INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Adversarial Optimization for Dictionary Attacks on Speaker Verification

Mirko Marras[1], *Paweł Korus*[2,3], *Nasir Memon*[2], *Gianni Fenu*[1]

[1]University of Cagliari, Cagliari, Italy
[2]New York University, New York, U.S.A.
[3]AGH University of Science and Technology, Krakow, Poland
{mirko.marras, fenu}@unica.it, {pkorus, memon}@nyu.edu

Abstract

In this paper, we assess vulnerability of speaker verification systems to dictionary attacks. We seek master voices, i.e., adversarial utterances optimized to match against a large number of users by pure chance. First, we perform menagerie analysis to identify utterances which intrinsically hold this property. Then, we propose an adversarial optimization approach for generating master voices synthetically. Our experiments show that, even in the most secure configuration, on average, a master voice can match approx. 20% of females and 10% of males without any knowledge about the population. We demonstrate that dictionary attacks should be considered as a feasible threat model for sensitive and high-stakes deployments of speaker verification.

Index Terms: Adversarial Examples, Authentication, Biometrics, Dictionary Attacks, Speaker Verification.

1. Introduction

Biometric technologies provide user profiling services based on physical and behavioral traits. In a lot of use cases, they offer a better user experience than traditional practices [1, 2, 3, 4]. While voice is one of the most analyzed biometric sources in diverse applications [5, 6], its usage for authentication shows vulnerability to impersonation attacks [7], e.g., spoofing [8], replay [9], synthesis [10] and transformation [11]. These attacks require speech samples of the victim, and their collection can greatly vary in difficulty based on the person's public presence.

In this study, we demonstrate the feasibility of dictionary attacks on the voice modality, which allows for targeting large populations without specific knowledge of the individuals or their speech models. Such attacks, recently demonstrated for fingerprints [12, 13], rely on the necessary usability-security trade-offs in mass deployments (e.g., only partial scans of multiple independent finger impressions) and stand in stark contrast to the prevailing individual-targeted attacks [14]. The widely-known menagerie analysis already hints at this vulnerability - it shows that certain individuals act as *wolves* and can match a lot of users. This suggests the existence of a potentially large family of *master voices* (MVs) which match large populations by chance with high probability. Such MVs don't necessarily correspond to a particular person's voice, or even to human speech.

Our results show that adversarial attacks on modern speaker verification systems allow for effectively seeking MVs that generalize between user populations. In our experiments, even in the most conservative setting, on average, a MV could match ≈10% males and ≈20% females within a single presentation attempt. This suggests that dictionary attacks should be considered as a valid threat model for speaker verification systems.

2. Related Work

Speaker Recognition. Speaker recognition has recently undergone a revolution thanks to deep-learned acoustic representations [15]. Best hand-crafted solutions rely on Gaussian mixture models (GMMs) [16] trained on low dimensional feature vectors, joint factor analysis (JFA) [17], or i-Vectors [18]. Modern systems learn to extract effective acoustic feature vectors from one of the last layers of deep neural networks (DNNs) trained for standard or one-shot speaker classification. The most prominent examples include d-Vectors [19], c-Vectors [20], x-Vectors [21], VGGVox-Vectors [22], and ResNet-Vectors [23]. Researchers also explore end-to-end training of such systems [24].

Speaker *verification* aims to confirm or refute the expected identify of the speaker based on an enrolled speech model. The user is asked to provide several samples of his speech, and the utterances are then stored as a collection of acoustic feature vectors. Depending on the verification policy, the presented input may be compared with all of the collected vectors [25], or with a single combined vector [19, 20]. While this makes the system more robust and usable, it may compromise the overall security.

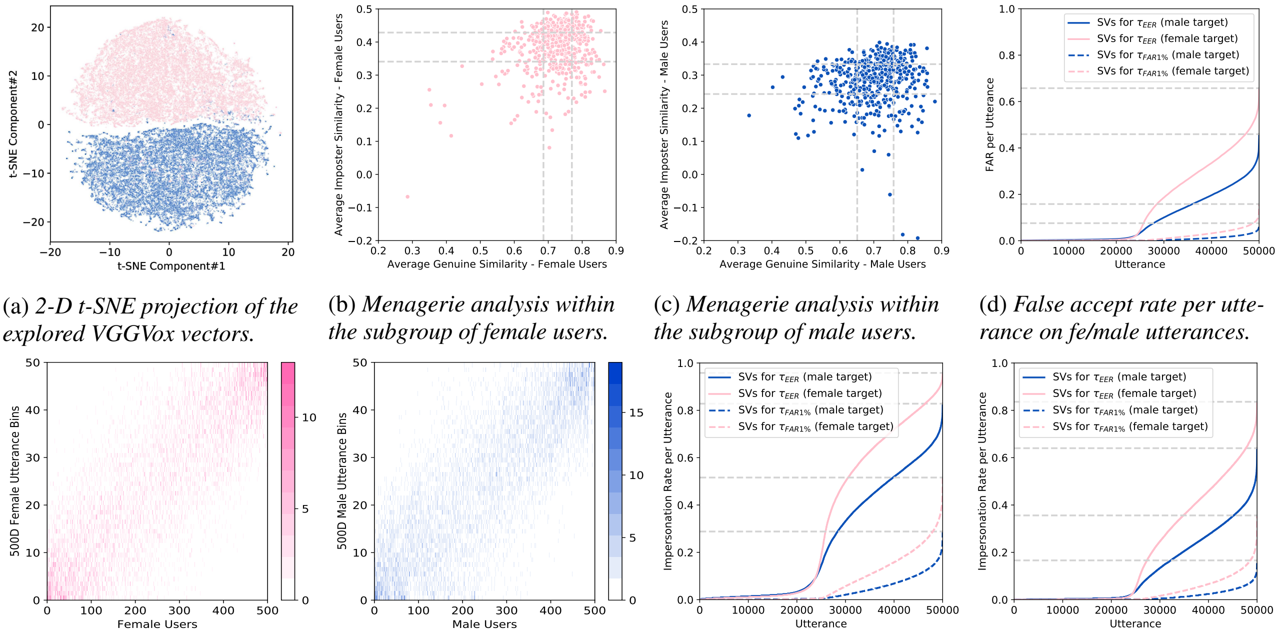
Adversarial Machine Learning. End-to-end optimization of complex systems has spawned a new class of attacks based on *adversarial perturbations*, which involve direct optimization of the model inputs to trigger prediction errors [26]. Such adversarial examples tend to be perceptually indistinguishable from benign inputs, but contain carefully crafted noise which triggers misclassification. The problem has recently attracted massive interest of the machine learning and security communities [27, 28]. Known defenses range from detection of adversarial inputs [29] to training of more secure models [30].

Currently, however, the bulk of the research focuses on computer vision [31]. Adversarial attacks on audio focused on speech recognition [32, 33, 34], and aim to trigger malicious behavior, e.g., through the use of hidden voice commands [35]. Susceptibility of speaker verification has gained attention only in the past year, and currently mirrors the spoofing attack, where a specific user is targeted by the adversarial example [36]. We believe it is imperative to consider a broader spectrum of attacks, including dictionary attacks against larger populations.

3. Problem Statement

Let $A \subset \mathbb{R}^*$ denote the domain of audio waveforms with unknown length. We consider a traditional two-step processing pipeline with an intermediate visual acoustic representation $S \subset \mathbb{R}^{k \times *}$ (e.g., a spectrogram), and an explicit feature extraction step which produces fixed-length representations in $D \subset \mathbb{R}^e$. We denote the respective stages as $\mathcal{F} : A \rightarrow S$ and $\mathcal{D} : S \rightarrow D$. Given a *verification policy* p, a *decision threshold* τ, and N *enrolled utterances* per user, a speaker verification

http://dx.doi.org/10.21437/Interspeech.2019-2430



(a) *2-D t-SNE projection of the explored VGGVox vectors.*

(b) *Menagerie analysis within the subgroup of female users.*

(c) *Menagerie analysis within the subgroup of male users.*

(d) *False accept rate per utterance on fe/male utterances.*

(e) *Female user ranking based on utterance false accept rates.*

(f) *Male user ranking based on utterance false accept rates.*

(g) *Impersonation rate per utterance with Any10 policy.*

(h) *Impersonation rate per utterance with Avg10 policy.*

Figure 1: *Exploratory menagerie analysis aimed to investigate the existence of master voices on the sampled population $A_{VC2-Train}$.*

system can be defined as a function:

$$v_{p,\tau} : D \times D_u^N \to \{0, 1\} \tag{1}$$

which compares an input feature vector d from an unknown user with a set of enrolled feature vectors $d_u^1, ..., d_u^N$ from user u to confirm or refute the speaker's identity (1 and 0, respectively). We consider two verification policies p, *AnyN* [25] and *AvgN* [19, 20], which rely on a similarity function $\mathcal{S} : D \times D \to \mathbb{R}$:

$$v_{p,\tau} = \begin{cases} any\left(\{\mathcal{S}(d, d_u^i) > \tau : i \in 1, \ldots, N\}\right) & \text{if p} = AnyN \\ \mathcal{S}\left(d, \frac{1}{N}\sum_{i=1}^{N} d_u^i\right) > \tau & \text{if p} = AvgN \end{cases}$$

Hence, finding master voices becomes an optimization problem, which aims to find audio waveforms maximizing the following objective function given a population of users U:

$$\tilde{a} = \underset{a}{\operatorname{argmax}} \underset{u \in U}{\mathbb{E}} \left[v_{p,\tau}\left(\mathcal{D}(\mathcal{F}(a)), D_u^N\right)\right] \tag{2}$$

3.1. Verification System and Data Sets

In this study, we leveraged the VoxCeleb data sets [22, 23] one of the largest corpora for speaker verification and identification. It includes unconstrained speech of celebrities, extracted from public videos, and featuring diverse acoustic environments.

All waveforms were single-channel, 16-bit recordings sampled at 16 kHz. As acoustic feature extraction $\mathcal{F}$, we computed magnitude spectrograms with a sliding window of size $k = 512$ samples, and a stride of 160 samples. We applied the Hamming window of 512 samples. Then, mean and variance normalisation was performed on every frequency bin. As feature extractor $\mathcal{D}$, we used the VGGVox model [22] pre-trained on the train portion of the first version of the data set $A_{\text{VC1-Train}}$ (1,211 speakers) and validated on the test portion of the same data set $A_{\text{VC1-Test}}$. The model extracts a $e = 1024$-dimensional representation from each $512 \times *$ spectrogram. As similarity function $\mathcal{S}$, we used the cosine similarity. When we applied a policy, we sampled $N = 10$ utterances per user for enrollment. The selected value N can represent a good trade-off between long and short enrolment lists employed by the existing literature.

For our experiments on master voices, we used the second version of VoxCeleb [23]. From it, we sampled two disjoint populations, i.e., $A_{\text{VC2-Train}}$ for exploration and training, and $A_{\text{VC2-Test}}$ for testing. Each included 1,000 speakers, equally divided between the sexes. Each individual was represented by 50 utterances, leading to the total of 50,000 utterances with duration between 4 and 10 seconds for each population.

The selected verification pipeline achieves an Equal Error Rate (EER) of 8% on utterance-utterance pairs from the validation set $A_{\text{VC1-Test}}$ (consistent with results reported in [22]), which increases to 11.2% on our sampled population $A_{\text{VC2-Train}}$. Based on the latter evaluation, we chose two global decision thresholds for future experiments: $\tau_{\text{EER}} = 0.53$ and $\tau_{\text{FAR1\%}} = 0.74$, which correspond to the Equal Error Rate and False Acceptance Rate (FAR) of 1%, respectively.

4. Exploratory Menagerie Analysis

Our first step was to perform exploratory menagerie analysis to assess prevalence of naturally occurring wolves [37] - potential candidates for master voices. We conducted the experiments on $A_{\text{VC2-Train}}$ for male and female speakers separately, since they exhibit distinct characteristics, which is confirmed in feature domain $\mathcal{D}$ (see Fig. 1a for a 2-D t-SNE projection [38]).

For each user, we first computed the average *genuine score*, which represents how well they match against themselves, and the average *imposter score* indicating how well they match against others (Fig. 1b and 1c). Each point in the scatter plots

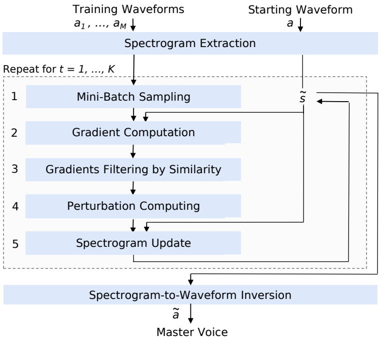


Figure 2: *The proposed approach for generating a master voice.*

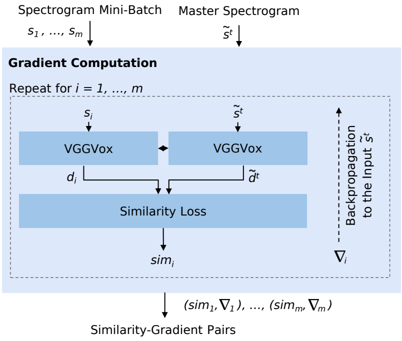


Figure 3: *The core of the Gradient Computation module.*

represents a user. Intuitively, to find good master voices, we are interested in people in the top of the graphs since they cause a disproportionate number of false acceptances. Interestingly, the model revealed a bias against women, who exhibit significantly greater intra-sex average imposter similarity than men.

To investigate this phenomenon, we resorted to utterance-level analysis. In Fig. 1d, we show false acceptance rates for utterance-to-utterance matching targeting a specific gender. The x axis is ordered by FAR, which leads to nearly no errors at the beginning, and a sudden deterioration in the middle, which corresponds to unlikely erroneous intra-sex matches. On the end, the plot clearly shows the existence of outliers, and significant differences between males and females. The plot also reveals that, with the commonly used equal-error-rate threshold, the female wolves can match over 60% of their peers' utterances.

We also assessed the consistency of individual speakers to produce easily confusible utterances. The results are shown in Fig. 1e and 1f for female and male speakers, respectively. For each gender, we ranked the utterances belonging to that gender by decreasing FAR and we grouped them in groups of 500 utterances based on their position in the ranking (i.e., the top-500 utterances with highest FAR belong to the first group and so on). For each user, we then counted how many of her/his utterances belong to each group. Users who generate high FARs have a lot of utterances in the top groups (bottom-left of the plots), while users with less impersonation power have utterances in the last groups (top-right of the plots). While some users are prone to produce high FARs, their utterances are scattered across groups. Hence, we can conclude that while individual speaker properties matter, there is a content-related component which inhibits attacks in challenge-response scenarios.

Finally, we assessed the impact of different verification policies (Fig. 1g and 1h). For each enrolled utterance and user u, we match other users based on their enrollment set, verification policy, and decision threshold τ. With the *Any*10 policy, we observed utterances capable of impersonating between 80% and 90% of the users for τ_{EER} and between 20% and 35% for $\tau_{FAR1\%}$. The results were only slightly worse for the *Avg*10 policy, where we observed impersonation rates between 60% and 80% for τ_{EER} and between 10% and 25% for $\tau_{\text{FAR1\%}}$.

The results indicate that naturally occurring wolves are good candidates for MVs[1]. However, speech content and background noise have strong influence as well. Thus, we explore adversarial perturbations for seeking effective MVs.

5. Master Voice Optimization

The goal of the optimization process is to generate a *master voice waveform* which maximizes the expected FAR (2). Given an existing *seed waveform* a and a set of M training waveforms A_{Train} of a large user population U, we seek $\tilde{a}$, which maximizes $\sum_{u \in U} v_{p,\tau}\left(\mathcal{D}(\mathcal{F}(\tilde{a})), D_u^N\right)$. We will later show empirically that, although the optimization relies on a given population, the results are fully generalizable to unseen individuals.

5.1. Optimization Procedure

Our optimization process seeks adversarial perturbations $\tilde{s}$ of the selected *input spectrogram* $s = \mathcal{F}(a)$, and relies on an iterated stochastic gradient descent (Fig. 2). By slightly perturbing the input spectrogram, we are able to make it more and more similar to an increasing number of training spectrograms, and bias the verification choices the system makes towards higher FAR. The following steps are repeated in every iteration t:

1. **Mini-Batch Sampling**. We sample a batch of m spectrograms $S_{\text{batch}} \leftarrow \{\mathcal{F}(a) : a \in A_{\text{Train}}\}$ with $m << M$.
2. **Gradient Computation**. We pair the current iteration of the input spectrogram $\tilde{s}^t$ and the batch spectrograms $\{(\tilde{s}^t, s_i) : s_i \in S_{\text{batch}}\}$ and feed them to the Siamese network for comparison (Fig. 3). We compute gradients w.r.t. $\tilde{s}^t$ and feed them to the next step for filtering.
3. **Gradients Filtering by Similarity**. We discard gradients obtained from target examples with similarity outside a certain range $[\mathcal{S}_{min}, \mathcal{S}_{max}]$. This prevents seeking futile optimization directions, i.e., users who we already match, who we have negligible matching chances.
4. **Perturbation Computation**. We compute the adversarial perturbation $\tilde{\nabla}^t$ as follows:

$$\tilde{\nabla}^t = \max\left(\frac{\alpha}{N}\sum_i \nabla_i, \tilde{\nabla}_{min}\right) \tag{3}$$

   where $\tilde{\nabla}_{min}$ is the minimum perturbation, α is the learning rate, and ∇_i is the gradient from i-th filtered pair.
5. **Spectrogram Update**. We update the current estimate of the input spectrogram as follows:

$$\tilde{s}^{t+1} = \tilde{s}^t + \tilde{\nabla}^t \tag{4}$$

The process is repeated until the gain in FAR is higher than γ. We then get a *master voice waveform* $\tilde{a}$ by inverting its optimized input spectrogram via the Griffin-Lim algorithm [39].

[1] We acknowledge huge impact of the decision thresholds. However, while a 1% FAR may seem to be excessive, even state-of-the-art models cannot guarantee acceptable TPRs for stricter thresholds.

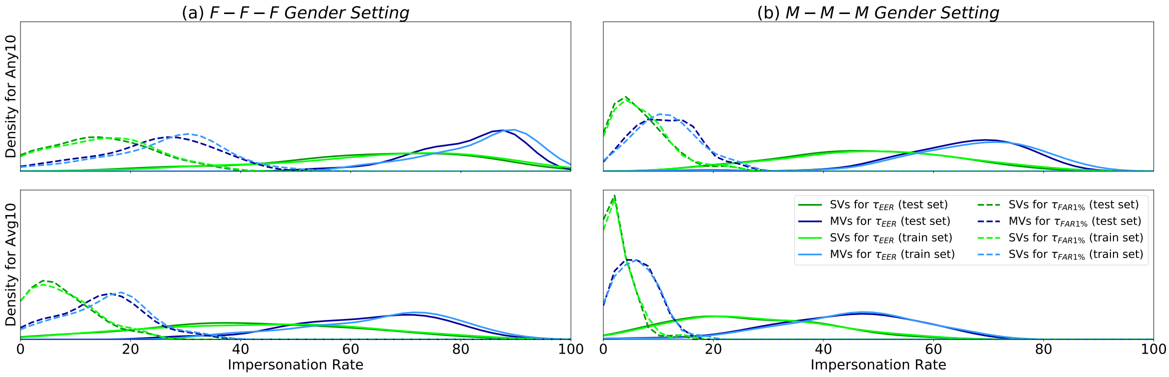


Figure 4: *The distribution of impersonation rates within the population $A_{\mathit{VC2\text{-}Test}}$ for F-F-F and M-M-M gender settings.*

Table 1: *Average impersonation rates of Seed (SVs) and master voices (MVs) on the population $A_{\mathit{VC2\text{-}Test}}$ [%] for diverse seed, training and testing genders (M : Male and F : Female).*

Gender			EER Threshold				FAR1% Threshold			
Seed	Train	Test	Any10		Avg10		Any10		Avg10	
			SVs	MVs	SVs	MVs	SVs	MVs	SVs	MVs
M	**M**	**M**	*47.2*	*65.0*	*27.2*	*44.9*	*6.7*	*16.3*	2.91	*9.6*
M	**F**	**M**	46.6	17.0	26.3	6.9	6.8	1.9	*3.5*	0.7
F	**M**	**M**	3.4	29.3	0.9	14.8	0.3	2.6	0.1	1.2
F	**F**	**M**	3.5	1.5	0.9	0.4	0.3	0.3	0.1	0.1
M	**M**	**F**	4.7	2.1	1.3	0.4	0.3	0.2	0.1	0.0
M	**F**	**F**	5.2	41.9	1.7	23.6	1.8	6.1	1.4	3.3
F	**M**	**F**	62.7	28.3	41.5	15.2	*15.4*	5.6	*8.1*	2.5
F	**F**	**F**	*63.44*	*80.98*	*41.67*	*62.66*	14.39	*34.23*	7.27	*20.78*

5.2. Evaluation Scenario and Results

Our optimization starts from a seed voice (SV) and aims to improve its impersonation capabilities. To measure the improvement that our optimization can achieve from an arbitrary SV, we sample SVs from the population $A_{\text{VC2-Train}}$ while controlling their initial impersonation power. For each gender, we ordered the corresponding 25,000 utterances according to their inherent impostor score for the *Any*10 verification policy and threshold τ_{EER}. We sampled 200 utterances from uniformly distributed percentiles in the population. Finally, we used our optimization procedure, to yield 100 master voices (50:50 for males and females, respectively) optimized for intra-sex matching (i.e. training only on utterances from the same gender in $A_{\text{VC2-Train}}$), and 100 master voices optimized for inter-sex matching (i.e. training only on utterances from the opposite gender in $A_{\text{VC2-Train}}$). We assess their impersonation rates on a separate testing data set $A_{\text{VC2-Test}}$ (disjoint population of 1,000 people).

Table 1 compares the impersonation rates for seed and master voices for different seed, training and testing genders, verification policies and thresholds. The reported results correspond to the test population $A_{\text{VC2-Test}}$, with people unseen at the time of optimization. On average, we can improve the seed impersonation rate by 20 and 10 percentage points for τ_{EER} and $\tau_{\text{FAR1\%}}$, respectively. For the least secure setting with the *Any*10 policy and threshold τ_{EER}, on average, a MV can impersonate 80% of females and 65% of males. In the most secure configuration, the *Avg*10 policy and $\tau_{\text{FAR1\%}}$, on average, a MV can still impersonate 20% of females and 10% of males.

Regarding gender, we observed a significant improvement in the impersonation rates when the same sex is chosen for seed, training and testing samples (i.e., M-M-M and F-F-F settings). Moreover, when the training and the testing gender are the same (i.e., F-M-M and M-F-F settings), the results seem to be good, independent from the seed gender - except for *Avg*10 policy at $\tau_{\text{FAR1\%}}$. This means that the added noise makes it possible to use perturbed utterances to impersonate users of the opposite gender. In contrast, settings with different training and the testing genders (i.e., M-F-M, F-F-M, M-M-F, F-M-F) led to poor results, highlighting how relevant the training gender is on the optimization. Female MVs seem to be more powerful than male MVs. This results from the gender bias observed on VGGVox.

In Fig. 4 we show the distribution of impersonation rates in the populations of seed and master voices for F-F-F and M-M-M gender settings. The probability of finding an utterance with high impersonation rate is low in SVs (green lines), while it significantly increases in MVs (blue lines). This means that MVs produce high impersonation rates independently from the starting utterance and, thus, from the speaker, the speech and the environment. The difference in impersonation rates between train (lighter colors) and test populations (darker colors) is negligible, so MVs generalize well across populations.

6. Conclusions and Future Work

We performed the first analysis of speaker verification systems in the context of recently reported dictionary attacks on biometrics. Based on the obtained results, we can conclude that:

1. Speech seems susceptible to dictionary attacks, and both speaker and speech content affect impersonation rates.
2. Adversarial optimization can be used to significantly increase impersonation capabilities of arbitrary inputs.
3. The gender bias of a verification model increases the difference in exposure to dictionary attacks across genders.
4. Master voices generalize well across populations and are robust to spectrogram computation and inversion.
5. Dictionary attacks should be considered as a valid threat model for speaker verification systems.

In next steps, we will investigate master voice transferability and generative models usage for master voice generation.

7. Acknowledgements

Marras thanks Sardinia Regional Government for financial support (P.O.R. Sardegna F.S.E. Oper. Prog. of Auton. Region of Sardinia, E.S.F. 2014–2020, Axis III, TG 10, PoI 10ii, SG 10.5). This work has been supported by Italian Ministry of Education, University and Research under ”iLearnTV” Project (DD n.1937 5.6.2014, CUP F74G14000200008 F19G14000910008).

8. References

[1] A. K. Jain and A. Kumar, "Biometrics of next generation: An overview," *2nd Gener. Biometrics*, vol. 12, no. 1, pp. 2–3, 2010.

[2] G. Fenu and M. Marras, "Leveraging continuous multi-modal authentication for access control in mobile cloud environments," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 331–342.

[3] G. Fenu, M. Marras, and M. Meles, "A learning analytics tool for usability assessment in moodle environments," *Journal of e-Learning and Knowledge Society*, vol. 13, no. 3, 2017.

[4] G. Fenu and M. Marras, "Controlling user access to cloud-connected mobile applications by means of biometrics," *IEEE Cloud Computing*, vol. 5, no. 4, pp. 47–57, 2018.

[5] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[6] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero, "Leveraging cognitive computing for multi-class classification of e-learning videos," in *Eur. Semant. Web Conf.* Springer, 2017, pp. 21–25.

[7] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[8] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[9] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification: a study of technical impostor techniques," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[11] Y. Stylianou, "Voice transformation: a survey," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3585–3588.

[12] A. Roy, N. Memon, and A. Ross, "Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017.

[13] P. Bontrager, A. Roy, J. Togelius, and N. Memon, "Deep master prints: Generating masterprints for dictionary attacks via latent variable evolution," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. IEEE, 2018.

[14] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2017.

[15] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2016, pp. 1–6.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proc. Interspeech 2011*, 2011, pp. 2341–2344.

[19] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[20] Y.-h. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez, and C. Parada, "Locally-connected and convolutional neural networks for small footprint speaker recognition," in *Proc. Interspeech 2015*, 2015.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[24] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[25] M. Ivanova, S. Bhattacharjee, S. Marcel, A. Rozeva, and M. Durcheva, "Enhancing trust in eassessment-the tesla system solution," in *Proceedings of 2018 Technology Enhanced Assessment Conference*, 2018.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[28] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," vol. 1, no. 2, p. 3, 2018.

[29] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[31] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[32] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," in *Annual Conf. on Neural Information Processing Systems (NIPS)*, 2017.

[33] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[34] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 103–117.

[35] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium*, 2016, pp. 513–530.

[36] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1962–1966.

[37] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220–230, 2010.

[38] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.

[39] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.