Evaluation Framework for Context-aware Speaker Recognition in Noisy Smart Living Environments

Gianni Fenu, Roberta Galici and Mirko Marras

Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, 09124 Cagliari, Italy

Abstract

The integration of voice control into connected devices is expected to improve the efficiency and comfort of our daily lives. However, the underlying biometric systems often impose constraints on the individual or the environment during interaction (e.g., quiet surroundings). Such constraints have to be surmounted in order to seamlessly recognize individuals. In this paper, we propose an evaluation framework for speaker recognition in noisy smart living environments. To this end, we designed a taxonomy of sounds (e.g., home-related, mechanical) that characterize representative indoor and outdoor environments where speaker recognition is adopted. Then, we devised an approach for off-line simulation of challenging noisy conditions in vocal audios originally collected under controlled environments, by leveraging our taxonomy. Our approach adds a (combination of) sound(s) belonging to the target environment into the current vocal example. Experiments on a large-scale public dataset and two state-of-the-art speaker recognition models show that adding certain background sounds to clean vocal audio leads to a substantial deterioration of recognition performance. In several noisy settings, our findings reveal that a speaker recognition model might end up to make unreliable decisions. Our framework is intended to help system designers evaluate performance deterioration and develop speaker recognition models more robust to smart living environments.

Keywords

Deep Learning, Security, Speaker Recognition, Speaker Verification, Noisy Environments, Sound Taxonomy.

1. Introduction

Speech is a more natural way of interacting with devices than tapping screens. This form of interaction is receiving more and more attention, with voice-enabled services being used in every aspect of our lives. Speaker recognition analyzes the identity of an individual before accessing to a service. Unlike speech recognition, which detects spoken words, speaker recognition inspects patterns that distinguish one person's voice from another [1]. Recognizing the identity of a speaker becomes crucial in different scenarios. For instance, voice-enabled devices (e.g., assistants, smartphones) allow home owners to turn on lights, unlock doors, and listen to music seamlessly [2]. These recognition abilities can prevent unauthorized individuals from using devices without the owner's permission and can provide evidence needed to personalize user's experiences with these devises, even outside the domestic borders [3, 4, 5]. Moreover, speaker recognition can make lives of older adults' and people with special needs easier and safer [6]. Hence, it is imperative to study and devise data-driven speaker recognition models that can improve human quality of life.

State-of-the-art speaker recognition matchers exhibit impressive accuracy, especially when the voice quality is reasonably good [7]. For this reason, they implicitly or explicitly impose constraints on the environment, such as being stationary and quiet. Conventionally, speaker matchers are trained to classify vocal examples under idealistic conditions but are expected to operate well in real-world challenging situations. However, their performance sharply degrades when audios with substantial background sounds (e.g., traffic) are encountered. Enhancing voice data is demanding since related algorithms do not often explicitly attempt to preserve biometric cues in the data [8, 9, 10]. Existing robust speaker models are being trained on data which do not cover various levels of interfering sounds and different sound types [11, 12]. Hence, several questions concerning how much and under which background sounds speaker recognition performance degrades and how each type of sound impacts on the mechanics of these matchers remain unanswered.

Our study in this paper is hence organized around these directions and aims to perform an extensive performance analysis of deep speaker recognition matchers in a range of noisy living environments. To this end, we designed and collected a taxonomy of sounds (e.g., footsteps, laughing) that characterize representative living ambients where speaker recognition is finding adoption. Then, we depicted an approach that allows us to simulate challenging noisy conditions in

Proceedings of the CIKM 2020 Workshops, October 19-20, Galway, Ireland. Stefan Conrad, Ilaria Tiddi (Eds.)

EMAIL: fenu@unica.it (G. Fenu); r.galici1@studenti.unica.it (R. Galici); mirko.marras@unica.it (M. Marras)
ORCID: 0000-0003-4668-2476 (G. Fenu); 0000-0003-1989-6057 (M. Marras)

raw vocal audios by adding sounds of our taxonomy, according with the environment under consideration. Finally, we experimented with a public dataset, originally collected in controlled environments, and two state-of-the-art speaker recognition models, to inspect the impact of background noisy sounds on their performance. Our contribution is threefold:

- We design a taxonomy of ambient sounds tailored to speaker recognition research, and we provide a dataset of recordings with labeled sound sources for each category in our taxonomy.

- We propose an evaluation framework for speaker recognition benchmarking, enabling easier and faster simulation of indoor and outdoor noisy environments in (clean) vocal audios[1].

- Given a large vocal dataset, we perform an extensive analysis on the impact of the sounds in our taxonomy on the performance of two state-of-the-art speaker recognition matchers.

Our experiments showed that, even when the background sound volume is low, speaker recognition systems undergo a substantial deterioration of accuracy. Only in case of nature-related sounds (e.g., chirping, wind), the sound impact is negligible. Certain environmental settings lead to error rates five to ten times higher than error rates achieved in ideal conditions.

The rest of this paper is organized as follows. Section 2 depicts an overview of related works. Then, our taxonomy and the simulation framework are described in Section 3. Section 4 presents our experiments. Finally, Section 5 provides insights for future work.

2. Related Work

Our research lies at the intersection among three perspectives, namely studies which analyze the impact of background sounds on recognition, audio enhancement algorithms aimed to improve data quality, and speaker recognition approaches which seek to classify noisy vocal data with no pre-processing.

2.1. Explorative Analysis in Noisy Environments

Explorative analyses simply investigate how noisy environments influence speaker recognition performance. For instance, Qian et al. [13] studied the low-level noisy optimization task by means of evolutionary algorithms. The authors found that a *Bitwise* noise can fundamentally affect recognition patterns during evaluation and, thus, might make it harder to deploy matchers in the real world. Differently, Ko et al. [14] focused on a performance comparison between acoustic models trained with and without simulated far-field speech under a real far-field voice dataset. Their experiments showed that acoustic models trained on simulated far-field led to significantly lower error rates in both a distant- and close-talking scenarios. In [15], the authors presented a feature learning approach, referred as to *e-vector*, that can capture both channel and environment variability. Recently, Vincent et al. [16] analyzed the performance of speaker recognition matchers on the *CHiME3* dataset, which consists of real recordings in noisy environments. Finally, Donahue et al. [17] analyzed the benefits resulting from training a speaker recognition matcher with both clean speech data and fake speech data created by means of a generative adversarial network.

Though this research has greatly expanded our understanding, past works focused on low-level noises (e.g., Bitwise) or did not specifically control how and under which ambient sounds the performance degrades. We argue that different background sounds may lead to fundamentally different impacts and, thus, a clear understanding on the extent of this impact is lacking.

2.2. Input Audio Quality Enhancement

Existing literature included audio enhancement algorithms that aim to provide audible improvements in a sound, without degrading the quality of the original recording. This type of strategy well fits with the forensic context, where audios may have some kind of background sound disturbance or sound artifact that may interfere with the voice of interest. Examples of audio enhancement methods are removing static noise, eliminating phone-related interference, and clearing up random sounds (e.g., dogs barking, bells ringing). For instance, Hou et al. [8] proposed a convolution-based audio-visual auto-encoder for speech enhancement through multi-task learning. In [9], the authors investigated how to improve speech/non-speech detection robustness in very noisy environments, including stationary noise and short high-energy noise. Similarly, Afouras et al. [10] proposed an audio-visual neural network able to isolate a speaker's voice of interest from other sound interference, using visual information from the target speaker's lips. However, the designed methods do not often attempt to preserve biometric cues in the data and depend on the nature of the sound, which varies according to the context. Hence,

[1] Code, data, pre-trained models, and documentation are publicly available at https://mirkomarras.github.io/dl-voice-noise/.

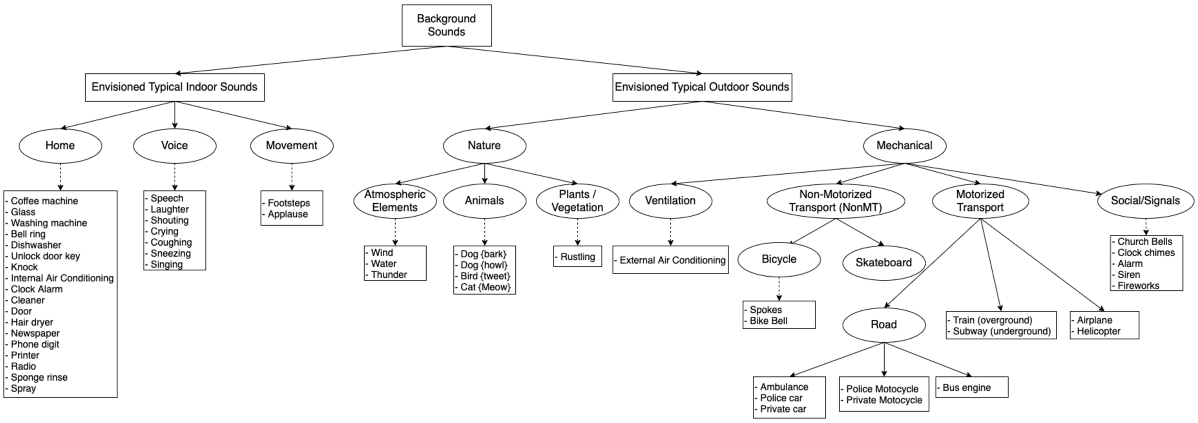

Figure 1: Our taxonomy of sounds characterizing representative environments where speaker recognition is adopted.

our framework becomes a key asset to study recognition performance on background sounds against which countermeasures have been under-explored.

2.3. Robust Speaker Recognition

Speaker recognition matchers have traditionally relied on Gaussian mixture models [18], joint factor analysis [19], and IVectors [20]. Recently, speaker matchers achieved impressive accuracy thanks to speaker embedding representations extracted from Deep Neural Networks trained for (one-shot) speaker classification. Notable examples include CVectors [21], XVectors [11], VGGVox- and ResNet-Vectors [12]. Moreover, Kim et al. [22] proposed a deep noise-adaptation approach that dynamically adapts itself to the operational environment. Existing approaches in this area do not make any assumption on the training and testing data, which come from various noisy situations. Therefore, there is no fine-grained control on how these systems perform in specific noisy applicative scenarios and the noisy situations are limited by the variety of recordings including into the considered dataset.

3. The Proposed Framework

Our framework is composed by a *dataset* of sounds categorized according to our pre-defined taxonomy and a *toolbox* that simulates background living scenarios.

3.1. Sound Taxonomy for Speaker Recognition in Living Scenarios

Collecting utterances produced in various living scenarios while keeping trace of the background sounds in each utterance is challenging and time-consuming. Hence, being able to combine utterances recorded in quiet environments with the background sounds of the considered scenario represents a viable alternative. The first step to put this idea into practice consists of collecting sounds from a wide range of sources and organizing them in a hierarchical taxonomy.

Noisy sounds research is challenging due to the lack of labeled audio data. Past works collected sounds from specific environments and resulted in commercial or private datasets. Recent contributions have provided publicly available datasets of environmental recordings [23]. On top of these collections, many studies have been carried out on sound classification [24]. Being designed for sound classification tasks, existing taxonomies can not directly be applied nor combined to simulate scenarios of speaker recognition. For instance, they often include a few classes, sounds of marginal interest (e.g., gun shots), and are organized according to the sound type. Conversely, for our purposes, a taxonomy should be designed with situational and contextual elements in mind (e.g., grouping sounds based on the ambient where they frequently appear).

To address these issues, we proposed a compilation of environmental sounds from over 50 classes. The selected sound clips were constructed from recordings available on the above-mentioned urban sound taxonomies and on *Freesound*[2], with a semi-automated pipeline. Specifically, we first identified a representative set of scenarios/environments where speaker recognition is actively used nowadays, and then we filtered out the categories of sounds included into existing taxonomies that are of marginal interest for the selected scenarios (e.g., fire engine). Then, we introduced new sound categories that help to model speaker recognition scenarios whose sounds are not present in ex-

[2] https://freesound.org/

isting sound taxonomies (e.g., dishwasher, footsteps). The included classes were arbitrarily selected with the goal of maintaining balance between major types of sounds characterizing the selected scenarios and of considering the limitations in the number and diversity of available sound recordings. Freesound was queried for common terms related to the considered scenarios. Search results were verified by annotating fragments that contain events associated with a given scenario. Sounds are grouped in two major categories pertaining to indoor and outdoor contexts[3]:

- **Indoor** category including sounds divided into three different categories: `Home` (e.g., TV, washing machine), `Voice` (e.g., chatting, laughing) and `Movement` (e.g., footsteps, applause).
- **Outdoor** category with sounds divided into two categories: `Nature` contains different types of sounds, such as atmospheric elements (e.g., rain, wind), animal sounds (e.g., dogs, cats, birds), and sounds associated to plants and vegetation (e.g., leaves). `Mechanical` includes sounds produced by ventilation, motorized transports (e.g, cars, trains), non-motorized transports (e.g., bicycles), and other signals (e.g., church bells).

The collected audios were converted to a unified format (16 kHz, mono, wav) to facilitate their processing with existing audio-programming packages. These sounds were arranged into the taxonomy in Figure 1 based on the above-mentioned considerations.

3.2. Toolbox for Background Living Scenario Simulation

Our taxonomy is proposed to facilitate the simulation of real-world applicative contexts in vocal audio. Thus, on top of this taxonomy, a way to combine vocal audio and background sounds is needed. To this end, we propose a Python toolbox that can simulate an applicative context into a vocal audio. Specifically, we define an *applicative context* has a set of one or more sound entries taken from the taxonomy. Each entry includes a string identifier associated to the sound category to include (e.g., `Home` or `Voice`), a floating-point number that specifies the volume level of that sound in the current context, and a floating-point number in [0,1] representing the probability of adding that sound into a vocal example. Given a context defined as above and a list of vocal audios where that context should be simulated, a routine changes each vocal audio by adding to it the combination of sounds included into the context definition, with their given volume and probability. For each sound category, the sound to add can be specified or randomly chosen. Our toolbox and our definition of context provide the necessary level of flexibility to simulate real-world scenarios created from all the combinations of the taxonomy's sounds.

4. Experiments

In this section, we assess how much and under which background sounds speaker recognition performance degrades, how each background sound type impacts on model mechanics, and how much the volume level of the background sounds leads to models which provide less accurate predictions. In fact, how each noisy context influences the behavior of state-of-the-art architectures, such as VGGVox and XVector, still remains under-explored, since their effectiveness has been often evaluated under ideal conditions, with background-sound unlabelled vocal audios, or the same type of vocal audios (e.g, from interviews).

4.1. Seed Human Voice Dataset

Given its large scale and its wide adoption in the literature, we simulated applicative contexts into the vocal data belonging to the *VoxCeleb-1* dataset [12]. This collection consists of short utterances taken from video interviews published on Youtube, including speakers from a wide range of different ethnicities, accents, professions and ages, fairly balanced with respect to their gender (i.e., 55% of males). The dataset is split into development and test sets having disjoint speakers. The development set has $1,211$ speakers and $143,768$ utterances, while the test set consists of 40 speakers and $4,874$ utterances. Our study leveraged trial pairs provided by the authors together with the *VoxCeleb-1* data[4].

4.2. Benchmarked Models

Our analysis benchmarks two state-of-the-art speaker recognition architectures: VGGVox [12] and XVector [11]. They have received great attention in recent years, and this motivated us to deepen their robustness in noisy

[3] Our preliminary analysis considers two disjoint sets of indoor and outdoor sounds, leaving settings that cross-link sound entities within a graph-based taxonomy as a future work.

[4] Due to the large amount of comparisons to simulate all contexts, our study focused on $1,000$ out of $37,702$ VoxCeleb-1 trial pairs and leaves as a future work the extension to the larger VoxCeleb-2. Here, we are more interested in understanding matchers robustness against background sounds, so the accuracy gains with larger datasets would not substantially affect the findings of our analysis.

environments. VGGVox is based on the *VGG-M Convolutional Neural Network*, with modifications to adapt to the audio spectrogram input. The last fully-connected layer is replaced by two layers, a fully connected layer with support in the frequency domain and an average pooling layer with support on the time domain. XVector is a *Time Delay Neural Network*, which allows neurons to receive signals spanning across multiple frames. Given a filterbank, the first five layers operate on speech frames, with a small temporal context centered at the current frame. Then, a pooling layer aggregates frame-level outputs and computes mean and standard deviation. Finally, two fully-connected layers aggregate statistics across the time dimension.

4.3. Model Training and Testing Details

The code, implemented in Python, ran on a NVIDIA GPU. The audios were converted to single-channel, 16-bit streams at a 16kHz sampling rate. We used 512-point Fast Fourier Transforms. VGGVox received spectrograms of size $512x300$, while XVector received filterbanks of size $300x24$. Both representations were generated in a sliding window fashion using the hamming window of width $25ms$ and step $10ms$, and normalized by subtracting the mean and dividing by the standard deviation of all frequency components. Each model was trained for classification on *VoxCeleb-1* dev set using *Softmax*, with batches of size 64. To keep consistency with respect to the original implementations of VGGVox and XVector, we used the *Adam* optimizer, with an initial learning rate of 0.001, decreased by a factor of 10 every 10 epochs, until convergence. For testing, we considered speaker embeddings of size 512. The choice of the architectural parameters was driven by the original model implementations, without any specific adaptation, given that we are interested in benchmarking the original models in noisy environments rather than tuning/arranging the parameters to align with our goals.

4.4. Speaker Recognition Protocols

Given a pretrained model, a set of trial verification pairs, and a target context, the protocol worked as follows. For each trial pair in the set, we assumed that the first audio represented the enrolled utterance ideally collected in controlled environments, while the second audio was the probe provided in the target context. Hence, the first audio remained unchanged. The second audio was changed by adding sounds that characterize the target context, as explained in Section 3.2. For both the enrolled and the changed audios, the acoustic representations were extracted and fed into that pretrained model to get the speaker embeddings. The *Cosine* similarity between the speaker embeddings was calculated. This process was repeated for each trial pair in the set. Finally, given the resulting similarity scores and the verification labels (i.e., 0 for different-user pairs, 1 for same-user pairs), the Equal Error Rate (EER) under that context was computed. The entire protocol was repeated with different background sound volume levels, treated as percentages, assumed to be in $[0, 0.05, 0.10, 0.20, 0.30, 0.50, 1, 1.5]$ (e.g., 1 means that the original volume is kept, 0.5 means that the volume of the background sound is reduced by 50%).

This protocol was carried out on 25 contexts composed by either single categories of the third and fourth levels of our taxonomy and their combination.

4.5. Experimental Results

Given the considered taxonomy contexts, the *VoxCeleb-1* trial pairs, and two pre-trained speaker recognition models, we followed the protocol in Section 4.4 to calculate the EERs at various background sound volumes.

Indoor. Table 1 and 2 report the EERs under various combinations of indoor sounds. It can be observed that `Voice` is the individual sound category that leads to the highest degradation in performance, with $27 - 30\%$ of EER at the 1.5 volume ratio. `Home` and `Movement` showed a similar impact when the volume level was below 1.0, while the former brought more negative effects with volume ratios higher than 1.0. When two or more sound categories were combined, EERs easily turned out to values above 15%, especially in scenarios where both `Home` and `Voice` sounds were present. XVector's performance substantially decreased as soon as sounds were added, while VGGVox showed a more robust behavior against background sounds. It might be possible that the changes introduced by the background sound in the spectrograms fed into VGGVox had a lower influence on the recognition pattern learnt by the convolutional layers during training. On the other hand, with XVector, the temporal context employed at each layer of the network might be highly influenced by the changes introduced into the filterbanks through the background sound addition.

Nature Outdoor. Table 3 and 4 report the EERs obtained when nature-related sounds were added. These sounds showed different degradation patterns from each other, compared with indoor-related sounds. It can be observed that models were robust against `Plants Vegetation` at any volume. Conversely, sounds coming from the `AtmosphericElements` category led to

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
Home	4.60	5.80	7.80	10.00	12.50	19.30	26.40
Movement	5.00	6.40	7.90	12.60	14.20	16.70	20.50
Voice	8.10	12.20	17.30	18.20	19.40	24.20	27.90
Home-Movement	5.00	6.50	13.00	15.90	23.20	30.00	32.20
Home-Voice	10.10	14.00	20.00	23.10	26.00	34.60	39.10
Movement-Voice	11.00	15.70	14.60	17.00	18.80	23.90	29.40
Home-Movement-Voice	**14.00**	**17.00**	**22.90**	**26.90**	**31.80**	**39.30**	**42.20**

Table 1

VGGVox - Indoor Scenario. EERs achieved by VGGVox under an Indoor Scenario. Bold values show the highest EER at each volume level. VGGVox led to an EER of *2.20%* when no sounds were added to the vocal file.

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
Home	10.49	13.90	19.59	20.99	26.30	32.19	36.40
Movement	12.60	17.80	23.70	26.00	30.70	28.50	30.30
Voice	15.40	18.10	20.59	24.20	25.00	32.49	30.60
Home-Movement	13.50	**30.30**	**36.19**	**36.70**	**39.60**	37.60	**49.09**
Home-Voice	17.10	24.09	26.90	29.70	35.60	41.10	40.60
Movement-Voice	**21.79**	27.90	33.59	34.19	37.90	36.80	39.80
Home-Movement-Voice	21.59	**30.30**	**36.19**	**36.70**	**39.60**	**44.20**	46.09

Table 2

XVector - Indoor Scenario. EERs achieved by XVector under an Indoor Scenario. Bold values show the highest EER at each volume level. XVector led to an EER of *6.35%* when no sounds were added to the vocal files.

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
Animals	4.40	5.70	8.20	10.60	15.50	22.70	32.50
AtmosphericElements	5.80	8.70	14.30	18.10	25.70	30.80	40.00
PlantsVegetation	3.10	3.00	3.00	3.50	3.90	6.60	8.30
Animals-AtmosphericElements	**7.50**	11.20	**18.60**	**24.60**	31.00	**41.30**	**48.60**
Animals-PlantsVegetation	3.70	6.00	8.40	11.90	16.60	27.40	35.00
AtmosphericElements-PlantsVegetation	5.60	9.00	15.00	17.80	24.20	31.90	41.40
Animals-AtmosphericElements-PlantsVegetation	7.00	**11.60**	18.30	23.80	**31.80**	38.60	43.70

Table 3

VGGVox - Nature Outdoor Scenario. EERs achieved by VGGVox under a Nature Outdoor Scenario. Bold values show the highest EER at each volume level. VGGVox led to *2.20%* of EER when no sounds were added to the vocal files.

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
Animals	7.09	10.39	15.60	16.40	20.90	28.20	30.80
AtmosphericElements	13.20	19.29	28.90	31.59	34.50	42.00	43.90
PlantsVegetation	5.50	5.50	7.79	8.20	9.89	12.90	17.30
Animals-AtmosphericElements	**17.00**	22.19	30.79	**35.90**	38.60	43.20	44.69
Animals-PlantsVegetation	9.29	9.79	17.10	19.70	24.50	31.30	35.00
AtmosphericElements-PlantsVegetation	14.10	21.69	26.00	32.69	35.60	43.80	45.69
Animals-AtmosphericElements-PlantsVegetation	15.20	**22.79**	**31.70**	35.09	**38.90**	**46.19**	**47.59**

Table 4

XVector - Nature Outdoor Scenario. EERs achieved by XVector under a Nature Outdoor Scenario. Bold values show the highest EER at each volume level. XVector led to an EER of *6.35%* when no sounds were added to the audio files.

the worst EERs, with 40 – 43% of EER at the highest volume level. Models suffered from the combination of `Animals` and `AtmosphericElements` sounds (44–48% of EER reached at a volume ratio of 1.5). Compared with indoor scenarios, both VGGVox and XVector showed here similar degradation patterns. This behavior might be justified by the intrinsic properties and characteristics of the nature sounds included into our taxonomy, which are shorter and less deafening.

Mechanical Outdoor. Table 5 and 6 show us that mechanical outdoor sounds led to substantial negative impacts on the model performance, except in case of `SocialSignals` sounds, compared with indoor and

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
MotorizedTransport	3.10	4.50	8.70	10.80	16.80	26.10	35.10
NonMT	28.00	27.00	28.40	28.90	25.70	30.10	30.60
SocialSignals	7.00	7.50	9.40	11.00	11.60	16.10	20.20
Ventilation	20.30	20.10	20.90	22.20	25.70	29.80	32.30
MotorizedTransport-NonMT	26.60	27.40	29.70	32.70	31.70	39.30	44.90
MotorizedTransport-SocialSignals	8.10	9.20	14.00	16.70	22.60	32.10	38.00
MotorizedTransport-Ventilation	20.60	22.70	22.70	25.60	30.70	37.40	44.10
NonMT-SocialSignals	30.20	30.10	29.40	28.60	30.00	36.00	37.80
NonMT-Ventilation	35.60	**38.00**	36.10	39.00	40.50	**41.60**	43.60
SocialSignals-Ventilation	21.40	19.90	25.20	27.00	29.90	35.50	40.20
MotorizedTransport-NonMT-SocialSignals-Ventilation	**37.60**	36.30	**38.50**	**38.30**	**42.40**	10.00	**48.50**

Table 5
VGGVox - Mechanical Outdoor Scenario. EERs achieved by VGGVox under a Mechanical Outdoor Scenario. Bold values show the highest EER at each volume level. VGGVox led to an EER of *2.20%* when no sounds were added to the audio files.

Sound Combination	Volume						
	0.05	0.10	0.20	0.30	0.50	1.00	1.50
MotorizedTransport	9.30	13.30	20.99	27.20	32.49	40.00	42.10
NonMT	47.50	48.00	**53.50**	46.90	**53.30**	50.60	44.39
SocialSignals	10.20	8.69	13.80	14.50	19.40	24.70	29.90
Ventilation	33.59	32.30	34.30	34.80	39.70	44.90	49.80
MotorizedTransport-NonMT	49.80	48.69	45.30	51.70	47.59	48.40	50.40
MotorizedTransport-SocialSignals	11.79	19.49	22.49	29.50	35.80	43.99	45.69
MotorizedTransport-Ventilation	31.79	32.49	36.90	42.00	43.30	49.10	47.50
NonMT-SocialSignals	49.90	50.50	51.60	49.40	49.20	49.40	48.00
NonMT-Ventilation	**52.40**	49.50	50.10	**49.90**	48.30	**51.30**	**51.60**
SocialSignals-Ventilation	39.70	38.70	38.40	42.20	40.50	44.69	50.30
MotorizedTransport-NonMT-SocialSignals-Ventilation	50.40	**49.90**	51.50	48.40	49.40	49.70	49.70

Table 6
XVector - Mechanical Outdoor Scenario. EERs achieved by XVector under a Mechanical Outdoor Scenario. Bold values show the highest EER at each volume level. XVector led to an EER of *6.35%* when no sounds were added to the audio files.

nature outdoor settings. Even at low volume levels, it is important to notice that `Ventilation` sounds caused substantial degradation in EERs, and this effect was amplified when two or more sound categories were combined. Among the most degraded settings, combining `NonMT` and `Ventilation` led to EERs of 35 – 50% even at a volume ratio of 0.1. Outdoor sounds coming from the `Nature` and `Mechanical` categories seemed to lead to more overlapping decision boundaries than indoor sounds. For instance, while being composed by a different combination of sound types, both the `MotorizedTransport-NonMT` and `NonMT-SocialSignals` settings showed similar EERs at a volume level higher than 0.4. It follows that mixing outdoor sounds can hamper more speaker recognition, and each type of outdoor sound significantly impacts on model effectiveness. Similarly to the indoor scenario, VGGVox was more robust than XVector, possibly due to its depth in terms of layers.

Based on our results, under the considered settings, speaker recognition matchers do not appear adequately reliable. The impact of background sounds on performance depends on the context and the sound.

5. Conclusions and Future Work

In this paper, we proposed a taxonomy of labeled background sound recordings for speaker recognition research in noisy environments. Then, we devised a simulation framework of indoor and outdoor contexts in vocal audios. Finally, we assessed the impact of the taxonomy sounds on the performance of two speaker recognition models. Based on the results, indoor sounds have a lower impact than outdoor sounds, and outdoor scenarios that involve mechanical sounds are the most challenging, even at low background sound volumes.

Our work opens to a wide range of research directions. We plan to enrich the taxonomy with more categories and audios organized into an ontological representation. We will extend our analysis to other models (e.g., ResNet) and languages beyond English. We will also inspect how background sounds and respective scenarios affect the internal model dynamics (e.g., speaker embeddings). Naturally, we will leverage our framework to devise audio enhancement methods able to deal with the sounds of our taxonomy and to design novel approaches for more robust speaker recognition.

Acknowledgments

This work has been partially supported by the Sardinian Regional Government, POR FESR 2014-2020 - Axis 1, Action 1.1.3, under the project "SPRINT" (D.D. n. 2017 REA, 26/11/2018, CUP F21G18000240009).

References

[1] J. H. Hansen, T. Hasan, Speaker recognition by machines and humans: A tutorial review, IEEE Signal processing magazine 32 (2015) 74–99.

[2] A. S. Tulshan, S. N. Dhage, Survey on virtual assistant: Google assistant, siri, cortana, alexa, in: Intern. Sym. on Signal Processing and Intelligent Recognition Systems, 2018, pp. 190–201.

[3] H. Feng, K. Fawaz, K. G. Shin, Continuous authentication for voice assistants, in: Proc. of the Annual International Conference on Mobile Computing and Networking, 2017, pp. 343–355.

[4] M. Schmidt, P. Braunger, A survey on different means of personalized dialog output for an adaptive personal assistant, in: Adjunct Publication of the Conference on User Modeling, Adaptation and Personalization, 2018, pp. 75–81.

[5] M. Marras, P. Korus, N. D. Memon, G. Fenu, Adversarial optimization for dictionary attacks on speaker verification, in: Proc. of the Annual Conference of the International Speech Communication Association, ISCA, 2019, pp. 2913–2917.

[6] A. Pradhan, K. Mehta, L. Findlater, Accessibility came by accident: Use of voice-controlled intelligent personal assistants by people with disabilities, in: Proc. of the CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13.

[7] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, The 2016 nist speaker recognition evaluation., in: Interspeech, 2017, pp. 1353–1357.

[8] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, H.-M. Wang, Audio-visual speech enhanc. using multimodal deep conv. neural networks, IEEE Transactions on Emerging Topics in Computational Intelligence 2 (2018) 117–128.

[9] A. Martin, L. Mauuary, Robust speech/nonspeech detection based on lda-derived parameter and voicing parameter for speech recognition in noisy environments, Speech communication 48 (2006) 191–206.

[10] T. Afouras, J. S. Chung, A. Zisserman, The conversation: Deep audio-visual speech enhancement, arXiv preprint arXiv:1804.04121 (2018).

[11] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-indep. speaker verification., in: Interspeech, 2017, pp. 999–1003.

[12] A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, Computer Speech & Language 60 (2020).

[13] C. Qian, Y. Yu, Z.-H. Zhou, Analyzing evolutionary optimization in noisy environments, Evolutionary computation 26 (2018) 1–41.

[14] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, S. Khudanpur, A study on data augm. of reverberant speech for robust speech recogn., in: 2017 IEEE Inter. Conference on Acoustics, Speech and Signal Processing, IEEE, 2017, pp. 5220–5224.

[15] X. Feng, B. Richardson, S. Amman, J. R. Glass, An environmental feature representation for robust speech recognition and for environment identification., in: INTERSPEECH, 2017, pp. 3078–3082.

[16] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, R. Marxer, An analysis of environment, microphone and data simulation mismatches in robust speech recognition, Computer Speech & Language 46 (2017) 535–557.

[17] C. Donahue, B. Li, R. Prabhavalkar, Exploring speech enhancement with generative adv. networks for rob. speech recogn., in: IEEE Inter. Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5024–5028.

[18] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Dig. Sig. Processing 10 (2000) 19–41.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. on Audio, Speech, and Language Processing 19 (2011) 788–798.

[20] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, M. W. Mason, I-vector based speaker recognition on short utterances, in: Proc. Interspeech 2011, 2011, pp. 2341–2344.

[21] Y.-h. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez, C. Parada, Locally-connected and convolutional neural networks for small footprint speaker recognition, in: Proc. Interspeech 2015, 2015, pp. 1136–1140.

[22] S. Kim, B. Raj, I. Lane, Environmental noise embeddings for robust speech recognition, arXiv preprint arXiv:1601.02553 (2016).

[23] J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound, in: Proc. of ACM Intern. Conf. on Multimedia, 2014, pp. 1041–1044.

[24] A. Brown, J. Kang, T. Gjestland, Towards standardization in soundscape preference assessment, Applied acoustics 72 (2011) 387–392.