.al* [17] analyzed the features of each facial component to distinguish identical twins from the same database in [16]. They also analyzed the possibility of using facial marks to distinguish identical twins. This work confirmed the challenge of recognizing identical twins merely based on appearance. Besides face biometric, some other biometrics are proposed to distinguish identical twins, such as fingerprint [5, 2], palmprint [6], ear [7] and iris [8, 2]. Sun *et al.* [2] reported the equal error rate for 4-finger fusion was 0.49%, and 2-iris fusion was also 0.49. Nejati *et.al* [7] used sift flow to capture the shape difference between different ears and uses the flow field to distinguish twins. They reported the accuracy using ear biometric was higher than 0.90. Zhang *et al.* [3, 4] proposed

to use behavior level biometric to distinguish identical twins. Their assumption was that twins may look like, but they behave different caused by their different training experience. Zhang *et al.* [3] firstly utilized the Right-Cauchy tensor to compute the deformation on fixed facial expression and used the deformation pattern of those expression motions for recognition. Besides, they also tried to use external motion, such as head motion, as biometric in [4] to recognize twins. They computed the abnormality of external motions, and then weighted more on more abnormal motions in order to enlarge the difference. They reported a very good recognition accuracy for those behavior based biometrics. Nevertheless, unlike face biometric, all those above non-face biometrics are intrusive and not very natural, they are not suitable for twin recognition in the wild. We are looking for a biometric, which is not only discriminating but also common and easy to use.

2.2. Audio based Speaker Verification

Audio signal provides several levels of information. Primarily, audio signal conveys the words or message being spoken, and on a secondary level, it also conveys information about the identity of the speaker included in audio waves [10]. Depending upon the application, the general area of audio biometric is divided into two specific tasks: verification and identification. In verification, the goal is to determine from a audio sample if a person is whom he or she claims. In speaker identification, the goal is to determine which one of a group of known audios best matches the input audio sample. Furthermore, in either task the audio can be constrained to text dependent (*i.e.*the speaker is required to talk same phrase) and text independent (*i.e.* the speaker can talk different phrase). Various approaches are proposed for either group. For instances, Douglas *et al.* [10] and Sinith *et al.* [9] proposed to use Mel Frequency Cepstral Coefficients and Gaussian Mixture Model to solve text independent identification problem. Dupont *et al.* [18] and Dean *et al.* [19] tried to use hidden Markov model to model the distribution of the speaker spectral shape from audio sample and claimed the identity using maximum likelihood of the posterior probabilities belonging

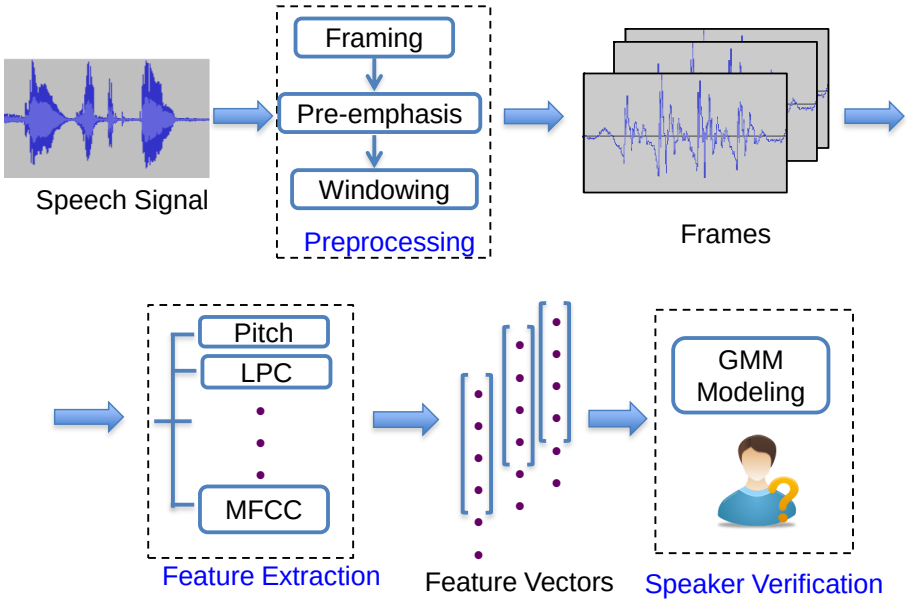


Figure 1: Flowchart of twin verification using audio

to different classes. Shi *et.al.* [20] analyzed the effects of uncertainty in the phase of audio signals and indicate that a large amount of phase uncertainty can benefit for both text independent speaker recognition and text dependent speaker recognition. Further, Nakagawa *et.al* [11] proposed to use Gaussian Mixture Model to fuse Mel Frequency Cepstral Coefficients and phrase feature for both text independent and dependent identification. As mentioned earlier, those works focused only on general population, while twins were overlooked by them. It is urgent for the biometric society to fill the potential security hole of audio based recognition system caused by twins.

3. Proposed Approach

The proposal of our twin verification method is illustrated in Figure 1. Input audio is first preprocessed to generate frames of audio signals. Then comprehensive acoustic features are extracted from each frame. Finally GMM models are learned from these features for speaker verification. The audio preprocess-

ing and feature extraction are implemented using openSMILE [21] which is an open-source program for general audio signal processing and low-level acoustic feature extraction.

3.1. Audio Feature Extraction

The first step of audio preprocessing is framing, which divides a stream of audio signal into successive overlapping frames. The sample rate of the audio is 44.1 kHz, and the frame size is set to 23 milliseconds with 50% overlap. The energy in the high frequencies is boosted in each frame (i.e., pre-emphasis) to compensate for the nonlinear nature of human voice that more energy is located at lower frequencies. A Hamming window is utilized to smooth out the discontinuities at the beginning and the end of each frame. Silent frames are then filtered out where the voicing probability are lower than a certain threshold (0.4 in our experiments).

After preprocessing, six audio features are extracted from the frames. The extracted audio features in time domain and frequency domain are as follows: Pitch [22], Frame Energy, Linear Prediction Coefficients (LPC) [23], Spectral Centroid, Spectral Rolloff, and Mel Frequency Cepstral Coefficients (MFCC) [24].

Pitch is a high-level perceptual property of the audio that allows the ordering on a frequency-related scale, which is calculated as the fundamental frequency (F0). Internally openSMILE uses the Sub-Harmonic-Summation (SHS) method [25] to identify the fundamental frequency from the harmonic structure. The Energy of each frame is calculated as the root-mean-square of the sample points in the frame. LPC is the coefficients of the linear predictive coding from each frame. Linear predictive coding is to estimate a linear coefficients which can represent a audio sample point by a linear combination of several past samples. In our work, the number of coefficients is set to 8. For frequency domain features, Fourier transform is applied to each frame. Spectral Centroid is to use the centroid of the magnitude spectrum as feature. Spectral Rolloff is to use the frequency below which 75% of the magnitude distribution is concentrated

as feature. In addition, we also extract MFCC, which is a more complex and comprehensive perceptual feature and is widely used in most speaker recognition methods. It first maps the powers of the spectrum onto the *mel scale* using a group of triangular overlapping windows (i.e., a filter bank). Psychophysical studies have shown that human perception of sound frequency does not follow a linear scale. The mel scale is a non-linear perceptual scale that converts a given frequency f in Hz into *mel*s using the following approximation formula:

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}). \tag{1}$$

Then the mel-frequency cepstrum is obtained by taking the discrete cosine transform (DCT) of the logs of the powers at each mel frequency in the filter bank. Finally, MFCCs are the amplitudes of the resulting cepstrum. In our work, 21 coefficients are used for MFCC.

3.2. Modeling using GMM

For each subject, his/her identity-dependent acoustic spectral distribution is modeled as a weighted sum of M component densities given by the equation

$$p(x) = \sum_{i=1}^{M} w_i b_i(x) \tag{2}$$

where x is the D-dimensional feature vector (In our case, it contains Pitch, LCP and MFCC), $b_i(x)$ is the component density and w_i is the mixture weight. Each component density is represented as a Gaussian distribution of the form

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Delta_i|^{1/2}} \exp\{-\frac{1}{2}(x - \mu_i)'\Delta_i^{-1}(x - \mu_i)\} \tag{3}$$

with mean vector μ_i and covariance matrix Δ_i. The sum of mixture weights w_i equals to 1. For convenience, we represent mean vectors, covariance matrices and mixture weights in a model vector $\Gamma = \{w_i, \mu_i, \Delta_i\}, i = 1, ..., M$. Therefore, each speaker is represented by his/her model Γ.

Given the training data in the gallery, we use Expectation Maximization algorithm [26] to estimate the Γ for each subject. In the verification phase, given a test feature vector ψ, and the hypothesized speaker S, we aim to check

whether the hypothesized identity is the same as the classified identity. We state this task as a basic hypothesis test between two hypotheses:

H0: ψ is from the hypothesized speaker S.

H1: ψ is not from the hypothesized speaker S (*i.e.*, ψ is from the twin sibling of hypothesized speaker S).

The optimum classification to decide between these two hypotheses is through the likelihood ratio (LR) given by

$$LR = \frac{p(\psi|H0)}{p(\psi|H1)} \tag{4}$$

If $LR > \epsilon$, we accept H0; otherwise, we reject $H0$. Here, ϵ is the threshold, $p(\psi|H0)$ is the probability density function for the hypothesis subject S for the observed feature vector ψ, and $p(\psi|H1)$ is the probability density function for not being the hypothesis subject S for the observed feature vector ψ.

4. Experiments

4.1. Data Collection

We collected a twins audio-visual database at the Sixth Mojiang International Twins Festival held on 1 May 2010 in China. It includes Chinese, Canadian and Russian subjects for a total of 39 pairs of twins. Several examples can be seen in Figure 2. As we can see, each subject looks very similar to the other silbling of the twins. During our data collection, we know that even for their parents, sometimes it is hard to differentiate among their twin children purely by appearance difference.

For each subject, there were at least three audio recordings, each around 30 seconds. The audio texts of those recordings were different. For the first recording, the subjects were required to count the number from one to ten. For the second recording, the subjects were reading a paragraph; For the third recording, the subjects were reciting a poem. In each recording, the text was repeated three times by the subject. The recordings were then cut into three clips, each of which contains one repetition of the text. In all, we used 702 audio

Figure 2: Some image examples of identical twins.

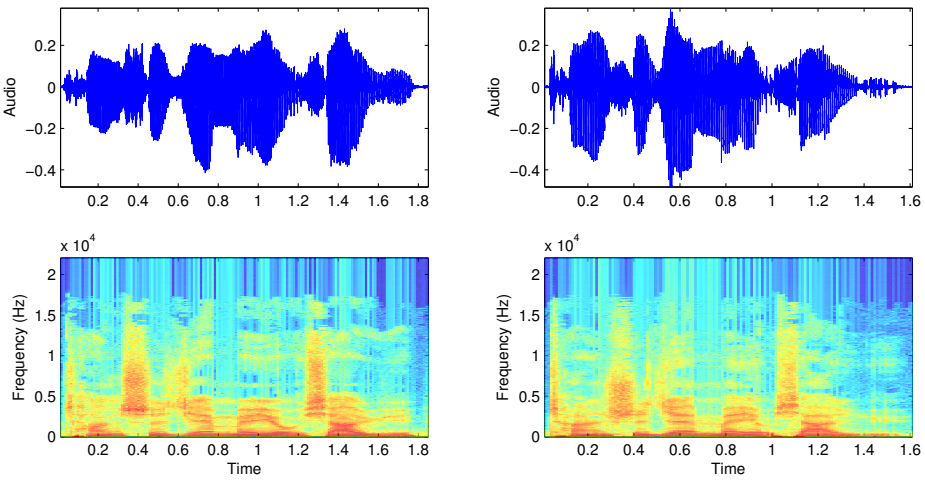


Figure 3: Example audios (first row) and corresponding spectrograms (second row) of identical twin reading the same text. Each column is from one sibling of the twins.

clips in our experiments, 9 audio clips of 3 different texts for each subject.An example is shown in the Figure 3. We are making this twin audio database publicly available[3].

4.2. Text-Dependent Verification

In the experiments for text-dependent verification, the database was divided into 117 sets according to the texts and number of twins. Each set contained 3 repetitions of the same text by each pair of twins. We used two of the audio clips as gallery for training for each subject and the remaining one as a probe for testing. Before training, we first framed all the audio clips and extracted 10 features from each frame. These features include: Voice probability, Pitch [22], Energy, LPC [23], Spectral Centroid, Spectral Rolloff, Spectral Flux, MFCC [24] and spectral statistics including spectral variance, skewness, kurtosis, slope, and the positions of spectral maximum and minimum. The voice probability characterizes the probability of frame to be text spoken by people. The voice probability is low during the breaks of the texts. By controlling the threshold

[3]This database can be obtained via email to lijing@comp.nus.edu.sg

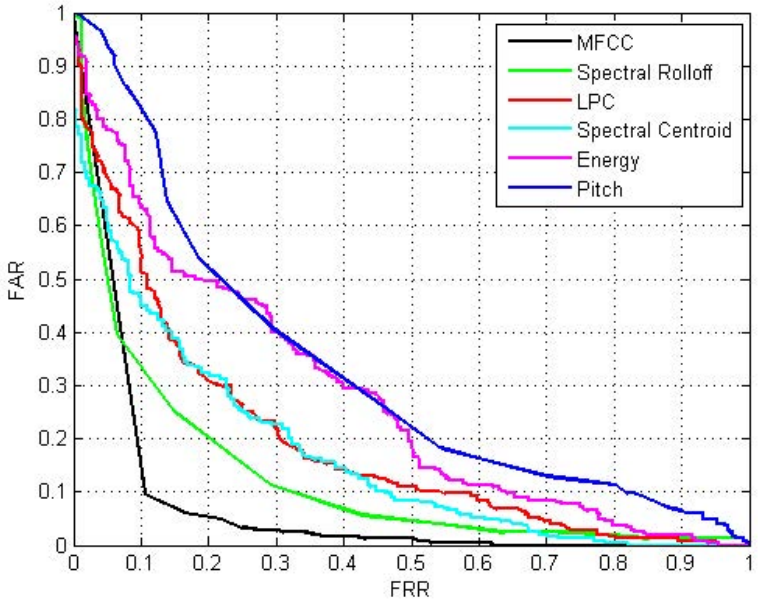
θ for the voice probability, we discarded unreliable frames that contain noises instead of human voice. After the filtering, a Gaussian Mixture Model of each subject was learned in all sets. The covariance matrix of all GMMs was assumed to be diagonal, whereas the number of components for GMMs changes with the feature types. In the testing phase, each probe was tested on the two GMMs trained on the corresponding pair. The number of components for GMMs was optimized on the test sets for better performance.

The twin verification performance is evaluated in terms of Twin Equal Error Rate (Twin-EER) which Twin False Acceptance Rate (Twin-FAR) meets the False Rejection Rate (FRR). The Twin-FAR is the ratio between the times that twin imposter is recognized as genuine and the total number of imposter. FRR is the ratio between the times that genuine is recognized as imposter and the total number of the genuine. In our experiments, the FRR-FAR curve was obtained by controlling the threshold ϵ for LR in Equation 4.

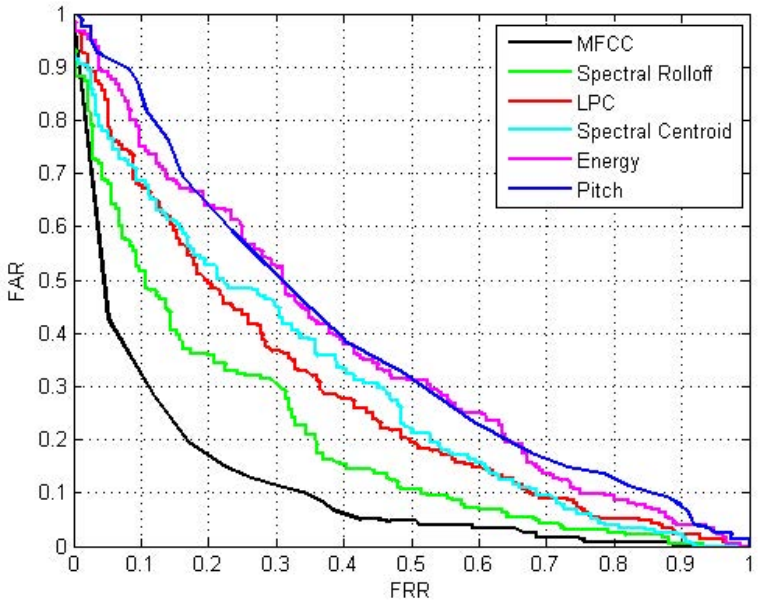
The result for text-dependent verification is shown in Figure 4(a). As can be seen, MFCC is the most discriminative feature with the equal error rate 0.1, followed by Spectral Rolloff with 0.2.

4.3. Text-Independent Verification

In the experiments for text-independent verification, the database was divided into 39 sets according to the texts and pairs of twins. Each set contains 3 repetitions of 3 different texts by each pair of twins. We used the clips of 2 texts as gallery to train GMM for each subject and the clips of remaining one text as probes. The raw data were framed and filtered by the same methods as in text-dependent verification section. Then the Gaussian Mixture Model for each subject was learned from the features of 6 clips with the same assumption that the covariance matrix is diagonal. In the testing phase, 3 clips of the third text by one subject were tested on the two corresponding GMMs. The final FRR-FAR curve was obtained by controlling threshold ϵ. In the same way, the number of components for GMMs was optimized on the test sets for better performance.



(a) Text-Dependent Accuracy



(b) Text-Independent Accuracy

Figure 4: Performance comparison for text-dependent and text-independent audio verification

The result for text-independent verification is shown in Figure 4(b). Among all the features, MFCC is the most discriminative one with error rate of 0.18, followed by Spectral Rolloff, LPC, Spectral Centroid, Energy, and Pitch in sequence. The performances for most features are greater than 0.4, except Spectral Flux.

4.4. Appearance based performance

For baseline comparison, we choose three traditional facial appearance approaches, Eigenface, Local Binary Pattern and Gabor, to test the performance of using appearance to distinguish between identical twins. For each twin subject, we randomly selected 8 images. The images were then registered by eye positions detected by STASM [27] and resized to to 160 by 128. For Eigenface, we vectorized gray intensity in each pixel as feature and performed PCA to reduce the dimension. For LBP, we divided the image into 80 blocks. For each block, we extracted the 59-bins histogram. For Gabor, we used 40 Gabor (5 scales, 8 orientation) filters and set the kernel size for each Gabor filter to 17 by 17. A PCA was performed to reduce the feature dimension for LBP and Gabor. The experimental result is shown in Figure 5. From this figure, we can see that identical twins indeed pose a great challenge to appearance based approach. The General-EER of Gabor for general population is around 0.12, while Twin-EER is significantly larger than 0.33. We can also see that there is no huge difference between Intensity, LBP and Gabor for twin verification. The Twin-EERs for them are 0.35 (Intensity), 0.34 (LBP) and 0.34 (Gabor), separately.

4.5. Discussion

We compared audio performance with appearance performance. We can see a significant improvement in performance of audio based approach from Figure 4. In all, there are 2 audio features that surpass the performance of appearance features in both text-dependent and text-independent tests: MFCC and Spectral Rolloff. The Twin-EER for MFCC in text-independent test is 0.18

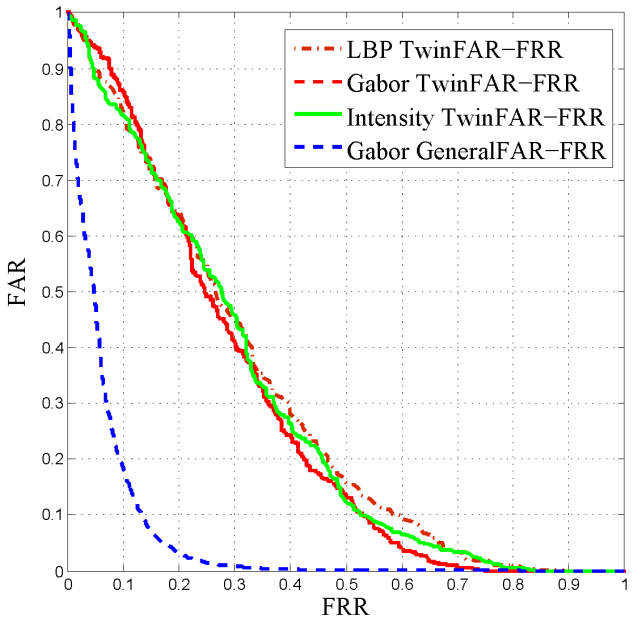


Figure 5: Appearance Accuracy

and the Twin-EER for MFCC in text-dependent test is 0.10. Twin-ERRs for Spectral Rolloff in text-dependent test and text-independent test are 0.20 and 0.30. Moreover, based on the experimental results in [18], the General-EER for speaker verification on general population is around 0.05, which is much smaller than the best (0.18) in twins database. The difference may come from three aspects: 1) Insufficient training data in our experiments. In our case, we only use one audio recording around 30 seconds as training, and the audio text is very simple and sometime duplicated. Therefore, it cannot cover the entire audio spectral pattern. 2) The audio spectral pattern for identical twins may have some overlaps. Identical twins share the same genetic code, therefore their voice may share some similarity. 3) Our audio recording is not collected in noise-free environment, the ambient sound may also degrade our performance.

We also compared the performance of both text-dependent and text-independent verification. As shown in Figure 4, this results indicate that when twins are speaking same text, they can be better distinguished. This actually is not sur-

prising, because when speaking same contents, the intra-variation of the audio features from same subjects are becoming smaller, which are beneficial for classification.

We also compared the performance of various audio features. The performance of the features generally follows the order: MFCC, Spectral Rolloff, LPC, Energy and Pitch. MFCC is the most discriminative feature for the audio-based twin verification, while Pitch and Engery are the worst. Pitch represents the major spectral frequency of the subject and usually the largest difference of pitch feature is from gender, speaking contents and dialect, while for identical twins, they usually have same dialect and gender. This may cause the low discrimination ability of Pitch. For Energy feature, it reflects the energy expressed in the audio wave, therefore it is very sensitive to the magnitude of speaking volume. This feature is easy to fail when subjects are talking different contents, different rhythm and different mood. For LPC, it is a linear coding of audio signal in time domain. The linear assumption may be not enough to model to identity information in the audio signal. Moreover, the analysis is performed on time domain instead of frequency domain, thus it may not robust against with environment noise. MFCC and spectral Rolloff performed both captured the frequency property of the audio. Compared with spectral Rolloff, MFCC computed the statistics of spectral in different frequency range and encoded different frequency ranges into a vector, while spectral Rolloff only captured the global spectral statistics.

4.6. Fusion of Appearance Features and Audio Feature

In the algorithm, we combined the appearance and audio features to improve the twin verification accuracy to develop a multimodal system for twin recognition. We chose Gabor and LBP to represent appearance feature and MFCC to represent audio feature. We picked those features, because these features performed the best in each category verified in our previous experiment. We fused these two biometrics at two levels: feature level and confidence level. The feature level is to extract a new feature and then perform classification. The

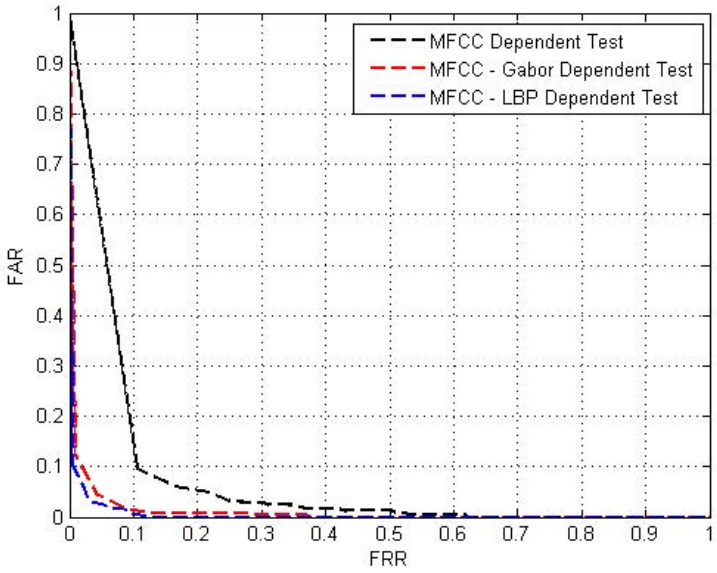
confidence level is to render a decision about the identity of an individual by consolidating the scores from multiple biometrics.

To fuse the appearance and audio features at feature level, we concatenated the normalized appearance features to the normalized MFCC vectors. For normalization, we employed whiten PCA for Gabor, LBP and MFCC. After feature normalization, similar to Section 3.2, we used GMM to model the distribution of extracted feature and perform classification using likelihood ratio. We carried out both text-dependent verification and text-independent verification in our test. The final FRR-FAR curve was obtained by controlling threshold ϵ. The experimental result is shown in Figure 6. From this figure, we can see that both text-dependent and text-independent verifications are greatly improved to the error rate of 0.04 and 0.07. Moreover, fusion of MFCC and LBP outperforms the fusion of MFCC and Gabor.

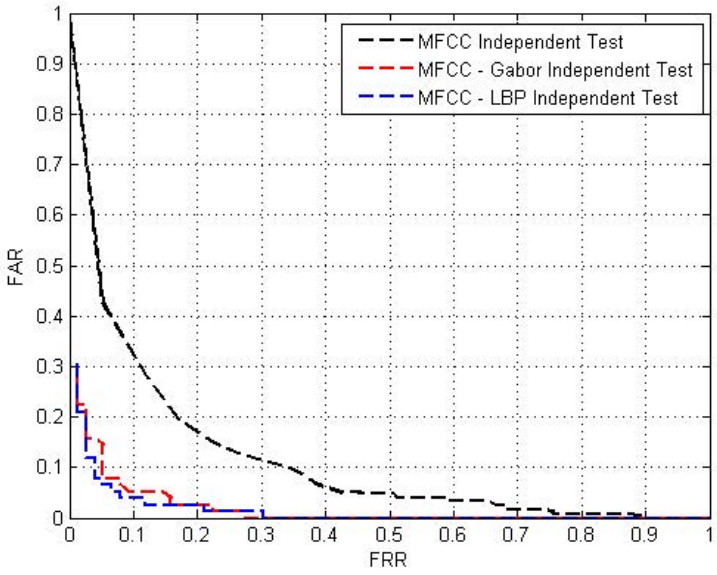
For confidence level fusion, given a probe and a claim identity, we computed the appearance confidence and audio confidence respectively. The new confidence was computed as $p(\psi|H0) = \alpha p_{Gabor/LBP}(\psi|H0) + (1-\alpha)p_{MFCC}(\psi|H0)$. Then, we computed the likelihood ratio as Equation 4 which was then compared against the pre-set threshold ϵ. If $LR > \epsilon$, they are genunines; otherwise they are imposters. We conducted the experiments on the whole database and set the α for the best of test performance in our dataset. The performance is showed in Figure 7. From this figure, we can see that the error rate for text-dependent verification decreases from 0.1 to 0.08 and error rate for text-dependent verification decreases from 0.18 to 0.15.

4.7. Discussion

We can see that twin can be well distinguished by fusion of appearance and audio features. The best equal error rate is less than 0.05, which is a significant improvement compared with appearance-based approach and audio-based approach alone. As far we know, this is the best performance achieved for twin recognition. Moreover, among different strategies, feature level fusion performs better than confidence level, and this is not surprising. Similar to

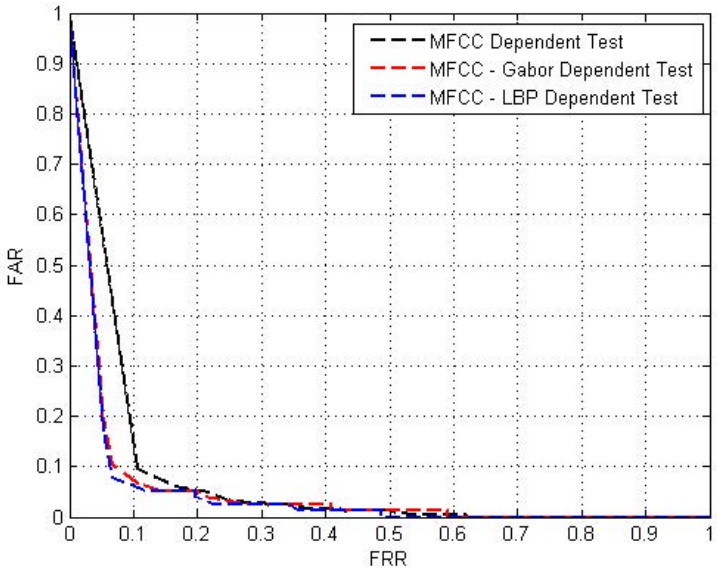


(a) Text-Dependent Test

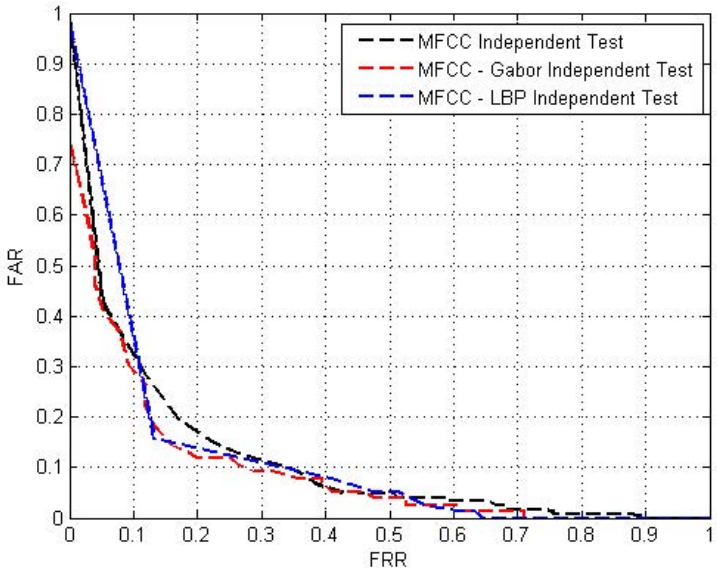


(b) Text-Independent Test

Figure 6: Performance of Feature Level Fusion



(a) Text-Dependent Test



(b) Text-Independent Test

Figure 7: Performance of Confidence Level Fusion

previous audio features, still the text-dependent setting can generate better performance than text-independent setting.

5. Conclusion and Future Work

In this work, we collect a moderate size of identical twins database including appearance and audio. The database contains several images and 3 recordings of different contents repeated three times of each pair of twins. We propose to use Gaussian Mixture Model to model the audio spectral pattern for verification. We examine the performances of several different vocal features. Among these features, MFCC has the most discrimination ability. The results verify that audio biometric can be used to distinguish between identical twins and it is significantly better than traditional facial appearance features, including EigenFace, LBP and Gabor. We further prove that the accuracy can be improved via fusion of audio biometric and facial appearance. Our results show that feature level fusion outperforms the confidence level fusion.

In future, we would like to test the robustness of our audio proposal, including the length of training data and environment noise. Even though our current result is very promising, we still hope to collect a larger twin database for our research. We also intend to test the scalability of our audio proposal. Finally, we look forward building a multimodal biometric system to which can work well for general population but also can prevent the evil twin attack.

References

[1] J. Martin, H. Kung, T. Mathews, D. Hoyert, D. Strobino, B. Guyer, S. Sutton, Annual summary of vital statistics: 2006, Pediatrics.

[2] Z. Sun, A. Paulino, J. Feng, Z. Chai, T. Tan, A. Jain, A study of multibiometric traits of identical twins, SPIE.

[3] L. Zhang, N. Ye, E. M. Marroquin, D. Guo, T. Sim, New hope for recognizing twins by using facial motion, in: WACV, IEEE, 2012, pp. 209–214.

[4] L. Zhang, K. Ma, H. Nejati, L. Foo, T. Sim, D. Guo, A talking profile to distinguish identical twins, Image and Vision Computing.

[5] A. Jain, S. Prabhakar, S. Pankanti, On the similarity of identical twin fingerprints, Pattern Recognition (2002) 2653–2663.

[6] A. Kong, D. Zhang, G. Lu, A study of identical twins' palmprints for personal verification, Pattern Recognition (2006) 2149–2156.

[7] H. Nejati, L. Zhang, T. Sim, E. Martinez-Marroquin, G. Dong, Wonder ears: Identification of identical twins from ear images, ICPR (2012) 1201–1204.

[8] J. Daugman, C. Downing, Epigenetic randomness, complexity and singularity of human iris patterns, Proceedings of the Royal Society of London (2001) 1737.

[9] M. Sinith, A. Salim, K. Gowri Sankar, K. Sandeep Narayanan, V. Soman, A novel method for text-independent speaker identification using mfcc and gmm, in: ICALIP, IEEE, 2010, pp. 292–296.

[10] D. A. Reynolds, R. C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models, Speech and Audio Processing, IEEE Transactions on 3 (1) (1995) 72–83.

[11] S. Nakagawa, L. Wang, S. Ohtsuka, Speaker identification and verification by combining mfcc and phase information, Audio, Speech, and Language Processing, IEEE Transactions on 20 (4) (2012) 1085–1095.

[12] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 6645–6649.

[13] M. A. Turk, A. P. Pentland, Face recognition using eigenfaces, in: CVPR, IEEE, 1991, pp. 586–591.

[14] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, Computer Vision-ECCV 2004 (2004) 469–481.

[15] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, Image processing, IEEE Transactions on 11 (4) (2002) 467–476.

[16] P. Phillips, P. Flynn, K. Bowyer, R. Bruegge, P. Grother, G. Quinn, M. Pruitt, Distinguishing identical twins by face recognition, in: FG 2011.

[17] B. Klare, A. Paulino, A. Jain, Analysis of facial features in identical twins, in: Biometrics (IJCB), 2011 International Joint Conference on, IEEE, 2011, pp. 1–8.

[18] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, Multimedia, IEEE Transactions on 2 (3) (2000) 141–151.

[19] D. Dean, S. Sridharan, T. Wark, Audio-visual speaker verification using continuous fused hmms, in: Proceedings of the HCSNet workshop, 2006, pp. 87–92.

[20] G. Shi, M. M. Shanechi, P. Aarabi, On the importance of phase in human speech recognition, Audio, Speech, and Language Processing, IEEE Transactions on 14 (5) (2006) 1867–1874.

[21] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the Munich versatile and fast open-source audio feature extractor, in: Proceedings of the international conference on Multimedia, ACM, 2010, pp. 1459–1462.

[22] R. J. Zatorre, A. C. Evans, E. Meyer, A. Gjedde, Lateralization of phonetic and pitch discrimination in speech processing, Science 256 (5058) (1992) 846–849.

[23] B. S. Atal, S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, The Journal of the Acoustical Society of America 50 (1971) 637.

[24] B. Logan, et al., Mel frequency cepstral coefficients for music modeling, in: International Symposium on Music Information Retrieval, Vol. 28, 2000, p. 5.

[25] D. J. Hermes, Measurement of pitch by subharmonic summation, The journal of the acoustical society of America 83 (1) (1988) 257–264.

[26] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society (1977) 1–38.

[27] S. Milborrow, F. Nicolls, Locating facial features with an extended active shape model, ECCV (2008) 504–513.