Celal Bayar University Journal of Science

Noise-Robust Spoofed Speech Detection Using Discriminative Autoencoder

Gökay Dişken[1]*, Zekeriya Tüfekçi[1]

[1] Adana Alparslan Türkeş Science and Technology University, Department of Electrical-Electronics Engineering, Adana, Türkiye
[2] Çukurova University, Department of Computer Engineering, Adana, Türkiye
* gdisken@atu.edu.tr
* Orcid: 0000-0002-8680-0636

Received: 20 June 2022
Accepted: 8 June 2023
DOI: 10.18466/cbayarfbe.1132319

Abstract

Audio spoof detection gained the attention of the researchers recently, as it is vital to detect spoofed speech for automatic speaker recognition systems. Publicly available datasets also accelerated the studies in this area. Many different features and classifiers have been proposed to overcome the spoofed speech detection problem, and some of them achieved considerably high performances. However, under additive noise, the spoof detection performance drops rapidly. On the other hand, the number of studies about robust spoofed speech detection is very limited. The problem becomes more interesting as the conventional speech enhancement methods reportedly performed worse than no enhancement. In this work, i-vectors are used for spoof detection, and discriminative denoising autoencoder (DAE) network is used to obtain enhanced (clean) i-vectors from their noisy counterparts. Once the enhanced i-vectors are obtained, they can be treated as normal i-vectors and can be scored/classified without any modifications in the classifier part. Data from ASVspoof 2015 challenge is used with five different additive noise types, following a similar configuration of previous studies. The DAE is trained in a multicondition manner, using both clean and corrupted i-vectors. Three different noise types at various signal-to-noise ratios are used to create corrupted i-vectors, and two different noise types are used only in the test stage to simulate unknown noise conditions. Experimental results showed that the proposed DAE approach is more effective than the conventional speech enhancement methods.

Keywords: Deep learning, denoising autoencoder, i-vector, spoofing detection

1. Introduction

Vulnerability of automatic speaker detection systems against spoofed speech is an important drawback for practical usage of these systems. Early studies showed that conventional features such as mel-frequency cepstral coefficients (MFCC) are not suitable for detection of synthetic speech signals [1]. Organizations such as the ASVspoof challenges provided a common database to the researches, which accelerates the awareness and number of studies on the spoof detection problem [2–4].

Many different solutions to the aforementioned problem have been proposed. One of the most efficient features against synthetic speech is the constant-Q cepstral coefficients (CQCC) [5], which served as a baseline method with Gaussian Mixture Model (GMM) classifier. Phase-based features, which are usually neglected in speaker/speech recognition, were found to be beneficial for spoof detection [6], [7]. For modeling/classifying the extracted features traditional methods such as GMMs and i-vectors [3] could be preferred. However, recent studies include deep learning architectures such multilayer perceptron (MLP) [8], convolutional neural networks (CNN) [9], recurrent neural networks (RNN) [10].

Compared to the baseline systems such as MFCC−GMM, most of the mentioned studies achieved superior performances for detecting synthetic and/or replayed speech data. On the other hand, under additive noise, the detection performance drops rapidly as shown in [11] using the ASVspoof 2015 database. Several different feature types were investigated with GMM and i-vector backends, and to reduce the noise effects, conventional speech enhancement methods such as spectral subtraction and Wiener filtering were employed. However, speech enhancement methods further deteriorate the detection performance. Similar observations were made in [12] for additive noise and reverberation, using MLP classifier, and score level fusion of different features led to an improved performance. Complex deep learning architectures yielded to the most robust systems so far, considering the noisy ASVspoof 2015 database. In [13], a CNN system was examined with different combinations of noise aware training, masking, and RNN. The most robust system was observed with CNN+Mask+RNN combination. In [14], gated recurrent convolutional networks were used with modified group delay features and spectrogram obtained via short-time Fourier Transform. A noise mask was also estimated with a CNN mask estimator, where the purpose is to mask the noise dominant time-frequency bins and use the speech dominant bins in further processes. Identity vectors were observed after stacking the network outputs, and a probabilistic linear discriminant analysis was used for classification.

Despite the various methods proposed so far for spoof detection, studies on robustness against the additive noise are very limited. Conventional speech enhancement methods are ineffective, and the proposed robust deep learning-based systems are computationally demanding. In this work, discriminative denoising autoencoder (DDAE) [15] with i-vector inputs are investigated to verify if it can provide a robust system, or fails as speech enhancement methods did. One of the advantages of the i-vectors is that variable length utterances can be expressed as fixed low dimensional vectors [16]. This property is also important for neural network systems as they require fixed length inputs. For instance, to use the CNN architecture, speech files are truncated or padded to a selected length [9]. Using the i-vectors, the information loss or adding redundant information can be avoided. The DAE networks mainly used for restoring the noisy/corrupted features, and can be found in several speech related studies such as [17], [18]. The original (clean) version of the inputs is given as the DAE targets for training the system. In the discriminative DAE, class labels are also provided so that the network can also learn the differences between classes. Following the previous robustness studies mentioned before, the ASVspoof 2015 database is used in this work with five different noises from NOISEX-92 [19] and QUT-Noise [20] databases. The experimental results proved that the proposed approach is more robust than the conventional speech enhancement methods.

2. Denoising Autoencoders with I-vectors

2.1. I-vector

Text-independent speaker verification enjoyed the advantages of the i-vectors in the last decade. Some of the advantages that i-vector introduced are fixed dimensional representation, modeling both speaker and channel variabilities in the same space, channel compensation in the i-vector space, opportunity to use support vector machine and linear discriminant analysis based classifiers. The conventional i-vector extraction scheme consists of training a GMM with a high number of mixtures (named as universal background model (UBM)) and a low rank matrix called total variability matrix. Let M be a speaker and channel independent GMM supervector and expressed as,

$$M = m + T\omega \qquad \textbf{(2.1)}$$

where m is the mean supervector from the UBM, and ω is a random vector with normal distribution, and T is the total variability matrix (also known as i-vector extractor). For each utterance, maximum a posterior estimate of ω is the i-vector.

Besides their high performances in speaker recognition, i-vectors were also used in spoofed speech detection [21], [22], however their performance is poor under additive noise [11]. Contrary, robust speaker recognition studies can be found in the literature [15], [18], [23]–[27]. The main disadvantages of these methods are the increased computational demand, and poor performance for short duration utterances [27]. The average utterance length for the ASVspoof 2015 data is 3.5 seconds, which is very short even for the clean data. Although the results are not given in this work, preliminary experiments showed that using the MAP denoising methods of [25], [27] do not introduce any robustness to spoof detection system, and the main reason for this result may be the limited amount of training data and their short durations. Therefore, DDAE is used in this work to exploit the nonlinear relation between the noisy and clean i-vectors, and to avoid extreme computational load (as the conventional i-vector extraction framework is followed, and once the DDAE is trained, evaluation is fast as in the typical neural networks). The idea of using i-vectors as the inputs of a neural network was also applied in different speech related areas [17], [28], [29]. Hence, it may be worth exploring a similar approach for synthetic speech detection.

For the classification of the i-vectors, several choices are possible but cosine distance scoring is used in this work due to their higher performances for short durations [16]. Equation 2 shows the cosine distance scoring where the i-vectors of two classes as human (ω_{hum}) and spoof (ω_{spo}) are compared against a test i-vector (ω_{test}), and Equation 3 shows the cosine distance formula.

$$score = \cos(\omega_{hum}, \omega_{test}) - \cos(\omega_{spo}, \omega_{test}) \qquad \textbf{(2.2)}$$

$$cos(\omega_a, \omega_b) = \frac{\langle \omega_a, \omega_b \rangle}{\|\omega_a\| \|\omega_b\|} \qquad \textbf{(2.3)}$$

2.2. Discriminative denoising autoencoder

The DDAE used in this work follows the similar architecture of [15]. Different from the DAEs, DDAE network is trained to denoise and classify the inputs at the same time. Therefore, class-specific information will be added to the enhanced features. Another aspect of the DDAE is that there is no assumption on the noise's distribution in the i-vector space [15].

The DDAE consists of two MLPs. The first MLP represents the DAE part, and has two hidden layers with 500 nodes each. Although the details will be given in the experimental setup section, the i-vectors used in this work are 100-dimensional, hence the input of the DDAE. Therefore, the output layer of the first MLP has 100 nodes, as they represent the enhanced i-vectors. Then, those nodes act as the inputs of the second MLP, which also has hidden layer with 500 nodes, and an output layer with two nodes, representing the human and spoof classes. ReLU activation is used after each layer except the last output layer. Similarly, dropout with a 0.5 probability is applied to each layer except the last output.

The first MLP (DAE part) has the mean square error as the cost function as given below,

$$minMSE = \frac{1}{S}\sum_{i=1}^{S}||\omega_i - \widehat{\omega_i}||^2 \qquad \textbf{(2.4)}$$

where $\widehat{\omega}$ is the enhanced i-vector output, an ω d is the target i-vector. S is the total number of training data. It should be noted that the enhanced outputs are taken from the output of the first MLP, before applying dropout or ReLU for the next layer. The cost function of the second MLP is the cross-entropy error between the predicted class and the target class as given in Equation 5,

$$CE = \frac{1}{S}\sum_{i=1}^{S}\sum_{k=1}^{K} l_i^k \log\left(o_i^k\right) \qquad \textbf{(2.5)}$$

where $o_i^k$is the predicted probability and l_i^k is the ground truth probability of the ith training sample being a member of kth class, respectively. Combining these two cost functions with a suitable weight (i.e. $0 \le \alpha \le 1$), a multi-task training can be achieved by minimizing the total cost as shown in Equation 6. The complete architecture is illustrated in Figure 1.

$$\text{Total cost} = \min(1-\alpha)MSE + \alpha CE \qquad \textbf{(2.6)}$$

3. Experiments

3.1. Database description

To examine the robustness of the proposed system, ASVspoof 2015 database is chosen, which consists of synthetic speech attacks. The number of utterances for each partition of the database is given in Table 1. For the spoofed speech, 10 different attacks are available. Five of those attacks (S1 – S5) are included in every subset, the other five (S6 – S10) are only presented in the evaluation set, hence called unknown attacks. S3, S4, and S10 attacks are based on speech synthesis methods. S3 and S4 share the same underlying algorithm but generated with different amounts of data. S10 is a unit selection-based algorithm and considered the hardest to detect for this database. The other attacks are voice conversion attacks, using different algorithms to enhance diversity. More detailed explanation for each attack type can be found in [4]. The detection performance of unknown attacks (evaluation set) can be related to the generalization capacity of the systems.

For the additive noises, car, babble, and white noises are chosen from NOISEX-92 database, and street and café noises are selected from QUT-NOISE database, following the configurations of the previous robust spoofed speech detection studies [12], [30]. For all of the data, sampling rate is 16 kHz. Also, café and car noise are not included in the training to simulate unseen noise conditions.

Table 1. Partitions of ASVspoof 2015 database

Subset	Number of utterances	
	Human	Spoof
Train	3750	12625
Development	3497	49875
Evaluation	9404	184000

To train the DDAE in a multi-condition manner, a random noise signal (white, babble, or street) is added to the utterances at random SNR levels in the range of 0 dB to 20 dB, with 5 dB steps. While adding the noise, a random starting point is chosen, and a segment equal to the clean signal is extracted. Hence, each clean signal can be considered to be distorted with a unique (to some degree) noise signal. As seen in Table 1, the number of human utterances is less than the spoofed utterances. This may result in an imbalance while training the DDAE. A similar problem is solved by clustering the spoofed i-vectors in [21], where equal error rate (EER) was decreased to 10% from 30.71%. In this work, instead of decreasing the spoof data, human data is corrupted three times while the spoof data is corrupted only one time, following the mentioned process. In the end, a total of 11250 clean-noisy i-vector pairs are created for the human data, and 12625 clean-noisy i-vector pairs are created for the spoof data.

The EER is used as the performance metric, as it is a widely used metric for speaker verification and spoof detection tasks. It defines the operating point where the false acceptance (verifying a spoofed speech) rate and the false rejection (rejecting a genuine speech) rate are equal to each other.

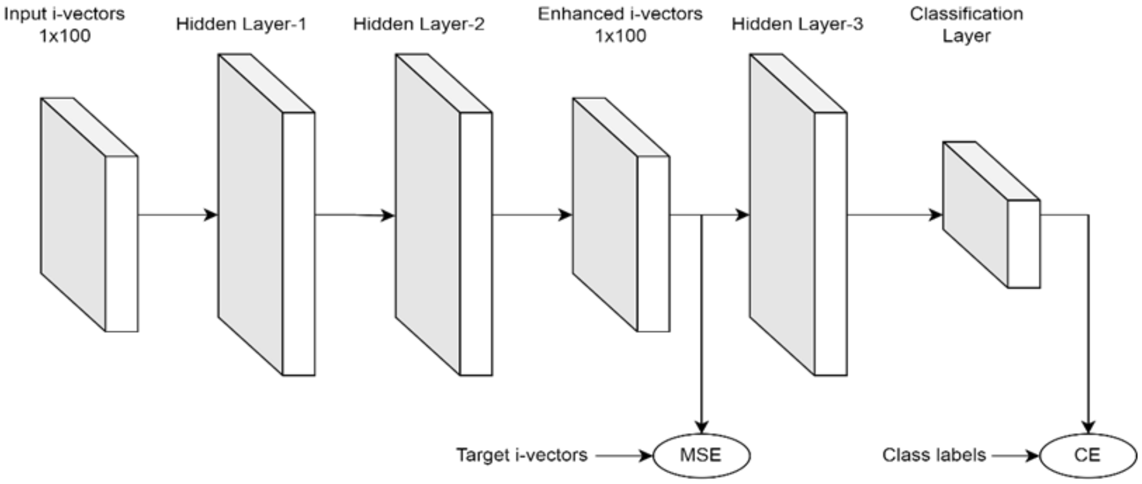

Figure.1: DDAE architecture for denoising the corrupted i-vectors.

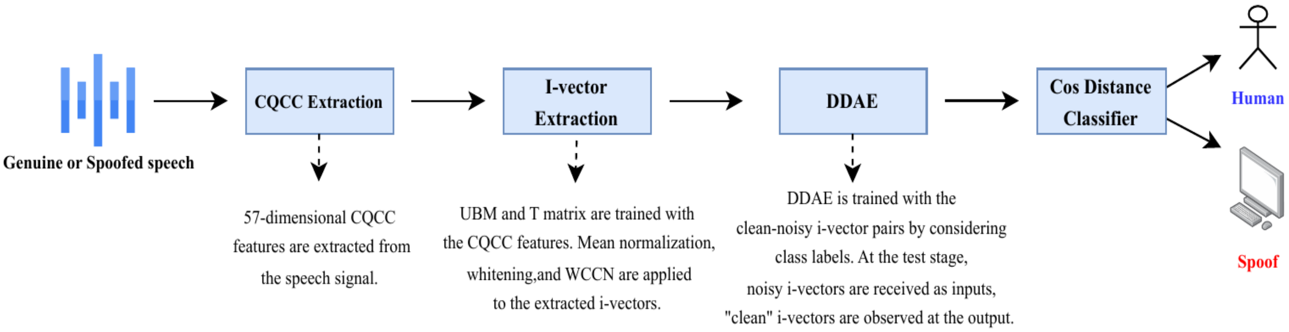
Genuine or Spoofed speech
CQCC Extraction
I-vector Extraction
DDAE
Cos Distance Classifier
Human
Spoof
57-dimensional CQCC features are extracted from the speech signal.
UBM and T matrix are trained with the CQCC features. Mean normalization, whitening,and WCCN are applied to the extracted i-vectors.
DDAE is trained with the clean-noisy i-vector pairs by considering class labels. At the test stage, noisy i-vectors are received as inputs, "clean" i-vectors are observed at the output.

Figure.2: Block diagram of the proposed system.

3.2. CQCC and i-vector extraction

For the feature extraction, the recipe of [3] is followed. 19 -dimensional CQCC features are extracted from utterances, then delta and acceleration coefficients are appended to the static coefficients. A UBM with 64 mixtures is trained using the training partition of ASVspoof 2015 data. Total variability matrix with 100 factors is trained with the same data. After the extraction of the i-vectors, mean normalization, whitening, and within class covariance normalization are applied. To represent human and spoof classes, single i-vectors are obtained by averaging the respective clean i-vectors of each class. Length normalization is also included, which also simplifies the cosine distance scoring to the numerator part of Equation 3.

CQCC extraction is done using the baseline codes provided by the ASVspoof challenge organizers. All i-vector related processes are implemented in MATLAB via MSR Identity Toolbox [31].

3.3. DDAE training

Once the clean-noisy i-vector pairs are obtained, the DDAE can be trained using these pairs and class labels. It is also found that including clean-clean i-vector pairs in the training improved the performance on the clean test data. This may be due to the fact that if the network did not see any clean data while training, it assumes that every input is noisy, hence reduced performance for clean data. Learning rate is chosen as 1e-4, and Adam optimizer is used. The network is trained for 500 epochs with an early stop option if the training loss does not decrease for 10 epochs. The network related operations (creating the network, training, extracting enhanced i-vectors, testing) are realized with PyTorch. Also, while training the network only two labels (human and spoof) is used, as more labels (indicating different spoof attacks S1 – S5) did not bring any improvements. Python programming language is used for the DDAE training and EER calculations. The DDAE is trained on a GTX 1070 TI GPU.

3.4. Results

Table 2 shows the results for the development set. As expected, the EER increases while the SNR decreases. Besides, the proposed system achieved a good performance for the clean conditions. Similarly, Table 3 shows the results for the evaluation set. Detection performance of the S2 attack was poor compared to the others in both development and evaluation sets. Another interesting result was observed for the S8 attack in clean condition, which produced more EER than the S10 attack. In general S10 attack was considered to be the most detrimental, which is a unit selection-based speech synthesis method [4]. A similar result was also obtained with multi-condition trained CQCC – GMM in [30].

In Table 4, the results of the proposed system are compared with previous studies. Although it is not possible to make an exact comparison due to the randomness while adding the noise, and also at the training stages, at least a general idea can be developed. Also, instead of analyzing each noise type, SNR level, and spoof attack, only the average values were considered to an excessive table size. Methods selected for the comparison are multi-conditionally trained CQCC – GMM and FBANK – GMM systems of [30]. On average, the results indicated that the proposed DDAE system generally performed better than the multi-conditionally trained GMM systems with a few exceptions such as car noise at 20 dB and 10 dB SNR levels, and street noise at 10 dB and 0 dB SNRs. On the other hand, a more advanced network architecture is necessary to compete with the state-of-the-art system reported in [14].

Table 2. EER (%) results for the development set

Noise Type	SNR (dB)	S1	S2	S3	S4	S5	Average
Clean		0.05	0.36	0.23	0.18	0.28	0.22
Seen Conditions							
Babble	20	11.94	18.82	10.16	10.23	8.67	11.94
	10	24.43	31.55	16.06	16.6	45.64	20.86
	0	37.56	42.66	28.02	28.63	27.39	32.85
White	20	18.68	32.14	16.94	17.34	23.17	21.65
	10	24.26	36.63	21.99	21.64	27.99	26.50
	0	37.12	48.16	26.6	25.95	43.38	36.24
Street	20	12.48	19.54	9.61	9.33	10.41	12.27
	10	23.15	29.57	14.31	14.9	18.58	20.10
	0	35.21	40.4	23.57	23.35	28.01	30.11
Unseen Conditions							
Car	20	0.59	2.42	1.53	1.46	1.03	1.41
	10	2.67	7.11	3.88	3.85	2.44	3.99
	0	8.1	16.06	7.58	7.57	5.83	9.03
Cafe	20	16.88	24.3	14.08	14.31	16.25	17.16
	10	26.35	33.45	21.18	21.56	24.03	25.31
	0	38.24	42.89	34.31	34.11	35.05	36.92

Table 3. EER (%) results for the evaluation set under different noise configurations.

Noise Type	SNR (dB)	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Average
Clean		0.09	0.53	0.23	0.18	0.21	0.26	0.13	7.97	0.09	5.12	1,48
Seen Conditions												
Babble	20	13.02	19.99	9.55	9.61	9.12	13.99	9.95	17.27	12.01	23.01	13.75
	10	24.92	31.32	14.48	14.69	15.96	22.42	20.01	20.62	22.89	30.77	21.81
	0	37.45	40.93	26.07	26.19	26.3	32.2	35.37	30.34	36.03	38.45	32.93
White	20	18.26	30.57	15.21	14.93	21.64	27.25	19.81	21.94	19.16	21.74	21.05
	10	24.04	34.38	19.83	19.47	26.21	32.86	24.25	25.89	24.1	25.86	25.67
	0	35.62	44.31	24.7	24.24	40	43.18	37.74	33.63	38.94	31.07	35.34
Street	20	12.84	20.32	9.05	8.77	10.47	14.31	10.46	20.14	14.47	20.63	14.15
	10	23.57	30.12	14.62	14.67	19.12	24.43	18.8	23.53	25.09	27.24	22.12
	0	34.38	39.56	21.48	21.5	27.1	32.88	29.19	28.52	32.82	35.32	30.27
Unseen Conditions												
Car	20	0.57	2.53	1.31	1.18	0.65	0.93	0.33	8.35	0.74	8.01	2.46
	10	2.79	7.79	3.78	3.45	2.08	3.67	2	12.58	3.66	11.94	5.37
	0	9.31	17.72	7.34	7.29	6.08	10.44	7.35	14.99	10.97	19.32	11.08
Cafe	20	16.89	24.63	13.2	13.19	15.72	20.62	13.94	21.96	18.16	24.19	18.25
	10	25.86	32.89	19.9	19.74	23.56	28.76	21.35	25.62	26.4	31.47	25.55
	0	37.68	41.58	32.07	32.46	33.8	37.74	34.12	34.9	36.98	40.45	36.18

Table 4. Comparison of different systems for the evaluation data based on average EER (%) for K=Known, U=Unknown, A=All data.

		Proposed System			CQCC-GMM[30]			FBANK-GMM[30]		
Noise Type	**SNR (dB)**	**K**	**U**	**A**	**K**	**U**	**A**	**K**	**U**	**A**
Clean		0.24	2.71	1.48	0.1	0.9	0.5	3.2	8.6	5.6
Seen Conditions										
Babble	20	12.25	15.24	13.75	18.2	18.3	18.3	14.5	17.4	16.0
	10	20.27	23.34	21.81	33.9	33.6	33.8	18.1	20.5	19.3
	0	31.38	34.47	32.93	44.6	44.0	44.3	29.6	31.1	30.3
White	20	20.12	21.98	21.05	46.8	44.6	45.7	17.0	19.1	18.0
	10	24.78	26.59	25.67	48.9	48.1	48.5	23.7	24.5	24.1
	0	33.77	36.91	35.34	49.3	48.9	49.1	30.8	31.6	31.2
Street	20	12.29	16.00	14.15	22.7	22.3	22.5	14.5	17.9	16.2
	10	20.42	23.81	22.12	37.5	36.3	36.9	18.7	21.0	19.8
	0	28.80	31.74	30.27	46.1	45.4	45.8	29.1	28.3	28.7
Unseen Conditions										
Car	20	1.24	3.67	2.46	0.9	2.7	1.8	10.2	18.1	14.2
	10	3.97	6.77	5.37	4.3	5.6	4.9	14.3	19.7	17.0
	0	9.54	12.61	11.08	13.0	13.0	13.0	21.6	23.8	22.7
Cafe	20	16.72	19.77	18.25	30.7	30.1	30.4	17.9	20.5	19.2
	10	24.39	26.72	25.55	42.1	41.3	41.7	21.9	23.4	22.6
	0	35.51	36.83	36.18	47.5	47.1	47.3	40.8	38.5	39.6

4. Discussion

The results given in the previous subsection verify that the proposed approach effectively reduced the noise effects. Compared to the multi-conditionally trained systems and conventional speech enhancement methods (which were reported to be even more harmful than no enhancement at all), the DDAE system performed better except for a few cases. On the other hand, a more complex deep learning architecture delivered state-of-the-art performance for the ASVspoof 2015 data and the given noise types.

Analyzing the results, there may be some possible modifications to further increase the performance of the proposed systems. As stated previously, data imbalance can affect the i-vector performance. Although the number of i-vectors for the DDAE training was some sort of balanced due to the augmented human data, the imbalance may have already altered the i-vector extraction process. Therefore, using more balanced data for i-vector extraction could lead to more accurate i-vectors, which may eventually lead to a better trained DDAE system.

Another issue is the short length utterances. As discussed previously, i-vectors are known to be less effective for short durations [18]. Solving this problem may increase the DDAE performance, and also may give the opportunity to use other robust i-vectors frameworks such as [25].

A masking strategy could be developed for the i-vectors. Inspired by the state-of-the-art system, a mask for separating the noise and speech parts may boost the detection performance. Although the complex gated recurrent convolutional network achieved impressive results, the masking features almost halved the EER [14]. This verifies the importance of masking the noise before the classification occurs. Although the DDAE approach tries to compensate for the corrupted features, including a mask during the training process will likely to increase the performance. Two possible applications of the masking will be investigated in the future works. One of them is using the mask before the i-vector extraction, hence obtaining cleaner i-vectors and use the DDAE for both enhancing and classifying them. The other is applying the mask in the i-vector space, which will require some knowledge or assumption of how the noise signal can be interpreted in the i-vector space (such as the normal distribution assumption in [27].

As the i-vectors deliver state-of-the-art performances for speaker recognition, using them for spoof detection could be beneficial. Instead of using different systems for spoof detection, speaker recognition, and noise robustness, i-vectors could be used at each stage. At least, i-vectors could be a common part to create a more straight-forward system (combinations of i-vectors – PLDA, i-vectors – DDAE, etc.). As a final note, robust PLDA classifiers can also be achieved with multi-condition training or SNR aware training, whose robustness will be higher than the simple cosine scoring used in this paper. So, investigating those modifications in future studies is expected to be beneficial for both research and practical purposes.

5. Conclusion

In this work, robust synthetic speech detection was achieved with DDAE network and i-vector inputs. The network consists of two MLPs, where the first MLP acts as a DAE. ASVspoof 2015 data were used in the experiments with human and spoof classes. Five different noise types at three different SNR levels were used at the test stage. The experimental results showed that the proposed DDAE system can deliver a better performance than multi-conditionally trained GMMs, with CQCC and filterbank features. Also, for clean test signals, the proposed network achieved sufficient performance (especially for the development set where the average EER was 0.22%). On the other hand, there was a performance gap between the state-of-the-art system and the proposed one. The possible reasons for the limited performance and opportunities for further improvements were discussed. Considering this work and previous literature, robust spoofed speech detection requires more complex systems, and masking noise dominant features are highly effective. Although the conventional speech enhancement methods were found to be ineffective, enhancing the noisy i-vectors with DDAE can offer alternative solutions.

Acknowledgement

This work was supported by TUBITAK under project number 121E057.

Author's Contributions

Gökay Dişken: Organized the datasets and the experiments, wrote the manuscript, analyzed the results.
Zekeriya Tüfekci: Supervised the experiments, assisted in designing the network structure, analyzed the results.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, vol. 2, pp. 1698–1701.

[2] A. Nautsch *et al.*, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 3, no. 2, pp. 252–265, Apr. 2021.

[3] H. Delgado *et al.*, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 296–303.

[4] Z. Wu *et al.*, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.

[5] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, Sep. 2017.

[6] J. Yang and L. Liu, "Playback speech detection based on magnitude-phase spectrum," *Electron. Lett.*, vol. 54, no. 14, pp. 901–903, Jul. 2018.

[7] A. T. Patil, H. A. Patil, and K. Khoria, "Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection," *Comput. Speech Lang.*, vol. 72, no. 1, p. 101301, Mar. 2022.

[8] J. Yang, R. K. Das, and N. Zhou, "Extraction of Octave Spectra Information for Spoofing Attack Detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2373–2384, Dec. 2019.

[9] C. Zhang, C. Yu, and J. H. L. Hansen, "An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 684–694, Jun. 2017.

[10] S. Scardapane, L. Stoffl, F. Rohrbein, and A. Uncini, "On the use of deep recurrent neural networks for detecting audio spoofing attacks," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 3483–3490, 2017.

[11] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Commun.*, vol. 85, pp. 83–97, Dec. 2016.

[12] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "An Investigation of Spoofing Speech Detection Under Additive Noise and Reverberant Conditions," in *INTERSPEECH 2016*, 2016, pp. 1715–1719.

[13] A. Gómez Alanís, A. M. Peinado, J. A. Gonzalez, and A. Gomez, "A Deep Identity Representation for Noise Robust Spoofing Detection," in *Interspeech 2018*, 2018, pp. 676–680.

[14] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1985–1999, Dec. 2019.

[15] S. Mahto, H. Yamamoto, and T. Koshinaka, "i-Vector Transformation Using a Novel Discriminative Denoising Autoencoder for Noise-Robust Speaker Recognition," in *Interspeech 2017*, 2017, pp. 3722–3726.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[17] W. Rao *et al.*, "Neural networks based channel compensation for i-vector speaker verification," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016, pp. 1–5.

[18] H. Yamamoto and T. Koshinaka, "Denoising autoencoder-based speaker feature restoration for utterances of short duration," in *Interspeech 2015*, 2015, pp. 1052–1056.

[19] A. Varga and H. J. M. Steeneken, "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 Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[20] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Interspeech 2015*, 2015, pp. 3456–3460.

[21] C. Zhang *et al.*, "Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5035–5039.

[22] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint Speaker Verification and Antispoofing in the i-Vector Space," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 821–832, Apr. 2015.

[23] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4042–4046.

[24] D. Ribas and E. Vincent, "An Improved Uncertainty Propagation Method for Robust I-Vector Based Speaker Recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6331–6335.

[25] W. Ben Kheder, D. Matrouf, M. Ajili, and J.-F. Bonastre, "A Unified Joint Model to Deal With Nuisance Variabilities in the i-Vector Space," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 633–645, Mar. 2018.

[26] W. Ben Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, and P.-M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4190–4194.

[27] W. Ben Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, and M. Ajili, "Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition," *Comput. Speech Lang.*, vol. 45, pp. 104–122, Sep. 2017.

[28] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[29] W. Wang, W. Song, C. Chen, Z. Zhang, and Y. Xin, "I-vector features and deep neural network modeling for language recognition," *Procedia Comput. Sci.*, vol. 147, pp. 36–43, 2019.

[30] Y. Qian, N. Chen, H. Dinkel, and Z. Wu, "Deep Feature Engineering for Noise Robust Spoofing Detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1942–1955, Oct. 2017.

[31] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," *IEEE Speech Lang. Process. Tech. Comm. Newsl.*, pp. 1–4, 2013.