Hearing versus Seeing Identical Twins

Li Zhang, Shenggao Zhu, Terence Sim, Wee Kheng Leow, Hossein Najati and Dong Guo

School of Computing
National University of Singapore
Singapore, 117417
{lizhang,shenggao,tsim,leowwk}@comp.nus.edu.sg,dnguo@fb.com,

Abstract. Identical twins pose a great challenge to face recognition systems due to their similar appearance. Nevertheless, even though twins may look alike, we believe they speak differently. Hence we propose to use their voice patterns to distinguish between twins. Voice is a natural signal to produce, and it is a combination of physiological and behavioral biometrics, therefore it is suitable for twin verification. In this paper, we collect an audio-visual database from 39 pairs of identical twins. Three types of typical voice features are investigated, including Pitch, Linear Prediction Coefficients (LPC) and Mel Frequency Cepstral Coefficients (MFCC). For each type of voice feature, we use Gaussian Mixture Model to model the voice spectral distribution of each subject, and then employ the likelihood ratio of the probe belonging to different classes for verification. The experimental results on this database demonstrate a significant improvement by using voice over facial appearance to distinguish between identical twins. Furthermore, we show that by fusion both types of biometrics, recognition accuracy can be improved.

Keywords: identical twins; verification; fusion; Gaussian Mixture Model

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 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. The challenge using facial appearance to distinguish 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, such as expressions and head motion [3] to distinguish between identical twins. Zhang *et al.* [3] proposed to use exception reporting model to model the head motion abnormality to differentiate twins. They reported the verification accuracy was over 90%, but their algorithm was very sensitive to subject behavior consistence and strongly relied on accurate tracking algorithm. Several researchers showed encouraging results by using fingerprint [4, 2], palmprint [5], ear [6] and iris [7, 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 of the discriminating ability of those biometrics, those biometrics require the cooperation of the subject. Therefore, it is desirable to identify twins in a natural way. In this paper, we propose to utilize voice biometric to distinguish between identical twins and compare voice biometric with facial appearance. Voice is non-intrusive and natural, it does not require explicit cooperation of the subject and is widely available from videos captured by ordinary cam-corders. To the best of our knowledge, we are the first to investigate voice and appearance biometrics at the meantime.

Voice signal usually conveys several levels of information. Primarily, voice signal conveys the words or message being spoken, but on a secondary level, it also conveys information about the identity of the speaker [8]. Voice biometric tries to extract the identity information from the voice and uses it for speaker recognition. Generally speaking, the speaker recognition can be divided into two specific tasks: speaker verification and speaker identification. In speaker verification, the goal is to establish whether a person is who he/she claims to be; whereas in speaker identification, the goal is to determine the identity (name or employee number) of the unknown speaker. In either task the speech can be further divided into text dependent (*i.e.*the speaker is required to talk same phrase) and text independent (*i.e.* the speaker can talk different phrase). Douglas *et al.* [8] and Sinith *et al.* [9] proposed to use Mel Frequency Cepstral Coefficients and Gaussian Mixture Model to solve text independent identification problem for general population, i.e. non twins. Dupont *et al.* [10] and Dean *et al.* [11] tried to use hidden Markov model to model the distribution of the speaker spectral shape from voice sample and claimed the identity using maximum likelihood of the posterior probabilities belonging to different classes. Both these works demonstrated that the identity of speaker can be well recognized via their voices under the condition that voice samples were in good quality and the gallery size was small, i.e. the number of subjects is small. This conclusion, in turn, brings new hope to use voice biometric to differentiate identical twins, because to distinguish between identical twins, the number of involved subjects was very small *i.e. the number of twins siblings*. In this paper, we are trying to answer those questions as follows:

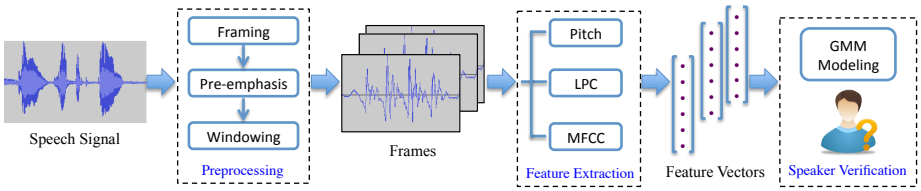


Fig. 1. Flowchart of twin verification using voice

1. **Can voice be used to distinguish between identical twins? Is it better than appearance based approach? If it is, which voice feature is the best for identical twins?**
2. **Can we combine facial appearance with speech to improve accuracy?**

Our work can be divided into three parts: 1) we firstly collected a twin audio-visual database with 39 pairs of identical twins and test the discriminating ability of facial appearance to distinguish between identical twins by using Eigenface [12], Local Binary Pattern [13] and Linear Discriminating Analysis on Gabor wavelet features (Gabor) [14]. 2) We propose to use Gaussian Mixture Model to estimate the spectral shape of each twin subject, an then use the ratio of the probabilities belonging to different twin subjects for verification. Three types of voice features are used: Pitch, LPC and MFCC. 3) We use confidence level fusion to combine the Gabor and MFCC to improve accuracy.

2 Twin Verification using GMM

2.1 Preprocessing and Feature Extraction

The proposal of our twin verification can be seen in Figure 1. The first step of preprocessing is framing which is to divide audio into successive overlapping frames. The frame size is set to 23 milliseconds in our work, with 50% overlap. The energy in the high frequencies is boosted in each frame to compensates 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 the frame. Since silent frames may exist in the speech signal, we filter out these frames using a simple thresholding method. The threshold θ indicates the probability of containing human voice in this frame. If θ is larger than the threshold, we keep this frame; otherwise we throw it away. In our experiments, we set the threshold t0 0.4.

After preprocessing, various acoustic features can be extracted from the frames. We select three kinds of features for testing and comparison purpose, which are Pitch [15], Linear Prediction Coefficients (LPC) [16], and Mel Frequency Cepstral Coefficients (MFCC) [17]. Pitch is a perceptual property of the voice that allows the ordering on a frequency-related scale. MFCC is to map the

powers of the frame spectrum onto the mel scale and then uses amplitudes of discrete cosine transform of the list of mel scale as feature. LPC is the coefficients of the linear predictive coding from the frames. In our work, the MFCC coefficient number is set to 13 and the predictor order (i.e., the number of LPC coefficients) is set to 8.

2.2 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

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

where x is the D-dimensional feature vector (In our case, it is 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

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

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

Given the training data in the gallery, we use Expectation Maximization algorithm [18] 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 same to classified identity. We state this task as a basic hypothesis test between two hypotheses:

H0: ψ is from the hypothesized twin 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

$$LR = \frac{p(\psi|H0)}{p(\psi|H1)} \tag{3}$$

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.

3 Experiments

3.1 Data and Performance Evaluation

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

Fig. 2. Some image examples of identical twins

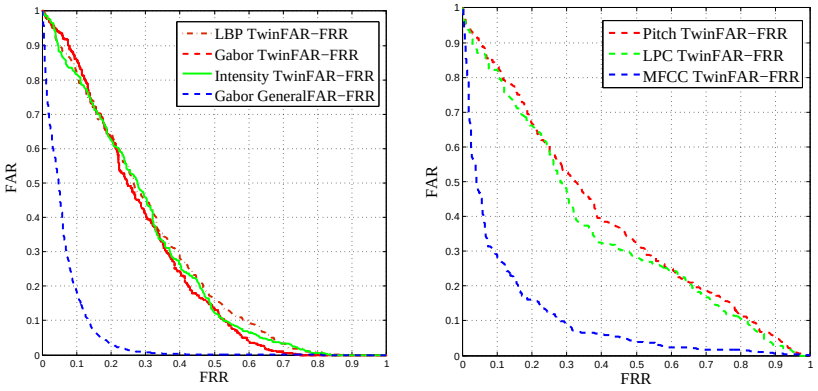
Russian subjects for a total of 39 pairs of twins. Several examples can be seen in Figure 2. For each subject, there are at least three audio recordings, each around 30 seconds. The talking content of those recordings are different. For the first recording, the subjects are required to count the number from one to ten; For the second recording, the subjects are reading a paragraph; For the third recording, the subjects are reciting a poem.

The twin verification performance is evaluated in terms of Twin Equal Error Rate(Twin-EER) which Twin False Accept Rate(Twin-FAR) meets the False Reject Rate (FRR). The Twin-FAR is the ratio between the times that twin imposter is recognized as genuine with the total number of imposter. FRR is the ratio between the times that genuine is recognized as imposter with the total number of the genuine. We also introduce General Equal Error Rate(General-EER) where General False Accept Rate(General-FAR) meets the FRR. The General-FAR is the ratio between the times that general imposter is recognized as genuine with the total number of the non-twin imposter. The purpose of introducing General-FAR is to compare the verification accuracy between twins with non-twins to see the challenge brought by twins.

3.2 Performance of Appearance and Audio Based Approach

We chose 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 select 8 images. The images are then registered by eye positions detected by STASM [19] 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 extract 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 is performed to reduce the feature dimension for LBP and Gabor. The experimental result is shown in Figure 3(a). 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 0.122, 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 0.352 (Intensity), 0.340 (LBP) and 0.338 (Gabor), separately.

For voice based twin verification, we use one of the audio recordings as gallery to train the GMM for each subject. Then, the remaining audio recordings are



(a) Appearance Accuracy (b) Voice Accuracy

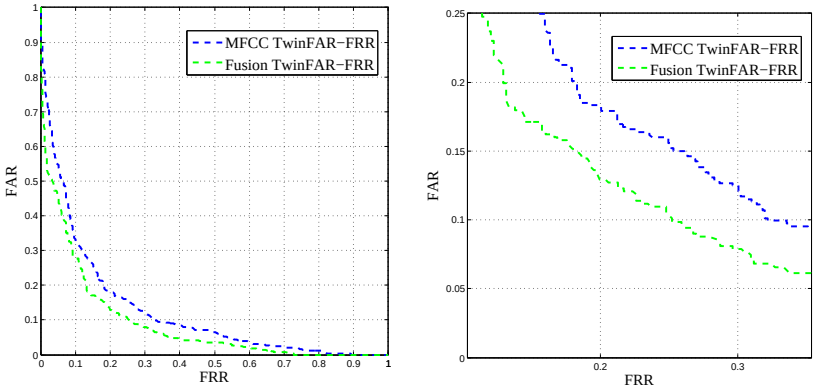
Fig. 3. Performance comparison between facial appearance and voice biometric

used as probe. For each recording, we divide it into three parts, and each part is acted as single probe. During GMM training, the covariance matrix is assumed to be diagonal and the number of Gaussians is set to 4 for Pitch, 4 for LPC and 5 for MFCC. The number of gaussian is optimized on the test set for better performance. The experimental result is showed in Figure 3(b). Compared with Figure 3(a), it can be clearly seen that twins can be better distinguished via voice than appearance. The Twin-EER for MFCC is 0.171, which is significantly better than appearance (the best for appearance is 0.338). However, not all voice features are better than appearance. The Twin-EERs of pitch and LPC (0.394 for Pitch and 0.366 for LPC) are even larger than appearance based approach. This shows that Pitch and LPC is not discriminating enough for twins.

Moreover, based on the experimental results in [10], the General-EER for speaker verification on general population is around 0.05, which is much smaller than the best (0.171) 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 talking content is very simple and sometime duplicated. Therefore, it may cannot cover the entire voice spectral pattern. 2) The voice spectral pattern for identical twins may have some overlap. Identical twins share the same genetic code, therefore their voice may share some similarity. 3) Our audio recording is not collected in very clean environment, the environment sound may also degrade our performance. The General-EER reported by [10] was obtained at clean recording room.

4 Fusion of Gabor and MFCC

In this section, we combine the appearance and speech to improve the twin recognition accuracy. We choose Gabor as feature to represent appearance feature; we choose MFCC as feature to represent voice feature. The reason for our



(a) Fusion of Gabor and MFCC (b) Zoom-in Result for Figure 4(a)

Fig. 4. Performance of Fusion of Gabor and MFCC

choice is trivial, because these two features perform the best in each category in our previous experiment. In multimodal systems, there are three levels of fusion when combining two biometrics. The first is fusion at the feature extraction level. The features for each biometric modality are formed into a new feature. The second is fusion at the confidence level. Each biometric provides a similarity score, and these scores will be combined together to assert the veracity of the claimed identity. The third fusion is at decision level. Each biometric will make one decision and final decision is made based on those decisions.

In our proposal, we use the second fusion strategy. Given a probe and a claim identity, we compute the Euclidean distance of Gabor, denoted as GD, and the likelihood ratio against the claimed identity, denoted LR in Equ 3, separately. The final similarity, FS, is computed as the weighted sum of GD and LR, denoted as $FS = \alpha GD + (1 - \alpha)LR$. Then, we compare the FS against the pre-set threshold ϵ. If $FS > \epsilon$, we accept; otherwise we reject. We conducted the experiments on the whole database, and the performance is showed in Figure 4. From this figure, we can see that when α is set to 0.415, by fusion of Gabor and MFCC the Twin-EER decreases from 0.171(MFCC) to 0.160. We set the α for the best of test performance in our dataset.

5 Conclusion and Future Work

In this work, we collect a moderate size of identical twins database including appearance and voice. We propose to use Gaussian Mixture Model to model the voice spectral pattern and use the ratio of likelihood from two different classes for verification. The experimental results verify that voice biometric can be used to distinguish between identical twins and it is significantly better than traditional facial appearance features, including EigenFace, LBP and Gabor; Among various

voice features, MFCC has the most discriminating ability. We further prove that the accuracy can be improved via fusion of voice biometric and facial appearance.

In future, we would like to test the robustness of our voice 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 voice 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. Martin, J., Kung, H., Mathews, T., Hoyert, D., Strobino, D., Guyer, B., Sutton, S.: Annual summary of vital statistics: 2006. Pediatrics (2008)
2. Sun, Z., Paulino, A., Feng, J., Chai, Z., Tan, T., Jain, A.: A study of multibiometric traits of identical twins. SPIE (2010)
3. Zhang, L., Ye, N., Marroquin, E.M., Guo, D., Sim, T.: New hope for recognizing twins by using facial motion. In: WACV, IEEE (2012) 209–214
4. Jain, A., Prabhakar, S., Pankanti, S.: On the similarity of identical twin fingerprints. Pattern Recognition (2002) 2653–2663
5. Kong, A., Zhang, D., Lu, G.: A study of identical twins' palmprints for personal verification. Pattern Recognition (2006) 2149–2156
6. Nejati, H., Zhang, L., Sim, T., Martinez-Marroquin, E., Dong, G.: Wonder ears: Identification of identical twins from ear images. ICPR (2012) 1201–1204
7. Daugman, J., Downing, C.: Epigenetic randomness, complexity and singularity of human iris patterns. Proceedings of the Royal Society of London (2001) 1737
8. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using gaussian mixture speaker models. Speech and Audio Processing, IEEE Transactions on **3**(1) (1995) 72–83
9. Sinith, M., Salim, A., Gowri Sankar, K., Sandeep Narayanan, K., Soman, V.: A novel method for text-independent speaker identification using mfcc and gmm. In: ICALIP, IEEE (2010) 292–296
10. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. Multimedia, IEEE Transactions on **2**(3) (2000) 141–151
11. Dean, D., Sridharan, S., Wark, T.: Audio-visual speaker verification using continuous fused hmms. In: Proceedings of the HCSNet workshop. (2006) 87–92
12. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: CVPR, IEEE (1991) 586–591
13. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. Computer Vision-ECCV 2004 (2004) 469–481
14. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. Image processing, IEEE Transactions on **11**(4) (2002) 467–476
15. Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A.: Lateralization of phonetic and pitch discrimination in speech processing. Science **256**(5058) (1992) 846–849
16. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America **50** (1971) 637
17. Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval. Volume 28. (2000) 5

18. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society (1977) 1–38
19. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. ECCV (2008) 504–513