Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Effects of waveform PMF on anti-spoofing detection for replay data - ASVspoof 2019

Itshak Lapidot[1,2], *Jean-François Bonastre*[2]

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel
[2]Avignon University, LIA, France

itshakl@afeka.ac.il, jean-francois.bonastre@univ-avignon.fr

Abstract

On speaker recognition identity impersonation recent challenges, we have observed that the *probability mass function* (PMF) of the waveform of genuine speech differs significantly from the PMF of identity theft extracts. In our previous works, we presented the analysis of the influence of the waveform on the *logical access* (LA) spoofing condition, where the spoofing extracts are composed of synthesized or converted speech. In this work we extend our analysis to *physical access* (PA) condition, which focuses on replayed speech. We show that for replayed data, changes in PMF significantly influence performance in terms of spoofing detection. Next, we suggest a way to reduce the observed gap between bona fid and replayed speech waveforms. By analogy with the process used to gaussianize the acoustic parameters, we propose a spoofing speech *genuinization*, which moves the PMF of the spoofing speech near the PMF of the genuine speech. The proposed *genuinization* process is assessed on the ASVspoof 2019 challenge datasets, using the baseline system provided by the challenge organization. In terms of spoofing detecion *equal error rate* (EER), both *linear frequency Cepstral coefficient* (LFCC) and *constant Q cepstral coefficients* (CQCC) features based systems lead to better results when applied on non-genuanized replayed data. On the other hand, when the spoofing detection systems are trained on genuanized data, the results on genuanized replayed data are very good compared to the results obtained without applying genuinization on the data. As observed previously in LA case, the performance is not consistent and it opens problematic questions on generalization capabilities of anti-spoofing systems.
Index Terms: anti-spoofing, waveform, probability mass function (PMF), CQCC, LFCC, GMM, genuinization.

1. Introduction

In recent years, there has been growing interest in the sensitivity of speaker recognition to spoofing attacks and in the development of spoofing countermeasures [1, 2, 3, 4, 5, 6]. In the area of voice authentication, the most common threats are speech synthesis, voice conversion and replay of recorded utterances . Generally, countermeasures are made up of an additional system capable of separating authentic speech and spoofing speech, regardless of the type of spoofing attack. Feature extraction is one the main differences between the different approaches proposed in the literature ([7, 8, 9, 10]). The most promising features seem to be *constant Q cepstral coefficients* (CQCC) [7] which are a non-linear extension of the *linear frequency cepstral coefficients* (LFCC). Other proposed features are mainly based on short-term spectral conversion (e.g., *mel–frequency cepstral coefficients* (MFCC) and CQCC) that ignore the time domain. Only a few exceptions take into account the time domain and even in those cases, it is only used as a preprocessing step followed by a short-term spectral analysis. For example, [11] filters the voice excitation source in order to estimate the residual signal and uses it together with the frequency domain information inside a *Gaussian mixture model* (GMM)-based classifier and [12] applies cochlear filtering and nerve spike density to perform short-term spectral analysis.

It is not surprising that spectral analysis is widely used in the field of countermeasures, since spectral characteristics are also commonly used in many speech conversion systems [13, 14, 15, 16] or in synthesis systems [17, 18].

The lack of interest in time domain information for speech synthesis, speech conversion or for countermeasure is more surprising, as the time domain information is well known for its richness, particularly, but not exclusively, for voice quality parameter estimation and pathological voice assessment [19, 20, 21, 22, 23, 24]. It seems straightforward that at least voice quality parameters should be important for separation between genuine speech and spoofing speech. Ignoring the time domain is apparently more linked to the intrinsic difficulty of this type of approach than to a lesser interest in it.

In [25] and [26], a simple way to use the temporal domain based on entropy parameters has been proposed to detect speech overlap between two speakers. In [27], a similar approach was applied to evaluation of database adequacy. In both cases it was found that this simple representation of information in the time domain provides interesting and valuable information, that is not captured by conventional approaches that are based on short-term spectra. Following the same path, we proposed in [28] a *probability mass function* (PMF) representation of a waveform. Significant differences between the PMF of the spoofing speech and the PMF of the genuine speech were observed. In the same article, a process inspired by the Gaussianization of MFCCs proposed by [29] and noted by analogy, *genuinization*, was also proposed in order to reduce the gap between the PMF of spoofing speech and the PMF of genuine speech. This *genuinization* process works on the waveform amplitude level. It was evaluated on the ASVspoof 2019 challenge [30] *logical access* (LA) conditions.

This article mainly proposes an extension of [28]. It wishes to evaluate the interest of PMF of the waveforms in the case of replay attacks in the context of the ASVspoof 2019 challenge [30].

The rest of the paper is organized as follows: Section 2 describes the databases that are supplied by ASVspoof challenge; in 3, PMF of genuine and spoofing speech are presented and compared; the *genuinization* process is shortly described in section 4; experiments and results are presented in section 5, while

10.21437/Odyssey.2020-44

section 6 presents some general comments on the experiments. Section 7 concludes the paper.

2. Databases

The experiments presented in this work are carried out using the following ASVspoof 2019 data sets: genuine (*bona fide*), logical access (LA - speech synthesis and voice conversion techniques) and physical access (PA - replayed speech). A summary of the different datasets is presented in Tables 1 and 2. It is important to mention that the conditions corresponding to physical access (PA) are simulated both for the acoustic conditions of the recorded room (27 different conditions) and for the replay devices (9 different configurations) . In terms of duration, for both LA and PA, most of recordings are in the range of $1-6$ seconds.

Table 1: Logical condition databases.

	#Speakers		#Utterances	
Subset	**Male**	**Female**	**Bona fide**	**Spoof**
Training	8	12	2,580	22,800
Development	8	12	2,548	22,296

Table 2: Physical condition databases.

	#Speakers		#Utterances	
Subset	**Male**	**Female**	**Bona fide**	**Spoof**
Training	8	12	5,400	48,600
Development	8	12	5,400	24,300

3. PMFs of genuine and spoofing speech

In order to calculate a given waveform PMF, we count how many occurrences are there for each possible sample values and store the results in a histogram. As the audio files are quantized with 16 bits per sample, a 2^{16} bins histogram is extracted. Then, each bin is divided by the total number of samples to obtain the corresponding probability for that bin. If a PMF is calculated on several audio recordings, the histograms are simply accumulated before calculating the bin probabilities. No *voice activity detection* (VAD) is applied: a PMF is extracted using all samples, including non-speech segments.

Figures 1 and 2 respectively show the PMFs of the training files for the LA condition (synthesized or converted speech) and the training files for the PA condition (replayed speech). It is clear that the PMF corresponding to the spoofing data shows a much more pronounced peak near the origin, compared to the PMF of genuine speech. This effect is even more evident for the PA condition than for the LA condition.

Several research works like [31] show experimentally that the non-speech parts are important for the task of detection of identity theft, at least within the framework of the ASVspoof challenge, but they do not provide a theoretical explanation for this phenomenon. Following this discovery, the VAD is generally not used in the field of spoofing detection. To confirm or contradict this generally accepted fact, we repeat the previous experiments for the LA condition, but this time we calculate the PMF after applying a VAD, therefore only on the parts of the recordings that contain speech. Hence, we compute the PMF over the non-speech parts only. We select a very simple energy VAD, using the same approach as in [32] and [33].

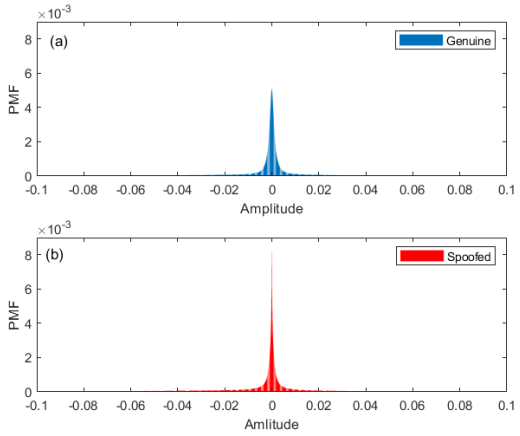

Figure 1: Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech (no VAD).

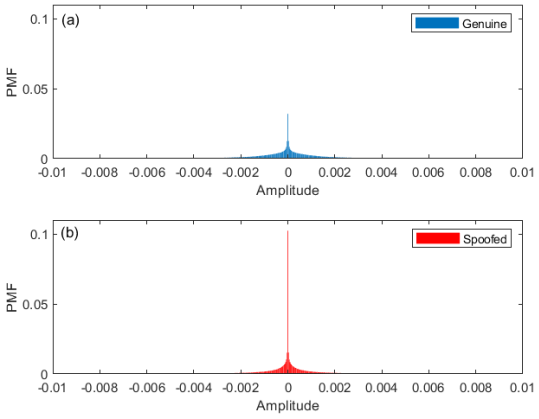

Figure 2: Waveform amplitude PMFs for physical condition, train set, Genuine (a) and Spoofing (b) speech (no VAD).

Figure 3 presents the PMFs computed using the speech parts only (as opposed to the PMF in figure 1 where no VAD is applied). Figure 4 shows the waveform PMFs computed only using the non-speech parts. It is clear that when the PMFs are calculated only on speech parts, little difference is observed between genuine speech and spoofing speech, while a significant difference is observed in the case of PMFs calculated only on the non-speech part. Figures 5 and 6 present the same experiments based on physical access (PA) conditions. The same effects are observed.

It confirms that non-speech is of great importance for the ASVspoof challenge but asks the question whether this conclusion is really based on the characteristics of speech or on other factors. It seems to us that the latter hypothesis is more probable as speech characteristics are certainly more linked to speech parts of the recordings than to non-speech parts...

4. *Genuinization* process

Figures 1 and 2 show that the PMF of genuine speech and the PMF of spoofing speech differ strongly, for both LA and PA conditions. In order to reduce the gap between the PMFs, we

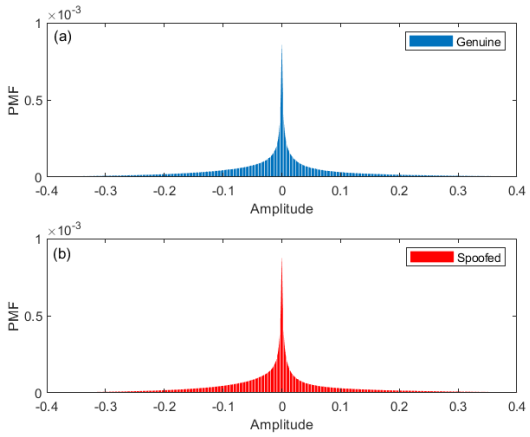

Figure 3: Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech, speech part only (after applying VAD).

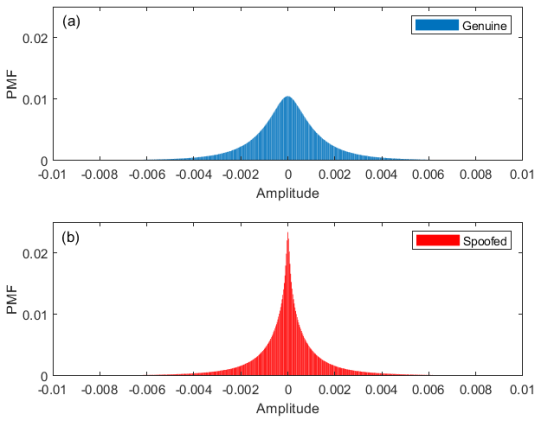

Figure 4: Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech, non-speech part only (after applying VAD).

wish to transform the spoofing speech signal samples in order to obtain a PMF that will be as close as possible to the PMF of genuine speech. Of course, this *genuinization* process should not modify the other aspects of the spoofing speech. Let $p_x^g(k)$ be the PMF of the genuine speech waveform, where g means genuine; x is a discrete random variable $x \in \{1, \ldots, 2^{16}\}$; k is the value assigned to x (the actual signal's amplitude is $s(n) = -1 + k \cdot 2^{-15}$). The *cumulative distribution function* (CDF) is $F_x^g(k) = \sum_{q=1}^{k} p_x^g(q)$. Then, for each spoofing speech signal $s(n)$ a PMF $p_x^s(k)$ (s for the spoofing signal) is calculated, followed by the its CDF $F_x^s(k)$. The *genuinization* algorithm is described in Algorithm 1.

5. Experiments using *Genuinization*

This section presents different experiments using the proposed *genuinization* process. Figure 7 shows the PMF of spoofing speech before and after *genuinization* for the LA condition (a narrow band of amplitudes is underlined to facilitate comparisons). The PMF of genuine speech is also given for comparison. No VAD process is applied and all the samples are used. A significant correction of the PMF is observed. We performed a subjective assessment of the spoofing files by listening to several recordings. This evaluation indicates that poor quality spoofing files lose quality after *genuinization* whereas for high quality recordings, no degradation is observed. However, this

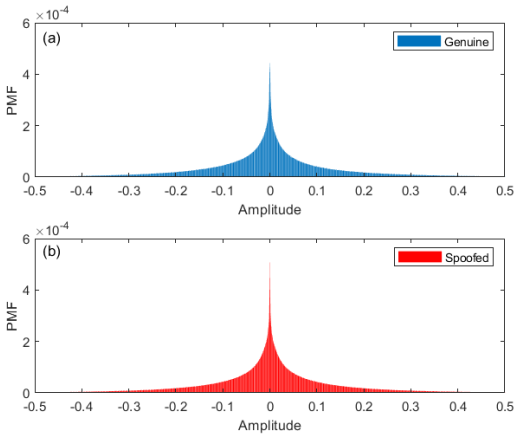

Figure 5: Waveform amplitude PMFs for physical condition, train set, Genuine (a) and Spoofing (b) speech, speech part only (after applying VAD).

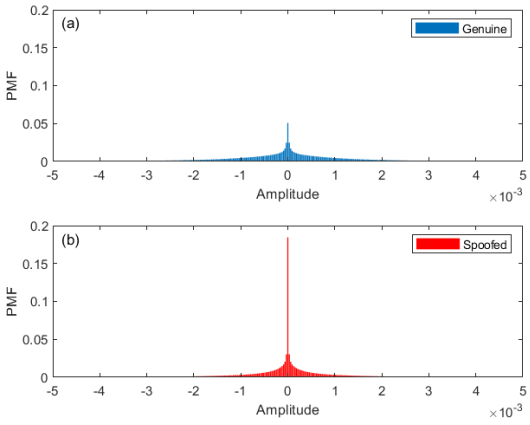

Figure 6: Waveform amplitude PMFs for physical condition, train set, Genuine (a) and Spoofing (b) speech, non-speech part only (after applying VAD).

Algorithm 1 Genuanization algorithm

Require:
Given a spoofing file, $s(n)$ ▷ $n = 1, \ldots, N$
Be the genuinized file, $\hat{s}(n)$
Genuine CDF $F_x^g(k)$ ▷ $k \in 1, \ldots, 2^{16}$
Spoofing file CDF $F_x^s(k)$
for $k := 1$ **to** N **step** 1 **do**
  Set $k = [s(n) + 1]\, 2^{15}$.
  Find $q^* = \arg\max_q \{F_x^g(q) \leq F_x^s(k)\}$
  Set $\hat{s}(n) = -1 + 2^{-15} \cdot q^*$
Return: $\hat{s}(n)$

does not guarantee that *genuinization* is capable of improving spoofing performance or making it more difficult to detect.

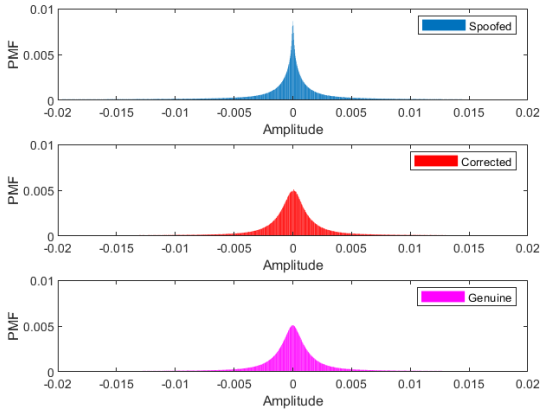

Figure 7: Waveform amplitude PMFs for logical condition, train set, original spoofing speech (upper), spoofing speech after genuinized (middle) and genuine speech (bottom).

In the rest of this section, we assess the effect of the *genuinization* on spoofing speech detection performance. The challenge baseline system is used and two feature sets are evaluated, LFCC and CQCC. The GMMs have 512 mixture components. Several variants are explored: the *genuinization* is applied to the training data, to the test data or to both; the *genuinization* is performed only on the non-speech parts (the parameter estimation is done on the non-speech parts and only these parts of the signals are transformed) or uses the complete utterances. Spoofing speech detection performance is evaluated on ASVspoof 2019 development dataset in terms of EER [30].

To help the reader, Table 3 presents a summary of the experimental conditions.

Table 3: Experimental setup. It shows for each experiment whether the *genuinization* is applied (y) or not (n), respectively, on train and test data. The NS column indicates if all the samples are used for the *genuinization* (n) or if the *genuinization* is focused on Non Speech (NS) only (y).

Table number	Train [y/n]	Test [y/n]	NS [y/n]
4	n	n	n
4	n	y	n
5	y	n	n
5	y	y	n
6	n	n	n
6	n	y	y
7	n	n	n
7	n	y	n
8	y	n	n
8	y	y	n

5.1. LA condition

Table 4 summarizes the results when the *genuinization* is applied or not during the test (on spoofing records only), but not during the training phase. With CQCC features, applying *genuinization* increases the spoofing detection error rates by a factor of 10. It might indicate that CQCC features, which work about 9 times better than LFCC without *genuinization*, are related to the waveform amplitude information. For LFCC, the EER decreases from 2.7% without *genuinization* to 1.29% with *genuinization*. The latter result may suggest that the LFCC features are loosely linked to the time domain waveform information.

Table 4: Spoofing detection performance using original or *genuinized* test files for LA conditions.

		Original	Genuinized
LFCC	**EER [%]**	2.709	1.291
CQCC	**EER [%]**	0.394	3.219

The table 5 presents a similar experience using *genuinization* for the training phase. For CQCC, the spoofing detection system displays an EER at 43% with the original test data. When *authentication* is also used for test files, the EER approaches 0%. For LFCC, the use of *genuinization* on the training set results in a decrease in performance (around 34.4% EER) but it also approaches zero when *genuinization* is applied to the test spoofing recordings. The two results seem to confirm our hypotheses on the role of waveforms for MFCC and CQCC.

Table 5: Spoofing detection performance using *genuinization* during the training, for LA condition. The results when textit-genuinized is applied or not on test files are provided.

		Original	Genuinized
LFCC	**EER [%]**	34.379	0.048
CQCC	**EER [%]**	43.477	0.007

In section 5 we raised the question on the use of non-speech parts for spoofing detection. The following experiments aim to assess whether the *genuinization* process is sensitive to the question of speech/non-speech. Such sensitivity would support the hypothesis that the good results shown using non-speech parts are more of an artifact of the evaluation process, rather than a true characteristic of speech or spoofing. To assess the sensitivity to non-speech segments, we suggest working only on non-speech parts: the parameters of the *genuinization* (the target CDF) are trained using non-speech parts only and the *genuinization* is applied only to non-speech parts during the test phase.

Table 6 presents the results when the *genuinization* is performed on non-speech data only (it is similar to the table 4, except for the emphasis on non-speech parts). For CQCC, the spoofing detection EER is multiplied by a factor of 4 when non-speech *genuinization* is applied (from 0.39% to 3.37% EER). For the LFCC, a relative increase of about 25% in the EER is observed when non-speech *genuinization* is applied (from 2.7% to 3.37% EER). These results confirm that the CQCC are more linked to waveform information than the LFCC are. Moreover the results indicate that the good performance obtained in the literature using non-speech parts for spoofing detection is mainly due to an experimental artifact .

5.2. PA condition

This section is devoted to the effect of *genuinization* on the physical access condition (PA), the replay attack condition. Figure 8 shows the corresponding PMFs for genuine speech,

Table 6: Spoofing detection performance using non-speech only *genuinization* of test data for LA conditions.

		Original	Genuinized
LFCC	**EER [%]**	2.709	3.374
CQCC	**EER [%]**	0.394	1.577

spoofing records without *genuinization* and with *genuinization*, computed on all samples (VAD is not applied). When we apply the *genuinization* to the replayed data, we observe an unexpected phenomenon, with an astonishing behavior close to the null amplitude (attention, the maximum value on the vertical axis is not the same for the three plots). It looks like no sample is assigned to zero during the *genuinization* process. It can be explained by looking at the CDFs as shown in Figure 9. Since the probability that the spoofing signal has the value zero is much higher than for the authentic signal, the assignment $q^* = \arg\max_q \{F_x^g(q) \leq F_x^s(k)\}$ (as in the algorithm 1) never has the value of $q^* = 2^{15}$ (which gives $s(n) = -1 + q^* \cdot 2^{-15} = 0$) but only a higher value. This explains why in Figure 8, middle plot, the right side has a "humped back", equivalent to the probability of $s(n) = 0$. In order to overcome this phenomenon, we propose to perturb the spoofing signal a little bit and to apply the algorithm 1 on this perturbed signal. The result of this approach is illustrated in Figure 10. It can be seen that, now, the genuinized PMF is similar to the PMF of the genuine speech.

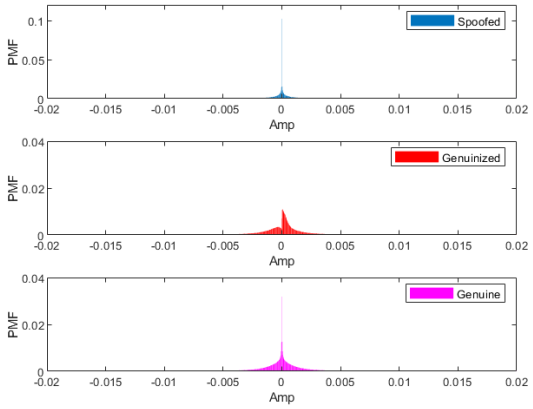

Figure 8: Waveform amplitude PMFs for PA condition, train set, original spoofing speech (upper), spoofing speech after genuinized (middle) and genuine speech (bottom).

We also examine the effect of applying *genuanization* to the training data for physical access condition (PA), following the same process as for the logical access (LA) condition, using all samples (VAD is not applied). The results are presented in Table 7 and in Table 8. The use of the *genuinization* on the test data reduces the EER for the two feature sets, LFCC (from 11.96% to 8.41% EER) and CQCC (from 9.87% to 8.52%EER). To train the spoofing model after *genuinization* yields a large loss in spoofing detection performance when the *genuinization* is not applied on test data (from 12% to 30% EER for LFCC and from 10% to 20% EER for CQCC). Still, the same model allows for a marked improvement in performance when the *genuinization* is also applied to the test data

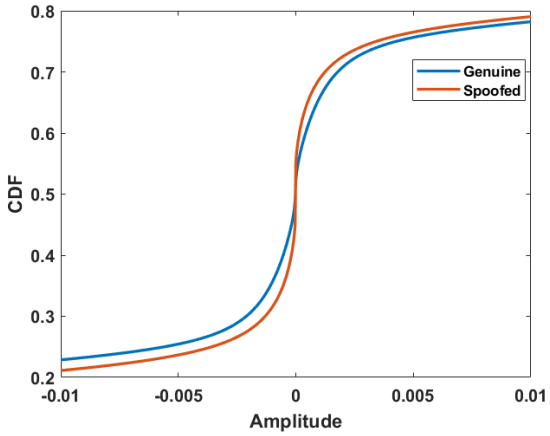

Figure 9: The CDFs of the genuine and spoofing speech for PA condition.

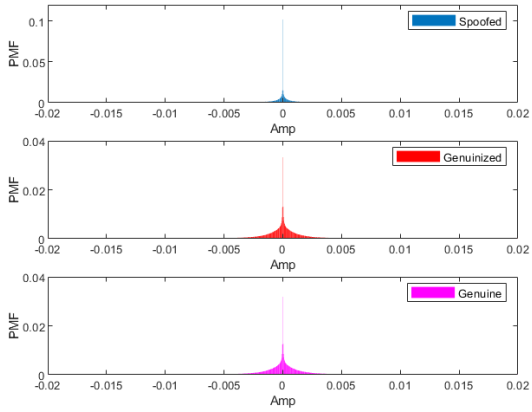

Figure 10: Waveform amplitude PMFs for PA condition, train set, original spoofing speech (upper), spoofing speech after genuinized (middle) of the perturbated signal and genuine speech (bottom).

(1% EER for LFCC and 0.1% EER for CQCC).

Table 7: Spoofing detection performance using original or *genuinized* test files for PA conditions.

		Original	Genuinized
LFCC	**EER [%]**	11.96	8.41
CQCC	**EER [%]**	9.87	8.52

Table 8: Spoofing detection performance using *genuinization* to train spoofing model and original or *genuinized* test files for PA conditions.

		Original	Genuinized
LFCC	**EER [%]**	30.22	0.98
CQCC	**EER [%]**	19.54	0.09

6. Comments

The time domain is generally overlooked in conventional spoofing/anti-spoofing approaches, which focus mainly on the frequency domain. This is mainly due to the intrinsic difficulty of methods in the time domain compared to those based on the frequency domain. We aim to overcome this limitation since certain statistics of speech could be more easily highlighted in the time domain than in the frequency domain.

This article follows on our previous work, [28], where we proposed a relatively simple approach in the time domain, which exploits the *probability mass function* (PMF) of the waveform amplitudes. We have shown that for both ''sides of the coin", spoofing and anti-spoofing, more attention should be paid to the time domain. This is evident in the differences in waveform amplitude PMFs that we highlighted. Based on the PMF differences, we have proposed a simple *genuinization* method to transform the spoofing speech in order to reduce the waveform amplitudes PMF gap between genuine and spoofing speech. We showed that it worked very well for LA conditions.

In the present work, we first extended our interest to physical access (PA) conditions. We observed that the probability of the replayed speech signal at $s(n) = 0$ is exceptionally high. To overcome this phenomenon, we perturbed slightly the speech amplitude before the CDF calculation. After this modification, our *genuinization* approach works as expected on the PA condition (we applied the same protocol on the LA conditions and it did not harm results).

Furthermore, when we looked at the performance of ASVspoof 2019 challenge baseline system using LFCC or CQCC features, we found that the baseline system was vulnerable to time domain signal changes brought about by the *genuinization* process. For LA condition, the CQCC seem to be more sensitive to the waveform amplitude information than the LFCC. On the other hand, for PA conditions, we observed that the LFCC were more sensitive than the CQCC. Even if we are not able to generalize our findings to all anti-spoofing systems, it seems reasonable to assume that at least some of them are also vulnerable to these time domain changes. It can also be assumed that other features may also be sensitive to time domain manipulations.

It is surprising that restricting the signal to the speech parts (via the use of a VAD) negatively affects performance during ASVspoof challenges. It seemed interesting to better understand the cause for this phenomenon, so part of the work presented in this article was devoted to this point. This probably relates to the differences we observed in the low amplitude part of the PMF between authentic speech and spoofing speech, since this area is mainly associated with non-speech events. We have also shown that focusing on non-speech only for the *genuinization* affects the performance (the *genuinization* is trained using non-speech segments only and is applied to non-speech parts only). This raises the question regarding the information used by current spoofing and anti-spoofing systems. Certainly, more attention should be paid to the non-speech (and low energy) parts when generating the spoofing files using voice synthesis or voice conversion technologies.

Another problem is related to the evaluation process of the ASVspoof 2019 challenge itself, concerning the PA condition. In order to control the replay conditions, "recording" and "replay" are simulated. This may be the cause of the form of the replayed PMF. The synthesis procedure generally takes into account the spectral density, which is the same as the autocorrelation information (that is to say the second order statistics). It does not care about the waveform sample distribution and the PMF can differ considerably from the real replayed PMF, while the spectral density remains very close.

7. Conclusion

To conclude, this work extends our previous work carried out on the LA condition, to the PA condition. A simple approach in the time domain, the waveform amplitude distribution, is used to compare genuine speech versus spoofing speech. Similar effects are observed using PA data as those we had found for LA conditions. Even though the proposed method is simple, it clearly shows that time domain information cannot be ignored for spoofing as well as for spoofing countermeasures.

In the real world, during replay attacks, the spoofing system does not have access to the data after the replay process, so it is impossible to apply *genuinization* in the PA conditions. Still, using the same principle, it seems to us that applying a *genuinization* pre-processing before replaying the signals is interesting and should be studied carefully.

This work also answers an open question regarding the common use of non-speech signal zones during various anti-spoofing challenges. It shows that the gain observed while working on non-speech parts is more an artifact than a fundamental characteristic.

In addition, this last point illustrates that it is not possible to rely solely on performance measurement and that it is essential to use an explainable approach in the field of spoofing of voice authentication and its countermeasures.

8. Acknowledgement

This work was supported by the ANR–JST CRES VoicePersonae project and was mainly done during I. Lapidot's sabbatical stay in LIA Avignon University.

9. References

[1] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *INTERSPEECH*, 2007.

[2] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *INTERSPEECH*, 2012.

[3] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, *Speaker recognition anti-spoofing*. Book Chapter in "Handbook of Biometric Anti-spoofing", Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014, 06 2014.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH 2015, September 6-10, 2015, Dresden, Germany*, 2015.

[5] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *ODYSSEY 2016, The Speaker and Language Recognition Workshop, June 21-24, 2016, Bilbao, Spain*, Bilbao, SPAIN, 06 2016.

[6] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: The automatic speaker verification spoofing

and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[7] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, vol. 45, pp. 516 – 535, 2017.

[8] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 632–643, June 2017.

[9] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, 2015.

[10] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, April 2016.

[11] T. B. Patel and H. A. Patil, "Significance of source–filter interaction for classification of natural vs. spoofed speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 644–659, June 2017.

[12] ——, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 618–631, June 2017.

[13] Z. Wu, T. Kinnunen, C. E. Siong, and H. Li, "Text-independent f0 transformation with non-parallel data for voice conversion," in *INTERSPEECH*, 2010.

[14] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65 – 82, 2017.

[15] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," *Speech Communication*, vol. 96, pp. 1 – 9, 2018.

[16] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, March 2019.

[17] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130 – 153, 2015.

[18] A. Khodabakhsh, A. Mohammadi, and C. Demiroglu, "Spoofing voice verification systems with statistical speech synthesis using limited adaptation data," *Computer Speech and Language*, vol. 42, pp. 20 – 37, 2017.

[19] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *proceedings of ICASSP-93*, 1993, pp. 554–557.

[20] D. D. Deliyski, "Acoustic model and evaluation of pathological voice production," in *EUROSPEECH*, 1993.

[21] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, no. 2, pp. 131–138, 1996. [Online]. Available: https://doi.org/10.1016/0167-6393(95)00040-2

[22] A. N. C. Christer Gobl, "Amplitude-based source parameters for measuring voice quality," in *Voice Quality: Functions, Analysis and Synthesis*, 2003.

[23] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[24] P. G. Vilda, R. Fernández-Baíllo, M. V. R. Biarge, V. N. Lluis, A. Á. Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759–781, 2009. [Online]. Available: https://doi.org/10.1016/j.specom.2008.09.005

[25] O. Ben-Harush, I. Lapidot, and H. Guterman, "Entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *Proceedings of Interspeech 2009*, 2009.

[26] O. Ben-Harush, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, Sep. 2009, pp. 1–6.

[27] I. Lapidot, H. Delgado, M. Todisco, N. Evans, and J.-F. Bonastre, "Speech database and protocol validation using waveform entropy," in *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, September 2-6, 2018, Hyderabad, India*, Hyderabad, INDIA, 09 2018.

[28] I. Lapidot and J.-F. Bonastre, "Effects of Waveform PMF on Anti-Spoofing Detection," in *Interspeech 2019*. Graz, Austria: ISCA, Sep. 2019, pp. 2853–2857.

[29] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *ODYSSEY 2001 -The Speaker and Language Recognition Workshop*, Crete, Greece, June 2001.

[30] "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," Tech. Rep., 01 2019.

[31] G. Valenti, H. Delgado, M. Todisco, N. Evans, and L. Pilati, "An end-to-end spoofing countermeasure for automatic speaker verification using evolving recurrent neural networks," in *ODYSSEY 2018, The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, FRANCE, June 26-29 2018.

[32] O. Ben-Harush, O. Ben-Harush, I. Lapidot, and H. Guterman, "Initialization of iterative-based speaker diarization systems for telephone conversations," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 414 –425, feb. 2012.

[33] I. Lapidot and J.-F. Bonastre, "Generalized viterbi-based models for time-series segmentation applied to speaker diarization," in *ODYSSEY 2012 -The Speaker and Language Recognition Workshop*, 2012.