Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 668–675

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Speech spoofing countermeasures based on source voice analysis and machine learning techniques

Raoudha Rahmeni[a,*], Anis Ben Aicha[(1,2)b], Yassine Ben Ayed[c]

[a] *University of Sfax, National School of Engineers of Sfax (ENIS), Tunisia*
[b] *University of Carthage, [(1)] Higher School of Communications of Tunis (SUPCOM), LR11TIC04 COSIM research Laboratory, [(2)] Faculty of Sciences of Bizerte (FSB), Tunisia*
[c] *University of Sfax, Higher Institute of Computer Sciences and Multimedia (ISIMS), Tunisia*

Abstract

Automatic speaker verification (ASV) [7] systems are susceptible to malicious attacks. It discredit the performance of a standard ASV system by increasing its false acceptance rates. This paper presents a new countermeasure for the protection of automatic speaker verification systems from spoofed signals. The new countermeasure is based on the analysis of a sequence of acoustic feature vectors using the glottal inverse filtering. In the proposed method, speech is decomposed into a glottal source signal and model the vocal tract filter through glottal inverse filtering. The IAIF desriptors are constructed and are used as features. Support Vector Machines (SVM) classifier and Extreme learning machine (ELM) are used to classify the obtained features as genuine or spoofed. It is hoped that the proposed method can help to detect the genuine speech from the spoofed one.

Keywords: Speech spoofing ; Anti-spoofing countermeasures ; Glottal flow ; Machine learning

1. Introduction

Over the last years, a lot of progress has been seen in the study of voice biometrics. Automatic speaker verification system (ASV) [3, 4] have become increasingly popular. It's widely acknowledged to be vulnerable to spoofing attacks such as impersonation , replay attacks, voice conversion and speech synthesis. All provoke significant increases in a binary decision which is made for accepting or rejecting a claimed identity. It is only almost recently that the community has investigated spoofing countermeasures. Thus, various spoofing

* Raoudha Rahmeni ,Anis Ben Aicha, Yassine Ben Ayed. Tel.: +216-955-831-05,+216-526-391-14,+216-956-394-98
E-mail address: raoudha.rahmeni@gmail.com,anis.benaicha@supcom.tn,yassine.benayed@gmail.com

10.1016/j.procs.2019.09.222

countermeasures are elaborated either for dedicated attacks or claimed to be generally applicable. This paper presents a new countermeasure which aims to provide a more universal spoofing countermeasure which is less dependent on prior knowledge, i.e. not specific to a given attack. It is based on characteristics of a sequence of features vectors captured using IAIF [8] method. To limit the scope of the current research, we only focus on SS and VC attacks. In this paper we will first present the glottal flow parameters extraction, and then we detail the motivation and the idea of the proposed countermeasure. in Section 4, we will present the details of the used database. In Section 5 we will analyse the gottal flow parameters. Section 6 will describe the experimental results and finally, Section 7 will conclude the paper.

2. Glottal flow parameters extraction

Speech is a result of filtering the glottal flow using the vocal tract cavities, and converting the resulting velocity flow into pressure at the lip.The glottal flow is estimated from the speech signal contrary to the conventional extraction of filter based features such as MFCC parameters, which is carried out asynchronously. The estimation and parameterization of the glottal source involves the processing of speech frames whose duration is proportional to the pitch period. In order to get an initial estimate of the vocal tract, the method Iterative Adaptive Inverse Filtering (IAIF) [9] is used and is a semi-automatic inverse filtering method developed by Paavo Alku. It takes a speech pressure signal as input and generates an estimate of the corresponding glottal flow signal [10][12]. Using the IAIF method , the speech spectrum can be estimated as a low-order all-pole process, which is computed pitch-asynchronously over several fundamental periods, from the glottal excitation. As shown in fig 1, IAIF [11], [13] is composed of two main phases. First one, a first-order all-pole model is estimated to get a preliminary estimate of the glottal flow. This contribution is then removed from the original speech signal using inverse filtering. Second, to estimate the contribution of the vocal tract, a higher-order all-pole model was computed on the inverse-filtered signal. So to have the glottal flow signal, the original speech signal is filtered with the inverse of the estimated vocal tract filter. The method operates in two repetitions, but to get a more accurate estimate of the glottal contribution, an imperceptibly higher order model replace the first order all-pole model. We obtain finally as an output of the IAIF method , the glottal flow signal and the vocal tract estimated. As we say before, analysis of voice production with inverse filtering comprises usually two stages: In the first one, the compunting of the estimates of the glottal flow. The second stage of the analysis, the parameterisation stage. The features that are extracted from the glottal signal are divided into three groups: time domain, frequency domain, and those that represent the variations of the fundamental frequency.

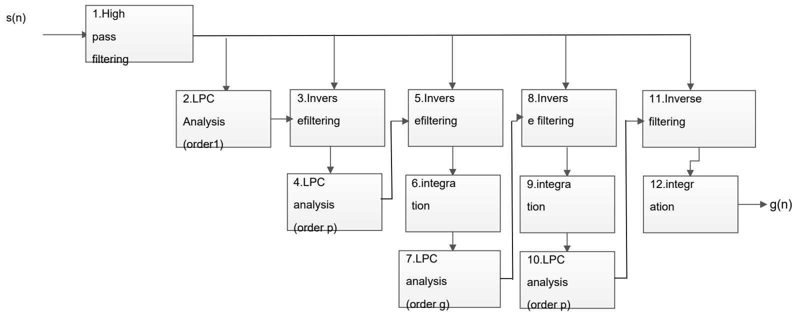

Fig. 1: Block diagram of the IAIF method

3. Idea and motivation

In this work we present a novel countermeasure against spoofing using voice conversion and speech synthesis techniques. The main idea of the current work is based on 2 steps. The first step is the extraction of features from the speech utterance. So, as shown in ,the speech utterance is windowed and framed into frames no longer than 30 ms. We take only the voiced frame to apply the IAIF method. For each frame, we compute teh IAIF descriptors and than we concatenate them to form one matrix with row represent the descriptors and the column represent the glottal flow. After extrating the features we need to clean it.These cleaned features are used as input of second step. To identify the spoofed speech from the genuine, SVM and ELM techniques are used as a classification techniques. We apply teh same process for the MFCC features [15], [16], [17]. The proposed method is based on analysis of IAIF method and MFCC features. We think that with concatenating the IAIF descriptors and the MFCC features we can obtain a good result.

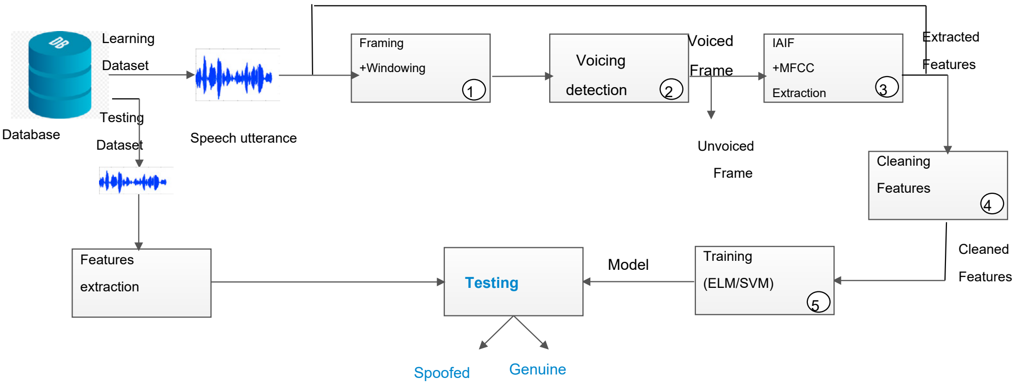

Fig. 2: Flowchart of the proposed idea.

4. Used material

The speech parameterizations of interest were evaluated on the SAS corpus. This database is taken from the ASVspoof challenge during the 2015 edition of INTERSPEECH in Dresden (Germany). Its well-understood and widely-used the challenge to support the countermeasure performance and the assessments of vulnerabilities to spoofing . The database was developed using a different 10 voice conversion and speech synthesis systems. So we obtain a spoofed speech which is modified from the original genuine speech data. Genuine speech is recorded from 106 human speakers (45 male and 61 female) without any modification, and without significant channel or background noise effects. The entire dataset is subdivised into three subsets respectively training, development and evaluation in Table 1.

Table 1: 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

4.1. Training subset

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

4.2. 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.

4.3. 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.

5. Glottal flow parameters analysis

In voice production studies, the selection of the parameterisation method of the glottal flow is a fundamental part. So we can successfully quantify voice production with inverse filtering to know the different alternatives available in the parameterisation stage and to know on which features of the glottal flow these measures focus. Now it's possible to select the "best" numerical measure that reflects the behaviour of the voice source in the experiment. The most aboveboard method to parameterise the glottal flow waveform obtained by inverse filtering is to extract certain critical time-spans and amplitude values from the timedomain flow signal or from its first derivative.In our work, we use the time based golttal flow parameters classical like opening quotient (OQ). It is defined as the ratio of the opening phase length to the total length of the glottal cycle (or period) and it is inversely proportional to the intensity of the voice.

$$OQ = \frac{T_{01} + T_c}{T} \tag{1}$$

Some authors consider two parameters for the OQ, OQ_1 and OQ_2, which are defined as

$$OQ_1 = \frac{T_{01+Tc}}{T} \tag{2}$$

and

$$OQ_2 = \frac{T_{02+Tc}}{T} \tag{3}$$

where T_{02} corresponds to the time between the inflection point and maximum. The Speed quotient (SQ) is defined as the ratio between the time interval of the opening phase and the time interval of the closing phase and it is calculated as

$$SQ = \frac{T_{01}}{T_c} \tag{4}$$

Some authors consider two parameters for the speed quotient, SQ_1 and SQ_2, as defined respectively.

$$SQ1 = \frac{T - T_c}{T_c} \tag{5}$$

and

$$SQ2 = \frac{T_{01}}{T_c} \tag{6}$$

The closing quotient (CIQ) is defined as the ratio between the time interval of the closing phase and the time interval of one complete cycle of the vocal folds. The CIQ is calculated as

$$CIQ = \frac{T_c}{T} \tag{7}$$

the parameter amplitude quotient (AQ) is defined as the ratio between the maximum value of the glottal signal, which is denoted by A_v and the minimum of the derivative of the flow, which is denoted by d_{peak}=min $\frac{d_{g(t)}}{dt}$ during a glottal cycle and it is calculated as

$$AQ = \frac{A_v}{d_{peak}} \tag{8}$$

and we have the normalized amplitude quotient (NAQ) which is the ratio between the AQ and total time length of the glottal pulse (T). The NAQ is calculated as

$$NAQ = \frac{AQ}{T} \tag{9}$$

After obtaining these descritors we need to normalize an clean it to use it as input of training phase as mentioned in fig 3. We neglagte the value which is deviate from the values of other observations, abnormally low or high using the rule 1.5 x interquartile gap. So we obtain only a cleaned descriptors.

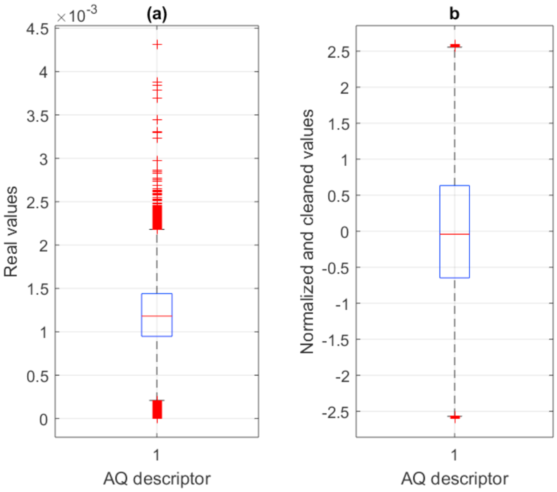

Fig. 3: Cleaned AQ descriptor. (a) real values of AQ obtained by IAIF (b) normalized and cleaned values of AQ.

6. Experimental results

For our classification experiments, we opted for an SVM [5] for its discriminant properties and an ELM. SVM [6], or Support Vector Machines, is originally developed and widely used for pattern recognition or classification. Extreme learning machine [19], [18] is a type of neural network. Its specificity is to have only one layer of hidden nodes, where the weights of the hidden node connection entries are randomly distributed and never updated. Our database is divided into two sets. 80% of speech utterances are reserved for the training and the remainder are used as test dataset. After obtaining IAIF's descriptors, we use the SVM for classififcation. As we know the SVM is based on kernel like lienar, RBF , polynomial. We remark that the best result obtained is for the polynomial kernel As shown in fig 4.

On the other hand,use use the ELM. We fix the activation function and than we variate the number of cells in hidden neurons. We compare between the MFCC features and the IAIF descriptors as illustared in

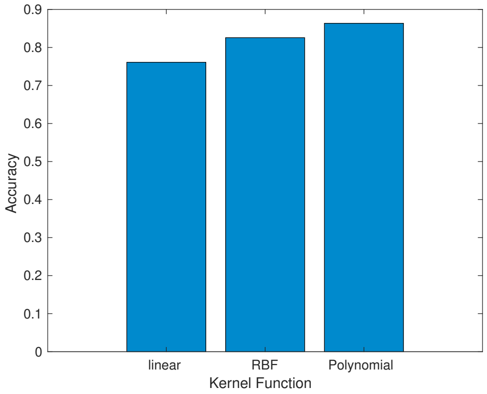

Fig. 4: Classification using Kernel's SVM and IAIF

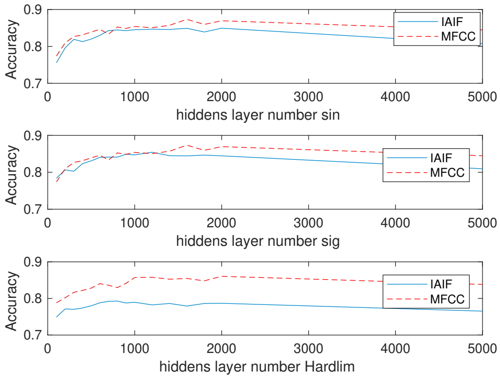

Fig. 5: Comparaison between mfcc and iaif with fixing kernel of ELM

fig 5. We recognize that the best results are obtained for the MFCC features.The best result is obtained when the number of cells in the hidden layer for ELM is **1800** for both the activation function **sig** and **sin** and is **2000** for the activation function **hardlim**. The results of classification for the data set using MFCC features by ELM and SVM method are presented in Table 2. Showing that SVM is superior to ELM method in accuracy. The best results are obtained when S2 are used with an accuracy reached 0.9329.

Table 2: CLASSIFICATION RESULTS USING MFCC DESCRIPTORS OVER SVM AND ELM .

subset	Accuracy	
	ELM	SVM
G,S1	0.5006	0.8630
G,S2	0.5135	**0.9329**
G,S3	0.5121	0.9189
G,S4	0.4949	0.9208
G,S5	0.4921	0.8754

Table 3: CLASSIFICATION RESULTS USING IAIF DESCRIPTORS OVER SVM AND ELM .

subset	Accuracy	
	ELM	SVM
G,S1	0.8407	**0.8635**
G,S2	0.8302	0.7152
G,S3	0.7437	0.5912
G,S4	0.7522	0.5980
G,S5	0.7748	0.7225

Table 4: CLASSIFICATION RESULTS USING IAIF DESCRIPTORS OVER SVM AND ELM .

subset	Accuracy	
	ELM	SVM
G,S1	0.7888	0.9005
G,S2	0.9064	**0.9578**
G,S3	0.7843	0.9050
G,S4	0.7908	0.9092
G,S5	0.7811	0.8731

Let's talk about the IAIF descriptors.As shown in Table 4.As the MFCC features, SVM is more efficient than ELM in term of accuracy. It has the best precision . So we obtain the best results when S1 are used with an accuracy reached 0.8635. Now we combine the IAIF descriptors and the MFCC features then we apply the SVM and the ELM. As illstrated in Table 4, we obtain the best result when we apply S2 for an accuracy equal to 0.9578.

7. Conclusions

This paper reports a new countermeasure for the protection of automatic speaker verification (ASV) systems from spoofing.We used 2 types of features (the IAIF descriptors and the MFCC features). The glottal flow parameters were extracted and analysedThese features are used as SVM inputs ans ELM inputs. Our system performed verry well on the combination between both of features with an Accuracy 0.9578 in the case of the SVM classification.

References

[1] T. Kinnunen *et al*, "A Spoofing Benchmark for the 2018 Voice Conversion Challenge: Leveraging from Spoofing Countermeasures for Speech Artifact Assessment," *arXiv preprint arXiv:1804.08438*, 2018.

[2] C. Chih-Chung and L. Chih-Jen, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp.1–27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109-116, 1992.

[9] K. E. Cummings and M. A. Clements, “Analysis of the glottal excitation of emotionally styled and stressed speech," *Journal of Acoustic Society of America*, vol. 98, no. 2-3, pp. 88-98, 1995.

[10] D.G.Childers, “Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127-138, 1995.

[11] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit “Detection of glottal closure instants from speech signals: A quantitative review ," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994-1006, 2012.

[12] T. Drugman, P. Alku, A. Alwan, B. Yegnanarayana, “Glottal source processing: from analysis to applications," *Computer Speech and Language*, vol. 98, no. 5, pp. 1117-1138, 2014.

[13] H. Auvinen, T. Raitio, M. Airaksinen, S. Siltanen, B. Story, P. Alku, “Automatic glottal inverse filtering with theMarkov chain Monte Carlo method," *Computer Speech and Language*, vol. 28, no. 5, pp. 1139-1155, 2014.

[14] K. Sri Rama Murty, B. Yegnanarayana, “Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition," *IEEE SIGNAL PROCESSING LETTERS*, vol. 13, no. 1, pp. 1139-1155, 2006.

[15] J. Chen , K. K. Paliwal, M. Mizumachi and S. Nakamura, “Robust mfccs derived from differentiated power spectrum," *Eurospeech*, vol. 2, no. 1, pp. 577-582, 2001.

[16] M. Kalamani, M. Krishnamoorthi, R.S. Valarmathi, “Delta Mel Frequency Cepstral Coefficient based Feature Extraction Algorithm for Continuous Tamil Speech Recognition," *Digital Signal Processing*, vol. 9, no. 9, pp. 577-582, 2017.

[17] M.A.Hossan, S.Memon, M.A.Gregory, “A novel approach for MFCC feature extraction," *IEEE fourth international conference on signal processing and communication systems*,vol. 2, no. 1, pp. 1-5, 2010.

[18] G.-B.Huang, C.-K.Siew, “Extreme Learning Machine with Randomly Assigned RBF Kernels," *ICARCV*,vol. 11, no. 1, pp. 489-501, 2004.

[19] G.-B.Huang, Q.-Y.Zhu, C.-K.Siew, “Extreme learning machine: Theory and applications," *Neurocomputing*,vol. 70, no. 1, pp. 489-501, 2006.