# An investigation on voice mimicry attacks to a speaker recognition system

Donato Impedovo[1,2], Annalisa Longo[1], Tonino Palmisano[1], Lucia Sarcinella[1,2], and Davide Veneto [2]

[1] *University of Bari Aldo Moro - Department of Computer Science, Via E. Orabona n.4,70125, Bari, IT*
[2] *Digital Innovation srl, Via E. Orabona n.4 (c/o Department of Computer Science),70125, Bari, IT*

**Abstract**

Voice mimicry is the act in which imitators reproduce the vocal characteristics of another person. It can be considered to be an attack to a speaker recognition system. This work evaluates a speaker identification system under mimicry attacks: the goal is to point out how the accuracy of the system changes depending on the various real scenarios could occur. For this purpose, a GMM-UBM model and an I-Vector have been implemented and tested over dataset of Italian language imitations. Tests have been performed different audio lengths and different use cases. Use cases also take into consideration some possible countermeasures.

**Keywords**

Mimicry Attacks, Voice Recognition, Speaker Recognition, Voice.

## 1. Introduction

Voice has always been one of the most widely used biometrics as a distinctive and measurable feature for the recognition of users in terms of biometric security [1]. The study of speaker recognition focuses more on who is speaking than on what is said, and it can be categorized into two macro-categories, speaker identification and speaker verification. In the first case, the goal is to establish the identity of the speaker; in this case, the system performs a 1:N match between the sample under analysis and all the known models (i.e. users) and then determines which of these is the most similar through the issuance of a score . In the second case, the task is focused on verifying, precisely, if the speaker is whom he/she claims to be, so the system performs a 1:1 match between the sample and the declared model and, depending on whether the score exceeds a certain threshold, the system will issue a boolean value [2]. A further distinction is between text- dependent and text-independent systems. The first case adopts the same text/sentence during testing and training [3], the second refers to a process in which there is no constraint on the text to be pronounced [4].

An important topic that is crucial in these years is about the security of the biometric systems, indeed these systems can be prone to various attacks [5]. In the case of speaker verification/identification systems, replay attacks, speech synthesis, voice conversion and mimicry can be considered [6]. Mimicry is probably the simplest and most common approaches that consists in imitating the voice of another person to attack the system. The attacker tries to imitate the timbre and prosody of the voice without the use of special technologies [7]. This problem has several implications and can occur in many different situations. In fact, it is also connected to the phenomenon of scam and cyberbullying. In most cases a malicious/bully can imitate the victim's voice with the aim to obtain information from unsuspecting people or to mock the imitated person by means of vocal recordings/calls. A speaker recognition system could potentially contribute to identify these actions in multiple scenarios.

*ITASEC'22: Italian Conference on Cybersecurity, June 20–23, 2022, Rome, Italy*
EMAIL: email1@mail.com (A. 1); email2@mail.com (A. 2); email3@mail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)

This work is focused on the task of speaker identification with a text-independent approach. More specifically, this study focuses on voice mimicry attacks to analyze the vulnerability of speaker identification systems, depending on the various scenarios that may arise, whether it is under attack or not.

The main contributions of this study are:

- Identify and test a set or reals attack scenarios and possible countermeasures.
- Compare two state-of-the-art speaker identification systems: a Gaussian Mixture Model - Universal Background Model (GMM-UBM) recognizer and an I-vector recognizer.
- The creation of an Italian language Imitation dataset for the aim of the study.

The paper is divided into the following sections. The section 2 describes the previous state-of-the-art research in the field of speaker identification and mimicry attacks. The section 3 describes the methods and the approach used in this work. The section 4 explains the dataset and the experiments that have been performed and finally in the section 5 and 6 will be discussed the result obtained during the testing phase and the conclusion.

## 2. Related Work

The amount of related works is not extensive highlighting that this is a relevant emerging topic. Yee Wah Lau et al. [8] have experimented a simple mimicry attack on a speaker recognition system to assess its vulnerability: they conclude that if the voice of the impostors is remarkably similar to the voice of the targets speaker the authentication fails. This research has also shown that repeated attempts by an impostor to mimic a target speaker's voice can allow him to obtain a voice much more similar to the target and contribute to the degraded performance.

More recent studies have focused on two widely used speaker recognition systems, one based on a Gaussian Mixture Model - Universal Background Model (GMM-UBM) [9] and the other on an I-Vector classifier [10]. Hautamäki et al. have involved professional imitators in this experiment and the most important aspect for this work is that all the study has been done on a Finnish Language Imitations Dataset created by the authors. They compare the performance of two state-of-the-art systems, a GMM-UBM recognizer and an I-Vector recognizer, testing both the systems, first on genuine voice, as baseline and then on mimicked voice. The result obtained showed that the professional impersonator didn't degrade the performance of the system tested, however there was only a slightly increase of the false acceptance rate for the I-vector system compared to the GMM-UBM [11].

Other interesting research, [12] compared the performance of three different systems GMM-UBM, an I-Vector with cosine similarity and an I-Vector with a probabilistic linear discriminant analysis (PLDA), under mimicry attack. Similar to the previous work it has been used a Finnish language dataset of imitators and imitations acquired from non-expert human listener. The study has showed that the GMM-UBM slightly increased the EER under mimicry attacks, but the other two systems based on the I-vector increased for two times the EER.

Vestman et al. proposed another type of work based on impersonation slightly different from the previous. They used a two ASV system one publicly available based on I-vector and PLDA, and one closed source ASV system based on x-Vector. The aim of this research is to perform a similarity search, in a speech corpus, between recruited attackers and potential target speaker with the first ASV system, then test the impersonators and the most similar voice to target speaker, founded by the first system, on the second ASV system. The research highlights the impersonators don't affect the performance of the ASV system, but an ASV system that attacks another ASV system can be potentially dangerous [13].

## 3. Methods

In this work two systems have been tested. The first is based on GMM-UBM models and the second on the I-Vector model.

*Gaussian Mixture Model – Universal Background Model*

Systems based on Gaussian Mixture Model – Universal Background Model are widely used in the field of speaker recognition due to their easily implementation, the low computational cost compared to the other technics, and the excellent result that can be achieved [9].

A Gaussian Mixture Model (GMM) is a "parametric probability density function represented as a weighted sum of Gaussian component densities" [14]. At the basis of the GMM-UBM system there is the Universal Background Model (UBM), which is a GMM estimated from a large speech dataset. The purpose of the UBM is to model the general feature space distribution of speech. Then with a Maximum a Posteriori (MAP) adaptation from the UBM is possible to obtain the target speaker GMM models.

After the generation of the GMM target speaker models the system can recognize "target speaker" or "ubm" in case of "no target speaker" this because the UBM model can acts like an impostor hypothesis model.

Finally, the verification score of the system is the log-likelihood ration between the test utterance generated from the speaker and that generated by the UBM [9].

*I-Vector*

Another widely used approach is based on the identity vector (I-Vector). At the base of this approach there is the idea that the MAP adaptation performed only on the mean vectors will result in a super-vector of concatenated means.

Given m the super-vector of UBM means, T a low-rank matrix that defines the total variability space and $\phi$ a standard distribution vector. The super-vector M of the segment GMM can be calculated adapting the means of the UBM, so M can be written as:

$$M = m + T\phi \tag{1}$$

The standard distribution $\phi$ is used as the extracted i-vector, while the T-matrix is calculated with an expectation-maximization (EM) algorithm from a development dataset. Usually on the i-vector can be applied a post-processing algorithm like radial Gaussianization, in this way the i-vector can better follow the Gaussian assumptions used in the UBM model. Finally for measuring the similarity between two utterances represented by their corresponding i- vectors can be used the cosine similarity [12].

These two systems have been selected because of their large use in many on-the-shelf application [2]. However, in both cases, common preliminary common operations are carried out as described in the following.

### 3.1. Pre-Processing and Feature Extraction

The audios have been all converted to .Wav format, re-sampled to 8 kHz and switched from stereo to mono channel [11]. Feature extraction have been performed in 5 different sub-steps.

1. First, a pre-emphasis filter has been applied to the signal to enhance the high frequencies of the spectrum, reduced by the speech production process,
2. The signal have been divided into successive 25-millisecond frames with 10-millisecond overlaps by frame blocking and hamming windowing. This process is intended to minimize frequency discontinuity. Since the speech signal varies slowly over time or is a quasi-stationary signal, speech must be examined over a sufficiently short period [15],
3. A Voice Activity Detection (VAD) filter have been applied: this allowed the removal of all superfluous parts (in most cases silence and/or background noise) from the signal, selecting only those discriminating components and allowing the correct speaker to be identified,

4. A RASTA (Relative Spectral Filtering) filter have been applied to eliminate frequencies that are different from the normal change in the voice signal, such as frequencies affected by background noise recorded together with the voice signal,
5. Finally, the MFCC features have been extracted. In detail, for each audio file, a feature vector composed of 20 MFCC coefficients have been extracted with a filter bank composed of 26 triangular filters. The choice of using 20 coefficients was made after some considerations: withmore coefficients, the performance worsens because they would represent rapid changes in signal energy, not representative of the individual's vocal characteristics. On the other hand, fewer would not have enough information to represent the voice adequately. In addition, 20 MFCC delta-features were also computed from the 20 MFCC coefficients for a vector size of40 values. During the feature extraction phase, it was also chosen to replace the first value of the vector, the cepstral coefficient at position zero, usually with a null value, with the log of the entire energy component [16].

## 3.2. Training

Two different approaches have been used for training depending on the specific system.

Regarding the GMM-UBM, Model Adaptation was chosen adapting a previously pre-trained UBM of the Italian language to the speaker's feature vector. The adaptation has been performed using the Maximum a Posteriori (MAP) algorithm, which as the name may suggest, will maximize the a posteriori probability that, given a recording, the correct speaker is selected. It consists of a first phase in which the feature vectors are mapped with probabilistic ratios, inserting them into the UBM mixtures, in this case, 512. Then, the mixtures are adapted using the new data, referring to the "relevance factor ", which quantifies the new data to be analyzed in a single mixture to balance the contribution made by the latter to perform the adaptation. Finally, the testing phase is characterized by matching the feature vector extracted from the system and the model of a speaker, calculating the LLR (Log-Likelihood Ratio).

Concerning the i-vector system, the UBM model has been used to calculate the Total Variability Matrix or TV Matrix, which represents a matrix in which the i-vectors will be extracted, then it will use a function with rank equal to 400 and number of iterations fixed at 20 to realize the actual TV Matrix. The next step will proceed to obtain the statistics for each type of audio, specifically test or training. Once the statistics are obtained, it will proceed with the actual extraction of the i-vector for each speaker and for each test segment, thus obtaining a unique and discriminating model. The resulting super-vectors with reduced dimensions are then the i-vectors, representing the result of the mapping carried out in the first phases. In the last instance, it will calculate the cosine similarity or the cosine of the angle between the adjacent vectors that are compared between the vector representing the speaker enrollment, then the model, and the vector representing the speaker test.

## 4. Data and Experiments

The speech material used in this work consists of 22 Italian celebrities from the world of politics and entertainment. The audios were extracted from interviews, shows, or performances available on online platforms. These 22 identities represent the genuine set of users which will be attacked by imitators. In addition, 17 imitators were chosen and their original voices as well as imitations speech of the 22 genuine users were collected. Finally, there are 24 impersonators representing attacks in the system. As can be guessed, the "imitation" relationship between famous individuals and imitator is n-to-n since multiple imitators can imitate each famous individual, and one imitator can imitate multiple famous individuals. For each speaker, both genuine and impostors, one audio of the duration of 5 minutes have been extracted, while for the audios of the imitations, their duration varies from 40 seconds to 5 minutes.

**Table 1**
Italian Imitation Dataset

| Genuine | Impostors (with their original voice) | Impersonation attacks |
|---|---|---|
| 22 users | 17 users | 24 audios |

In addition to the dataset created, it has been used another dataset for the UBM model training as already described in the previous section, in detail is an extensive Italian repository, the Common Voice Corpus 6.1, consisting of 5729 voices, female, male and uncleared identity (54% male and 14% female and 23% uncleared identity) at different ages (from 18 to 79 years), 158 hours of speech and about 80.000 files [17].

The model has been trained on the 25% of the entire dataset this because the training phase of the UBM model has taken a large amount of time (more than 10 hours).

For the testing phase, audio files were chunked into 1-second and 5-seconds files to analyze if and how, the system behaves by varying the length of the audio.

Tests have been evaluated in terms of precision, recall, and F1 score.

Different use cases have been considered as reported in Table 2. Cases 1, 3, 6, and 7 represent situations in which the system is tested only considering speakers of which it is aware. Cases 2 represent the situation in which the imitator (unknown to the system) performs a mimicry attack. Case 4 is referred to the situation in which the imitator is known to the system as a genuine user, but he/she performs a mimicry attack on another genuine user. Case 5 refers to situations in which the imitator act as a genuine user as well as an impostor.

The last two cases are referred to the possibility to add imitation models.

**Table 2**
Use cases

| Use case | UBM | Models know to the system | Test | Meaning |
|---|---|---|---|---|
| 1 | Pre-trained | Genuine | Genuine | The baseline system. |
| 2 | Pre-trained | Genuine | Genuine + Imitations | The baseline system under mimicry attack. |
| 3 | Pre-trained | Genuine + Imitators | Genuine + Imitators | The baseline system including imitators which does not act as imitators at testing. |
| 4 | Pre-trained | Genuine + Imitators | Genuine + Imitations | The baseline system including imitators under mimicry attack. |
| 5 | Pre-trained | Genuine + Imitators | Genuine + Imitations + Imitators | The baseline system including imitators which act as imitators (mimicry attack) and genuine users at testing. |

| | | | | |
|---|---|---|---|---|
| 6 | Pre-trained | Genuins + Imitations | Genuine | The system explicitly aware of imitations. |
| 7 | Pre-trained | Genuins + Imitations | Genuine + Imitations | The system explicitly aware of imitations under mimicry attack |

## 5. Result and Discussion

Tables 2 report results obtained for the GMM-UBM system referred, respectively, to audio duration of 1s and 5s.

**Table 3**
Results on GMM-UBM approach at 1 second.

| | | Test 1s | | |
|---|---|---|---|---|
| Use case | | F1 Score | Precision | Recall |
| 1 | The baseline system | 91,53% | 94,39% | 89,52% |
| 2 | The baseline system under mimicry attack | 76,79% | 70,87% | 86,69% |
| 3 | The baseline system including imitators which does not act as imitators at testing | 88,37% | 92,09% | 86,39% |
| 4 | The baseline system including imitators under mimicry attack | 80,91% | 79,72% | 85,65% |
| 5 | The baseline system including imitators which act as imitators (mimicry attack) and genuine users at testing | 79,43% | 76,90% | 84,70% |
| 6 | The system explicitly aware of imitations | 91,78% | 97,15% | 87,45% |
| 7 | The system explicitly aware of imitations under mimicry attack | 77,88% | 77,56% | 73,18% |

**Table 4**

Results on GMM-UBM approach at 5 second.

| | | Test 5s | | |
|---|---|---|---|---|
| Use case | | F1 Score | Precision | Recall |
| 1 | The baseline system | 99,42% | 99,70% | 99,17% |
| 2 | The baseline system under mimicry attack | 84,86% | 80,50% | 96,17% |
| 3 | The baseline system including imitators which does not act as imitators at testing | 99,37% | 99,71% | 99,06% |
| 4 | The baseline system including imitators under mimicry attack | 88,84% | 86,48% | 95,86% |
| 5 | The baseline system including imitators which act as imitators (mimicry attack) and genuine users at testing | 89,10% | 85,46% | 97,16% |
| 6 | The system explicitly aware of imitations | 99,53% | 99,60% | 99,09% |
| 7 | The system explicitly aware of imitations under mimicry attack | 88,29% | 95,65% | 87,13% |

The baseline system under attack shows a performance degradation of 15% in both tests (1s and 5s) if the impostor is an outsider (use case 2). A performance degradation of 8% is observed if the impostor is an insider (use case 4 vs use case 3). The situation in which the system is trained on some possible mimicry attack (use case 5) does not differ significantly in performance from the previous case.

Considering the tests performed at 1 second and at 5 seconds, it can be seen that the length of the audio segment affects the performance of the system, in general it can be stated that a wider duration strongly decreases successful attacks. However, the general performance trend along the different use cases is independent by the audio duration. The performance worsens considerably (cases 2, 5, 7) when the system is attacked and does not know speakers or knows only partially imitations (cases 6 and 7) or imitators (cases 3,4,5).

Tables 6 and 7 reports results obtained for the I-Vector model.

**Table 5**
Results on I-Vector approach at 1 second.

| Test 1s | | | | |
|---|---|---|---|---|
| Use case | | F1 Score | Precision | Recall |
| 1 | The baseline system | 92.13% | 93.49% | 91.21% |
| 2 | The baseline system under mimicry attack | 71.56% | 64.24% | 87.64% |
| 3 | The baseline system including imitators which does not act as imitators at testing | 89.25% | 90.23% | 88.90% |
| 4 | The baseline system including imitators under mimicry attack | 78.79% | 76.94% | 85.87% |
| 5 | The baseline system including imitators which act as imitators (mimicry attack) and genuine users at testing | 76.06% | 70.59% | 86.73% |
| 6 | The system explicitly aware of imitations | 91.99% | 94.43% | 89.92% |
| 7 | The system explicitly aware of imitations under mimicry attack | 88.34% | 89.78% | 87.66% |

**Table 6**
Results on I-Vector approach at 5 second.

| Test 5s | | | | |
|---|---|---|---|---|
| Use case | | F1 Score | Precision | Recall |
| 1 | The baseline system | 99.50% | 99.85% | 99.17% |
| 2 | The baseline system under mimicry attack | 80.43% | 73.66% | 95.61% |
| 3 | The baseline system including imitators which does not act as imitators at testing | 99.80% | 99.91% | 99.70% |
| 4 | The baseline system including | 87.56% | 84.55% | 95.78% |

| | | | | |
|---|---|---|---|---|
| | imitators under mimicry attack | | | |
| 5 | The baseline system including imitators which act as imitators (mimicry attack) and genuine users at testing | 87.08% | 81.81% | 97.49% |
| 6 | The system explicitly aware of imitations | 99.80% | 99.92% | 99.69% |
| 7 | The system explicitly aware of imitations under mimicry attack | 95.50% | 98.18% | 95.37% |

Considering the result achieved with the i-vector system, it can be seen that the baseline system under attack shows a performance degradation of 30% in 1s test and 25% in 5s test if the impostor is an outsider (use case 2). A performance degradation of 14% in both tests (1s and 5s) if the impostor is an insider (use case 4 vs use case 3). The situation in which the system is trained on some possible mimicry attack (use case 5) does not differ significantly in performance from the previous case.

Finally, comparing the I-vector results, the same trend obtained in the GMM-UBM model in the various use cases is also reflected in this model. However, the I-vector system suffer from attacks of a slightly higher performance degradation.

## 6. Conclusion

In this work, a study of the vulnerability of speaker identification systems against voice mimicry attacks is presented. The study conducted presents a new Italian mimicking dataset, two very widespread models are implemented and tested on this dataset: the GMM-UBM, and I-vector.

The result obtained using the GMM-UBM system show that the baseline system under attack has a considerable degradation of the performance especially if the system doesn't know or partially know the imitators or the imitations. The same trend has be observed for the I-vector system, even if with slightly increased degradation respect to the GMM-UBM. Comparing the I-vector results, the same trend obtained in the GMM-UBM model in the various use cases is also reflected in this model. However, the I-vector system suffer from attacks of a slightly higher performance degradation.

Performance degradation also depends upon the fact if impostor is another genuine user known by the system or not.

Concerning the length of the audio, longer audio files are able to report higher performances then shorter ones, however, the general performance degradation trend along the different use cases is independent by the audio duration.

In future studies, it will be possible to extends this work with other state-of-the-art technics like artificial neural networks (ANN), extend the Italian mimicry dataset with additional users and compare the result between GMM-UBM, I-vector and ANN.

## 7. Acknowledgements

This work has been supported by the Italian Ministry of Education, University and Research within the PRIN2017 - BullyBuster project - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms.

## 8. References


[1] Kabir, M. M., Mridha, M. F., Shin, J., Jahan, I., & Ohi, A. Q. (2021). A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. IEEE Access (Volume: 9).

[2] Hanifa, R. M., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. Computers & Electrical Engineering, 90, 107005.

[3] El-Moneim, S. A., Sedik, A., Nassar, M. A., El-Fishawy, A. S., Sharshar, A. M., Hassan, S. E., El-Samie, F. E. (2021). Text-dependent and text-independent speaker recognition of reverberant speech based on CNN. International Journal of Speech Technology.

[4] Torfi, A., Dawson, J., & Nasrabadi, N. M. (2018). Text-Independent Speaker Verification Using 3D Convolutional Neural Networks. 2018 IEEE International Conference on Multimedia and Expo (ICME).

[5] Impedovo, D., & Pirlo, G. (2018). Automatic Signature Verification in the Mobile Cloud Scenario: Survey and Way Ahead. IEEE Transactions on Emerging Topics in Computing.

[6] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., & Alegre, F. (2015). Spoofing and countermeasures for speaker verification: A survey. Speech Communication Volume 66, 130-153.

[7] Desai, K. S., & Pujara, H. (2016). Speaker recognition from the mimicked speech: A review. 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).

[8] Lau, Y. W., Wagner, M., & Tran, D. (2004). Vulnerability of speaker verification to voice mimicking. Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing.

[9] Saakshar, K., Pranathi, K., Gomathi, R., & Sivasangari, A. (2020). Speaker Recognition System using Gaussian Mixture Model. 2020 International Conference on Communication and Signal Processing (ICCSP).

[10] Ibrahim, N., & Ramli, D. (2018). I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction. 22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.

[11] Hautamäki, R., Kinnunen, T., Hautamäki, V., & Laukkanen, A.-M. (2013). I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. Interspeech.

[12] Hautamaki, R. G., Kinnunen, T., Hautamaki, V., & Laukkanen, A.-M. (2015). Automatic versus human speaker verification: The case of voice mimicry. Speech Communication 72.

[13] Vestman, V., Kinnunen, T., Hautamäki, R. G., & Sahidullah, M. (2020). Voice mimicry attacks assisted by automatic speaker verification. Computer Speech & Language, 59, 36-54.

[14] Li, S. Z., & Jain, A. (2009). Gaussian Mixture Models. Encyclopedia of Biometrics.

[15] Benesty, J., Sondhi, M., Huang, Y., & Greenberg, S. (2008). Springer Handbook of Speech Processing. Springer.

[16] Leu, F.-Y., & Lin, G.-L. (2017). An MFCC-Based Speaker Identification System. 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA).

[17] Mozilla. (2020). Datasets. Retrieved from Mozilla Common Voice: https://commonvoice.mozilla.org/en/datasets