Multilingual Voice Impersonation Dataset and Evaluation

Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch

Norwegian University of Science and Technology, Gjøvik 2815, Norway
{hareesh.mandalapu,raghavendra.ramachandra,christoph.busch}@ntnu.no

Abstract. Well-known vulnerabilities of voice-based biometrics are impersonation, replay attacks, artificial signals/speech synthesis, and voice conversion. Among these, voice impersonation is the obvious and simplest way of attack that can be performed. Though voice impersonation by amateurs is considered not a severe threat to ASV systems, studies show that professional impersonators can successfully influence the performance of the voice-based biometrics system. In this work, we have created a novel voice impersonation attack dataset and studied the impact of voice impersonation on automatic speaker verification systems. The dataset consisting of celebrity speeches from 3 different languages, and their impersonations are acquired from YouTube. The vulnerability of speaker verification is observed among all three languages on both the classical i-vector based method and the deep neural network-based x-vector method.

Keywords: Biometrics · Speaker recognition · Voice impersonation · Presentation Attack.

1 Introduction

Biometric authentication for providing access to information, devices, and networks have been used in security applications for many years. Speaker recognition is one of the modalities that has been prominently used as biometrics for the last few decades. Though computational intelligence has advanced, biometric systems are still vulnerable in the authentication of individuals. In voice-based person verification, there are emerging new ways of attacks every day. The popular speaker verification vulnerabilities are voice impersonation, audio replay attack, speech synthesis, and voice conversion. Though speech synthesis and voice conversion can cause severe impact, these attacks can only be performed with certain access to the biometric system. The conventional physical access attacks can only be performed by voice impersonation or replay attacks.

Voice impersonation is discussed to be having minimal impact on speaker recognition systems when compared to other kinds of attacks [17]. However, studies have shown that a professional impersonator having enough training on the target's speech can perform a successful attack [2][3]. It is also a simple way of attacking a voice-based biometric system. By adjusting the vocal cords, an

impersonator can mimic a target speaker's voice. Though it has been observed that it is difficult to impersonate untrained target's voices, well-known impersonators after multiple attempts can successfully attack a speaker recognition system. Automatic speaker verification and the vulnerability evaluation have multiple dependencies like text, language, and channel effects [6]. After considering the issues mentioned above, there is a requirement of research work in fully understanding the effect of impersonation with all the dependencies.

In this work, two popular speaker recognition systems are evaluated over the effect of impersonation. We have included three different languages with no text-dependency and various channel data to accommodate the previously mentioned dependencies of automatic speaker verification. Further, this work is organized as follows. A literature review on the previous studies on voice impersonation is presented in Section 2. In Section 3, the impersonation dataset captured is mentioned with details. The Automatic speaker Verification methods chosen for our experiments and trained dataset used are discussed in Section 4. Section 5 explains the experiments performed and results obtained in impersonation vulnerability evaluation. The conclusion of this work and future directions are presented in Section 6.

2 Related Work

In the initial works, amateur impersonators were used in performing attacks. Lau et al. [8] have performed experiments on the YOHO dataset, which contains 138 speakers. Two subjects acted as impersonators, and the vulnerability of the speaker recognition system towards such mimicry attack was verified. Upon performing multiple attempts, it was observed that an impostor could perform an attack if the impostor has the knowledge about enrolled speakers in the database [9]. In [10], Mariéthoz et al. assessed the vulnerability of state-of-the-art text-independent speaker verification system based on Gaussian mixture models (GMMs) to attacks conducted by a professional imitator. It was observed that the GMM based systems are robust to mimicry attacks.

Farrús et al. [2] performed experiments on prosodic features extracted from voices of professional impersonators to perform mimicry attacks on speaker identification systems. The increase in acceptance rates was observed when imitated voices are used for testing. Panjwani et al. [12] proposed a generic method and used crowd-sourcing for identifying impersonators. The GMM-UBM based method displays an increase in impostor attack presentation match rate (IAPMR) when using professional impersonators. Hautamäki et al. [3] used three modern speaker identification systems to test the case of voice mimicry. It has been observed that the EER values for GMM-UBM based method are decreased but increased for two other i-vector based methods.

The ASVspoof (Automatic Speaker Verification spoof) challenges are a series of evaluations focus on improving countermeasures to attacks on speaker verification systems. Voice conversion and speech synthesis attacks are the primary focus in the first ASVspoof challenge [17]. The Second ASVspoof challenge is

evaluated for countermeasures to different kinds of replay attacks [7]. The recent challenge in this series includes both physical (replay attacks) and logical access (voice conversion, speech synthesis) attacks [16]. Impersonation attacks are not considered in any series of these competitions, mentioning that impersonation's relative severity is uncertain. However, the attacks discussed in these series assumed to have access to the biometric system. For example, the audio sample's digital copy is necessary to perform replay attacks, and logical access attacks need access into the system where the digitally manufactured copy of utterance is presented. Impersonation is a physical access attack on voice-based biometrics that does not require any access to the biometric system, which makes it an interesting research topic for this study.

It was observed in most of these methods that voice impersonation has a considerable impact on speaker verification systems, but all these methods possess certain challenges, which are observed as follows.

- There is no publicly available impersonation attack dataset similar to other attacks like replay, voice conversion, and speech synthesis. Also, there is a requirement of professional impersonators to compose a dataset.
- State-of-the-art speaker verification systems are not employed in the evaluation.
- The text-dependent methods are used to perform an attack, which is not a generalized scenario.
- The impact of language and channels are not discussed in the previous evaluations.
- Standard protocols were not used to evaluate the impact of impersonation.

The following contributions are made in this paper to address the challenges mentioned above.

- A dataset of bona fide and impersonator samples is created from YouTube videos for three different languages, which will be made publicly available (similar to VoxCeleb dataset).
- Three different languages, text-independent speeches, and multiple channel data are captured in the dataset.
- Extensive experiments are carried out on one classical and one state-of-the-art speaker verification systems in three different languages.
- Results are presented following ISO/IEC standards for biometric system performance evaluation and presentation attack detection.

3 Voice Impersonation Dataset

The dataset of bona fide speeches and corresponding impersonated speeches are acquired in a process similar to that of the VoxCeleb database. The easiest way to obtain this type of attack dataset is by looking for popular people and their impersonators' speeches that are uploaded to YouTube. In this work, three languages are chosen as per the authors' knowledge: English and two Indian languages: Hindi and Telugu. Multi-lingual data samples also help us to understand

the impact of language used in training data on ASV systems. The bona fide speakers and their well-known impersonators are carefully selected from different subjects in each language. The speakers include political figures and actors.

The bona fide speeches are taken from the interview videos of the target speakers. The impersonation speeches are obtained from YouTube videos of television shows and performances by mimicry artists ranging from amateurs to professionals. The speeches are manually annotated and segmented to individual speakers without any loss in the quality of audio. The speech samples with dominating background noise like applause and music are ignored. The number of speakers and utterances for each language in this dataset is presented in Table 1.

Table 1. Details of impersonation attack dataset.

Language	No. of speakers	Bona fide utterances	Impersonation utterances
English	15	506	411
Hindi	15	768	449
Telugu	15	677	549

4 Vulnerability of ASV systems to Voice Impersonation

The impact of voice impersonation on automatic speaker verification (ASV) systems are verified by performing a presentation attack on the ASV methods using impersonation samples. The initial step in this process is to acquire voice impersonation samples for a set of speakers. Due to the lack of professional impersonators for several speakers, and based on the authors' knowledge of target speakers, we have chosen an obvious way of obtaining impersonation samples from YouTube and included three different languages.

4.1 Training Dataset

In our work, we have used the pre-trained models [1] from Kaldi toolkit [13]. The models are trained on verification split of the VoxCeleb1 and entire VoxCeleb2 dataset [11]. The training dataset is a part of the VoxCeleb dataset, which is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. The main reasons for choosing VoxCeleb trained model are a huge variety of speakers and samples in the dataset (more than 1 million samples and over 7200 speakers) and also the similarity to our dataset of mimicry samples from YouTube. The training dataset contains speech from speakers of a wide variety of cultures, accents, professions, and ages. The details of dataset is presented in Table 2 and 3.

[1] VoxCeleb Models: http://kaldi-asr.org/models/m7

Table 2. Details of the verification split of VoxCeleb1 dataset

VoxCeleb1	Dev Set	Test Set
No.of Speakers	1211	40
No.of Videos	21,819	677
No.of Utterances	148,642	4,874

Table 3. Details of VoxCeleb2 dataset

VoxCeleb2	Dev Set	Test Set
No.of Speakers	5994	118
No.of Videos	145,569	4911
No.of Utterances	1,092,009	36,237

4.2 Automatic Speaker Verification (ASV) Systems

The next step is to obtain ASV systems to examine vulnerability due to voice impersonation. We chose two different methods for this purpose 1. a classical i-Vector based system and 2. a state-of-the-art deep neural network-based x-vector method.

I-vector Method The I-vector based automatic speaker verification method is the state-of-the-art approach proposed in [1]. I-vectors are the low dimensional representation of a speaker sample that is estimated using Joint Factor Analysis (JFA), which models not only the channel effects but also information about speakers. With the help of i-vector extraction, a given speech utterance can be represented by a vector, which includes total factors. The channel compensation in i-vectors is carried out in a low-dimensional total variability space. In this method, we have employed probabilistic linear discriminant analysis (PLDA) [14] to train the speaker models. The trained PLDA models are then used to compute the log-likelihood scores of the target samples to verify the speaker.

X-Vector Method The deep learning and end-to-end speaker verification approaches are the recent popular methods replacing handcrafted methods. The x-vector based speaker verification is one of the latest approaches using deep neural network (DNN) embeddings [15]. This approach uses trained DNN to differentiate speakers by mapping their variable-length utterances to a fixed-dimensional embedding called as x-vectors. A large amount of training data is one of the biggest challenges in this approach. Therefore, data augmentation with added noise and reverberation is used to increase the size of training data.

In the implementation of ASV methods, we have used the pre-trained Universal Background Models, i-vector extractor, x-vector extractor, and speaker recognition codes from Kaldi [2].

[2] Kaldi GitHub: https://github.com/kaldi-asr/kaldi

5 Experimental Results and Discussion

The test set of the VoxCeleb1 dataset is used to verify the performance of obtained ASV methods using pre-trained models. The results of ASV methods on the VoxCeleb1 test set are in Table 4. The thresholds used for attack samples matching bona fide samples are from this test set evaluation.

Table 4. Performance of ASV methods on VoxCeleb1 test set

ASV method	EER (%)
i-vector method	5.3
x-vector method	3.1

The performance of the speaker recognition systems is evaluated using the standardised metrics from ISO/IEC on biometric performance [4]. In addition the Equal Error Rate is reported. The Equal Error Rate (EER) is the rate at which false match rate (FMR) and false non-match rate (FNMR) are equal. The detection error trade-off (DET) curve is used to plot the relationship between the false match rate (FMR) and the false non-match rate (FNMR) for zero-effort impostors and impersonation attacks. Further, the impostor attack presentation match rate (IAPMR) is calculated for each language in two ASV methods. Impostor attack presentation match rate (IAPMR) is the proportion of impostor attack presentations using the same PAI species in which the target reference is matched [5]. In this case, it is the percentage of impersonation attack samples when matched with target speakers above the threshold, which is set by the test set for each ASV system.

5.1 Equal Error Rate (EER) comparison

Table 5. Equal Error Rate (EER%) values of zero-effort impostors and impersonation attacks for the ASV methods on each language

Language	Scenario	i-vector method	x-vector method
English	zero-effort impostors	5.99	3.83
	impersonation attacks	12.94	11.10
Hindi	zero-effort impostors	7.88	5.72
	impersonation attacks	11.17	12.22
Telugu	zero-effort impostors	4.84	3.86
	impersonation attacks	5.57	4.77

The EERs (%) are presented in Table 5 for both zero-effort impostors and impersonation presentation attacks in order to compare the vulnerability caused by voice impersonation on ASV methods. The zero-effort impostors' evaluation is performed with no targeting attacks, whereas the presentation attacks are evaluated by presenting attack samples targeting corresponding speakers. It is important to remember that the zero-effort impostor scores are computed by targeting one speaker on other speakers only in the same language. However, the impersonation samples of one speaker are intended only to target that particular speaker. The IAPMR values that are presented show how many attack samples are matched with target speakers' bona fide samples.

The results show that the increase in the EER (%) values when impersonation attacks are performed. The vulnerability due to the voice impersonation can be seen in both ASV methods. Although the x-vector method has better performance without any attacks (in zero-effort impostors), it can be seen that the vulnerability due to impersonation is similar to i-vector based method. This raises the point that impersonation attacks have an impact even on an advanced deep neural network-based approach similar to the classical method. The comparison of the impact of impersonation attack among different languages deduces some important points. It is interesting to see that the impact is high in the English language when compared to other languages. The reason for this could be that the language in the training dataset is English. This makes ASV methods to recognize the English impersonators more efficiently than other languages.

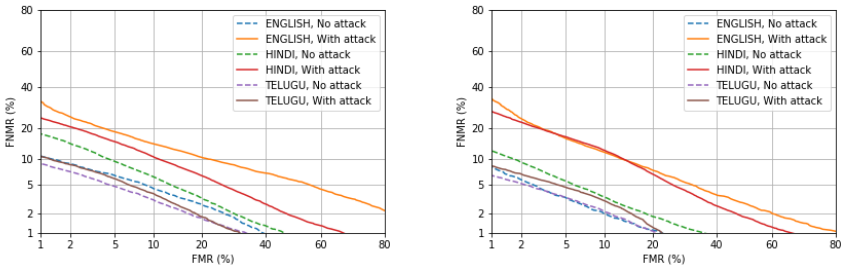
5.2 FMR vs FNMR comparison

The False match rates versus false non-match rate comparisons show the performance of a biometric system by examining the rate of mismatches in both bona fide and impostor samples. We have fixed the false match rate at 0.001 for each case of zero-effort impostors and attacks, then obtained thresholds to compute the false non-match rate. This shows the number of bona fide samples that are not allowed into the system with a fixed allowance of impostors into the system.

Table 6. False non-match rate (FNMR %) of zero-effort impostors and impersonation attacks when False match rate is at 0.001 (i.e. FMR = 0.1%) on each language.

Language	Scenario	i-vector method	x-vector method
English	zero-effort impostors	18.23	16.36
	impersonation attacks	51.93	66.22
Hindi	zero-effort impostors	27.43	22.63
	impersonation attacks	37.29	44.74
Telugu	zero-effort impostors	15.34	12.04
	impersonation attacks	18.31	14.55

The increase in the amount of bona fide samples that result in false match is observed in all languages when attacks are performed. The highest number of mismatches can be seen in the English language in x-vector based method, where more than 66% of FNMR is observed. Further, the DET curves in Figure 1 shows the FMR versus FNMR of two methods in different languages with and without attacks. The increase in error rates can be seen among all systems when the impersonation attack is carried out among all three languages.



(a) i-vector method (b) x-vector method

Fig. 1. Detection Error Tradeoff (DET) curves of the ASV methods with and without impersonation attacks..

5.3 IAPMR evaluation

Table 7. IAPMR (%) values of the impersonation attacks.

Language	i-vector method	x-vector method
English	62.87	58.14
Hindi	46.97	53.87
Telugu	33.43	41.90

The IAPMR values in Table 7 show the percentage of impersonation attack samples that are matched with bona fide samples in each language. The classical i-vector based method has 62.87% of attacks matched in English, which is a considerable amount showing the reasonable impact of voice impersonation on the ASV method. The state-of-the-art x-vector method accepts 58.14% of the samples. This displays a high vulnerability of the ASV method towards impersonation even on the state-of-the-art methods. For other languages Hindi and

Telugu, IAPMR values are lower, which shows the language dependency of the speaker recognition. It is interesting to see that the x-vector method has a higher impact than i-vector method in Hindi and Telugu, unlike English. This impact can also be due to the dependency on the language used in training, which is English.

6 Conclusion

Impersonation attack have been considered as an obvious way of attacking an automatic speaker verification system. In this work, we have studied previous works on voice impersonation evaluation, and a novel dataset of voice impersonation is created. The dataset is captured in a similar way of the VoxCeleb data capturing mechanism in three different languages. The vulnerability of voice impersonation as an attack is examined on a classical and another state-of-the-art speaker recognition systems. The state-of-the-art speaker recognition method is based on a deep neural network-based method that resembles the current technology. Experiments are performed, and evaluations are carried out using ISO/IEC standards with EER, FMR/FNMR, and IAPMR metrics. The results show that the voice impersonations make the ASV methods vulnerable, with many attacks being accepted by the system. It is also interesting to see the vulnerability variation among different languages. The future works on this topic will examine the specific characteristics of the impersonator that are useful in making a successful attack on ASV methods. Also, choosing a training dataset with different languages to examine the language dependency of ASV methods and working on speaker-specific features, like residual phase, to avoid the vulnerability caused by impersonation.

References

1. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (May 2011). https://doi.org/10.1109/TASL.2010.2064307
2. Farrús Cabeceran, M., Wagner, M., Erro Eslava, D., Hernando Pericás, F.J.: Automatic speaker recognition as a measurement of voice imitation and conversion. The Intenational Journal of Speech. Language and the Law **1**(17), 119–142 (2010)
3. Hautamäki, R.G., Kinnunen, T., Hautamäki, V., Laukkanen, A.M.: Automatic versus human speaker verification: The case of voice mimicry. Speech Communication **72**, 13–31 (2015)
4. ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework. International Organization for Standardization and International Electrotechnical Committee (March 2006)
5. ISO/IEC JTC1 SC37 Biometrics: ISO/IEC FDIS 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. International Organization for Standardization (2017)

6. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech communication **52**(1), 12–40 (2010)
7. Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.A.: The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017)
8. Lau, Y.W., Tran, D., Wagner, M.: Testing voice mimicry with the yoho speaker verification corpus. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. pp. 15–21. Springer (2005)
9. Lau, Y.W., Wagner, M., Tran, D.: Vulnerability of speaker verification to voice mimicking. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004. pp. 145–148 (Oct 2004). https://doi.org/10.1109/ISIMP.2004.1434021
10. Mariéthoz, J., Bengio, S.: Can a professional imitator fool a gmm-based speaker verification system? Tech. rep., IDIAP (2005)
11. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: A large-scale speaker identification dataset. In: Lacerda, F. (ed.) Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017. pp. 2616–2620. ISCA (2017). https://doi.org/10.21437/Interspeech.2017, http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0950.html
12. Panjwani, S., Prakash, A.: Crowdsourcing attacks on biometric systems. In: Symposium On Usable Privacy and Security (SOUPS 2014). pp. 257–269 (2014)
13. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF, IEEE Signal Processing Society (2011)
14. Prince, S., Li, P., Fu, Y., Mohammed, U., Elder, J.: Probabilistic models for inference about identity. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 144–157 (2012)
15. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333 (April 2018). https://doi.org/10.1109/ICASSP.2018.8461375
16. Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., Lee, K.A.: Asvspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441 (2019)
17. Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H.: Spoofing and countermeasures for speaker verification: A survey. speech communication **66**, 130–153 (2015)