Blind source separation by long-term monitoring: A variational autoencoder to validate the clustering analysis

Published Version:
De Salvio, D., Bianco, M.J., Gerstoft, P., D'Orazio, D., Garai, M. (2023). Blind source separation by long-term monitoring: A variational autoencoder to validate the clustering analysis. THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, 153(1), 738-750 [10.1121/10.0016887].

Availability:
This version is available at: https://hdl.handle.net/11585/914596 since: 2024-09-10

Published:

DOI: http://doi.org/10.1121/10.0016887

Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis

Domenico De Salvio,[1,2,a] Michael J. Bianco,[2] Peter Gerstoft,[2] Dario D'Orazio,[1] and Massimo Garai[1]

[1]*Department of Industrial Engineering (DIN), University of Bologna, Viale del Risorgimento 2, Bologna, 40136, Italy*

[2]*NoiseLab, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92037, USA*

Noise exposure influences the comfort and well-being of people in several contexts, such as work or learning environments. For instance, in offices, different kind of noises can increase or drop the employees' productivity. Thus, the ability of separating sound sources in real contexts plays a key role in assessing sound environments. Long-term monitoring provide large amounts of data that can be analyzed through machine and deep learning algorithms. Based on previous works, an entire working day was recorded through a sound level meter. Both sound pressure levels and the digital audio recording were collected. Then, a dual clustering analysis was carried out to separate the two main sound sources experienced by workers: traffic and speech noises. The first method exploited the occurrences of sound pressure levels via Gaussian Mixture Model and K-means clustering. The second analysis performed a semi-supervised deep clustering analyzing the latent space of a Variational autoencoder. Results show that both approaches were able to separate the sound sources. Spectral matching and the latent space of the variational autoencoder validated the assumptions underlying the proposed clustering methods.

[a]domenico.desalvio2@unibo.it

I. INTRODUCTION

A common metric for sound monitoring is represented by the A-weighted continuous equivalent level $L_{A,eq}$. Deeper statistical representations of acoustic monitoring are provided by percentile levels, i.e. the 95% SPL (Yadav *et al.*, 2021). However, the $L_{A,eq}$ does not show any detail about the acoustic scene (Green and Murphy, 2020). Further, the assessment of background noise levels through percentiles relies on temporal assumptions. The need of going beyond the $L_{A,eq}$ has been addressed especially in passive acoustic monitoring. In works concerning ecology and underwater acoustics, for instance, the assessment of the ambient noise levels is carried out through the probability density of the power spectral density (Merchant *et al.*, 2013, 2015; Parks *et al.*, 2009). The separation of sound sources would allow more detailed analyses of sound contexts. This ability can improve monitoring and design of several contexts resulting in the achievement of more comfortable spaces.

Workplaces are one of the most lived-in spaces by people. The achievement of a comfortable environment is important for both well-being and productivity. In offices, the latter are deeply influenced by noises. Individual perceptions can be affected by the nature of sound (Koskela *et al.*, 2014). The performances can either decrease or increase. It has been shown that both high or low frequency noises can improve cognitive tasks (Alimohammadi and Ebrahimi, 2017). The most important factor related to workers' comfort concerns the speech intelligibility. Thus, the most distracting noise for workers is represented by colleagues' speech (Ellermeier *et al.*, 2001; Haapakangas *et al.*, 2020). The NF S31-199 and the ISO 22955:2021 highlight the importance of assessing the noise levels at workstations depending on the activity carried out in the office. In particular, the ISO provides a survey for employees to rank the level of annoyance of several noise sources. This new approach deeply affects the design of open-plan offices (Harvie-Clark *et al.*, 2021; ISO 22955, 2021; NF S31-199, 2016). Thus, the ability of separating the noise sources is fundamental. However, there is a lack of ability in the technical praxis about measuring sound sources in real-world contexts.

Measurement techniques based on the statistical probability densities of SPLs were used to monitor noise contributions in classrooms. These methods are based on machine learning algorithms. Machine learning is the study of algorithms that improve their performance through the experience (Mitchell, 1997). The applications of these techniques frequently involve statistical methods and their use is rapidly increasing in acoustics (Bianco *et al.*, 2019). Long-term monitoring allow to collect large amounts of data. Thus, sound level meter measurements can be exploited using statistical methods. To analyze the collected SPLs, clustering techniques can find pattern among data. The multimodal SPLs' occurrences curve were exploited via Gaussian Mixture Model and K-means clustering to separately assess the noise due to the HVAC systems, the noise produced by the students, the teachers' speech, and the signal-to-noise ratios (D'Orazio *et al.*, 2020; Hodgson *et al.*, 1999; Wang and Brill, 2021). An application of the Gaussian Mixture Model in five offices was made to measure the human activity noise levels (Dehlbæk *et al.*, 2016). Then, two algorithms were proposed as unsupervised methods to separate and identify the mechanical noise and the human activity during working hours (De Salvio *et al.*, 2021).

Blind source separation is a major issue addressed not only in machine learning but in deep learning too. This is a type of machine learning based on artificial neural networks that learns representations of data with multiple level of abstraction (LeCun *et al.*, 2015). Inspired by the cocktail party effect, i.e. the ability of humans to focus the auditory attention to one speaker filtering other stimuli (Bronkhorst, 2000), the need of extracting the single source from a mixture of signals lies in many useful applications such as speech, music, and environmental audio processing (Vincent *et al.*, 2018). In the framework of the acoustic source separation, the concept of *deep clustering* was introduced. Deep clustering refers to the ability of performing clustering through deep learning algorithms (Hershey *et al.*, 2016). One of the most popular category to perform this is represented by the autoencoders. These kind of network performs a non-linear mapping of the data through an encoder and a decoder. The first maps the function to be trained, the second learns how to reconstruct the original data (Min *et al.*, 2018). Applications of autoencoders in acoustics concerned speech enhancement and clustering of geophysical data (Jenkins *et al.*, 2021; Lu *et al.*, 2013; Ozanich *et al.*, 2021).

A variational autoencoder is a deep generative model that forces the latent code of autoencoders to follow a predefined distribution (Min *et al.*, 2018). It has the same architecture of autoencoders, high-dimensional data are encoded into a low-dimensional latent space (Kingma and Welling, 2014). The ability of parametrizing data through a probability distributions gained broad attention in the deep learning community. Successful applications of variational autoencoders concern speech enhancement, blind source separation, and sound source localization in reverberant spaces (Bianco *et al.*, 2021; Leglaive *et al.*, 2019; Neri *et al.*, 2021).

The present work deals with the blind source separation through a sound level meter long-term monitoring. Basing on the methods proposed in previous work (De Salvio *et al.*, 2021), a dual analysis of the same phenomenon is proposed. A sound level meter recorded both the sound pressure levels and the digital audio of the working activity inside an office. Then, two clustering analyses were performed. The first exploited the two machine learning algorithms earlier mentioned, i.e. the Gaussian Mixture Model and the K-means clustering; the second performed a deep clustering analysis through a variational autoencoder. The goal is to identify and separately measure the main sound sources experienced by workers during the activity with both approaches.

II. THEORETICAL BACKGROUND

A. Clustering techniques

Clustering algorithms look for pattern in data (Bishop and Nasrabadi, 2006). Data are gathered in different clusters basing on their similarity. This kind of process is very useful when a great amount of unlabelled data is available. The task of clustering is finding useful properties among data, called features, which allow the data to be labelled. Different algorithms use different criteria to find similarity in data, i.e. shaping clusters. This study used two algorithms: the Gaussian Mixture model, and the K-means clustering.

1. Gaussian Mixture Model

Gaussian Mixture model (GMM) is a model-based clustering technique (McLachlan and Peel, 2004). A probabilistic model recovers the original general distribution. The latter is described as a linear combination of Gaussian curves. Given a set of N independent observations $X = \{x_1, ..., x_N\}$, the density $f(x_i)$ is:

$$f(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k) \tag{1}$$

where K are the number of components, $\mathcal{N}(x_i|\mu_k, \sigma_k)$ represents a Normal distribution with mean μ_k and covariance σ_k, and π_k is the *mixing proportion* or *weights*, that is:

$$0 \leq \pi_k \leq 1 \qquad (k = 1, ..., K) \tag{2}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{3}$$

The most common approach to fit mixtures of distributions is represented by the maximum likelihood (ML). ML means that, given a set of observations, the assumed statistical model is the most probable. The likelihood function $\mathcal{L}$ of a mixture of univariate normal distributed heteroscedastic components is defined as:

$$\mathcal{L}(x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}. \tag{4}$$

In the present study, the local ML are found via the iterative Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977).

2. K-means clustering

K-means clustering (KM) is a distance-based clustering technique (Bishop and Nasrabadi, 2006). It aims to shape a number of K clusters given a set of independent observations $X = \{x_1, ..., x_N\}$. Data are gathered minimizing the squared error Euclidean distance between the empirical mean c_{k_i}, called *centroid*, of a cluster k_i and the data points in the cluster. The squared error J is defined as:

$$J(k_i) = \sum_{x_{k_i} \in k_i} ||x_{k_i} - c_{k_i}||^2. \tag{5}$$

The goal is to minimize the sum of the squared error over all K clusters:

$$J(K) = \sum_{i=1}^{K} J(k_i). \tag{6}$$

K-means minimizes the objective function $J(K)$ through an iterative process. The main steps of the iterations are:

1. Selection of an initial partition of data into K clusters.
2. Generation of a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centres.

After the first step, steps 2 and 3 are repeated until convergence (Jain, 2010).

B. Model selection

An important issue in data clustering concerns the optimal number of clusters in data. For some classes of algorithms, such as GMM and KM, the number of clusters has to be specified before running the iterative process. Estimating the number of clusters is an open problem (Aggarwal and Reddy, 2014). Several metrics allow to find the most likely number of clusters with different approaches. Here, four metrics were used to assess the models' number of components, i.e. sound sources, in the collected data, as next.

1. Davies-Bouldin

The Davies-Bouldin index assesses similarity among clusters through the ratio of within- and between-cluster distances (Davies and Bouldin, 1979).

The within-to-between cluster distance ratio for the clusters k_i and k_j is defined as:

$$D_{i,j} = \frac{\bar{d}_{x_{k_i}} + \bar{d}_{x_{k_j}}}{d_{c_{k_i},c_{k_j}}} \tag{7}$$

where

$$\bar{d}_{x_{k_i}} = \frac{1}{n_{k_i}} \sum_{x_{k_i} \in k_i} |x_{k_i} - c_{k_i}| \tag{8}$$

is the average distance between each point in the cluster k_i and its centroid and n_{k_i} is the size of the cluster. Similarly, $\bar{d}_{x_{k_j}}$ is defined for the cluster k_j. The Euclidean distance between the centroids of both clusters is:

$$d_{c_{k_i},c_{k_j}} = (|c_{k_i} - c_{k_j}|^2)^{1/2}. \tag{9}$$

Then, with K as the number of clusters, the Davies-Bouldin index DB is defined as:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \{D_{i,j}\}. \tag{10}$$

The optimal model is represented by the smallest value obtained in equation 10.

2. *Gap statistic*

Gap statistic was introduced by Tibshirani et al. and formalizes the "elbow" method (Tibshirani *et al.*, 2001). The latter is a common empirical approach to find the best number of clusters by visualizing and assessing the highest decrease of the error measurement among models. The Gap criterion estimates the elbow by finding the largest gap value between the within-cluster dispersion of the model and the expected within-cluster dispersion of a reference distribution.

Let $d_{x_{k_i},x_{k'_i}}$ be the distance between observations x_{k_i} and $x_{k'_i}$ belonging to the same cluster k_i. The within-cluster dispersion is defined as:

$$W_K = \sum_{i=1}^{K} \frac{1}{2n_{k_i}} D_{k_i} \tag{11}$$

where n_{k_i} is the number of data in the cluster k_i, and D_{k_i} is:

$$D_{k_i} = \sum_{x_{k_i},x_{k_{i'}} \in k_i} d_{x_{k_i},x_{k_{i'}}}. \tag{12}$$

the pairwise distances of all points in the cluster k_i.

Then, the Gap value is defined as:

$$Gap(K) = \mathbb{E}_r^*\{\log(W_K)\} - \log(W_K). \tag{13}$$

where $\mathbb{E}_r^*$ is the expectation under a sample size r from the reference distribution. In the present study, the expected within-cluster dispersion of the reference distribution is evaluated via Monte Carlo sampling. The reference distribution is represented by a uniform distribution. The optimal model is represented by the highest value obtained in equation 13.

3. Calinski-Harabasz

The Calinski-Harabasz index measures the similarity of data points in clusters through the ratio between the separation and the cohesion of the model (Caliński and Harabasz, 1974). It is also know as *variance ratio criterion*. The separation SS_B is measured through the inter-cluster dispersion, i.e. the weighted sum of the Euclidean squared distances between the centroids of a clusters and the centroid of the whole dataset. It is defined as:

$$SS_B = \sum_{i=1}^{K} n_{k_i} ||c_{k_i} - C||^2 \tag{14}$$

where n_{k_i} is the number of observations in the cluster k_i, c_{k_i} is the centroid of the cluster k_i, and C is the centroid of the whole dataset.

The cohesion SS_W is measured through the intra-cluster dispersion, i.e. the sum of the Euclidean squared distances between each observation and the centroid of the same cluster. It is defined as J(K) in equation 6:

$$SS_W = J(K) = \sum_{i=1}^{K} \sum_{x_{k_i} \in c_{k_i}} ||x_{k_i} - c_{k_i}||^2 \tag{15}$$

where x_{k_i} is a data point in the cluster k_i.

Then, the Calinski-Harabasz index CH is defined as:

$$CH = \frac{SS_B}{SS_W} \frac{N - K}{K - 1} \tag{16}$$

The optimal model is represented by the highest value obtained in equation 16.

4. Silhouette coefficient

The silhouette coefficient is a graphical quantitative evaluation of the degree of separation among clusters (Rousseeuw, 1987). Given two data points x_{k_i} and $x_{k_{i'}}$ in the cluster k_i, the within-cluster mean distance, i.e. the similarity, between x_{k_i} and the other $x_{k_{i'}}$th points in the same cluster is defined as:

$$a(i) = \frac{1}{|n_{k_i}| - 1} \sum_{x_{k_i}, x_{k_{i'}} \in k_i} d_{x_{k_i}, x_{k_{i'}}}. \tag{17}$$

The dissimilarity between x_{k_i} and the other x_{k_j}th points belonging to the cluster k_j, is defined as the mean distance between x_{k_i} and x_{k_j}. Hence, the shortest distance between x_{k_i} and the other points of other clusters is defined as:

$$b(i) = \min \frac{1}{|n_{k_j}|} \sum_{x_{k_i} \in k_i, x_{k_j} \in k_j} d_{x_{k_i}, x_{k_j}}. \tag{18}$$

The cluster with the lowest dissimilarity is defined as "neighbor" and represents the second best choice for k_i. The silhouette value $s(i)$ is defined as:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i). \end{cases} \tag{19}$$

It can be deduced that $-1 \leq s(i) \leq 1$. Thus, x_{k_i} is deemed properly clustered if $s(i)$ is close to 1, and wrongly clustered if close to -1. In case $s(i)$ is close to 0, either k_i or k_j represent a good choice for x_{k_i}. If $s(\bar{i})$ is the mean of each $s(i)$, the silhouette coefficient SC can be defined as:

$$SC = \max_K \bar{s}(K) \tag{20}$$

where K is the number of clusters. The SC is defined only for a number of clusters $K > 1$. The optimal model is represented by the highest value obtained in equation 20.

C. Variational Autoencoder

The variational autoencoder (VAE) is a way to realize inference and learning in probabilistic models and was introduced by Kingma and Welling (Kingma and Welling, 2014; Kingma *et al.*, 2019). From a deep learning perspective, a VAE has the same architecture of autoencoders. Thus, it is made by an encoder and a decoder. Both are connected by a latent space. One of the most important qualities of VAEs concerns their ability of describing observations through a probabilistic approach in the latent space. Like classical autoencoders, a VAE tries to reconstruct output from input. Thus, it learns a latent variable model for its input data.

The encoder is represented by a neural network. Its aim is to output a latent hidden representation z of the input x with weights and biases θ. Typically, the latent space has a lower dimension with respect to the input size. Thus, it can be deduced that the encoder learns a compressed representation of the input data according to the distribution $q_\theta(z|x)$. In the present study, the input $x \in \mathbb{R}^{m_1 \times m_2}$ and its latent representation $z \in \mathbb{R}^n$ and the distribution $q_\theta(z|x)$ is represented by a Gaussian probability density.

The decoder is a neural network as well. Typically, it has a mirrored architecture of the encoder. Its aim is to reconstruct the input sampling only from the compressed representation of the latent space z. Thus, it outputs parameters to the probability distribution of data with weight and biases ϕ. The decoder process is denoted by the distribution $p_\phi(x|z)$. The latter is represented by a standard Normal distribution $\mathcal{N}(0, 1)$ with mean 0 and variance 1.

The whole process is assessed by the evidence lower bound (ELBO) loss function. For a datapoint x_i, it is defined as:

$$l_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[log p_\phi(x_i|z)] + D_{KL}(q_\theta(z|x_i)||p_\phi(z)) \qquad (21)$$

where the first term is called *reconstruction loss* and it is represented by the expected negative loglikelihood of the ith datapoint. It describes the amount of information lost for the reconstruction through the whole process. The expectation is calculated with respect to the encoder's distribution over the representations. The second term is called *regularizer term* and it is represented by the Kullback-Leibler divergence between the two distributions q_θ and p_ϕ (Kullback and Leibler, 1951). Thus, it describes how the two distributions are close one each other. The $\sum_{i=1}^{N} l_i$ is the total loss evaluated over the whole dataset of N datapoints.

III. EXPERIMENTAL SETUP

The case study is represented by a small office with 3 workers placed in 3 different workstations. The monitoring was conducted after the COVID-19 emergency. Hence, people wore face masks. The type of work carried out in the office is collaborative. Thus, there is high interaction between colleagues. The analysis is based on two recordings of the same event: the sound pressure levels (SPLs) and the digital audio. The sound level meter acquired octave band filtered (125 – 4000 Hz) sound pressure levels every 0.1 seconds. The digital audio was recorded with a sample rate of 51.2 kHz and a depth of 32 bit. These recordings represent the raw data used in the experimental process. Figure 1 shows the plan of the office and the arrangement of the workstations besides the placement of the sound level meter. The sound level meter collected about 6 hours of working activity in the office.

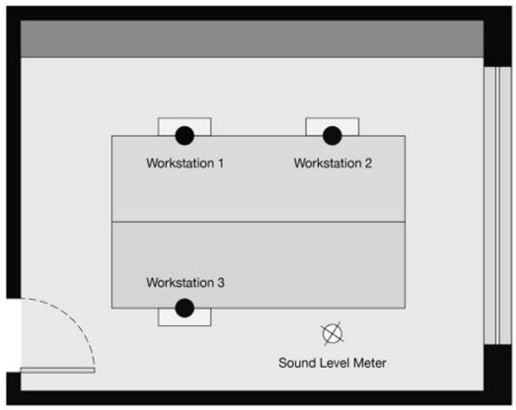


FIG. 1. (Color online) Plan of the office under study.

The air conditioning system was turned off during the measurement and the window is exposed towards an highly busy road. Thus, the sound environment can be described as made by two kinds of sound sources: the traffic and the speech. The room has volume of about 60 m^3 with no acoustic treatments and can be considered as a ”lightly damped” environment. The acoustical properties of the office, in particular the reverberation time and the façade sound insulation, are shown in Section IV A 2. Figure 2 shows a sample of the data used. The waveform on the top represents a 10-minute recording, the time series of SPLs in the middle is used in the machine learning approach, the spectrograms at the bottom are exploited for the deep learning process.

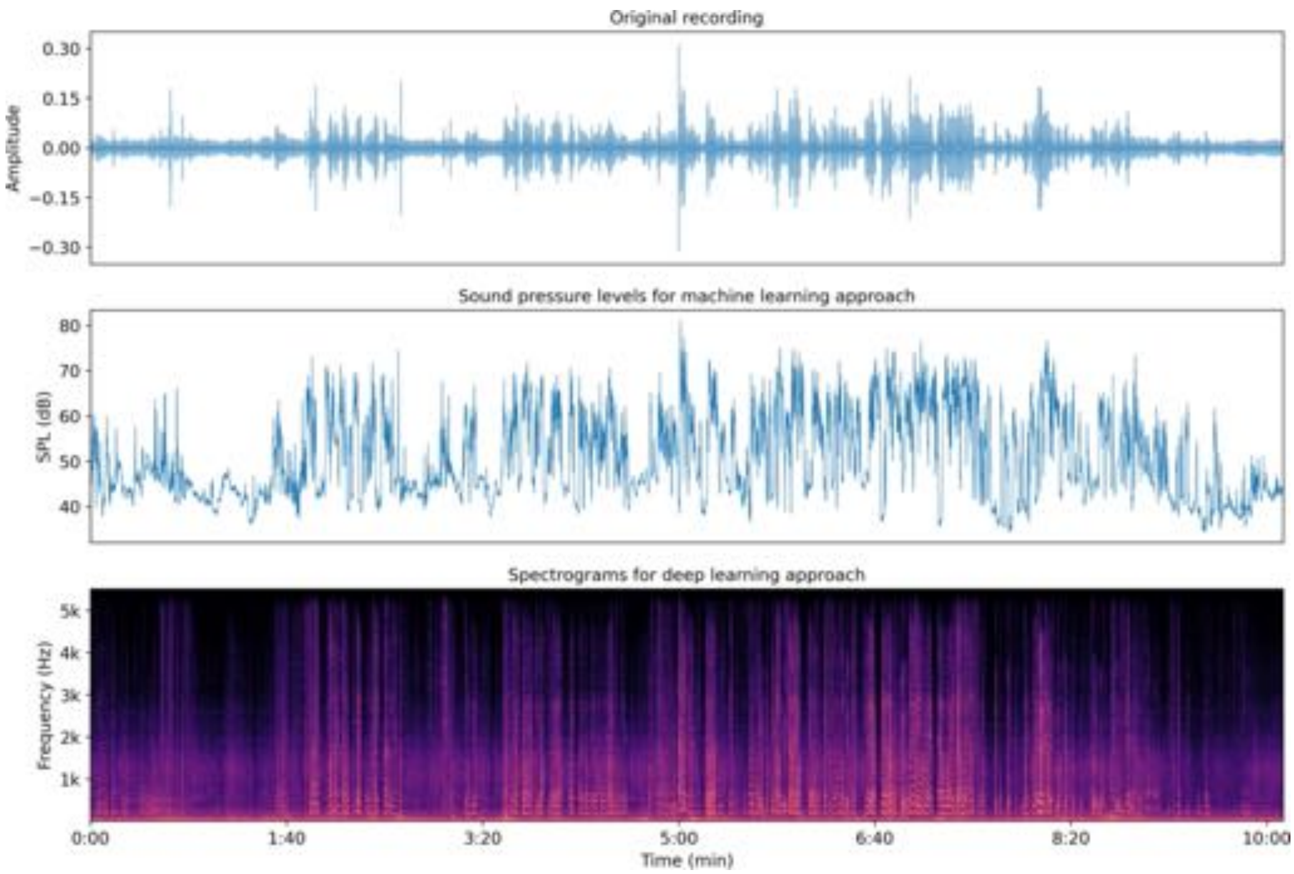


FIG. 2. (Color online) Example of the data used in this study. On the top, a sample of 10 minutes recording. In the middle, the sound pressure levels obtained in the same sample. This constitutes one of the databases for the machine learning approach. On the bottom, the spectrograms obtained by the same sample used for the deep learning approach.

A. GMM and KM analyses

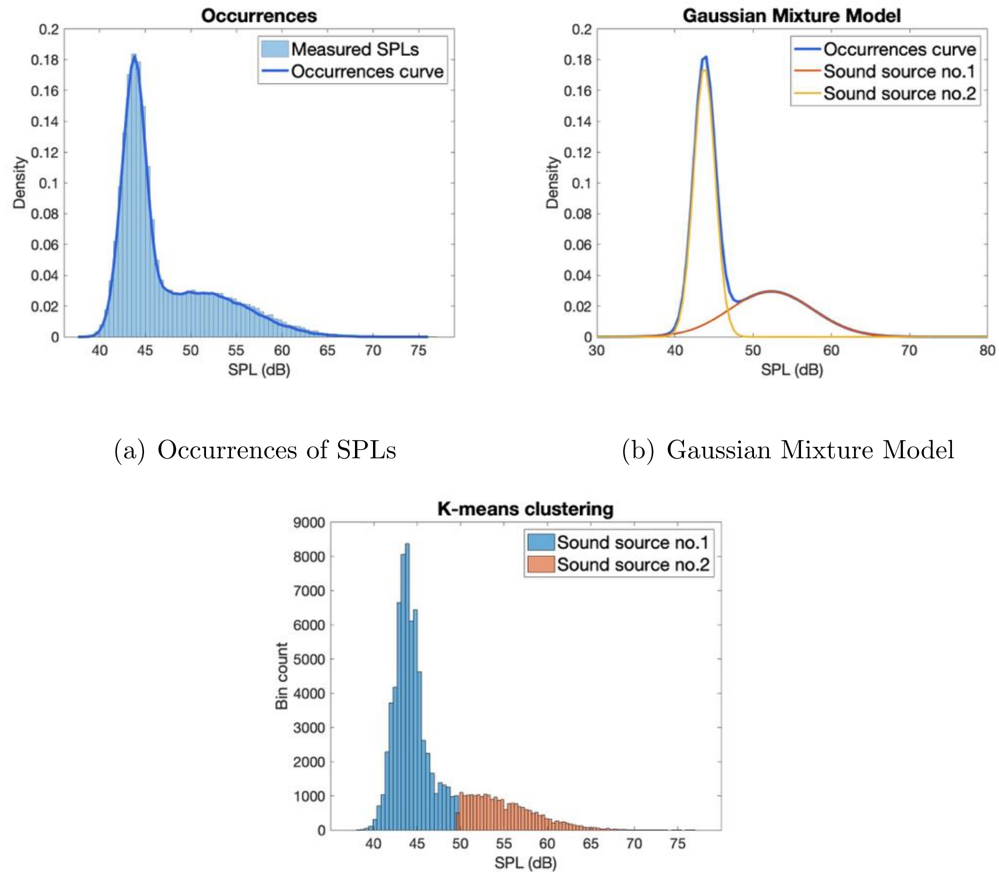
The procedure of the clustering analyses via GMM and KM follows the same flow described in a previous work (De Salvio *et al.*, 2021). Once the distribution of the SPLs occurrences is obtained, the number of clusters to look for in the collected data has to be set first. Thus, models with 2 up to 10 components, i.e. sound sources, for both GMM and KM were used as candidates. Each metric used for searching the most likely number of clusters finds similarity among clusters according to its own approach. Hence, the metric's highest value represents the best model for Silhouette, Calinski-Harabasz, and Gap statistic, while the metric's lowest value represents the best choice for Davies-Bouldin coefficient. In this study, the the majority rule was used to obtain the optimal number of clusters. Thus, the most frequent number obtained comparing each metric represents the number of active sound sources during the event.

Figure 3 shows the type of distribution used for the clustering. On the left, the SPLs occurrences are collected and analyzed through the normalized occurrences distribution. In the middle, an example of processing via GMM and on the right, an example of processing via KM.

Following a brief summary of the procedure:

- **Step 1:** Clustering analysis performed over several candidate models.
- **Step 2:** Selection of the best model among candidates.
- **Step 3:** Spectral analysis and sources labelling according to statistical or distance metrics.

The three steps are applied for each different clustering algorithm we want to use. Basing on the acoustic task, any SPLs' frequency band can be used to complete the process. In this study, the whole procedure is carried out in the range from 125 up to 4000 Hz octave bands. Hence, each occurrences' curve is analyzed looking for the most likely number of clusters. Both GMM and KM were set to repeat the iterative process with different initial values. In general, using different starting points typically results in a solution that is a global minimum (Jain *et al.*, 1999).



(a) Occurrences of SPLs

(b) Gaussian Mixture Model

(c) K-means clustering

FIG. 3. (Color online) Machine learning approach: example of a sound level meter's measurement processing. The figure on the left shows the occurrences distribution of the measured SPLs. The distribution can be processed via GMM (at the center) and KM clustering (on the right).

After the step 2, each model for each octave band can be collected. The means of each Gaussian component obtained by the GMM represent the SPLs of each source. Similarly, the centroids of each cluster obtained via KM represent the SPLs of each sound source. Thus, the spectra can be reconstructed.

Labelling the sound sources, i.e. linking each spectra to each bunch of clusters found in each octave band, exploits the temporal parameters of the clusters. The dispersion of data can be associated to the temporal behavior of the sound sources. Dense clusters represent nearly stationary noises, while spread data refer to a random source. Being the machine learning approach an unsupervised analysis, this step is performed after the optimal model is selected and depends on the clusters' features given by the algorithm. Concerning the GMM, a cluster's standard deviation (SD) equal or greater than 5 dB refers to a speech source. Values lower than 5 dB describe a mechanical source. In fact, preliminary studies show that this value is deemed as a good threshold to separate continuous sound sources from human-related noises (Leonard and Chilton, 2019; Olsen, 1998). Regarding the KM, the temporal properties of the sound sources are described by the square root of the average intra-cluster Euclidean distance (AICD) of data points. Similarly to the SD, lower values are associated to continuous noises, otherwise to human noises.

B. VAE analysis

The digital audio recording has been chunked in 1-second length samples to obtain the dataset for the analysis through the VAE. Power spectrograms of each chunk were used as input for the network. A pre-processing procedure has been carried out before feeding the encoder. The audio has been resampled at 11025 Hz to make the input comparable to the octave band range (125-4000 Hz) used in the cluster analysis. Moreover, visualizing the spectrograms, no useful information were found above 5 kHz. Short-time Fourier Transforms (STFT) with a segment length of $N_{FFT} = 256$ and an overlap area of 50% were used to obtain the spectrograms. With these values each audio chunk is processed with an FFT with a physical length of about 20 ms. Thus, it can be deemed that in each FFT only one sound source is detected. A minMAX normalization has been applied to each spectrogram to have all the amplitude values in [0,1] range. Overall, the dataset contained about 23k samples.

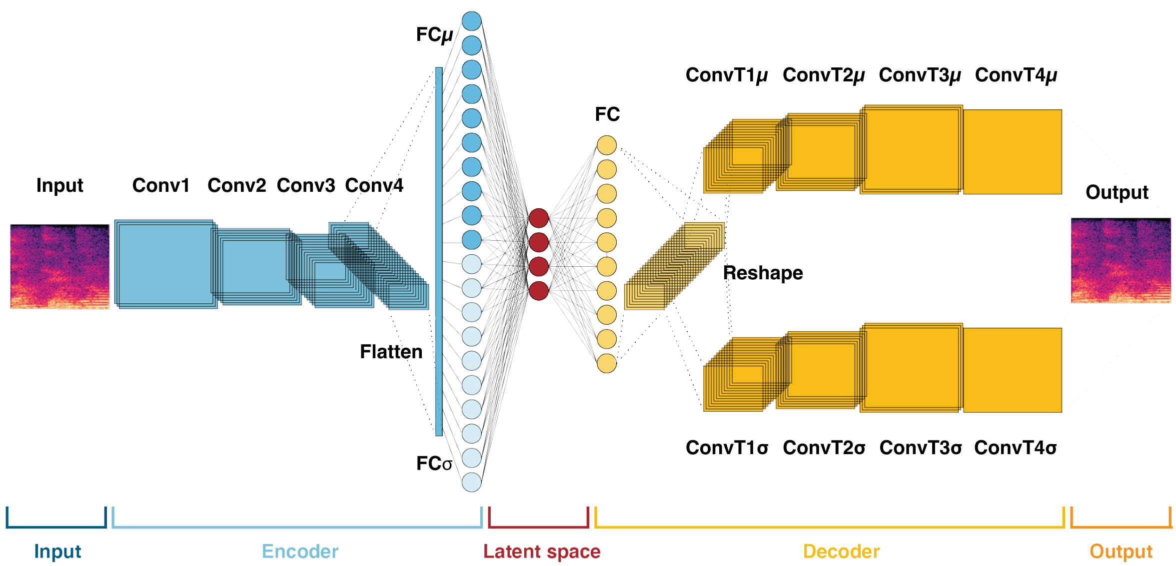


FIG. 4. (Color online) Architecture of the VAE. The encoder is constituted by four convolutional layers, shown in light blue. The latent space is shown in red and the decoder is represented by the yellow blocks.

Samples of 1-second length can be easily listened. Then, the dataset has been manually labelled listening each sample of the recording in three classes: traffic, speech, and unclassified sounds. The latter category was useful to label all the samples where the main sound source was represented by impulsive noises like slammed doors. It is worth noting that, during the labelling process, audio chunks containing only whispers were labelled as speech. This choice can create uncertainties on the dataset's labels. At the end of the labelling

TABLE I. Architecture of the variational autoencoder. The type of layers and their properties, like input shape, filters, kernel size, the activation functions, and the output size are shown.

	Layer	Input shape	Filters	Kernel size	Activation	Output shape
Input	Reshape	[128,87]	–	–	–	[1,128,87]
Encoder	Convolutional (stride = 2)	[1,128,87]	16	[3,3]	ReLU	[16, 64, 44]
	Convolutional (stride = 2)	[16, 64, 44]	32	[3,3]	ReLU	[32, 32, 22]
	Convolutional (stride = 2)	[32, 32, 22]	64	[3,3]	ReLU	[64, 16, 11]
	Convolutional (stride = 2)	[64, 16, 11]	128	[3,3]	ReLU	[128, 8, 6]
	Flatten	[128,8,6]	–	–	–	[6144]
	Fully connected mu	[6144]	–	–	–	[30]
	Fully connected sigma	[6144]	–	–	–	[30]
Latent space	Fully connected	[30]	–	–	–	[30]
Decoder	Fully connected	[30]	–	–	ReLU	[6144]
	Reshape	[6144]	–	–	–	[128, 8, 6]
	Transpose convolutional mu (stride = 2)	[128,8,6]	128	[3,3]	ReLU	[64, 16, 11]
	Transpose convolutional mu (stride = 2)	[64, 16, 11]	64	[3,3]	ReLU	[32, 32, 22]
	Transpose convolutional mu (stride = 2)	[32, 32, 22]	32	[3,3]	ReLU	[16, 64, 44]
	Transpose convolutional mu (stride = 2)	[16, 64, 44]	16	[3,3]	ReLU	[1,128,87]
	Transpose convolutional sigma (stride = 2)	[128,8,6]	128	[3,3]	ReLU	[64, 16, 11]
	Transpose convolutional sigma (stride = 2)	[64, 16, 11]	64	[3,3]	ReLU	[32, 32, 22]
	Transpose convolutional sigma (stride = 2)	[32, 32, 22]	32	[3,3]	ReLU	[16, 64, 44]
	Transpose convolutional sigma (stride = 2)	[16, 64, 44]	16	[3,3]	ReLU	[1,128,87]
Output	Reshape mu	[1,128,87]	–	–	Tanh	[128,87]
	Reshape sigma	[1,128,87]	–	–	Sigmoid	[128,87]

process, the dataset had more than 12k traffic samples and about 10.5k speech samples. Only 139 samples were labelled as unclassified. The dataset can be considered balanced. The 80% of the dataset was used for the training set, the remaining 20% for the test set.

The VAE was built in Pytorch. The input size of the spectrograms is 128×87. The encoder is made by four strided convolutional layers (stride = 2). Then, a flatten layer allows the convolutional layers to be linked to the fully connected layers. As highlighted in Section II C, a VAE maps the input to a multivariate latent distribution. The distribution used in the present work is the Gaussian distribution. Parameters of the distribution, the mean and standard deviation, i.e. mu and sigma, are the outputs of the encoder. For this reason, the fully connected layer of the encoder is doubled. Here are the inputs for the 30-dimensional latent space. The decoder uses a different distribution, the prior on the latent distribution. The architecture of the decoder depends on the parameters needed to specify the multivariate generative distribution. In the present work, two parameters are needed: the mean and the variance. Thus, the decoder is made by four strided transposed convolutional layers (stride = 2) for both parameters. The Pyro library was used for the stochastic variational inference. Then, spectrograms are reconstructed reshaping the output of mean and variances obtained by the decoder. Non-linearities are activated through $ReLU$ functions for all layers except for the output parameters. The decoder is parametrized according to a standard Normal distribution $\mathcal{N}(0, 1)$. Thus, a $Tanh$ activation function is used for the output of means and a $Sigmoid$ activation function for the the output of variances. The VAE was trained using a batch size of 32, the Adam optimizer and θ and ϕ weights were updated with a learning rate equal to 1×10^{-5}. Figure 4 shows a graphical scheme of the VAE's architecture. The light blue and yellow layers represent, respectively, the encoder and the decoder. Both are linked by the latent space represented with the red fully connected layer. Details about the architecture of the whole network are listed in Table I. Here, the type, the input size, the number of filters, the kernel size, the activation functions, and the output size are shown for each layer. Training stopped after 400 epochs since not relevant improvements of the loss function on the test dataset were detected.

IV. RESULTS

A. Machine Learning results

1. *Source separation via GMM and KM*

Previous work (De Salvio *et al.*, 2021) used only one metric per each algorithm to assess the optimal number of cluster. Moreover, the elbow technique made the analysis influenced by the operator's choice. Thus, the step 2 described in section III A is different. The comparison among different metrics makes the analysis more robust and inclined to the automation. Table II shows the results of model selection. Silhouette (SC), Davies-Bouldin (DB), Gap statistic (GS), and Calinski-Harabasz (CH) coefficients were used to assess the most likely number of clusters for each octave band (125-4000 Hz) and the A-weighted continuous level $L_{A,eq}$. Concerning GMM, the model selection metrics found that the optimal number of clusters is equal 2 according to the majority rule. This is true for SC and DB for each octave band and $L_{A,eq}$. Different results were found only for GS in the 125 Hz octave band and for CH in the 500 and 4000 Hz octave bands. The same analysis was carried out for KM. Here, SC, DB, and GS found an optimal number of clusters equal to 2 for each occurrences curve analyzed. Completely different results were shown by CH that found 6 clusters in each octave band and $L_{A,eq}$ as the best model. Overall, comparing all metrics, the number of active sources in the office is 2. These results are consistent with the expectations. The main sound sources experienced during a common working day by employees were speech and traffic, indeed.

Figure 5 shows the reconstructions of the spectra of both sound sources. Then, the plots in the middle and on the bottom show the relative spectra compared with references from standards. Blue lines show results for GMM, red lines for KM. In the relative analyses, yellow lines show reference spectra. To compare the reconstructed one with reference, each relative spectrum is obtained by setting the 1 kHz octave band to 0 dB. Table III shows the quantitative results obtained via clustering analysis.

Both algorithms showed very similar qualitative results. Spectra have the same tendencies, indeed. The most noticeable difference concerns the peak of the speech spectra. It is detected in the 500 Hz octave band for KM while in the 250 Hz octave band for GMM. With respect to previous work, low frequencies seem to be easier to separate in this case for both algorithms (De Salvio *et al.*, 2021). This may be due to the different background noise, the traffic outside the office instead of a mechanical noise inside the same space.

Concerning the traffic noise, the reference is represented by the normalized spectrum shown in EN 1793-3 (EN 1793-3, 1997). It is worth noting that the reference spectrum refers to free field conditions. Thus, acoustical properties of the room and the facade's insulation can affect the shape of the results. The shape of the traffic spectra seem to be

TABLE II. Analysis of the most likely number of clusters in the measured SPLs. Results are shown per each metric, octave band from 125 up to 4000 Hz, and the continuous A-weighted level $L_{A,eq}$. Metric abbreviations refer to silhouette (SC), Davies-Bouldin (DB), Gap statistic (GS), and Calinski-Harabasz (CH) coefficients. Majority rule's row show the optimal number of clusters used to run both GMM and KM algorithms.

GMM							
Metric	Frequency octave band (Hz)						$L_{A,eq}$
	125	250	500	1k	2k	4k	
SC	2	2	2	2	2	2	2
DB	2	2	2	2	2	2	2
GS	5	2	2	2	2	2	2
CH	2	2	4	2	2	5	2
Majority rule							
No. Sources	2	2	2	2	2	2	2

KM							
Metric	Frequency octave band (Hz)						$L_{A,eq}$
	125	250	500	1k	2k	4k	
SC	2	2	2	2	2	2	2
DB	2	2	2	2	2	2	2
GS	2	2	2	2	2	2	2
CH	6	6	6	6	6	6	6
Majority rule							
No. Sources	2	2	2	2	2	2	2

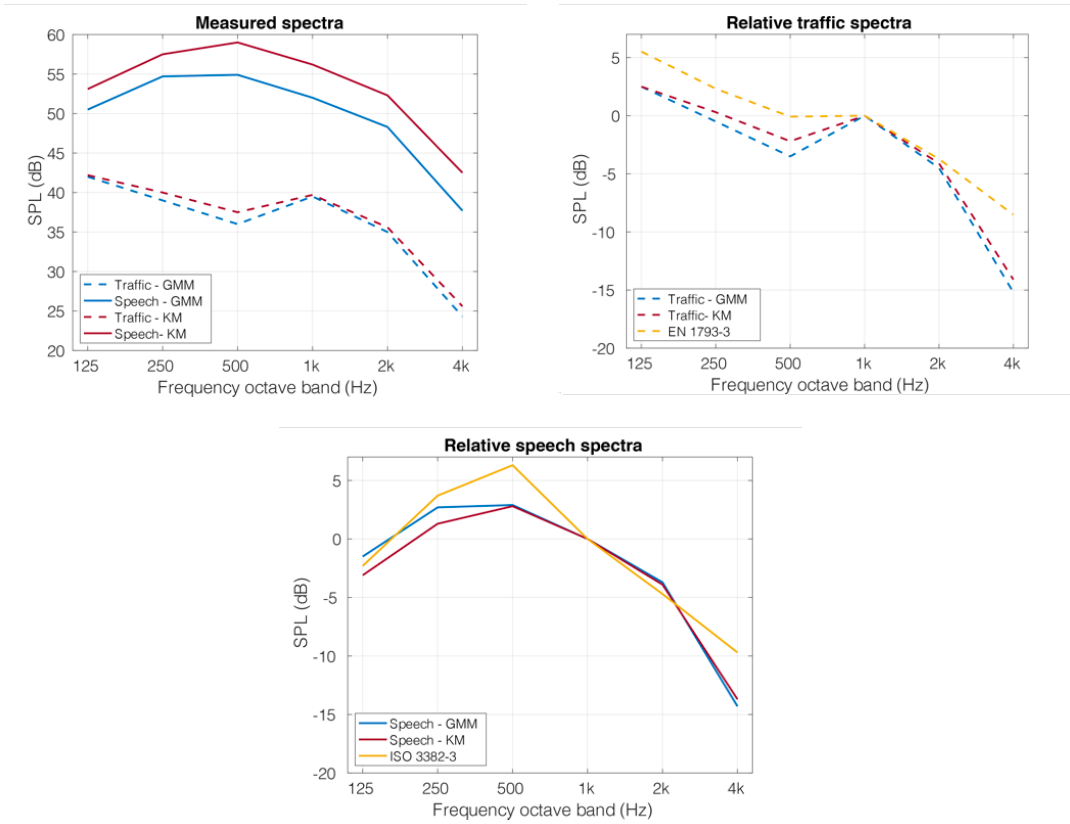


FIG. 5. (Color online) Results of clustering analyses. On the top: reconstruction of the spectra from 125 up to 4000 Hz. Blue and red lines represent the spectra reconstructed respectively via GMM and KM. Dashed and solid lines represent respectively the traffic and the speech spectra. In the middle and on the bottom: relative spectra of traffic and speech spectra compared with references curves. Traffic reference is taken from EN 1793-3, speech reference is taken from ISO 3382-3.

very similar. The most noticeable difference concerns the 500 Hz octave band. However, both low-frequencies emitted at slow speeds and the 1 kHz frequencies emitted at free-flow speed seem to be accurately detected (Can *et al.*, 2010).

TABLE III. SPLs of each sound source obtained via GMM and KM. Standard deviations SD for GMM and average intra-cluster distance AICD for KM are reported.

Source	Frequency octave band (Hz)						$L_{A,eq}$
	125	250	500	1k	2k	4k	
GMM							
Traffic	42.0	39.0	36.0	39.5	35.0	24.3	42.5
SD	3.0	3.0	3.3	4.3	4.0	3.7	3.5
Speech	50.5	54.7	54.9	52.0	48.3	37.7	57.5
SD	5.8	7.1	9.1	8.9	8.5	8.7	7.8
KM							
Traffic	42.2	40.0	37.5	39.7	35.6	25.6	43.3
AICD	2.9	3.7	4.5	4.2	4.2	4.3	4.0
Speech	53.1	57.5	59.0	56.2	52.3	42.5	60.8
AICD	4.1	5.2	6.4	6.3	6.1	6.3	5.8

The ISO 3382-3 shows the reference speech spectrum of a directional source at a distance of 1 m in free field from the speaker (ISO 3382 - 3, 2012). This is the reference for the speech source; the related spectra obtained via clustering have similar tendencies as shown on the bottom of Figure 5. Differences can be referred to several factors. The first concerns the influence of the acoustical properties of the room. As noticed for the traffic noise, the ISO spectrum is evaluated at a distance of 1 m from the source in free field. As opposed to previous work, slight differences concern low frequencies for speech sources. However, these can be due to the change of the spectrum in noisy environments and the measurement uncertainty at low frequencies, especially in the 125 and 250 Hz octave bands (Leembruggen *et al.*, 2016; Rindel *et al.*, 2012). Directivity of the source can affect spectra tendencies too. In the present study, there are 3 speakers in 3 different positions. Thus, the overall directivity of the measured source cannot be considered the same as the reference. Moreover, at low frequencies, modal effects could have affected the results since the sound level meter was used only in one position.

Both sources show a drop concerning the 4 kHz octave band. This may be attributed to the acoustical properties of the room since higher frequencies can be strongly affected by their interactions with surfaces and furnitures in the room.

Further considerations can be made regarding the clusters size. This is described by the SD and the AICD; both are shown in brackets in Table III. The physical meaning associated to SD and AICD is the temporal randomness of the source. Mechanical sources produce the same SPLs occurrences depending on their mechanical cycle, indeed. This results in low SDs for continuous sources because the corresponding Gaussian curve will be narrow. On the contrary, a human-related noise produce higher SDs. The traffic noise can be deemed in the middle of these two categories of noise sources. It does not have the same continuity of a mechanical device but it has specific spectral properties. Moreover, the road has to be busy to be detected in a long-term monitoring because the occurrences curve has to be affected by the noise source. Thus, traffic can be deemed more continuous than the speech but not like a mechanical source. These considerations are confirmed by the results obtained. Traffic SDs lie in the range 3.0 - 4.3 dB for each octave band. Previous work showed mechanical SDs due to the HVAC system in the range 0.9 - 3.9 dB. Thresholds analyses deserve detailed studies in future works. However, all non-human sound sources were confirmed to be under the threshold of 5 dB (De Salvio *et al.*, 2021).

The absolute spectra shown on the top of Figure 5 point out differences between SPLs of the two methods that can be related to the homoscedasticity of data, i.e. constant variances of data. GMM can be considered as a general case of KM. The two algorithms show the same results only if the homoscedasticity condition is fulfilled (MacKay, 2003). This is shown in Table III. SPLs are the same for GMM and KM when SD and AICD are almost equal, e.g. in the 125 and 1000 Hz octave bands of the traffic source. This result confirms that AICD is a reliable metric to assess the shape of the cluster. It has to be noted that the size of clusters can be affected by the type of clustering performed by the algorithm, especially for large SDs. GMM is a soft clustering algorithm, i.e. it can assign data points to more than one cluster with proportional weights. KM performs hard clustering, instead, i.e. assign each data point to one and only one cluster (Bishop and Nasrabadi, 2006). The GMM's fuzziness can affect the resulting SD of clusters associated with random sound sources.

2. Hints on the influence of the office's acoustical properties

As noted in previous section IV A 1, the acoustical properties of the office influence the spectra obtained via GMM and KM. Thus, the reverberation time T_{20} and the façade sound level difference $D_{2m,nT}$ were measured respectively according to the precision method described in the ISO 3382-2 and the global method of the ISO 16283-3 (ISO 16283 - 3, 2016; ISO 3382 - 2, 2008). Measurements' results are shown in Figure 6. Solid and dashed lines show respectively the T_{20} and $D_{2m,nT}$ tendencies in octave bands from 125 up to 4000 Hz.

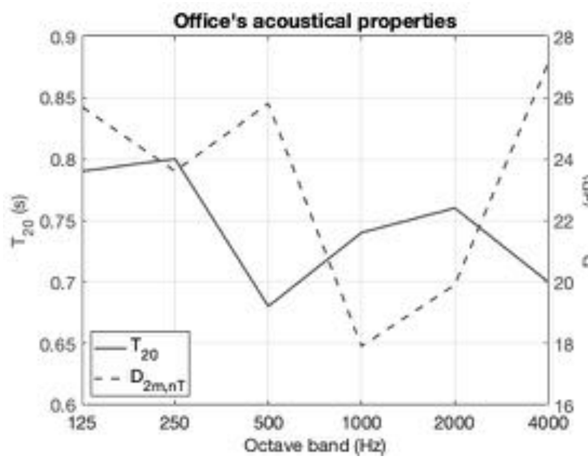


FIG. 6. Acoustical properties of the office under study. The reverberation time T_{20} is shown on the left axis, the façade sound level difference $D_{2m,nT}$ on the right axis.

The office has a reverberation time averaged on the mid frequencies of 500-1000 Hz of about 0.72 s. The environment can be deemed as "live" because there are no acoustic treatments. There is a reverberation's drop in the 500 Hz band maybe due to two steel closets. The façade has an average insulation of about 22 dB at the mid frequencies of 500 and 1000 Hz. The drop of $D_{2m,nT}$ in the 1 kHz band is due to the coincidence effect of the window glass.

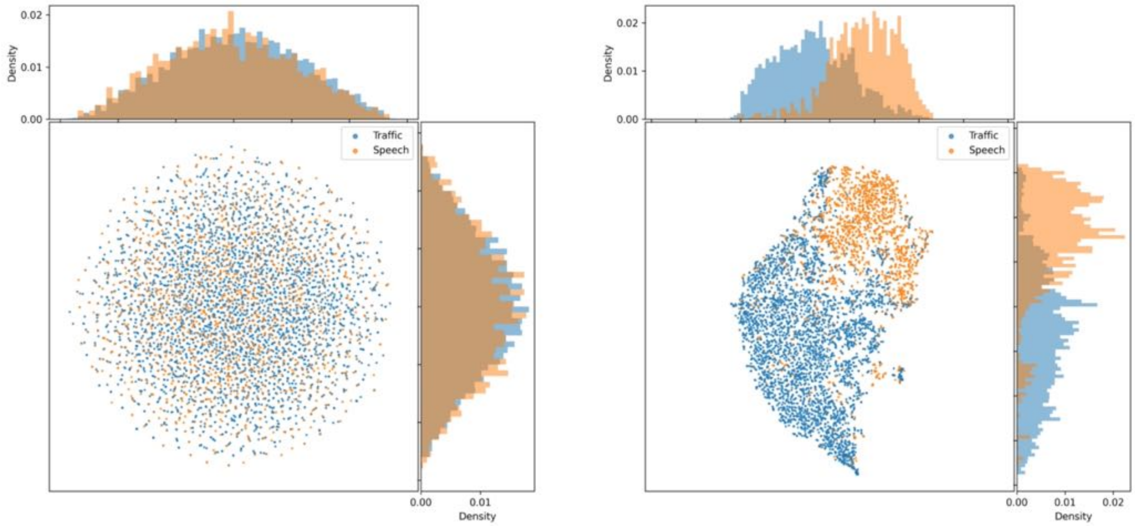
The tendencies of measurements' results in Figure 6 may bring preliminary insights about the comparison of measured and reference spectra shown in Figure 5. Traffic and speech spectra seem to be related to the tendency of the T_{20}. In fact, the drop in the 500 Hz octave band is visible in both sources. Further, the reverberation time has its minimum value in the same band, as well as one of the highest values of the façade insulation. The energy of both sources in the 4 kHz octave band seems to be affected respectively by the T_{20} for the speech and by the $\mathrm{D}_{\mathrm{2m,nT}}$ for the traffic. Thus, a preliminary analysis of the room's acoustics seems to support the results obtained through the machine learning approach.

B. Deep learning results

1. Latent space

The clustering analysis carried out through the machine learning approach has been based on assumptions and spectral matching. The discussion of these evidences depends on the operators' knowledge. Hence, it is useful to find an objective method to either confirm or not the goodness of using GMM and KM. A semi-supervised analysis via deep learning allows the results to be directly evaluated. This is possible because the audio recording can be listened. Further, the latent space of a VAE is able to perform a clustering analysis. Thus, the deep and the machine learning approaches can be compared. The difference between the two approaches is due to the labelling step. In the machine learning approach, the step was made at the end of the process, in the deep learning approach, the data were previously labelled. Thus, the latent space of the VAE aims to be a qualitative tool to assess the machine learning approach.

Figures 7(a) and 7(b) show the latent distributions of respectively the untrained and trained network. Because the dimension of the latent space is equal to 30, a 2D t-stochastic



(a) Untrained network

(b) Trained network

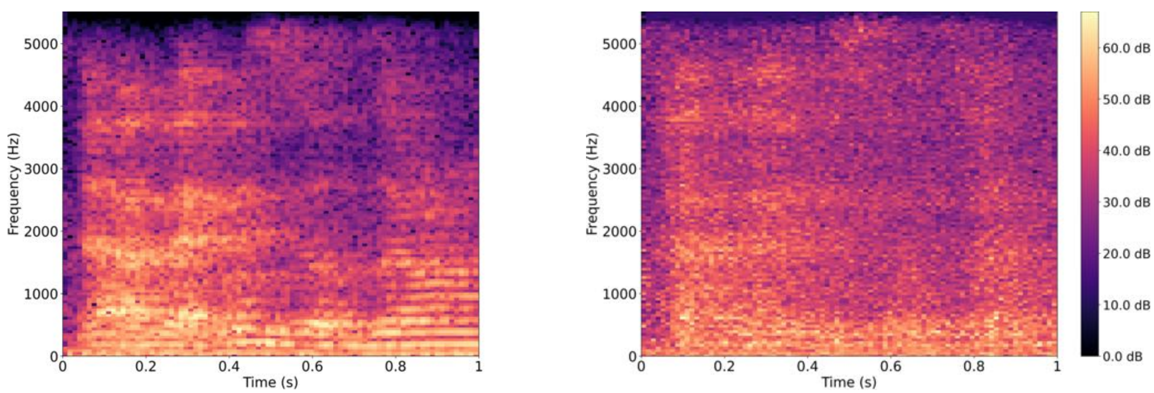
FIG. 7. (Color online) Latent space of the untrained 7(a) and trained 7(b) VAE. Histograms show the x and y projections of the density distributions of the data. Blue and orange dots and histograms represent respectively the traffic and the speech data.

neighbor embedding (t-SNE) visualization was used (Van der Maaten and Hinton, 2008). This is a dimensionality reduction technique commonly used to visualize high-dimensional data. The t-SNE algorithm evaluates similarity between pairwise instances in both high and low dimensional space. Then, through a cost function, the similarities are optimized. Figures 7(a) and 7(b) are obtained with a perplexity equal to 30.

Data in the latent space are represented basing on their categorical label. The untrained latent space in Figure 7(a) shows a circular distribution of data since it is perfectly described by a Gaussian distribution (Connor *et al.*, 2021). However, there is no categorical separation among data, i.e. blue and orange dots are mixed up. Figure 7(b) shows the results of the training. After the network has learnt the latent representation of the input data, the latent space shows a clear separation of the two categories. Clusters are well-defined. On the sides, histograms show the 1D projection of the plot along the main axes. These distributions help to assess whether the two clusters in the 2D plot are overlapped or not. In the present case, histograms of the trained network show that the two clusters are close but not overlapped. Thus, clusters are well-separated too. The VAE is able to identify and separate the two sound sources through a Gaussian latent space. Different densities within clusters may refer to further properties, e.g. timbre, not considered in the categories taken into account in this study. Uncertainties on data distributions, i.e. speech frames in the traffic cluster and vice versa, can be attributed to the manual labelling. For instance, whispers can be manually labelled as speech but classified by the network as traffic.

2. The reconstruction of the spectrograms

The aim of these approaches is to measure different sound sources. Thus, the reconstruction of the audio recording can be post-processed to achieve sound level meter measurements. An example of the comparison between the original input and its reconstruction obtained via VAE is shown in Figures 8(a) and 8(b). The reconstruction is blurred and this is common in VAEs (Neri *et al.*, 2021). The blur does not allow a quantitative analysis through the audio recording. From an energy point of view, the reconstruction has lost resolution in the frequency domain, especially at low frequencies, where the fundamental frequencies of the speech lie. At the same time, low energy areas in the mid and high frequencies (around 3000 and 4000 Hz) show higher amplitudes in the reconstruction with respect to the original spectrogram. Reconstructed samples are highly noisy. Thus, a reconstruction of the sound level meter measurement through the reconstructed spectrograms would not be reliable. However, this loss of information concerns not only the reconstructed data but the original too. The heavy preprocessing needed to obtain a fast network results in low resolution audio samples that cannot be considered reliable for a sound level meter measurement. In other words, the preprocessing step itself adds further uncertainty to the results.



(a) Original sample

(b) Reconstructed sample

FIG. 8. (Color online) Example of original 8(a) and reconstructed 8(b) magnitude spectrograms obtained through the VAE.

VAEs can identify underlying structures of data. With respect to standard autoencoders, they push the latent code to follow a predefined distribution (Min *et al.*, 2018). In the present study, the VAE uses an isotropic Gaussian distribution as prior. The Gaussian representation of the two sound sources is the common thread among GMM, KM, and VAE. The ability of identifying the two sound sources through all the methods used in this work leads to deem reasonable to describe sound sources in long-term monitoring with Gaussian distributions. Further, the goal of these methods is measuring the single contribution of each sound source in mixtures obtained in real-world conditions. Thus, clustering techniques seem to provide more reliable methods than the VAE. This is mainly due to two factors. The first concerns the ability of GMM and KM to perform blind source separation without particular pre-processing steps on the measured data. Analyses are carried out directly on the SPLs occurrences. The second factor concerns the need of deep learning approach of recording audio in work contexts. This can arise privacy issues, one of the most important aspects on the application of big data approaches in real contexts (Kelleher and Tierney, 2018). On the contrary, clustering techniques provide simple and smooth applications for measuring sound environments. As stated in Section IV A 1, GMM can be considered as a generalization of KM. Recollecting the better performance in the step concerning the optimal number of clusters, GMM seems to be the most reliable method to perform blind source separation of sound level meter data. Further studies have to deal with the quantitative aspects of these methods.

V. CONCLUSIONS

In this study, the blind source separation methods carried out via clustering algorithms have been qualitatively validated through a deep learning approach. A dual analysis was performed. The first exploits the occurrence curve of the SPLs through GMM and KM, the second uses the audio recording through a VAE. The goal of both analyses was to separately measure the two main sound sources that describe the sound context inside the office selected as case study: the traffic due to the busy nearby road and the speech of workers.

Clustering algorithms confirmed the robust results obtained in previous works and the reliablity in the separation of spectra in mixtures, identifying both sources. The reliability was assessed through a spectral matching. Relative spectral tendencies in free-field conditions were taken from standards and used as reference curves with respect to the results obtained. Taking into account the experimental conditions, such as reverberation effects on the spectra, it is possible to assess the reliability of cluster analysis in each octave band.

The deep clustering analysis performed by the encoder of the VAE into its latent space was analyzed. The two categories manually labelled were represented by the VAE as two well-defined and separated clusters. Thus, the VAE learned different features from the two sound sources. However, this technique cannot be used to measure the separated sound sources because of the heavy preprocessing on the audio data led to noisy spectrograms.

The ability of measuring sound components in real-world conditions represent an essential issue in sound contexts analyses. The dissection of complex sound environments leads to a deeper understanding of the interactions among sound sources and heavy improvements on the acoustic design processes. The results obtained by the VAE validate the assumptions and the observation made in the assessment of the clustering analyses. However, the validation concerns only the qualitative results. Further studies have to examine the quantitative results obtained by these methods because the goal is to provide a reliable automated analysis of measured data.

Aggarwal, C. C., and Reddy, C. K. (**2014**). "Data clustering," Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra .

Alimohammadi, I., and Ebrahimi, H. (**2017**). "Comparison between effects of low and high frequency noise on mental performance," Applied Acoustics **126**, 131–135.

Bianco, M. J., Gannot, S., Fernandez-Grande, E., and Gerstoft, P. (**2021**). "Semi-supervised source localization in reverberant environments with deep generative modeling," IEEE Access **9**, 84956–84970.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. (**2019**). "Machine learning in acoustics: Theory and applications," The Journal of the Acoustical Society of America **146**(5), 3590–3628.

Bishop, C. M., and Nasrabadi, N. M. (**2006**). *Pattern recognition and machine learning*, **4** (Springer).

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acustica united with Acustica **86**(1), 117–128.

Caliński, T., and Harabasz, J. (**1974**). "A dendrite method for cluster analysis," Communications in Statistics-theory and Methods **3**(1), 1–27.

Can, A., Leclercq, L., Lelong, J., and Botteldooren, D. (**2010**). "Traffic noise spectrum analysis: Dynamic modeling vs. experimental observations," Applied Acoustics **71**(8), 764–770.

Connor, M., Canal, G., and Rozell, C. (**2021**). "Variational autoencoder with learned latent structure," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 2359–2367.

Davies, D. L., and Bouldin, D. W. (**1979**). "A cluster separation measure," IEEE transactions on pattern analysis and machine intelligence (2), 224–227.

De Salvio, D., D'Orazio, D., and Garai, M. (**2021**). "Unsupervised analysis of background noise sources in active offices," The Journal of the Acoustical Society of America **149**(6), 4049–4060.

Dehlbæk, T. S., Brunskog, J., Petersen, C. M., and Marie, P. (**2016**). "The effect of human activity noise on the acoustic quality in open plan offices," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, Vol. 253, pp. 4117–4126.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22.

D'Orazio, D., De Salvio, D., Anderlucci, L., and Garai, M. (**2020**). "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods," Applied Acoustics **169**, 107448.

Ellermeier, W., Eigenstetter, M., and Zimmer, K. (**2001**). "Psychoacoustic correlates of individual noise sensitivity," The journal of the acoustical society of America **109**(4), 1464–1473.

EN 1793-3 (**1997**). "Road traffic noise reducing devices - Test method for determining the acoustic performance - part 3: Normalized traffic spectrum" (European Committee for Standardization).

Green, M., and Murphy, D. (**2020**). "Environmental sound monitoring using machine learning on mobile devices," Applied Acoustics **159**, 107041.

Haapakangas, A., Hongisto, V., and Liebl, A. (**2020**). "The relation between the intelligibility of irrelevant speech and cognitive performance—a revised model based on laboratory studies," Indoor air **30**(6), 1130–1146.

Harvie-Clark, J., Bourdeau, E., Chevret, P., and Brocolini, L. (**2021**). "How will iso 22955 affect designs for open plan offices?," ACOUSTICS 2021 .

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (**2016**). "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 31–35.

Hodgson, M., Rempel, R., and Kennedy, S. (**1999**). "Measurement and prediction of typical speech and background-noise levels in university classrooms during lectures," The Journal of the Acoustical Society of America **105**(1), 226–233.

ISO 16283 - 3 (**2016**). "Acoustics - Field measurement of sound insulation in buildings and of building elements—part 3: Façade sound insulation," (International Organization for Standardization, Geneva, Switzerland).

ISO 22955 (**2021**). "Acoustics - Acoustic Quality of Open Office Spaces," (International Organization for Standardization, Geneva, Switzerland).

ISO 3382 - 2 (**2008**). "Acoustics - Measurement of room acoustic parameters — part 2: Reverberation time in ordinary rooms" (International Organization for Standardization, Geneva, Switzerland).

ISO 3382 - 3 (**2012**). “Acoustics - Measurement of room acoustic parameters — part 3: Open-plan offices” (International Organization for Standardization, Geneva, Switzerland).

Jain, A. K. (**2010**). “Data clustering: 50 years beyond k-means,” Pattern recognition letters **31**(8), 651–666.

Jain, A. K., Murty, M. N., and Flynn, P. J. (**1999**). “Data clustering: a review,” ACM computing surveys (CSUR) **31**(3), 264–323.

Jenkins, W. F., Gerstoft, P., Bianco, M. J., and Bromirski, P. D. (**2021**). “Unsupervised deep clustering of seismic data: Monitoring the ross ice shelf, antarctica,” Journal of Geophysical Research: Solid Earth **126**(9), e2021JB021716.

Kelleher, J. D., and Tierney, B. (**2018**). *Data science* (MIT Press).

Kingma, D. P., and Welling, M. (**2014**). “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Kingma, D. P., Welling, M. *et al.* (**2019**). “An introduction to variational autoencoders,” Foundations and Trends® in Machine Learning **12**(4), 307–392.

Koskela, H., Maula, H., Haapakangas, A., Moberg, V., and Hongisto, V. (**2014**). “Effect of low ventilation rate on office work performance and perception of air quality–a laboratory study,” Proceedings of Indoor Air 673–675.

Kullback, S., and Leibler, R. A. (**1951**). “On information and sufficiency,” The annals of mathematical statistics **22**(1), 79–86.

LeCun, Y., Bengio, Y., and Hinton, G. (**2015**). “Deep learning,” nature **521**(7553), 436–444.

Leembruggen, G., Verhave, J., Feistel, S., Holtzem, L., Mapp, P., Sato, H., Steinbrecher, T., and Van Wijngaarden, S. (**2016**). "The effect on sti results of changes to the male test-signal spectrum," Proc. IOA **38**, 78–87.

Leglaive, S., Girin, L., and Horaud, R. (**2019**). "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 101–105.

Leonard, P., and Chilton, A. (**2019**). "The lombard effect in open plan offices," Proceedings of the Institute of Acoustics, Milton Keynes, United Kingdom .

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (**2013**). "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, Vol. 2013, pp. 436–440.

MacKay, D. J. (**2003**). *Information theory, inference and learning algorithms* (Cambridge university press).

McLachlan, G. J., and Peel, D. (**2004**). *Finite mixture models* (John Wiley & Sons).

Merchant, N. D., Barton, T. R., Thompson, P. M., Pirotta, E., Dakin, D. T., and Dorocicz, J. (**2013**). "Spectral probability density as a tool for ambient noise analysis," The Journal of the Acoustical Society of America **133**(4), EL262–EL267.

Merchant, N. D., Fristrup, K. M., Johnson, M. P., Tyack, P. L., Witt, M. J., Blondel, P., and Parks, S. E. (**2015**). "Measuring acoustic habitats," Methods in Ecology and Evolution **6**(3), 257–265.

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (**2018**). "A survey of clustering with deep learning: From the perspective of network architecture," IEEE Access **6**, 39501–

39514.

Mitchell, T. M. (**1997**). *Machine Learning* (McGraw-Hill, New York).

Neri, J., Badeau, R., and Depalle, P. (**2021**). "Unsupervised blind source separation with variational auto-encoders," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 311–315.

NF S31-199 (**2016**). "Acoustique - Performances acoustiques des espaces ouverts de bureau (Acoustics - Acoustic performance for open-plan offices)" (Association française de normalisation, France).

Olsen, W. O. (**1998**). "Average speech levels and spectra in various speaking/listening conditions," American Journal of Audiology **7**(2), 21–25.

Ozanich, E., Thode, A., Gerstoft, P., Freeman, L. A., and Freeman, S. (**2021**). "Deep embedded clustering of coral reef bioacoustics," The Journal of the Acoustical Society of America **149**(4), 2587–2601.

Parks, S. E., Urazghildiiev, I., and Clark, C. W. (**2009**). "Variability in ambient noise levels and call parameters of north atlantic right whales in three habitat areas," The Journal of the Acoustical Society of America **125**(2), 1230–1239.

Rindel, J. H., Christensen, C. L., and Gade, A. C. (**2012**). "Dynamic sound source for simulating the lombard effect in room acoustic modeling software," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, Vol. 2012, pp. 954–966.

Rousseeuw, P. J. (**1987**). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics **20**, 53–65.

Tibshirani, R., Walther, G., and Hastie, T. (**2001**). "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(2), 411–423.

Van der Maaten, L., and Hinton, G. (**2008**). "Visualizing data using t-sne.," Journal of machine learning research **9**(11).

Vincent, E., Virtanen, T., and Gannot, S. (**2018**). *Audio source separation and speech enhancement* (John Wiley & Sons).

Wang, L. M., and Brill, L. C. (**2021**). "Speech and noise levels measured in occupied k–12 classrooms," The Journal of the Acoustical Society of America **150**(2), 864–877.

Yadav, M., Cabrera, D., Kim, J., Fels, J., and de Dear, R. (**2021**). "Sound in occupied open-plan offices: Objective metrics with a review of historical perspectives," Applied Acoustics **177**, 107943.