On the contribution of the voice texture for speech spoofing detection

1st Raoudha Rahmeni
University of Sfax
National School of Engineers of Sfax (ENIS)
Sfax, Tunisia
raoudha.rahmeni@gmail.com

2nd Anis Ben Aicha†‡
University of Carthage
† Higher School of Communications
of Tunis (SUPCOM), COSIM Laboratory
‡ Faculty of Sciences of Bizerte
Tunisia
anis.benaicha@supcom.tn

3rd Yassine Ben Ayed
University of Sfax
Higher Institute of Computer
Science and Multimedia (ISIMS)
MIRACL Laboratory
Sfax, Tunisia
yassine.benayed@gmail.com

***Abstract*—Automatic speech verification (ASV) consists in the implementation of automatic algorithm to measure and asses human biometric parameters. Serious vulnerabilities are emphasized concerning spoofing attacks. In this paper, we propose to investigate the possible contributions of voice textures to detect spoofing attack. Voice texture is a recent concept of voices characterization using an overall sound homogeneity. Well known spoofing attacks are based on text to speech (TTS), voice conversion (VC) and replay techniques. According to the nature itself of the spoofed voices, their textures are different from those of genuine speeches. The concept of the texture is well developed in the context of image processing. Local binary patterns (LBP) is one of the famous visual descriptor of the images. LBP was adapted to be used as speech texture descriptor. LBP coding is applied to all input genuine and spoofed signals. after that, histogram of the LBP descriptors is constructed and is used as features. Support Vector Machines (SVM) classifier is used to classify the obtained features as genuine or spoofed. From the experimental results, it is observed that the proposed method could increase the difference between genuine and spoofed speech.**

***Index Terms*—ASV, Speech spoofing, Voice texture, Machine learning**

I. Introduction

Automatic speaker verification (ASV) technologies are based on biometric approaches to verify the human speaker identity. Such techniques have reaches amount of maturity to be adopted for mass-market adoption. Many products are developed such as efficient authentication in smart phone or e-commerce for example. ASV could be efficiently used in many other applications such as dialling, telephone banking, telephone shopping and password reset systems where users access a service remotely from any location.

Even ASV technique presents a low cost and convenient approach for human identification, the reliability remains a concern [5], [9]. In fact, ASV based systems are vulnerable to spoofing attacks. We mention impersonation, replayed speech, synthesized speech (SS) and voice conversion (VC) [10]. It was shown that these attacks provoke significant increases in the false acceptance rate of state-of-the-art ASV systems. In order to limit the scope of the current research, we only focus on SS and VC attacks.

Recently, many researchers are interested to investigate how much ASV systems are vulnerable to spoofing attacks. Hence, various spoofing countermeasures are elaborated either for dedicated attacks or claimed to be generally applicable [11]. Generally, spoofing countermeasures are in fact pattern recognition task. From the speech utterance, features capturing the artefacts generated by spoofing attacks are captured. These features are used in training phase in order to develop a model able to distinguish between genuine and spoofed speech [10]–[12].

In the current research, we propose a new features to characterize genuine and spoofed speeches. In fact we are inspired from image texture analysis. In the field of image processing, image texture designates the spatial arrangement of pixel intensity or colour [13]. Due to the nature itselves of spoofed speeches generated either by SS or CV techniques, some artefacts can be detected. We think that the repartition of speech components in time-frequency domain are not the same when regarding genuine and spoofed speeches. The spectrogram is one way to represent signals in both time and frequency domain. Such representation can reveal the texture difference. We propose to investigate the potential of the texture to detect the differences spatially artefacts generated by SS or CV algorithms. In the present work we have used as texture method analysis the well-known Local binary patterns (LBP) [7] . LBP is a visual descriptor used for classification of images and it is found to be a powerful feature for texture classification.

The remainder of the paper is organized as following. Section II presents the LBP foundation. In section III, we detail the motivation and the idea of the current framework. In section IV, the used database is presented. Section V presents an analysis of the speech texture in terms of LBP. Section VI studies the pertinent LBP features. Section VII is reserved to experimental results.

II. Conventional LBP for image texture measure

In the real world textures are not uniform [1] , due to variations in orientation, scale, or other visual appearance. The gray scale invariance is important due to fitful illumination

978-1-7281-1292-3/19/$31.00 ©2019 IEEE

or great within class variability. The basic idea of the LBP approach [2] is the independence of the local differences of the central pixel and its neighbours and the central pixel itself. The LBP method was first proposed by Ojala et al [7], [8] to encode the pixel-wise information in textured images [3]. Images are probed locally by sampling greyscale values at a central point $x_{0,0}$ and p points $x_{r,0},...,x_{r,p-1}$ spaced equidistantly around a circle of radius p centered at $x_{0,0}$. So p is a spaced pixels on a circle of radius $r(r>0)$ that form a circularly symmetric neighbor set. If the coordinates of x are $(0,0)$ then the coordinates of x_r are given by $(-R\sin(\frac{2\pi p}{P}),R\cos(\frac{2\pi p}{P}))$ .The gray values of neighbors which do not drop down exactly in the center of pixels are estimated using the interpolation. In the conventional LBP approach [7], [8], the image pixels are first designated as a binary class by thresholding the difference between the center pixel and its neighbours using the function $s(x)$.

$$s(x)=\left\{\begin{array}{ccc} s(x) & = & 1 \; when \; x \; \geq \; 1 \\ s(x) & = & 0 \; when \; x \; < \; 0 \end{array}\right\} \tag{1}$$

The neighboring labels are concatenated and used a unique descriptor for each pattern. For example, 01010100 is uniform patterns. The histogram of the uniform patterns in the whole image is used as the feature vector. It has been proven to be effective for both face recognition and facial expression recognition applications.

Given an $N\times M$ image I, let $LBP_{p,r}(i,j)$ be the identified LBP pattern of each pixel (i, j) by the following equation.

$$LBP_{p,r}=\sum_{n=0}^{p-1}s(x_{r,n}-x_{0,0})2^n \tag{2}$$

then the whole texture image is represented by an histogram vector $\underline{\mathrm{h}}$ of length K.

$$\underline{h}(k)=\sum_{i=1}^{N}\sum_{j=1}^{M}\delta(LBP_{p,r}(i,j)-k) \tag{3}$$

where $0\leq k\leq K-1$, and $K=2^p$ is the number of all the LBP codes. nevertheless, the basic LBP operator produces rather long histograms (2^p distinct values), and it becomes an incurable problem to estimate $\underline{h}$ due to the overwhelming dimensionality of $\underline{h}$ with large p.

III. Motivation and ideas

We are interested in the spoofing attacks based on synthesized and conversion techniques. The spoofed speech is therefore generated artificially using synthesized or conversion algorithms. The genuine human speech is well complicated and non-stationary signal. In the literature, the antispoofing developed techniques are based principally on the comparison of the genuine and the spoofed speeches [4]. More specifically, the main architecture of developed techniques is based on two steps. The first one is the extraction of some features from the speech utterance. The second one is the use of classification techniques to identify the spoofed speech from the genuine one [5], [6].

As mentioned before the terms of texture is well defined and large investigated in the field of image processing. However, for the current framework, we are interested in the voice signal which is one dimension signal. And hence, as a first attempt, we have thought to adapt the voice signal to the context of image processing. In another world, we seek to transform and represent the speech utterance in two dimensions instead of one. By nature itself of spoofed speech, it will be possible to identify some artefacts in the new representation so it will be possible to discriminate the genuine speech from spoofed one. The whole principle of the idea is depicted in Fig. 1.

As speech is a non-stationary signal, there is no meaning to have a statistical study of it. Hence, the speech utterance is framed into frames no longer than 30 ms, where it is possible to assume it as stationary. For each frame, we compute the power spectrum density (PSD). The whole PSD obtained from the speech are concatenated to form one matrix with N rows representing the time index of frames and K columns representing the frequency index. The process is the same as the spectrogram. However, here we control the overlapping of the frames. The amplitude of the obtained spectrogram is then represented in logarithmic scale. The idea behind, is that the weak frequency components will not be neglected. The obtained spectrogram is represented in the Figure 2. As we can remark, there is a specific texture of the spectrogram. The main idea of the current work stems from the analysis of the texture represented by the spectrogram. We think that with appropriate tools of textures analysis, it will be possible to discriminate genuine speech from spoofed ones. We use here the LBP technique to characterize obtained spectrogram. For this purpose, we rescale the spectrogram to get a range from 0 to 255 which corresponds to the image grey levels. The LBP transformation is applied according to the method elaborated in the section II. From the LBP image, we compute the LBP features vector.

IV. Used material

The ASVspoof challenge is taken during the 2015 edition of INTERSPEECH in Dresden (Germany). It gives a database based on “SAS corpus” which is a previous corpus and it was developed using a distanctive 10 voice conversion and speech synthesis systems. The countermeasure performance and the assessments of vulnerabilities to spoofing are supported by this challenge. The dataset consist of genuine and spoofed speech and is divided into three subsets (training,development and evaluation). 106 human speakers (45 male and 61 female) give genuine speech recorded with no modification, background noise effects or significant channel. On the other hand the original genuine speech is modified by using some voice conversion and speech synthesis algorithms to obtain spoofed speech. The number of speakers in each subset is illustrated in Table IV-C.

A. Training subset

The systems which distinguish between genuine and spoofed speech are learned or trained by using audio from the

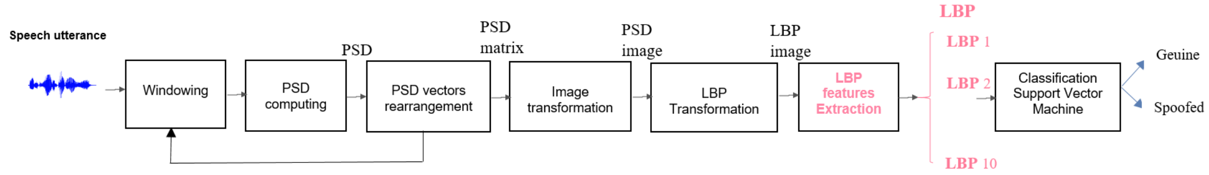

Fig. 1. Flowchart of the proposed idea.

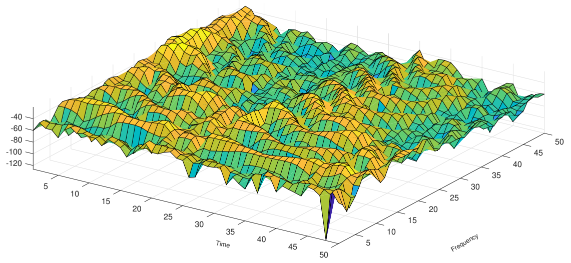

Fig. 2. 3D representation of the spectrogram of genuine voice utterance.

training dataset. This dataset includes spoofed and genuine speech from 25 speakers (10 male, 15 female).the spoofed utterance is developed using two speech synthesis Which are implemented with the hidden Markov model and three voice conversion algorithms whiche are based on frame selection, spectral slope shifting and an available voice conversion toolkit within the Festvox system.

B. Developement Subset

The spoofing detection algorithms are developed using the audio of the developement dataset which used for the optimisation and the design countermeasures. It includes both genuine and spoofed speech from a subset of 35 speakers (15 male, 20 female). Talking about the spoofed speech is developed using one of the five spoofing algorithms which are used for the training dataset.

C. Evaluation Subset

The evaluation data is comprised a genuine and spoofed utterances collected from 46 speakers (20 male, 26 female). The same recording conditions for genuine speech are used the training, development and the evaluation sets. However, spoofed data are developd using different spoofing algorithms. The same algorithms to generate the developement dataset are used in addition to others which reffered to as unknown spoofing algorithms.

V. "Voice texture" analysis in terms of LBP

We aim to emphasis at the starting that all speeches utterances are normalized by their standard deviation. Hence, treated speeches are all unit energy. The experimental protocol

TABLE I
Number of non-overlapping target speakers and utterances in the training, development and evaluation datasets.

subset	Speakers		Utterances	
	male	female	Genuine	Spoofed
Training	10	15	3750	12625
Developement	15	20	3497	49875
Evaluation	20	26	9404	184000

depicted in figure 3 is conducted to compute the spectrogram and the LBP transform of the spectrograms. The two sequences are related to the same person. Figure 3. a represents the histogram computed from the spectrogram of the two speeches sequences genuine and spoofed respectively. As we can remark, the two histograms are different. We can at least expect such results, since the histogram computed from the spectrogram represents the frequency of the apparition of a certain grey level in the spectrogram. This quantity depends from the phonemes of the speech. So, it doesnt reveal anything about the nature of the analysed speech. However it leads as to more investigate the distribution and the articulation between different frequency components. This task can be observed on the LBP image computed from the spectrogram. LBP transformation of the spectrogram overcomes the grey level distribution and represents only the texture. Figure 3.b, represents the histogram computed from the LBP images of both genuine and spoofed spectrograms. As we can see, a noticeable difference can be picked up. This is means clearly that is possible to consider the LBP image computed from the spectrogram as a feature to discriminate between genuine and spoofed speech.

VI. Analysis of extracted LBP features

LBP transformation transform the spectrogram to an image when the grey levels of pixels are replaced by an LBP coefficient related to the local texture of the considered pixel. We propose to proceed as Ojala in [14] to compute from the whole LBP image a set of features reflecting the texture of the image. The LBP image is devided into cells. Then the LBP histogram is computed for each cell. The LBP features are obtained by computing L1 norm of histograms cells. We represent in figure 4 the two LBP features relative to the genuine and spoofed speech. As we expected there is some differences between the two LBP features. This finding joint

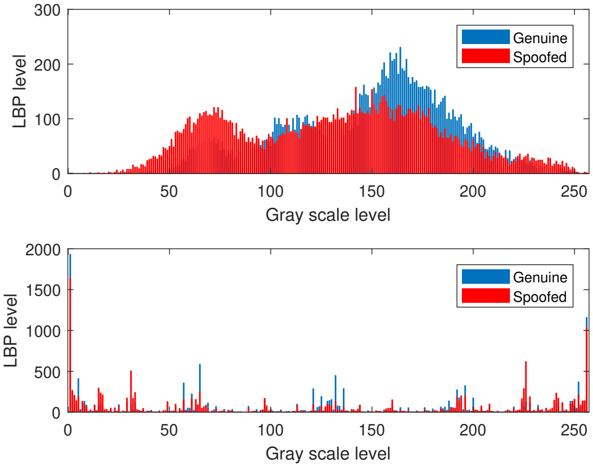

Fig. 3. Representation of histograms : (a) Spectrum based and (b) LBP based

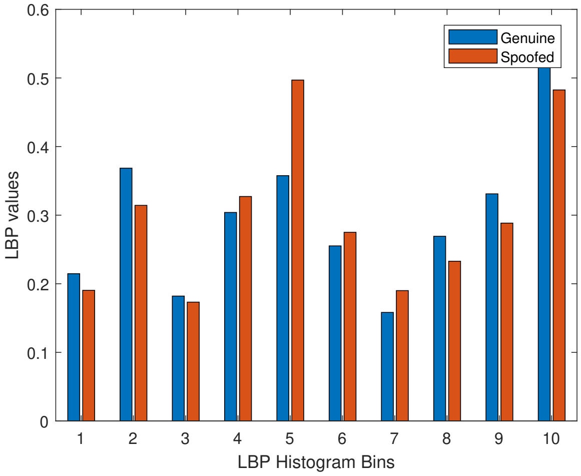

Fig. 4. LBP features of both genuine and spoofed speeches.

those observed from figure 3. Thus, it is possible to use LBP features as discriminant features to verify spoofing speech.

The two sets of LBP features are analysed using boxplot toolbox. It is a graphical representation of criteria that shows lowest value, highest value, median value, and the size of the first and the third quartile in the same graph [15]. This box plot is used in order to discard descriptors with low separation ability between the genuine speeches and the spoofed ones. In fact, obvious differences are immediately apparent.

Figure 5 represents the boxplots of the 10 extracted features of both genuine and spoofed speeches. Discarded features are those which their boxplots overlap. In such situation, genuine LBP feature and spoofed LBP feature range almost in the same interval. These kind of features are considered non discriminant. As example the variability of LBP6 is almost the same for the two cases genuine and spoofed speeches. We expect that such features will not be a discriminant one. On the other hand features such LBP3 vary differently for the genuine and spoofed speeches. We expect that such feature will be a discriminant one.

VII. Experimental results

For our classification experiments, we opted for an SVM for its discriminant properties. It is one of the most widely used data learning tools in recent years. SVMs, or Support Vector Machines, is originally developed and widely used for pattern recognition or classification [16], [17]. The database is divided into two sets. 80% of speech utterances are reserved for the training and the remainder are used as test dataset. As first experiments, we feed the SVM module with only one features. The classification results are presented in the table II . We remark that LBP3 feature perform more better that LBP6 feature. This joins again our finding when boxplot is used. We have proposed then to combine the best features in order to find the best combination in terms of accuracy. The Table III represents the tested combination of LBP features. The best results are obtained when LBP3 are used with an accuracy reached 0.7167. we remark the same accuracy value obtained as LBP3 when we combine the features (LBP3.8.5.2.9) ,(LBP3.8.5.2.9.10.4) and (LBP3.8.5.2.9.10.4.1).

We like to mention that the objective of the current framework is to prove that it is possible to exploit the “speech texture as a discriminant feature for speech spoofing verification. As a first attempt we have used the LBP technique as “speech texture measure. We are conscious and we recognize that more sophisticated techniques can be developed to characterize and measure the “speech texture. We think that with such tools it will be possible to reach more accurate results.

TABLE II
Classification result using SVM over LBP features.

	Accuracy
LBP1	0.6000
LBP2	0.6167
LBP3	0.7167
LBP4	0.6167
LBP5	0.6667
LBP6	0.5333
LBP7	0.6000
LBP8	0.6833
LBP9	0.6167
LBP10	0.6167

TABLE III
The tested combination of LBP features.

	Accuracy
LBP3	0.7167
LBP3,LBP8	0.5000
LBP3,LBP8,LBP5	0.6167
LBP3,LBP8,LBP5,LBP2	0.6000
LBP3,LBP8,LBP5,LBP2,LBP4	0.7167
LBP3,LBP8,LBP5,LBP2,LBP4,LBP9	0.6170
LBP3,LBP8,LBP5,LBP2,LBP4,LBP9,LBP10	0.7167
LBP3,LBP8,LBP5,LBP2,LBP4,LBP9,LBP10,LBP1	0.7167
LBP3,LBP8,LBP5,LBP2,LBP4,LBP9,LBP10,LBP1,LBP7	0.7000
LBP3,LBP8,LBP5,LBP2,LBP4,LBP9,LBP10,LBP1,LBP7,LBP6	0.6833

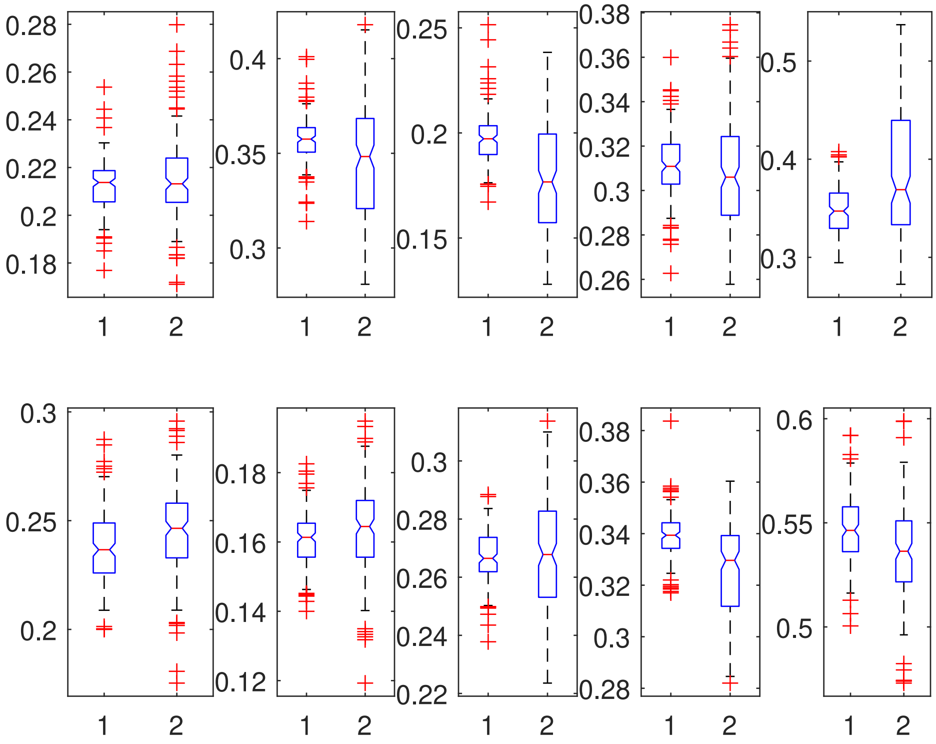

Fig. 5. LBP Boxplots. From the left top corner to the bottom right corner : LBP1...LBP10

VIII. Conclusion

In the current paper the usefulness of the "speech texture is investigated. Inspired from image field, we have computed the texture of the speech in time-frequency domain. The textures of both genuine and speech signals were analysed. It is found that the texture can be useful to be used as discriminant feature for spoofed speech verification. The LBP technique is used to extract features from transformed spectrogram. LBP features are used as SVM inputs. Experimental results proves the validity of the idea.

References

[1] DC. He and L. Wang, "Texture Unit, Texture Spectrum, And Texture Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, pp. 509 – 512, 1990.

[2] M. Arya, N. Mittal and G. Singh, "Texture-based feature extraction of smear images for the detection of cervical cancer," *IET Computer Vision*, 2018.

[3] M. Kiechle, M. Storath, A. Weinmann and M. Kleinsteuber, "Model-based learning of local image features for unsupervised texture segmentation," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1994–2007, 2018.

[4] A. Poddar,M. Sahidullah and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91-101, 2017.

[5] W. Z. Yamagishi *et al*, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588-604, 2017.

[6] C. Hanili, "Data selection for i-vector based automatic speaker verification anti-spoofing," *Digital Signal Processing*, vol. 72, pp. 171-180, 2018.

[7] T. Ojala, M. Pietikinen and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29 no. 1, pp. 51-59, 1996.

[8] M. Pietikinen and T. Ojala, "Texture analysis in industrial applications," *Image Technology*, pp. 337-359, Springer, 1996.

[9] T. Kinnunen *et al*, "Assessing the limits of replay spoofing attack detection," *ASVspoof 2017 challenge*, 2017.

[10] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311-5327, 2016.

[11] Z. Wu, N. Evens, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communications*, vol. 66, pp. 130-153, 2014.

[12] M. Sahidullah, T. Kinnunen and C. Hanilci, "A comparison of features for synthetic speech detection," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[13] R. M. Haralick, and K. Shanmugam, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610-621, 1973.

[14] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution Gray Scale and Rotation Invariant Texture Classification With Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.

[15] Y. Sun and M. G. Genton, Functional boxplots, Journal of Computational and Graphical Statistics, vol. 20, no. 2, pp. 316-334, 2011.

[16] B. Scholkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond," *MIT press*, 2001.

[17] A. Ben Aicha, "Noninvasive Detection of Potentially Precancerous Lesions of Vocal Fold Based on Glottal Wave Signal and SVM Approaches," in *Procedia Computer Science*, vol. 126, pp. 586-595, 2018.