A Comprehensive Exploration of Noise Robustness and Noise Compensation in ResNet and TDNN-based Speaker Recognition Systems

Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonastre, Sandipana Dowerah, Romain Serizel, Denis Jouvet

A Comprehensive Exploration of Noise Robustness and Noise Compensation in ResNet and TDNN-based Speaker Recognition Systems

1st Mohammad MohammadAmini
LIA (Laboratoire Informatique d'Avignon)
University of Avignon
Avignon, France
mohammad.mohammadamini@univ-avignon.fr

2nd Driss Matrouf
LIA (Laboratoire Informatique d'Avignon)
University of Avignon
Avignon, France
driss.matrouf@univ-avignon.fr

3rd Jean-François Bonatsre
LIA (Laboratoire Informatique d'Avignon)
University of Avignon
Avignon, France
jean-francois.bonastre@univ-avignon.fr

4th Sandipana Dowerah
CNRS, Inria, Loria
Universite de lorraine
Nancy, France
sandipana.dowerah@loria.fr

5th Romain Serizel
CNRS, Inria, Loria
Universite de lorraine
Nancy, France
romain.serize@loria.fr

6th Denis Jouvet
CNRS, Inria, Loria
Universite de lorraine
Nancy, France
denis.jouvet@inria.fr

***Abstract*—In this paper, a comprehensive exploration of noise robustness and noise compensation of ResNet and TDNN speaker recognition systems is presented. Firstly the robustness of the TDNN and ResNet in the presence of noise, reverberation, and both distortions is explored. Our experimental results show that in all cases the ResNet system is more robust than TDNN. After that, a noise compensation task is done with denoising autoencoder (DAE) over the x-vectors extracted from both systems. We explored two scenarios: 1) compensation of artificial noise with artificial data, 2) compensation of real noise with artificial data. The second case is the most desired scenario, because it makes noise compensation affordable without having real data to train denoising techniques. The experimental results show that in the first scenario noise compensation gives significant improvement with TDNN while this improvement in Resnet is not significant. In the second scenario, we achieved 15% improvement of EER over VoiCes Eval challenge in both TDNN and ResNet systems. In most cases the performance of ResNet without compensation is superior to TDNN with noise compensation.**

***Index Terms*—Speaker recognition, ResNet, Additive noise, Reverberation, Robustness.**

I. Introduction

Speaker recognition systems authenticate the identity of the claimed users from their speech utterances. The state-of-the-art speaker recognition systems mostly use DNNs to extract a fixed-size compact representation from variable-length speech utterances known as speaker embedding or x-vector. Achieving more robust embeddings is a core task in the speaker recognition systems. Since the emergence of TDNN x-vector system until now, several speaker embedding architectures are proposed. Among them, TDNN [1], CNN [2], ResNet [3], and VGGVox [3] systems are commonly used.

The robustness of the DNN-based speaker recognition (SR) systems in general and specifically their robustness against environment variabilities such as additive noise, reverberation, and far-distance recording device has made them more promising. Several strategies such as data argumentation [4], and noise compensation [5], [6] are explored to make the TDNN-based SR systems more robust against noise and reverberation and other variabilities. The previous research shows the weakness of TDNN-based SRs against noise and reverberation distortions. In [6] it is shown that in the presence of noise and reverberation using a compensation technique before scoring (statistical or DAE) can bring the performance of x-vector system closer to clean situation.

In this paper, firstly we explore the robustness of ResNet speaker embedding system in the presence of additive noise, reverberation and both distortions. In all cases the robustness of ResNet system is compared with TDNN system. In our experiments, the impact of artificial and real noises and reverberation is examined. The second part of our work is dedicated to noise compensation in both systems. The noise compensation is done with a stacked denoising autoencoder that in our previous research tested successfully in the domain of TDNN x-vectors [4]. In the current paper, noise compensation is extended to ResNet based system and noise compensation in real environments. In our work two scenarios are explored:

- compensation of artificial noise with artificial data
- compensation of real noise with artificial data

The goal of artificial noise compensation is usually to have different experimental set up to test the performance of denoising techniques or to simulate hypothetical situation where the speaker recognition system would be used. The second scenario, doing real noise compensation with artificial data is the most demanded in the field of speaker recognition.

Because in this approach without having a real training data for denoising and just by data simulation it is possible to reduce the impact of noise and other distortions (in real situations). In our experiments we explore this situation with x-vectors extracted from both TDNN and ResNet systems. In this paper we address the difficulties of real noise compensation by using only simulated noises in training data.

In the following parts of this paper, firstly the related works are reviewed in section 2. The system configuration is described in section 3. The experiments setup is explained in section 4 and results are discussed in section 5.

II. Related works

The robustness of a speaker recognition system is treated in different parts of the system including signal processing, robust feature extraction, robust speaker modeling, compensation techniques in the speaker modeling level and adaptations in the scoring tools. In this section, the related works are reviewed. The reviewed works are in the two main categories: compensation techniques at the speaker modeling level (i.e., x-vector level or speaker embedding) and robust speaker models extractors (DNNs).

In numerous works, the researchers have tried to propose more robust speaker embedding system. In the domain of DNNs several architectures are proposed. D-vector is among the earliest speaker embedding systems that the DNN assigns speaker identity to input frames [8]. This system becomes more efficient with x-vector system that works at the segment level instead of frame level [1]. Due to the success of the x-vector TDNN system, different variations for this system are proposed. E-TDNN [10], that considers longer temporal context and FTDNN [11] that reduce the weights of each TDNN layer to the multiplication of two smaller matrices, are among two variants of the original TDNN x-vector system. ResNet is an another architecture that is is widely used for speaker modeling [9]. In [13] a combination of LSTM and ResNet is proposed. In [7] several works are reviewed that can be categorized into input features, loss function and pooling operations modifications. In aforementioned works, there is no direct treating of noise and reverberation and the general improvement of the system for both noisy and clean environments is targeted.

There are also few works that tried to handle noise at the speaker embedding level. In [14] it is shown that training the speaker embedding network with a specific noisy data doesn't have significant impact on the performance of SR system. They conclude that adding more diversity and using more data is more important. In [15] a VoiceID loss function was proposed that uses the feedback from the speaker modeling system to generate a ratio mask. In [21] an adversarial strategy was proposed to make the speaker embedding more robust against noise. In the standard x-vector extractors, after the embedding layer, a DNN speaker classifier is optimized. In this work, a second classifier is trained adversarially that accepts the type of noise in the output. In another work, a GAN based speaker embedding proposed that uses a binary discriminator to discriminate noisiness of the x-vector alongside the speaker recognition classifier [22].

In general, most of the speaker embeddings try to give more robust representation of speech utterances for both clean and noisy environments. Targeting the noise, reverberation and other variabilities at the x-vector level is another approach to make embeddings more robust against distortions. In [5] statistical i-MAP and several DAEs are used to do a transformation between noisy and clean x-vectors. In [6] two configurations are proposed in the case of having more than one distortion. In [4], it was shown that however we can make the SR systems more robust by using data augmentation techniques, but using denoisng techniques still bring us closer to the performance of clean environments. These compensation techniques are used in the TDNN x-vector system.

To the best of our knowledge, noise compensation and the behavior of acoustic noise is not explored in the ResNet speaker embedding systems in the previous work. To fill this gap we explore the robustness of ResNet against different distortions and study the effectiveness of noise compensation in the case of both artificial and real noises and reverberation.

TABLE I
The proposed ResNet-34 architecture. Last row, N is the number of speakers. The dimensions are (Frequency×Channels×Time). The input is comprised of 60 filter banks from speech segments. During training we use a fixed segment length of 400.

Layer name	Structure	Output
Input	–	$60 \times 400 \times 1$
Conv2D-1	3×3, Stride 1	$60 \times 400 \times 64$
ResNetBlock-1	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$, Stride 1	$60 \times 400 \times 128$
ResNetBlock-2	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 128 \end{bmatrix} \times 4$, Stride 2	$30 \times 200 \times 128$
ResNetBlock-3	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 256 \end{bmatrix} \times 6$, Stride 2	$15 \times 100 \times 256$
ResNetBlock-4	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 3$, Stride 2	$8 \times 50 \times 512$
Pooling	–	8×256
Flatten	–	2048
Dense1	–	256
Dense2 (Softmax)	–	N
Total	–	–

III. System Configuration

In section 3.1 the architecture of the ResNet speaker embedding network is presented. In section 3.2 the integration of compensation module with the SR system is presented.

A. ResNet and TDNN architecture

The ResNet speaker embedding used in this paper is a variant based on ResNet [12]. The ResNet model for extracting x-vectors is made of three parts: *ResNet Blocks*, a *statistics-level* layer, and *segment-level* representation layer.

The *frame-level* component is based on the well-known ResNet topology. The *statistics-level* component converts a variable length speech signal into a single fixed-size vector.

The statistics-level is composed of one layer: the statistics-pooling, that aggregate the output vectors of the DNN and computes their mean and standard deviation. The *segment-level* component maps the segment-level vector to speaker classes. The mean and standard deviation are concatenated together and they are sent to additional hidden layers and the output softmax layer with speaker labels. The ResNet is trained with ArcFace softmax loss function to classify the speakers. More details about the architecture of the Resnet system is presented in Table I.

In our experiments we used the TDNN architecture introduced in [1].

B. Compensation module

The compensation module performs a transformation between noisy and clean x-vectors. The compensation module tries to remove the impact of noise from x-vectors. We used a stacked DAE that was introduced in [4]. The Stacked DAE is composed of several DAE blocks. The noisy x-vectors fed to the first DAE. The next DAE block receives (the output of its predecessor block) concatenated with the difference between noisy x-vectors and the output of the previous block. In the hidden layers the Relu activation function is used, and the output layer is Linear. The network is trained with stochastic gradient descent.

IV. EXPERIMENTS SETUP

A. Speaker embedding training

The x-vector extractors are trained on Voxceleb2 corpus. In order to increase the diversity of the acoustic conditions in the training set, the MUSAN corpus was used for data augmentation [17]. Also, a RIR pool used for data reverberation [1]. The x-vector systems is trained on MFCC features with 25 frame length, and ResNet is trained on 60 Fbanks with 25 frame length.

B. Test and enrollment

In our experiments we used two datasets: Fabiole [23] and VoiCes [18]. The Fabiole protocol is used to evaluate the robustness of the system against simulated noise and reverberation. In the Fabiole protocol, we have 130 hundred speakers in the enrollment and 30 speaker for the test. The number of test files is 6870. In both protocols, one file is used per speaker in the enrollment. The Voices protocol is used to evaluate the robustness against real noise and reverberation. VoiCes dataset has train and test parts. The test part was created from 1320 clean files coming from Librispeech (100 speakers) and the train part is recorded from 2583 files coming from Librispeech (200 speakers). We used 300 files, each file belonging to one speaker for enrollment and 3603 remained files are used as test utterances. In all experiments with VoiCes the far microphone (mic 05) and the rooms (room2, room3) with more reverberation are chosen. The details of protocols are presented in Table II.

TABLE II
THE BASELINE SYSTEM.

Protocol	Test	Enroll	Trials
Fabiole	6870	130	893k
Voices	3603	300	1080k

TABLE III
ROBUSTNESS AGAINST DIFFERENT DISTORTIONS (EER)

Distortion	Protocol	ResNet	TDNN
Clean	Fabiole	6.27	15.21
	Voices	0.89	1.25
Noise	Fabiole [SNR 0-5]	8.28	17.83
	Fabiole [SNR 5-10]	7.43	16.58
	Fabiole [SNR 10-15]	6.87	15.95
Reververation	Fabiole	9.75	18.20
	Voices room 2	1.24	2.53
	Voices room 3	2.6	6.68
Reververation and Noise	Fabiole [SNR 0-5]	12.48	21.47
	Voices room 2	1.24	3.71
	Voices room 3	2.6	6.69

C. Back-end

The TDNN system is evaluated with PLDA back-end. The PLDA is trained with 200k utterances extracted from Voxceleb. The ResNet is evaluated with cosine distance.

V. RESULTS AND DISCUSSION

A. Exploring the Robustness of TDNN and ResNet

In this section the results are presented. In all experiments the ResNet is compared with TDNN x-vector system. The results for different situations are presented in Table III.

Baseline. In the baseline experiment, there is no noise and reverberation.

Additivie noise. Because there is no dataset with just additive noise, the systems are evaluated with simulated additive noise. The experiments are done with Fabiole prtocol. The Freesound [24] noises are added to the clean speech with Pyacoustics [1] tool. In three experiments, different SNRs are tested.

Reverberation. In another experiment, we explored the robustness of both systems against reverberation. In this case we tested the systems with both real and simulated reverberation. The protocol of adding reverberation is described in [6]. For real reverberation, we used recorded files of room2 and room3 in Voices that are recorded without noise.

Additive noise and reverberation. In this case, the systems are tested with both real and simulated noise and reverberation.

If we compare the baseline results in Table III with the presence of distortions, we see that the ResNet has relative robustness in the presence of noise and reverberation. For example, in the worst case for SNR between 0 and 5 and reverberation on Fabiole protocol, the EER is 12.48 while with TDNN in the clean environment the EER is 15.21. With

[1] https://github.com/timmahrt/pyAcoustics

TABLE IV
USING SIMULATED NOISES FOR NOISE COMPENSATION (EER)

System	Clean	Noisy	Denoised
TDNN	15.21	21.47	18.00
Resnet	6.27	12.48	12.18

VoiCes protocol in the presence of noise and reverberation in room 3 the EER is 2.6 while for TDNN system the EER is 6.69. Just in the case of noisy and reverberation for Fabiole protocol the performance of ResNet degrades significantly. One possible reason behind this degradation in Fabiole comes from this fact that in Fabiole there are 1720 files shorter than 4 seconds. In all other experiments ResNet shows a relative robustness against noise and reverberation. For example, in the VoiCes protocol in the clean position the EER is 0.89 but in the presence of severe noise and reverberation it is 2.6.

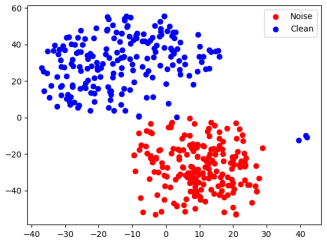
B. Noise compensation

In the second group of experiments we did noise compensation in both TDNN and ResNet system in the presence artificial and/or real noises and reverberation. During noise compensation, two scenarios are considered. The results for each scenario are described in this subsection.

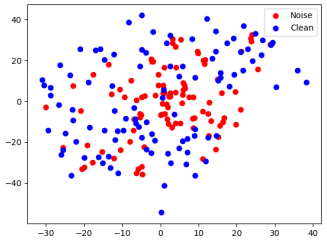
1) Artificial noise compensation with artificial training data: In our experiments, we used pairs of noisy/x-vectors to train DAE. The training pairs are constructed from Voxceleb. The noisy version is prepared by adding Freesound noises and RIR files with Pyacoustics. In the training data, there are about 5 million pairs of noisy/clean x-vectors. In the noisy version, there are one or both additive noise and reverberation distortions. The noises and RIR files used to prepare the training data are different from those that are used for test protocols.

After doing transformation on noisy test files with the trained DAE, we observed a small gain in terms of EER for ResNet system. For example in the presence of additive noise and reverberation in Fabiole protocol the EER reduce from 12.48 to 12.18, while in TDNN it reduce from 21.74 to 18.03. The results are shown in Table IV.

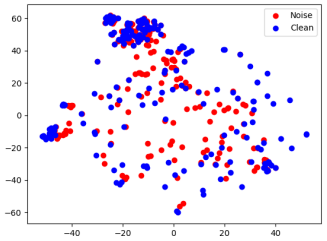
To have a profound interpretation of this phenomena we did a visualisation of noisy and clean x-vectors with t-SNE. The visualisation shows that ResNet x-vectors remain in the same space and the noise and reverberation doesn't have a big impact on them. In this experiment we chose a random noisy x-vector and its 1000 closest neighbors. The chosen vectors are plotted alongside their correspondent clean version. The t-SNE is trained with the both clean and noisy x-vectors. This experiment shows that noisy x-vectors in ResNet system are not separable and far from their clean version. But in TDNN system, there is a significant shift between noisy and clean x-vectors. This phenomena explains that there is no big difference between noisy and clean versions of x-vectors to be compensated by denoisers. This is in conformity with obtained EER in noisy and reverbant environment leveraging ResNet system (Fig. 1). However, the small residual noise in Resnet is not trainable with DAEs, we don't know whether we have arrived to the limit of doing noise compensation in this system or if it is possible to do noise compensation in ResNet x-vectors.



(a) TDNN noisy



(a) TDNN denoised



(b) ResNet

Fig. 1. t-SNE visualisation of TDNN and ResNet

2) Real noise compensation with artificial training data: In order to train the DAE for real noise and reverberation compensation, the same training data used that was prepared already for artificial noise compensation. In this experiment the standard VoiCes protocols introduced in [18] are used. The results before and after noise compensation are shown in table V. Intuitively, having a simulated training data that matches better with the real noisy test data, gives better results. During data simulation, we tried to prepare another training data by fine-tuning several parameters such as room size, sound-absorption, and microphone distance. We observed that creating a training simulated data by fixing these parameters doesn't bring more improvement. The experiments show that having more diversified data with different parameters such as

TABLE V
REAL NOISE COMPENSATION WITH ARTIFICIAL NOISY TRAINING DATA (EER)

	Voices Eval		Voices Dev	
System	**Noisy**	**Denoised**	**Noisy**	**Denoised**
TDNN	4.44	3.80	7.89	7.28
Resnet	1.37	1.15	5.10	5.04

random microphone distance and having different room sizes increases the chance of capturing the given noise in the test situation.

VI. CONCLUSION

In this paper, we explored the noise robustness of two state-of-the-art speaker recognition systems, TDNN and ResNet. We have shown through our experiments that the system based on ResNet is much more robust to noise (additive noise and reverberation) than the TDNN. Also, real and artificial noise compensation is done in both systems. The most unexpected result is that the compensation techniques (based on DAE) gives a marginal improvement in the case of artificial noises with ResNet, while the improvement is significant in TDNN system. Despite this finding, the ResNet system remains more efficient than the TDNN, with or without noise, with or without compensation. However we found a degree of improvement in the case of real noise compensation in both TDNN and ResNet systems, we had shown that a precise simulation of real situation is the main challenge of doing real noise compensation with artificial training data. The objective of future work is handling noise and reverberation in speaker embedding level in order to avoid the limitations of noise compensation in x-vector space.

REFERENCES

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[2] D. Cai, Z. Cai and M. Li, "Deep Speaker Embeddings with Convolutional Neural Network on Supervector for Text-Independent Speaker Recognition," 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018

[3] Arsha Nagrani and Joon Son Chung and Weidi Xie and Andrew Zisserman, Voxceleb: Large-scale speaker verification in the wild, omputer Speech Language, 2020.

[4] Mohammad MohammadAmini and Driss Matrouf, "Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments," 2020 28th European Signal Processing Conference (EUSIPCO), 2021, pp. 1-5, doi: 10.23919/Eusipco47968.2020.9287690.

[5] Mohammad Mohammadamini, Driss Matrouf, Paul-Gauthier Noé. (2020) Denoising x-vectors for Robust Speaker Recognition. Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 75-80, DOI: 10.21437/Odyssey.2020-11

[6] Mohammad Mohammadamini, Driss Matrouf, Jean-Francois Bonastre, Romain Serizel, Sandipana Dowerah, and Denis Jouvet, "Compensate multiple distortions for speaker recognition systems," in 2021 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 141–145

[7] Zhongxin Bai, Xiao-Lei Zhang,, Speaker recognition based on deep learning: An overview, Neural Networks, Volume 140,, 2021,, Pages 65-99,, ISSN 0893-6080, , https://doi.org/10.1016/j.neunet.2021.03.004.

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependentspeaker verification, in: 2014 IEEE International Conference on Acous-tics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4052–4056.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recog-nition, in: Proceedings of the IEEE conference on computer vision andpattern recognition, 2016, pp. 770–778

[10] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, S. Khudanpur, Speaker recognition for multi-speaker conversations using x-vectors, in: ICASSP 2019-2019 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP), IEEE, 2019.

[11] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, S. Khu-danpur, Semi-orthogonal low-rank matrix factorization for deep neuralnetworks., in: Interspeech, 2018, pp. 3743–3747.

[12] Zeinali, H., Wang, S., Silnova, A., Matejka, P., Plchot, O.: BUT system descriptionto voxceleb speaker recognition challenge 2019. CoRRabs/1910.12592(2019),http://arxiv.org/abs/1910.12592

[13] J. Zhou, T. Jiang, Z. Li, L. Li, Q. Hong, Deep speaker embedding extraction with channel-wise feature responses and additive supervision-softmax loss function, Proc. Interspeech 2019 (2019) 2883–2887

[14] Minh Pham, Zeqian Li, and Jacob Whitehill, How Does Label Noise Affect the Quality of Speaker Embeddings?, INTERSPEECH 2020.

[15] Suwon Shon, Hao Tang, James Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in INTERSPEECH 2019, Graz, Austria, 2019

[16] Waad Ben Kheder, Driss Matrouf. Pierre-Michel Bousquet. Jean-François Bonastre, Moez Ajili, "Fast 79 i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition," Computer Speech Language, vol. 45, pp. 104-122, 2017.

[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "MUSAN: A Music, Speech, and Noise Corpus," 28 10 2015. [Online]. Available: https://arxiv.org/abs/1510.08484.

[18] Colleen Richey, Maria A.Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, Karl Ni, Voices Obscured in Complex Environmental Settings (VOICES) corpus, 2018, arXiv:1804.05053.

[19] BBC, "BBC," BBC, [Online]. Available: http://bbcsfx.acropolis.org.uk/

[20] Brignatz V., Duret J., Matrouf D., Rouvier M. (2021) Language Adaptation for Speaker Recognition Systems Using Contrastive Learning. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham.

[21] J. Zhou, T. Jiang, Q. Hong and L. Li, "Extraction of Noise-Robust Speaker Embedding Based on Generative Adversarial Networks," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1641-1645, doi: 10.1109/APSIPAASC47483.2019.9023295.

[22] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang and B. Xia, "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6196-6200, doi: 10.1109/ICASSP.2019.8683828.

[23] Moez Ajili, Jean-François Bonastre, Juliette Kahn, Solange Rossato, and Guillaume Bernard. 2016. FABIOLE, a Speech Database for Forensic Speaker Comparison. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 726–733, Portorož, Slovenia. European Language Resources Association (ELRA).

[24] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In Proceedings of the 21st ACM international conference on Multimedia (MM '13). Association for Computing Machinery, New York, NY, USA, 411–412. DOI:https://doi.org/10.1145/2502081.2502245