THEORETICAL VULNERABILITIES IN MAP SPEAKER ADAPTATION

Tetsushi OHKI, Akira OTSUKA

National Institute of Advanced Industrial Science and Technology (AIST)
2-3-26 Aomi, Koto-ku, Tokyo, JAPAN

ABSTRACT

We analyze the theoretical vulnerability of maximum a posteriori(MAP) speaker adaptation, which is widely used in practical speaker recognition systems. First, we proved that there exist a set of feature vectors, what are called wolves, which can impersonate almost all the registered speakers with probability asymptotically close to 1 with at most two trials. Second, our experiment shows that the wolves with appropriate parameters achieved 0.99 of successful impersonation rate on *Spear* speaker recognition toolkit with ATR speech database.

Index Terms— biometrics, speaker recognition, wolf attack, impersonation, MAP adaptation

1. INTRODUCTION

Security of biometric authentication systems is conventionally measured by the false acceptance rate (FAR), which gives the average error rate, or equivalently successful impersonation rate when imposters present their biometric information. However, it is also possible that imposters present more prospective biometric information taking advantage of the environmental conditions and matching algorithms. Une et al. [1] and Inuma et al. [2] proposed the wolf attack and its success probability (WAP), which is a new measure that takes such attacks into account. A wolf is the worst biometric information that can impersonate any registered users with a probability much higher than FAR.

The subject of this paper is to investigate the vulnerability of practical automatic speaker verification(ASV) systems, specifically, maximum a posteriori (MAP) adaptation based ASV systems, which derive the speaker model by updating the well-trained parameters in the Universal Background Model (UBM) via adaptation. There are many previous works involving spoofing through voice synthesis [3, 4, 5] and voice conversion [6, 7, 8, 9, 10]. All of these works have not evaluated the wolf attack. The main contributions of this work are as follows:

1. We proved the existence of an asymptotically universal wolf which can impersonate all registered speakers in MAP adaptation based ASV algorithms with the usual linear scoring techniques.
2. Our wolf attack can achieve 99% of successful impersonation rate (WAP) at most two trials even when speaker model has a large number of Gaussian components. This is experimentally examined using open source speaker recognition toolkit *Spear*[1] [11] along with ATR multi-speech database [12].

2. PREVIOUS WORKS

The name, wolf, came from the well-known Doddington Zoo [13], which refers to a biometric information which tends to impersonate others. The wolf attack probability is the maximum false acceptance rate when a single live or fake biometric information is presented [2, 1]. In other words, the wolf attack probability (WAP) gives the upper-bound of successful impersonation probability for any subversive presentation without stealing the victim's biometric feature. Une et al. [1] showed WAP=1.0 for a finger vein authentication algorithm [14]. Ohki et al. [15] experimentally showed WAP=0.92 for vector quantization-based ASV systems when up to 3 trials are allowed for failed attempts. Other previous results show the vulnerabilities in more general classifiers. Kryszczuk and Drygajlo [16] reported that likelihood-ratio-based face classifiers are easily spoofed by the face images to which white Gaussian noise is superimposed. The noise contaminated face images give feature vectors with extremely small likelihood values to a claimed user model and the world model (UBM) simultaneously. They reported [16] "above 50% of noise almost every claim is accepted." Ohki and Otsuka [17] investigated more general vulnerability in the likelihood-baed classifiers where the likelihood values are not necessarily small. They focused on the approximation error between UBM and true distribution of over the feature space. Their experiments showed WAP=0.6 for the GMM-UBM ASV system [18].

In the context of speaker verification, Reynolds [19] showed that maximum a posteriori (MAP) adaptation based Gaussian mixture modeling provides high-accuracy speaker verification even under the situation where a few samples are available to create each speaker model. MAP adaptation derives each speaker model from UBM. Thus, it inherently

[1] http://pypi.python.org/pypi/bob.spear

provides homogeneous models among the individual speaker models and UBM. In the following section, we investigate the theoretical vulnerability in the MAP adaptation-based ASV systems [19].

3. VULNERABILITY IN MAP ADAPTATION

3.1. Map Adaptation

Given a UBM as M-mixture of D-dimensional Gaussian distribution over feature space $\mathcal{X}$ s.t.

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

where w_i are the mixture weights sum to unity, $p_i(\cdot)$ are the Gaussian probability functions given below with μ_i and Σ_i being the mean and the covariance matrix respectively.

Let $\mathcal{U}$ be a set of speakers. A training vector from the hypothesized speaker $u \in \mathcal{U}$ is denoted by $X^u = \{x_1^u, \ldots, x_T^u\}$. A speaker model $\widehat{p}(x|u)$ is adapted from the UBM as follows.

For mixture i in the UBM, we compute

$$\Pr(i \,|\, x_t^u) = \frac{w_i p_i(x_t^u)}{\sum_{j=1}^{M} w_j p_j(x_t^u)}. \tag{2}$$

Using this $\Pr(i \,|\, x_t^u)$ and x_t^u, we compute the weight, mean and variance parameters for the expectation step.

$$n_i = \sum_{t=1}^{T} \Pr(i \,|\, x_t^u) \tag{3}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \,|\, x_t^u) x_t^u \tag{4}$$

$$E_i(xx^\top) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \,|\, x_t^u) x_t^u x_t^{u\top} \tag{5}$$

where $E_i(\cdot)$ computes the expectation weighted by $\Pr(i \,|\, x_t^u)$. In the adaptation step, we update the old UBM statistics to create adapted parameters with the following equations.

$$\begin{aligned}
\widehat{w_i^u} &= [\alpha_i^w n_i/T + (1-\alpha_i^w) w_i]\gamma \\
\widehat{\mu_i^u} &= \alpha_i^m E_i(x) + (1-\alpha_i^m)\mu_i \\
\widehat{\Sigma_i^u} &= \alpha_i^v E_i(xx^\top) + (1-\alpha_i^v)(\Sigma_i + \mu_i\mu_i^\top) - \widehat{\mu^u}_i \widehat{\mu^u}_i^\top
\end{aligned}$$

Reynolds et al. [19] introduced the adaptation coefficients $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ which control the balance between old and new estimates. The scale factor γ is introduced to ensure the mixture weights sum to unity. $\alpha_i^v = 0$ is called linear scoring where covariance matrices are invariant such that $\widehat{\Sigma_i^u} = \Sigma_i$. It is widely used for its significant reduction in computation. With the MAP adapted parameters for the individual model $(\widehat{w_i^u}, \widehat{\mu_i^u}, \widehat{\sigma_i^u})$, the model for the speaker u is given as

$$\widehat{p}(x|u) = \sum_{i=1}^{M} \widehat{w_i^u}\, \widehat{p}_i(x|u), \tag{6}$$

where $\widehat{p}_i(x \,|\, u) = \mathcal{N}(x \,|\, \mu_i^u, \Sigma_i^u)$.

3.2. Likelihood Ratio

Likelihood-ratio-based decision is the following two-class hypothesis test. According to Neyman-Pearson lemma [20], this is known as the most powerful.

$$L_u(x) = \frac{p(x \,|\, u)}{p(x)} \tag{7}$$

where we assume $p(x \,|\, \overline{u}) \approx p(x)$ with $\overline{u} = \mathcal{U} \setminus \{u\}$ for a sufficiently large speaker set $\mathcal{U}$.

In the MAP adaptation, each speaker model is adapted from the UBM learning the training vector X^u. Thus, the real decision is made by the following equation using Eq. (6),

$$\widehat{L_u}(x) = \frac{\widehat{p}(x \,|\, u)}{p(x)}. \tag{8}$$

3.3. Wolf in Linear Scoring

We define a wolf as a biometric feature which gives higher acceptance rate, say $\delta \geq \mathrm{FAR}$, than the normal features.

Definition 1. *(δ-wolf) Given a probability distributions $p(x)$ representing UBM and $p(x \,|\, u)$ representing speaker models for all $u \in \mathcal{U}$, and given a threshold $t > 0$, for any $\delta \geq \mathrm{FAR}$, a feature vector x is said to be δ-wolf if it satisfies*

$$\mathrm{AR}_t(x) = \Pr[\widehat{L_u}(x) > t] \geq \delta \tag{9}$$

where probability is taken over $u \in \mathcal{U}$.

Lemma 1. *Given a feature vector $v \in \mathcal{X}$ with $\|v\| = 1$ and a speaker $u \in \mathcal{U}$, compute the maximum standard deviation $\widehat{\sigma^u}$ and σ projected on the feature vector v as follows.*

$$\widehat{\sigma^u}^{-2} = \min_{1 \leq i \leq M} v^\top \widehat{\Sigma_i^u}^{-1} v \tag{10}$$

$$\sigma^{-2} = \min_{1 \leq i \leq M} v^\top \Sigma_i^{-1} v \tag{11}$$

If $\widehat{\sigma^u} > \sigma$ is satisfied, then for any threshold $t > 0$, there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$,

$$\widehat{L_u}(\lambda v) > t. \tag{12}$$

Proof. (Sketch) We take the log of $\widehat{L_u}$, which leads to

$$\log \widehat{L_u}(\lambda v) \approx \frac{-\lambda^2}{2} \left(\frac{1}{\widehat{\sigma^u}^2} - \frac{1}{\sigma^2} \right) + o(1). \tag{13}$$

This is a convex function of λ under the condition $\widehat{\sigma^u} > \sigma$. Hence the lemma. □

Lemma 1 holds for the general MAP adaptation. It implies that injecting sufficiently large vectors in a speaker voice sample will make impersonation successful regardless of the threshold. This result cannot directly apply to the linear scoring case where $\widehat{\sigma^u} = \sigma$ for all $u \in \mathcal{U}$. Interestingly, however, similar analysis can lead to 0.5-wolf as stated below.

Theorem 1. *In the case of linear scoring in MAP adaptation, there exists $\lambda_t > 0$ for all threshold $t \in \mathbb{R}$ such that exactly half of the feature vectors in a set $\{x \in \mathcal{X} \mid \|x\| \geq \lambda_t\}$ are 0.5-wolf. Namely,*

$$\Pr[\widehat{L_u}(x) > t] \geq 0.5 \tag{14}$$

where the probability is taken over $u \in \mathcal{U}$. More precisely, every vector $x \in \{y \in \mathcal{X} \mid \|y\| \geq \lambda_t\}$ satisfies the following.

$$\Pr[\widehat{L_u}(x) > t] + \Pr[\widehat{L_u}(-x) > t] = 1 \tag{15}$$

Proof. From the definition, $\widehat{\Sigma_i^u} = \Sigma_i$ for all $u \in \mathcal{U}$ in linear scoring. Apparently, $\widehat{\sigma^u} = \sigma$ holds for all $u \in \mathcal{U}$ from Eq.(10) and Eq.(11). Given v, let i give the minimum in Eq. (11). For brevity, we write

$$\sigma^{-2} = \widehat{\sigma^u}^{-2} = v^\top \Sigma_i^{-1} v. \tag{16}$$

Let $x = \lambda v$ with $\lambda > 0$. $\log \widehat{L_u}(\lambda v)$ can be written as follows.

$$\begin{aligned}
\log \widehat{L_u}(\lambda v) &\approx \frac{-(\lambda v - \widehat{\mu_i^u})^\top \Sigma_i^{-1} (\lambda v - \widehat{\mu_i^u})}{2} \\
&\quad - \frac{-(\lambda v - \mu_i)^\top \Sigma_i^{-1} (\lambda v - \mu_i)}{2} + o(1) \\
&= -(\widehat{\mu_i^u} - \mu_i)^\top \Sigma_i^{-1} (\lambda v - \frac{\mu_i + \widehat{\mu_i^u}}{2}) + o(1) \\
&= -\lambda (\widehat{\mu_i^u} - \mu_i)^\top \Sigma_i^{-1} v + o(1)
\end{aligned} \tag{17}$$

This is a linear function of λ. Set λ_t as[1]

$$\lambda_t > \max_{u \in \mathcal{U}} \left| \frac{\log t + o(1)}{(\mu_i - \widehat{\mu_i^u})^\top \Sigma_i^{-1} v} \right|, \tag{18}$$

Then, for all $x \in \{y \in \mathcal{X} \mid \|y\| \geq \lambda_t\}$, each $u' \in \mathcal{U}$ satisfies $\widehat{L_{u'}}(x) > t$ or $\widehat{L_{u'}}(-x) > t$. Thus, a vector x divides $\mathcal{U}$ into

$$\mathcal{U}_x^+ = \{u' \mid \widehat{L_{u'}}(x) > t\} \text{ and } \mathcal{U}_x^- = \{u' \mid \widehat{L_{u'}}(-x) > t\}.$$

It is easily proved that $\mathcal{U} = \mathcal{U}_x^+ \cup \mathcal{U}_x^-$ and $\mathcal{U}_x^+ \cap \mathcal{U}_x^- = \emptyset$. Thus,

$$\Pr[\widehat{L_u}(x) > t] = \frac{|\mathcal{U}_x^+|}{|\mathcal{U}|}, \quad \Pr[\widehat{L_u}(-x) > t] = \frac{|\mathcal{U}_x^-|}{|\mathcal{U}|} \tag{19}$$

where the probability is taken over $u \in \mathcal{U}$. Eq. (19) directly implies Eq. (15). Further, we see the complementary relation in $\mathcal{U}_x^+ = \mathcal{U}_{-x}^-$ and $\mathcal{U}_x^- = \mathcal{U}_{-x}^+$. This implies that exactly half of the set satisfies Eq. (14). Hence the theorem. □

[1] We assume $(\mu_i - \widehat{\mu_i^u})^\top \Sigma_i^{-1} v \neq 0$ for all $u \in \mathcal{U}$ for brevity of proof. From performance, $\mu_i \neq \widehat{\mu_i^u}$ for all u. Choosing orthogonal v is negiligible.

4. EXPERIMENT

4.1. Wolf Attack

While the discussion on the theoretical vulnerability of an ASV system, we assume the ideal ASV system where all acoustic equipment and environment have ideal characteristics so that an attacker can control a feature vector received by the system completely as intended. Namely, the attacker can input an arbitrary biometric feature to a matching algorithm.

From Theorem 1, a feature vector $x \in \mathcal{X}$ with sufficiently larger norm than λ_t is a 0.5-*wolf*. One can create such a wolf feature by sampling a feature vector x from a training set, say X, and scale it by a fixed large constant C. Our wolf feature vector $V_W \in \mathcal{X}$ is simply constructed as follows:

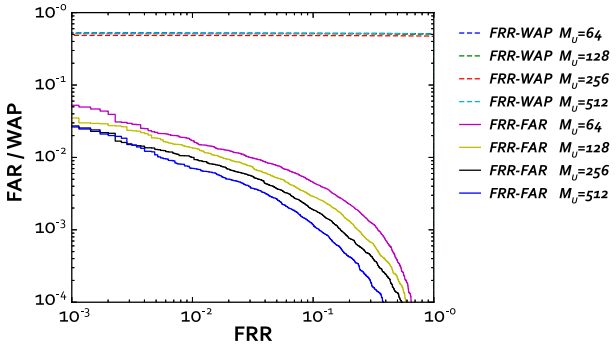
$$V_W = Cx \quad \text{for every sampled } x \in X \tag{20}$$

We define two types of attack samples as a sequence of feature vectors of length F. The first type is called "full-frame wolf," which is comprised of V_W in every occurrence of a feature vector. The second type is called "single-frame wolf," which is comprised of the real feature vectors but exactly one of the occurrences is replaced with V_W. In our experiment, we selected the feature vector x randomly from UBM mean vectors. Our experiment shows that $C \leq 1000$ achieved sufficiently high WAP at all settings. Furthermore, we examined the impact of double trials using both positive V_W and negative $-V_W$ wolf features as described in 3.3. This is only applicable if the targeted ASV system allows multiple trials.

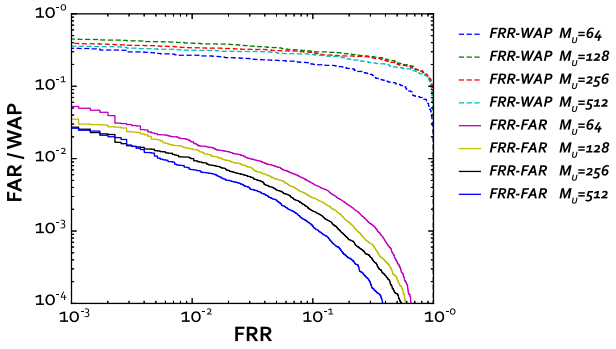
4.2. Experimental Setup

We evaluated the proposed wolf attack using the speech data included in the ATR multi-speaker database [12]. The database is a collection of short time 16kHz sampled utterances of 294 male speakers and 410 female speakers. For each speaker there are 50 utterances of approximately 4 seconds. We used 200 male and 200 female to keep the ratio of male to female speaker to the same for overall evaluation. We conduct our wolf attack experiments on *Spear*, an open source toolbox for speaker recognition. *Spear* implements a set of complete speaker recognition toolchain including MAP adaptation based GMM-UBM algorithm. *Spear* filters out the non-speech part using VAD algorithm [21] and calculates 60 coefficients MFCC for each frame.

The evaluation was conducted using 200 speakers as the UBM training set and 100 speakers as the development set and remaining 100 speakers as the evaluation set. For the development and the evaluation set, we separated each speaker's 50 utterances to 15 utterances for speaker model training and 35 utterances for score calculation. UBM model order M_U was varied in the range 64, 128, 256, and 512 components. Since the speaker model is derived by updating the well-



(a) full-frame attack ($C = 100$, $F = 400$)



(b) single-frame attack ($C = 1000$)

Fig. 1: Wolf attack probability in proposed wolf attack. DET curve of wolf attacks drawn by FRR and WAP were labelled as "FRR-WAP". DET curve of baseline system drawn by FRR and FAR was labeled as "FRR-FAR".

trained parameters in the UBM via adaptation, speaker model order is always equivalent to the UBM model order.

We evaluated two types of wolf attack described in 4.1. For the full-frame attack, we set the frame length F to 400 frames. Constant factor C is set to 100 for full-frame attack and 1000 for single-frame attack, respectively.

4.3. Experimental Results

Fig.1(a), (b) shows the wolf attack probability calculated in our proposed wolf attack along with baseline performance. In this evaluation, DET curve was drawn between the FRR and WAP instead of FAR. Note that DET curve indicates better performance when it is closer to the lower-left corner of the graph. In other words, wolf attack is more efficient when it is closer to the upper-right corner of the graph. As shown in both figures, we can see that the proposed wolf achieved clearly high impersonation performance.

Let $t = t_{EER}$ be the threshold that can achieve Equal Error Rate (EER) for the baseline performance. Table 1 indicates EER, WAP and $\overline{\text{WAP}}$ when threshold t was set to t_{EER}. $\overline{\text{WAP}}$ is calculated using the development set. In Table 1, we also show the WAP in case the targeted system allows at most two trials. In this experiment, imposter uses V_W and $-V_W$ for the two trials.

Table 1: List of EER, WAP, $\overline{\text{WAP}}$ (threshold t was fixed to t_{EER}). second trial shows a WAP at most two trials.

		UBM / Speaker Model Order			
		64	128	256	512
		Baseline Performance			
EER		0.014	0.012	0.009	0.008
	trial	full-frame attack			
WAP	first	0.52	0.51	0.48	0.51
	second	0.99	0.98	0.99	0.99
$\overline{\text{WAP}}$	first	0.52	0.50	0.51	0.51
	second	0.99	0.99	0.99	0.98
	trial	single-frame attack			
WAP	first	0.28	0.39	0.34	0.31
	second	0.46	0.54	0.69	0.53
$\overline{\text{WAP}}$	first	0.28	0.38	0.37	0.29
	second	0.47	0.57	0.69	0.49

As shown in Table 1, the full-frame attack can achieve almost 50% of WAP regardless of the number of Gaussian components. Moreover, it achieved 99% of WAP when targeted system allows at most two trials. These results confirms the existence of 0.5-*wolf* proved in the Theorem 1. Additionally, we can see that single-frame attack achieved 28% to 39% in the first trial and 46% to 69% of WAP in the second trial. This fact means that an attacker can impersonate to the targeted system with quite high probability using the voice which contains only single frame (10ms) of wolf sample.

5. CONCLUSION

In this study, we investigated theoretical vulnerabilities in MAP adaptation based speaker recognition system. Our remarkable result is that the proposed wolf attack achieved 99% of successful impersonation with at most two trials on *Spear* toolkit and ATR speech database even when the speaker model has a large number of Gaussian components. Since we assumed the ideal attacker in this study, it is left to investigate how to apply our wolf attack to microphone or transmission level spoofing attacks such as voice conversion attacks. Possible countermeasures may include that UBM takes larger variance than individual speaker models in MAP adaptation.

6. REFERENCES

[1] M. Une and A. Otsuka, "Wolf attack probability: a new security measure in biometric authentication systems," *Lee, S.-W., Li, S.Z.(eds.) ICB 2007, LNCS*, vol. 4642, pp. 396–406, 2007.

[2] M. Inuma and H. Otsuka, A. Imai, "Theoretical framework for constructing matching algorithms in biometric authentication systems," in *Proc. Third IAPR/IEEE International Conference on Biometrics (ICB 2009)*. Sept. 2009, vol. LNCS 5558, pp. 293–300, Springer-Verlag.

[3] Takashi Masuko Keiichi, Keiichi Tokuda, and Takao Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. the International Conference on Spoken Language Processing*, 2000, pp. 302–305.

[4] P. L. De Leon, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[5] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[6] P. Z. Patrick, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using alisp: Indexation in a client memory," in *Proc. 2005 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2005, vol. 1, pp. 17–20.

[7] D. Matrouf, J. F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Proc. 2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, May 2006, vol. 1.

[8] T. Kinnunen, Z. Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," March 2012.

[9] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept 2013, pp. 1–8.

[10] Z. Wu, A. Larcher, K. A. Lee, E. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints.," in *Proc. Interspeech*, 2013.

[11] E. Khoury, L. El Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on Bob," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[12] "ATR Promotions Inc. multi speaker database: SDB," `http://www.atr-p.com/sdb.html`.

[13] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheeps, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Proc. ICSLD 98*, Nov. 1998.

[14] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles," in *Proc. the 9th IAPR Conference on Machine Vision Applications*, May 2005, pp. 347–350.

[15] T. Ohki, S. Hidano, and T. Takehisa, "Evaluation of wolf attack for classified target on speaker verification systems.," in *Proc. International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2012, pp. 182–187.

[16] Krzysztof Kryszczuk and Andrzej Drygajlo, "Addressing the vulnerabilities of likelihood-ratio-based face verification," in *Proc. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2005, pp. 426–435.

[17] T. Ohki and A. Otsuka, "Theoretical vulnerability in likelihood-ratio-based biometric verification," in *Proc. 2014 IEEE International Joint Conference on Biometrics (IJCB)*, Sept 2014, pp. 1–8.

[18] Douglas A Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[19] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[20] Jerzy. Neyman and E. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Trans. the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[21] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7229–7233.