



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

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

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

De Salvo, 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>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

**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 Garai¹

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

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

^adomenico.desalvio2@unibo.it

16 I. INTRODUCTION

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

28 Workplaces are one of the most lived-in spaces by people. The achievement of a comfort-
29 able environment is important for both well-being and productivity. In offices, the latter
30 are deeply influenced by noises. Individual perceptions can be affected by the nature of
31 sound (Koskela *et al.*, 2014). The performances can either decrease or increase. It has been
32 shown that both high or low frequency noises can improve cognitive tasks (Alimohammadi
33 and Ebrahimi, 2017). The most important factor related to workers' comfort concerns the
34 speech intelligibility. Thus, the most distracting noise for workers is represented by col-
35 leagues' speech (Ellermeier *et al.*, 2001; Haapakangas *et al.*, 2020). The NF S31-199 and
36 the ISO 22955:2021 highlight the importance of assessing the noise levels at workstations

37 depending on the activity carried out in the office. In particular, the ISO provides a survey
38 for employees to rank the level of annoyance of several noise sources. This new approach
39 deeply affects the design of open-plan offices ([Harvie-Clark *et al.*, 2021](#); [ISO 22955, 2021](#); [NF
40 S31-199, 2016](#)). Thus, the ability of separating the noise sources is fundamental. However,
41 there is a lack of ability in the technical praxis about measuring sound sources in real-world
42 contexts.

43 Measurement techniques based on the statistical probability densities of SPLs were used
44 to monitor noise contributions in classrooms. These methods are based on machine learning
45 algorithms. Machine learning is the study of algorithms that improve their performance
46 through the experience ([Mitchell, 1997](#)). The applications of these techniques frequently
47 involve statistical methods and their use is rapidly increasing in acoustics ([Bianco *et al.*,
48 2019](#)). Long-term monitoring allow to collect large amounts of data. Thus, sound level meter
49 measurements can be exploited using statistical methods. To analyze the collected SPLs,
50 clustering techniques can find pattern among data. The multimodal SPLs' occurrences curve
51 were exploited via Gaussian Mixture Model and K-means clustering to separately assess the
52 noise due to the HVAC systems, the noise produced by the students, the teachers' speech,
53 and the signal-to-noise ratios ([D'Orazio *et al.*, 2020](#); [Hodgson *et al.*, 1999](#); [Wang and Brill,
54 2021](#)). An application of the Gaussian Mixture Model in five offices was made to measure the
55 human activity noise levels ([Dehlbæk *et al.*, 2016](#)). Then, two algorithms were proposed as
56 unsupervised methods to separate and identify the mechanical noise and the human activity
57 during working hours ([De Salvio *et al.*, 2021](#)).

58 Blind source separation is a major issue addressed not only in machine learning but in
59 deep learning too. This is a type of machine learning based on artificial neural networks
60 that learns representations of data with multiple level of abstraction (LeCun *et al.*, 2015).
61 Inspired by the cocktail party effect, i.e. the ability of humans to focus the auditory attention
62 to one speaker filtering other stimuli (Bronkhorst, 2000), the need of extracting the single
63 source from a mixture of signals lies in many useful applications such as speech, music, and
64 environmental audio processing (Vincent *et al.*, 2018). In the framework of the acoustic
65 source separation, the concept of *deep clustering* was introduced. Deep clustering refers
66 to the ability of performing clustering through deep learning algorithms (Hershey *et al.*,
67 2016). One of the most popular category to perform this is represented by the autoencoders.
68 These kind of network performs a non-linear mapping of the data through an encoder and
69 a decoder. The first maps the function to be trained, the second learns how to reconstruct
70 the original data (Min *et al.*, 2018). Applications of autoencoders in acoustics concerned
71 speech enhancement and clustering of geophysical data (Jenkins *et al.*, 2021; Lu *et al.*, 2013;
72 Ozanich *et al.*, 2021).

73 A variational autoencoder is a deep generative model that forces the latent code of autoen-
74 coders to follow a predefined distribution (Min *et al.*, 2018). It has the same architecture
75 of autoencoders, high-dimensional data are encoded into a low-dimensional latent space
76 (Kingma and Welling, 2014). The ability of parametrizing data through a probability distri-
77 butions gained broad attention in the deep learning community. Successful applications of
78 variational autoencoders concern speech enhancement, blind source separation, and sound

79 source localization in reverberant spaces (Bianco *et al.*, 2021; Leglaive *et al.*, 2019; Neri
80 *et al.*, 2021).

81 The present work deals with the blind source separation through a sound level meter
82 long-term monitoring. Basing on the methods proposed in previous work (De Salvio *et al.*,
83 2021), a dual analysis of the same phenomenon is proposed. A sound level meter recorded
84 both the sound pressure levels and the digital audio of the working activity inside an office.
85 Then, two clustering analyses were performed. The first exploited the two machine learning
86 algorithms earlier mentioned, i.e. the Gaussian Mixture Model and the K-means clustering;
87 the second performed a deep clustering analysis through a variational autoencoder. The
88 goal is to identify and separately measure the main sound sources experienced by workers
89 during the activity with both approaches.

90 II. THEORETICAL BACKGROUND

91 A. Clustering techniques

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

98 **1. Gaussian Mixture Model**

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

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

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

$$0 \leq \pi_k \leq 1 \quad (k = 1, \dots, K) \quad (2)$$

105

$$\sum_{k=1}^K \pi_k = 1. \quad (3)$$

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

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

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

112 **2. K-means clustering**

113 K-means clustering (KM) is a distance-based clustering technique (Bishop and Nasrabadi,

114 2006). It aims to shape a number of K clusters given a set of independent observations
115 $X = \{x_1, \dots, x_N\}$. Data are gathered minimizing the squared error Euclidean distance
116 between the empirical mean c_{k_i} , called *centroid*, of a cluster k_i and the data points in the
117 cluster. The squared error J is defined as:

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

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

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

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

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

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

125 B. Model selection

126 An important issue in data clustering concerns the optimal number of clusters in data.
127 For some classes of algorithms, such as GMM and KM, the number of clusters has to be
128 specified before running the iterative process. Estimating the number of clusters is an open
129 problem (Aggarwal and Reddy, 2014). Several metrics allow to find the most likely number

130 of clusters with different approaches. Here, four metrics were used to assess the models'
 131 number of components, i.e. sound sources, in the collected data, as next.

132 **1. Davies-Bouldin**

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

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

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

136 where

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

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

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

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

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

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

142 **2. Gap statistic**

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

149 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
 150 k_i . The within-cluster dispersion is defined as:

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

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

$$D_{k_i} = \sum_{x_{k_i}, x_{k'_i} \in k_i} d_{x_{k_i}, x_{k'_i}}. \quad (12)$$

152 the pairwise distances of all points in the cluster k_i .

153 Then, the Gap value is defined as:

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

154 where \mathbb{E}_r^* is the expectation under a sample size r from the reference distribution. In
 155 the present study, the expected within-cluster dispersion of the reference distribution is
 156 evaluated via Monte Carlo sampling. The reference distribution is represented by a uniform
 157 distribution. The optimal model is represented by the highest value obtained in equation
 158 13.

159 **3. Calinski-Harabasz**

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

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

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

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

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

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

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

$$CH = \frac{SS_B}{SS_W} \frac{N - K}{K - 1} \quad (16)$$

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

173 **4. Silhouette coefficient**

174 The silhouette coefficient is a graphical quantitative evaluation of the degree of separation

175 among clusters (Rousseeuw, 1987). Given two data points x_{k_i} and $x_{k_{i'}}$ in the cluster k_i , the
 176 within-cluster mean distance, i.e. the similarity, between x_{k_i} and the other $x_{k_{i'}}$ th points in
 177 the same cluster is defined as:

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

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

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

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

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

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

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

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

189 C. Variational Autoencoder

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

198 The encoder is represented by a neural network. Its aim is to output a latent hidden
199 representation z of the input x with weights and biases θ . Typically, the latent space has a
200 lower dimension with respect to the input size. Thus, it can be deduced that the encoder
201 learns a compressed representation of the input data according to the distribution $q_\theta(z|x)$.
202 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
203 distribution $q_\theta(z|x)$ is represented by a Gaussian probability density.

204 The decoder is a neural network as well. Typically, it has a mirrored architecture of the
205 encoder. Its aim is to reconstruct the input sampling only from the compressed representa-
206 tion of the latent space z . Thus, it outputs parameters to the probability distribution of data
207 with weight and biases ϕ . The decoder process is denoted by the distribution $p_\phi(x|z)$. The
208 latter is represented by a standard Normal distribution $\mathcal{N}(0, 1)$ with mean 0 and variance
209 1.

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

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

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

220 III. EXPERIMENTAL SETUP

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

229 of the office and the arrangement of the workstations besides the placement of the sound
230 level meter. The sound level meter collected about 6 hours of working activity in the office.

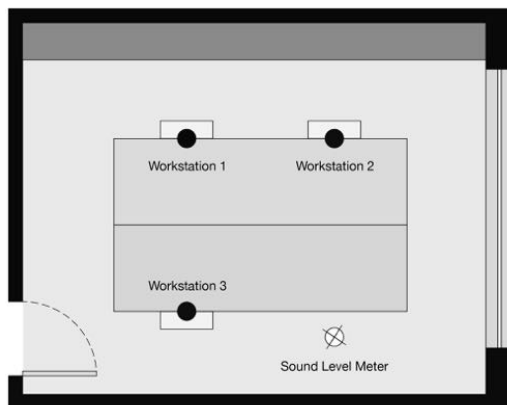


FIG. 1. (Color online) Plan of the office under study.

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

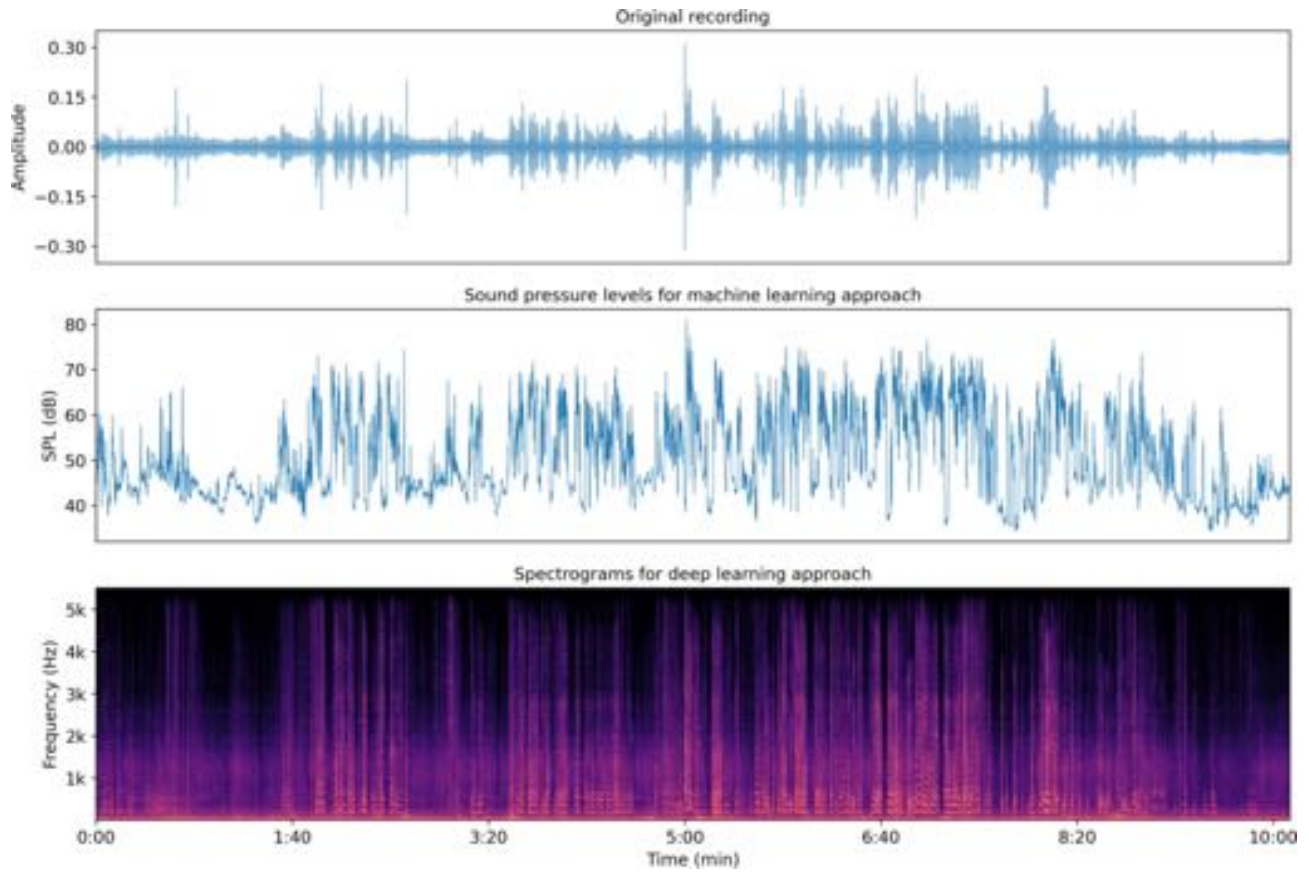


FIG. 2. (Color online) Example of the data used in this study. On the top, a sample of 10 minutes recording. In the middle, the sound pressure levels obtained in the same sample. This constitutes one of the databases for the machine learning approach. On the bottom, the spectrograms obtained by the same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

240 A. GMM and KM analyses

241 The procedure of the clustering analyses via GMM and KM follows the same flow de-
 242 scribed in a previous work (De Salvio *et al.*, 2021). Once the distribution of the SPLs
 243 occurrences is obtained, the number of clusters to look for in the collected data has to be
 244 set first. Thus, models with 2 up to 10 components, i.e. sound sources, for both GMM

245 and KM were used as candidates. Each metric used for searching the most likely number of
246 clusters finds similarity among clusters according to its own approach. Hence, the metric's
247 highest value represents the best model for Silhouette, Calinski-Harabasz, and Gap statistic,
248 while the metric's lowest value represents the best choice for Davies-Bouldin coefficient. In
249 this study, the the majority rule was used to obtain the optimal number of clusters. Thus,
250 the most frequent number obtained comparing each metric represents the number of active
251 sound sources during the event.

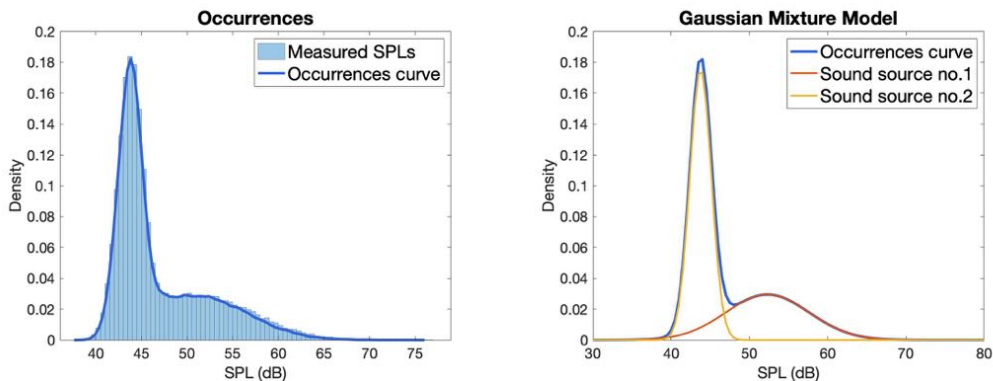
252 Figure 3 shows the type of distribution used for the clustering. On the left, the SPLs
253 occurrences are collected and analyzed through the normalized occurrences distribution. In
254 the middle, an example of processing via GMM and on the right, an example of processing
255 via KM.

256 Following a brief summary of the procedure:

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

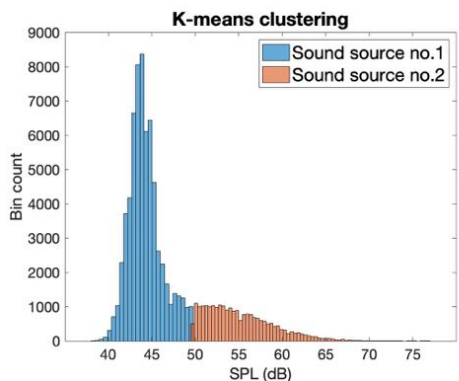
261 The three steps are applied for each different clustering algorithm we want to use. Basing
262 on the acoustic task, any SPLs' frequency band can be used to complete the process. In
263 this study, the whole procedure is carried out in the range from 125 up to 4000 Hz octave
264 bands. Hence, each occurrences' curve is analyzed looking for the most likely number of
265 clusters. Both GMM and KM were set to repeat the iterative process with different initial

266 values. In general, using different starting points typically results in a solution that is a
 267 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).

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

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

285 **B. VAE analysis**

286 The digital audio recording has been chunked in 1-second length samples to obtain the
287 dataset for the analysis through the VAE. Power spectrograms of each chunk were used
288 as input for the network. A pre-processing procedure has been carried out before feeding
289 the encoder. The audio has been resampled at 11025 Hz to make the input comparable to
290 the octave band range (125-4000 Hz) used in the cluster analysis. Moreover, visualizing the
291 spectrograms, no useful information were found above 5 kHz. Short-time Fourier Transforms
292 (STFT) with a segment length of $N_{\text{FFT}} = 256$ and an overlap area of 50% were used to obtain

293 the spectrograms. With these values each audio chunk is processed with an FFT with a
 294 physical length of about 20 ms. Thus, it can be deemed that in each FFT only one sound
 295 source is detected. A minMAX normalization has been applied to each spectrogram to have
 296 all the amplitude values in $[0,1]$ range. Overall, the dataset contained about 23k samples.

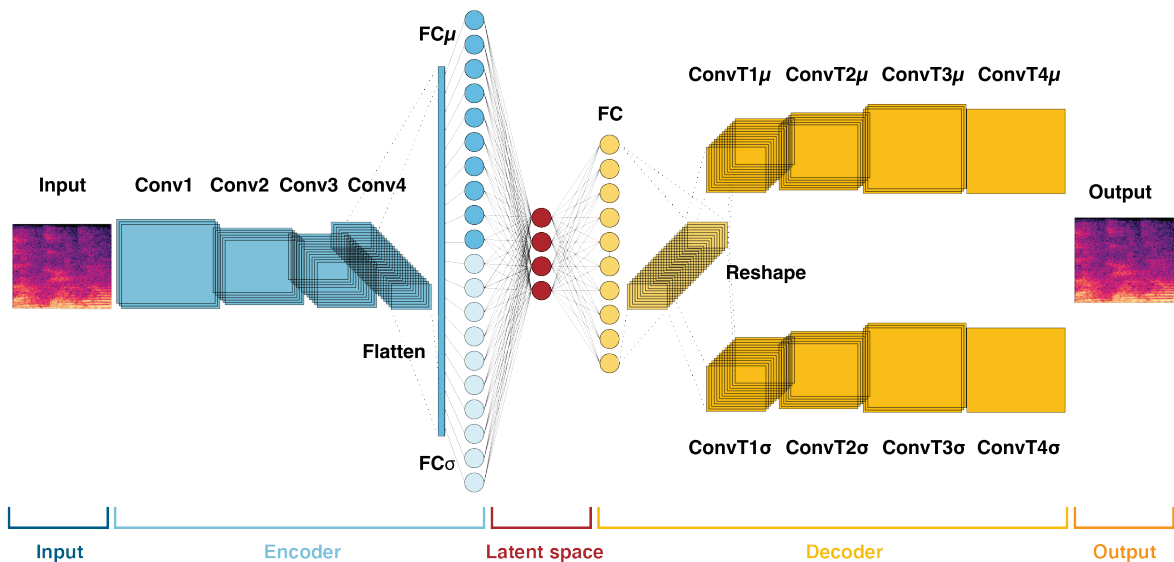


FIG. 4. (Color online) Architecture of the VAE. The encoder is constituted by four convolutional layers, shown in light blue. The latent space is shown in red and the decoder is represented by the yellow blocks.

297 Samples of 1-second length can be easily listened. Then, the dataset has been manually
 298 labelled listening each sample of the recording in three classes: traffic, speech, and unclas-
 299 sified sounds. The latter category was useful to label all the samples where the main sound
 300 source was represented by impulsive noises like slammed doors. It is worth noting that,
 301 during the labelling process, audio chunks containing only whispers were labelled as speech.
 302 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]

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

306 The VAE was built in Pytorch. The input size of the spectrograms is 128×87 . The
307 encoder is made by four strided convolutional layers (stride = 2). Then, a flatten layer
308 allows the convolutional layers to be linked to the fully connected layers. As highlighted in
309 Section II C, a VAE maps the input to a multivariate latent distribution. The distribution
310 used in the present work is the Gaussian distribution. Parameters of the distribution, the
311 mean and standard deviation, i.e. μ and σ , are the outputs of the encoder. For this
312 reason, the fully connected layer of the encoder is doubled. Here are the inputs for the
313 30-dimensional latent space. The decoder uses a different distribution, the prior on the
314 latent distribution. The architecture of the decoder depends on the parameters needed to
315 specify the multivariate generative distribution. In the present work, two parameters are
316 needed: the mean and the variance. Thus, the decoder is made by four strided transposed
317 convolutional layers (stride = 2) for both parameters. The Pyro library was used for the
318 stochastic variational inference. Then, spectrograms are reconstructed reshaping the output
319 of mean and variances obtained by the decoder. Non-linearities are activated through *ReLU*
320 functions for all layers except for the output parameters. The decoder is parametrized
321 according to a standard Normal distribution $\mathcal{N}(0, 1)$. Thus, a *Tanh* activation function
322 is used for the output of means and a *Sigmoid* activation function for the the output of
323 variances. The VAE was trained using a batch size of 32, the Adam optimizer and θ and
324 ϕ weights were updated with a learning rate equal to 1×10^{-5} . Figure 4 shows a graphical

325 scheme of the VAE’s architecture. The light blue and yellow layers represent, respectively,
326 the encoder and the decoder. Both are linked by the latent space represented with the
327 red fully connected layer. Details about the architecture of the whole network are listed in
328 Table I. Here, the type, the input size, the number of filters, the kernel size, the activation
329 functions, and the output size are shown for each layer. Training stopped after 400 epochs
330 since not relevant improvements of the loss function on the test dataset were detected.

331 IV. RESULTS

332 A. Machine Learning results

333 1. Source separation via GMM and KM

334 Previous work (De Salvo *et al.*, 2021) used only one metric per each algorithm to assess
335 the optimal number of cluster. Moreover, the elbow technique made the analysis influenced
336 by the operator’s choice. Thus, the step 2 described in section III A is different. The
337 comparison among different metrics makes the analysis more robust and inclined to the
338 automation. Table II shows the results of model selection. Silhouette (SC), Davies-Bouldin
339 (DB), Gap statistic (GS), and Calinski-Harabasz (CH) coefficients were used to assess the
340 most likely number of clusters for each octave band (125-4000 Hz) and the A-weighted
341 continuous level $L_{A,eq}$. Concerning GMM, the model selection metrics found that the optimal
342 number of clusters is equal 2 according to the majority rule. This is true for SC and DB
343 for each octave band and $L_{A,eq}$. Different results were found only for GS in the 125 Hz
344 octave band and for CH in the 500 and 4000 Hz octave bands. The same analysis was

345 carried out for KM. Here, SC, DB, and GS found an optimal number of clusters equal to 2
346 for each occurrences curve analyzed. Completely different results were shown by CH that
347 found 6 clusters in each octave band and $L_{A,eq}$ as the best model. Overall, comparing all
348 metrics, the number of active sources in the office is 2. These results are consistent with
349 the expectations. The main sound sources experienced during a common working day by
350 employees were speech and traffic, indeed.

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

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

363 Concerning the traffic noise, the reference is represented by the normalized spectrum
364 shown in EN 1793-3 (EN 1793-3, 1997). It is worth noting that the reference spectrum
365 refers to free field conditions. Thus, acoustical properties of the room and the facade's
366 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

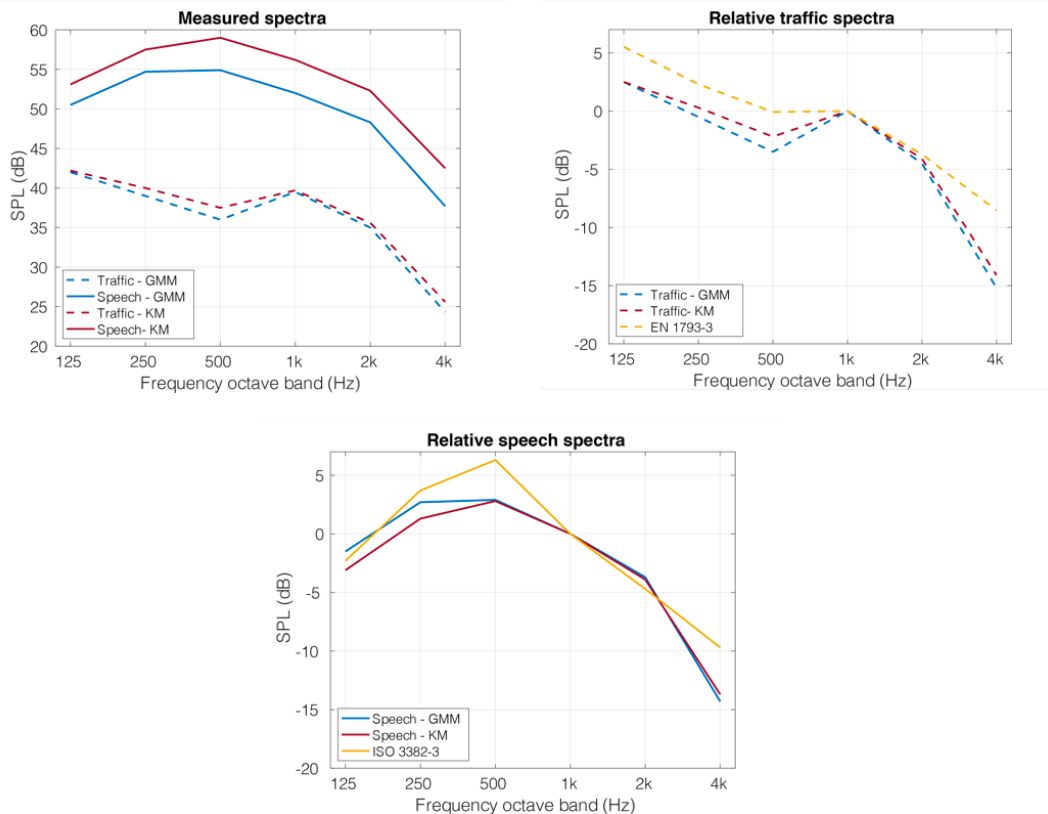


FIG. 5. (Color online) Results of clustering analyses. On the top: reconstruction of the spectra from 125 up to 4000 Hz. Blue and red lines represent the spectra reconstructed respectively via GMM and KM. Dashed and solid lines represent respectively the traffic and the speech spectra. In the middle 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.

367 very similar. The most noticeable difference concerns the 500 Hz octave band. However,
 368 both low-frequencies emitted at slow speeds and the 1 kHz frequencies emitted at free-flow
 369 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

370 The ISO 3382-3 shows the reference speech spectrum of a directional source at a distance
371 of 1 m in free field from the speaker (ISO 3382 - 3, 2012). This is the reference for the
372 speech source; the related spectra obtained via clustering have similar tendencies as shown
373 on the bottom of Figure 5. Differences can be referred to several factors. The first concerns

374 the influence of the acoustical properties of the room. As noticed for the traffic noise, the
375 ISO spectrum is evaluated at a distance of 1 m from the source in free field. As opposed
376 to previous work, slight differences concern low frequencies for speech sources. However,
377 these can be due to the change of the spectrum in noisy environments and the measurement
378 uncertainty at low frequencies, especially in the 125 and 250 Hz octave bands ([Leembruggen](#)
379 [et al., 2016](#); [Rindel et al., 2012](#)). Directivity of the source can affect spectra tendencies
380 too. In the present study, there are 3 speakers in 3 different positions. Thus, the overall
381 directivity of the measured source cannot be considered the same as the reference. Moreover,
382 at low frequencies, modal effects could have affected the results since the sound level meter
383 was used only in one position.

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

387 Further considerations can be made regarding the clusters size. This is described by the
388 SD and the AICD; both are shown in brackets in [Table III](#). The physical meaning associated
389 to SD and AICD is the temporal randomness of the source. Mechanical sources produce
390 the same SPLs occurrences depending on their mechanical cycle, indeed. This results in low
391 SDs for continuous sources because the corresponding Gaussian curve will be narrow. On
392 the contrary, a human-related noise produce higher SDs. The traffic noise can be deemed in
393 the middle of these two categories of noise sources. It does not have the same continuity of a
394 mechanical device but it has specific spectral properties. Moreover, the road has to be busy
395 to be detected in a long-term monitoring because the occurrences curve has to be affected

396 by the noise source. Thus, traffic can be deemed more continuous than the speech but not
397 like a mechanical source. These considerations are confirmed by the results obtained. Traffic
398 SDs lie in the range 3.0 - 4.3 dB for each octave band. Previous work showed mechanical
399 SDs due to the HVAC system in the range 0.9 - 3.9 dB. Thresholds analyses deserve detailed
400 studies in future works. However, all non-human sound sources were confirmed to be under
401 the threshold of 5 dB (De Salvio *et al.*, 2021).

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

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

415 As noted in previous section IV A 1, the acoustical properties of the office influence the
416 spectra obtained via GMM and KM. Thus, the reverberation time T_{20} and the façade

417 sound level difference $D_{2m,nT}$ were measured respectively according to the precision method
 418 described in the ISO 3382-2 and the global method of the ISO 16283-3 (ISO 16283 - 3, 2016;
 419 ISO 3382 - 2, 2008). Measurements' results are shown in Figure 6. Solid and dashed lines
 420 show respectively the T_{20} and $D_{2m,nT}$ tendencies in octave bands from 125 up to 4000 Hz.

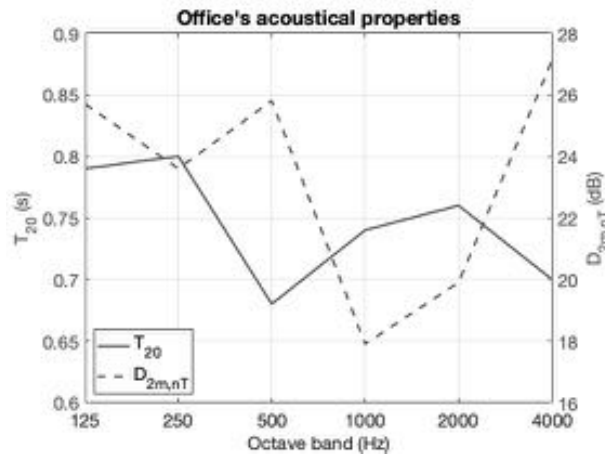


FIG. 6. Acoustical properties of the office under study. The reverberation time T_{20} is shown on the left axis, the façade sound level difference $D_{2m,nT}$ on the right axis.

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

427 The tendencies of measurements' results in Figure 6 may bring preliminary insights about
 428 the comparison of measured and reference spectra shown in Figure 5. Traffic and speech
 429 spectra seem to be related to the tendency of the T_{20} . In fact, the drop in the 500 Hz octave

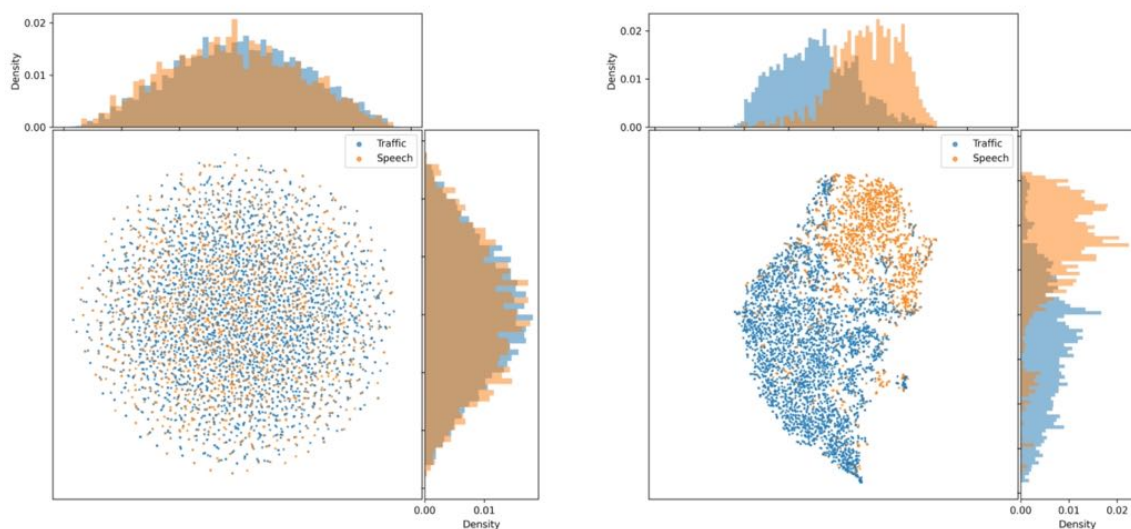
430 band is visible in both sources. Further, the reverberation time has its minimum value in
431 the same band, as well as one of the highest values of the façade insulation. The energy of
432 both sources in the 4 kHz octave band seems to be affected respectively by the T_{20} for the
433 speech and by the $D_{2m,nT}$ for the traffic. Thus, a preliminary analysis of the room’s acoustics
434 seems to support the results obtained through the machine learning approach.

435 **B. Deep learning results**

436 *1. Latent space*

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

448 Figures 7(a) and 7(b) show the latent distributions of respectively the untrained and
449 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.

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

455 Data in the latent space are represented basing on their categorical label. The untrained
 456 latent space in Figure 7(a) shows a circular distribution of data since it is perfectly described
 457 by a Gaussian distribution (Connor *et al.*, 2021). However, there is no categorical separation
 458 among data, i.e. blue and orange dots are mixed up. Figure 7(b) shows the results of the
 459 training. After the network has learnt the latent representation of the input data, the latent

460 space shows a clear separation of the two categories. Clusters are well-defined. On the sides,
461 histograms show the 1D projection of the plot along the main axes. These distributions help
462 to assess whether the two clusters in the 2D plot are overlapped or not. In the present case,
463 histograms of the trained network show that the two clusters are close but not overlapped.
464 Thus, clusters are well-separated too. The VAE is able to identify and separate the two
465 sound sources through a Gaussian latent space. Different densities within clusters may refer
466 to further properties, e.g. timbre, not considered in the categories taken into account in this
467 study. Uncertainties on data distributions, i.e. speech frames in the traffic cluster and vice
468 versa, can be attributed to the manual labelling. For instance, whispers can be manually
469 labelled as speech but classified by the network as traffic.

470 2. *The reconstruction of the spectrograms*

471 The aim of these approaches is to measure different sound sources. Thus, the reconstruc-
472 tion of the audio recording can be post-processed to achieve sound level meter measurements.
473 An example of the comparison between the original input and its reconstruction obtained
474 via VAE is shown in Figures 8(a) and 8(b). The reconstruction is blurred and this is com-
475 mon in VAEs (Neri *et al.*, 2021). The blur does not allow a quantitative analysis through
476 the audio recording. From an energy point of view, the reconstruction has lost resolution in
477 the frequency domain, especially at low frequencies, where the fundamental frequencies of
478 the speech lie. At the same time, low energy areas in the mid and high frequencies (around
479 3000 and 4000 Hz) show higher amplitudes in the reconstruction with respect to the original
480 spectrogram. Reconstructed samples are highly noisy. Thus, a reconstruction of the sound

481 level meter measurement through the reconstructed spectrograms would not be reliable.
 482 However, this loss of information concerns not only the reconstructed data but the original
 483 too. The heavy preprocessing needed to obtain a fast network results in low resolution audio
 484 samples that cannot be considered reliable for a sound level meter measurement. In other
 485 words, the preprocessing step itself adds further uncertainty to the results.

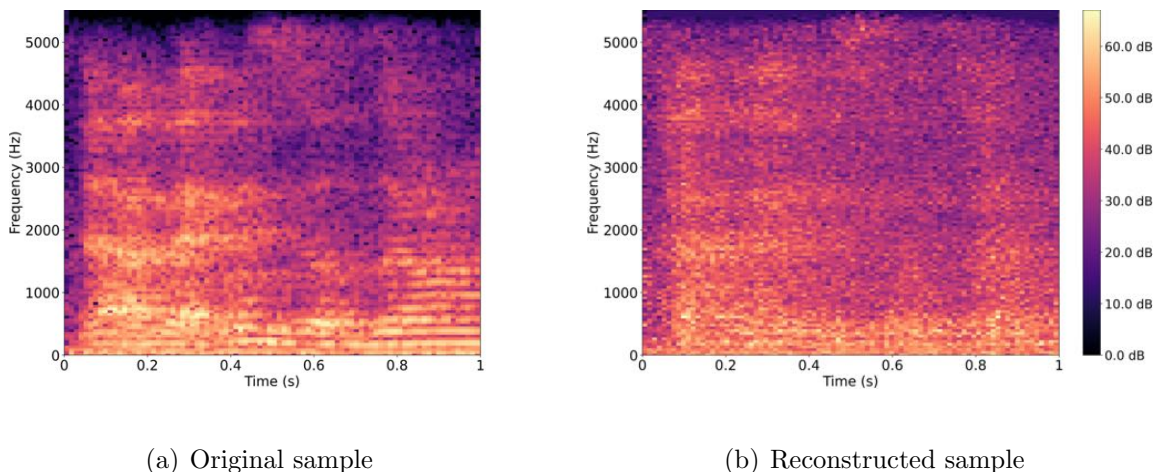


FIG. 8. (Color online) Example of original 8(a) and reconstructed 8(b) magnitude spectrograms obtained through the VAE.

486 VAEs can identify underlying structures of data. With respect to standard autoencoders,
 487 they push the latent code to follow a predefined distribution (Min *et al.*, 2018). In the
 488 present study, the VAE uses an isotropic Gaussian distribution as prior. The Gaussian
 489 representation of the two sound sources is the common thread among GMM, KM, and VAE.
 490 The ability of identifying the two sound sources through all the methods used in this work
 491 leads to deem reasonable to describe sound sources in long-term monitoring with Gaussian
 492 distributions. Further, the goal of these methods is measuring the single contribution of each
 493 sound source in mixtures obtained in real-world conditions. Thus, clustering techniques seem

494 to provide more reliable methods than the VAE. This is mainly due to two factors. The first
495 concerns the ability of GMM and KM to perform blind source separation without particular
496 pre-processing steps on the measured data. Analyses are carried out directly on the SPLs
497 occurrences. The second factor concerns the need of deep learning approach of recording
498 audio in work contexts. This can arise privacy issues, one of the most important aspects on
499 the application of big data approaches in real contexts (Kelleher and Tierney, 2018). On the
500 contrary, clustering techniques provide simple and smooth applications for measuring sound
501 environments. As stated in Section IV A 1, GMM can be considered as a generalization
502 of KM. Recollecting the better performance in the step concerning the optimal number of
503 clusters, GMM seems to be the most reliable method to perform blind source separation of
504 sound level meter data. Further studies have to deal with the quantitative aspects of these
505 methods.

506 V. CONCLUSIONS

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

513 Clustering algorithms confirmed the robust results obtained in previous works and the
514 reliability in the separation of spectra in mixtures, identifying both sources. The reliability

515 was assessed through a spectral matching. Relative spectral tendencies in free-field condi-
516 tions were taken from standards and used as reference curves with respect to the results
517 obtained. Taking into account the experimental conditions, such as reverberation effects on
518 the spectra, it is possible to assess the reliability of cluster analysis in each octave band.

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

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

532

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

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

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

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

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

545 Bronkhorst, A. W. (2000). “The cocktail party phenomenon: A review of research on
546 speech intelligibility in multiple-talker conditions,” *Acta Acustica united with Acustica*
547 **86**(1), 117–128.

548 Caliński, T., and Harabasz, J. (1974). “A dendrite method for cluster analysis,” *Communi-*
549 *nications in Statistics-theory and Methods* **3**(1), 1–27.

550 Can, A., Leclercq, L., Lelong, J., and Botteldooren, D. (2010). “Traffic noise spectrum
551 analysis: Dynamic modeling vs. experimental observations,” *Applied Acoustics* **71**(8),
552 764–770.

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

556 Davies, D. L., and Bouldin, D. W. (1979). “A cluster separation measure,” *IEEE transac-*
557 *tions on pattern analysis and machine intelligence* (2), 224–227.

558 De Salvio, D., D’Orazio, D., and Garai, M. (2021). “Unsupervised analysis of background
559 noise sources in active offices,” *The Journal of the Acoustical Society of America* **149**(6),
560 4049–4060.

561 Dehlbæk, T. S., Brunskog, J., Petersen, C. M., and Marie, P. (2016). “The effect of human
562 activity noise on the acoustic quality in open plan offices,” in *INTER-NOISE and NOISE-*
563 *CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, Vol.
564 253, pp. 4117–4126.

565 Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). “Maximum likelihood from in-
566 complete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B*
567 (Methodological) **39**(1), 1–22.

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

571 Ellermeier, W., Eigenstetter, M., and Zimmer, K. (2001). “Psychoacoustic correlates of
572 individual noise sensitivity,” *The journal of the acoustical society of America* **109**(4),
573 1464–1473.

574 EN 1793-3 (1997). “Road traffic noise reducing devices - Test method for determining the
575 acoustic performance - part 3: Normalized traffic spectrum” (European Committee for
576 Standardization).

577 Green, M., and Murphy, D. (2020). “Environmental sound monitoring using machine learn-
578 ing on mobile devices,” *Applied Acoustics* **159**, 107041.

579 Haapakangas, A., Hongisto, V., and Liebl, A. (2020). “The relation between the intelligibil-
580 ity of irrelevant speech and cognitive performance—a revised model based on laboratory
581 studies,” *Indoor air* **30**(6), 1130–1146.

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

584 Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). “Deep clustering: Dis-
585 criminative embeddings for segmentation and separation,” in *2016 IEEE International
586 Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 31–35.

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

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

593 ISO 22955 (2021). “Acoustics - Acoustic Quality of Open Office Spaces,” (International
594 Organization for Standardization, Geneva, Switzerland).

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

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

600 Jain, A. K. (2010). “Data clustering: 50 years beyond k-means,” *Pattern recognition letters*
601 **31**(8), 651–666.

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

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

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

608 Kingma, D. P., and Welling, M. (2014). “Auto-encoding variational bayes,” in *2nd Interna-*
609 *tional Conference on Learning Representations, ICLR 2014 - Conference Track Proceed-*
610 *ings*.

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

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

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

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

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

622 Leglaive, S., Girin, L., and Horaud, R. (2019). “Semi-supervised multichannel speech
623 enhancement with variational autoencoders and non-negative matrix factorization,” in
624 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Pro-*
625 *cessing (ICASSP)*, IEEE, pp. 101–105.

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

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

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

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

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

636 Merchant, N. D., Fristrup, K. M., Johnson, M. P., Tyack, P. L., Witt, M. J., Blondel, P., and
637 Parks, S. E. (2015). “Measuring acoustic habitats,” *Methods in Ecology and Evolution*
638 **6**(3), 257–265.

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

641 39514.

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

643 Neri, J., Badeau, R., and Depalle, P. (2021). “Unsupervised blind source separation with
644 variational auto-encoders,” in *2021 29th European Signal Processing Conference (EU-*
645 *SIPCO)*, IEEE, pp. 311–315.

646 NF S31-199 (2016). “Acoustique - Performances acoustiques des espaces ouverts de bu-
647 reau (Acoustics - Acoustic performance for open-plan offices)” (Association française de
648 normalisation, France).

649 Olsen, W. O. (1998). “Average speech levels and spectra in various speaking/listening con-
650 ditions,” *American Journal of Audiology* **7**(2), 21–25.

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

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

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

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

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

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

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

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

672 Yadav, M., Cabrera, D., Kim, J., Fels, J., and de Dear, R. (2021). “Sound in occupied open-
673 plan offices: Objective metrics with a review of historical perspectives,” *Applied Acoustics*
674 **177**, 107943.