IberSPEECH 2022
14-16 November 2022, Granada, Spain

The BioVoz Project: Secure Speech Biometrics by Deep Processing Techniques

Antonio M. Peinado, Alejandro Gómez-Alanís, José A. González-López,, Ángel M. Gómez, Eros Roselló, Manuel Chica-Villar, José C. Sánchez-Valera, José L. Pérez-Córdoba, Victoria Sánchez

Dpt. Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada, Spain

{amp,agomezalanis,joseangl,amgg,erosrosello,manuelc,svjosecarlos,jlpc,victoria}@ugr.es

Abstract

Currently, voice biometrics systems are attracting a growing interest driven by the need for new authentication modalities. The BioVoz project focuses on the reliability of these systems, threatened by various types of attacks, from a simple playback of prerecorded speech to more sophisticated variants such as impersonation based on voice conversion or synthesis. One problem in detecting spoofed speech is the lack of suitable models based on classical signal processing techniques. Therefore, the current trend is based on the use of deep neural networks, either for direct attack detection, or for obtaining deep feature vectors to represent the audio signals. However, these solutions raise many questions that are still unanswered and are the subject of the research proposed here. These include what spectral or temporal information should be used to feed the network, how to compensate for the effect of acoustic noise, what network architecture is appropriate, or what methodology should be used for training in order to provide the network with discriminative generalization capabilities. The present project focuses on the search for solutions to the aforementioned problems without forgetting a fundamental issue, little studied so far, such as the integration of fraud detection in the whole biometrics system.
Index Terms: antispoofing, voice biometrics, speaker verification, robustness, secure biometrics, embeddings, deep neural networks.

1. Introduction

In a digital society in which interaction with information systems is increasingly frequent, either to access them or to carry out all kinds of formalities, procedures and transactions, there is a clear need to incorporate authentication techniques that increase the security of this interaction while keeping naturalness and ease of use. Of the various forms of authentication, those based on the use of biometric physical (fingerprint, iris, face, voice, etc.) or behavioral (gait, speech, etc.) features. In this field, voice is presented as a form of biometric characterization that is particularly suitable and natural for transactions also carried out by voice, while complementing other modalities (fingerprints, iris, etc.).

The underlying technology of any voice biometrics system is automatic speaker verification (ASV). In recent years, interest in ASV systems has increased substantially due to their commercial applications [1]. However, these systems are susceptible to malicious attacks that could breach their security, as an impostor can fool the system using any of these four types of attacks [2]: (i) interpretation attack (imitating the voice of a legitimate user), (ii) replay attack (using recorded voice of a legitimate user), (iii) speech synthesis (by text-to-speech conversion), and (iv) voice conversion (transformation of the impostor's voice to resemble the genuine target voice). Commonly studied attacks are replay, conversion and synthesis attacks. Likewise, attacks can be physical, when they are captured by microphone, or logical, when they are injected directly into the system, bypassing the microphone. While replay attacks are commonly executed physically, conversion and synthesis attacks are usually applied logically. As with other types of biometric techniques, the scientific community has recently begun to develop countermeasures, or anti-spoofing techniques, to detect these attacks on ASV systems [3]. The importance of developing technologies that address this problem is reflected in a number of challenges associated with the Interspeech'2015/17/19/21 congresses [4] [5][6][7][8].

All this recent interest in the security of voice biometrics systems has grown in parallel to the spectacular development of machine learning algorithms in general and deep neural networks in particular. Thus, the field of signal and voice processing is undergoing an unprecedented transformation to which speaker verification and antispoofing techniques are no strangers. For this reason, given the difficulty in finding models based on classical signal processing that discriminate between genuine signals and attacks, the use of neural networks as highly discriminative detectors and/or feature extractors, with the capacity to generalize to attacks not seen during training, has recently been proposed [9][10][11]. The present proposal also adopts this type of approach, conceptually represented in Figure 1 for the case of a standalone system (independent of the ASV system). It must be taken into account that this is a conceptual diagram, so different blocks may be combined in practical implementations.

The BioVoz project is focused on the following aspects of anti-spoofing (also referred to as presentation attack detection, PAD, in the ISO/IEC 30107 standard):

1. Improvement of the architecture of the standalone anti-spoofing system for feature extraction (spectral and/or embeddings), consolidating the work initiated by the research team in the last years.

2. Training methodology of the standalone anti-spoofing system: development of cost functions and data augmentation techniques that allow the network to generalize and detect new attacks not seen in training.

3. Integration of ASV and anti-fraud systems into a single system in which both parts collaborate.

While the first objective aims to consolidate the work done by the research team in recent years, the other two represent very little explored and innovative lines of work, which we consider crucial for the translation of these technologies to secure applications.

10.21437/IberSPEECH.2022-53

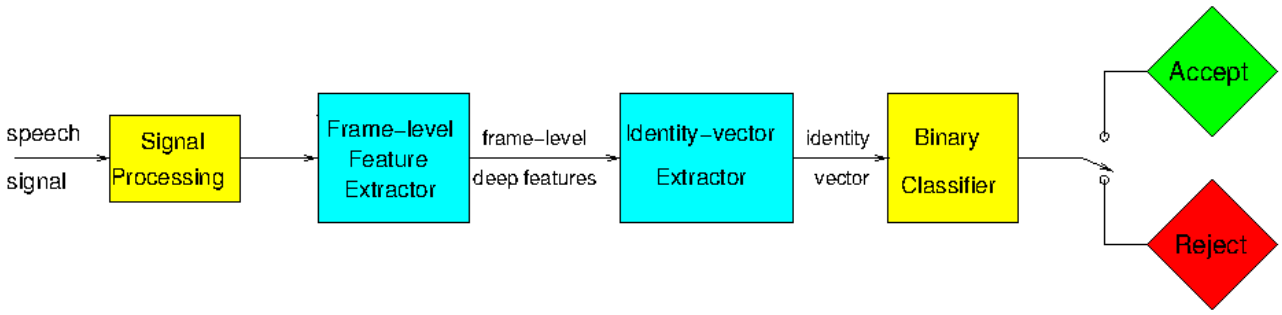

Figure 1: *Conceptual diagram of a standalone antispoofing system.*

2. State of the art in anti-spoofing

The following subsections review the state of the art on the different aspects related to this proposal.

2.1. Feature extraction for anti-spoofing

2.1.1. Spectral features

The first issue to be considered is the extraction of voice features (typically spectral) to feed the system and serve as the basis for the extraction of deep features representing each of the analysis frames. In addition to classical features based on filter banks (FBANK), and Mel-scale cepstral coefficients (MFCC), other types of specific parameters have been proposed for spoofing attack detection: constant-Q cepstral coefficients (CQCC) [12], choclear filter cepstral coefficients and instantaneous frequency features (CFCC-IF) [13] and long-term spectral statistics (LTSS) [14]. These three types of features have proven to be more effective when used with classical classifiers such as LDA (linear discriminant analysis), GMM (Gaussian mixtures models) or SVM (support vector machines). Despite the variety of existing works in search of suitable spectral features, there have been few studies specifically using signal phase as in the case of CFCC-IF mentioned above or in [15], where it is shown that relative-phase shift (RPS), modified group delay (MGD) or cosine phase (CosPhase) parameters provide meaningful performance improvements.

2.1.2. End-to-end systems

As an alternative to the use of separate magnitude and phase characteristics, other formats can be considered. Thus, a particularly interesting way to avoid this separation is to work in end-to-end mode, i.e. directly on the audio samples as in [16], where the use of a CNN (convolutional neural network) for signal processing is proposed, or in [17], where a new type of CNN called SincNet, which implements an optimizable filter bank using gradient descent techniques, is proposed. End-to-end models have undergone a spectacular development in recent years in speech applications, where the so-called Wavenet model [18] for speech synthesis has gained considerable popularity. Another possibility is derived from the use of spectral information as a complex sequence to be processed by the network. Although networks capable of handling complex data [19] are not widely used at the moment, their potential usefulness in signal processing systems is evident [43].

2.1.3. Deep feature extraction

One methodology to ensure that the anti-spoofing system is able to generalize and detect attacks not even seen during training is based on the extraction of deep features or embeddings [20][21], i.e., discriminative features extracted by deep neural networks, understood as those that are extracted from one of the intermediate layers that make up the neural network. This type of technique is known in the literature as distance metric learning [22][23], and finds its application in those problems where it is necessary to define some kind of distance between pairs of signals, so that the representations obtained for signals of the same class are closer (in the space of the defined metric) than those of different classes.

For the extraction of frame-level features (see Fig. 1), deep neural networks with multiple fully connected layers (deep neural networks, DNN) [10][21], as well as convolutional networks [21][24] have been applied. Once the deep raster features have been obtained, it is necessary to combine all of them to obtain a single identity vector. For this purpose, several methodologies have been proposed such as obtaining the average [10] or a weighted average using an attention model [25], or the application of recurrent neural networks (RNN) that allow the learning of long temporal dependencies [24][26][27]. Specifically, this last technique has been proposed by this research team and it is the one that presents the best performance in detecting synthesis spoofing and voice conversion attacks on ASVspoof-2015 data. Also, there are networks that directly perform the process of extracting the spoofing identity vectors from the spectral information. This is the case of [21], where a single RNN sweeps the incoming sequence, of [11], which combines a CNN with an RNN to provide the identity vector, or [14], where the identity vector is obtained from signal spectral statistics.

A research line of interest for the present project is based on the application of attention models to the extraction of deep raster features. Attention models have been initially proposed to increase the accuracy of classification systems in which an alignment between the input sequence to the network and the target sequence is required since both sequences have different lengths. At the moment it is a topic little explored in anti-spoofing systems [28] or ASV [25].

2.1.4. Feature extraction for noisy environments

A problem inherent to all automatic systems working with speech signals is that of noisy environments that reduce their performance. Anti-spoofing systems are no exception [15][29]. Despite this, there are so far few works that address it. Thus, in [29] a noisy version of the ASVspoof 2015 challenge database

was developed with which in [10] it is shown that neural networks can also be very effective in extracting suitable discriminative features even in noisy conditions. This research team has also proposed a technique based on the use of noise masks that provide useful information to the neural network about the noise present at each point in the time-frequency space [24][27].

An interesting unsupervised alternative to combat the effect of noise is based on the use of bottleneck features. Although this type of features are very little studied in the field of anti-spoofing, they have been tested in robust ASV under different modalities such as multitask training [30], generative adversarial networks (GAN) [31] or variational autoencoders [32].

3. Training methods for anti-spoofing

One of the most relevant aspects when trying to train neural networks is the selection of a training methodology and a cost function appropriate to the problem at hand. A typical cost function in classification problems is the cross-entropy. In the case of anti-spoofing with deep features the most common is to use cross entropy to perform a multiclass training considering the genuine speech class and the classes corresponding to each of the spoofing attacks present in the training set [21].

However, in both ASV and anti-spoofing, other training techniques to obtain more discriminative features, using cost functions that incorporate distances or similarity measures, have been proposed. Thus, for example, Siamese networks use a contrastive cost function (acting on two networks that share parameters) that attempts to maximize the distance between features of different classes at each iteration [33]. Networks based on a triplet loss criterion [34] take this idea further by simultaneously (in the same iteration) trying to maximize the distance between features of different classes and minimize the distance between features of the same class. This project research team has already work on these methods, proposing an effective generalization of triplet-loss based on the use of kernels [35]. Another alternative training methodology to the previous ones is that of multi-task training, whereby the network tries to learn according to different training criteria in order to exploit the synergies between them [30].

On the other hand, training techniques that perform data augmentation from adversarial examples have also emerged in recent years as one of the main lines of research [36]. Adversarial learning techniques allow finding model weaknesses by generating artificial adversarial examples that closely resemble real examples. Moreover, these adversarial examples can be used to retrain the system (by performing data augmentation), which thus learns to defend itself against such adversarial attacks. Because an anti-spoofing system must be very robust to both seen and unseen attacks, it is imperative to defend it against different types of adversarial attacks [37][44].

4. ASV-PAD integration

PAD systems must be integrated with the ASV system to form the complete voice-based biometric system. However, this integration is not sufficiently explored. There are two types of integration: (1) integration at the score level; and (2) integration at the deep feature level (embeddings).

Integration at the score level combines the scores obtained independently by the ASV and anti-spoofing systems. In turn, at this level of integration there are two possible approaches [Sahidullah16] represented in Figure 2: (a) cascade integration, and (b) score fusion. In cascaded integration, the logical access

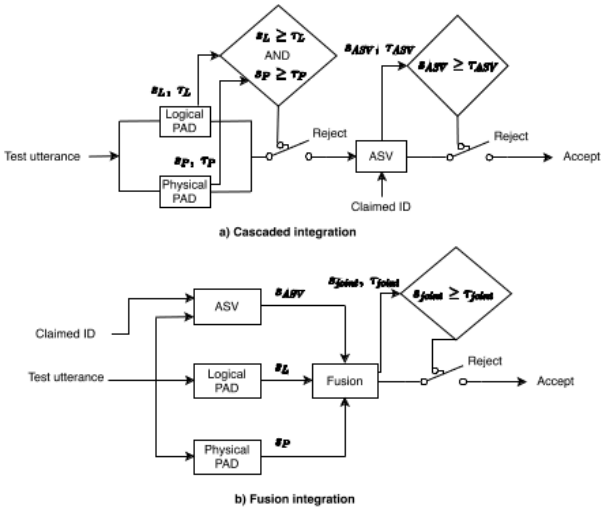

Figure 2: *Block diagrams of two different possibilities for score-level ASV/Antispoofing integration: (a) cascade integration, (b) fusion integration.*

(LA) and physical access (PA) anti-spoofing systems usually precede the ASV system to decide whether the speech signal is spoofed. If no voice spoofing is detected, the voice signal is processed by the ASV system to decide whether the identity claimed by the speaker is true. On the other hand, in score fusion, the scores of the different systems are combined by logistic regression to obtain a single integration score (Sjoint), which decides whether the identity claimed by the speaker is true and there is no impersonation. This type of integration has also been recently explored on a two-dimensional score space using Gaussian mixture models [40].

For integration at the deep feature (or embedding) level, the embeddings extracted by ASV and anti-spoofing systems is processed jointly, in order to take advantage of the fact that they share the genuine speech subspace. There are only two papers that have explored this type of embedding [41][42]. In [41], 1) a two-stage probabilistic linear discriminant analysis (PLDA) classifier is proposed, where in the first stage a PLDA classifier is trained using only embeddings extracted from genuine speech, while in the second one a new means vector is estimated, 2) a new subspace of supplanted speech is added to the same PLDA, and 3) only embeddings extracted from supplanted speech are used for training. More recently, in [42] a multi-tasking neural network (ASV and anti-spoofing tasks) is optimized using an adaptation of the triplet loss to detect the three types of classes existing in embedding: (1) genuine voice of the speaker claiming the correct identity, (2) genuine voice of an intruder claiming another identity, and (3) spoofed speech.

5. Project objectives

The main objectives of this project are summarized in the following points:

1. Development of feature extraction methods for standalone anti-spoofing systems. Several directions are considered here:

   (a) Determination of a suitable signal representation for speech anti-spoofing, with special attention to the incorporation of the signal phase since, al-

though commonly discarded in speech processing, it is considered very relevant in the detection of attacks on ASV systems.

(b) Development of an ASV+PAD-integrated identity embedding extraction network, trained as a whole with a suitable criterion. This may provide a simple and powerful feature extractor which also allows incorporating available information about the acoustic environment to provide robustness against noise. The new network architecture should incorporate the analysis capability of convolutional networks for feature extraction and the sequential processing capability of recurrent ones. Possibilities to be reckoned with are the following: incorporating attention models that help to identify those parts of the audio sequence that contain the most relevant information about the attack or, as an alternative to supervised features, the application of bottleneck features based on GANs or variational autoencoders.

(c) Study of alternative data or network formats. Although the previous points focus on the use of spectral information in modulus/phase format, given the peculiarity of phase information, other possibilities will be considered alternatively, such as feature extraction networks being able to process the spectra as complex sequences, or to extract features directly from the signal samples (end-to-end systems).

2. Development of training methods suitable for anti-spoofing. Although the extraction of deep features is usually approached by solving a classification problem with neural networks (which implies the use of cross-entropy or similar classical cost functions), the fact is that the criterion with which they are obtained should take into account that the ultimate goal of the extraction is not only to achieve a high discriminative capacity but also to generalize against attacks not foreseen during training. Thus, we consider relevant the study of new cost functions, such as the contrastive function used in Siamese networks or triplet loss functions, which, incorporating error measures based on distances, simultaneously provide discrimination and generalization capabilities. We also consider the adaptation of other novel training methodologies not explored in anti-spoofing such as the combination of several cost criteria under multi-task structures.

3. Integration of the countermeasure system with the associated ASV system. Although voice anti-spoofing technologies are closely linked to ASV, the integration of both systems has been little studied. We think that it is important not only to look for the optimal working point (from the individual scores provided by both systems) for a given application, as is the case in any verification system, but that both subsystems can integrate and "collaborate" to obtain a more accurate final decision. Ultimately, this final decision is unique (it consists of accepting or rejecting the claimed identity), so it seems logical to develop systems that address the biometric problem jointly.

6. Experimental framework

The experimental framework to be adopted will be that of the ASVspoof challenges that have taken place to date (2015/17/19/21). Each challenge has defined a database with subsets for training, validation and test, a baseline system that serves to obtain reference results (only ASVspoof 2017), and performance measures.

7. Expected results & Information

The expected results of this project can be summarized in the following two points:

1. Improvement of the security of the new modalities of authentication through voice biometrics. To this end, we hope to achieve highly reliable methods for detecting phishing attacks. Since these systems must work in conjunction with ASV, another expected outcome of the project is to obtain a more robust and efficient integration methodology for ASV and anti-spoofing.

2. Advances in the state of the art associated to deep learning and, in particular, to deep neural networks as essential facilitating techniques for extraction and processing of complex signals such as speech. We hope that this project will help to improve our understanding of different related aspects such as what information we should process and how we should present it, what network architecture is suitable to process a certain type of data, and how we should train the network to extract a highly discriminative and, at the same time, generalizing representation.

More details and information about ongoing activities and obtained results can be found at *http://sigmat.ugr.es/proyectos/biovoz/*.

8. Acknowledgement

This work was supported by the FEDER/Junta de Andaluca-Consejera de Transformacin Econmica, Industria, Conocimiento y Universidades Project PY20_00902.

9. References

[1] Brett Beranek, "Does voice answer the call for a better customer experience?", *Biometric Technology Today*, Vol. 2018, Issue 5, May 2018, Pages 7-9.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, y H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, no. 4, pp. 788-798, 2015.

[3] Z. Wu et al., "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, 2016.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", *Proc. Interspeech 2015*.

[5] Tomi Kinnunen, Md Sahidullah, Hector Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection", *Proc. Interspeech 2017*.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. Aik Lee: "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", *Proc. of Interspeech 2019*.

[7] ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge. *http://www.asvspoof.org/.*

[8] Yamagishi et al: "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection", *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 16 September, 2021.

[9] N. Chen, Y. Qian, H. Dinkel, B. Chen and K. Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge," *Proc. Interspeech, 2015*.

[10] Y. Qian, N. Chen, H. Dinkel, and Z. Wu, "Deep Feature Engineering for Noise Robust Spoofing Detection,' *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(10), pp. 1942-55, 2017.

[11] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 684–694, 2017.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," *Proc. Odyssey*, pp. 249252, 2016.

[13] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech,' *Pr. Interspeech15*, pp. 2062-66.

[14] H. Muckenhirn, P. Korshunov, M. Magimai-Doss and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 11, pp. 20982111, 2017.

[15] C. Hanilci, T. Kinnunen, M. Sahidullah, A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83-97, 2016.

[16] H. Muckenhirn, M. Magimai-Doss, S. Marcel: "End-to-End Convolutional Neural Network-based Voice Presentation Attack Detection". *IEEE International Joint Conference on Biometrics (IJCB)*, 2017.

[17] M. Ravanelli, Y. Bengio, "Speaker recognition from raw waveform with Sincnet," *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021-1028, 2018.

[18] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio." *ArXivabs/1609.03499*, 2016.

[19] H. G. Zimmermann, A. Minin, and V. Kusherbaeva, Comparison of the complex valued and real valued neural networks trained with gradient descent and random search algorithms, *European Symposium on Artificial Neural Networks*, 2011.

[20] S. Yadav, A. Rai: Learning discriminative features for speaker identification and verification. *Proc. of Interspeech 2018*.

[21] Y. Qian, N. Chen, and K. Yu, "Deep Features for automatic spoofing detection,' *Speech Communication*, vol. 85, pp. 43–52, 2016.

[22] B. Kulis, Metric learning: A survey, *Foundations and trends in machine learning*, 5(4), 287-364, 2012.

[23] E. Hoffer, N. Ailon. Deep metric learning using triplet network." *International Workshop on Similarity-Based Pattern Recognition*, Springer, Cham, (pp. 84-92), 2015.

[24] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, A. Gomez, "A Deep Identity Representation for Noise Robust Spoofing Detection, *Proc. Interspeech*, 2018.

[25] K. Okabe, T. Koshinaka and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding, *Proc. Interspeech*, 2018.

[26] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez and A. M. Gomez: "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection". *Proc. of Interspeech'2019*, September 15-19, Graz, Austria.

[27] Alejandro Gomez-Alanis, Antonio M. Peinado, Jose A. Gonzalez, Angel M. Gmez: "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection". *IEEE Trans. On Audio Speech and Language Processing*, vol. 27, no. 12, pp. 1985-1999, Dec. 2019.

[28] J. Li et al, "Attention-Based LSTM Algorithm for Audio Replay Detection in Noisy Environments." *Applied Sciences* 9 (2019): 1539.

[29] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant condition," *Proc. Interspeech*, 2016.

[30] Y. Qian et al: Noise and metadata sensitive bottleneck features for improving speaker recognition with non-native speech input. *Proc. Interspeech 2016*.

[31] Hong Yu1, Zheng-Hua Tan, Zhanyu Ma, Jun Guo: Adversarial Network Bottleneck Features for Noise Robust Speaker Verification. *Proc. of Interspeech 2017*.

[32] J. Chien and C. Hsu: "Variational manifold learning for speaker recognition". *Proc. ICASSP*, 2017.

[33] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric", *Proc. Interspeech*, 2018.

[34] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, and I. Kremnev, "Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition, *Proc. Interspeech*, 2018.

[35] Alejandro Gomez-Alanis, Jose A. Gonzalez, Antonio M. Peinado: "A Kernel Density Estimation Based Loss Function and Its Application to ASV-Spoofing Detection", *IEEE Access*, June 2020.

[36] X. Yuan, P. He, Q. Zhu and X. Li, Adversarial Examples: Attacks and Defenses for Deep Learning.*IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805-2824, 2019.

[37] S. Liu, H. Wu, H. Lee and H. Meng, Adversarial Attacks on Spoofing Countermeasures of Automatic Speaker Verification.*Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[38] H. Wu, S. Liu, H. Lee and H. Meng, Defense against adversarial attacks on spoofing countermeasures of ASV,*Proc. ICASSP*, 2020.

[39] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, Z.-H. Tan, Integrated spoofing countermeasures and automatic speaker verification: an evaluation on ASVspoof 2015, *Proc. of Interspeech 2016*, San Francisco, USA, Sept. 2016.

[40] M. Todisco, H. Delgado, K. Aik Lee, M. Sahidullah, N. Evans, Tomi Kinnunen, J. Yamagishi: Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. *Proc. of Interspeech 2018*.

[41] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu and S. Marcel, Joint speaker verification and anti-spoofing in the i-vector space, *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821-32, 2015.

[42] J. Li, M. Sun, X. Zhang and Y. Wang, Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss, *IEEE Access*, vol. 8, pp. 7907-7915, 2020.

[43] K. Osako, R. Singh, B. Raj: Complex recurrent neural networks for denoising speech signals. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[44] H. Wu, S. Liu, H. Lee and H. Meng, Defense against adversarial attacks on spoofing countermeasures of ASV,*Proc. ICASSP*, 2020.