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

Effects of waveform PMF on anti-spoofing detection

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

In the context of detection of speaker recognition identity impersonation, we observed that the waveform *probability mass function* (PMF) of genuine speech differs from significantly of of PMF from identity theft extracts. This is true for synthesized or converted speech as well as for replayed speech. In this work, we mainly ask whether this observation has a significant impact on spoofing detection performance. In a second step, we want to reduce the distribution gap of waveforms between authentic speech and spoofing speech. We propose a *genuinization* of the spoofing speech (by analogy with *Gaussianisation*), i.e. to obtain spoofing speech with a PMF close to the PMF of genuine speech. Our *genuinization* is evaluated on ASVspoof 2019 challenge datasets, using the baseline system provided by the challenge organization. In the case of *constant Q cepstral coefficients* (CQCC) features, the *genuinization* leads to a degradation of the baseline system performance by a factor of 10, which shows a potentially large impact of the distribution os waveforms on spoofing detection performance. However, by ''playing" with all configurations, we also observed different behaviors, including performance improvements in specific cases. This leads us to conclude that waveform distribution plays an important role and must be taken into account by anti-spoofing systems.

Index Terms: anti-spoofing, waveform, probability mass function (PMF); CQCC, LFCC, GMM.

1. Introduction

In recent years the sensitivity of speaker recognition to spoofing attacks and the development of spoofing countermeasures raised an increasing interest, [1, 2, 3, 4, 5, 6]. In the field of voice authentication area, the most common threats come from replaying recorded utterances, voice synthesis and voice conversion. Associated countermeasures are generally composed of a specific additional system capable of separating true exmaple of speech and spoofing examples, regardless of the type of spoofing attacks. Different approaches are applied, [7, 8, 9, 10]. One of the main differences between these approaches (as well as between speaker recognition and spoofing detection) is related to the feature extraction. Different features were proposed for anti-spoofing systems [9]. The most promising seem to be *constant Q cepstral coefficients* (CQCC) [7] which are a non-linear extension of the *linear frequency Cepstral coefficients* (LFCC). Most of the proposed features are based on short-term spectral conversion (e.g., *mel–frequency cepstral coefficients* (MFCC) and CQCC) and ignore the time domain. There are few exceptions, and even when the time domain is in interesting, it is only used as a pre-processing step followed by short-term spectral analysis. [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; [12] applies cochlear filtering and nerve spike density perform a short-term spectral analysis.

Spectral features are commonly used not only for countermeasures, but also in many speech conversion systems [13, 14, 15, 16] and synthesis algorithms [17, 18].

This apparent lack of interest in time domain information is surprising as 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 are important for genuine vs spoofing speech separation. If time domain is mainly ignored in spoofing countermeasure, this is certainly more related to the intrinsic difficulty of time-domain approaches than to a lack of information at this level.

To overcome this limitation, we wish to start by exploiting simple representations of time-domain related information. In previous works, we explored the entropy of waveform coefficients. In [25] and [26] it has been shown that entropy parameters can be applied successfully to detect the overlap of speech between two speakers. In [27], we successfully applied a similar approach to database assessment. In both cases, it was found that this simple or oversimple representation of time domain information provided interesting information, clearly omitted by conventional approaches based on short-term spectra-based.

In some of the experiments on the detection of speech falsification, we examine, somewhat by chance, the global *probability mass functions* (PMFs) of the genuine speech recordings versus the spoofing speech recordings (for all the cases, synthesized, converted or replayed). We were surprised by the big differences observed. This work is based on this observation and further explores the role of waveform amplitudes' PMFs in spoofing speech detection. We propose to correct the imbalanced observed between the PMFs of genuine speech and spoofing speech. For this, we propose a process inspired from the *Gaussianization* of the MFCC features proposed by [28], applied at the waveform coefficient level and noted by analogy *genuinization*. We examine the effect of our *genuinization* process when it applies to different types of spoofing speech and genuine speech. We also investigate the behavior of *genuinization* on high and low energized part of the speech signal and on spoofing detection system training set. In this work we use the ASVspoof 2019 challenge [29] train and the development sets as well as the baseline system provided by the challenge.

2. Databases

As presented in the previous section, we use in this article the data composed of the ASVspoof 2019 challenge [29] genuine (*Bona fide*) and logical conditions (speech synthesis and voice conversion techniques). A summary of the different datasets is presented in Table 1.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2607

Table 1: *Logical condition databases.*

	#Speakers		#Utterances	
Subset	**Male**	**Female**	**Bona fide**	**Spoof**
Training	8	12	2,580	22,800
Development	8	12	5,400	48,600

We use also training data of the Physical condition (replayed speech) but only in order to compute and show the corresponding waveform PMF, in the next section. In this challenge, the conditions were simulated both for the recorded room acoustic conditions (27 different conditions) and for replay devices (9 different configurations). The Physical train dataset is summarized in Table 2. For both Logical and Physical conditions, the vast majority of the recordings is $1-6$ seconds in duration.

Table 2: *Physical condition train databases.*

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

3. PMFs of genuine and spoofing speech

The audio files of ASVspoof 2019 have 16 bits per sample. To compute a given PMF, we take all the audio files and extract the corresponding 2^{16} bins frequency histogram. Figures 1 and 2 show the PMF of, respectively, training files of logical condition (synthesized or converted speech) and training files of physical condition (replayed speech). In both cases, PMF of the genuine speech (Bona fide) is also provided for comparison purposes. All PMFs are calculated using all speech samples (speech activity detection is not applied). It appears clearly that spoofed data PMF has a much sharper pick close to the origin. This effect is more accentuated for physical condition than for synthesized and converted speech (logical condition).

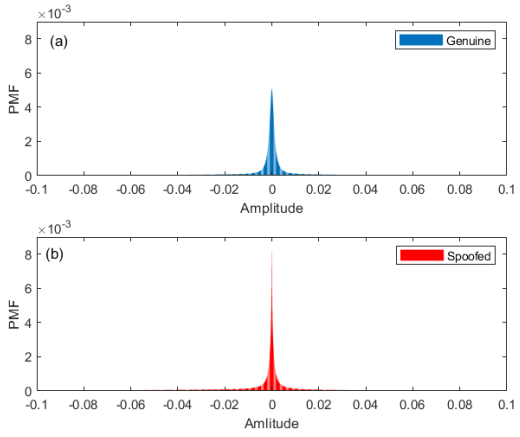

Figure 1: *Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech (no VAD).*

In spoofing detection domain, non-speech parts are known to be informative [30] so usually *voice activity detection* is not

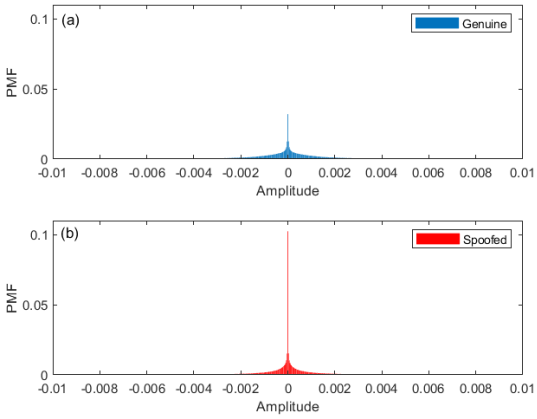

Figure 2: *Waveform amplitude PMFs for physical condition, train set, Genuine (a) and Spoofing (b) speech (no VAD).*

applied. To confirm this fact, we display in Figure 3 information similar to that in Figure 1 but with PMFs computed only on speech parts when Figure 4 proposes the PMFs of the non-speech parts (both experiments are using the same VAD process). We apply a very simple energy VAD, using the same approach as in [31] and [32]. When there is almost no difference between genuine and spoofing speech PMFs for the speech part, a significant difference is observed for the non-speech part. This observation confirms that non-speech is of great importance for current spoofing speech detection systems and raises the question whether this detection is really based on speech characteristics or on non-speech artifacts.

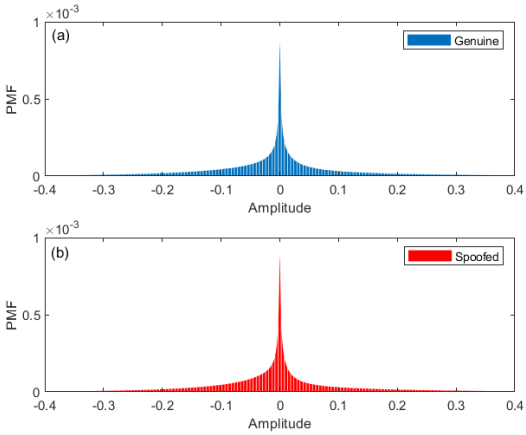

Figure 3: *Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech, speech part only (after applying VAD).*

4. *Genuinization* process

With the two PMFs, one for genuine speech and the other for a single recording of spoofing speech, the *genuinization* process wishes to correct the amplitudes of the spoofing speech samples to obtain a PMF as close as possible to PMF of genuine speech. Given the waveform amplitudes PMF of the genuine

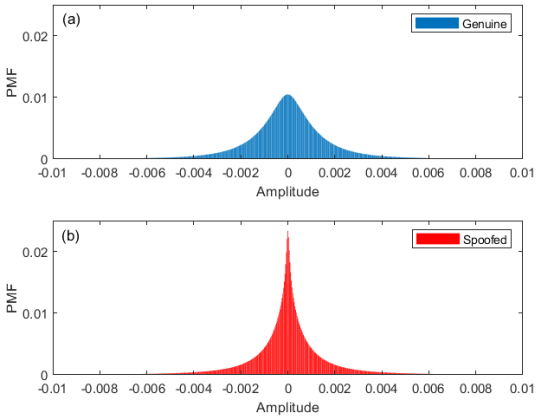

Figure 4: *Waveform amplitude PMFs for logical condition, train set, Genuine (a) and Spoofing (b) speech, non speech part only (after applying VAD).*

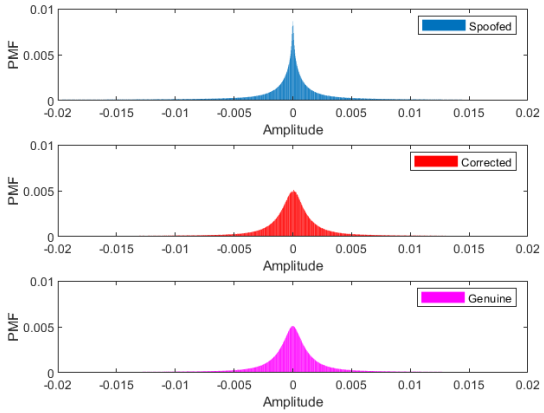

Figure 5: *Waveform amplitude PMFs for logical condition, train set, original spoofing speech (upper), spoofing speech after genuinized (middle) and genuine speech (bottom).*

speech $p_x^g(k)$, where g means genuine; x is a discrete random variable $x \in \{1,\ldots,2^{16}\}$; k is the value that assigned to x (the actual signal's amplitude is $s(n) = -1 + k \cdot 2^{-15}$. Next, the *cumulative distribution function* (CDF) is calculated, $F_x^g(k) = \sum_{q=1}^{k} p_x^g(q)$. For each spoofing speech signal $s(n)$ a PMF $p_x^s(k)$ (s for the spoofed signal) is calculated, followed by the corresponding CDF $F_x^s(k)$. The *genuinization* algorithm is then applied, as described in 1.

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 $q^* = \arg\limits_{q} \{F_x^g(q) = F_x^s(k)\}$
  Set $\hat{s}(n) = -1 + 2^{-15} \cdot q^*$
Return: $\hat{s}(n)$

5. Experiments using *Genuinization*

The following experiments are done on all samples, without using a VAD process. Figure 5 presents, for logical condition, the comparison of the PMFs of genuine speech, spoofing speech before *genuinization* and after *genuinization*. The Figures emphasize a narrow band of amplitudes to facilitate comparisons. The correction of the PMF looks significant. However, this does not mean that this process is able to improve the quality of the spoofing speech or make it more difficult to detect. Subjectively, by listening to several recordings, we think that spoofing files of poor quality sound even worse after *genuinization* while for high quality recordings, no degradation is perceived. As far as genuine files are concerned, the application of *genuinization* has no noticeable effect, according to our subjective opinions.

In the following paragraph, we evaluate the effect of *genuinization* for spoofing speech detection, using the provided baseline system. Two feature sets are evaluated, LFCC and CQCC. The GMMs have 512 mixture components. Spoofing speech detection performance is computed on ASVspoof 2019 development dataset in terms of EER and min-tDCF [29].

Table 3: *Spoofing detection performance using original or genuinized test files.*

		Original	Genuinized
LFCC	**EER [%]**	2.709	1.291
	min-tDCF	0.0663	0.0367
CQCC	**EER [%]**	0.394	3.219
	min-tDCF	0.0112	0.0992

Table 4: *Spoofing detection performance using genuinization to train spoofing model and original or genuinized test files.*

		Original	Genuinized
LFCC	**EER [%]**	34.379	0.048
	min-tDCF	0.6304	0.0015
CQCC	**EER [%]**	43.477	0.007
	min-tDCF	0.8910	0.0001

Table 3 shows the results with spoofing GMM model learned on original training data. With CQCC features, applying *genuinization* to spoofing test data increases the EER by 10 times. It seems to indicate that CQCC features, which are performing about 9 times better than LFCC without *genuinization*, are also more linked to waveform amplitudes information. For LFCC, *genuinization* seems to have a positive effect, with an EER downsized from 2.7% (no *genuinization*) to 1.29% (with *genuinization*). The latter result tends to say that LFCC features are less linked to time domain waveform information, so *genuinization* corresponds here to a classical feature normalization step. As the physical condition experiments do not show the same behavior, more experiments with different conditions are required before we can propose a definitive conclusion.

Table 4 presents the results of a similar experiment, but after applying *genuinization* to train the spoofing GMM. For CQCC, the EER is 43% without and about 0% with *genuinization* applied on the files. For LFCC, applying *genuinization* on spoof-

ing model destroys performance on original test data where the EER drops close to 0 when applied on *genuinization* spoofing test data. Both results tend to confirm our hypotheses about the role of waveform amplitudes for MFCC and CQCC.

We aim in the next paragraphs to evaluate whether our *genuinization* process is sensitive to the speech/non-speech question. To evaluate this sensitivity, we propose to compute our *genuinization* parameters on non-speech parts only, so only the non-speech parts of the signals are used to compute the CDF of the genuine speech on the training set. Then, for the spoofing files of the train and development sets, the procedure described in Algorithm 1 is classically applied (in other words, we changed the targeted distribution, which is now computed only on non-speech data).

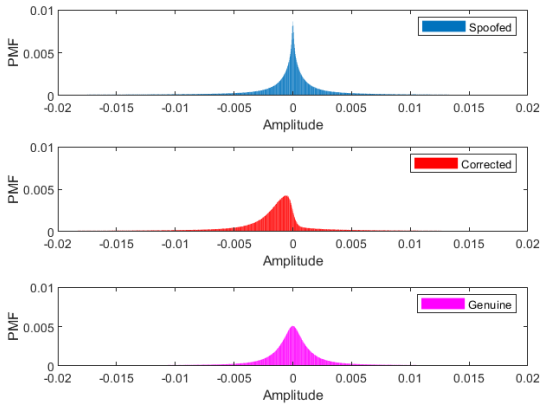

Figure 6: *Waveform amplitude PMFs for logical condition, train set, original spoofing speech (upper), spoofing speech after non-speech-based genuinized (middle) and genuine speech (bottom).*

Figure 6 presents the PMFs of genuine speech, original spoofing speech and spoofing speech after this "non speech-only" *genuinization*. The PMF after *genuinization* is far from being identical to the genuine speech PMF. It is very clear when Figure 6 is compared with Figure 5, where the PMF after *genuinization* was clearly closer to the genuine speech PMF. This is explained by the heavier tails of spoofing speech, particularly the left tail, as showed in Figure 7.

Table 5 presents an experiment close to the one presented in Table 3 but where the *genuinization* parameters are learnt on non-speech data. The results confirm our previous findings for CQCC, with an EER multiplied by a factor of 4. For LFCC, the results are less clear where for the non-speech case we see a degradation of the EER by about 1.2.

Table 5: *Spoofing detection performance using original or genuinized test files, with genuinization parameters estimated on non speech parts.*

		Original	**Genuinized**
LFCC	**EER [%]**	2.709	3.374
	min-tDCF	0.0663	0.0941
CQCC	**EER [%]**	0.394	1.577
	min-tDCF	0.0112	0.0431

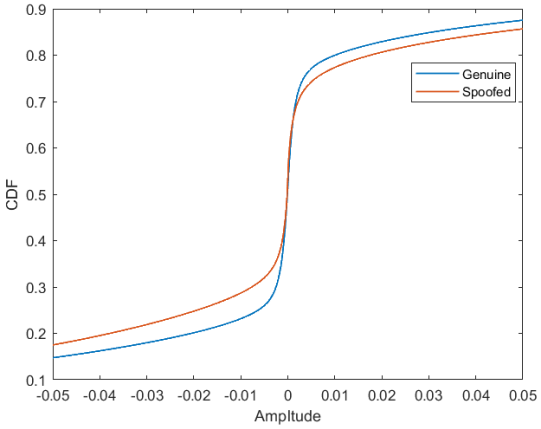

Figure 7: *CDFs of the Logical conditions train set waveform amplitudes for the Genuine and Spoofing speech.*

6. Conclusions

In this work, we investigated the role of time-domain information both for speaker recognition spoofing attacks and spoofing countermeasures questions. Our interest is motivated by the fact that classical spoofing/anti-spoofing approaches usually neglect this domain and focus mainly on the frequency domain.

To start with a simple approach in the time domain, we proposed to exploit the PMFs of the waveform amplitudes. We have shown that for both problems, spoofing and anti-spoofing, more attention should be paid to time domain since large differences in waveform amplitudes PMFs have been observed between genuine and spoofing speech.

Based on this finding, we haved proposed a simple *genuinization* method capable of transforming the spoofing speech to reduce the waveform amplitudes PMF gap between genuine and spoofing speech.

Furthermore, when we examined the performance of ASVspoof 2019 challenge baseline system using LFCC or CQCC features, we found that the system was vulnerable to the time domain changes achieved using *genuinization* process. This is especially true for CQCC features, which appear very sensitive to waveform amplitudes information. We can not generalize it to all anti-spoofing systems, but we can assume that at least some of them are also vulnerable to time domain changes.

Another observation is that the largest PMF difference between genuine and spoofing speech takes place in low amplitudes area, which is mainly linked to non-speech. This result may explain why VAD is rarely used in anti-spoofing systems.

To conclude, to the best of our knowledge, it is one of the first works which uses the waveform amplitudes distribution to analyze genuine speech versus spoofing speech. Even if our method is simple, maybe oversimple, it allows us to open a large number of questions. We also plan to examine the effects on the replayed speech and see how our *genuinization* idea can be applied as pre-processing, before replaying the data, in order to reduce the gap between genuine speech and replayed speech.

7. Acknowledgement

This work was supported by the ANR–JST CRES VoicePersonae project

8. 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. D. Leon, *Speaker Recognition Anti-spoofing*. London: Springer London, 2014, pp. 125–146.

[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, Automatic Speaker Verification Spoofing and Countermeasures Challenge, colocated with INTERSPEECH 2015, September 6-10, 2015, Dresden, Germany*, Dresden, GERMANY, 09 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, NÂ°4), June 2017*, 02 2017.

[7] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech 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 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 parkinsonâ€™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] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *ODYSSEY 2001 -The Speaker and Language Recognition Workshop*, Crete, Greece, June 2001.

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

[30] 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.

[31] 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.

[32] 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.