# Speaker Verification Performance Degradation against Spoofing and Tampering Attacks

*Jesús Villalba, Eduardo Lleida*

Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida}@unizar.es

## Abstract

In this paper, we evaluate the performance of current state of the art speaker verification (SV) systems against some examples of spoofing and tampering attacks. We understand as spoofing the fact of impersonating another person using, for instance, a recording of his voice. On the contrary, we call tampering to the alteration of somebody's voice in order to prevent being detected by a SV system. These techniques can produce important performance degradations. We show that, for the EER operating point, spoofing can produce false aceptance rates of 68% and tampering misses rates of 50%. This is critical in some security applications which makes necessary to develop methods to detect manipulated speech signals.

**Index Terms**: speaker verification, forgery, disguise, spoofing, tampering, JFA.

## 1. Introduction

Current state of the art speaker verification systems (SV) have achieved great performance due, mainly, to the appearance of the GMM-UBM [1] and Joint Factor Analysis (JFA) [2] approaches. However, this performance is usually measured in conditions where impostors do not do any effort to disguise their voices to be similar to any true target speaker and where a true target speaker does not try to modify his voice to hide his identity. That is what happens in NIST evaluations [3]. Therefore, the purpose of this paper is to evaluate SV on this kind of adverse situations.

We have classified the possible attacks to SV as spoofing and tampering. Spoofing is the fact of impersonating another person using different techniques like voice transformation or playing of a recording of the victim. On the other side, tampering is the alteration of somebody's voice to prevent being detected by SV. There are multiple techniques for voice disguise, in [4] authors do a study of voice disguise methods and classify them into electronic transformation or conversion, imitation, and mechanical and prosodic alteration. In [5] an impostor voice is transformed into the target speaker voice using a voice encoder and decoder. More recently, in [6] an HMM based speech synthesizer with models adapted from the target speaker is used to deceive a SV system. In [7] the effects of speaking while grasping a pencil in the teeth are studied. In this work, we focus on low technology techniques like replay attack or putting a handkerchief over the mouth.

This paper is organized as follows. Section 2 explains the spoofing and tampering methods that have been studied in this work. Section 3 describes the experiments and databases, used to measure the perfomance degradation due to these attacks, and shows the results we have got. Finally, in section 4 we show some conclusions.

## 2. Spoofing and tampering methods

### 2.1. Replay attack spoofing

A replay attack consists of an impostor that try to impersonate another person using a recording of his voice. The impersonator could get this recording by several means. One of them would be surreptitiously doing a far field recording of the victim using the microphone of a smartphone or a laptop. Another option could be even getting it from the internet if the victim is a public person. Then, the impostor just needs to replay the sentence using a loudspeaker. This is one of the main weaknesses of SV. Especially text independent systems, that accept that the users say whatever they want. It has mayor importance for applications such as telephone access to bank accounts or admission to restricted areas in a high security facility.

### 2.2. Cut and paste spoofing

The usual approach of commercial SV systems to prevent replay attacks is the use of text dependent systems. The user, that wants to be authenticated, is asked to utter a given sentence that is different for every access attempt. In this case, the SV checks both, the speaker identity and whether the uttered sentence is correct. In this manner, the robustness of the system against attacks is highly increased given that it is unlikely that the asked sentence is among the ones previously recorded by the impostor.

However, this method is not unfailing. If the impostor would have access to fair amount of data from the legitimate user he could be able to build the requested sentence using pieces of different recordings. This is what is call cut and paste spoofing attack. Currently, anybody without any particular expertise can do this with the audio editing programs available in the market.

### 2.3. Handkerchief tampering

What we have called handkerchief tampering consists of covering the speaker's mouth with a handkerchief together with the hand between the mouth and the microphone making a shell. This implies an important distortion on the spectral distribution of the speech signal. Current, state of the art SV use mainly spectral based features (MFCC, PLP, LPCC) so the performance of those systems can be greatly affected. This technique can be applied to cheat the systems of law enforcement agencies that search for criminals into phonecalls.

### 2.4. Nasalization tampering

This kind of tampering consists of obstructing the nostrils while the user is speaking. In this way, the sound wave is reflected back along the nasal cavity interfering with the wave in the pharynx. At certain frequencies both waves cancel each other introducing anti-resonances in the vocal tract transfer function. Like in the previous case, this can modify the spectrum of the signal in such a way that a person could not be detected.

## 3. Experiments

### 3.1. Speaker verification system

We have used a SV system based on JFA [2] to measure the performance degradation. Feature vectors of 20 MFCC (C0-C19) plus first and second derivatives are extracted. After frame selection, features are short time Gaussianized as in [8]. A gender independent Universal Background Model (UBM) of 2048 Gaussians is trained by EM iterations. Then 300 eigenvoices $v$ and 100 eigenchannels $u$ are trained by EM ML+MD iterations. Speakers are enrolled using MAP estimates of their speaker factors ($y$,$z$) so the speaker means super vector is given by $M_s = m_{UBM} + vy + dz$. Trial scoring is performed using first order Taylor approximation of the LLR between the target and the UBM Models like in [9]. Scores are ZT Normalized and calibrated to log-likelihood ratios by linear logistic regression using FoCal package [10] and the SRE08 trial lists. We have used telephone data from SRE04, SRE05 and SRE06 for UBM and JFA training, and score normalization.

### 3.2. replay attack spoofing

*3.2.1. Database*

We have used a database consisting of 5 speakers. Each speaker have 4 groups of signals:

- Originals: Recorded by a close talk microphone and transmitted by telephone channel. There are 1 train signal and 7 test signals. They are transmitted through different telephone channels: digital (1 train and 3 test signals), analog wired (2 test signals) and analog wireless (2 test signals).
- Microphone: Recorded simultaneously with the originals by a far field microphone.
- Analog Spoof: The microphone test signals are used to do a replay attack on a telephone handset and transmitted by an analog channel.
- Digital Spoof: The microphone test signals with replay attack and transmitted by a digital channel.

We have used these signals to create 35 legitimate target trials, 140 non spoof non target, 35 analog spoofs and 35 digital spoofs. The training signals are 60 seconds long and the test signals 5 seconds approximately.

*3.2.2. Results*

We have got an EER=0.71% using the non spoofing trials only. In Figure 1 we show the score distribution of each trial dataset. There is an important overlap between the target and the spoof dataset. If we would choose the EER operating point as decision threshold we would accept 68% of the spoofing trials. Table 1 presents the score degradation statistics between a legitimate utterance and the same utterance after the spoofing processing (far field recording, replay attack). The average degradation is only around 30%. However, it has a big dispersion with some spoofing utterances getting a higher score than the original ones.

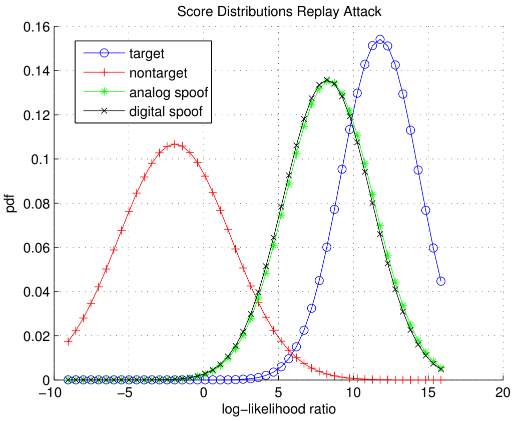

Figure 1: *Score distributions of the replay attack database*

Table 1: *Score degradation due to replay attack*

| | | Mean | Std | Median | Max | Min |
|---|---|---|---|---|---|---|
| Analog | Δscr | 3.38 | 2.42 | 3.47 | 9.70 | -1.26 |
| | Δscr/scr (%) | 29.00 | 19.37 | 28.22 | 70.43 | -10.38 |
| Digital | Δscr | 3.52 | 2.30 | 3.37 | 9.87 | -1.68 |
| | Δscr/scr (%) | 30.29 | 18.92 | 29.52 | 77.06 | -16.74 |

### 3.3. Cut and Paste spoofing

*3.3.1. Database*

The cut and paste database consists of three phases:

- Phase 1+Phase2: it has 20 speakers. It includes landline (T) signals for training, non spoof tests and spoofs tests; and GSM (G) for spoofs tests.
- Phase 3: it has 10 speakers. It includes landline and GSM signals for all training and testing sets.

Each phase has three sessions:

- Session 1: it is used for enrolling the speakers into the system. Each speaker has 3 utterances by channel type of 2 different sentences (F1,F2). Each sentence is around 2 seconds long.
- Session 2: it is used for testing non spoofing access trials and has 3 recordings by channel type of each of the F1 and F2 sentences.
- Session 3: it is made of different sentences and a long text that contain words from the sentences F1 and F2. It has been recorded by a far field microphone. From this session several segments are extracted and used to build 6 sentences F1 and F2 that will be used for spoofing trials. After that, the signals are played on a telephone handset and transmitted through a landline or GSM channel. In this manner, these utterances include cut and paste and replay attack processing.

*3.3.2. Results*

We have done separate experiments using phase1+2 and phase3 datasets. For phase1+2, we train speaker models using 6 landline utterances, and do 120 legitimate target trials, 2280 non

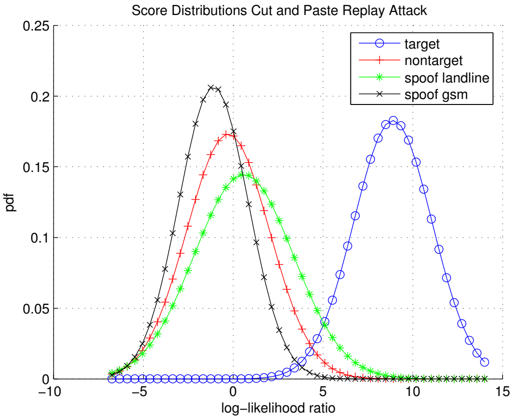

Figure 2: *Score distributions of cut+paste phase 1+2*

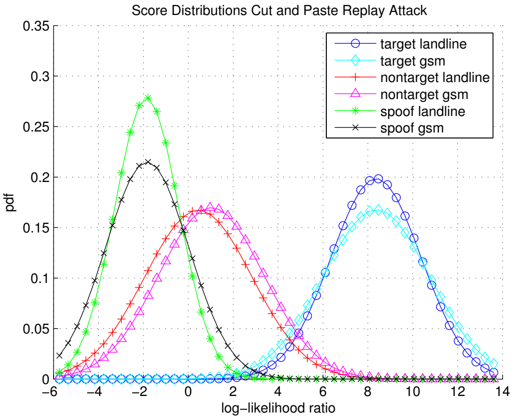

Figure 3: *Score distributions of cut+paste phase 3*

spoof non target, 80 landline spoofs and 80 GSM spoofs. For phase 3, we train speaker models using 12 utterance (6 landline + 6 GSM), and do 120 legitimate target trials (60 landline + 60 GSM), 1080 non spoof non target (540 landline + 540 GSM) and 80 spoofs (40 landline + 40 GSM). Using non spoof trials we have got and EER=1.66% and EER=5.74% for phase1+2 and phase3 respectively. Figures 2 and 3 show the score distributions for each of the databases. Table 2 shows the score degradations statistics due to the spoofing processing. The degradation is calculated by speaker and sentence type, that is, we calculate the difference between the average score of the clean sentence $Fx$ of a given speaker and the average score of the spoofing sentences $Fx$ of the same speaker. We can appreciate that the degradation is more strong in this case than in the database with replay attack only. Even for phase 3, the spoofing scores are lower than the non target scores. This means, the processing used for creating the spoofs can modify the channel conditions in a way that makes the spoofing useless. We think that this is affected too by the length of the utterances. It is known that when the utterances are very short Joint Factor Analysis cannot do proper channel compensation. If the channel component were well estimated the spoofing scores should be higher.

## 3.4. Handkerchief tampering

### 3.4.1. Database

This database consists of 10 speakers with 3 sessions:

Table 2: *Score degradation due to cut+paste replay attack*

| | | | Mean | Std | Median | Max | Min |
|---|---|---|---|---|---|---|---|
| Phase1+2 | T | Δscr | 8.29 | 3.87 | 7.96 | 17.89 | 1.41 |
| | | Δscr/scr (%) | 90.53 | 31.64 | 90.72 | 144.88 | 27.46 |
| | G | Δscr | 9.98 | 2.96 | 9.56 | 18.517535 | 5.40 |
| | | Δscr/scr (%) | 111.94 | 18.03 | 109.437717 | 159.69 | 80.41 |
| Phase3 | T | Δscr | 10.21 | 2.51 | 9.76 | 17.78 | 6.86 |
| | | Δscr/scr (%) | 123.06 | 18.47 | 117.54 | 180.38 | 95.60 |
| | G | Δscr | 10.21 | 3.32 | 10.19 | 18.36 | 4.65 |
| | | Δscr/scr (%) | 121.63 | 19.50 | 119.39 | 167.15 | 92.67 |

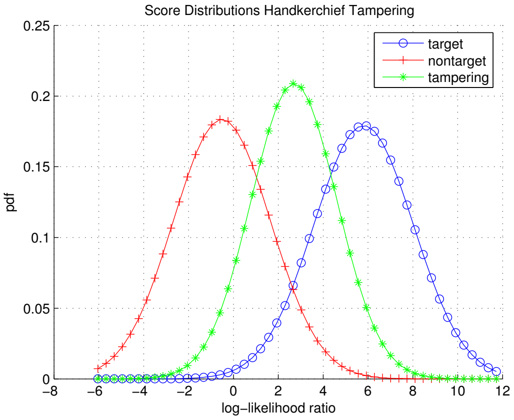

Figure 4: *Score distributions of the handkerchief tampering database*

- Session 1: training speaker models. Around 12 seconds of speech by speaker.
- Session 2: clean test signals. They are 120 short segments of around 3 seconds length.
- Session 3: tampering test signals. Another 120 short segments repeating the sentences of session 2.

With this database, we can do 120 target trials, 120 tampering trials and 1080 non target trials.

### 3.4.2. Results

We have got an EER=6.66% using non tampering trials only. Figures 4 and 5 show the score distributions of each of the trial subsets and the $P_{miss}$ and $P_{fa}$ versus the decision threshold. These figures evidence the great loss of performance that tampering can produce. For the EER operating point, we would reject 50% of true speakers with tampering. Table 3 presents the score degradation statistics between a clean sentence and itself with tampering.

Table 3: *Score degradation due to handkerchief tampering*

| | Mean | Std | Median | Max | Min |
|---|---|---|---|---|---|
| Δscr | 3.10 | 2.20 | 2.90 | 10.37 | -0.88 |
| Δscr/scr (%) | 52.80 | 32.80 | 56.05 | 120.19 | -31.25 |

## 3.5. Nasalization tampering

### 3.5.1. Database

The database consists of 52 speakers. It includes read and spontaneous speech recorded over a GSM channel. Speech segments can have 60, 90 or 120 seconds. Clean segments of 120 seconds have been used for speaker enrollment and the rest for testing. We have 198 clean targets trials, 165 tampering trials and 10098 non target trials.

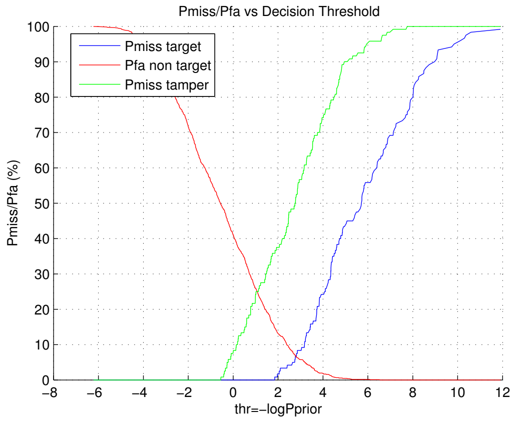

Figure 5: *Pmiss/Pfa vs. decision threshold for the handkerchief tampering database*

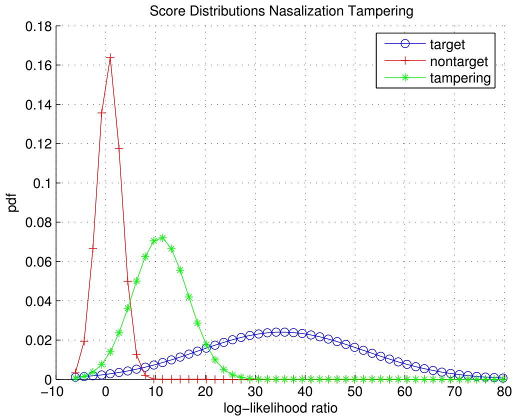

Figure 6: *Score distributions of the nasalization tampering database*

### 3.5.2. *Results*

For this database, we have got an EER=4.54% using clean trials only. Figures 6 and 7 show the score distributions of each of the trial subsets and the $P_{miss}$ and $P_{fa}$ versus the decision threshold. Table 4 presents the score degradation statistics due to the tampering. In this case the tampering and clean sentences are different, so we calculate the degradation as the difference between the average score of the clean recordings of a given speaker and the average score of his tampering recordings. The score degradation is quite big, however the error rates seem less affected having 10% of rejection for the EER operating point. Perhaps, this is due to the bigger length of the utterances that allows better intersession compensation.

Table 4: *Score degradation due to nasalization*

| | Mean | Std | Median | Max | Min |
|---|---|---|---|---|---|
| Δscr | 27.03 | 11.65 | 28.27 | 51.72 | 4.51 |
| Δscr/scr (%) | 68.69 | 13.63 | 69.21 | 96.85 | 11.13 |

## 4. Conclusions

In this paper, we have evidenced the vulnerability of state of the art SV systems to several kinds of spoofing and tampering attacks. For this purpose, we have used databases specifically created to evaluate each type of attack. We have seen that spoofing trials, although having lower scores than the legitimate ones, can produce score distributions high enough to get big acceptance rates. This can be a serious threat for security applications such as authenticating a remote client to give him access to a bank account. On the other side, tampering attacks like nasalization or using a handkerchief to modify your voice can produce low verification scores. This means that applications such as search of criminals in telephone recordings by law enforcement agencies could be easily overcome.

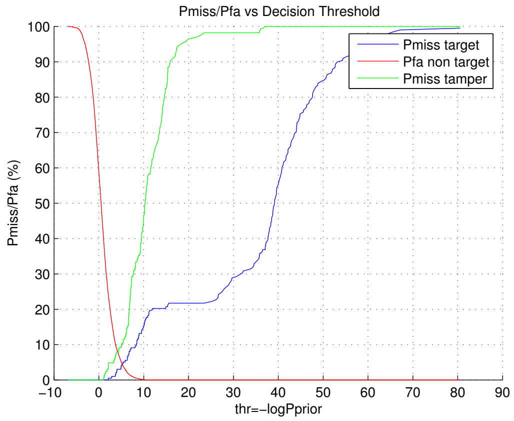

Figure 7: *Pmiss/Pfa vs. decision threshold for the nasalization tampering database*

In order to increase speaker verification systems robustness to this kind of attacks, methods to detect when a signal has been manipulated should be investigated. In the future, we will drive our efforts to this task. However, detecting all kind of manipulations that can be done to a signal can be complicated.

## 5. References

[1] 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, Jan. 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A Study of Interspeaker Variability in Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[3] “http://www.itl.nist.gov/iad/mig/tests/sre/2010/ NIST_SRE10_evalplan.r6.pdf.”

[4] P. Perrot, G. Aversano, and G. Chollet, “Voice disguise and automatic detection:review and perspectives,” *Lecture Notes In Computer Science*, pp. 101–117, 2007.

[5] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, “Voice Forgery Using ALISP: Indexation in a Client Memory,” in *Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, pp. 17–20.

[6] P. L. De Leon, M. Pucher, and J. Yamagishi, “Evaluation of the vulnerability of speaker verification to synthetic speech,” in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[7] R. M. D. Figueiredo and H. D. S. Britto, “A report on the acoustic effects of one type of disguise,” *Forensic Linguistics*, vol. 3, no. 1, pp. 168–175, 1996.

[8] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Oddyssey Speaker and Language Recognition Workshop*, Crete, Greece, 2001.

[9] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, “Comparison of scoring methods used in speaker recognition with Joint Factor Analysis,” in *ICASSP ’09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4057–4060.

[10] N. Brummer, “http://sites.google.com/site/nikobrummer/ focalbilinear.”