Vulnerability of Speaker Verification to Voice Mimicking

Yee Wah Lau, Michael Wagner and Dat Tran

School of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia

ABSTRACT

In this paper, we consider mimicry, a simple technology form of attack requiring a lower level of expertise to investigate whether a speaker recognition system is vulnerable to mimicry by an impostor without using any assistance of other technologies. Experiments on 138 speakers in the YOHO database and two people who played a role as imitators have shown that an impostor can attack the system if that impostor knows a registered speaker in the database who has very similar voice to the impostor's voice.

1. INTRODUCTION

Biometrics offers greater security than traditional methods in person recognition. In particular, voice recognition technology produces relatively low to medium error rate and it has a high public acceptance rate due to unobtrusive nature [1]. Voice dialing, banking over a telephone network, database access services, security control for confidential information, and remote access of computers are important applications of speaker recognition technology. Reducing the false acceptance error rate of impostors has been investigated. Instead, one of the biometric recognition technologies, the computer audio-based speaker recognition system still has a potential of security threat concerns [10][14].

There have been some reports on impostor attack to speaker recognition system. For instance, the sensitivity to computer by voice-altered impostor using trainable speech synthesis technology and the imposture using synthetic speech against speaker verification based on spectrum and pitch were investigated [5][8][9]. In those reports, attacks are categorised as *organised approach* using the substitution method. Speech synthesis system is required to alter the impostor voice in order to attack the system. However, this may not be practical in an operational system where minimum-processing time is required. Therefore, in this paper, we consider mimicry, a simple technology form of attack requiring a lower level of expertise to investigate whether the speaker recognition is vulnerable to mimicry without using any assistance of other technologies. We used a Gaussian mixture model (GMM)-based speaker verification system and the YOHO speaker database for investigation.

2. GAUSSIAN MIXTURE MODELS

Let $X = \{x_1, x_2, ..., x_T\}$ be a set of T vectors, each of which is a d-dimensional feature vector extracted by digital speech signal processing. Assuming a statistical independence between these vectors, the probability of the set X given the model λ can be calculated as follows

$$\log P(X \mid \lambda) = \sum_{t=1}^{T} \log P(x_t \mid \lambda) \tag{1}$$

Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities, which is a weighted sum of c component densities, given by the equation

$$P(x_t \mid \lambda) = \sum_{i=1}^{C} w_i N(x_t, \mu_i, \Sigma_i) \tag{2}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i, i = 1,..., c, are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$, i = 1,..., c, are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\right\}}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \tag{3}$$

In training the Gaussian mixture model (GMM), these parameters are estimated such that in some sense, they best match the distribution of the training vectors. The most widely used training method is the maximum likelihood (ML) estimation. The following re-estimation formulas are used to estimate GMM model parameters [12]

$$\overline{w_i} = \frac{1}{T} \sum_{t=1}^{T} P(i \mid x_t, \lambda) \tag{4}$$

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} P(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{T} P(i \mid x_t, \lambda)} \tag{5}$$

$$\overline{\Sigma}_i = \frac{\sum_{t=1}^{T} P(i \mid x_t, \lambda)(x_t - \overline{\mu}_i)(x_t - \overline{\mu}_i)'}{P(i \mid x_t, \lambda)} \quad (6)$$

3. SPEAKER VERIFICATION

Let λ_0 be the claimed speaker model and λ_i be a model representing all other possible speakers, i.e. impostors. Let $P(X \mid \lambda_0)$ and $P(X \mid \lambda_i)$ be the likelihood functions of the claimed speaker and impostors, respectively. For a given input utterance X and a claimed identity, a claimed speaker's score $S(X)$ is used as follows

$$S(X) \begin{cases} > \theta & accept \\ \leq \theta & reject \end{cases} \quad (7)$$

where θ is the decision threshold and the score $S(X)$ is calculated as [11]

$$S(X) = \log P(X \mid \lambda_0) - \log\left[\frac{1}{B}\sum_{i=1}^{B} P(X \mid \lambda_i)\right] \quad (8)$$

B is the number of background speaker models used to represent the population close to the claimed speaker.

4. EXPERIMENTS

4.1 Speech Database

The YOHO corpus was designed for speaker verification systems in office environments with limited vocabulary. There are 138 speakers, 106 males and 32 females. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97 pronounced as "twenty-one", "ninety-seven", and spoken continuously in sets of three, for example "36-45-89", in each utterance. There are four enrolment sessions per speaker, numbered 1 through 4, and each session contains 24 utterances. There are also ten verification sessions, numbered 1 through 10, and each session contains 4 utterances. All waveforms are low-pass filtered at 3.8 kHz and sampled at 8 kHz.

4.2 Speech Processing and Algorithmic Issues

Speech processing was performed using HTK V2.0, a toolkit [16] for building hidden Markov models (HMMs). The data were processed in 32 ms frames at a frame rate of 10 ms. Frames were Hamming windowed and pre-emphasized. The basic feature set consisted of 12th-order mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames.

GMMs are initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e. $[\sigma_k]_{ii} = \sigma_k^2$ and $[\sigma_k]_{ij} = 0$ if $i \neq j$, where σ_k^2, $1 \leq k \leq K$ are variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [11]. This constraint places a minimum variance value $\sigma_{\min}^2 = 10^{-2}$ on elements of all variance vectors in the GMM in our experiments.

Each speaker was modelled by using 96 training utterances in four enrolment sessions without end-point detection. Error rates therefore were not too low to allow meaningful comparisons between the different background speaker modelling methods for speaker verification. GMMs were trained in text-independent mode.

4.3 Impostors

In this experiment, a male and a female were selected to be imitators to mimic speakers' voice in the YOHO database. Their first language is Chinese and second language is English. Both the two imitators have been in Australia for more than 7 years. They are both amateur imitators and do not have any knowledge of mimicry. They selected a speaker in the YOHO database, listened to that speaker's voice and speaking style, then they attempt to mimic the selected speaker, recorded their voice, save to their database and then use that database to test the system and get the false acceptance error rate using the selected speaker as the claimed speaker.

3.3 Recording Sessions

Firstly, the imitator was required to naturally speak 40 utterances with the same vocabulary used in the YOHO verification sessions. These utterances were used to measure the average similarity scores between the imitator and the same-gender subset of the 138 speaker set in the YOHO database. Let X be the verification data set of the imitator and λ_i be the model of the ith speaker in the YOHO database, the similarity score was the following log-likelihood function

$$S(X) = \log P(X \mid \lambda_i) \quad (9)$$

Based on the 138 scores collected, we chose the closest, intermediate and furthest speakers having highest, medium and lowest scores, respectively.

As those target speakers were identified, the imitator started recording. For each of the 3 target speakers, there were four recording sessions numbered 1 through 4. For each of the four sessions, the imitator listened each of the 40 utterances of the corresponding target speaker once at a time then repeat that utterance with his/her voice and record it. We performed the four recording sessions to consider the improvement of the imitator's mimicry skill after each session. In total, there were 480 (40 utterances x 4 sessions x 3 target speakers) utterances recorded for each of the two imitators.

All the recordings were recorded using a high quality microphone and saved to 16-bit wave files and sampled at 8 kHz. The environment for recording is in a laboratory where the system was set up in a corner of the lab with a medium to low-level noise from the adjoining rooms, air-conditioning system, people walking through the corridor, and the fan in the Sun workstation. The same speech processing in Session 4.2 was applied to process the imitators' voice database.

4. EXPERIMENTAL RESULTS

We used the false acceptance error rate to measure the mimicry level. The higher the false acceptance error rate is, the better the mimicry is. Figure 1 shows false acceptance error rates versus threshold for the female imitator (the imitator 1). We can see that the error rate for session 4 is highest (e.g. at the threshold $\theta = 0.5$) since the imitator 1 has 4 times for listening to improve her mimicry.

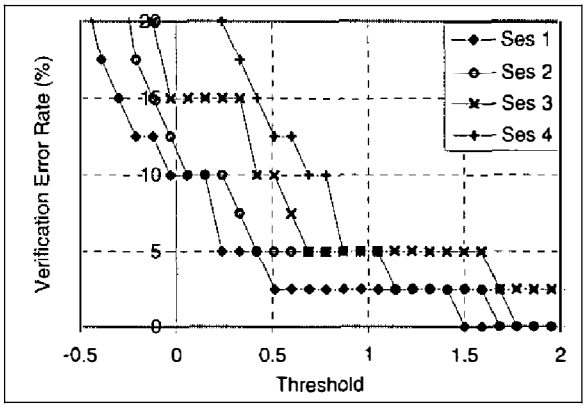

Figure 1: False acceptance error rates (in %) versus threshold for the female imitator (imitator 1).

The three target speakers found for the imitator 1 were the speakers 192 (closest), 160 (intermediate), and 146 (furthest). In Figure 2, the imitator could provide a good mimicry. Consider the error rates at the threshold $\theta = 1.0$. The false acceptance rate obtained by all impostors in the YOHO database is 0%, but the imitator could produce a false acceptance rate of 5%. If the system is preset to the equal error rate threshold, approximately 0.5 as seen in Figure 2, then the imitator 1 could achieve up to 15%. This means that if the imitator 1 logs on as the speaker 192 to the system preset with $\theta = 0.5$, the imitator has 6 out of 40 times (30%) to be accepted by the system.

However as shown in Figure 3, if the imitator logs on to the system as the speaker 135, the imitator 1 cannot be accepted since the false acceptance error rate is lower than the average error rate produced by all impostors in the YOHO database. A similar result was obtained for the furthest target speaker 136.

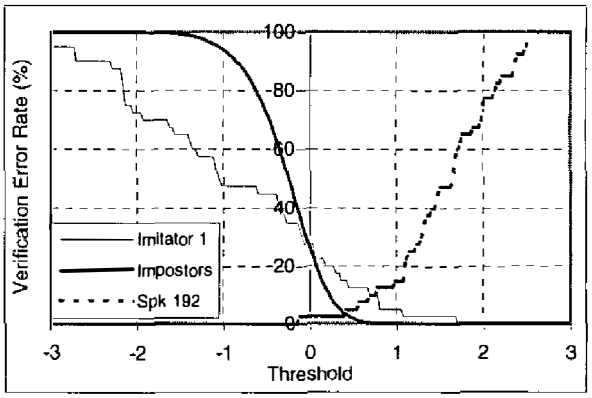

Figure 2: False acceptance error rates (in %) versus threshold for the imitator 1, all impostors and the speaker 192.

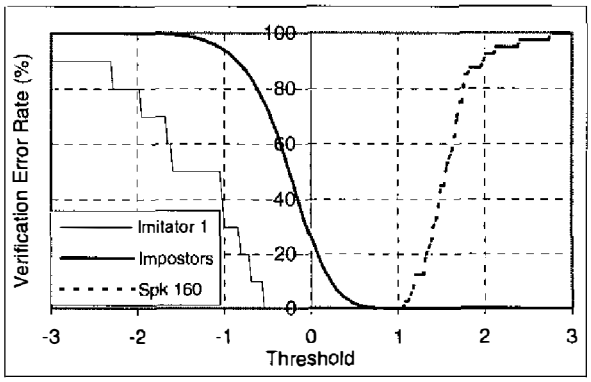

Figure 3: False acceptance error rates (in %) versus threshold for the imitator 1, all impostors and the speaker 160.

For the imitator 2, the three target speakers found were the speakers 239 (closest), 151 (intermediate), and 140 (furthest). A high false acceptance rate was produced by the imitator 2 in Figure 4 when he logged on to the system as the speaker 140. With the threshold $\theta = 1.0$, the false acceptance rate obtained by all impostors in the YOHO database is 0%, but the imitator 2 could produce a false acceptance rate of 20%. If the system is preset to the equal error rate threshold, approximately 0.5, then the imitator 2 could achieve up to 35%.

However as shown in Figure 5, the imitator cannot attack the system when he logs on to the system as the speaker 151. A similar result was obtained for the furthest target speaker 239.

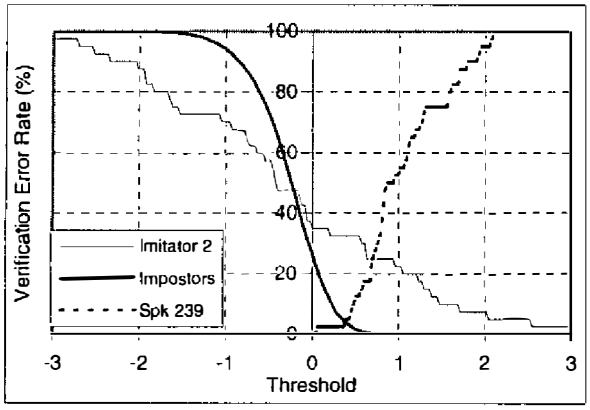

Figure 4: False acceptance error rates (in %) versus threshold for the imitator 2, all impostors and the speaker 239.

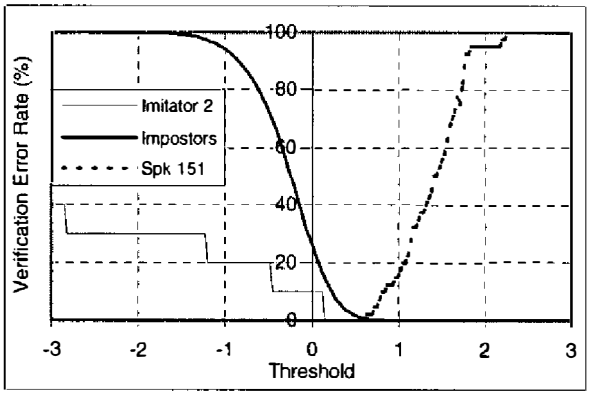

Figure 5: False acceptance error rates (in %) versus threshold for the imitator 2, all impostors and the speaker 151.

5. CONCLUSION

We have considered the mimicry, a simple technology form of attack requiring a lower level of expertise to investigate whether a speaker recognition system is vulnerable to mimicry by an impostor without using any assistance of other technologies. Experimental results showed a fact that a normal person can get a high chance to attack the system if that person knows the closest speaker in the database. If that person does not know who is closest, but if the speaker database is small and the list of speaker names is disclosed, then that person can attack the system by using each speaker name at a time to log on. Doing this way, the person can find the closest speaker and get more chance to attack the system. A notable effect is that choosing speakers in the database whom are close to the client's voice and use their recording for impostor attempts will outperform the re-synthesis they have tried in their experiment [4].

6. REFERENCES

[1] J. Campbell, "Testing with the YOHO CD_ROM voice verification corpus" ICASSP'95, 1995, Vol.1 pp. 341-344.

[2] R.O. Duda and P.E. Hart, "Pattern classification and scene analysis", John Wiley & Sons, 1973.

[3] S. Furui, "Recent advances in speaker recognition", Pattern Recognition Letters, 18, pp. 859-872, 1997.

[4] S. Furui, "An overview of speaker recognition technology", in Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[5] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, M. Plumpe, "Recent Improvements on Microsoft's Trainable Text-to-Speech System: Whistler", in Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing. Munich, Germany. Apr. 1997.

[6] X.D. Huang, Y. Ariki, and M.A. Jack, "Hidden Markov models for speech recognition", Edinburgh University Press, 1990.

[7] B. H. Juang, "The Past, Present, and Future of Speech Processing", IEEE Signal Processing Magazine, 15:3, pp. 24-48, May, 1998.

[8] J Lindberg & M. Blomberg, "Vulnerability in speaker verification. A study of technical impostor techniques" Proc of Eurospeech 99, 1211-1214, 1999.

[9] T. Masuko, K. Tokuda, and T. Kobayashi, "Impostors using Synthetic Speech Against Speaker Verification Based on Spectrum and Pitch", Proc. ICSLP 2000.

[10] B.L. Pellom, Hansen J.H.L., "An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Impostors", IEEE ICASSP-99: Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp.837-840, 1999.

[11] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, vol. 17, pp. 91-108, 1995.

[12] D.A. Reynolds, "A Gaussian mixture modeling approach to text-independent Speaker Identification", PhD thesis, 1993.

[13] T. Satoh, T. Masuko, T. Kobayashi, K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis", Proc. 7th European Conference on Speech Communication and Technology, EUROSPEECH, vol.2, pp.759-762, Aalborg, Denmark, 2001.

[14] M. Takashi, T. Keiichi, K. Takao "Imposture using synthetic speech against speaker verification based on spectrum and pitch", In ICSLP-2000, vol. 2, 302-305, 2000.

[15] D. Tran and M. Wagner, "Fuzzy Gaussian Mixture Models for Speaker Recognition", Australian Journal of Intelligent Information Processing Systems (AJIIPS), vol. 5, no. 4, pp. 293-300, 1998.

[16] P. C. Woodland et. al., "Broadcast news transcription using HTK", in Proceedings of ICASSP, USA, 1997.