Uludağ University Journal of The Faculty of Engineering, Vol. 29, No. 1, 2024 RESEARCH

DOI: 10.17482/uumfd.1311113

INCREASING ROBUSTNESS OF I-VECTORS VIA MASKING: A CASE STUDY IN SYNTHETIC SPEECH DETECTION

Barış AYDIN [*]
Gökay DİŞKEN [**]

Received: 07.06.2023; revised: 22.02.2024; accepted: 15.03.2024

Abstract: Ensuring security in speaker recognition systems is crucial. In the past years, it has been demonstrated that spoofing attacks can fool these systems. In order to deal with this issue, spoof speech detection systems have been developed. While these systems have served with a good performance, their effectiveness tends to degrade under noise. Traditional speech enhancement methods are not efficient for improving performance, they even make it worse. In this research paper, performance of the noise mask obtained via a convolutional neural network structure for reducing the noise effects was investigated. The mask is used to suppress noisy regions of spectrograms in order to extract robust i-vectors. The proposed system is tested on the ASVspoof 2015 database with three different noise types and accomplished superior performance compared to the traditional systems. However, there is a loss of performance in noise types that are not encountered during training phase.

Keywords: deep learning, convolutional neural network, spoof detection, speaker recognition, robust features

I-vectorlerin Maskeleme Yoluyla Dayanıklılığının Arttırılması: Sentetik Konuşma Tespitinde Bir Vaka Çalışması

Öz: Konuşmacı tanıma sistemleri için güvenlik hayati önem taşımaktadır. Geçtiğimiz yıllarda, sahte konuşma saldırılarının bu sistemleri kandırabildiği ortaya konmuştur. Bu durumu önlemek amacı ile sahte konuşma tespit sistemleri geliştirilmiştir. Bu tür sistemler bazı durumlarda oldukça yüksek performans sergilese de, gürültü altında performansları kötüleşmektedir. Geleneksel konuşma iyileştirme yöntemleri performansı artırmak bir yana, daha da kötüleştirmektedir. Bu çalışmada, konvolüsyonel sinir ağı yapısı kullanılarak elde edilen maskenin gürültü etkisini azaltmaktaki performansı incelenmiştir. Maske, spektrogramın gürültülü bölgelerini bastırmakta ve bu spektrogramdan elde edilen i-vectorleri gürbüz hale getirmekte kullanılmıştır. ASVspoof 2015 veri tabanı ve üç farklı gürültü tipi ile gerçekleştirilen testlerde önerilen sistemin geleneksel sistemlerden daha üstün olduğu gösterilmiştir. Ancak eğitim aşamasında karşılaşılmayan gürültü tiplerinde performans kaybı olmaktadır.

Anahtar Kelimeler: derin öğrenme, evrişimli sinir ağları, sahte konuşma tanıma, konuşmacı tanıma, gürbüz öznitelikler

[*] Adana Alparslan Türkeş Science And Technology University, Faculty of Engineering, Department of Electrical and Electronic Engineering, 01250, Sarıçam/ADANA
[**] Adana Alparslan Türkeş Science And Technology University, Faculty of Computer and Informatics, Department of Artificial Intelligence Engineering, 01250, Sarıçam/ADANA
Corresponding Author: Barış Aydın (barisaydin@atu.edu.tr)
This is the expanded version of the paper titled "Robust Synthetic Speech Detection using Masked Spectrogram-Based i-vectors," which was presented orally at the ELECO'2022 Symposium and invited for evaluation in the Uludağ University Journal of The Faculty of Engineering by the symposium organizing committee.

1. INTRODUCTION

Speaker recognition refers to identifying individuals based on their voices by utilizing the physical differences in vocal production organs. In addition to these physical differences, each speaker has a unique speaking style, including a certain accent, rhythm, intonation style, pronunciation pattern, word choice, etc. (Kinnunen et al., 2010) Because of the uniqueness of the voice to the individual, speaker recognition systems are used in various fields, including telephone banking (HSBC, 2017), e-commerce (Find Biometrics, 2018), and forensic science (Find Biometrics, 2018). With increasing usage areas, it is crucial to prevent potential attacks that could be carried out by malicious people on these systems.

Possible attack types on a speaker recognition system include synthesizing the speaker's voice, using various software to transform the attacker's voice into the target person's voice, imitating the target speaker's voice, and using a pre-recorded voice of the target speaker (Find Biometrics, 2018), (Find Biometrics, 2018), (Hanilçi et al., 2016), (Gomez-Alanis et al., 2019)).

In recent years, various organizations such as ASVspoof have increased awareness and research in the field of spoofed speech detection ((Evans et al., 2013), (Alegre et al., 2013), (Sizov et al., 2015), (Evans et al., 2013), (Wu et al., 2014), (Wu et al., 2015), (Dutoit et al., 2007)). In particular, systems that utilize deep learning algorithms can achieve highly successful results (Wang et al., 2021), (Jung et al., 2022). On the other hand, additive noise, which is one of the biggest problems in speech-related systems, reduces the success rate in spoofed speech detection (Hanilçi et al., 2016). There are limited studies on robust fake speech detection ((Hanilçi et al., 2016), (Gomez-Alanis et al., 2019), (Gomez-Alanis et al., 2018)). Specifically, low performance of traditional speech enhancement methods (Wiener filter, spectral subtraction, etc.) makes the problem even more challenging (Hanilçi et al., 2016). However, much more successful results can be achieved with complex deep learning systems (Gomez-Alanis et al., 2019). These types of systems use different methods for feature extraction and classification than those used in traditional speaker recognition systems. Therefore, these systems are focused solely on the problem of noise and spoofed speech detection.

EER (Equal Error Rate) is a typical statistic for measuring the performance of spoof speech detection systems. It is defined as the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR). This criterion shows the degree to which systems are able to discriminate between synthetic and real speech.

Studies have indicated that the emergence of diverse speech synthesis (SS) and voice conversion (VC) methodologies has rendered speech-based biometric systems exceedingly susceptible to spoofing assaults. According to (Diyopsi et al., 2017), this circumstance may result in a rise in false acceptance rates, making countermeasures against spoofing attacks necessary. EER is important in assessing the effectiveness of these measures since a low EER value shows that the systems can successfully distinguish between spoof and genuine speech.

(Hassan et al., 2021) suggests combining spectral features like MFCC, GTCC, Spectral Flux, and Spectral Centroid to create a synthetic speech detector. In order to train a biLSTM to categorise the speech, the fused feature set attempts to capture differences between real and synthetic signals.

Using the ASVspoof 2019 LA dataset, the system demonstrated efficacy in identifying voice conversion and synthetic speech attacks.
Novel speech features for improved detection of spoofing attacks (Dipjyoti et al., 2015) presents new speech features for spoofing attack detection that are based on alternate frequency-warping methods. When tested against the ASVspoof 2015 corpora, the features—which were computed using formant-specific block transformation—perform better than previous methods in differentiating between natural and synthetic speech, achieving 0% equal error rates on a variety of spoofing attack tasks.

(Nugroho et al., 2022) applies a Deep Neural Network (DNN) approach, achieving significant performance with a model accuracy rate of 96.5%, precision of 97.3%, recall of 96.5%, and an F1 Measure of 96.7%. The study underscores DNN's robustness in fake speech detection, processing extensive data.

A one-class learning anti-spoofing system is introduced by (Zhang et al., 2020) to identify unknown synthetic voice spoofing assaults. Using an angular margin to distinguish spoofing assaults in the embedding space and compacting bona fide speech representation, the system achieves an EER of 2.19% on the ASVspoof 2019 Challenge, outperforming previous approaches.

The effectiveness of high dimensional magnitude and phase-based features for detecting spoofed speech is examined in (Xiao et al.,2015)'s study. Advances in text-to-speech (TTS) and voice conversion (VC) technologies pose a serious threat to automatic speaker verification (ASV) systems. Through the use of two magnitude-based and five phase-based characteristics in combination with multilayer perceptron analysis, the research was able to detect known spoofing assaults in the ASVspoof 2015 challenge with a low equal error rate (EER) of 0.29%. With an EER of 5.23%, the detection performance for unknown spoofing kinds was less successful, underscoring the need for additional study to increase the method's generalizability to novel and undiscovered spoofing approaches.

The ASSERT system is reviewed in the publication "LARIHS ASSERT Reassessment for Logical Access ASVspoof 2021 Challenge" by (Benhafid et al. 2021), with an emphasis on improving the detection of logical access spoofing attacks. Thinner SENet backbones with new activation functions and the use of advanced features and loss algorithms are among the improvements. The success of these changes in spoofing detection was demonstrated by the reevaluated system's 60% improvement in min-tDCF for the ASVspoof 2019 evaluation, which marked a considerable improvement over the original.

In (Dişken, 2023), a novel differential convolutional neural network generates finer noise masks based on directional changes of activations. These masks, combined with linear filterbank magnitudes, are inputted into various spoof detection systems, including PLDA with x-vectors, Emphasized Channel Attention, ECAPA-TDNN, and LCNN with LSTM layers. Experiments on the ASVspoof 2015 dataset show that the LCNN-LSTM network with noise masks achieves superior performance, with an average Equal Error Rate (EER) of 2.67% for known noise types and 3.10% for unknown noise types. Clean ASVspoof 2015 data has an EER of 0.83%, while ASVspoof 2019 data under logical access conditions has a 2.6% EER.

The main purpose of the study is that while traditional methods use two separate systems with different features for spoof detection and speaker verification, the study can perform both tasks with a single system via i-vector. (Dehak et al., 2011). Thus, both speaker recognition and spoofed

speech detection using the same i-vectors can be possible. The mask obtained using a convolutional neural network (CNN) is utilized to reduce the effect of noise. This mask is applied on the noisy spectrogram, then, conventional i-vector extraction steps are followed. To test the proposed system, ASVspoof 2015 database and three different noise types (babble, white, car) are used based on previous studies in the literature((Hanilçi et al., 2016), (Gomez-Alanis et al.,2018), (Gomez-Alanis et al., 2019)). The results showed that the masking process increased the robustness of i-vector features and, unlike traditional methods, an improvement in performance is observed. There are few studies on this subject in the literature. Successful results can be obtained with deep learning models, by using more than one system. The goal of the study is to ensure the use of a single system instead of two different systems.

2. ROBUSTNESS VIA MASKING

The goal of masking is to improve the robustness against noise by distinguishing less reliable regions of the speech spectrum (more corrupted by noise) and more reliable regions (less affected by noise). Previous studies have shown that applying classic SNR-based masks for spoofed speech detection yields the best results in noisy scenarios (Gomez-Alanis et al., 2019). In the proposed system, the CNN structure used in (Gomez-Alanis et al., 2019) is preferred due to its high performance. The mask creation network is shown in Figure 1. Noisy spectrograms of 31 frames in length, consisting of 15 frames to the right and 15 frames to the left of the central frame are used as inputs to the CNN structure. Therefore, the output of the system (the last linear layer) indicates the signal-to-noise ratio (SNR) for the relevant frame. The sigmoid function is applied to these values to obtain mask values in the range of 0-1. Here, 0 represents completely noise, and 1 represents completely speech information.

The average noise shown in Figure 1 is calculated by averaging the first 10 frames of the corresponding noisy speech data. Typically, it is assumed that the first and last few frames of the speech signal contain only noise information. Therefore, the trained CNN structure has an explicit noise reference, instead of relying only on the spectrogram. During the training phase, the instantaneous SNR target presented to the CNN for each frame is calculated as follows:

$$SNR(\mathrm{t,f}) = 20.\log_{10}\frac{X(\mathrm{t,f})}{N(\mathrm{t,f})} \tag{1}$$

The notation (t, f) represents time-frequency partitions. X and N are the spectrograms of the clean speech and noise, respectively (generated using short-time Fourier transform (STFT)). The values of the target masks to be used in the training phase are obtained using an adjustable sigmoid function given in Equation 2.

$$m_k = \frac{1}{1-e^{-\alpha(SNR(t,f_k)-\beta)}} \tag{2}$$

Here, α controls the slope of the sigmoid, and β corresponds to the threshold commonly used to define Ideal Binary Masks (IBMs) (Wang et al., 2009). Combining these two equations, the Equation 3 is obtained, which calculates the target mask values as $\gamma = 20.\ \alpha/\log(10)$).

$$m_k = \frac{X^{\gamma}}{X^{\gamma} + e^{\alpha\beta}.N^{\gamma}} \tag{3}$$

Figure 1, an example spectrogram at the input of the network and the corresponding mask at the output can be seen. The generated mask is multiplied with the noisy spectrogram to obtain a spectrogram with reduced noise. After this stage, a simple speech activity detection system (Kinnunen et al., 2010) was used to discard frames containing silence or noise.

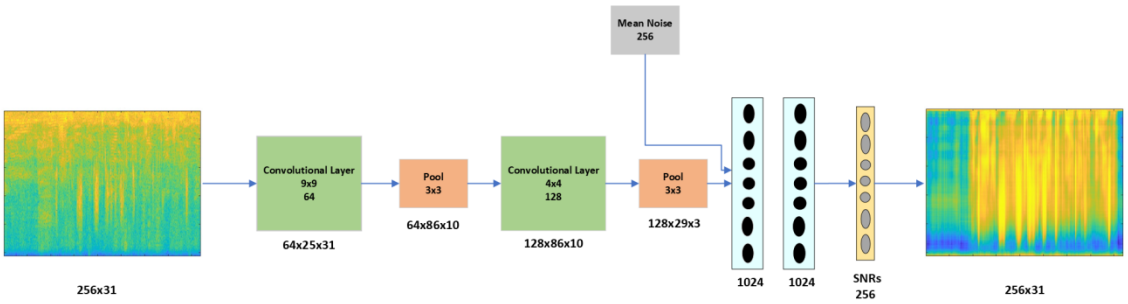

Figure 1:
CNN architecture

3. I-VECTORS

I-vectors are low and fixed dimensional representations of variable length audio data. This feature allows applications of various normalization techniques in a low-dimensional space (Varga et al., 1993). Following the traditional i-vector extraction steps given in (Sizov et al., 2015), the training of the universal background model (UBM) and the total variability matrix T is performed. A speaker and channel-independent Gaussian mixture model (GMM) supervector can be defined as follows:

$$M=m+T\omega \tag{4}$$

Here, m is the mean supervector taken from the UBM and ω is a randomly generated vector with a normal distribution. The i-vector is obtained by maximizing the posterior of ω for each audio file. Once the i-vector is extracted, various compensation and dimensionality reduction techniques such as within-class covariance normalization (WCCN), linear discriminant analysis (LDA), and length normalization can be applied ((Dehak et al., 2011), (Delgado et al. 2018)). The features obtained from audio data are used for GMM and UBM training via applying Mel-frequency cepstral coefficients (MFCC) or CQCC. MFCC extraction steps typically involve filtering the magnitude spectrogram obtained with STFT with triangular filters placed linearly on the Mel scale, taking the logarithm, and applying discrete cosine transform.

CQCC is based on constant-q transform which gives a variable resolution, providing greater frequency resolution for lower frequencies and enhanced temporal resolution for higher frequencies. CQCC usually performs better than MFCC for spoofed speech detection.

In the proposed study, robust i-vectors are created by using vector extraction of masked spectrograms with reduced noise effect. Apart from the masking process, all steps (MFCC extraction, UBM and T training) follow traditional methods. The obtained vectors are scored using classifiers such as cosine distance and probabilistic LDA (PLDA) to calculate performance.

4. EXPERIMENTAL SETUP

4.1 ASVSPOOF 2015

The proposed system was tested using the ASVspoof 2015 dataset (Wu et al., 2015). This dataset consists of non-overlapping training, development, and test subsets. The spoof speech attacks are included in this dataset. While there are 10 different attack algorithms, only five of them (S1-S5) are available in the training set. The remaining five (S6-S10) are only available in the test set.

Table 1: ASVSPOOF 2015 data distribution

ASVSPOOF 2015	Training Data	Test Data
Real Speech	3750	9404
Spoofed Speech	12625	184000

The attacks are listed as (Wu et al., 2015);

S1: A simplified voice conversion algorithm where the converted speech is generated by selecting frames from the target speech.

S2: A basic voice conversion method that adjusts only the first mel-cepstral coefficient to align the source spectrum slope with the target.

S3: A speech synthesis algorithm using a hidden Markov model-based system (HTS) with speaker adaptation techniques, requiring only 20 adaptation utterances.

S4: Similar to S3, but utilizing 40 adaptation utterances for potentially improved performance or adaptation quality.

S5: A voice conversion technique implemented using both the voice conversion toolkit and the Festvox system for transforming speech.

S6: A voice conversion approach based on joint density Gaussian mixture models and maximum likelihood parameter generation, focusing on maintaining global variance for naturalness.

S7: Similar to S6 but uses line spectrum pairs (LSP) instead of mel-cepstral coefficients for representing the spectrum, offering a different approach to spectrum conversion.

S8: A tensor-based voice conversion method that constructs a speaker space using a Japanese dataset, offering a novel approach to handling speaker characteristics.

S9: A voice conversion algorithm that employs kernel-based partial least squares (KPLS) for implementing a non-linear transformation function, simplifying the process by excluding dynamic information.

S10: A speech synthesis algorithm executed with the MARY Text-To-Speech system (MaryTTS), an open-source platform for generating speech.

For noise, 'car', 'white', and 'babble' noises from the Noisex-92 database (Varga et al., 1993) and 'cafe' noise from the QUT-Noise database (Dean et al., 2015) were used, based on similar studies in the literature ((Hanilçi et al., 2016), (Gomez-Alanis et al., 2019)). White, babble, and cafe noises were used for mask training. Car noise was only used in the tests to analyze the system's performance under noise types that it had not previously encountered.

4.2 CNN PARAMETERS

The CNN structure used to obtain the mask was trained with a learning rate of 3e-4 and binary cross entropy was chosen as the learning criterion. Only the noisy versions of real speech audios (3750 speech samples) were used as the training data. Each data was corrupted with random noise type selected from white, babble, and cafe noises at a random SNR values between 0 dB and 20 dB. Thus, multi-condition training was performed to prevent the system from focusing on a single noise type and level.

The CNN architecture depicted processes a noisy spectrogram through multiple layers to enhance speech. Initially, the input spectrogram—encapsulating a series of frames—is fed into the first convolutional layer, which extracts basic features like edges and patterns indicative of noise or speech characteristics. Subsequently, a pooling layer reduces the feature map's dimensionality, emphasizing the most salient features and making the network less sensitive to the exact positioning of features within the frames. A second convolutional layer then detects more complex features, combining the simpler patterns identified earlier. This is again followed by a pooling layer, which further condenses the data, preparing it for the final classification steps.

The last part of the network consists of fully connected layers culminating in a linear layer that computes the SNR values for the central frame. These SNR values undergo normalization through a sigmoid function, resulting in a binary mask that distinguishes between noise (0) and speech (1). An average noise reference, derived from the first noise-dominated frames, informs the network what noise looks like. This reference improves the network's ability to differentiate between noise and speech. The output mask from the CNN is then used to clean up the noisy input, and a speech activity detection system removes any remaining silent or noise-heavy frames, ensuring a clear speech output. Overall structure is shown in Figure 3.

4.3 I-VECTOR PARAMETERS

The first step in i-vector extraction is obtaining MFCC features. For this, the audio signal is divided into frames of 25 ms length with a frame step of 10 ms. The windowed frames are transformed with a 512-point FFT. The filter bank consists of 32 triangular filters. After discrete cosine transformation, 32 coefficients are used. In addition, delta and delta-delta features are added to obtain 96-dimensional features per frame.

The CQT is applied with a maximum frequency of Fmax = 8kHz. The number of octaves is 9. The number of bins per octave B is set to 96. Re-sampling is applied with a sampling period of 16 bins in the first octave. Resulting feature vectors are of dimension 19, excluding the C_0 coefficient (C_f 29 coefficients + C_0 for the original system).

The UBM consists of 512 Gaussian components and is trained only on real speakers in the training data. The 600-dimensional T matrix is trained using the entire training data (Hanilçi, 2018). After

obtaining i-vectors, whitening, WCCN, and length normalization are applied.The process is shown in Figure 2.

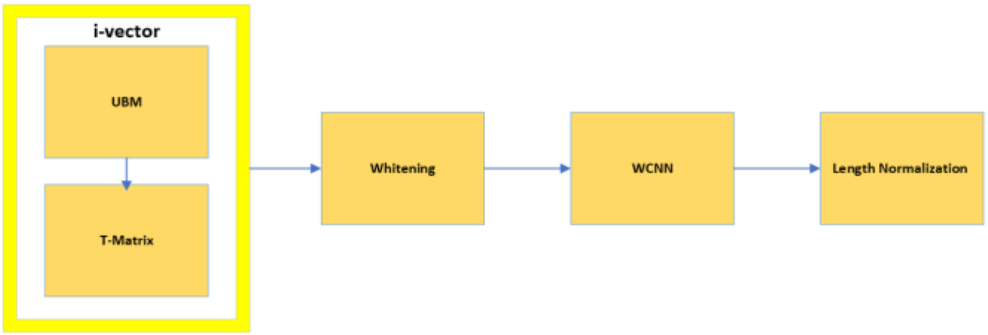

Figure 2:
Process after feature (MFCC/CQCC) extraction

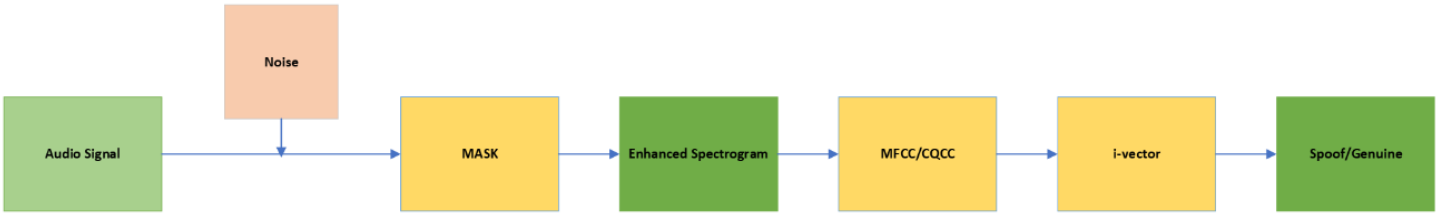

Figure 3:
Proposed System

5. RESULTS

Table 2 shows the performance of the proposed methods on the development data. The EER value represents the average EER values for five different attacks in the development set. For comparison, the results of MFCC-based i-vector without mask (Hanilçi et al., 2016) are also included in the table. As can be seen, i-vectors enhanced with masks provide significantly better performance than classical i-vectors.

As seen in the Table 2, the highest relative improvement in MFCC based i-vector system, with a rate of over 50%, was achieved with the cosine distance classifier at 20 dB level for the babble noise type for encountered noise type. The lowest improvement in MFCC based i-vector system was observed with the PLDA classifier at 20 dB level for car noise type with a 22% improvement rate. The results indicate that masking contributes to robustness compared to the same system without mask. Similar observations can be made for CQCC based system.

Table 2: EERs (%) for the development data

Noise Type	dB	Proposed System with mask				MFCC without mask (Hanilçi et al., 2016)	MFCC without mask (Hanilçi et al., 2016)
		MFCC based i-vector	MFCC based i-vector	CQCC based i-vector	CQCC based i-vector		
		COS	PLDA	COS	PLDA	COS	PLDA
White	0 dB	31.05	32.17	40.27	41.56	43.47	43.67
	10dB	25.32	26.60	32.49	33.47	36.35	37.87
	20dB	16.44	18.06	23.57	23.12	26.48	25.32
Babble	0 dB	30.80	31.09	39.67	37.76	45.71	45.82
	10dB	19.44	20.25	28.62	29.66	33.59	33.13
	20dB	10.34	11.49	18.78	19.62	20.94	20.65
Car	0dB	27.54	29.42	32.45	33.82	39.62	38.01
	10dB	24.16	25.38	23.72	24.66	33.67	32.50
	20dB	17.08	18.64	15.54	16.47	24	24.35

Table 3 and Table 4 show the results for the test data by using MFCC based i-vector system. The noticeable issue here is the low performance for the car noise type. Generally, this type of noise is considered to be the least disruptive ((Hanilçi et al., 2016), (Gomez-Alanis et al., 2019)). However, the masking performance is affected by this noise type because it was not used in the training stage of the mask. As evidence, the examples given on the logarithmic scale in Figure 4. The noisy spectrograms on the top belong to signals corrupted with babble 0 dB and car 20 dB.

Table 3. EERs (%) for known attacks of evaluation data (Proposed System/MFCC based i-vector with mask**)**

Noise Type	dB	S1		S2		S3		S4		S5	
		COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA
White	0	25.85	26.07	36,68	36.72	20.46	20.76	20.15	20.50	28.46	28.21
	10	16.96	18.46	25,96	27.13	6.80	7.38	6.88	7.48	18.05	18.72
	20	5.84	7.14	11.04	12.34	1.48	1.76	1.50	1.83	9.35	10.00
Babble	0	28.97	29.57	38.38	38.51	31.32	31.52	31.47	31.76	31.12	31.55
	10	18.54	20.63	29.96	31.26	16.05	16.97	16.49	17.53	19.59	20.99
	20	10.83	13.55	18.18	20.59	3.55	4.34	3.88	4.71	10.87	12.73
Car	0	28.97	35,95	38.94	42.54	18.48	22.26	19.35	23.24	22.59	26.04
	10	27.58	34.11	39.61	43.87	18.34	21.72	19.30	22.90	19.44	23.58
	20	21.85	26.33	35.40	38.92	9.54	11.48	9.85	11.84	17.44	21.05

Table 5 shows the performance of the proposed systems and the GMM method for the test data from (Hanilçi et al., 2016) for comparison. (Hanilçi et al., 2016) did not analyze the performance of i-vectors since GMM outperformed it. The average EER values are calculated for known attacks (S1-S5), and unknown attacks (S6-S9). Attack type S10 is difficult to detect and has reduced the system performance; therefore, it is not included in the average calculation as in (Hanilçi et al., 2016)

Table 4. EERs (%) for unknown attacks of evaluation data (Proposed System/MFCC based i-vector with mask**)**

Noise Type	dB	S6		S7		S8		S9		S10	
White	0	33.57	33.70	32.71	33.03	16.76	16.68	29.90	30.10	41.12	41.35
	10	22.55	23.29	23.91	25.28	6.60	6.91	21.92	22.98	35.16	35.40
	20	13.28	14.21	10.05	11.56	2.95	3.39	10.27	12.22	34.29	34.29
Babble	0	32.96	33.08	33.58	34.30	25.86	26	33.33	33.67	40,54	40.45
	10	21.69	22.90	23.24	25.39	16.65	17.53	24.44	26.35	39.84	40.06
	20	14.08	15.79	12.54	15.32	7.61	8.38	14.15	17.14	40.45	40.87
Car	0	26.63	29.95	29.84	34.95	29.42	33.28	40.82	43.21	45.21	47.91
	10	24.73	29.14	28.59	33.58	28.72	32	38.97	42.01	49.82	49.96
	20	22.83	26.83	23.67	27.63	17.53	19.01	29.05	33.33	50	50

The results indicate that the proposed approach is more effective for lower dBs. Also, the proposed system delivered the worst results for the unseen noise type (car). This result emphasizes the importance of creating balanced training data. More noise types are necessary to increase the generalization capacity of the network, as noise type may not be known a priori for a practical spoof detection system.

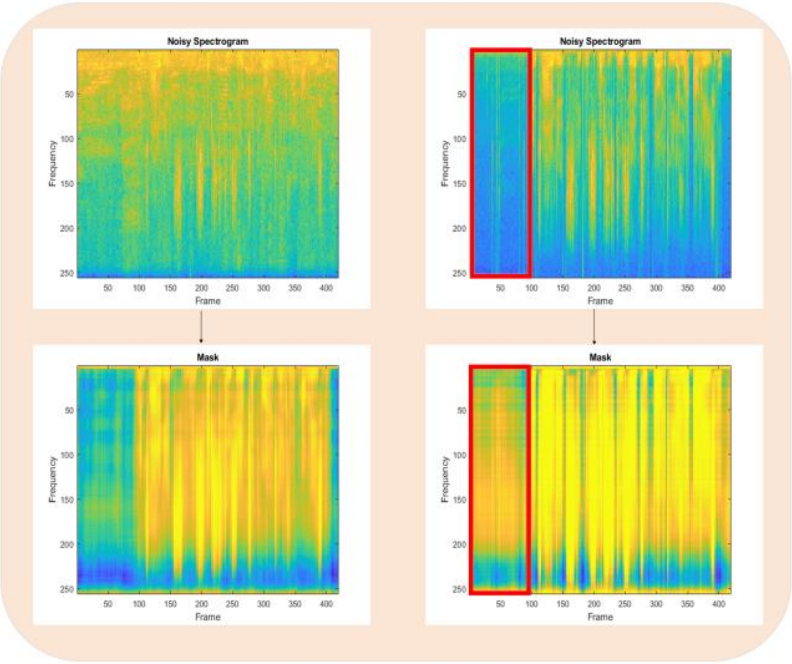

Figure 4:
Spectrogram of speech data corrupted with babble 0 dB (upper-left) and car 20 dB (upper-right), and their masks below

Table 5. Comparison of average EERs (%) for evaluation (K: Known attacks (S1-S5), U: Unknown attacks (S6-S9)

Noise Type	dB	MFCC based i-vector		CQCC based i-vector		MFCC (Hanilçi et al., 2016)		RPS (Hanilçi et al., 2016)		SCMC (Hanilçi et al., 2016)		MGD (Hanilçi et al., 2016)	
		K	U	K	U	K	U	K	U	K	U	K	U
		Cos	PLDA	Cos	PLDA								
White	0	26.32	26.45	28.23	28.37	35.07	39.66	44.56	46.64	43.73	42.27	44.42	45.88
	10	14.93	15.83	18.74	19.61	25.45	29.78	42.16	44.98	33.36	32.14	37.42	38.66
	20	5.84	6.61	9.13	10.34	16.43	17.94	38.53	40.62	19.92	15.40	27.25	36.24
Babble	0	32.25	32.58	31.43	31.76	33.54	28.40	40.81	40.66	29.74	25.13	37.59	40.77
	10	20.12	21.47	21.50	23.04	15.59	12.76	21.17	23.71	8.32	5.30	26.30	35.65
	20	9.46	11.18	12.09	14.15	7.48	6.49	6.09	10.62	2.15	1.39	14.20	23.55
Car	0	25.66	30.00	31.67	35.34	17.33	14.69	24.66	25.67	8.59	7.36	30.32	36.63
	10	24.85	29.23	30.25	34.18	7.31	6.03	5.28	9.93	2.16	1.67	15.99	24.44
	20	18.81	21.92	23.27	26.70	3.57	2.83	0.74	3.67	0.79	0.52	9.39	16.12

6. CONCLUSION

This study proposed noise mask based robust i-vector extraction and examined its performance for noisy spoofed speech detection tasks. The results showed that the mask structure is successful in reducing noise effects. This situation reveals that mask structures can be useful in an area where traditional speech enhancement methods have performance-decreasing effects (Hanilçi et al., 2016). The CNN-based mask, on the other hand, failed against a noise type that was not encountered in the training phase. This situation provides a clue about the necessity of increasing the diversity of noise in the database to prevent memorization.

The i-vector method was chosen due to its high performance for speaker verification. With the proposed method, the same i-vectors can be used for both speaker verification and spoof detection, in a robust manner. However, compared to the state-of-the-art systems in detecting synthetic speech under noise ((Gomez-Alanis et al., 2019), (Wang et al., 2021)), the proposed system was found to be far behind. A reason for this is the low performance of i-vectors in short audio recordings (Hanilçi, 2018), where the average data length in ASVspoof 2015 dataset is 3.5 seconds. Another reason is the other systems' complexities. For example, the study in (Gomez-Alanis et al., 2019) designed a system which consists of two different feature types, deep learning models for each feature, and an external classifier in addition to the CNN-based noise mask. Therefore, even though the masking approach performs better than the traditional methods, advanced architectures are necessary for achieving impressive results.

ACKNOWLEDGEMENT

This work was supported by TÜBİTAK (Project No: 121E057).

CONFLICT OF INTEREST

The authors confirm that there are no known conflicts of interest or any financial associations with any institution, organization, or individual that could have influenced the present work.

AUTHOR CONTRIBUTIONS

The contributions of the authors are as follows: both authors collaborated in preparing the data and coding. Barış Aydın was responsible for obtaining the test results and writing the article. Gökay Dişken analyzed and interpreted the results.

REFERENCES

1. Alegre, F., Amehraye, A. and Evans, N. (2013) A one-class classification approach to generalized speaker verification spoofing countermeasures using local binary patterns, *PInt. Conf. on Biometrics: Theory*, Applications and Systems (BTAS), IEEE, Washington DC, USA. doi: 10.1109/BTAS.2013.6712706

2. ASVspoof, (2014). ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. Available: https://www.asvspoof.org/asvSpoof.pdf Accessed: Dec 19, 2014

3. Benhafid, Z., Selouani, S. A., Yakoub, M. S., Amrouche, A. (2021) LARIHS ASSERT reassessment for logical access ASVspoof 2021 challenge. *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, Online*, 94-99. doi: 10.21437/ASVSPOOF.2021-15

4. Dean, D., Kanagasundaram, A., Ghaemmaghami, H., Hafizur, M., Sridharan, S. (2015) The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition, *Interspeech 2015*, International Speech and Communication Association, Dresden. doi: 10.21437/Interspeech.2015-685

5. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P. (2011) Front-End Factor Analysis for Speaker Verification, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798. doi: 10.1109/TASL.2010.2064307

6. Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K. A., Yamagishi, J. (2018) ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements, *Odyssey 2018 - The Speaker and Language Recognition Workshop*, ASVSpoof, Odyssey, 296-303. doi: 10.21437/Odyssey.2018-42

7. Dipjyoti, P., Monisankha, P., Goutam, S., (2015) Novel speech features for improved detection of spoofing attacks, *2015 Annual IEEE India Conference (INDICON)*, New Delhi, India, pp. 1-6, doi: 10.1109/INDICON.2015.7443805.

8. Dipjyoti, P., Monisankha, P., Goutam, S., (2017) Spectral features for synthetic speech detection. *IEEE journal of selected topics in signal processing*, 11.4: 605-617. doi: 10.1109/JSTSP.2017.2684705

9. Dişken, G. (2023) Differential convolutional network for noise mask estimation. Applied Acoustics, 211, 109568. doi: 10.1016/j.apacoust.2023.109568

10. Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., Stylianou, Y. (2007) Towards a voice conversion system based on frame selection, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),*. IEEE, Honolulu, USA,4, 513-516. doi: 10.1109/ICASSP.2007.366962

11. Evans, N., Yamagishi, J., and Kinnunen, T. (2013) Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols, and metrics, *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*..

12. Evans, N., Kinnunen, T., and Yamagishi, J. (2013) Spoofing and countermeasures for automatic speaker verification, *Interspeech 2013*, ISCA, Lyon, France, 925-929. doi: 10.21437/Interspeech.2013-288.

13. Find Biometrics (2018). Voicevault Biometrics to Protect Payments. Available: https://findbiometrics.com/voicevault-biometrics-toprotect-payments-25131/ (Accessed: Jun. 13, 2018)

14. Find Biometrics (2018). Morpho and Agnitio Partner, Bring Voice Biometrics to Criminal ID. Available: https:// findbiometrics.com/morpho-and-agnitio-partner-bring-voice-biometricsto-criminal-id-21261/ (Accessed: Jun. 13, 2018)

15. Gomez-Alanis, A., Peinado, A. M., Gonzalez, J. A., and Gomez, A. M. (2018) A Deep Identity Representation for Noise Robust Spoofing Detection, *Interspeech 2018*, International Speech and Communication Association, Haydarabad, 676-680. doi: 10.21437/Interspeech.2018-1909

16. Gomez-Alanis, A., Peinado, A. M., Gonzalez, J. A., and Gomez, A. M. (2019) A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection, *IEEE/ACM Transactions on Audio, Speech, And Language Processing*, 27(12), 1985-1999. Doi: 10.1109/TASLP.2019.2937413

17. Hanilçi, C. (2018) Data selection for i-vector based automatic speaker verification anti-spoofing, *Digital Signal Processing*, 72, 171-180. doi: 10.1016/j.dsp.2017.10.010 (Article)

18. Hanilçi, C., Kinnunen, T., Sahidullaha, M., Sizova, A. (2016) Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise, *Speech Communication*, 85, 83-97. doi: 10.1016/j.specom.2016.10.002

19. Hassan, F., Javed, A. (2021) "Voice Spoofing Countermeasure for Synthetic Speech Detection," *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2021, pp. 209-212, doi: 10.1109/ICAI52203.2021.9445238.

20. HSBC (2017). HSBC Voice ID Making Telephone Banking Safer Than Ever. Available: https://www.hsbc.co.uk/1/2/voice-id (Accessed: Dec. 29, 2017)

21. Jung, J., Heo, H., Tak, H., Shim, H., Chung, J. S., Lee, B. J., Yu, H. J., & Evans, N. (2022) AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks, *ICASSP 2022*, IEEE, Lyon, France. doi: 10.21437/Interspeech.2013-288

22. Nugroho, K., Winarno, E., (2022) Spoofing Detection of Fake Speech Using Deep Neural Network Algorithm, *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, pp. 56-60. doi: 10.1109/iSemantic55962.2022.9920401.

23. Sizov, A., Khoury, E., Kinnunen, T ., Wu, Z. and Marcel, S. (2015) Joint speaker verification and anti-spoofing in the i-vector space, IEEE Transactions on Information Forensics and Security, *IEEE Transactions on Information Forensics and Security*, 10(4), 821-832. doi: 10.1109/TIFS.2015.2407362

24. Xiao, X., Tian, X., Du, S., Xu, H., Chng, E., Li, H. (2015). Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge. In Interspeech (pp. 2052-2056). doi:10.21437/Interspeech.2015-465

25. Varga, A., Steeneken, H. J. M. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech

recognition systems, *Speech Communication*, 12(43), 247-251. doi: 10.1016/0167-6393(93)90095-3

26. Wang, D. L., Kjems, U., Perdersen, M. S., Boldt, J. B., Lunner, T. (2009) Speech intelligibility in background noise with ideal binary time-frequency masking, J. *Acoustical Soc. America*, 125(4), 2336–2347. doi: 10.1121/1.3083233

27. Wang, X., Yamagishi, J. (2021) A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection., *Interspeech 2021*, ISCA, Brno, Czech Republic, 4259-4263. doi: 10.21437/Interspeech.2021-702

28. Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T., King, S. (2015) SAS: A speaker verification spoofing database containing diverse attacks, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, South Brisbane, Queensland, Australia, 9(5), 4440-4444. doi: 10.1109/ICASSP.2015.7178810.

29. Wu, Z., Kinnunen, T., Evans, N., & Yamagishi, J. (2015). Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database, University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/298.

30. Zhang, C., Yu, C., Hansen, J. H. L. (2017) An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing, *IEEE Journal of Selected Topics in Signal Processing*, 11, 684-694, 2017.

31. Zhang, Y., Jiang, F., Duan, Z. (2020) One-Class Learning Towards Synthetic Voice Spoofing Detection, in *IEEE Signal Processing Letters*, vol. 28, pp. 937-941, doi: 10.1109/LSP.2021.3076358.