Research Article GU J Sci 36(4): 1553-1561 (2023) DOI: 10.35378/gujs.1062788

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Robust Spoofed Speech Detection with Denoised I-vectors

Gokay DISKEN

Adana Alparslan Turkes Science and Technology University, Electrical-Electronics Engineering Department, 01250, Adana, Turkey

Highlights

- This paper focuses on detection of synthetic speech under additive noise.
- Denoising autoencoder (DAE) network is used to obtain clean estimates of i-vectors.
- DAE – i-vector combination has not been applied to spoof detection previously.

Article Info

Received: 25 Jan 2022
Accepted: 06 Oct 2022

Keywords

Denoising autoencoder
Speaker verification
Spoof detection

Abstract

Spoofed speech detection is recently gaining attention of the researchers as speaker verification is shown to be vulnerable to spoofing attacks such as voice conversion, speech synthesis, replay, and impersonation. Although various different methods have been proposed to detect spoofed speech, their performances decrease dramatically under the mismatched conditions due to the additive or reverberant noises. Conventional speech enhancement methods fail to recover the performance gap, hence more advanced techniques seem to be necessary to solve the noisy spoofed speech detection problem. In this work, Denoising Autoencoder (DAE) is used to obtain clean estimates of i-vectors from their noisy versions. ASVspoof 2015 database is used in the experiments with five different noise types, added to the original utterances at 0, 10, and 20 dB signal-to-noise ratios (SNR). The experimental results verified that the DAE provides a more robust spoof detection, where the conventional methods fail.

1. INTRODUCTION

Speaker recognition systems, which verifies or identifies users from their utterances, achieved high performances with methods such as Gaussian mixture model – universal background model (GMM-UBM) [1], i-vectors [2], and x-vectors [3]. Usually, speaker verification systems' performances are examined by using the utterances from the real user and many other zero-effort imposter utterances. Using spoofing attacks instead of zero-effort imposters, the verification accuracy decreases rapidly [4–6].

Besides the spoofing attacks, environmental mismatch is another issue that affects the recognition performance. This type of mismatch usually occurs when the training data is obtained in a controlled environment such as a studio, office, etc., but the operating environment is different. Possible mismatch sources are the channel noises due to the different recording devices, additive noises such as car noises at outdoors, and reverberation [7–9].

The aforementioned problems, i.e. spoof detection and noise robustness, are addressed separately in general as the spoof detection is a relatively new research area, and none of the developed methods are effective against diverse spoofing attack types yet. Studies on the spoof detection gained momentum with the recent challenges and publicly available datasets of those challenges [5], [10–12]. The early investigations about spoof detection focused on the well-known features and classifiers used in speech processing literature [13,14]. Magnitude-based features dominate the speech/speaker recognition, but it is found that phase-based features are also effectively capturing spoofing information. For the classifiers, the i-vector approach, which provides state-of-the-art performance for text-independent speaker recognition, failed to compete with the conventional GMM method [15]. Besides the known features, several new features are also introduced for spoof detection. One of the most successful methods is the constant-Q cepstral coefficients

* e-mail: gdisken@atu.edu.tr

(CQCC) [16], which becomes a de facto standard. For the backend, deep learning based methods are used with various different architectures [17–21].

Only a few studies have considered noisy spoof detection [15,21,22]. In [15], several different features and classifiers are used for noise robust spoof detection on ASVspoof 2015 database, which includes synthetic speech attacks. Speech enhancement methods [23] such as spectral subtraction, Wiener filtering, minimum mean square error (MMSE), logarithmic MMSE are reported to further deteriorate the performance compared to the baseline GMM system without any enhancement. Therefore, the results of [15] proves the need of more advanced systems for robust spoof detection. In fact, [21,22] have shown that deep learning architectures incorporated with noise aware training and signal-to-noise masks are highly effective solutions, except their computational loads. Nevertheless, those deep architectures are used to extract identity vectors, which are then sent to the classifiers such as GMM, support vector machine, etc. Hence, they are highly specialized systems for spoof detection, and cannot share the same feature/classifier with a speaker recognition system.

In this work, i-vectors are used for noise robust spoof detection, where a Denoising Autoencoder (DAE) network is hired to estimate clean (denoised) i-vectors from their noisy counterparts. A similar approach is used in [24] for robust speaker recognition. To the best of the author's knowledge, robustness of the i-vectors has not been examined for spoof detection. Since the classical speech enhancement methods are not suitable for this task, and the deep learning solutions are not directly applicable within a speaker recognition system, it is worth exploring the performance of the proposed i-vector based system. The same i-vector can be used for both spoof detection and speaker recognition, without excessive increment on the computational load. Following the limited literature on the robust spoof detection, ASVspoof 2015 dataset is used in this study.

2. PROPOSED ROBUST SYSTEM

In this section, the conventional i-vector framework is briefly described. Then, the DAE network used in the experiments is introduced. All necessary parameters of the proposed system are also given in this section for completeness.

2.1. I-vector

I-vectors are fixed dimensional representation of variable length utterances. This property gives the opportunity of applying different normalization/enhancement technique in a low dimensional space. Following the conventional framework of [2], i-vector extraction mainly requires training of a UBM and a total variability matrix, *T*. A speaker and channel independent GMM supervector can be defined as

$$M = m + T\omega \tag{1}$$

where m is the mean supervector taken from the UBM, and ω is a random vector with normal distribution. The i-vector is obtained by the maximum a posterior estimate of ω for each utterance. Once the i-vectors are extracted, various compensation and dimensionality reduction techniques may applied such as within-class covariance normalization (WCCN), linear discriminant analysis (LDA), nuisance attribute projection (NAP) [2], length normalization [25]. For the scoring part, support vector machines, cosine distance, probabilistic LDA (PLDA), and two-covariance-based scoring are among the alternatives [2,26,27].

The proposed i-vector system follows the recipe of [10], where 19 dimensional CQCC features are extracted from each utterance, delta and acceleration coefficients are also appended. A UBM with 64 mixtures is trained using the train partition of ASVspoof 2015 data. The partitions of the database are given in Table 1, and further details are given in Section 3. Total variability matrix, *T*, has 100 factors. The extracted i-vectors are mean normalized, whitened, and WCCN is applied. Single i-vectors are constructed for human and spoofed speech by averaging the respective i-vectors of each class. Length normalization is also included.

For the scoring part, cosine distance scoring is employed. It can be computed using a target and a test i-vector. In this case, two target classes are available as human (ω_{hum}) and spoof (ω_{spo}), as described previously. A final score for a given test i-vector (ω_{test}) is calculated as

$$score = \cos(\omega_{hum}, \omega_{test}) - \cos(\omega_{spo}, \omega_{test}) \quad (2)$$

where *cos* is the cosine distance given below, and the denominator will be equal to one due to the length normalization

$$cos(\omega_a, \omega_b) = \frac{\langle \omega_a, \omega_b \rangle}{\|\omega_a\| \|\omega_b\|} \ . \quad (2)$$

2.2. Denoising i-vector

As stated in [24], the i-vector extraction process is a non-linear process. Hence, to overcome the non-linear effects of the noise in i-vector space, neural networks can be used. Although different approaches have been proposed to reduce the noise effects in the i-vector space [27,28], they require a lot of computational power, and do not perform well for utterances with short duration, which is the case for ASVspoof 2015 database. Therefore, a DAE network is preferred in this work, as it was proven to increase the robustness of speaker recognition systems [24]. The DAE network estimates a clean output based on the corrupted input, by learning a non-linear mapping between them [29].

The proposed DAE consists of two fully connected hidden layer with ReLU activation functions. Each hidden layer has 500 units, and dropout with 0.5 probability is added to each hidden layer to prevent overfitting. The inputs of the network are the noisy i-vectors, and the outputs are the denoised versions. Hence, both the input and the output dimensions are 100. The output layer is a regression layer, and the objective function of the network is to minimize the mean square error between the original clean i-vector and the denoised i-vector.

Following the related literature on noisy spoof detection [15,22,30], babble, car, and white noises are chosen from the Noisex-92 database [31], and street and café noises are chosen from the QUT-NOISE database [32]. Each utterance of the training data is corrupted with one of the noises from the Noisex-92 database chosen randomly. SNR levels are also selected randomly in the range of 0 to 20 dB with 5 dB steps. Street and café noises are only used in the test stage to simulate unseen environments. Also, in order to minimize the effects of unbalanced distribution of the training data, each of the utterances in the human partition is corrupted three times with a random noise type and SNR level as mentioned. As a result, a total of 11250 clean-noisy pairs are created for the human data, and 12625 clean-noisy pairs are created for the spoof data.

As the outputs of the DAE are assumed to be clean, the final scores are computed with the Equation 2, where i-vectors representing each class are observed by averaging, as discussed in the previous subsection. Since the noise effects are handled in the DAE, any kind of scoring process for i-vectors can be used without any modifications. In this work, the cosine distance is preferred because it is one of the basic methods that does not require training or any other information about the data. Hence, the robustness of the proposed system can be mainly attributed to the DAE. Therefore, whether the i-vector/DAE system increases the robustness of the spoof detection or not will be verified experimentally, without any support from front-end or back-end.

Table 1. *Partitions of ASVspoof 2015 database*

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

3. EXPERIMENTAL RESULTS

The spoof attack types of ASVspoof 2015 database are divided into 10 sub-parts. Five of them (S1–S5) are available in each partition, hence named as known attacks. The remaining parts (S6–S10) are only available in the evaluation partition, so called as unknown attacks. The generalization capacity of the spoof detection systems is also important to capture those unknown attacks. Attacks named as S1, S2, S5, S6, S7, S8, and S9 are voice conversion attacks. S3, S4, and S10 are speech synthesis attacks. S4 and S10 methods are trained with 40 utterances, and the other attacks are trained with 20 utterances. The sampling frequency of the dataset is 16 kHz. Also, there is no speaker overlap within training, development, and evaluation partitions. Average utterance duration is about 3.5 seconds, which is a drawback for the i-vectors as they tend to perform better with longer utterances. The original data do not include channel or background noise. Further details about the dataset (such as the algorithms of spoof attacks) can be found in [11].

The proposed DAE network was trained using the clean-corrupted pairs of the training data and validated on the development data. The baseline method to compare the results of the proposed approach is the CQCC–GMM method reported in [22]. The reasons for choosing this baseline are as follows;

- The proposed system shares the same feature type (CQCC).
- The GMM method is known to be superior to the i-vectors for spoof detection [15,22].
- Several speech enhancement methods such as spectral subtraction, Wiener filtering, MMSE, etc. are reported to be further detrimental than no enhancement at all for the GMM [15].
- Multi-condition training (i.e. pooling all noisy data in the training set) was used in the baseline system, which is similar to the proposed DAE network.

For completeness, CQCC – i-vector with cosine scoring [15] results for the development set are also given in order to verify the DAE's functionality. Comparison with the deep learning based methods are neglected due to the reasons mentioned in the Introduction section. In fact, the state-of-the-art system [21] consist of two feature sets (magnitude and phase based), a separately trained CNN for signal-to-noise mask estimation, gated recurrent convolutional neural networks for each feature set, and finally a PLDA classifier. This highly specialized system achieved impressive performances even under low SNR conditions. On the other hand, the proposed system in this study aims to increase the robustness of the i-vectors, as the same i-vectors can be used for both spoof detection and speaker recognition simultaneously. Therefore, it is conservative compared to the state-of-the-art.

The performance metric used in this study is equal error rate (EER). EER is the standard metric for assessing the performance of automatic speaker verification systems. It is also commonly used in spoof detection. EER is the point where false acceptance rate (misclassified spoof attacks) and false rejection rate (misclassified human speech) are equal. Lower EER value indicates higher accuracy. Once the scores for the test utterances are obtained, Bosaris Toolkit within the baseline codes for the ASVspoof 2015 data is used to calculate the EER, providing the labels for each utterance.

In Table 2, results for the development data are given in terms EER, which was computed by considering each individual attack type, then averaged over all attacks and shown under the Avg. column. Note that the street and café noises are unseen conditions for the proposed system, as they were not included in the DAE

training. Nevertheless, they perform relatively similar to the seen noise conditions. The best performance was achieved under the presence of car noise, due to the fact that it is a stationary noise and hence easier to detect noise statistics. Average results for the conventional CQCC – i-vector system are given in the last column. Note that café and street noises were not available for the conventional system. Comparing the average EERs, it is clear that the DAE network increased the robustness of the i-vectors.

Table 2. *EER (%) results for the development set under different noise configurations*

Noise Type	SNR (dB)	S1	S2	S3	S4	S5	Avg.	CQCC – i-vector [15]
Babble	20	9.75	16.22	10.16	10.58	8.61	11,06	27.63
	10	21.87	28.76	16.35	16.75	16.19	19,98	39.21
	0	38.05	42.95	27.12	27.74	28.38	32,85	46.20
White	20	15.75	26.98	15.98	16.15	25.75	20,12	41.55
	10	22.26	34.37	19.58	19.4	29.73	25,07	44.76
	0	30.29	44.62	24.4	24.03	41.79	33,03	48.27
Car	20	0.55	2.24	1.88	1.75	1.28	1,54	13.46
	10	2.47	6.65	4.04	4.11	2.87	4,03	25.53
	0	6.4	13.75	7.2	6.94	5.43	7,94	37.84
Cafe	20	14.95	22.38	14.8	15.46	16.83	16,88	-
	10	25.06	31.05	20.68	21.21	25.07	24,61	-
	0	37.52	41.23	29.83	29.63	34.62	34,56	-
Street	20	11.9	18.03	10.91	11.04	11.3	12,63	-
	10	23.22	28.68	17.83	18.51	19.87	21,62	-
	0	33.27	37.66	23.83	24.09	28.27	22,77	-

Table 3 shows the results for the evaluation set, and the results for the baseline method. It should be reminded that for the baseline method, car and street noises were the unseen conditions, to examine the performance under unseen stationary noise (i.e. car). However, eliminating the stationary noises are much easier than the non-stationary noises as seen in both Table 2 and Table 3. Comparing the performances under the non-stationary street noise, which was an unseen condition for the proposed system, and a seen condition for the baseline, it can be observed that the proposed DAE network is more robust even the noise statistics were not present in the training data. In general, the proposed system performed superior to baseline except for the car noises at 10 and 20 dB SNRs.

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

Proposed System													CQCC-GMM [22]
Noise Type	**SNR (dB)**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**Average**	
Babble	20	10.38	16.57	9.97	9.96	9.06	13.15	7.99	17.97	10.68	22.86	**12,85**	18.3
	10	21.74	27.75	14.87	14.75	15.86	22.09	17.84	20.62	23.68	28.8	**20,8**	33.8
	0	35.83	38.63	23.77	24.09	25.56	31.17	31.37	27.68	34.37	34.98	**30,74**	44.3
White	20	14.84	26.12	13.72	13.59	24.19	27.8	16.73	23.94	20.25	18.25	**19,64**	45.7
	10	23.06	32.77	17.44	17.2	28.8	33.35	22.95	25.65	26.53	21.81	**24,96**	48.5
	0	28.64	40.5	21.53	21.43	37.87	41.22	31.13	31.42	36.89	26.35	**31,69**	49.1
Car	20	0.78	2.54	1.96	1.78	1.19	1.46	0.32	7.52	0.79	14.42	3,27	**1.8**
	10	2.94	7.74	4.23	4.06	2.89	4.26	1.73	10.84	3.62	17.04	5,93	**4.9**
	0	7.92	16.21	7.26	7.23	5.98	9.69	5.61	13.46	8.76	23.01	**10,51**	13.0
Cafe	20	14.21	21.49	13.51	13.36	15.35	19.3	11.61	22.94	15.55	24.39	**17,17**	30.4
	10	24.13	30.18	19.07	19.16	23.55	27.82	18.32	26.18	26.07	30.52	**24,5**	41.7
	0	35.18	38.79	26.72	27.19	32.46	35.17	29.41	31.55	35.18	36.68	**32,83**	47.3
Street	20	12.05	18.84	10.39	10.01	11.04	14.55	9.33	20.76	13.21	23.8	**14,4**	22.5
	10	23.21	29.45	17.6	17.92	19.83	25.51	18.26	26.54	26.92	31.2	**23,64**	36.9
	0	32.37	37.72	22.58	22.84	27.87	32.39	28.21	31.82	35.91	36.2	**30,8**	45.8

4. DISCUSSION

The proposed i-vector/DAE based system is proven to be more robust than the CQCC-GMM baseline. One of the most important outcomes of this work is that even the GMMs are favored over the i-vectors for spoof detection, the presence of the additive noise may affect this preference. As the conventional speech enhancement methods are not adding any robustness to the GMM systems, their practical usage will be limited. The i-vectors, on the other hand, has the advantage of representing the variable utterances as fixed dimensional vectors. Using this property, the DAE network is effectively reducing the additive noise artifacts, as verified using the ASVspoof 2015 database.

Although the proposed system outperformed the baseline, its performance was not comparable to the more advanced deep learning architectures such as [21,22]. The main reason for that is the short duration of the utterances in the database, where the average is about 3.5 seconds [18]. Considering the fact that i-vectors perform better with long utterances (e.g. 30 seconds) [33], the performance of the proposed method becomes more impressive. Yet, more investigation is needed to further increase the performance, and close the gap between the deep learning based systems. Then, the same i-vectors can be used for both spoof detection and speaker verification under additive noise.

5. CONCLUSION

Robustness of the spoof detection systems against noise is an important issue for practical implementations and for the security of speaker recognition systems. The i-vectors fail to provide sufficient performance for spoof detection, especially due to the short duration utterances in the available databases. Contrary, in the presence of the noise, the performance of the conventional systems decreases rapidly as the SNR decreases, and classical speech enhancement methods make the situation even worse.

In this work, a DAE network is used to provide clean estimates from the noisy i-vectors. The proposed system is tested against five different additive noise types at three different SNR levels. Two of these noises were not present in the training data to simulate unseen conditions at the test stage. Out of 15 different noise configurations, the proposed system performed superior than the CQCC-GMM baseline except two cases. It is experimentally proved that the i-vector/DAE combination is more robust than the conventional speech enhancement methods with a GMM classifier. Future research efforts will focus on achieving a competitive performance compared to the advanced deep learning architectures.

FINANCIAL DISCLOSURE

This study was supported by TUBITAK under project no. 121E057.

CONFLICTS OF INTEREST

No conflict of interest was declared by the author.

REFERENCES

[1] Reynolds, D.A., Quatieri, T.F., and Dunn, R.B., “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, 10: 19–41, (2000).

[2] Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., and Ouellet, P., “Front-End Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech, and Language Processing, 19: 788–798, (2011).

[3] Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S., “Deep neural network-based speaker embeddings for end-to-end speaker verification”, in 2016 IEEE Spoken Language Technology Workshop (SLT), 165–170, (2016).

[4] Wu, Z., Chng, E.S., and Li, H., "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, 2: 1698–1701, (2012).

[5] Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T., and King, S., "SAS: A speaker verification spoofing database containing diverse attacks", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4440–4444, (2015).

[6] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H., "Spoofing and countermeasures for speaker verification: A survey", Speech Communication, 66: 130–153, (2015).

[7] Gales, M.F.J., and Young, S.J., "Robust speech recognition in additive and convolutional noise using parallel model combination", Computer Speech and Language, 9(4): 289–307, (1995).

[8] Fujimoto, M., and Riki, Y.A., "Robust speech recognition in additive and channel noise environments using GMM and EM algorithm", in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1: I-941–4, (2004).

[9] Zhao, X., Wang, Y., and Wang, D., "Robust Speaker Identification in Noisy and Reverberant Conditions", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4): 836–845, (2014).

[10] Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., and Yamagishi, J., "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements", in Odyssey 2018 The Speaker and Language Recognition Workshop, 296–303, (2018).

[11] Wu, Z., Yamagishi, J., Kinnunen, T., Hanilçi, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, and M., Delgado, H., "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge", IEEE Journal of Selected Topics on Signal Processing, 11(4): 588–604, (2017).

[12] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K.A, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in Interspeech 2019, 1008–1012, (2019).

[13] Sahidullah, M., Kinnunen, T., and Hanilçi, C., "A comparison of features for synthetic speech detection", in INTERSPEECH 2015, 2087–2091, (2015).

[14] Hanilçi, C., Kinnunen, T., Sahidullah, M., and Sizov, A., "Classifiers for Synthetic Speech Detection: A Comparison", in INTERSPEECH 2015, 2057–2061, (2015).

[15] Hanilçi, C., Kinnunen, T., Sahidullah, M., and Sizov, A., "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise", Speech Communication, 85: 83–97, (2016).

[16] Todisco, M., Delgado, H., and Evans, N., "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification", Computer Speech and Language, 45: 516–535, (2017).

[17] Gomez-Alanis, A., Peinado, A.M., Gonzalez, J.A., and Gomez, A.M., "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection", in Interspeech 2019, 1068–1072, (2019).

[18] Zhang, C., Yu, C., and Hansen, J.H.L., "An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing", IEEE Journal of Selected Topics on Signal Processing, 11(4): 684–694, (2017).

[19] Xiao, X., Tian, X., Du, S., Xu, Chng, E.S., and Li, H., "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge", in INTERSPEECH 2015, 2052–2056, (2015).

[20] Dua, M., Jain, C., and Kumar, S., "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems", Journal of Ambient Intelligence and Humanized Computing, 12(2): 1–16, (2021).

[21] Gomez-Alanis, A., Peinado, A.M., Gonzalez, J.A., and Gomez, A.M., "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12): 1985–1999, (2019).

[22] Qian, Y., Chen, N., Dinkel, H., and Wu, Z., "Deep Feature Engineering for Noise Robust Spoofing Detection", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(10): 1942–1955, (2017).

[23] Loizou, P.C., Speech enhancement: Theory and practice. Taylor & Francis, (2013).

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

[25] Garcia-Romero, D., and Espy-Wilson, C.Y., "Analysis of I-vector Length Normalization in Speaker Recognition Systems", in INTERSPEECH 2011, 249–252, (2011).

[26] Yamagishi, J., Lee, K.A. and Wang, L., "PLDA in the i-supervector space for text-independent speaker verification", EURASIP Journal on Audio, Speech, and Music Processing, 2014(29): 1-13, (2014).

[27] Ben Kheder, W., Matrouf, D., Bousquet, P.M., Bonastre, J.F., and Ajili, M., "Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition", Computer Speech and Language, 45: 104–122, (2017).

[28] Ben Kheder, W., Matrouf, D., Ajili, M., and Bonastre, J.F., "A Unified Joint Model to Deal with Nuisance Variabilities in the i-Vector Space", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(3): 633–645, (2018).

[29] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.A., "Extracting and composing robust features with denoising autoencoders", in Proceedings of the 25th international conference on Machine learning, 1096–1103, (2008).

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

[31] Varga, A., and Steeneken, H.J.M., "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(3): 247–251, (1993).

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

[33] Guo, J., Xu, N., Qian, K., Shi, Y., Xu, K., Wu, Y., and Alwan, A., "Deep neural network based i-vector mapping for speaker verification using short utterances", Speech Communication, 105: 92–102, (2018).