Compensate multiple distortions for speaker recognition systems

Mohammad Mohammadamini[1], Driss Matrouf[1], Jean-Francois Bonastre[1], Romain Serizel[2], Sandipana Dowerah [2], Denis Jouvet[2]

[1]LIA (Laboratoire Informatique d'Avignon), Avignon University

[2]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

Abstract*—The performance of speaker recognition systems reduces dramatically in severe conditions in the presence of additive noise and/or reverberation. In some cases, there is only one kind of domain mismatch like additive noise or reverberation, but in many cases, there are more than one distortion. Finding a solution for domain adaptation in the presence of different distortions is a challenge. In this paper we investigate the situation in which there is none, one or more of the following distortions: early reverberation, full reverberation, additive noise. We propose two configurations to compensate for these distortions. In the first one a specific denoising autoencoder is used for each distortion. In the second configuration, a denoising autoencoder is used to compensate for all of these distortions simultaneously. Our experiments show that, in the co-existence of noise and reverberation, the second configuration gives better results. For example, with the second configuration we obtained 76.6% relative improvement of EER for utterances longer than 12 seconds. For other situations in the presence of only one distortion, the second configuration gives almost the same results achieved by using a specific model for each distortion.

Keywords— additive noise, early reverberation, full reverberation, x-vector, denoising autoencoder

I. Introduction

In recent years, deep neural networks became the most commonly used approach for speaker modeling [1, 2]. Although DNN-based speaker recognition systems outperform their previous statistical systems such as i-vector, still their performance degrades in the presence of acoustic distortions, such as reverberation and additive noise. Using a huge amount of data makes the DNN based systems robust against noise and reverberation but in severe conditions the performance can degrade drastically [3]. Therefore, explicit denoising and dereverberation can make these systems more robust [4,5]. Approaches based on speech signal enhancement [6] and denoising techniques at the speaker modeling level are often proposed to reduce the impact of noise and reverberation on speaker recognition systems [7].

In the previous work [4,7] we proposed several Denoising Autoencoders (DAE) to reduce the effect of additive noise in x-vector space. The proposed DAEs estimate the clean x-vector from the noisy version. In this paper we want to extend the application of proposed denoising techniques to other distortions like early and full reverberation. Another contribution of our work is proposing two configurations for five environments, where there is one or more of these distortions: additive noise, early reverberation, full reverberation. Using data augmentation with noise and reverberation in training the speaker embedding network and denoising techniques in signal level, feature level and speaker modeling level are common approaches to make the system robust against noise and reverberation. It is important to specify that both strategies are important and none of them could be replaced by another one. Indeed, the proposed DAE tries to learn the relation between the x-vector affected by a given distortion and its clean version. Thus, it is a direct learning of the distortion effect, which would make the denoising system more efficient than multi-condition training [4]. The DAE uses more specific information about the distortion than what would be used in a data augmentation approach. Moreover, the DAE we propose is trained in x-vector space, which makes its training very quick, with respect to the x-vector extractor training.

In this article the additive noise and reverberation are simulated. The artificial reverberation is applied using image model [8]. Obviously, the use of test data from a real environment is preferable. Unfortunately, we do not have such a database at the moment. However, the simulation model we used has proven that is effective in generating real data [9].

The paper is organized as follows. Section II reviews related works. Section III describes the two proposed configurations (specific and general DAE) for handling different distortions. Section IV is devoted to experimental setup, and Section V presents and discusses the results.

II. Related works

Despite significant advances, the performance achieved by speaker recognition systems degrades dramatically when they are used in real applications where channel mismatch as well as environmental additive noise and reverberation are present separately or simultaneously. Many efforts have been put into modeling and compensating these distortions throughout the years by working on different levels. Reducing the negative effect of noise and reverberation could be treated in signal level or higher-levels (i.e., speaker modeling level).

In signal level, several works have been done to compensate for additive noise and/or reverberation. In [10] a system proposed for conditions in the presence of additive noise and reverberation. In this work, firstly the additive noise is removed through a binary masking estimated by a neural network. For reverberation the speaker model trained in reverberant condition. Also, some works are done at signal level to make the x-vector framework robust against noise and reverberation. In [11] LSTM and CNN networks are used to denoise the speech signal. The LSTM network is trained to reconstruct the clean log magnitude spectrum. The CNN network is used to denoise the short-time Fourier transform (STFT) magnitude spectrogram. The proposed LSTM with 20% to 30% relative improvement of EER, has given better results in comparison to other methods used in this research. In [6] a denoising autoencoder is used for joint compensation of additive noise and reverberation in the x-vector framework. In this research, the DAE reconstructs the clean version of the magnitude spectrum. In the best case a 30% percent improvement of EER has been reported. In [12] several masking-based beamformers used for denoising and dereverberation. The MVDR Rank1 beamformer gave the best results for the real RIRs, and the GEVBAN has given the best results with simulated RIRs. A DNN that supports acoustic beamforming and dereverberation is used in the frontend of x-vector framework [13]. The joint training of this network and x-vector network has improved the performance of speaker recognition with 40% percent improvement of EER for simulated RIRs and 25% improvement for real RIRs.

There are some works in speaker modeling level to compensate for a specific distortion. In [7], statistical i-map, and three derivations of denoising autoencoders presented to suppress the effect of additive noise in the x-vector domain. In [4] the importance of using denoising techniques alongside data augmentation was explored to make the x-vector system robust against additive noise. The 66% relative improvement of EER in the x-vector domain shows that noise compensation in speaker modeling level is very effective for speaker recognition systems. Siamese x-vector reconstruction was introduced in [5] to compensate for additive noise and sampling rate mismatch in the x-vector framework. It is shown that this technique is more effective to compensate sampling rate mismatch rather than additive noise. In the previous generation of speaker recognition systems (i.e., i-vector) the mapping from distorted to clean vectors is explored broadly. i-MAP [14] and joint i-MAP [15] are two statistical techniques used for noise compensation in the i-vector framework.

However, there is no specific research to compare the state-of-the-art techniques in signal level with speaker modeling level in the same experimental setup, the results reported by compensation techniques in speaker modeling level from previous works are more promising and working in speaker modeling level is easier than signal level. Also, the reviewed works done in speaker modeling level focused on one distortion. In the current paper we extend this approach for situations where there are different distortions.

III. System configuration

Our study will focus on 5 conditions and their combinations: Clean (D), additive Noise (N), Early reverberation (E), Full reverberation (F) and additive noise with Full reverberation (FN).

The reverberation is the sum of sound reflections arrived at a single point inside an acoustical enclosure. Early reflections which are called early reverberation arrive between 50-100ms after the arrival of the direct signal. The full reverberation is the next echoes that arrives to listener with longer delays [16].

In the next sections, the DAE(N), DAE (E), DAE (F), DAE (FN), DAE (N+E+F+FN) stands for experiments that the input of denoising autoencoder is noisy x-vector data, early reverberated x-vector data, the full reverberated x-vector data, the simultaneously noisy fully reverberated x-vector data and finally a combination of x-vectors for all distortions. The output of the DAE is always the clean x-vector data.

In this paper we compare two approaches in handling multi-acoustic distortions. In the first approach, after the x-vector network, a specific denoising autoencoder is used for each distortion. For clean x-vectors the scoring is done without passing them through a DAE. The details of this configuration are depicted in Figure 1. In this configuration, we assume that the type of distortion is known. We will show when the type of distortion is unknown, using a classifier can help to detect the kind of distortion automatically.

In the second configuration depicted in Figure 2, the compensation for different distortions is done by using one single DAE. In this configuration like the specific compensation, there is clean speech, distorted speech with additive noise, early reverberation, full reverberation, additive noise and full reverberation. With this system, it is not necessary to have previous information about the kind of distortion. As it is shown in results, the denoising autoencoder learns to compensate all those distortions simultaneously. In the case of clean speech, we show that, without any change in the system and without having

previous information about the environment, the denoising part does not have a negative effect on clean x-vectors.

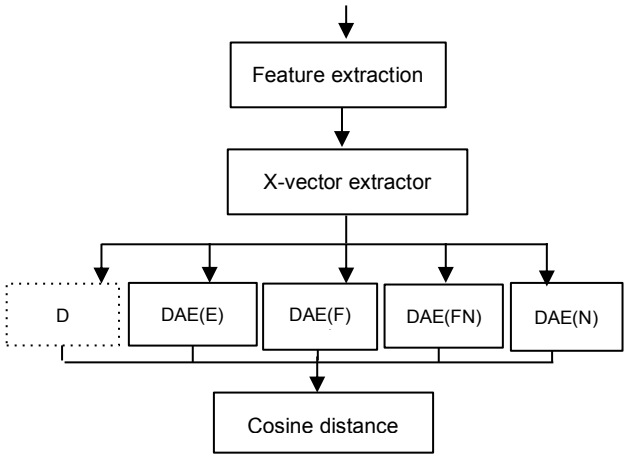

Figure 1. Using specific models for each distortion

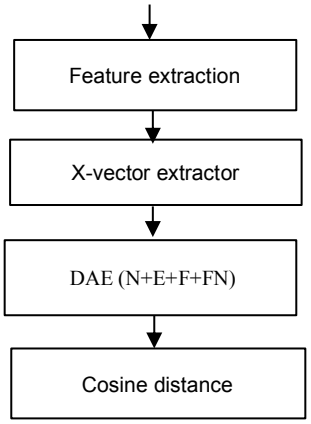

Figure. 2 General domain adaptation for different distortions

A. Noise and reverberation data simulation

To train the DAE for dereverberation or for denoising we need to have a set of x-vector pairs, distorted/clean. "Distortion" here corresponds to reverberation or additive noise. The reverberated version of the speech clips is obtained by convolving the original speech clips with room impulse responses (RIR) simulated with the pyroomacoustics[1]. The RIR were designed to simulate rooms whose dimensions are sampled randomly between [3m*4m*2.5m] and [6m*8m*3.5m]. The reverberation time for the rooms (RT_{60}) is drawn randomly between [200ms] and [600ms]. The microphone and the speech source are located at least 1m away from any wall. The microphone is at 0.5m height (to simulate a small robot on the floor) and the speech source height is drawn randomly in [1.6m, 1.9m] (to simulate a human standing). The distance between the speaker and the microphone is at least 1 meter. We generated 10000 RIR for training and 2000 RIR for the test. When considering only early reverberation the RIR is truncated to 50ms after the RIR first peak. When additive noise is present, the noise source clips are office noises collected from Freesound [17]. We collected 3275 clips for training and 1000 clips for the test. The training/test split is designed such that there is no overlap in terms of Freesound users between both the sets. The original noise clips are drawn randomly and convolved with a RIR from the same room as the speech clip. The noise source is located at least 1m away from any wall at a height of [1.6m, 1.9m]. The noise files are added with random SNR between [0,10].

IV. Experimental setup

The x-vector network introduced in [1] was used to create x-vectors for train and test data. The x-vector network has been trained with Voxceleb1, Voxceleb2 [2] and one million utterances From Voxceleb 1 and Voxceleb 2 augmented with different parts of Musan corpus (Noise, Babble, Speech) and real RIRs [18]. The x-vector extractor has been used to create clean, noisy and reverberant data to train DAE. The data simulation procedure is described in 3.1.

For each experiment firstly the reverberation and/or noise were added to Voxceleb2. Then pairs of distorted/clean x-vector were produced to train DAE. In the same manner, the data distorted for the test. In all cases the enrollment data is clean; because we assume that in real applications it is possible to have clean data for enrollment. In the experiments, we use deep stacked DAE introduced in [3]. The architecture (number of layers and neurons) is the same in all experiments. In the backend the cosine metric is used for scoring.

The FABIOL corpus was used for test and enrollment. In FABIOL corpus there are 6992 files from 130 speakers. Form 130 speakers in FABIOL, for 100 speakers there is a small number of utterances. We used these speakers just in enrollment. The utterances belonging to the remaining 30 speakers are randomly separated for the test and enrollment. In enrollment there are 3576 files and 3244 files are used for the test. Because the duration of files in FABIOL corpus spans from very short to long, the test files are separated into seven groups for each two seconds.

V. Results and discussions

The results obtained in the presence of each distortion are summarized in Table 1. When we are in a clean environment without noise or reverberation, we show that the use of DAE(F+N) gives almost the same (sometimes better) results as the baseline system. This is an interesting property of the proposed approach. If the x-vector belongs to the clean class, the scoring could be done directly. But

[1] https://github.com/LCV/pyroomacoustics

for general configuration presented in Fig 2, we don't care about the cleanness of the environment and even in the case of noise free and non-reverberant environments the test x-vector will be passed through the DAE. When we apply the DAE trained on noisy x-vectors on clean x-vectors, the output is still almost identical to the input. It means that for noise and reverberation free environments the system could be used without any modification.

In the case of additive noise, both specific models and general models were tested. As it is shown the results obtained by specific models are a little better. For example, for utterances longer than 12 seconds, the EER obtained by specific model is 1.91 but the EER obtained by general model is 2.23. For early reverberation and full reverberation distortions, the results obtained by general model for short segments are better than the specific model but the results obtained for longer utterances is almost the same. When there is additive noise and reverberation, in all cases the results achieved by the general model are better. For specific models the experiments are done directly without using a classifier. But to prove that it is possible to detect the type of distortion, we trained a feedforward neural network. The accuracy of the trained network is 81%. Even if we had a distortion classifier with 100% accuracy, the results show that it's better to use a general DAE instead of using a specific DAE for each distortion.

Table 1. The results obtained for different distortions (EER)

Environment	Duration (seconds)	[0,2]	[2,4]	[4,6]	[6,8]	[8,10]	[10,12]	[12,]
Dry signal (D)	D	9.89	5.08	3.62	2.44	1.77	1.85	1.65
	DAE (FN)	9.01	5.06	3.31	2.40	1.75	1.85	1.65
Additive noise (N)	N	14.58	11.16	8.83	9.12	8.05	7.97	6.63
	DAE (N)	10.48	6.08	3.56	3.27	2.26	1.91	1.91
	DAE (N+E+F+FN)	10.75	6.93	4.45	4.47	3.53	2.77	2.23
Early reverberation (E)	E	18.60	10.54	7.08	4.58	3.98	4.16	4.02
	DAE (E)	13.08	6.48	4.15	2.50	2.23	2.31	2.29
	DAE (N+E+F+FN)	11.57	5.67	3.85	2.96	2.58	2.77	2.29
Full reverberation (F)	F	20.01	11.96	8.01	6.15	4.88	5.09	5.23
	DAE (F)	9.05	5.03	3.26	2.48	2.21	2.31	2.23
	DAE (N+E+F+FN)	9.29	4.86	3.62	2.49	2.21	2.31	2.05
Additive noise and Full reverberation (FN)	FN	27.32	24.34	20.77	18.68	19.38	19.93	17.61
	DAE (FN)	14.24	9.93	7.37	4.97	3.96	3.68	4.53
	DAE (N+E+F+FN)	13.66	9.73	6.52	4.54	3.61	3.24	4.09

VI. Conclusion

In this paper we proposed two configurations for robust speaker recognition in the environments where there are several distortions. The systems act efficiently in environments with early reverberation, full reverberation, additive noise, additive noise and reverberation. To solve this problem, we proposed two configurations. In the first configuration we used a specific DAE for each distortion. In the second configuration, one DAE is used to learn all of these distortions simultaneously. The second configuration is simpler and gives the same or even better results than specific compensation for each distortion. We also showed that the speaker recognition performance doesn't decrease (with respect to the baseline) when the test data is clean, which is a nice property of the proposed denoiser.

Acknowledgement

This work has been supported by ROBOVOX ANR project.

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, 2018.

[2] J Arsha Nagrani, Joon Son Chung, Weidi Xie, Weidi Xie, Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language,* vol. 60, 2020.

[3] Ondrej Novotny, Oldrich Plchot, Pavel Matejka, Ladislav Mosner, Ondrej Glembek, "On the use of X-vectors for Robust Speaker Recognition," in *Odyssey 2018*, Les Sables d'Olonne, France, 2018.

[4] Mohammad Mohammadamini, Driss Matrouf, "Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments," in *EUSIPCO*, Amsterdam, 2020.

[5] Shai Rozenberg, Hagai Aronowitz, Ron Hoory, "Siamese x-vector reconstruction for domain adapted speaker recognition," in *INTERSPEECH*, Schenghai, China, 2020.

[6] Ondřej Novotný, Oldřich Plchot, Ondřej Glembek, Jan Honza Černocký, Lukáš Burget, "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition," *Computer Speech & Language,* no. 58, pp. 403-421, 2019.

[7] Mohammad Mohammadamini, Driss Matrouf, Paul-Gauthier Noé, "Denoising x-vectors for Robust Speaker Recognition," in *Odyssey 2020 The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020.

[8] R. Scheibler, E. Bezzam and I. Dokmanić, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 351-355, doi: 10.1109/ICASSP.2018.8461310.

[9] L. Wang, Z. Zhang, A. Kai and Y. Kishi, "Distant-talking speaker identification using a reverberation model with various artificial room impulse responses," *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, 2012, pp. 1-4.

[10] X. Zhao, Y. Wang and D. Wang, "Robust speaker identification in noisy and reverberant conditions," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[11] Sefik Emre Eskimeza, Peter Soufleris, Zhiya Duana, Wendi HeinIman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication,* vol. 99, pp. 101-113, 2018.

[12] Taherian, Hassan, Wang Zhong-Qiu, Wang, DeLiang, "Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments," in *INTERSPEECH*, 2019.

[13] Joon-Young Yang, Joon-Hyuk Chang, "Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification," in *INTERSPEECH*, 2019.

[14] Waad Ben Kheder, Driss Matrouf. Pierre-Michel Bousquet. Jean-François Bonastre, Moez Ajili, "Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition," *Computer Speech & Language,* vol. 45, pp. 104-122, 2017.

[15] Waad Ben Kheder, Matrouf Driss, Moez Ajili, Jean François Bonastre, "Probabilistic Approach Using Joint Clean and Noisy i-Vectors Modeling for Speaker Recognition," in *Interspeech*, 2016.

[16] Hu Y, Kokkinakis K. Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. *J Acoust Soc Am*. 2014;135(1):EL22-EL28. doi:10.1121/1.4834455

[17] Font, Frederic and Roma, Gerard and Serra, Xavier, "Freesound technical demo," ACMM, 2013.

[18] David Snyder, Guoguo Chen, Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015. [Online]. Available: https://www.groundai.com/project/musan-a-music-speech-and-noise-corpus/1. [Accessed 01 02 2020].