Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

MECCANICA E SCIENZE AVANZATE DELL'INGEGNERIA

Ciclo 35

Settore Concorsuale: 09/C2 – FISICA TECNICA E INGEGNERIA NUCLEARE

Settore Scientifico Disciplinare: ING-IND/11 – FISICA TECNICA AMBIENTALE

ADVANCED STATISTICAL MODELS FOR THE SEGREGATION, IDENTIFICATION AND MEASUREMENT OF COEXISTING SOUND SOURCES

Presentata da: Domenico De Salvio

Coordinatore Dottorato

Lorenzo Donati

Supervisore

Prof. Massimo Garai

Co-supervisori

Prof. Dario D'Orazio
Prof. Gian Luca Morini

Esame finale anno 2023

All models are wrong but some are useful
George Box

Sommario

L'interesse per i monitoraggi a lungo termine sta stabilmente crescendo grazie alla rilevante quantità di informazioni scientifiche e ingegneristiche che forniscono. Questo è dovuto alla crescita costante della capacità di immagazzinare e alla potenza computazionale della tecnologia per elaborare grandi quantità di dati. In questo scenario, il machine learning (ML) fornisce un'ampia famiglia di metodi statistici basati sui dati per processare ampi database. I metodi basati su ML permettono di identificare ed esplorare schemi complessi all'interno dei dati. Le analisi riguardanti il ML possono riguardare la ricerca di relazioni fra le caratteristiche e le etichette dei dati o fra le caratteristiche stesse. Le tecniche statistiche avanzate stanno raccogliendo attenzioni in molte branche dell'acustica. La possibilità di ottenere informazioni accurate attraverso questi nuovi metodi sta generando innovazioni nella pratica tecnica come le analisi del campo sonoro in spazi chiusi, l'identificazione del rumore generato da impianti tecnologici, lo studio del soundscape, e le nuove tendenze riguardanti le smart cities e la loro rete di sensori. Dunque, queste tecniche e le loro relative applicazioni nel campo dell'acustica rappresentano uno strumento innovativo per ottenere analisi del campo sonoro più accurate e robuste. Al giorno d'oggi la prassi comune delle misurazioni fonometriche si limita alla descrizione dei contesti sonori da un punto di vista energetico. Il livello equivalente L_{eq} rappresenta la metrica principale con il quale definire un ambiente acustico. Analisi più sofisticate possono prendere in considerazione i livelli statistici. Tuttavia, l'uso dei percentili acustici è basato su assunzioni temporali non sempre attendibili. L'approccio convenzionale non fornisce nessun dettaglio del fenomeno misurato. La necessità di andare oltre il L_{eq} è stata particolarmente affrontata dalle tecniche di monitoraggio acustico passivo. Ad esempio, studi riguardanti l'ecologia o l'acustica marina hanno stimato i livelli di rumore ambientale attraverso l'uso della densità di potenza spettrale. La capacità di valutare i contributi di diverse sorgenti di rumore porterebbe molteplici dettagli all'analisi degli ambienti acustici. Questa abilità migliorerebbe il monitoraggio, la diagnosi, e la progettazione in molteplici circostanze portando alla creazione di spazi acustici più confortevoli. Un approccio statistico, basato sullo studio delle occorrenze dei livelli di pressione sonora, porterebbe prospettive diverse nell'analisi

dei monitoraggi a lungo termine. Pochi lavori riguardanti l'acustica architettonica hanno indagato la possibilità di descrivere un contesto sonoro attraverso un'analisi statistica. L'illustrare una scena sonora attraverso il più probabile livello di pressione sonora, piuttosto che con porzioni di energia, ha fornito molte informazioni utili per la comprensione delle attività svolte durante le misurazioni. L'andamento delle mode statistiche delle occorrenze può catturare i comportamenti tipici di specifiche tipologie di sorgenti sonore. Infatti, la separazione non supervisionata delle sorgenti sonore è un argomento di forte interesse nel ML. Il presente lavoro vuole proporre un metodo basato su ML per identificare, separare e misurare sorgenti sonore in condizione di coesistenza in scenari reali. Tale metodo è basato su monitoraggi a lungo termine ed è indirizzato a tutti gli acustici che lavorano sull'analisi del rumore ambientale in molteplici contesti. Il metodo presentato è basato sull'analisi dei gruppi o cluster analysis. Lo scheletro principale di questo processo per analizzare l'attività in diversi spazi è costituito da due algoritmi: il Modello di Mistura Gaussiana e il K-means clustering. Il metodo è stato applicato di due contesti differenti: le aule universitarie e gli uffici. Nelle aule universitarie l'approccio statistico è volto a identificare la quantità di chiacchiericcio fra gli studenti, la student activity. Questa può essere ritenuta una metrica per valutare il grado di attenzione degli studenti durante le lezioni. Diverse lezioni sono state monitorate prima e dopo i lavori di ristrutturazione. La student activity è stata misurata attraverso tutti i metodi presenti in letteratura, oltre che con il metodo proposto nel presente lavoro. Il confronto fra i metodi evidenzia la differenza fra l'approccio convenzionale e quello proposto. Quest'ultimo descrive coerentemente gli andamenti delle funzioni di ripartizione e di densità di probabilità. Al contrario, i livelli equivalenti e percentili non mostrano alcuna corrispondenza con le caratteristiche percepibili delle curve. Inoltre, lo studio dell'effetto Lombard riscontrato dagli insegnanti e dagli studenti prima e dopo i lavori di ristrutturazione dimostrano un cambiamento nel comportamento degli studenti durante le lezioni. In questo caso il metodo proposto ha permesso un'analisi oggettiva, condotta tramite metriche oggettive, di comportamenti soggettivi. Dunque, può rappresentare una connessione fra le analisi condotte attraverso parametri oggettivi e quelle soggettive basate su sondaggi. Negli uffici il metodo è stato utilizzato per monitorare il rumore ambientale all'interno degli spazi di lavoro. L'analisi mira a separare i contributi di rumore dovuti all'attività umana da quelli dovuti agli impianti meccanici o al traffico. In questo contesto, la metrica più rappresentativa per valutare il comfort degli impiegati è l'intelligibilità del parlato, più specificatamente lo Speech Transmission Index. Quest'ultimo è principalmente influenzato da due fattori: le proprietà acustiche della stanza e il rumore di fondo. I risultati evidenziano l'importanza della separazione delle sorgenti sonore nei monitoraggi a lungo ter-

mine per valutare l'influenza del rumore di fondo in scenari reali. Approfondimenti preliminari circa l'affidabilità del metodo sono stati ottenuti valutando i differenti approcci fra i due algoritmi attraverso tecniche di spectral matching e analisi statistiche. Inoltre, lo studio delle caratteristiche descrittive dei modelli statistici utilizzati ha permesso la proposta di nuove metriche per la valutazione del grado di dinamicità del contesto misurato. Infine, una validazione qualitativa del metodo proposto è stata condotta attraverso un'analisi duale attraverso tecniche di machine e deep learning. Un autoencoder variazionale è stato utilizzato per ricavare una rappresentazione latente di un'intera giornata di lavoro all'interno di un ufficio ed ha dimostrato la capacità di una parametrizzazione gaussiana di riconoscere differenti tipologie di sorgenti sonore. Le considerazioni finali spiegano perché il Modello di Mistura Gaussiana rappresenta l'algoritmo migliore per separare le sorgenti sonore attraverso misurazioni fonometriche. Infine, la presente tesi vuole rappresentare la proposta dettagliata ma preliminare di un metodo statistico per analizzare monitoraggi a lungo termine. Studi futuri dovrebbero focalizzarsi sull'analisi quantitativa dei risultati, la determinazione delle incertezze del metodo, e la definizione di intervalli di valori di riferimento per identificare numericamente il tipo di sorgente misurato. Nonostante i casi studio presentati mostrino risultati preliminari, il metodo proposto si è dimostrato robusto ed affidabile nelle sue applicazioni e può rappresentare un importante strumento analitico per i tecnici acustici. La conoscenza dettagliata di una scena sonora ottenuta attraverso un approccio statistico permette diagnosi e progettazioni più accurate e affidabili per ottenere ambienti acustici più confortevoli.

Abstract

Long-term monitoring of acoustical environments is gaining popularity thanks to the relevant amount of scientific and engineering insights that it provides. The increasing interest is due to the constant growth of storage capacity and computational power to process large amounts of data. In this perspective, machine learning (ML) provides a broad family of data-driven statistical techniques to deal with large databases. ML-based methods allow to detect and explore complex patterns in data. Analyses can concern the discovery of relationships between features and data labels or among features themselves. Advanced statistical techniques are focusing attention on many acoustical branches. The chance of achieving accurate information through these new approaches is leading to innovations in technical practices such as the analysis of the sound field in enclosed spaces, the detection of noise due to technological systems, the study of the soundscapes, or the new tendencies concerning the smart cities and their sensor networks. Therefore, these techniques and their applications in acoustics represent an innovative tool to obtain a more accurate and valuable analysis of the sound fields.

Nowadays, the common praxis of sound level meter measurements limits the global description of a sound scene to an energetic point of view. The equivalent continuous level L_{eq} represents the main metric to define an acoustic environment, indeed. Finer analyses involve the use of statistical levels. However, acoustic percentiles – i.e., the sound pressure levels exceeded for a certain percentage of the measurement time – are based on temporal assumptions, which are not always reliable. The energetic approach does not provide any real detail of the measured phenomenon. The urge to move beyond the L_{eq} has been addressed, particularly in passive acoustic monitoring. Studies concerning ecology or underwater acoustics used the probability density of the power spectral density to estimate ambient noise levels, for example. The ability to evaluate noise contributions of different sources would bring greater details to the analysis of acoustic environments. This skill would enhance monitoring, diagnosis, and design in plenty of circumstances, leading to the creation of more comfortable spaces.

A statistical approach, based on the study of the occurrences of sound pressure levels, would bring a different perspective to the analysis of long-term monitoring. Few works in room acoustics investigated the chance to describe the sound context through statistical analyses. Depicting a sound scene through the most probable sound pressure level, rather than portions of energy, brought more specific information about the activity carried out during the measurements. The statistical mode of the occurrences can capture typical behaviors of specific kinds of sound sources. Blind source separation is a current ML relevant subject, indeed.

The present work aims to propose an ML-based method to identify, separate and measure coexisting sound sources in real-world scenarios. It is based on long-term monitoring and is addressed to acousticians focused on the analysis of environmental noise in manifold contexts. The presented method is based on clustering analysis. Two algorithms, Gaussian Mixture Model and K-means clustering, represent the main core of a process to investigate different active spaces monitored through sound level meters. The procedure has been applied in two different contexts: university lecture halls and offices.

In university lecture halls, the statistical approach aimed to identify the student activity, i.e. the chatting among students. The latter can be deemed a metric to assess to what extent the students are focused during lectures. Several lessons were monitored before and after renovation works measuring the student activity through all the methods used in the literature and the proposed one. The comparison among methods highlights the difference between the conventional and the proposed approaches. The latter describes consistently the tendencies of both cumulative and probability density functions. On the contrary, equivalent and statistical levels do not correspond to any detectable feature of the cumulative curve obtained from a sound level meter monitoring. Moreover, the study of the Lombard effect experienced by teachers and students before and after the acoustic treatments of the halls showed how the students' behavior changed. Thus, the proposed method provided detailed results through objective metrics to assess a subjective trend. It can represent a connection between measurable criteria and surveys.

In offices, the method was used to monitor the environmental noise of active workplaces. The analysis aims to separate the noise contributions of human activity from mechanical or traffic noise. Here, the main metric to evaluate the acoustic comfort of employees is represented by intelligibility, in the specific case the Speech Transmission Index. This is affected by two factors: the acoustic properties of the space and the background noise. Results highlighted the importance of blind source separation in long-term monitoring to assess the noise's influence on the intelligibility in active scenarios. Further, spectral matching and statistical discussions concerning

the different approaches of the two algorithms brought preliminary insights into the reliability of the proposed method, besides the proposal of features as metrics to assess the extent of the dynamic of the acoustical context. Finally, the qualitative validation of the proposed method was carried out through a dual analysis via machine and deep learning. A variational autoencoder learned a latent representation of the whole working day inside an office and proved the ability of a Gaussian parametrization to recognize the different kinds of sources. Final remarks explain the reasons why the Gaussian Mixture Model represents the best algorithm to perform a blind source separation through sound level meters.

Finally, the present thesis is intended as a detailed but preliminary proposal of a statistical method to analyze long-term monitoring. Further studies should deepen the quantitative analysis of the results, the determination of the method's uncertainty, and the definition of ranges of values to identify numerically the type of source. Despite the case studies presented here provide only preliminary results, the proposed method shows robust and reliable results in describing the acoustic scenario and it could represent an important analytical tool for acousticians. Detailed knowledge of the sound scene gained through a statistical approach would allow more accurate and reliable diagnoses and designs for more comfortable acoustic scenarios.

Table of contents

List of figures

List of tables

Introduction

In the era of big data, the appeal of artificial intelligence (AI) has been rapidly growing. The increase of computational power allows to exploit complex algorithms to accomplish different tasks, like classification and regression. Nowadays, AI represents a thriving field with several real-world applications and research topics. A subfield of AI is represented by machine learning (ML). One of its several definitions is [102]:

> A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

E, that gives to the program the ability to improve its P, is given by data. ML is data-driven and is able to explore and find more complex relationships within data than classical methods. AI history is notorious to be made by boom-and-bust cycles [150]. Its principles are not new and the advances are linked to breakthroughs during the years. For instance, the principal component analysis (PCA), one of the most common algorithms used to perform a dimensionality reduction in large datasets, was invented by Pearson in 1901 and developed by Hotelling in 1930s [110, 65, 53]. The computational model for neural networks (NNs) was developed in 1943 [96]. Then, the perceptron algorithm, the first artificial NN, was introduced in 1958 by Rosenblatt [125]. However, its interest decreased until the backpropagation algorithm was introduced [127]. In the time of this writing, AI is in a "third wave" begun in 2006 with the greedy layer-wise pretraining introduced by Hinton [50, 58].

In this new wave, ML has been leading advances in several fields, either humanistic or scientific. Progresses concern acoustics too, especially in signal processing methods where performances overcome traditional techniques [8]. Common limitations for many ML methods concern the need for large amounts of data to test and train the algorithms and the lack of interpretability of the resulting model, especially when the analysis deals with deep learning [78].

The main advances of ML in acoustics concern: source localization in reverberant and ocean environments, bioacoustics, blind source separation, and the classification

of environmental sounds. Each problem is addressed with different approaches. In fact, one of the basic rules in ML workflows concerns the *no free-lunch theorem* (NFL). It states that, within certain constraints, over the space of all possible problems, every optimization technique will perform as well as every other one on average [149]. Since ML can be outlined as an optimization problem, the theorem has its implications in the use of these algorithms. The most important is that there is no one optimum ML technique. Thus, there is no one model that works best for every kind of problem. NFL also affects typical criteria for what makes a "good" ML model, like avoiding overfitting or selecting the simplest reliable model. Given that there is not a single machine learning algorithm that works well for all potential prediction applications, both new algorithms and a deeper understanding of existing ones must be developed. NFL also proves that several machine learning methods should be tested for a particular predictive modeling problem. However, the consequences of the theorem are founded on the assumption that the algorithms choice is based on zero knowledge of the problem being solved. Thus, being familiar with the problem pave the way for the most effective algorithm selection.

Most ML techniques are often classified as either supervised or unsupervised learning [102]. With labeled input and output pairs, the goal of supervised learning is to learn a predicted mapping from inputs to outputs. These methods are the most popular type of learning. Examples of supervised algorithms are: linear regression, support vector machines, nearest-neighbor classifiers, neural networks. Without labelled data, the goal of unsupervised learning is to find meaningful patterns in the data. Data visualization, exploratory data analysis, anomaly detection, and feature learning are just a few of the practical applications of this type of learning [78]. Examples of unsupervised algorithms are: K-means clustering, Gaussian Mixture Model, principal component analysis, autoencoders. Unsupervised techniques are used also as a support tool to improve performances of supervised frameworks.

Cluster analysis, which looks for patterns in the data population based on similarity, is the most popular type of unsupervised learning. These algorithms frequently use iterative processes starting by guessing a set of clusters and then updating the subdivision to enhance both the diversity among clusters and the similarity within the same cluster. The way the similarity is measured depends on the algorithm used; it could be based on well-known definitions, e.g. Euclidean distance, or bespoke metrics. As seen from the NFL theorem, the problem's knowledge to solve allows to choose the best solution based on the task to accomplish.

Thesis motivations and objectives

Nowadays, the common technical praxis among acousticians is based on the energetic description of a sound context. Sound level meter measurements are used to collect different energetic metrics. The most common parameter used is the equivalent level L_{eq}, i.e. the root mean square level averaged over the whole recording time. However, being an energy average, only the highest levels affect significantly the results. L_{eq} is used to assess long-term sound exposure during representative measurement periods, such as activities to monitor. To eliminate the influence of the measurement duration, the sound exposure level SEL is used. Instead of averaging over the measurement period, a reference duration equal to 1 s is considered. L_{eq} is proportional to the average sound power, SEL indicates the total sound energy. Finer analyses are carried out through statistical sound pressure levels. Acoustical percentiles L_N are defined as the sound pressure levels exceeded for the N% of the measurement time, where N is the respective percentage to consider. For instance, L_{90} is the sound pressure level exceeded for the 90% of the period measured.

The current state-of-the-art does not allow technicians to carry out investigations about noise contributions of each kind of sound sources in real-world scenarios monitoring. It is common practice to associate different contributions with statistical levels. For instance, L_{90} is often associated with background noise in many contexts. However, this approach relies on temporal assumptions impossible to prove. The ability to evaluate different contributions would lay the foundations for more accurate acoustical diagnoses and designs in manifold scenarios.

Several works investigated different methods to separate sound sources in long-term monitoring. The most successful approaches used statistical methods rather than the conventional use of energetic metrics. The management of sound pressure levels as random variables enables the use of statistical models to describe the acoustical contexts. Thus, the scenario would be represented by the most probable sound pressure level. This would not be in contrast with the use of L_{eq}, which describes different characteristics of the measured phenomenon. Applications of statistical approaches were made in classrooms to separate the speech levels of teachers from the students chatting and the ventilation system noise [62, 128, 27]. Further, similar techniques were used to investigate the environmental noise in offices and hospitals [37, 56].

The present work proposes a measurement method to improve the ability of acousticians to identify and separate noise contributions from several coexisting sources in long-term monitoring. The analysis is based on real-world applications of clustering algorithms on sound pressure levels data. Based on the problem

to address, two unsupervised algorithms – the Gaussian Mixture Model and the K-means clustering – were selected to investigate the statistical features of long-term monitoring in complex environments. The performance comparison of both algorithms showed that the Gaussian Mixture Model is the best method to carry out statistical analyses of acoustical contexts. Finally, a variational autoencoder was useful in validating qualitatively all the assumptions at the basis of the unsupervised analysis.

Overview of the thesis

The thesis is made up of four chapters. The dissertation starts by presenting the theoretical background needed to address the whole work. Then, the developed method is proposed in a general description and applied in different real-world acoustical contexts for the evaluation of its reliability. Final remarks sum up all the insights obtained by the applications of the method and the open issues to address through future works.

Chapter 1

Chapter 1 contains an overview of the theoretical background of the present work. Unsupervised algorithms, i.e. Gaussian Mixture Model and K-means clustering, are outlined focusing on characteristics that are useful for the applications described in the following chapters. The relationship between the two algorithms is provided. Further, the model selection metrics are briefly presented to give a satisfactory overview of their different approaches to defining similarity among clusters. Hints about how neural networks work are provided to supply a basic idea of deep learning. Finally, the variational autoencoder is presented limitedly to what concern its use in the present work for the qualitative validation of the proposed method.

Chapter 2

Chapter 2 describes the method proposed in this work. Detailed descriptions provide each step to accurately reproduce the workflow. Starting from the data acquisition, the narrative provides explanations and recommendations about the setting of the sound level meter, the algorithms, and the meaning of the results. The method is generally presented and it can be used in any kind of context requiring a sound level meter measurement. Figures are based on the Gaussian Mixture Model because it is the recommended method. It is also easier to visualize.

Chapter 3

Chapter 3 shows the first applications of the proposed method in classrooms. Based on the literature, all the techniques to detect the student activity were used. Besides the Gaussian Mixture Model and the K-means clustering, the Peak detection was compared with the equivalent and statistical levels commonly used in praxis. Considerations about the differences among results are provided. Further studies were made after acoustical treatments made in the same lecture halls. Here, the change of the students' behavior was assessed through the measurement of the student activity before and after the renovation works. Active and passive treatments are described to provide a comprehensive context to understand the results.

Chapter 4

Chapter 4 provides details concerning the applications of the proposed method in offices. In this context, the analysis has been mainly focused on the method and its reliability. The first study investigates the source separation during working hours identifying the human and the mechanical noise components. Statistical and spectral insights about the results provide preliminary considerations about the reliability of the method. Then, a survey is presented over three further case studies to evaluate the adaptability of the analysis in different spaces. Moreover, preliminary insights about the use of the overlapping areas as acoustical features are provided. The last case study focuses on a dual analysis to validate the assumptions underlying the cluster analysis proposed by the method. A semi-supervised analysis via a variational autoencoder was carried out to assess the reliability of the unsupervised analysis. Final remarks explain why the Gaussian Mixture Model seems to be the most reliable algorithm among the others.

Main contributions

The main contributions of this thesis are:

- The proposal of a rigorous method to evaluate the different noise contributions of coexisting sound sources in real-world conditions. The goal is pursued by analyzing the accuracy of the outcomes from a quantitative and qualitative point of view (Chapters 2, 3, 4).
- The comparison among all the techniques used in literature to monitor student activity. Pros and cons of each method are reported (Chapter 3).

- The relationships and differences among conventional and proposed approaches describing the link between the probability and the cumulative density functions. The detection of the most probable sound pressure level corresponds to the zeros of the second derivative of the cumulative density function (Chapter 3).

- The analysis of the change of subjective human behavior as a consequence of acoustical treatments through objective metrics. The proposed method can be used as an objective tool to assess subjective attitude through the change of Lombard effect's slopes and correlations among occupancy and sound pressure levels (Chapter 3).

- The in-depth analysis of the statistical features of the speech signal, the mechanical noise, and the traffic noise through the variations of their standard deviations (Chapter 3, 4).

- A thorough analysis of the difference between the two unsupervised methods, i.e., Gaussian Mixture Model and K-means clustering, in long-term monitoring. Details about the meaning of using a hard or fuzzy algorithm to manage sound pressure levels are provided (Chapter 3, 4).

- The assessment of the Speech Transmission Index in real-word conditions. The analysis about to what extent the intelligibility is affected by the background noise is provided either the sound pressure levels are evaluated according to the conventional or the proposed approach (Chapter 4).

- The proposal of the overlapping areas as a metric to assess the amount of collaborative work according to ISO 22955. This has been conducted through a preliminary survey among different offices (Chapter 4).

- The latent representation of long-term monitoring conducted through a sound level meter. A deep clustering analysis to assess an entire working day was presented (Chapter 4).

- A comprehensive analysis about the ability of carrying out a measurement through a deep learning approach. Limitations about the accuracy of measurements and the pre-processing workflow needed for a deep learning analysis are provided. Further, explanations concerning the spectral uncertainties due to the latent representation of long-term monitoring are described (Chapter 4).

Chapter 1

Theoretical background

The method proposed in the present thesis is based on statistical analyses over long-term monitoring carried out through sound level meters. Data populations are represented by large amounts of sound pressure levels measured during the monitoring. Thus, data have no labels. The proposed method aims to find latent structures in data and corresponding insights to describe the physical characteristics of different coexisting sound sources. Because labels are not available in the database, the analysis is focused on unsupervised learning problems. These can involve cluster analysis, density estimation, or dimensionality reduction.

Cluster analysis is the general task of looking for patterns in data [9]. Data are gathered in different groups 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 of the data, called *features*, which allow the data to be labelled. Different algorithms use different criteria to find similarity in data, i.e. shaping groups called *clusters*. Main algorithms can be categorized in [1]:

Feature selection algorithms: used to remove noisy and irrelevant features in data.

Model-based algorithms: the basic concept is modelling data from a generative process. Once the generative process is chosen, the Expectation-Maximization algorithm estimates the model's parameters.

Distance-based algorithms: can be in turn divided in *flat* and *hierarchical*. The first use distance function and partitioning representative to shape clusters. The second use dendrograms to represent hierarchically the partitions.

Density- and grid-based algorithms: the amount of data points in a predetermined volume of the locality or a smoother kernel density estimate are two ways in which the density at any given position in the data space is defined.

Dimensionality reduction algorithms: given a matrix of data, these methods try to cluster rows and columns simultaneously to reduce the dimensionality of representation.

The choice of the algorithm depends on the data domain and the problem scenario. In this chapter, the theoretical background is limited to what concerns the tools used in the applications of the proposed method. All sections aim to provide the basic information to understand the details and the choices in the background of the present work. Section 1.1 describes the Gaussian Mixture Model and the K-means clustering method. These are the two main algorithms used to assess the proposed method in the present thesis. The qualitative validation of the method has been carried out through a variational autoencoder. Thus, some basic hints about how neural networks operate are provided in Section 1.3 to lay the foundations for understanding the variational inference used for the validation.

1.1 Unsupervised algorithms

1.1.1 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a model-based clustering technique [97]. A probabilistic model – known as *mixture distributions* – recovers the original structure of the initial data superimposing a linear combination of Gaussian distributions. The accuracy of the approximation to the initial data, i.e. the general distribution, is adjusted through the means, the covariances, and the weights of each component of the mixture. Given a set of N independent observations $x = \{x_1, ..., x_N\}$, the mixture of Gaussians $f(x)$ is expressed as:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{1.1}$$

where K is the number of components, μ_k are the means, Σ_k are the covariances, and π_k are the *mixing proportion* or *weights* of each component. After normalizing the density $f(x)$ and each Gaussian component, integrating Equation 1.1 with respect to x, then:

$$\sum_{k=1}^{K} \pi_k = 1. \tag{1.2}$$

Since both $f(\mathbf{x})$ and $\mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)$ are non-negative quantities, it is deduced that $\pi_k \geq 0$. Thus, combining the latter implication with Equation 1.2:

$$0 \leq \pi_k \leq 1 \qquad (k = 1,...,K). \tag{1.3}$$

The most common approach to fit mixtures of distributions is represented by the maximum likelihood estimation (MLE). The likelihood function $\mathcal{L}$ of a mixture of univariate normal distributed heteroscedastic components is defined as:

$$\mathcal{L}(x) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}. \tag{1.4}$$

A Gaussian mixture can be formulated in terms of discrete latent variables. To do this it is needed to introduce z, a K-dimensional binary random variable, in which a particular element z_k is equal to 1 and all other elements are equal to 0. This scheme is called 1 - of - K representation. Thus, $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. The joint distribution $f(x,z)$ is defined in terms of a marginal distribution $f(z)$ and a conditional distribution $f(x|z)$. The marginal distribution can be written as:

$$f(z) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{1.5}$$

The conditional distribution can be defined in the form:

$$f(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k,\Sigma_k)^{z_k}. \tag{1.6}$$

The joint distribution is obtained by $f(z)f(x|z)$. The marginal distribution of x is given by summing the joint distribution over all possible states of z. Thus, using Equations 1.5 and 1.6:

$$f(x) = \sum_z f(z)f(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k,\Sigma_k). \tag{1.7}$$

Hence, the marginal distribution of x has been written as a Gaussian mixture of the form seen in Equation 1.1. It follows that, since $f(x) = \sum_z f(x,z)$ for a given set of N observations $x = \{x_1,...,x_N\}$, for every data point x_n there is a corresponding latent variable z_n.

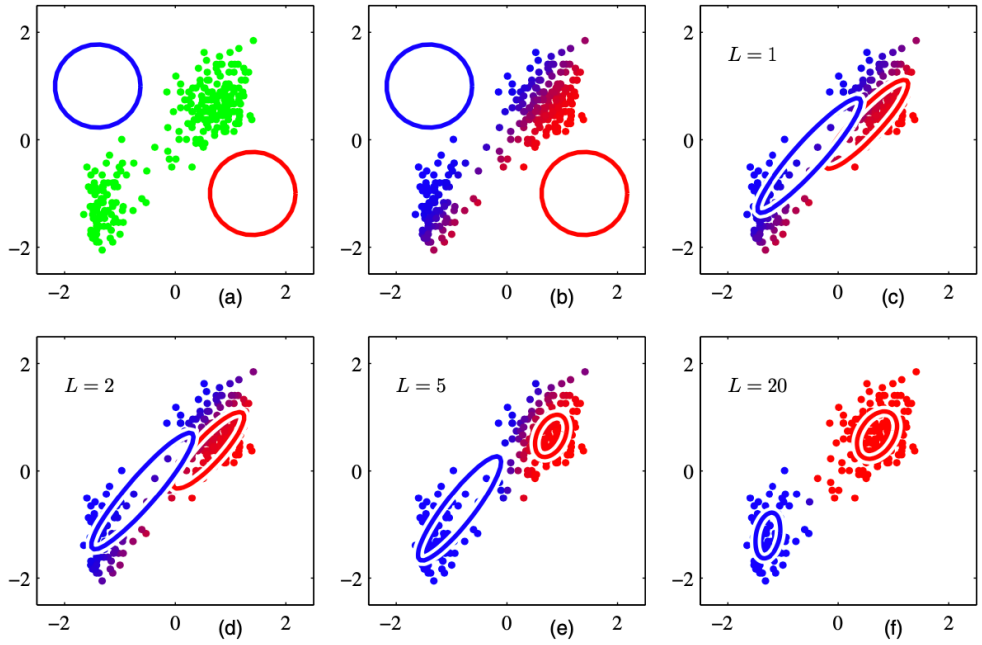


Fig. 1.1 Example of iterations carried out via the EM algorithm. From Bishop *Pattern recognition and machine learning* (p.437) [9].

The conditional probability of z given x can be defined using Bayes' theorem:

$$\gamma(z_k) = f(z_k = 1|x) = \frac{f(z_k = 1)f(x|z_k = 1)}{\sum_{j=1}^{K} f(z_j = 1)f(x|z_j = 1)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}. \quad (1.8)$$

Thus, γz_k is the posterior probability after x has been observed. It is possible to represent the responsibilities values $\gamma(z_{nk})$ per each data point x_n as well.

The formulation in terms of discrete latent variables allows us to work with a joint distribution $f(x,z)$ and motivate the use of the Expectation-Maximization (EM) algorithm [38]. This is a method used in the present work to find the local MLE. Defining N_k as the number of points assigned to cluster k, the EM algorithm is an iterative process, outlined as follows:

1. Initialize the hyperparameters: means μ_k, covariances Σ_k, and mixing coefficients π_k. Evaluate the initial value of the log-likelihood.

2. **E step**. Evaluate the conditional probabilities via Equation 1.8.

3. **M step.** Re-estimate the hyperparameters using the current conditional probabilities $\gamma(z_k)$.

4. Evaluate the log-likelihood $\log(\mathcal{L}(x))$ and check for convergence. If it is not satisfied, return to step 2.

Figure 1.1 shows an example of iterations carried out via the EM algorithm. The first plot 1.1(a) shows the dataset in green and two Gaussian components as blue and red circles. Plot 1.1(b) shows the initial E step where the conditional probabilities are calculated. Here, the same colors correspond to the relative cluster, i.e. red or blue. Shaded dots indicate data with significant probabilities of belonging either to the blue or the red cluster. Plot 1.1(c) shows the M step. Here, Gaussian means are moved to the means of the data. Plots from 1.1(d) up to 1.1(f) show 2, 5, and 20 EM iterations, respectively. Algorithm 1 shows the EM process in pseudocode.

Algorithm 1: EM algorithm for GMM.

Input: $x = \{x_1, ..., x_N\}$ data observation, K number of clusters
Output: $\pi = \{\pi_1, \ldots, \pi_k\}, \mu = \{\mu_1, \ldots, \mu_k\}, \Sigma = \{\Sigma_1, \ldots, \Sigma_k\}$

Initialize π, μ, Σ
// E-step
repeat
  for $m = 1 : M$ **do**
    for $n = 1 : N$ **do**
      for $k = 1 : K$ **do**
        $\gamma(z_k) = f(z_k = 1|x) = \frac{f(z_k=1)f(x|z_k=1)}{\sum_{j=1}^{K} f(z_j=1)f(x|z_j=1)} = \frac{\pi_k \mathcal{N}(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j,\Sigma_j)}$;
      end
    end
    // M-step
    for $k = 1 : K$ **do**
      $\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_k)x}{\sum_{n=1}^{N} \gamma(z_k)}$;
      $\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_k)(x-\mu_k)(x-\mu_k)^M}{\sum_{n=1}^{N} \gamma(z_k)}$;
      $\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_k)$;
    end
  end
until $\log(\mathcal{L}(x)) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\}$ *convergence is met*;

1.1.2 K-means clustering

K-means clustering (KM) is a non-probabilistic and distance-based clustering technique [1]. The standard algorithm was proposed by Lloyd to address the pulse-code modulation [90]. The optimization is focused on a distance metric. As a consequence, KM shapes a number of K clusters partitioning the data space into Voronoi cells.

Given a set of independent observations $x = \{x_1, ..., x_N\}$ of a random D-dimensional Euclidean variable x. The goal is to find both an assignment of data points to clusters

and a set of vectors $\{c_k\}$, called *centroids*, minimizing the squares of the distances of each data point to its closest centroid. Then, it is important to define $r_{nk} \in \{0,1\}$ where $k = 1,\ldots,K$. This is a set of binary indicator variables describing to which cluster the datapoint x_n has been assigned. Thus, if x_n has been assigned to the cluster k, then $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$. Then, the objective function can be defined as:

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - c_k||^2. \tag{1.9}$$

Equation 1.9 shows the sum of the distances' squares of each datapoint to its assigned vector of centroids $\{c_k\}$. As stated above, the goal is to minimize J. This is made first through the determination of r_{nk}. This term can be equal to 1 for whichever value of k corresponding to the closest centre. Thus:

$$r_{nk} = \begin{cases} 1 \text{ if } k = argmin_j||x_n - c_j||^2 \\ 0 \text{ otherwise.} \end{cases} \tag{1.10}$$

Keeping r_{nk} fixed, J can be minimized setting its derivative with respect to c_k to 0:

$$2\sum_{n=1}^{N} r_{nk}(x_n - c_k) = 0 \tag{1.11}$$

which is solved for c_k. Thus:

$$c_k = \frac{\Sigma_n r_{nk} x_n}{\Sigma_n r_{nk}}. \tag{1.12}$$

KM is a greedy algorithm. The minimization of the objective function J is known to be an NP-hard problem. Thus, it can converge to local minima through an iterative process. Figure 1.2 shows the same dataset seen in Figure 1.1 analyzed via KM. The first plot 1.2(a) shows the dataset in green and the initial choices for centroids c_1 and c_2 with blue and red crosses. Plot 1.2(b) shows the assignment of datapoints to the closer centroid. After the assignment, the centroids are recalculated as shown in 1.2(c). Plots from 1.2(d) to 1.2(i) show the subsequent iterations. Magenta lines indicate the boundaries of the Voronoi cells. Algorithm 2 describes the steps of the process.

After the first step, steps 2 and 3 are repeated until convergence [75]. Mathematical details about the proof of the convergence are described in [131].

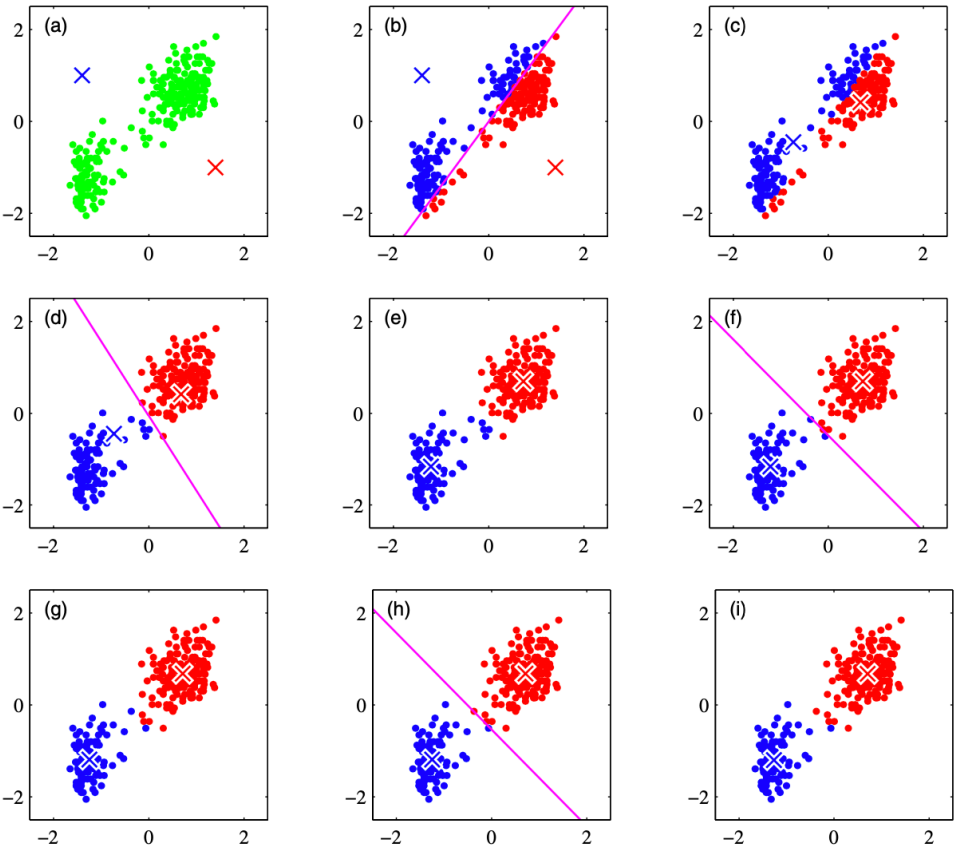


Fig. 1.2 Example of iterations carried out via the KM algorithm. From Bishop *Pattern recognition and machine learning* (p.426) [9].

Algorithm 2: K-means clustering

Input: $x = \{x_1, ..., x_N\}$ set of N data observations, K number of clusters
Output: Set of K clusters

1 Selection of an initial partition of data into K clusters;
2 **repeat**
3 Generation of a new partition by assigning each data x_{k_i} to its closest cluster center c_{k_i};
4 Compute new clusters centres c_{k_i}.
5 **until** *convergence criterion is met*;

1.1.3 Relationship between Gaussian Mixture Model and K-means

The EM algorithm for GMM and the KM algorithm are very similar when compared [9, 94]. The EM algorithm performs a soft assignment based on the posterior probability, in contrast to the KM approach, which does a hard assignment in which each data point is assigned to a single cluster. In reality, the KM may be derived as an example of an EM limit for Gaussian mixtures as follows.

Given a Gaussian Mixture Model with the components' covariance matrices defined by $\varepsilon\mathbf{I}$, where ε is a variance parameter shared among all components and $\mathbf{I}$ is the identity matrix. Thus:

$$f(x|\mu_k,\Sigma_k) = \frac{1}{(2\pi\varepsilon)^{1/2}} e^{-\frac{1}{2\varepsilon}||x-\mu_k||^2}. \tag{1.13}$$

Considering the EM algorithm for a Gaussian mixture with ε as a fixed constant instead of a parameter to be re-estimated, the posterior probability from Eq. 1.8 is:

$$\gamma(z_{nk}) = \frac{\pi_k e^{\frac{-||x_n-\mu_k||^2}{2\varepsilon}}}{\sum_j \pi_j e^{\frac{-||x_n-\mu_j||^2}{2\varepsilon}}}. \tag{1.14}$$

Considering the limit $\varepsilon \to 0$ and $\pi_k \neq 0$, the denominator will go to zero more slowly than the numerator. Thus, since $\gamma(z_{nk}) \to r_{nk}$, the assignment of data becomes hard, as in KM. Each data point is assigned to its closest mean. The re-estimation of μ_k made by EM and explained in Algorithm 1 reduces to the KM result. Moreover, the re-estimation of π_k comes down to setting the mixing coefficients equal to the portion of data points belonging to the cluster.

In conclusion, in the considered limit, the maximization of the expected log-likelihood becomes equivalent to the minimization of the distortion measure J for KM :

$$\mathbb{E}_z[\ln f(x,z|\mu,\Sigma,\pi)] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - c_k||^2 + const. \tag{1.15}$$

In this specific case, giving the GMM back the same results ok KM, the covariance matrices are not considered. Thus, clusters have a spherical shape.

1.2 Model selection metrics

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 [1]. 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.

1.2.1 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 [22]. 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:

$$SS_B = \sum_{i=1}^{K} n_{k_i} ||c_{k_i} - C||^2 \tag{1.16}$$

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):

$$SS_W = J(K) = \sum_{i=1}^{K} \sum_{x_{k_i} \in c_{k_i}} ||x_{k_i} - c_{k_i}||^2 \tag{1.17}$$

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

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

$$CH = \frac{SS_B}{SS_W} \frac{N-K}{K-1} \tag{1.18}$$

The optimal model is represented by the highest value obtained from Equation 1.18.

1.2.2 Davies-Bouldin

The Davies-Bouldin index assesses similarity among clusters through the ratio of within- and between-cluster distances [32].

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

$$D_{i,j} = \frac{\bar{d}_{x_{k_i}} + \bar{d}_{x_{k_j}}}{d_{c_{k_i}, c_{k_j}}} \tag{1.19}$$

where

$$\bar{d}_{x_{k_i}} = \frac{1}{n_{k_i}} \sum_{x_{k_i} \in k_i} |x_{k_i} - c_{k_i}| \tag{1.20}$$

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:

$$d_{c_{k_i},c_{k_j}} = (|c_{k_i} - c_{k_j}|^2)^{1/2}. \tag{1.21}$$

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

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j\neq i}\{D_{i,j}\}. \tag{1.22}$$

The optimal model is represented by the smallest value obtained from Equation 1.22.

1.2.3 Silhouette coefficient

The silhouette coefficient is a graphical quantitative evaluation of the degree of separation among clusters [126]. 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:

$$a(i) = \frac{1}{|n_{k_i}| - 1} \sum_{x_{k_i}, x_{k_{i'}} \in k_i} d_{x_{k_i}, x_{k_{i'}}}. \tag{1.23}$$

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:

$$b(i) = \min \frac{1}{|n_{k_j}|} \sum_{x_{k_i} \in k_i, x_{k_j} \in k_j} d_{x_{k_i}, x_{k_j}}. \tag{1.24}$$

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:

$$\begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i). \end{cases} \tag{1.25}$$

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 $\bar{s(i)}$ is the mean of each $s(i)$, the silhouette

coefficient *SC* can be defined as:

$$SC = \max_{K} \bar{s}(K) \tag{1.26}$$

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 from Equation 1.26.

1.2.4 Gap statistic

Gap statistic was introduced by Tibshirani et al. and formalizes the "elbow" method [138]. 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:

$$W_K = \sum_{i=1}^{K} \frac{1}{2n_{k_i}} D_{k_i} \tag{1.27}$$

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

$$D_{k_i} = \sum_{x_{k_i}, x_{k_{i'}} \in k_i} d_{x_{k_i}, x_{k_{i'}}}. \tag{1.28}$$

the pairwise distances of all points in the cluster k_i.

Then, the Gap value is defined as:

$$Gap(K) = \mathbb{E}_r^*\{\log(W_K)\} - \log(W_K). \tag{1.29}$$

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 from Equation 1.29.

1.3 Neural networks

A neural network (NN) is a group of connected nodes, called *neurons*. The simplest type of NN is represented by the feed-forward NNs, i.e. there are no loops in the network. This kind of NN is used here to present the basic elements that constitute a deep learning algorithm. A set of numerical input values is mapped by the single neuron into a single output value. A neuron is essentially a multi-input linear regression function. The major distinction is represented by the *activation function*, i.e., a function in which the output of the multi-input linear regression is passed through. The most important property of activation functions is that they map the output through non-linear functions. The non-linearity introduced in the process allows NNs to learn more complex relationships. One more fundamental element of NNs is represented by the *weights*. Each neuron has a weight associated with its connection and the weight is applied on the input received before the result is pushed in the activation function.

Figure 1.3 shows the architecture of a simple feed-forward fully connected NN. The organization of a NN is outlined by *layers*. The network shown in the picture is made by three layers: input, hidden layer, and output indicated in yellow, green and red, respectively. Labels $W_{i,j}$ on each connection indicate the weights applied on single neurons. Fully connected means that each neuron is connected to all neurons of the subsequent layer. If the NN has more hidden layers than one, the NN is called *deep NN*.

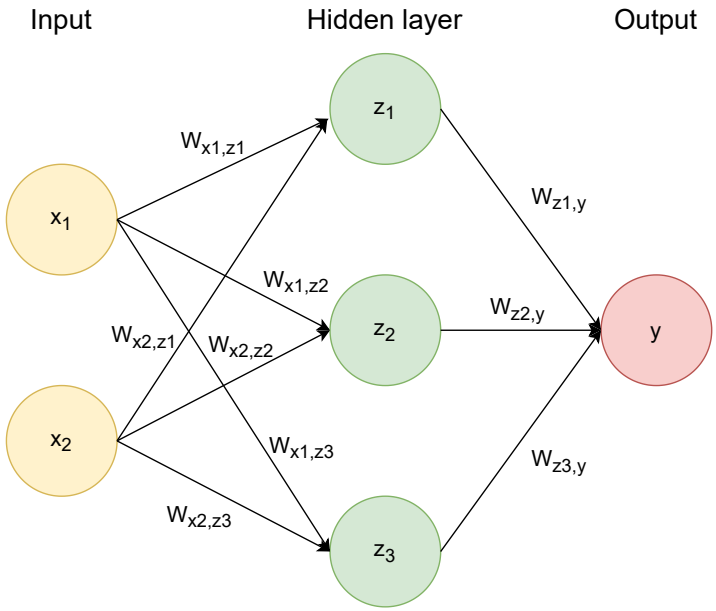


Fig. 1.3 Example of feed-forward fully connected NN.

The training of a NN aims to find the best weights for the network's connections. The adjustment of weights improves the accuracy in accomplishing the task. This is made by minimizing the error between predicted and expected values. The evaluation of the minimization process is made by means of a *loss function*. As long as the loss decreases, the learning keeps going.

Based on the architecture shown in Figure 1.3, the basis function of a NN in the hidden layer (1) can be expressed as:

$$z_j = h(\sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)}) \tag{1.30}$$

where $j = 1, \ldots, M$ is the number of linear combinations of the inputs $x = \{x_1, \ldots, x_n\}$; w_{ji} and w_{j0} are the weights and biases, respectively.

At the start of the training, the algorithm assigns random weights to the connections. Then, the *backpropagation algorithm* allows the weights to be updated in each neuron starting from the output layer and going back through the network [127]. The update does not aim to remove the error in training but to reduce it. This is due to the intent of generalizing the training on new instances.

As stated in the beginning of this chapter, a feed-forward fully connected NN has been presented as an example for introducing some basic concepts of deep learning. However, many other kinds of networks have been developed. One of the most popular architectures is the *convolutional NN* (CNN) [85, 86].

Convolutional layers

CNNs became very popular in computer vision for their ability to perform image recognition or classification. Since a spectrogram is defined by a matrix, it can be handled like an image. Hence, CNN achieved popularity in acoustics, too. A classical CNN architecture is constituted by two blocks: the feature learning and the classification.

Figure 1.4 shows an example of CNN. Grey blocks show the feature learning blocks. Here, the convolution layers learn the features through the *kernels*, indicated in light blue in the figure. A kernel is a matrix that slides over the input data. It performs an element-wise multiplication with the part of the input considered and then it sums up to obtain a single output. Kernel is the core of the convolutional operation. Given a 2D matrix as input data, every area the kernel slides over is subjected to this same procedure, which results in the creation of a second 2D matrix of features. In essence, the output features are the weighted sums of the input features that are roughly placed in the same area as the output pixel on the input layer (the

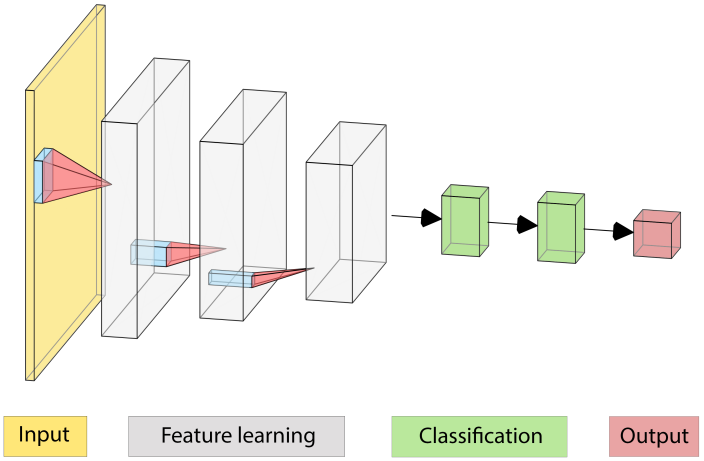


Fig. 1.4 Example of a CNN architecture. Grey blocks indicate the convolutional layers, green blocks the fully connected layers. Yellow and red boxes indicate input and output, respectively.

weights are the values of the kernel itself). As example, consider the computation of the discrete convolution between the 5x5 input feature map shown in the blue matrix of Figure 1.6 and the 3x3 kernel of values shown in Figure 1.5.

0	1	2
2	2	0
0	1	2

Fig. 1.5 Example of a 3x3 kernel of a convolutional layer. From "A guide to convolution arithmetic for deep learning", by V. Dumoulin and F. Visin [42].

It is worth to notice how the size of the kernel determines the number of input features that are combined to produce the output features map. The convolutional process limits significantly the number of parameters of the model and, consequently, the computational requirements. CNNs typically reduce the number of multiplications in comparison to fully connected NNs of a factor equal to 100.

Besides the input and the kernel sizes, the output size of a convolutional operation can be influenced by two properties that can be used in shaping the layer architecture: the stride s and the zero padding p. The first is the distance between two consecutive positions of the kernel. The second is the number of zeros concatenated at the extremes of the axes. These can be considered as hyperparameters of a CNN.

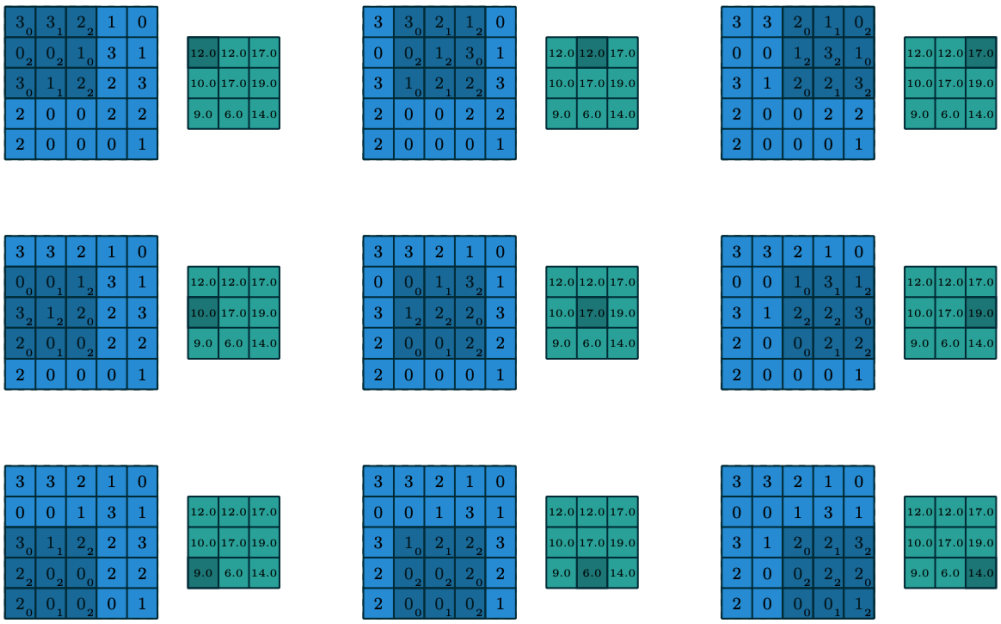


Fig. 1.6 Example of the computation of the discrete convolution between an input feature map and the 3x3 kernel shown in Figure1.5. From "A guide to convolution arithmetic for deep learning", by V. Dumoulin and F. Visin [42].

Figure 1.7 shows a convolutional operation with $s = 2$ and $p = 1$. It is noticeable how the output size in Figure 1.6 is smaller than the input map. Applying both s and p, the output size is equal to the input matrix. Thus, the output can be calculated as follows:

$$O_{height} = \frac{I_{height} - k_{height} + 2p}{s} + 1, \tag{1.31}$$

where the subscript *height* indicates the height of the output O, the input I, and the kernel k, respectively. The same can be done for the width if matrices or kernels are rectangular.

1.3.1 Variational Autoencoder

The variational autoencoder (VAE) is a way to realize inference and learning in probabilistic models and was introduced by Kingma and Welling [79]. 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. Figure 1.8 shows a general outline of a VAE's architecture.

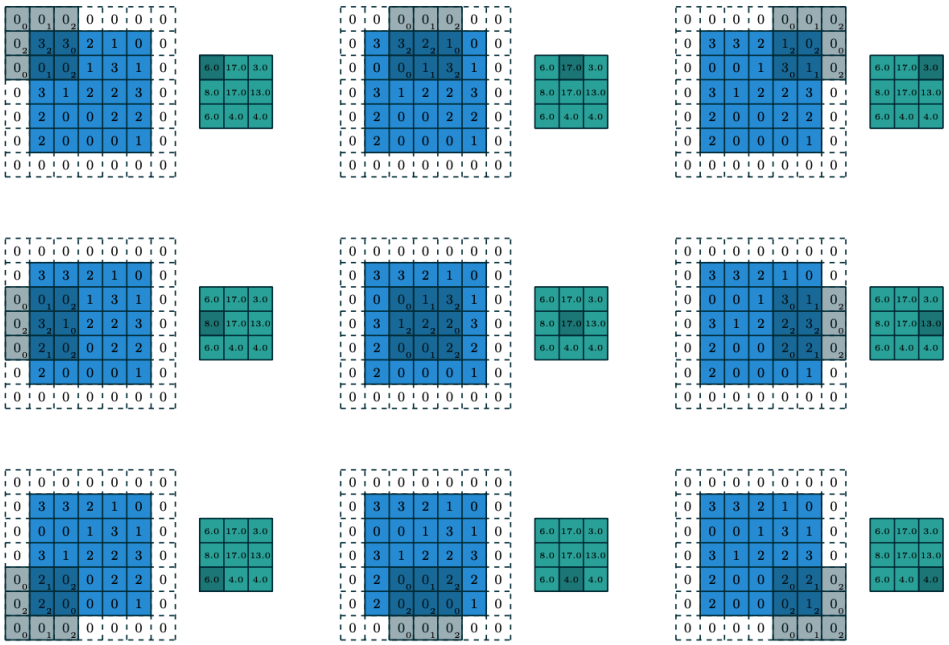


Fig. 1.7 Example of the computation of the discrete convolution between an input feature map and the 3x3 kernel shown in Figure1.5 with $s = 2$ and $p = 1$. From "A guide to convolution arithmetic for deep learning", by V. Dumoulin and F. Visin [42].

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 in comparison 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 is $x \in \mathbb{R}^{m_1 \times m_2}$ and its latent representation is $z \in \mathbb{R}^n$. 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:

$$l_i(\theta,\phi) = -\mathbb{E}_{z\sim q_\theta(z|x_i)}[log p_\phi(x_i|z)] + D_{KL}(q_\theta(z|x_i)||p_\phi(z)) \tag{1.32}$$

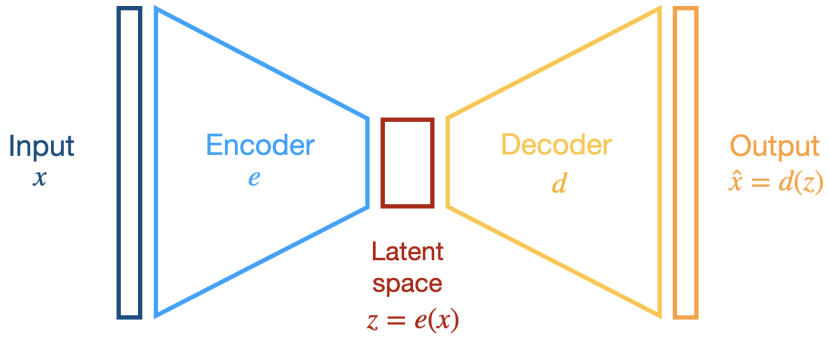


Fig. 1.8 General architecture of a VAE.

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 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. Thus, it describes how close the two distributions are to each other. Quantity $\sum_{i=1}^{N} l_i$ is the total loss evaluated over the whole dataset of N datapoints. Algorithm 3 shows the ELBO's optimization process.

Algorithm 3: Stochastic optimization of the ELBO [80].

Input: $\mathcal{D}$:Dataset
$q_\theta(\mathbf{z}|\mathbf{x})$: Inference model
$p_\phi(\mathbf{x}|\mathbf{z})$: Generative model
Output: ϕ, θ: Learned parameters

Initialize (ϕ, θ) **while** *Stochastic gradient descent (SGD) not converged* **do**
    $\mathcal{M} \sim \mathcal{D}$ (Random minibatch of data)
    $\varepsilon \sim p(\varepsilon)$ (Random noise for every datapoint in $\mathcal{M}$)
    Compute $\bar{\mathcal{L}}_{\phi,\theta}(\mathcal{M}, \varepsilon)$ and its gradients $\nabla_{\phi,\theta}\bar{\mathcal{L}}_{\phi,\theta}(\mathcal{M}, \varepsilon)$
    Update ϕ and θ using SGD optimizer
end

Chapter 2

Method

The present chapter describes the method proposed in the present work. The dissertation aims to generalize the method that will be detailed in further chapters where each case study will be addressed. Thus, this section represents the core of the present work and the common thread of each analysis carried out in different contexts. The method proposed would like to pave the way for a deep statistical processing of the data obtained through a sound level meter.

The flow of this chapter will show each detail of the method. Algorithm 4 shows the whole procedure in the form of a pseudocode at the end of this introduction. Each step will bring the data from its acquisition during the measurement up to their categorization to a specific sound source. Paragraphs follow the algorithm's steps with visual examples to better explain the process. Plots refer for the most to the Gaussian Mixture Model (GMM) despite the work concerning K-means clustering (KM) too. This has been done for two reasons: the first is that, based on the kind of data managed, GMM is visually easier to understand; the second is that, in the present work, GMM is the best algorithm to implement this method.

The entire procedure exploits SPLs and their occurrences during long-term monitoring. Thus, the method can be used with any kind of filter applied to the measurements. Algorithm 4 can be used with equivalent SPLs, band-filtered (one octave, third octave, etc.), and any kind of weights (A, B, C, D, or Z). All this processing on the data does not affect the aim of the method, i.e. finding patterns in realistic contexts to separate sound sources. The ability to exploit any kind of data obtained through a sound level meter would allow technicians to look for different results in different contexts.

The method presented here focuses only on the processing of SPLs. In one of the next case studies, a deep learning approach through a variational autoencoder will be used. However, the aim of measuring and not only classifying sound sources

and the use of digital audio recording does not make the deep learning an appealing tool for the technical applications presented here. Details about all these limits will be discussed in the corresponding case study. For these reasons, the deep learning approach will not be covered in the method section but will be used as a qualitative validation tool for the unsupervised analysis carried out via GMM and KM.

Algorithm 4: the method proposed in the present work.

```
% CA clustering algorithm
% CM candidate models
% SM selected models
% CH Calinski-Harabasz index
% DB Davies-Bouldin index
% SC Silhouette coefficient
% GS Gap statistic
% BM best model
Input: x_i short-time sound pressure levels, f(x_i) target distribution
Output: L_k, k = {1,...,N} set of N clusters
Initialize CA hyperparameters
Initialize L_k = -∞
// first step
for k = 1 : N do
    repeat
        set CA hyperparameters;
    until convergence criterion is met;
    CM(k, x_i)
end
// second step
for k = 1 : N do
    CH(CM(k, x_i));
    DB(CM(k, x_i));
    SC(CM(k, x_i));
    GS(CM(k, x_i));
end
SM(k̄) = {CH(k̄), DB(k̄), SC(k̄), GS(k̄)};
BM = mode(SM(k̄));
// third step
BM(k̄);
for j = 1 : k̄ do
    if labelling condition is satisfied then
        L_N̄;
    end
end
```

2.1 Data acquisition

Regardless of the context of the application, the proposed method is based on the acquisition of large amounts of SPLs. Thus, one or more microphones can be placed in the environment to carry out long-term monitoring. The sensors' placement can be based on the context, the activity, the geometry of the room, and the acoustical properties of the space. All these factors can influence the quantitative results obtained through the measurements and are at the discretion of the operator. A sound level meter with a calibrated class 1 microphone represents the most accurate way to measure SPLs. This kind of microphones was used in each case study addressed in the present work.

One of the first applications of a statistical management of SPLs from a sound level meter, in the meaning of exploiting the occurrences of SPLs, was proposed by Hodgson et al. [62]. In that work, several university lectures were recorded to measure the active sound sources: the ventilation noise, the student activity – i.e. the chatting among students –, and the speech levels of the teacher.

2.1.1 Time interval

The method proposed in the present work is based on the intuition of Hodgson. Different sources have different temporal variations. For instance, the speech has brakes among words and syllables. At the same time, the ventilation system is a quasi-steady noise and fills the pauses of the speech. Thus, an acquisition time frame faster than the speech allows to measure SPLs belonging to the background noise due to the ventilation system.

Figure 2.1 shows how a short acquisition time is able to catch different SPLs associable with different sound sources. In Figure 2.1a a 5-minutes long time history is shown. The recording was made during a university lecture through a sound level meter with an acquisition time of 0.1 seconds. Hence, it can be hypothesized that at least two sound sources were measured: the background noise and the speech of the teacher. A red dashed square in Figures2.1a, 2.1b, and 2.1c shows the time frame stretched and shown in the following figure. Thus, in these figures the time history is stretched from 5 minutes up to 1 second. In Figure 2.1d it is easy to notice how the sound level meter measured a lot of sound fluctuations. Differences between the maximum peak and the lowest values are about 10 dB. More in general, during a long-term monitoring carried out with an acquisition time that is short enough, any sound source can be detected through differences in temporal evolution, speed, and phase.

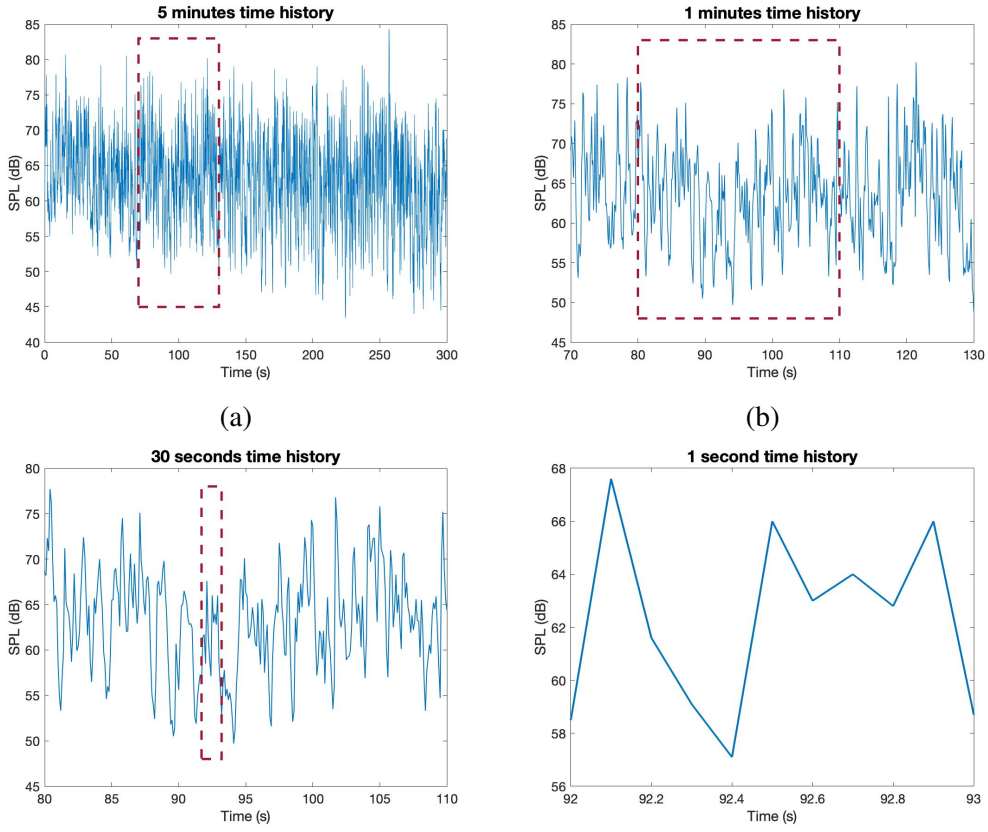
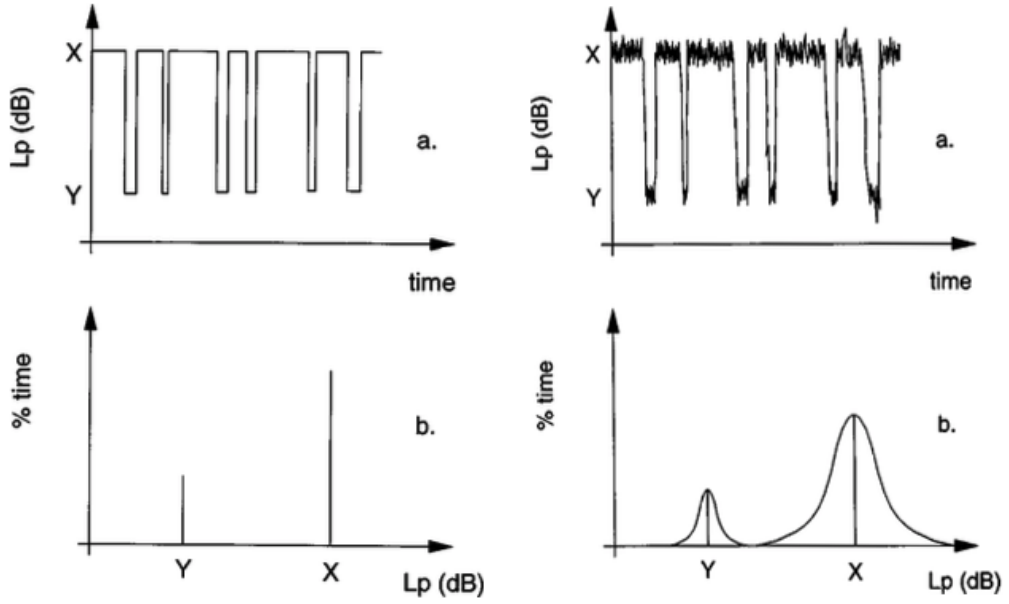


Fig. 2.1 Example of a SPLs time history recorded through a sound level meter. The recording was made during a university lecture. From the top left, a 5-minutes long recording (2.1a), 1-minute long close-up (2.1b), 30-seconds long close-up (2.1c), 1-second long close-up (2.1d). The red dashed squares indicate the frame zoomed in the following figure.

Understanding to what extent a time interval could be deemed as short enough is one of the issues to address in the development of the method. Hodgson used different time intervals depending on the frequency band. This was possible because he recorded an audio file. The acquisition time was set as ten cycles at the lowest frequency of interest. For instance, the time interval of the 50 Hz band was equal to 200 ms. Increasing the frequencies he decreased the length of the intervals up to 1 ms, corresponding to ten cycles of the 8 kHz band. This method optimizes the amount of data to manage.

The method proposed here wants to avoid audio recordings. Privacy is one of the most debated issues in machine learning applications. If technicians deal only with sound level meters and SPLs as outputs, then any sensitive information recording is prevented. Thus, the integration time used in each application of this work is equal



(a) (b)

Fig. 2.2 Example of time history and temporal frequency distribution of SPLs in a classroom. On the left, a idealized lecture. On the right, a more realistic lecture. From "Measurement and prediction of typical speech and background-noise levels in university classrooms during lectures", by M. Hodgson, R. Rempel, S. Kennedy [62].

to 100 ms. This value represents a usual period of time to measure a detailed sound scene catching sounds varying in time and impulsive events [142].

2.1.2 Length of the measurement

Besides understanding how short the time interval should be, it is important to understand how long the monitoring should be. Working with a statistical analysis, it is perceivable that the more data are obtained the better the analysis is made. However, it is important to get useful insights about the minimum length to make the obtained amount of data significant. As stated by Hodgson, considering a theoretical time history in which two sound sources emit each one a particular SPL with a sufficiently high background noise, the respective time history and the associated frequency – in a statistical meaning – distribution are shown in Figure 2.2a. In this case, the occurrences have two spikes. In a more realistic case, all the temporal variations of the sound sources ensure that the occurrence distributions have two different peaks but with non-zero widths, as shown in Figure 2.2b.

As stated in the introduction of the present work, machine learning deals with large amounts of data. Hence, many ML algorithms for processing data are based

on statistics. The ML-based analysis proposed here starts with the visualization of data to identifying the best approach to manage the obtained distribution. Datasets of long-term sound level meter monitoring are constituted by SPLs, i.e. 10 times the logarithm of the squared ratio of a given sound pressure to the reference sound pressure. Thus, the dB can be considered as a log-transformation of the sound pressure. As a consequence, it is expected that a SPL has a Normal tendency [14]. It follows that SPLs of a single sound source will follow a normal distribution as long as the monitoring lasts enough for the central limit theorem to hold. Based on the considerations above, the use of Gaussian distributions to fit the occurrences curve of a single sound source is consistent with the kind of data [62].

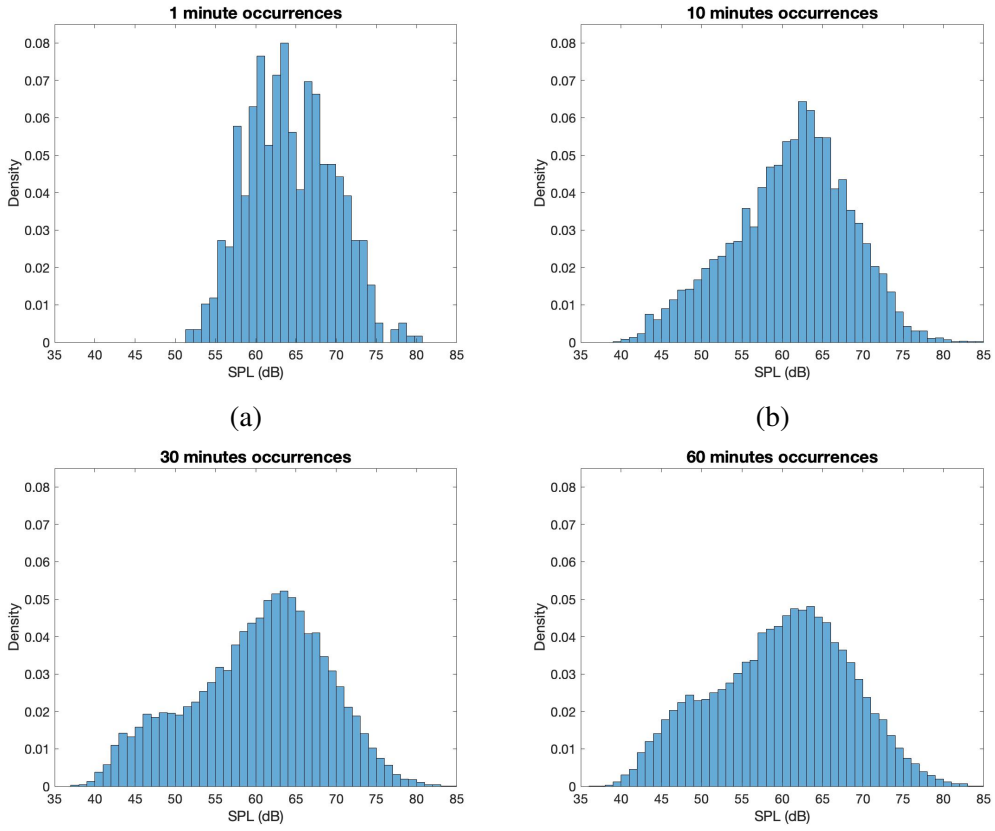


Fig. 2.3 Example of the occurrences collected during a long-term monitoring of a university lecture. From the top left, the occurrences distribution collected after 1 minute (2.3a), 10 minutes (2.3b), 30 minutes (2.3c), and 60 minutes (2.3d) of recording.

Figure 2.3 shows how the length of the monitoring affect the shape of the occurrence curve. Histograms are created with bins around 0.5 dB, which is about the uncertainty of a class 1 microphone. The recorded event is represented by a university lecture. Thus, it is expected to have at least two peaks represented by

the background noise on lower SPLs and the teacher's speech with higher SPLs. Figure 2.3a shows the occurrences obtained after a recording 1-minute long. Then, Figures 2.3b, 2.3c, and 2.3d show the occurrence curves recorded after 10, 30 and 60 minutes. Based on the intuition described above, it is possible to notice how 1 minute is enough to only achieve a preliminary idea on the tendency of the teacher's speech. The background noise becomes noticeable after 10 minutes and more clear after 30 minutes. The last plot shows how the occurrences related to the background noise increase. It is interesting to notice how the speech levels density decreases but the peak remains on the same x values. The peaks are one of the most important features of this curve because they are related to the corresponding SPL of the sound source. Moreover, long-term monitoring allows to ignore any kind of impulsive noise because, being an outlier, it would not affect the shape of the curve until it becomes continuous enough in time.

2.2 Data processing

The previous section discussed the way to collect as more data as possible through a sound level meter in a realistic environment. The aim is to achieve the best temporal resolution using the shortest time interval. Under the assumptions that single sound sources have Gaussian distributions, we expect an occurrence curve which is asymmetrical with different peaks. Any kind of skewness of the curve suggests the presence of moer sound sources with a low signal-to-noise ratio with respect to the visible peak. The higher the background noise is, the more distant the peaks are.

Starting from the Gaussian assumption, the present work investigates two different algorithms: the Gaussian Mixture Model (GMM) and the K-means clustering (KM). The reason for investigating two different algorithms lies in the homoscedasticity, i.e. the homogeneity of variance. As seen in Section 1.1.3, the GMM can be considered as a generalization of KM for small variances [94].

2.2.1 Step 1: Occurrence curve and candidate models

The data processing follows a few steps, which are presented here through visual examples. The first step concerns the collection of the occurrences. Figure 2.4 shows an example of occurrences. Here, the measured SPLs are shown in pale blue. The bins' visualization is enough for KM because it is a distance-based algorithm. Concerning the GMM, after collecting the SPLs, it is important to draw the envelope

of the distribution, i.e. the probability density function. This is shown with a blue line.

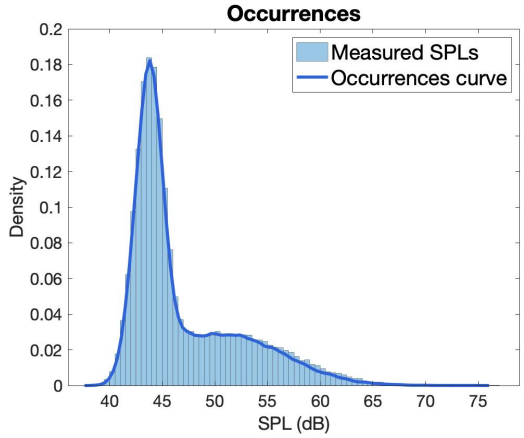


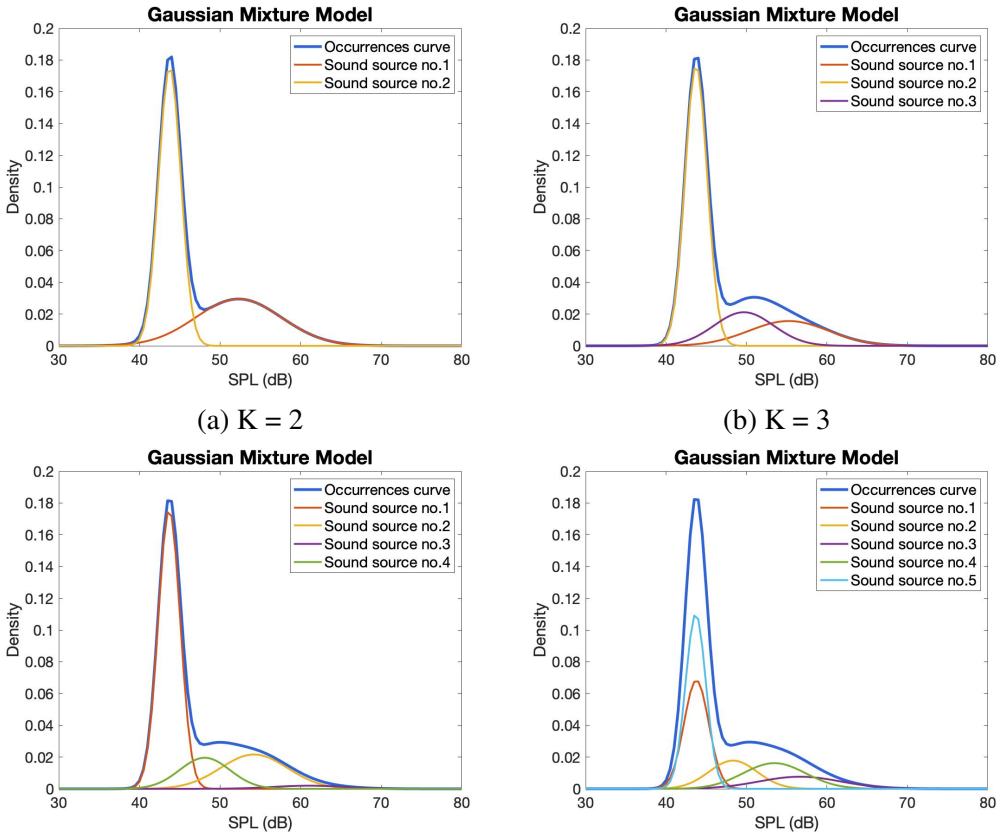
Fig. 2.4 Example of the occurrences collected from a long-term monitoring of an active office and the corresponding occurrence curve obtained.

In ML, a supervised problem, i.e., with labelled data, has a target that the model tries to predict. Knowing the target and exploiting the labels it possible to infer accuracy metrics to assess the goodness of the prediction. As stated in the previous chapter, GMM and KM perform unsupervised learning, i.e., data are not labelled. Hence, the aim is to look for patterns in the data. However, the number of clusters K has to be set before running the algorithms.

In this regard, to understand effectively how many clusters are in the data, the following step concerns the setting of the so-called *candidate models*. GMM and KM are run different times with a different number of clusters K. Both algorithms are set to repeat the iterative process multiple times using different set of initial values. This option prevents the algorithms for being trapped in undesirable local maxima. Further, the covariance type of GMM has to be set to let the components independent to adopt any position and shape [97]. Figure 2.5 shows different GMM candidate models. Starting from the upper left, the same occurrence curve shown in Figure 2.4 is processed via GMM with different values K from 2 up to 5. Each number of clusters represents the number of possible sound sources. The same procedure is followed with KM.

2.2.2 Step 2: Model selection

At this point, it is needed to choose the best model among candidates. As seen in the previous paragraph, an unsupervised analysis via GMM and KM cannot exploit an accuracy metric. Thus, it is needed to look for other types of measurements as



(a) K = 2 (b) K = 3

(c) K = 4 (d) K = 5

Fig. 2.5 Example of candidate models of active office long-term monitoring. The occurrence curve of Figure 2.4 has been processed via GMM for different numbers of clusters K. From the top left $K = 2$ (2.5a), $K = 3$ (2.5b), $K = 4$ (2.5c), $K = 5$ (2.5d).

indicators of performance. In cluster analysis, the most common way to assess the performance of the algorithm is to measure either the distinctiveness or the similarity between the clusters obtained. These two clusters' properties are commonly defined as the *cohesion* within each cluster and the *separation* among different clusters. In the present work, the selection of the best model is made through 4 different metrics: the Calinski-Harabasz index (CH), the Davies-Bouldin index (DB), the Silhouette coefficient (SC), and the Gap statistic (GS). The theoretical approaches of each of these metrics are described in Section 1.2.

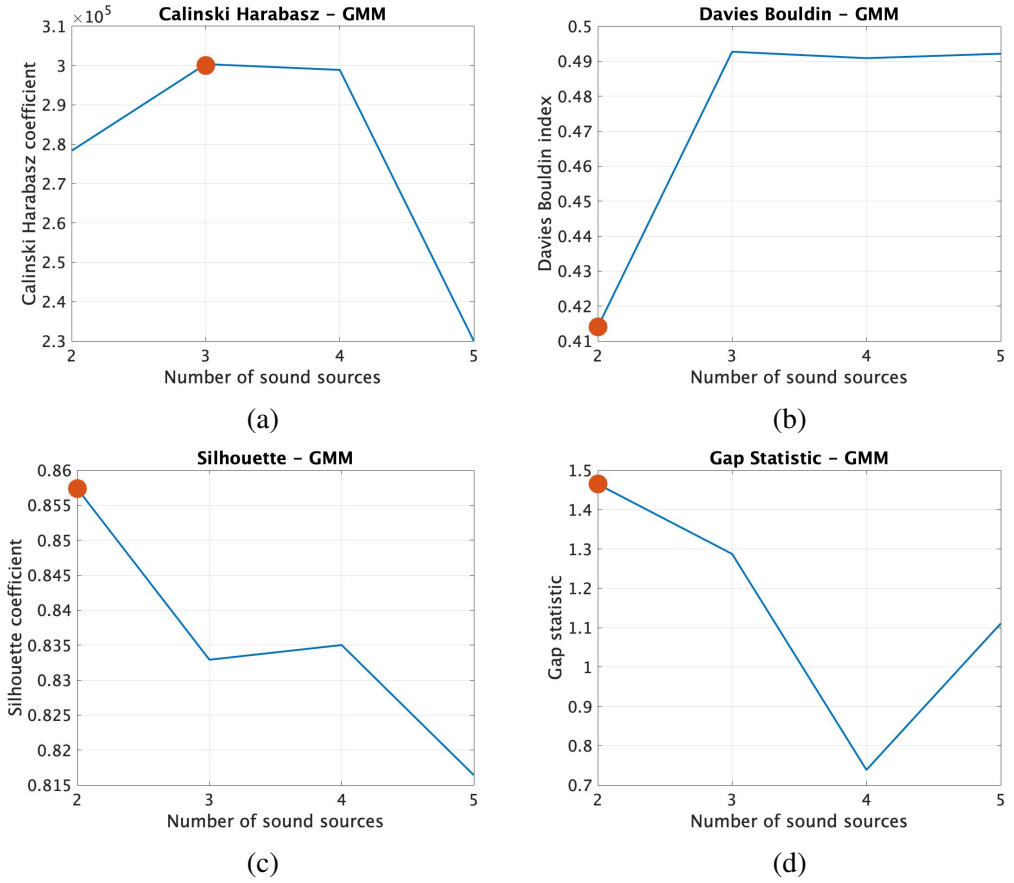
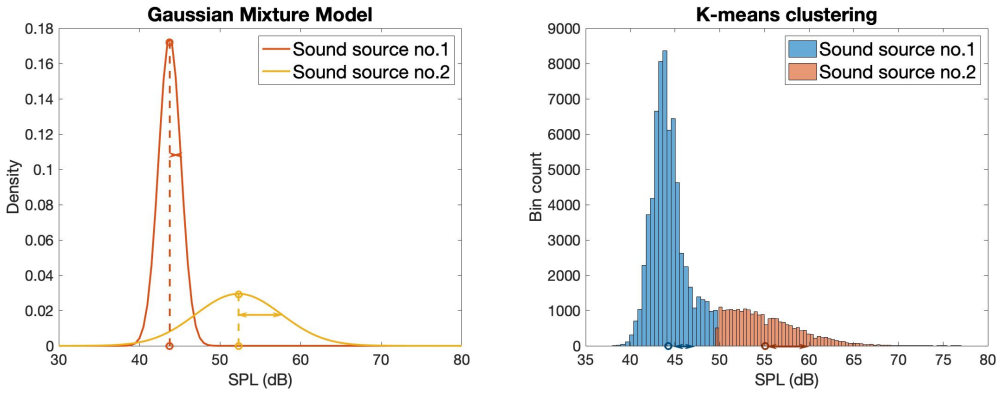


Fig. 2.6 Example of model selection. The candidate models shown in 2.5 have been evaluated via the Calinski - Harabasz index (2.6a), the Davies - Bouldin index (2.6b), the Silhouette coefficient (2.6c), and the gap statistic (2.6d). Red dots indicate the best model selected by the metric.

Considering that all these metrics assess the models in different ways, the best model is chosen according to the majority rule among the metrics. In case of a tie, the results of the model selection should be compared among all the spectral data. Achieving an optimal number of clusters different in only one third/octave band would refer to a masked source with a predominant energy contribution in that band. Thus, the most frequent model selected by the different criteria will be considered as



(a) (b)

Fig. 2.7 Example of selected clusters obtained via the GMM (2.7a) and the KM (2.7b). Dots and arrows indicate respectively means and standard deviations for the GMM, and centroids and average intra-cluster distances for KM.

the *best* model. Figure 2.6 shows the evaluation performed by means of the 4 metrics over the candidate models shown in Figure 2.5. The x-axes indicate the number of possible sound sources, i.e. the different candidates, and the red dot shows the best model according to that metric. CH gives back $K = 3$ (Figure 2.5b) as the best model among candidates, whereas DB, SC, and GS indicate that the best model is made by two clusters (Figure 2.5a). Thus, according to majority rule, the best model is $K = 2$.

2.2.3 Step 3: Labelling and spectra reconstructions

After the best model has been selected, it is possible to investigate the results of the clustering analysis. Following the example shown so far, the best model is represented by $K = 2$. Assuming that the same result has been achieved for KM, too, Figure 2.7 shows the best model selected via GMM and KM.

Concerning the GMM, on the left, the two Gaussian curves, i.e. the two clusters, represent the two sound sources found through the iterations of the EM algorithm. The use of well-known distributions is fundamental in unsupervised techniques. A Gaussian curve is univocally defined by two parameters: mean and variance. Through these parameters is possible to describe the clusters obtained via GMM, i.e., the sound sources. After starting from a bunch of unlabelled data, i.e. SPLs, it is possible at this point to characterize different groups with identifiable properties. In Figure 2.7a, for each Gaussian curve, we show its mean and standard deviation. The first are indicated with dots and the projection on the x-axis through a dashed line; the second are identified by arrows.

The two clusters obtained via KM are represented in pale blue and orange in Figure 2.7b. Despite their theoretical relationship shown in the previous chapter, GMM and KM have different foundations. GMM is a model-based algorithm with the model represented by the Gaussian distribution; KM is a distance-based algorithm, i.e. clusters are shaped optimizing specific distance metrics. Specifically, in this work, the metric chosen is the squared Euclidean distance. However, the features obtained via KM can be deemed as similar in comparison to means and standard deviations obtained via GMM. Keeping the analogy, it is possible to compare the mean to the so-called *centroid*, i.e. the centre of gravity of the cluster. The standard deviation is compared to the average intra-cluster distance (AICD), i.e. the average distance of each point-to-centroid belonging to the same cluster. On the x-axis of Figure 2.7b, the dots and the arrows indicate respectively the centroids and the AICDs of the two clusters.

Means and standard deviations are the features used to label the clusters obtained via GMM, centroids and AICDs are the features obtained via KM. Means and centroids represent whole clusters with values in dB. Thus, we can deem means and centroids as the representative SPLs of each sound source. Similarly, standard deviations and AICDs represent the temporal variability of a sound source. Tight Gaussian curves or short distances mean a high repeatability of a source's sound emission, e.g. mechanical noise due to a ventilation system. Next chapters will address in details the dB thresholds of standard deviations and AICDs to label mechanical, traffic, or human sources. More generally, these two quantities represent statistical – s.d. – and metrical – AICD – properties useful to understand and label the type of sound source measured. Table 2.1 sums up the features that is it possible to extract via GMM and KM.

Table 2.1 Features obtained via GMM and KM to label different sound sources.

Algorithm	**Feature**	**Physical meaning**
GMM	Mean	SPL of a sound source
	Standard deviation	Extent of randomness of a sound source
KM	Centroid	SPL of a sound source
	Average intra-cluster distance	Extent of randomness of a sound source

At this point the analysis is complete. However, it is not straightforward to understand whether a sound source could be a mechanical system or the traffic only assessing the standard deviation or the AICD. Moreover, technicians often work in the frequency domain. Thus, it becomes important to understand how this method works in frequency. As outlined in the introduction of this chapter, the method

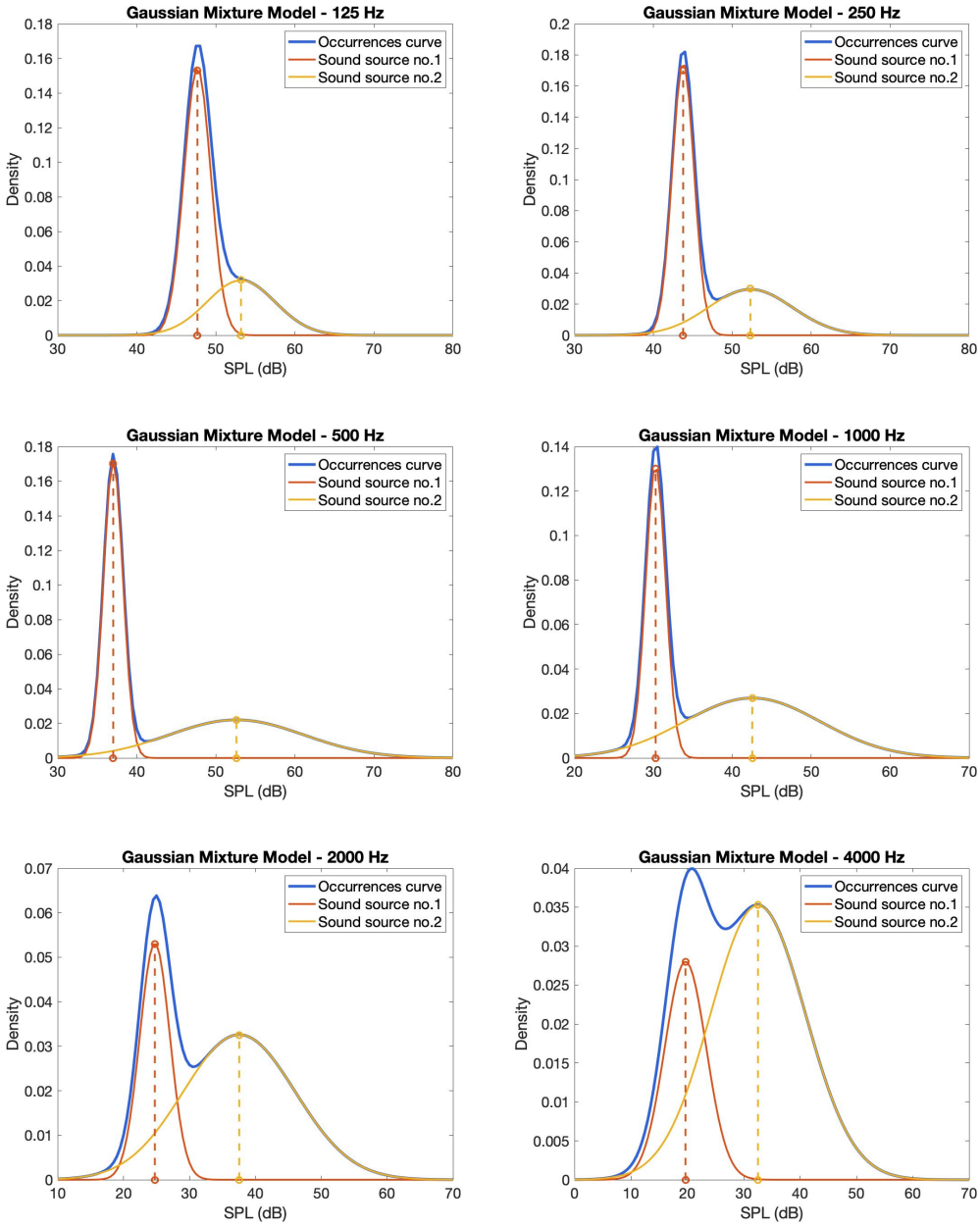


Fig. 2.8 Octave band occurrence curves of a long-term monitoring inside an active office processed via GMM from 125 Hz (on the top left) up to 4 kHz (on the bottom right). Dots indicate the means per each cluster.

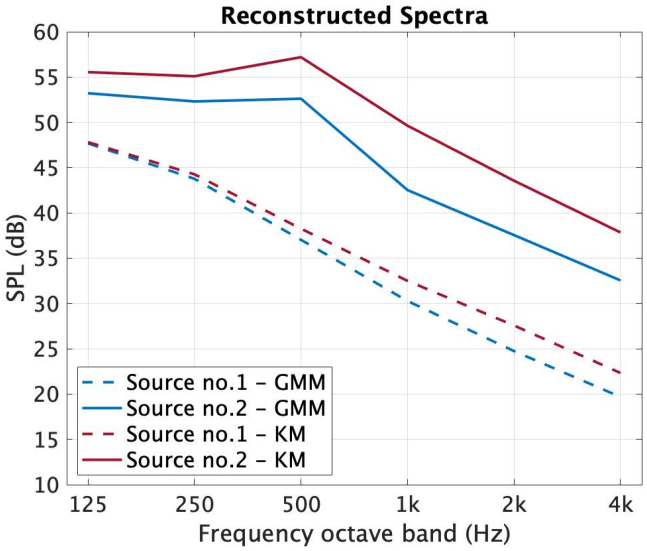


Fig. 2.9 Example of spectra reconstructed inside an active office. Red and blue lines show respectively the spectra obtained via GMM and KM. Dashed and solid lines show respectively the tendencies of source no.1 and source no.2.

proposed in this work can be used with any kind of SPL obtained through a sound level meter. Hence, it is possible to repeat the whole method per each frequency band under study. The end of the analysis will give back different SPLs and standard deviations or AICDs. Noticing similar feature values, it is possible to find a common thread across frequencies. Thus, it is possible to associate each SPL with the same source and reconstruct the spectrum.

Figure 2.8 shows the occurrence curve octave-band filtered and processed via GMM from 125 Hz up to 4 kHz. Blue lines represent the occurrence curves, the orange and yellow Gaussian distributions represent the two clusters obtained via GMM, source no.1 and source no.2. The dashed lines show the projections of the means on the x-axes. To simplify the visualization of the curves, the scales on both axes change. Looking at the curves, it is possible to notice how there are similarities among all the orange curves: low SPLs and small standard deviations. Moreover, the occurrence density remains high up to 500 Hz and starts to decrease in mid-high frequencies. All these clues suggest a mechanical behavior of the component in each octave band. Thus, sound source no.1 could be labelled as an example of mechanical source. Yellow curves are all similar as well: high SPLs and large standard deviations with an almost constant density of occurrences, about 0.03. These features represent a pretty random source with high temporal variations. Thus, sound source no.2 could be labelled as human voice. Collecting all SPLs per each octave band per each sound source it is possible to reconstruct both spectra. Figure 4.12 shows the reconstructions. Red and blue lines show, respectively, the results obtained via

GMM and KM. Dashed and solid lines show respectively the trends of source no.1 and source no.2. At first glance, the deductions about the mechanical and human voice seem to be confirmed by the tendencies of the spectra. However, details about spectral tendencies, values, and uncertainties are discussed in next chapters where different case studies are presented.

Chapter 3

Applications in classrooms

Keywords: *student activity, speech intelligibility, classroom acoustics, PA design.*

Classrooms and university lecture halls represent the first context in which test the statistical method. The reason lies in the need for realistic monitoring of the sound environment. In fact, acoustical comfort in learning space is the basic condition to provide an effective and successful learning process [81, 98]. Students comfort is strongly related to the acoustic condition of the environment. The communication quality allows students to be more focused during lectures and affects their cognitive tasks, besides improving the teachers' comfort as well decreasing their vocal effort [115, 20].

Speech is the main method of communication in learning spaces. Thus, the learning process broadly involves the speech intelligibility, i.e. the rating of the proportion of speech that is understood [69]. Basically, the intelligibility issue can be influenced by different factors:

- the frequency response of the room;
- the sound energy distribution throughout the space;
- the sound pressure level of emission;
- the background noise;
- if available, the frequency response and directivity of the public address (PA) system.

More generally, the speech signal is degraded by the path or the transmission channel between source and receiver. The assessment of speech intelligibility can be made through different metrics [18]. One of the most common way to measure

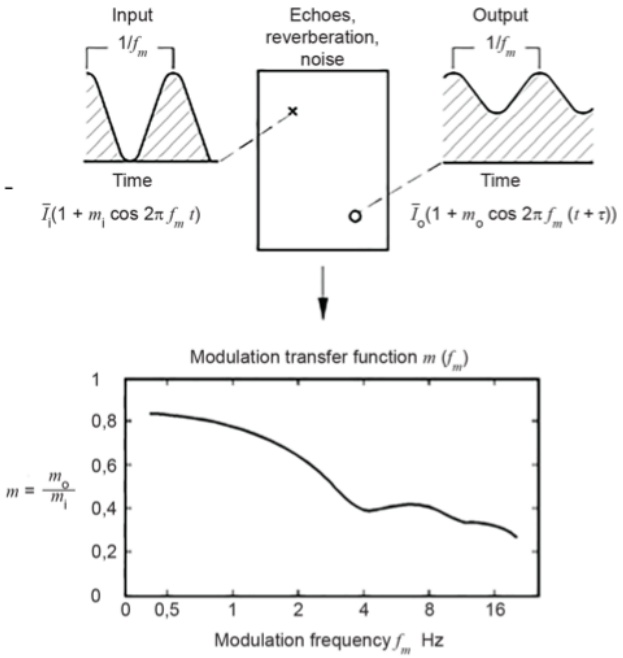


Fig. 3.1 Modulation transfer function – input/output comparison. Parameters m_1 and m_2 are the modulation depths of respectively input and output signals; $\bar{I}_1$ and $\bar{I}_0$ are the input and output intensities. From IEC 60268 (2020). Sound System Equipment – part 16: Objective rating of speech intelligibility by speech transmission index [69].

is the Speech Transmission Index (STI), described in IEC 60268-16 standard [69]. Introduced and developed by Houtgast and Steeneken, the STI is an objective method to assess the quality of the communication channel that may be affected by several acoustic and electro-acoustical distortions. The speech is outlined as a signal that fluctuates in time and in the intensity envelope of the sound. Slower fluctuations of the envelope describe word and sentence boundaries, faster fluctuations correspond to individual phonemes within words. Phonemes are considered as the unit element of the speech and connected discourse. Thus, the preservation of the intensity envelope results in achieving high intelligibility. STI assesses to what extent the intelligibility is preserved through a value between 0 and 1. Its calculation involves weighted contributions from seven octave frequency bands where the speech energy is significant. Summing, a modulation transfer function determines the degradation of the intensity envelope and the signals' fluctuations due to the transmission channel [67, 134, 135]. The STI model describes an ideal situation in which a talker with a standardised male speech spectrum is speaking with good articulation, i.e. with at a nominal word rate of about 3 to 4 syllables per second, and assumes listeners have normal hearing. However, humans alter the way they speak and hear according to

many factors, like the age, gender, native language and social relationship between talker and listener. Speech intelligibility may also be affected by pathologies such as speech and hearing disorders. For non-native speakers/listeners and for listeners with hearing loss it is possible to bring corrections according to IEC. Figure 3.1 shows how the speech signal is affected by the acoustic properties of the space. The upper part of the picture shows the input position in the room with a cross and the receiver with a circle. In the lower part, the modulation transfer function is evaluated for each modulation frequency in each octave band.

The signal-to-noise ratio (SNR) is the difference in dB between the signal and the background noise. Assuming that the signal is loud enough to neglect the SPL of emission, the intelligibility problem can be outlined as influenced by two factors: the acoustic properties of the room and the background noise [62]. In this view, the sound context of a learning space is represented by the SNR between the speech SPL of the teacher (SL) and the background noise. The latter is made up of several contributions: the mechanical systems, the traffic, the activities carried out in neighbouring spaces, and the noise generated by the students, called the student activity (SA) [62]. In some cases of lecture halls with high attendance, the SA overwhelms other noise contributions. Hence, the SNR is reduced to the difference between SL and SA.

The measurement of SA during lectures has been broadly debated among scholars. One of the first approaches concerned the assumption of the SL equal to the equivalent level L_{eq} and SA equal to the 90th statistical level L_{90}, i.e. the SPL exceeded the 90% of the measurement time [95, 83, 132, 13]. Hodgson et al. introduced the statistical assumption and the use of GMMs without considering the model selection step. The number of clusters to look for was chosen through a visual data exploration [62]. In the present chapter, the GMM with no model selection step has been called *Peak Detection* (PD). The same technique was used by Sato to evaluate the noise inside elementary classrooms. Here, there is no distinction between background noise and SA. Both sources were gathered in the same Gaussian curve [128]. Choi used PD as well to measure SAs, SLs, and SNRs in 15 different lectures in 11 university classrooms [27]. More recently, unsupervised algorithms used in machine learning improved the ability to investigate long-term monitoring. Besides the case studies presented in the next sections of this chapter, Wang and Brill used the K-means clustering algorithm to detect whether K-12 classrooms were active or not [147]. All these studies are summarized in Table 3.1. Here, as much information as possible about previous studies is shown, such as the school's grade, the hall type, the number of lessons and positions used during the measurements, the potential presence of a PA, the resulting SNRs and the standard deviations (s.d.), and the time interval.

Table 3.1 Comparison of SA measurement studies. For each study the grade of the school, the size of the room (classrooms have an occupancy of approximately one to fifty people and lecture halls host in the hundreds), the number of the rooms and the lessons, the number of measurement positions used, the analysis technique adopted, the signal-to-noise ratio (SNR), in dB, the standard deviation (s.d.), in dB and the length of window integration, in ms, are reported. Indents mean missing data from the cited studies. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

Ref.	Grade	Hall type	Rooms/Less.	Pos.	Method	P.A.	SNR (dB)	s.d. (dB)	W (ms)
[66]	High school	Classrooms	10/–	1	–	–	9.5	4.6	–
[95]	Elementary	Classrooms	12/–	1	PL	–	-4.5	–	–
[62]	University	Classrooms Lecture halls	11/18	3	PD	No	7.9	3.1	200 – 1
[128]	Elementary	Classrooms	27/27	4	PD	No	11.1	2.5	200
[83]	Elementary	Classrooms	4/–	9	PL	Yes No	13.0 2.0	–	850
[13]	Elementary	Classrooms	–/54	1	PL	No	–	–	50
[132]	Secondary	Classrooms	80/274	1	PL	No	14.8	4.6	–
[147]	University	Classrooms	220/–	–	KM	No	16.9	–	–
[112]	Elementary	Classrooms	46/59	$1 \div 2$	GM	Yes No	11.6 9.2	–	200
[27]	University	Classrooms	11/15	–	PD	No	7.7	2.4	200
[43]	University	Lecture halls	3/12	2	PL PD GM KM	Yes	18.1 16.3 15.3 15.4	3.2 2.5 3.1 3.2	100

The *Lombard effect* links the behavior of speakers, i.e. teachers, and students. This effect is a psychoacoustic involuntary tendency of speakers to increase the volume of their voice in noisy contexts [91]. This change includes not only loudness but also other acoustic features such as pitch, rate, and duration of syllables [136]. In learning spaces, all the sound sources and noise contributions are linked. With a high background noise due to the ventilation system, students could increase their levels of SA. Thus, teachers increase their SLs, i.e. their vocal effort [2]. All these reasons prove the need for a method to separate the different contributions of sound sources.

An SNR higher than 15 dB is considered necessary to neglect the influence of the background noise on the intelligibility [16, 62, 5, 59]. Beyond this value no improvements in quality occurred. However, in case of cognitive disabilities the advised SNR is equal to 20 dB [104, 45]. The use of a PA system represents one of the most common ways to ensure the achievement of these values, especially in large lecture halls. Nevertheless, the introduction of a PA does not override the need of a proper acoustic design of the space, even though it could help in avoiding the Lombard effect when used [148, 112]. Moreover, high occupations reduce the reverberation time because of the absorption of the students. Thus, it can be assumed that in large university lecture halls with low reverberation time, the intelligibility can depend only on SNR.

This chapter shows two applications of the method described in the previous sections. The main focus of the discussion concerns the measurement of SA. The understanding to what extent students are distracted during lectures can be deemed as a metric of the goodness of the rooms' acoustical properties. The first case study addresses the measurement of SA through different methods: equivalent and statistical levels – this method will be named PL – , PD, GMM, and KM. Preliminary discussions about the method will provide first insights about the difference among the techniques used by scholars. The SA can be considered as an objective metric to assess a subjective behavior of an entire group, i.e. the students. For this reason, the second case study will use SA as the main metric to evaluate the effectiveness of acoustic treatments in lecture halls. SA was measured in two of the three halls shown in the first case study before and after acoustic treatments and the PA redesign. Measurements were made only via GMM and KM and further discussions will address differences between methods and correlations among all the active sound sources during lectures besides spectral detailed studies.

3.1 Comparison among methods to measure student activity

3.1.1 Description of the halls

The first case study is represented by 3 historical lecture halls of the School of Literature and Philosophy of the University of Bologna. Students and teachers changed after each lesson. The halls were chosen because of their high occupancy and variability of lessons during the whole day. Hall I and Hall II are historical rooms with an amphitheater geometry and a volume equal to 1000 and 900 m^3 respectively; they have plastered walls and wooden seats and benches able to host up to 250 and 200 students respectively. Hall III has an approximately regular shoe-box shape with a volume of 850 m^3, except for the overhead coupled volumes between the ceiling and the false ceiling; its surfaces are plastered while seats are movable and made of plastic with a maximum occupancy of 170 students. Hall III hosts occasionally non-traditional teaching activity, like theatre rehearsal.

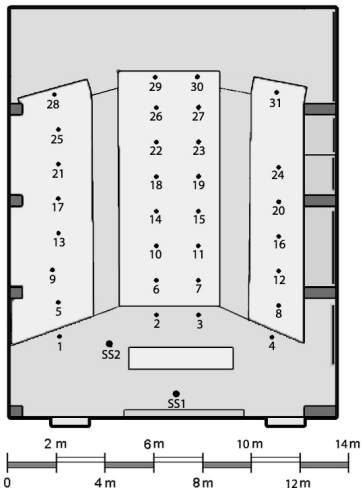
A preliminary measurement campaign was carried out in the halls under study aimed at qualifying the acoustical properties in empty state. Measurements were carried out following procedures and equipment according to ISO 3382-2 standard [72]. Monaural impulse responses were acquired with an Exponential Sine Sweep signal with a length of 512 K and sampled at 48 kHz. A high-SPL dodecahedron was used as the omnidirectional sound source [41]. The source was calibrated in the reverberation room according to ISO 3741 [74]. The variable occupancy by the students influence the total absorption area in the halls [26]. To consider the influences of students, the reverberation times in occupied condition were evaluated using the equation [139]:

$$T_{occ} = \frac{T_{unocc}}{1 + \frac{T_{unocc} C N \Delta A_{1p}}{0.16V}} \qquad (s) \tag{3.1}$$

where N is the maximum occupancy of the hall and C is the percentage of occupancy ($C = 1$ means a full occupied hall, $C = 0.8$ means an occupancy of 80%). ΔA_{1p} is the increase of the equivalent absorption area due to one person in m^2 *Sabine*. Values of ΔA_{1p} are taken from Appendix C of the Italian acoustic regulation for classrooms UNI 11532-2 [139]. Figure 3.2 shows pictures and plans of the lecture halls under study. Besides the graphical scale to give the proportions among geometries, the positions of source-receivers pairs is shown.

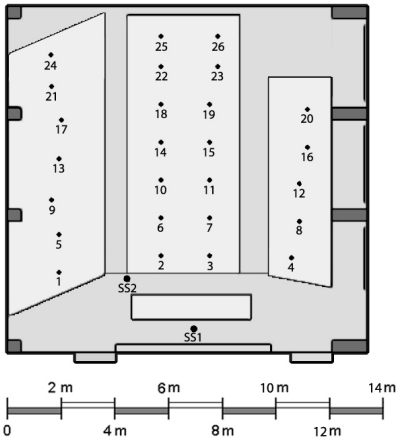
Table 3.2 shows a general overview of the properties of the halls. It shows both geometrical data – i.e. shape type, volume, maximum occupancy, and seating area –

(a) Hall I



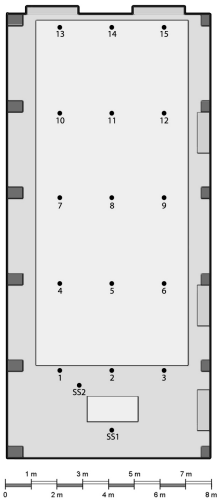
(b) Hall I – plan

(c) Hall II



(d) Hall II – plan

(e) Hall III



(f) Hall III – plan

Fig. 3.2 Pictures and plans of the three university lecture halls under study: Hall I (3.2a, and 3.2b), Hall II (3.2c, and 3.2d), and Hall III (3.2e, and 3.2f). Plans show sources and receivers positions.

and the reverberation in different occupancy configurations – i.e. unoccupied, 30%, 80%, and 100% of the maximum occupancy.

Table 3.2 Data overview of the lecture halls and ISO 3382-2 measurements results [72]; where: "V" is the volume, "N" the maximum occupancy, "S_A" is the audience area, "$T_{M,unocc}$" is the reverberation time in unoccupied condition, "$T_{M,occ\ 30\%}$" is the reverberation time in occupied condition at 30%, "$T_{M,occ\ 80\%}$" is the reverberation time in occupied condition at 80% and "$T_{M,occ\ 100\%}$" is the reverberation time in occupied condition at 100%. The subscript "M" means a value averaged over all the receivers in the octave bands of 500 – 1000 Hz. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

Hall	Type	V (m^3)	N	S_A (m^2)	$T_{M,\ unocc}$ (s)	$T_{M,\ occ\ 30\%}$ (s)	$T_{M,\ occ\ 80\%}$ (s)	$T_{M,\ occ\ 100\%}$ (s)
I	Amphitheater	1000	250	100	1.70	1.27	0.90	0.80
II	Amphitheater	900	200	100	1.72	1.34	0.95	0.90
III	Shoe-box	850	170	81	2.54	1.88	1.22	1.19

3.1.2 Measurement methods

The SA measurement method, as mentioned in the introduction of this chapter, is debated among scholars. To compare each method used in literature,12 university lectures were measured in 3 different halls. Two sound level meters were placed in the middle of the audience area on either side at a height of 1.2 m, i.e. the height of ears of a seated person. The positions were chosen maintaining a distance of at least 1 m from any surrounding surface. An operator attended the recorded lectures to report the activities and notice potential peculiarities in the dataset, e.g. peaks due to unexpected sounds. A-weighted short-time SPLs were recorded with an interval time of 100 ms.

Each lesson lasted about 90 minutes. Breaks and intervals among lectures were cut to focus the analysis only on the active lesson time. Thus the post-processing started from the selection of the active time from the time history. Figure 3.3 shows an example of selection of the active time after the lessons were recorded.

After cutting the time history, SPLs were collected in different datasets, one per each lesson. According to the literature, each dataset was explored with four methods categorized as *visual* or *blind*. Visual methods exploit the observations made by the operator during lessons to analyze the time history; blind methods look for patterns in data without any knowledge of the realistic environment. The categorization of the

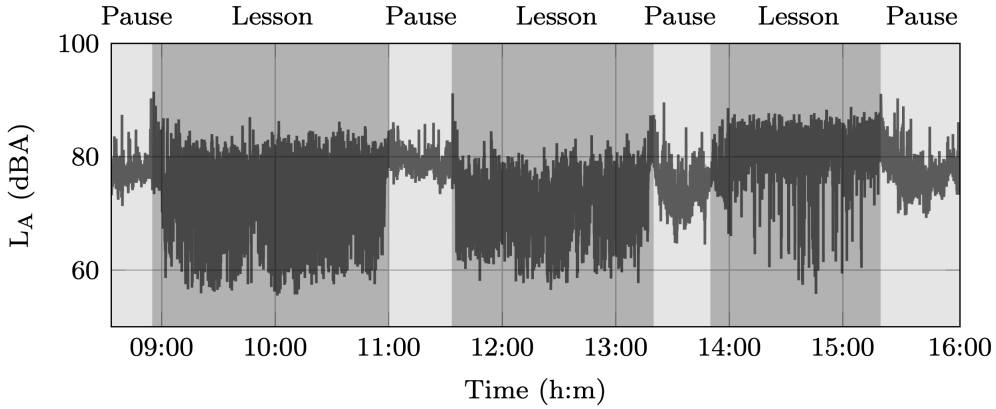


Fig. 3.3 Sample of the time history measured during a university lecture. Different shades of grey show the portions deemed as pause and lesson times. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

methods, recollecting their assumptions through a brief description, can be outlined as:

Visual methods

- **Percentile levels (PL)**: SA and SL are directly extracted from the sound level meter. The A-weighted equivalent level $L_{A,eq}$ is associated with the teachers' SL, and the acoustical 90th percentile level L_{90} with the SA. The term *acoustical percentile* refers to a different use of statistical levels in acoustics with respect to the traditional statistical language. The difference lies in the definition of percentile. In acoustics, a n-th percentile level is the SPL exceeded for the $n\%$ of the measurement time. In statistical terms it corresponds to the $100-n$ statistical percentile. For instance, the 90th acoustical percentile L_{90} corresponds to the 10th statistical percentile of the data distribution. Percentile levels are obtained from the whole dataset collected through the long-term monitoring.
- **Peak detection (PD)**: skipping the model selection step typical of data clustering, it is possible to assume a certain number of Gaussian curves to fit the original distribution of data. In this work, the multi-peak analysis and corresponding curve fitting was made via *OriginLab* software. Through the graphical interface it is possible to detect geometrical boundaries, i.e. peaks. With this approach the algorithm is forced to fit the data with Gaussian curves having the means nearby the given peak.

PL and PD use two different data distributions to perform calculations. The first uses the cumulative distribution function (cdf) of the measured SPLs, and the second uses the probability distribution function (pdf). For discrete variables, cdf and pdf are linked. The pdf can be expressed as the derivative of a cdf [49, 9]. Generally, the probability that x lies in the interval $(-\infty, z)$ is given by the cdf and is defined as:

$$P(z) = \int_{-\infty}^{z} p(x)dx. \tag{3.2}$$

The statistical rank r of a percentile q of N data observations is defined as:

$$r(q) = \frac{q}{100}(N+1). \tag{3.3}$$

Assuming that N is large enough, the acoustical percentile level L_q is equal to the rank of $100-q$. Defining $f(x)$ as the pdf of the observations, the value q can be expressed as:

$$q = P(x > L_q) = \int_{r(100-q)}^{\infty} f(x)dx. \tag{3.4}$$

Thus, according to Equation 3.2 and considering that *acoustical* percentile means doing a backward integration of the pdf, the cdf $g(x)$ can be expressed as:

$$g(x) = \int_{\infty}^{x} f(x')(-dx') = \int_{x}^{\infty} f(x')dx' = 1 - \int_{-\infty}^{x} f(x')dx'. \tag{3.5}$$

Blind methods

- **Gaussian Mixture Model (GMM)**: a model-based algorithm that describes a generic distribution as a sum of Gaussian curves. An iterative probabilistic process via Maximum Likelihood assigns data to each cluster. The Expectation-Maximization algorithm maximizes the likelihood function through iterations [38]. Each datapoint is assigned to clusters with probability weights. Clusters are defined by Gaussian curves and are shaped when convergence is reached. Theoretical details are described in Section 1.1.1.
- **K-means clustering (KM)**: a distance-based algorithm that shapes clusters grouping the closest points. In this work, an iterative process minimizes the squared Euclidean distance among data and centroids, i.e. the center of gravity of the cluster. Each iteration updates the redistribution

of data towards the nearest mean among data within the same cluster. The process is repeated until convergence is reached, i.e. when the centroids are no longer updated. Theoretical details are described in Section 1.1.2.

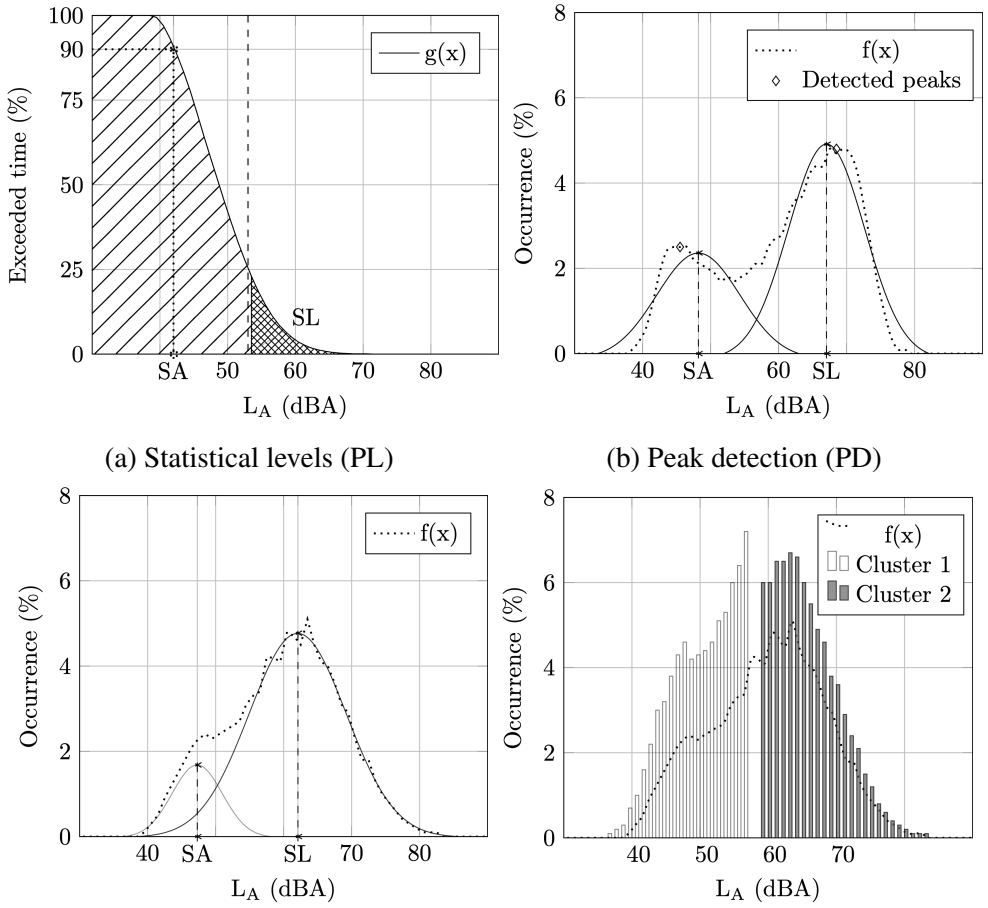
Figure 3.4 shows graphically the different approaches. Figure 3.4a shows the cdf indicated as $g(x)$. The equivalent level L_{eq} corresponds to the SL and is highlighted by the line that divides the two patch areas. The 90th statistical level corresponds to the SA highlighted by the dotted line. The L_{eq} divides the energy of the plot in two areas: the cross and diagonal patches. The energies within these two areas are equal. Thus, the PL's approach can be considered energetic-based and completely different from the other methods. PD, GMM, and KM use SPLs as random variables, indeed. However, despite the basic difference, PL is compared with others because it represents the common standard used by the technicians praxis. Figures 3.4b, 3.4c, and 3.4d show the PD, GMM, and KM approaches where the original pdf is indicated as $f(x)$. Concerning the PD, empty diamonds indicate the geometrical boundaries where the fitting is constrained. Figures 3.4c and 3.4d show the same pdf analyzed via GMM and KM.

The statistical levels used in PL were extracted using the post-processing commercial software 01 dB dBTrait. Supervised PD was made through the OriginPro software. Concerning the GMM, the maximum likelihood estimates of the parameters were derived via the EM algorithm, by using the `Mclust` function from the homonym R package. The algorithm to implement KM was the `kmeans` function of the stats R library.

3.1.3 Results and discussions

Each lesson lasted about 90 minutes and was managed individually. Thus, each dataset had an average of 52k samples. Half of the lectures were conducted by males and the other half by females. Hence, teachers' gender was not influential on the average values.

Figure 3.3 shows why it is important, even for blind techniques, to cut pauses during post-processing. The lack of SL brings students to raise their voice flattening the dynamic of the recording. This could result in influencing the statistics and results of a single lecture for two reasons. First, the recorded SPLs are quite high and they do not refer to SL, they are not consistent with the analysis. Secondly, the babble noise being more temporally steady, it could have a masking effect on the background noise to be recorded among sentences.



(a) Statistical levels (PL)

(b) Peak detection (PD)

(c) Gaussian Mixture Model (GMM)

(d) K-means clustering (KM)

Fig. 3.4 Methods to measure student activity (SA): Statistical levels (3.4a); Peak detection (3.4b), Gaussian Mixture Model (3.4c), K-means clustering (3.4d). From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

SA and SL values

Each dataset obtained from each sound level meter was processed via PL, PD, GMM, and KM. In regards to blind methods, the model selection step was skipped and the number of clusters was set equal 2 to easily compare the results among techniques. Then, SA and SL values were averaged over the two receivers to get one single value per each lesson. Table 3.3 shows all the results per each lesson and methods. Besides SA and SL values, the number of students during the lesson, the relative percentage over the maximum occupancy, and the corresponding hall where the lecture was conducted are shown. Standard deviations refer to the diffusion of results between the two receivers.

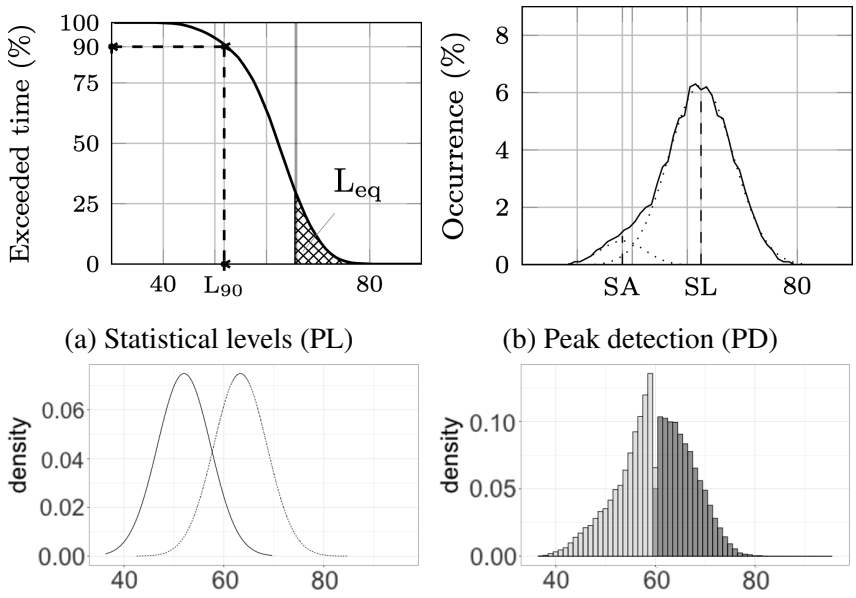
Almost all the lessons were conducted in a traditional way, i.e. with the teacher speaking from the desk. Lectures carried out differently from the traditional way were kept in the analysis basically for two reasons: they are representative of a different use of these spaces; they broaden the analysis of pros and cons of the different techniques.

In Hall I for most of the lessons, PL returns the lowest SA values and always the highest SLs. KM gives back always the highest SAs and GMM the lowest SLs for the majority of lectures. In Hall II, PD and GMM return the lowest SAs whereas PL the highest. GMM and PL produce the lowest and the highest SLs as well. In Hall III, the lowest SAs were measured by PD whereas the highest by KM. PD measured the lowest SLs too, whereas KM, PD, and PL measured the highest SLs. Mean tendencies show that PD and KM measured the lowest (51.0 dBA) and the highest (53.1 dBA) SAs, respectively; GMM and PL measured the lowest (67.2 dBA) and the highest SLs (70.5 dBA), respectively.

Following, three lessons are discussed since they were the “less traditional” among all. Lesson C was a meeting for internships. More persons talked from the desk point and the discussions were highly interactive with students. The measured SA involves the intentional speaking and the non-intentional speaking. However, only the latter type concerns to what extent it is possible to assess the focus of students. Figure 3.8 shows the four graphical results of lesson C. It could be assumed that intentional speaking of students does not overlap with the SL: teacher’s speech from PA and intentional speaking from students are not simultaneous. This condition is quite crucial: it brings significant differences among the results of each technique. If students and the teacher speak at the same time, the higher level overcomes the lower one. One more factor that could influence the results concerns the students’ proximity to the receivers. Considering a sound level meter 1.2 m height to simulate what a student perceives, the chatting nearby the microphone would record high

Table 3.3 Overview of the recorded lessons. For each lesson, the number of people, the percentage of occupancy, the corresponding room and the teacher gender are shown. Measured A-weighted values of student activity (SA), received speech level (SL) extracted through Percentile levels, Peak detection, Gaussian mixture and K-means clustering methods are reported. Values are averaged over the two receiver positions selected for the measurements performed during lessons. All values of SA and SL are in dBA. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

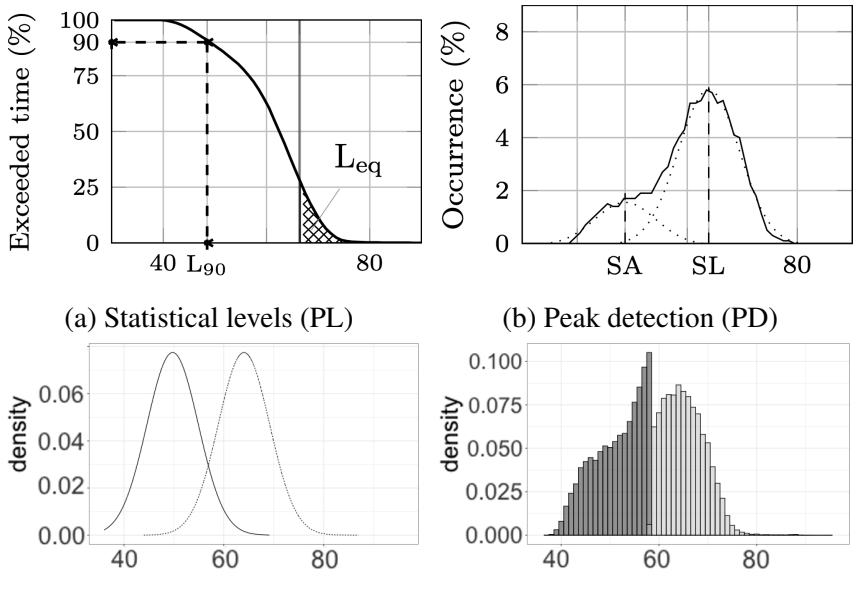
Lesson	Occupancy	(%)	Hall	PL		PD		GMM		KM	
				SA (s.d.)	SL (s.d.)	SA (s.d.)	SL (s.d.)	SA (s.d.)	SL (s.d.)	SA (s.d.)	SL (s.d.)
A	145	(60%)	I	48.0 (0.5)	69.9 (4.8)	48.0 (1.0)	65.1 (4.2)	48.2 (1.2)	65.0 (4.0)	52.2 (1.2)	68.3 (4.0)
B	200	(80%)	I	45.8 (0.6)	64.8 (4.9)	47.4 (1.4)	63.5 (4.8)	47.5 (1.5)	63.3 (4.6)	48.8 (1.3)	64.2 (4.5)
C	100	(50%)	I	53.0 (1.7)	68.9 (4.5)	51.1 (4.0)	65.8 (4.6)	53.3 (1.8)	66.3 (4.1)	55.8 (1.9)	68.4 (3.9)
D	150	(60%)	I	47.6 (1.3)	69.7 (4.6)	50.8 (3.0)	67.4 (4.9)	51.2 (2.0)	67.2 (4.5)	52.7 (1.9)	68.4 (4.4)
E	250	(125%)	II	52.1 (9.5)	72.3 (7.1)	47.9 (1.9)	68.2 (1.5)	48.4 (0.3)	67.5 (1.5)	49.1 (0.1)	68.0 (1.6)
F	160	(80%)	II	55.0 (8.1)	71.9 (4.0)	50.1 (1.9)	66.7 (1.8)	50.3 (1.5)	66.5 (1.5)	53.1 (0.2)	68.5 (1.0)
G	120	(60%)	II	61.6 (5.4)	78.6 (5.3)	62.0 (0.7)	75.8 (1.6)	61.0 (0.6)	75.5 (1.4)	55.7 (0.7)	74.9 (1.4)
H	150	(75%)	II	56.4 (7.0)	79.2 (3.6)	55.5 (0.6)	76.1 (1.3)	55.3 (0.1)	75.3 (0.8)	55.8 (0.0)	76.0 (0.8)
I	200	(100%)	II	58.8 (5.9)	74.6 (3.7)	53.6 (1.3)	68.1 (1.0)	53.4 (0.0)	68.0 (1.0)	56.5 (0.3)	69.7 (0.8)
J	110	(65%)	III	50.3 (2.1)	63.3 (2.3)	46.9 (2.0)	59.9 (2.3)	53.0 (2.0)	61.6 (2.2)	53.3 (2.2)	63.6 (2.4)
K	80	(50%)	III	47.5 (2.0)	67.8 (2.3)	50.4 (1.2)	68.2 (2.1)	50.6 (1.5)	67.6 (1.9)	50.6 (2.1)	67.7 (1.3)
L	175	(105%)	III	51.5 (1.8)	65.1 (1.7)	48.8 (2.3)	62.7 (2.4)	51.1 (2.4)	63.1 (2.4)	53.8 (1.8)	64.7 (2.2)
Mean				52.3 (3.8)	70.5 (4.1)	51.0 (1.6)	67.3 (2.7)	51.9 (1.2)	67.2 (2.5)	53.1 (1.2)	68.5 (2.4)



(a) Statistical levels (PL) (b) Peak detection (PD)

(c) Gaussian Mixture Model (GMM) (d) K-means clustering (KM)

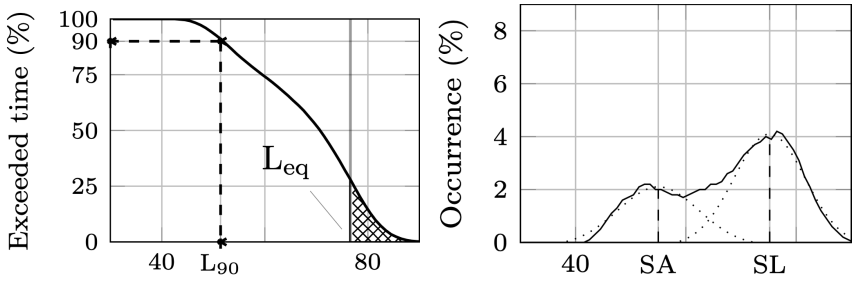
Fig. 3.5 Lesson C analyzed via PL (3.5a), PD (3.5b), GMM (3.5c), and KM (3.5d). From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].



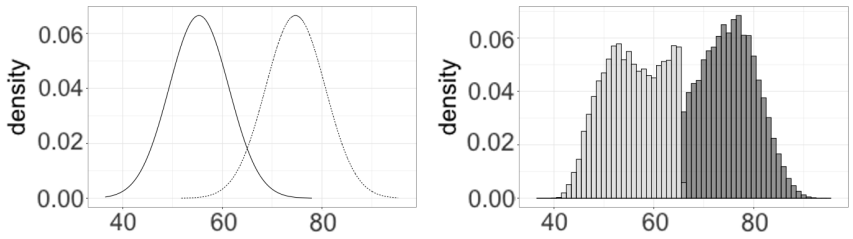
(a) Statistical levels (PL) (b) Peak detection (PD)

(c) Gaussian Mixture Model (GMM) (d) K-means clustering (KM)

Fig. 3.6 Lesson D analyzed via PL (3.6a), PD (3.6b), GMM (3.6c), and KM (3.6d). From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].



(a) Statistical levels (PL) (b) Peak detection (PD)



(c) Gaussian Mixture Model (GMM) (d) K-means clustering (KM)

Fig. 3.7 Lesson H analyzed via PL (3.7a), PD (3.7b), GMM (3.7c), and KM (3.7d). From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

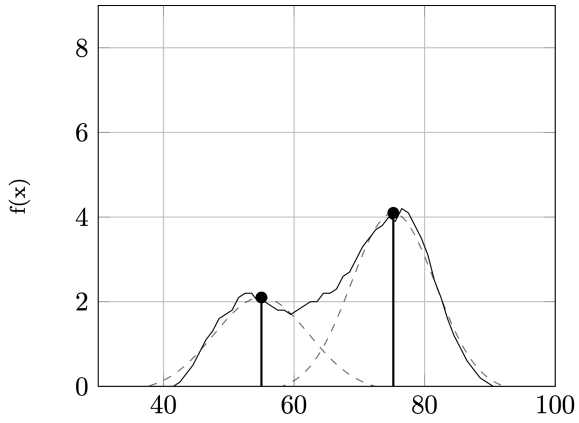
SPLs not belonging to SL. Nevertheless, it is reasonable to assume that the students chatting close to the receivers is not quantitatively proportional to the amount of SL over the whole lesson. Interactions between students and teacher may result in wider SL curves. During Lesson D a digital medium was streamed for a long portion of the lecture. It was transmitted through the PA as well as the voice of the teacher. Thus, the time history had fewer pauses and the detected SL was more continuous in time. Looking at the values of Table 3.3, SL values differ less than SA, indeed. PD and GMM return similar values of SA, whereas KM returns a higher value and PL a lower one. Figure 3.6 shows the four graphical results of lesson D. Teacher and students had a lot of interactions during lesson H. This led to higher attention paid by students, confirmed by the high SNRs values for each method. In this case, PL returns an SA value higher than the other methods. When the students' chatting is high, as seen in lesson D, PL seems to underestimate the SA; when the students are more quiet, the PL method seems to overestimate the SA. This could be linked to the different management of SPLs. PL treats SPLs as energy whereas the other methods as random variables. High standard deviations between sound level meters highlight the need for a more diffuse sound field, especially in Hall I. This confirms the complaints received by students and teachers about the acoustics of the halls under study. In the following sections, the impact of the acoustic treatments will be analyzed through the SA.

Differences among techniques

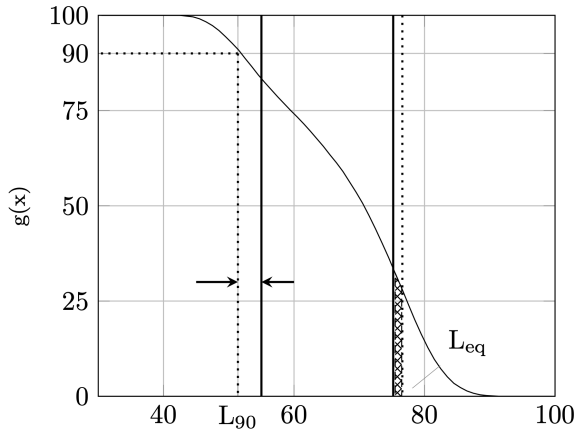
SA is spread throughout the space because the whole audience area contributes to it. Thus, it varies in time and space. However, considering the whole lesson time, it could be considered and treated as homogeneously distributed across the halls. As a consequence, low s.d. between the two sound level meters should confirm this assumption and give back homogeneous SA values. Table 3.3 shows PD, GMM and KM returning low s.d. values of SA, indeed. This is not true about PL, that seems to give back more diffusion between the two sound level meters.

The measurement of the teacher's voice is affected by the use of the microphone, too, i.e. distance and directivity. The voice may be heard as louder or higher in pitch depending on the PA coverage and its frequency response related to the acoustical characteristics of the space. Thus, due to the differences of the PA coverage, the SLs standard deviations (s.d) can be comparable only for the lessons carried out in the same lecture hall. According to the assumption above, SL values should have higher s.d. differences between the two sound level meters. In Hall I, SL measured high s.d. in a total range of 3.9 - 4.9 dBA. In Hall II, PL measured high s.d. for both SA (range 5.4 - 9.5 dBA) and SL (3.6 - 7.1 dBA). This behavior is not detected by the other methods that measured s.d. in the range 0.0 - 1.9 dBA for SAs and 0.8 - 1.8 dBA for SLs. Recalling that Hall I and Hall II are geometrically similar because of their amphitheatre shape (see 3.2, it seems that PL gives back more uncertainty with respect to the sound level meter placement throughout the space. In Hall III, which has a shoe-box shape, s.d. are quite similar among all the methods.

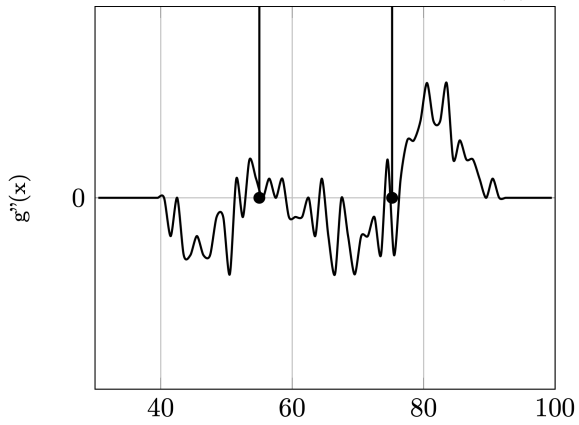
However, the PL technique is recognized as the technical praxis, so it is the most used among scholars [95, 83, 13, 132]. Figure 3.8 graphically shows the difference between the conventional approach of PL and the statistical one of the others. KM is slightly an exception because it does not exploit properly a probability distribution function. Given the same recording, Figure 3.8a shows the probability density function $f(x)$ and the relative dashed Gaussian curves obtained via GMM. Means of each component are highlighted by the two dots. Figure 3.8b shows the corresponding cumulative distribution function $g(x)$. The solid vertical lines are the projections of the means obtained in pdf. The dotted vertical lines highlight the 90th acoustical level L_{90} and the equivalent level L_{eq}. Here, the arrows and the dashed areas indicate the gap between the calculated means and statistical levels. It is interesting to notice that the two means correspond exactly to the inflection points of $g(x)$. This is confirmed by looking at the zeros of the numerical second derivative of $g(x)$ shown in Figure 3.8c.



(a) Probability density function $f(x)$



(b) Cumulative distribution function $g(x)$



(c) Second derivative of cdf $g''(x)$

Fig. 3.8 Relationship between the probability density function $f(x)$, the cumulative distribution function $g(x)$, and its numerical second derivative $g''(x)$. Solid lines show the projection of the means obtained via GMM over all the three plots. Dotted lines and dashed area in 3.8b show the 90th acoustical level L_{90} and the equivalent level L_{eq}. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

As mentioned in previous sections, GMM and KM give back the same outcomes only if the homoscedasticity, i.e. the same data variance, is fulfilled [94]. Although no probability assumption is usually mentioned, KM can be derived as maximum likelihood estimator of a fixed partition model of Gaussian clusters with equal within-cluster variances. According to such a model, $x_1;\ldots;x_n$ are independently drawn from $\mathcal{N}(\mu_{x_i\in c_k};\sigma^2), i=1,\ldots,n$, where $\mu_{x_i\in c_k}, k=1,\ldots,K$ are parameters given the cluster membership of x_i. This can be guessed by the results of Table 3.3. The gap between methods increases when the variances between the sound level meters are high for SL and low for SA, e.g. for lessons G and I. Moreover, the random initialization could lead to different local maxima solutions.

SNR and Lombard effect

Table 3.3 shows the mean SA and SL values measured over all the lessons and halls. Consequently, it is possible to extract the corresponding mean SNRs. PL measured about +18 dBA, PD about +16 dBA, and GMM and KM about +15 dBA. The last two are close to what was measured by Shield [132]. However, that work used the PL technique instead of GMM and KM. Thus, higher values can be attributed to the use of PA, unlike most of the earlier cited works (see Table 3.1). With respect to the studies where the PA was used [112, 83], it is important to notice that measurements were made in small classrooms in elementary grades instead of large university lecture halls.

Figure 3.9 shows the correlation between SA and SL. Thus, the trend of SNR. Different types of lines show the linear regressions for each method. PL, PD, GMM, and KM regressions are indicated by dotted, dash-dotted, solid, and dashed lines, respectively. Blind methods, i.e., GMM and KM, show similar tendencies. The offset between the two methods can be associated with reasons explained in the previous section about the differences among techniques, i.e., random initialization and heteroscedasticity. As assumed and expected, SA returned lower standard deviations. SL is strongly dependent on the PA. PL and PD show different slopes. It could be deemed that students set their own speech levels to not disturb the listening process and exploit the maximum intelligibility achievable in the space. In fact, regardless of the different approaches, the lowest measured SNR is equal to +15 dBA, the threshold above which the intelligibility is not affected by the background noise. This psychoacoustic behavior could be considered a sort of “inverse” Lombard effect [12]. With the slope values less than one, this inverse effect is more evident at low SA values and at high SA values.

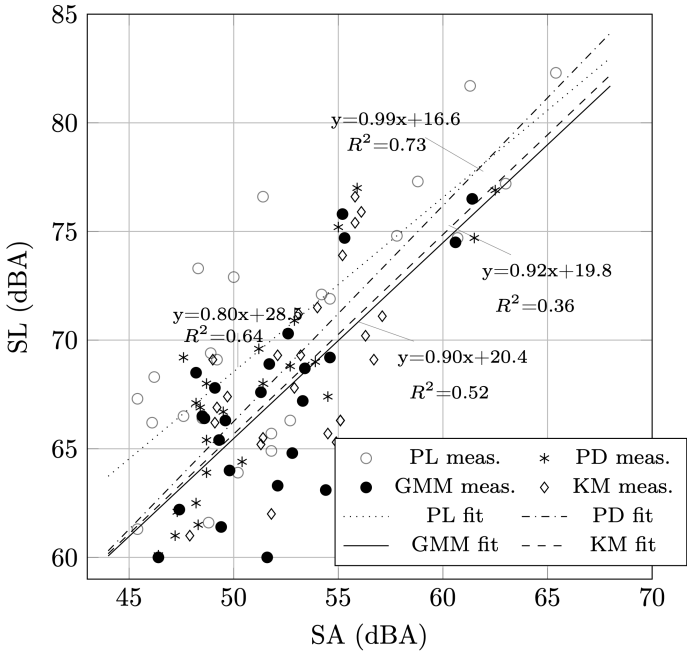


Fig. 3.9 Relationship between student activity (SA) and speech level (SL) measured during lessons. Each marker indicates a single lesson analyzed with the respective method. Regression lines refer to the whole dataset of lessons analyzed via each method. From "Measuring the speech level and the student activity in lecture halls: Visual-vs blind-segmentation methods" by D'Orazio et al [43].

Linking these results to the students' behavior, it is possible to notice that students are quieter at the beginning of the lesson and immediately after the break; they are noisier right before the break and at the end of the lesson. With the context being university halls, students are prone to listen the lesson because they are adults and the attendance is often non-mandatory. Earlier works show lower SNRs especially in lower grades, such as elementary or secondary schools (see Table 3.1). However, it can be because of non-acoustic reasons. Moreover, here the attendance is mandatory and the lessons are conducted in the same classroom for the entire day. Low SNR values in universities can be attributed to the size of classrooms – smaller than the present study, the HVAC noise turned on unlike the present study –, and the different interactions between teachers and students in university courses. All these reasons may influence the listening effort, increasing the student activity [93].

3.2 Design of active and passive acoustic treatments

Lecture halls need periodical renovations, even to meet local standard updates [4]. The enhancement of acoustic environments involves both objective and subjective

aspects. It has been shown that students prefer renovated spaces [111]. Despite the subjective impression and the listening effort depending more on reverberation than intelligibility besides other factors like sentence complexity, age, and linguistic abilities, the location of improvements is fundamental to properly increase the acoustic quality of classrooms [11, 144, 115]. In fact, in spaces with high intelligibility scores, high comprehension by the students is achieved even for low signal-to-noise ratios [28]. Teachers and students experienced and complained an acoustic discomfort in Hall I and Hall II, despite these rooms having been designed for learning purposes. Thus, active and passive acoustic treatments were designed. The goal was to increase the speech intelligibility according to the Italian standard UNI 11532-2 [139]. Improvements were designed based on the different shape, acoustical characteristics and specific use of each lecture hall. Designs involved both passive and active treatments. Recalling Table 3.2, Hall I has a rectangular plan, wooden and terraced seats that produce a typical amphitheatre space, reflective surfaces, and an articulated false ceiling. The volume is about 1000 m^3 and a maximum occupancy of about 250 students. Hall II has a rectangular plan, wooden and terraced seats, reflective surfaces, and a flat false ceiling. The volume is about 900 m^3 and the maximum occupancy is about 200 students. They vary essentially by the shape of the false ceilings, besides an extra volume at the rear part of the room in Hall I.

3.2.1 Passive treatments

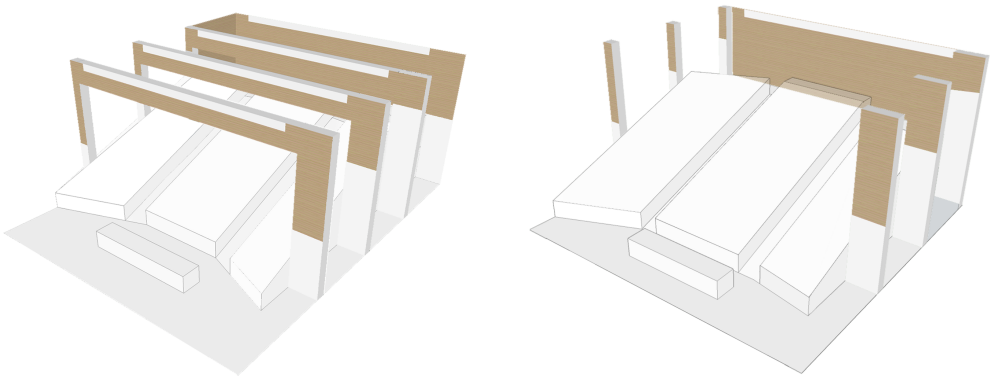
Passive acoustic treatments were designed to achieve the requirements provided by the national standard UNI 11532 [139]. Thus, the optimal reverberation time in occupied state was calculated. The revers formula allows to calculate the needed equivalent absorption area A, i.e. the amount and the properties of sound absorbing panels to introduce in each lecture hall. Nevertheless, adding the adequate A isn't enough to obtain a good speech intelligibility. The placement of the surfaces plays a key role in controlling the sound reflections to enhance the sound clarity and the sound energy distribution throughout the space. The placement could affect the early-to-noise ratio C_{50} with variations up to 4 dB, and the received speech levels with variations up to 3 dB values [118]. European standards guidelines suggest how to optimize the placement of the absorbing surfaces [139, 21, 39]. The ceiling should be kept reflecting enhancing the early reflections. It is possible to introduce absorbing material on the ceiling only along its edges. The rear wall represents the ideal area to absorb the late reflections.

Due to their high geometrical similarity, the design of treatments was the same for both halls. The passive treatments were made by means of wooden slat panels

Table 3.4 Absorption coefficients α of the passive acoustic treatments in octave bands from 125 up to 4000 Hz. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Material	α					
	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Wooden slat wall panels	0.03	0.33	0.73	0.89	0.85	0.77

with an air cavity 10 cm depth from the walls. They were placed on the rear wall of each hall and the overhanging beams and pillars, as shown in Figure 3.10. The corresponding absorption coefficients are shown in Table 3.4. After the renovation works, further measurements according to ISO 3382 [72] were carried out to assess the improvements achieved.



(a) Hall I (b) Hall II

Fig. 3.10 Passive acoustic treatments in Hall I and II: placement of the sound absorbing panels.

Monoaural receivers were used to acquire Exponential Sine Sweep – 512 K length sampled at 48 kHz – signals sent from an omnidirectional source. The latter was a high SPL custom dodecahedron calibrated in a reverberation chamber according to ISO 3741 [74]. Two source positions, one on the axis in the middle of the room behind the desk, and the other asymmetrical near the desk were used. Receivers were located homogeneously in the seating area. The same source-receiver positions were used before and after the restoration. The reverberation time in an unoccupied state has been decreased from 1.7 to less than 1.4 s for both halls. The early-to-late ratio C_{50} was improved of about +1.5 dB in both halls, the STI of +0.03 and +0.07 respectively for Hall I and II. All these data are summarized in Table 3.5.

Table 3.5 General and acoustic data of the halls under study before and after the restoration, respectively indicated as "before" and "after". Besides the shape of the inner space, it shows the volume "V", the maximum occupancy "N", the reverberation time in unoccupied state "T", the early-to-late index "C_{50}", the Speech Transmission Index STI and the equivalent absorption area A_0 of the lecture halls in unoccupied state. The subscript "M" states a value averaged over all the receivers in the octave bands of 500 – 1000 Hz, whereas "3" over the octave bands of 500 – 2000 Hz. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Hall	Shape	Volume (m^3)	Occupancy	T_M (s)		$C_{50,3}$ (dB)		STI		$A_{0,M}$ (m^2)	
		V	N	Before	After	Before	After	Before	After	Before	After
I	Amphitheater	1000	250	1.70	1.37	-2.8	-1.4	0.49	0.52	94	117
II	Amphitheater	900	200	1.72	1.38	-2.4	-1.0	0.47	0.54	84	105

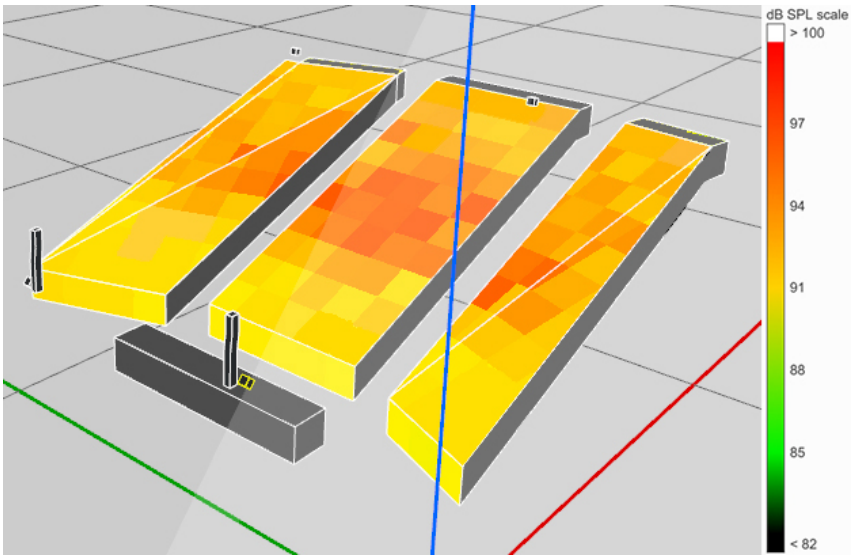
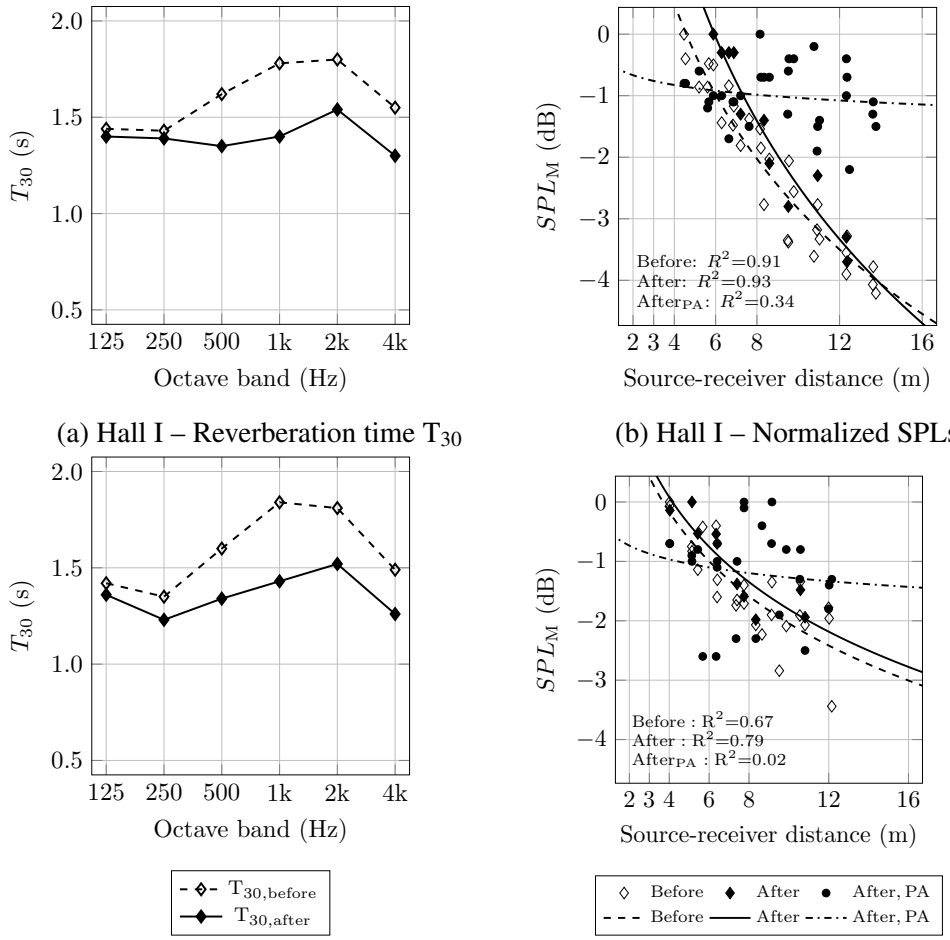


Fig. 3.11 Placement of the line array system. This was the same for Hall I and II. Colored squares indicate the receiving SPL over the audience area.

3.2.2 Active treatments

The PA (Public Address) system represents a crucial element for speech intelligibility, especially in large lecture halls. It is not possible to achieve an STI equal to 0.6 only with passive treatments and without a speech reinforcement system in such high volumes. The passive treatments and the intelligibility parameters obtained after works were assessed through numerical models made in Odeon Room Acoustics Software. However, all the values were considered precautionary because: the source was set as omnidirectional; the numerical model did not consider the PA system but only the teachers' voice as a source.



(a) Hall I – Reverberation time T_{30}

(b) Hall I – Normalized SPLs

(c) Hall II – Reverberation time T_{30}

(d) Hall II – Normalized SPLs

Fig. 3.12 Room acoustic properties of the two halls before and after the renovation works. On the left, the reverberation time as function of frequency octave band are shown. On the right, the normalized sound pressure levels (SPL) as function of the source-receiver distance is shown. Values of SPL have to be taken only qualitatively because they do not refer to an omnidirectional source when PA is considered. The reference level of the normalized SPLs is the position 4 meters far from the omnidirectional source before the renovation. At this distance the cylindrical wave of the line array is shaped. The subscript "M" states a value averaged over all the receivers in the octave bands of $500 - 1000$ Hz. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

The PA systems have been replaced with line arrays (L-Acoustic Syva) located behind the teachers' desk and supplementary loudspeakers (L-Acoustic 5XT) as fillers for the first and the last rows of the audience area. Choosing line arrays means, from a theoretical point of view, to exploit a cylindrical wave throughout the space. Thus, the decay with respect to an omnidirectional source is lower doubling the source-receiver distance.

The orientation of the fillers behind the desk was optimized to increase the self-monitoring of teachers, too. The high similarity of the halls allowed to design the same treatments for both. Due to the different sensitivity of the loudspeakers, the gain and the time delays between line arrays and fillers were set to reach a homogeneous coverage of the direct sound in the seating area with the help of a further numerical model made in Soundvision software. Here, the high directivity – vertical axis 25°, horizontal axis 130-140° in the range 250 - 2000 Hz – of line arrays was considered, unlike in Odeon, to assess a more realistic sound diffusion obtained over the audience. The goal of the PA design is to achieve an adequate coverage over the audience, with SPL differences among seats of less than 3 dB. The first rows have been underestimated because they can rely on the direct speech from teachers. Figure 3.11 shows the diffusion degree obtained with the line array system.

After the placement of the new PA, further measurements were carried out to assess the reliability of the acoustic design. Figure 3.12 shows the frequency behavior of reverberation time T_{30} and the spatial decay of normalized sound pressure levels (SPL) before and after the treatments for each hall. Besides the dodecahedron, the SPL decay has been evaluated using the PA as source. Passive treatments have a small influence on the spatial decay of a spherical sound source, i.e. the natural voice or traditional loudspeaker; hence, it should be noted that a line array shows a lower slope compared to a dodecahedron. SPLs measurements confirm the homogeneous coverage over the audience area. However, while a small dependency on the distance is observed (students seated in the first and the last rows receive similar useful energy), there are significant variations of SPLs on the same row of the audience due to the horizontal directivity of the line arrays. This specific coverage influences the placement of the sound level meters used for SA and SL measurements, as discussed in the following sections.

3.3 Measurement of student activity and speech levels before and after acoustic treatments

After the renovation works, 9 lessons in Hall I and II were monitored with the same setting and method seen in Section 3.1.2. Receivers were placed in the same position as before the works to minimize the differences between the two measurements. It is important to notice that all the lectures were measured in a pre-COVID19 scenario. Results were compared and shown with 9 lessons measured before the renovation in Table 3.6. Both measurements, before and after treatments were carried out with two sound level meters. Thus, the results shown are averaged over the two receivers.

Lectures from A to I in the upper part of the Table refer to previous outcomes obtained in [43] and were measured before the restoration; lectures from J to R refer to the measurements carried out after the works. The means are shown at the end of each series of lectures. As seen in Table 3.3, the standard deviations of the outcomes obtained by the two receivers are shown in brackets.

The measured A-weighted SA and SL values before the treatments lie respectively in the range of 47.5 – 61 dB and 63.3 – 75.5 dB for GMM, 48.8 – 56.5 dB and 64.2 – 76 dB for KM, 45.8 – 61.6 dB and 64.8 – 79.2 dB for percentile and equivalent levels. The measured SA and SL levels after the treatments lie respectively in the ranges 47.2 – 53.9 dB and 59 – 72.1 dB for GMM, 49.7 – 54.1 dB and 61.2 – 72.7 dB for KM, 45.9 – 53.3 dB and 61.1 – 74.4 dB for percentile and equivalent levels. Before the restoration work, the standard deviations between the two receivers, respectively for SA and SL, lie in the ranges 0 – 2 dB and 0.8 – 4.6 dB for GMM, 0 – 1.9 dB and 0.8 – 4.5 dB for KM, 0.5 – 9.5 dB and 3.6 – 7.1 dB for percentile and equivalent levels. Concerning the measured s.d. of SA and SL after the treatments, values lie respectively in the ranges 0.3 – 3.1 dB and 0.1 – 1.9 dB for GMM, 0.2 – 1.5 dB and 0.1 – 2.1 dB for KM, 0.4 – 1.1 dB and 0 – 3 dB for percentile and equivalent levels. The measured A-weighted mean values of SA and SL and their standard deviations in brackets before the treatments are respectively 52.1 (1.0) dB and 68.3 (2.6) dB for GMM, 53.3 (0.8) dB and 69.6 (2.5) dB for KM, 53.1 (4.4) dB and 72.2 (4.7) dB for percentile and equivalent levels. The same parameters measured after the treatments are respectively 50.8 (1.0) dB and 65.6 (0.8) dB for GMM, 51.7 (0.8) dB and 67.6 (1.1) dB for KM, 50 (0.7) dB and 67.8 (1.1) dB for percentile and equivalent levels.

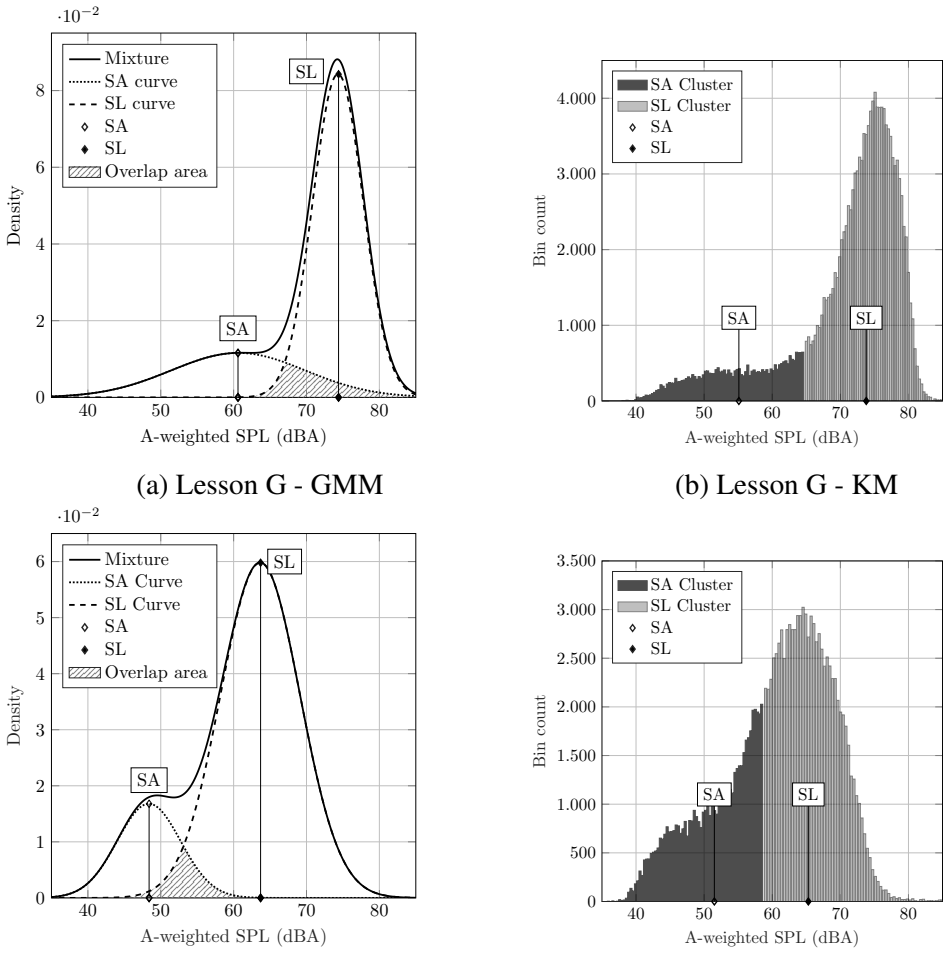
Hodgson measured in Canadian university classrooms and lecture halls SA and SL values in the range of 30 – 50.2 dB and 43 – 59 dB with average values of 41.9 dB and 50.8 dB, respectively [62]. Choi did not detect SA in Korean university classrooms. However, average values of noise and SL were 43.7 dB and 51.4 dB, respectively [27]. However, all the lessons measured in the cited previous works were conducted without the support of the PA. Also, the volumes of Hall I and II – 1000 and 900 m^3, respectively – are bigger than the average of the cited works, ranging from 110 to 957 m^3 [62] and from 188 to 343 m^3 [27]. Without considering the outliers, the occupancy in the halls can be deemed similar before and after the treatments.

3.3.1 Signal-to-noise ratio and evaluations of the renovation works

The intelligibility problem can be outlined as made up by two main factors: the acoustical properties of the space, specifically the reverberation time, and the signal-

Table 3.6 Overview of the recorded lessons. For each lesson, the corresponding room, the number of people, the percentage of occupancy, the equivalent absorption area taking into account the contribution of the people, and the ratio between the equivalent absorption area in occupied and empty states are shown. Measured A-weighted values of student activity (SA), received speech level (SL) extracted through, Gaussian Mixture Model GMM, K-means clustering KM, equivalent continuous L_{eq} and percentile levels L_{90} are reported. Values are averaged over the two receiver positions. All values of SA and SL are in dBA.The subscript "M" states a value averaged over all the receivers in the octave bands of $500-1000$ Hz. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Lesson	Hall	Occupancy (%)	$A_{occ,M}(m^2)$	$A_{occ,M}/A_{0,M}$	GMM		KM		L_{90} (s.d.)	L_{eq} (s.d.)
					SA (s.d.)	SL (s.d.)	SA (s.d.)	SL (s.d.)		
A	I (Before)	145 (60%)	156	1.66	48.2 (1.2)	65.0 (4.0)	52.2 (1.2)	68.3 (4.0)	48.0 (0.5)	69.9 (4.8)
B	I (Before)	200 (80%)	179	1.90	47.5 (1.5)	63.3 (4.6)	48.8 (1.3)	64.2 (4.5)	45.8 (0.6)	64.8 (4.9)
C	I (Before)	100 (50%)	137	1.46	53.3 (1.8)	66.3 (4.1)	55.8 (1.9)	68.4 (3.9)	53.0 (1.7)	68.9 (4.5)
D	I (Before)	150 (60%)	158	1.68	51.2 (2.0)	67.2 (4.5)	52.7 (1.9)	68.4 (4.4)	47.6 (1.3)	69.7 (4.6)
E	II (Before)	250 (125%)	190	2.26	48.4 (0.3)	67.5 (1.5)	49.1 (0.1)	68.0 (1.6)	52.1 (9.5)	72.3 (7.1)
F	II (Before)	160 (80%)	152	1.81	50.3 (1.5)	66.5 (1.5)	53.1 (0.2)	68.5 (1.0)	55.0 (8.1)	71.9 (4.0)
G	II (Before)	120 (60%)	135	1.44	61.0 (0.6)	75.5 (1.4)	55.7 (0.7)	74.9 (1.4)	61.6 (5.4)	78.6 (5.3)
H	II (Before)	150 (75%)	164	1.95	55.3 (0.1)	75.3 (0.8)	55.8 (0.0)	76.0 (0.8)	56.4 (7.0)	79.2 (3.6)
I	II (Before)	200 (100%)	188	2.24	53.4 (0.0)	68.0 (1.0)	56.5 (0.3)	69.7 (0.8)	58.8 (5.9)	74.6 (3.7)
Mean	Before	164	162	1.82	52.1 (1.0)	68.3 (2.6)	53.3 (0.8)	69.6 (2.5)	53.1 (4.4)	72.2 (4.7)
J	I (After)	130 (50%)	172	1.47	51.3 (1.1)	64.6 (0.7)	52.7 (0.7)	66.1 (0.5)	49.7 (0.8)	66.5 (0.4)
K	I (After)	185 (75%)	195	1.67	49.9 (0.4)	69.6 (0.4)	50.9 (0.4)	70.2 (0.3)	48.1 (0.5)	71.5 (0.1)
L	I (After)	130 (50%)	172	1.47	47.2 (0.5)	64.6 (0.5)	50.9 (0.6)	70.2 (0.5)	45.9 (0.4)	65.8 (0.4)
M	I (After)	80 (30%)	151	1.29	49.9 (0.4)	59.0 (0.3)	51.0 (0.4)	61.2 (0.2)	48.3 (0.5)	61.1 (0.0)
N	I (After)	190 (75%)	197	1.68	52.7 (0.4)	68.9 (0.1)	51.0 (0.2)	69.0 (0.1)	48.8 (0.5)	70.3 (0.1)
O	II (After)	110 (55%)	168	1.60	53.9 (3.1)	72.1 (1.1)	54.1 (1.3)	72.7 (2.0)	53.3 (1.0)	74.4 (2.4)
P	II (After)	125 (65%)	175	1.67	51.4 (0.8)	60.8 (1.3)	50.4 (0.9)	62.0 (2.1)	46.7 (0.7)	62.0 (1.3)
Q	II (After)	120 (60%)	173	1.65	48.8 (0.3)	66.4 (1.1)	49.7 (0.9)	69.8 (2.1)	58.4 (0.5)	71.6 (3.0)
R	II (After)	95 (50%)	161	1.53	52.1 (1.8)	64.9 (1.9)	54.4 (1.5)	67.0 (2.1)	51.4 (1.1)	67.5 (2.3)
Mean	After	129	174	1.57	50.8 (1.0)	65.6 (0.8)	51.7 (0.8)	67.6 (1.1)	50.0 (0.7)	67.8 (1.1)



(a) Lesson G - GMM

(b) Lesson G - KM

(c) Lesson D - GMM

(d) Lesson D - KM

Fig. 3.13 Lesson G (on top) and lesson D (on bottom). On the left, the Gaussian mixtures and the respective components are shown. The solid lines indicate the probability density function recorded during the lecture. The dotted and the dashed lines show respectively the Gaussian curves associated to the student activity (SA) and speech level (SL). On the right, the recorded SPLs are shown as a function of their occurrence distribution. The SA and SL clusters are shown respectively in dark and light grey. Markers indicate the mean values (for GMM) and the centroids (for KM) which identify the SPL of each sound source. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

to-noise ratio (SNR), i.e. the difference between the signal to understand and the background noise. During lectures both of these factors change, the first is influenced by the occupancy, the second by SA and SL fluctuations. As mentioned before, in lecture halls with high attendance, the signal-to-noise ratio SNR can be defined as the difference between SA and SL, indeed.

Comparing the techniques, the SA and SL means decreased after the treatments per each methods. It could be assumed that a quieter environment has been achieved, in particular for teachers. The average SL decreases by about 2.3 dB for the unsupervised methods, GMM and KM, and about 4.4 dB for the L_{eq}. The SA decreased, respectively by an average of about 1.4 dB for the unsupervised methods and 3.1 dB for L_{90}. Concerning the analysis after the restoration works (lessons from J to R), it is noticeable that, in most lectures, KM gives back higher SA values whereas the L_{90} the lower ones. Concerning SL, L_{eq} returns the highest values in most lectures, whereas the GMM is the lowest in all cases. Thus, the comparison among techniques seems to confirm the results obtained in previous sections. Equivalent continuous levels return the highest SL, statistical levels return the lowest SA after the works. This could mean that the human chatter does not fit with the 90th acoustical percentile of exceeded time.

It was proved that GMM and KM converge to the same results only if the homoscedasticity is reached, i.e. all random variables have the same variance [94, 6]. This condition cannot be fulfilled, in particular in the case of human noise which has a high temporal variability. Inside classrooms, SA shows a larger statistical distribution than SL, especially when teachers use a PA. The amplification chain - from microphone to loudspeakers - can provide tools that compress the dynamic of the voice, reducing the variance of SL close to its mean value.

Further considerations can be found in the initial hypotheses of the two algorithms. The mean difference between GMM and KM regards the cluster distribution of data. In GMM, a single datapoint can belong to more than one cluster with an assigned probability whereas in KM this is not possible. In this latter technique a single datapoint can be assigned only to one cluster. The ability to assign one point to one or more clusters is the difference between hard and soft clustering [8].

The two factors, homoscedasticity and the hard/soft clustering assumption, are visualized in Figure 3.13. In Figures 3.13a and 3.13c, solid lines show the recorded probability density functions and the two Gaussian components - dotted for SA and dashed for SL - for two lessons (G and D). The data points belonging to the fuzzy borders, i.e. the data in common between the two clusters weighted by an assigned probability, are highlighted with the dashed patch. The overlapping area is a function of the number of clusters and the variances of the Gaussian curves; thus the larger the variance, the larger the area is. Noticing the shape of the SA curve, it can be seen how large its variance is, unlike the SL curve; thus it shows the effect of heteroscedasticity on the mean values. On the right, in Fig. 3.13b and 3.13d, the same lessons are plotted after being post-processed via KM. The clusters associated with SA are in dark grey, whereas those associated with SL are in light

grey. Here, the distinction of the clusters is sharp, each data point belongs to one and only one cluster. On the y-axis the bin count is shown rather than the density to preserve the shape of the total distribution. The consequences of the fuzzy borders and the heteroscedasticity of the clusters are particularly evident in lesson G. Here, the SA difference between GMM and KM is about 5.3 dB. Conversely, it is worth noting that the difference of SL of the same lesson for both GMM and KM is 0.6 dB. Concerning lesson D, the difference of SA calculated via GMM and KM is 3.1 dB and is 1.8 dB regarding SL.

Concerning the standard deviations, their means are quite different, especially for L_{90} and L_{eq}, which shows an SA decrease of 3.7 dB and 3.6 dB for SL. The means obtained by the unsupervised methods measured lower decreases but just for SL, respectively of 1.8 dB for GMM and 1.4 dB for KM. However, looking at the single lessons, it is clear how the s.d. of SA are lower after the works. The lack of difference of the mean values is due to the outliers, indeed. The decrease of s.d. may suggest that the sound field after the renovations is more diffuse and more homogeneous coverage is achieved with the redesign of PA.

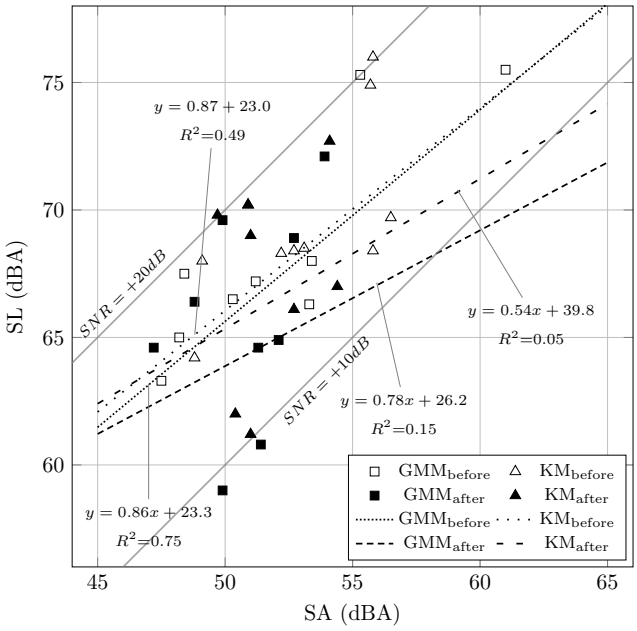


Fig. 3.14 Relationship between student activity (SA) and speech level (SL) measured values via GMM and KM. Empty and solid markers indicate respectively before and after acoustic treatments. Each marker refers to a whole lesson. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Figure 3.14 shows the correlation between SA and SL before and after works. The trends measured before the works are shown with dotted and dashed black

lines for GMM and KM, respectively. Despite the R-squared coefficients, higher for GMM ($R^2 = 0.75$) and lower for KM ($R^2 = 0.49$), tendencies are very similar. After the renovation works, the correlation decreased, showing less dependency between the noise (SA) and the signal (SL). Looking at the R-squared values, 0.15 for GMM and 0.05 for KM respectively, it is possible to notice how measured data are more homogeneous, unlike before the works. The difference can be ascribed to the different spread of the sound energy between a traditional PA and a line array, i.e. the use of a cylindrical wave. The energy emitted by a traditional PA decreases significantly with the distance. This means that the first rows are exposed to higher sound levels than the last. Thus, it is likely to have different SA levels throughout the audience area. A high-performance system, such as a line array, provides high dynamical speech signals. This can bring either improvements or impairments to the comprehension of the students. Nevertheless, the dynamic can be controlled through compressors or limiters to keep the speech levels in a short range. Thus, an optimal criterion to set a homogeneous SL in the hall may be thought to control the rise of SA.

When the rooms are equipped with a PA, teachers can experience feedback of their own voice or not. It depends on both the acoustic properties of the environment and the coverage of the loudspeakers. The latter, because of their position, act as monitors for teachers without triggering a Larsen effect. Moreover, in the halls under study, two supplementary fillers cover the area near the teacher to reach an adequate self-monitoring level.

A recent paper on unsupervised methods in K-12 classrooms stated that these measured values could not be reliable to make considerations about the Lombard effect [147]. In this case, due to the lack of dosimeters, proximity effects between teachers' mouth and microphone and variations of gains of the PA are ignored. Thus, for the same reasons, remarks about the Lombard effect are avoided. Further studies may address the repetition of this kind of analysis, including dosimeters on teachers and varying the thresholds and the ratio of compressors and the levels of limiters. Nevertheless, since the signal-to-noise ratios are always measured between 10 and 20 dB, it is reasonable to assume a sort of self-adjustment of noise levels by students.

3.3.2 Effects of occupancy on the acoustics of a lesson

The occupancy plays a key role in the dynamical context of the acoustics of a lesson. Generally in school grades, the amount of students is fixed in each room. In universities, it continuously varies during lectures. Thus, the equivalent absorption area within university lecture halls changes during the day. The absorption area in

occupied condition, i.e. during the lecture, is one of the main features affecting all the parameters proposed by Hodgson et al in their predictive model for SA [62]. Moreover, the influence of the students depends, besides the acoustical properties of the room, by their distribution through space [25, 26]. The sound absorption provided by people is higher if the room has chairs made by reflective materials, e.g., wood. Thus, students seated heterogeneously in the audience area could create different absorption concentrations throughout the space. Closer or larger seats influence the exposure of people surfaces providing less or more sound absorption. However, Hall I and II have the seats distributed as terraces. Hence, the exposed area of each person does not depend by the percentage of occupancy and their spatial distribution. The importance of attendance in each lesson is highlighted by the ratio between the equivalent absorption areas in the occupied and empty state in Table 3.6. In large lecture halls, the presence of students can double the absorption area in some cases, strongly affecting the reverberation time instead of small classrooms, such as in secondary school [23].

The drop of correlation between SA and SL shown in Figure 3.14 has led the analysis to focus on the relationships among the occupancy, the SNR, and its components, SA and SL, respectively. These are shown in Figure 3.15. As already noticed, similar occupancies were measured before and after the works, barring the outliers. White and black lines represent the tendencies before and after the restoration in all plots, respectively. Starting from the top, the relationship between occupancy and the signal-to-noise ratio is shown. The enhancement of the acoustic conditions of the halls seems to have made the correlation more sensitive. The SNR increases linearly with the occupancy, indeed. This could mean either that the bigger the audience, the quieter the environment or the bigger the audience, the higher the speech level used by the teacher. It is well known that a SNR equal or greater to 15 dB does not affect the intelligibility scores and values around 20 dB are considered as “ideal” targets in classrooms [17, 10, 61]. Before the works, the SNR did not seem to correlate with the occupancy, and it is confirmed by the R-squared coefficients. After the works, the relationship has become clearer but not so tight. However, it can be deduced that SNR is more dependent on the number of students. Moreover, an occupancy of about 120 students seems to be a sort of threshold for the behaviour of listeners after the works. Crowded lectures seem to trigger a psychoacoustical effect which leads the students to achieve the best SNR without affecting the intelligibility. Lower occupancies keep the SNR lower than 15 dB whilst higher occupancies reach SNR values greater than 15 dB. In correspondence with the occupancy of about 120 people, the records are more variable, and the measured SNRs span from about 9 to 20 dB. None of the recorded lessons exceeds the value of 20 dB, as expected for the

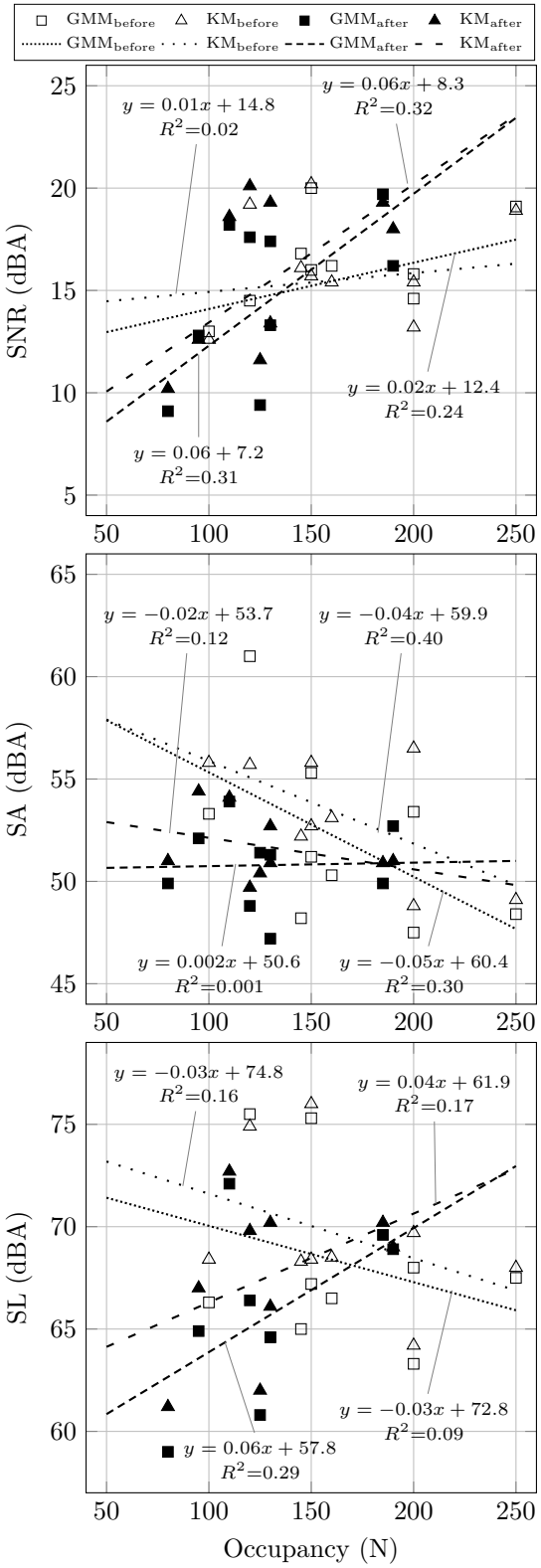


Fig. 3.15 Relationship between occupancy and signal-to-noise ratio (SNR), student activity (SA), and speech level (SL) measured values via GMM and KM. Empty and solid markers indicate respectively before and after acoustic treatments. Each marker refers to a whole lesson. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

above reasons. Results suggest that in large lecture halls, despite the optimal acoustic characteristics, a PA is necessary to easily achieve an SNR of +15 dB, especially in the back of the room [59, 60].

In the middle of Fig. 3.15, the plot shows the relationship between occupancy and SA. After the acoustic treatments, the SA seems to keep a constant tendency regardless of the number of students attending the lectures. Thus, it seems that the SA is independent of the occupancy. This is particularly true for GMM, less for KM, which preserves a descending tendency as measured before the treatment, even though with different slopes and smaller R-squared coefficients. Thus, it could mean that before the works, students control their noise depending on their number. Quieter environments as a result of acoustic treatments of classrooms were measured by Oberdörster and Tiesler in primary school [106].

On the bottom, Fig. 3.15 shows the relationship between occupancy and SL. Here, the acoustical treatments seem to have achieved the greatest effect. The tendencies changed directions, indeed. This means that teachers tend to increase their voice levels with rising occupancy. However, it could not be related to a vocal effort because all lectures were carried out with a PA system. The reasons to explain this behaviour could be various. High occupancies before the restoration helped to reach lower reverberation, more similar to values obtained after the treatment. Furthermore, SA tendencies before the works decreased with respect to the occupancy (shown in the middle of Fig. 3.15). Thus, combining reverberation and noise conditions, the more crowded the room the quieter the environment was. Before and after tendencies of the relationship between SL and occupancy cross each other in correspondence with an occupancy of about 150 students. The main differences between before and after states are noticeable below this value. Tendencies after works change slopes mainly because lower SL values were measured with half-empty halls. Nevertheless, the R-squared coefficients are not high enough to state a tight correlation between the two parameters. However, it is interesting to notice how the SNR depends more from SL than SA after works.

Furthermore, it should be noted that the direct field emitted by a traditional loudspeaker drops in a few meters compared to the early and late reflections. A line array allows the direct field to reach the back of the hall instead. This means that the early reflections are almost absent because of the high directivity of the source, whereas the late reflections are the same, regardless of the system's technology. This could justify the increase of SL after the PA redesign in Figure 3.15. Before the PA redesign, SL decreased with occupancy maybe because the absorption increased, especially on the late reflections. After the installation of the line array, most of the speech energy in highly occupied conditions is in the direct contribution. Thus,

the absorption due to the occupancy has a greater impact on SL values before the renovation than after. The ratio between the equivalent absorption area in occupied and empty state has a higher average value before (A_{occ}/A_0 =1.82) than after (A_{occ}/A_0 =1.52) the works, as shown in Table 3.6.

3.3.3 Spectral analysis of the measured sources

Spectral analysis can bring further insights about the way the clustering works. It is expected to achieve two speech signals during lectures, one associate with SA and the other with SL. The aid of the PA turned out to be very useful in assessing the spectral distribution of the clusters. Indeed, it is expected to have an SL spectrum more anechoic-like, being made almost totally of direct energy, and thus very similar to the reference standards [69, 3, 73]. To obtain the spectra, the clustering was broadened over the octave bands 125 - 4000 Hz range. Figure 3.16 shows the average relative spectra over all the measured lessons. On the left, the plot presents the outcomes of SA (dashed line) and SL (solid line) obtained before the renovation. On the right, the plot shows the results obtained after the works. Relative spectra were evaluated by setting the 1 kHz octave band equal to 0 dB. According to the standard tendencies, the spectra obtained are associable to the speech. Being produced by a single source focused in a particular point – i.e. the loudspeakers positions –, SL has a sharp trend in agreement with [62] and [147]. SA is slightly different. This is expected because it is more spread and affected by the noise which modifies the shape, especially in the low frequencies where the greatest uncertainties are [87, 124].

Different results regard the spectra obtained after the works. They are more flattened even for SL, as shown on the right of Fig. 3.16. It is worth recalling that treatments regarded the redesign of the PA besides the surfaces; thus, SL seems to be deeply affected by the equalization of the new loudspeakers. However, the most interesting result concerns the shape of SA from the middle to the high frequencies 1-4 kHz. In fact, despite the differences underlined above about the PA, the shapes of the spectra from 125 up to 500 Hz are quite similar before and after the treatments. From 1 kHz up, the behaviour of SA is not as expected since the PA like SL cannot influence it. In these frequencies, the clustering seems to be less reliable. The reasons why this happened could be speculated on multiple levels. On the side of the algorithms, if the peaks of the occurrence curves are not so clear, it may be difficult to characterize the difference between the two sources. On the side of the renovation works, the treatments regarded mainly the absorption of mid and high frequencies (see Table 3.4). It may be possible that the formants of the speech are more affected losing more energy by the treatments than the fundamentals. On the

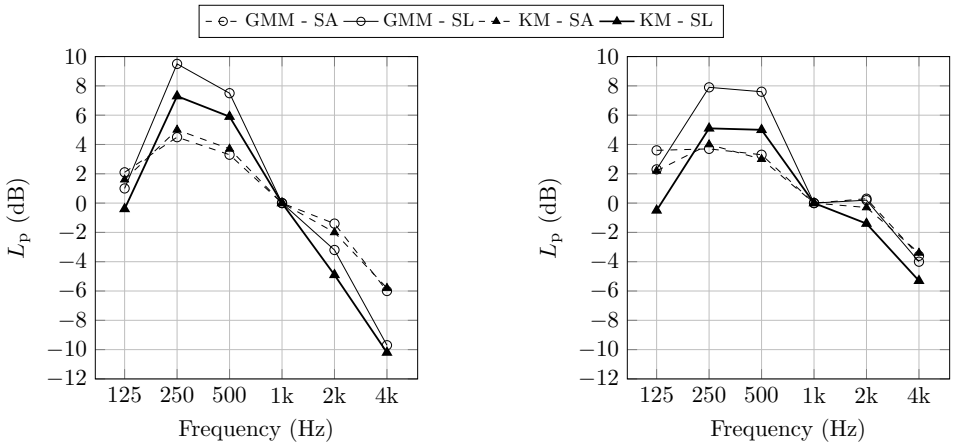


Fig. 3.16 Average relative spectra of student activity (SA) and speech levels (SL) obtained via GMM and KM. On the left SA and SL obtained before the acoustic treatments of the halls are shown, on the right the values obtained after works. Values are averaged over all measured lectures. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

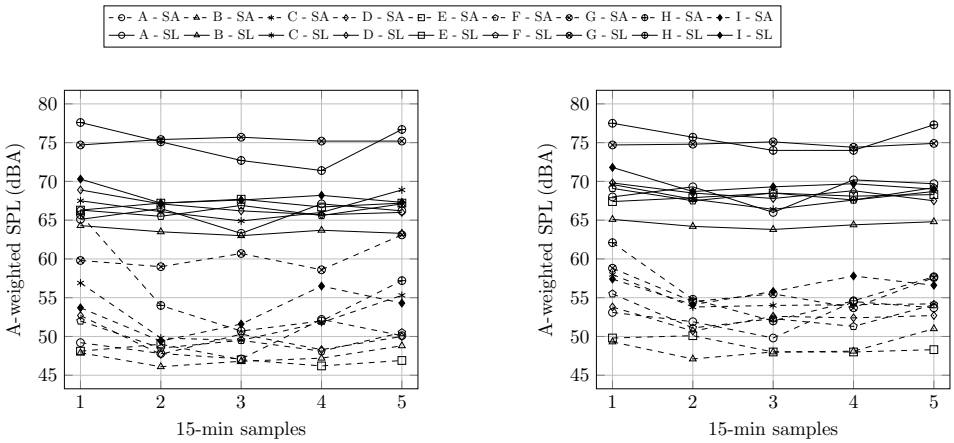
side of the students' behaviour, it has been noticed in previous sections how quieter environments have been achieved; thus it could mean that the detection of SA is more difficult.

3.3.4 Variations of SA and SL during lectures

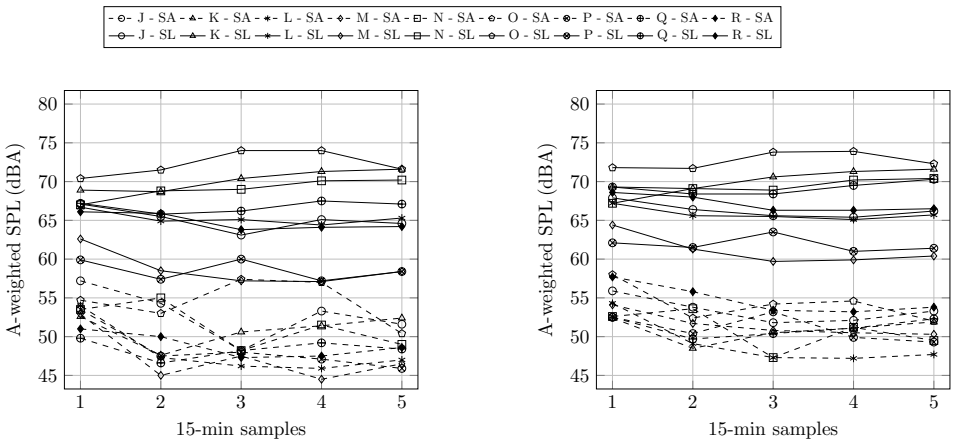
A power analysis allowed to increase the statistical significance of the study. First, each lesson has been analyzed by slots of 15 minutes, increasing the sample size up to 45. Then, the *huge* effect size of about 3.8 calculated for SA and SL populations allows the analysis to reach a significance level of $p < 0.001$ and statistical power of 100%.

The data sample augmentation led to the analysis of the temporal fluctuations of SNR during lectures. Irrelevant speech noise, i.e. SA in this case, can affect speech intelligibility with informational or energetic masking [19, 116]. Thus, it is important to consider to what extent the SNR varies during lessons [133].

The temporal trends of each lecture before and after the treatments for GMM and KM are shown in Fig. 3.17. SL are shown with solid lines, SA with dashed lines. The results between the methods seem to be quite consistent with similar tendencies. Exceptions are represented by lesson G before the treatments and lesson O after the restoration. Differences in this kind of analysis are strictly related to the shape of the occurrence curve and the consequent variance of data, as seen in Figure 3.13a. In these two lessons – G and O – the occurrence curves are particularly



(a) Before - GMM (b) Before - KM



(c) After - GMM (d) After - KM

Fig. 3.17 15-minutes samples of student activity (SA) and speech level (SL) for each lecture and each algorithm before and after the acoustic treatments of the halls. SA and SL are indicated respectively with dashed and solid lines. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Table 3.7 Correlation matrix among the main parameters of this study before and after the acoustic treatments of the halls. The main parameters are: the global, i.e. evaluated on the L_{eq} occurrences, student activity (SA) and speech level (SL) for both methods GMM and KM, occupancy N, and equivalent absorption area in occupied state A_{occ}. Before and after correlations are indicated respectively in regular and bold. From "Effectiveness of acoustic treatments and PA redesign by means of student activity and speech levels" by De Salvio et al [35].

Correlation coefficients - Before/**After**						
	SA - GMM	SA - KM	SL - GMM	SL - KM	N	A_{occ}
SA - GMM	1	-	-	-	-	-
SA - KM	0.85/**0.66**	1	-	-	-	-
SL - GMM	0.85/**0.61**	0.65/**0.21**	1	-	-	-
SL - KM	0.81/**0.57**	0.71/**0.29**	0.97/**0.98**	1	-	-
N	-0.48/**0.13**	-0.52/**-0.54**	-0.32/**0.34**	-0.36/**0.14**	1	-
A_{occ}	-0.44/**0.07**	-0.38/**-0.54**	-0.31/**0.33**	-0.32/**0.16**	0.90/**0.99**	1

skewed. This results in a SA curve with a high variance and a heteroscedasticity broadly pronounced. The small differences obtained for lesson O in Table 3.6 can be due to the average calculated on the whole recording and the smaller standard deviation measured between the two sound level meters after the acoustic treatments. Despite both techniques producing similar results, the KM seems to be less sensitive to variations. This stability may be due to the sharp borders of the KM algorithm.

Correlation coefficients before and after the renovation have been evaluated between SA and SL for both unsupervised methods besides occupancy and equivalent absorption area in occupied state. Table 3.7 shows the correlation matrix. Before and after correlations are indicated in regular and bold, respectively. The first interesting results concern the decrease of the correlation among SA and SL for both techniques GMM and KM. This is particularly evident for KM, which lowers the correlation from 0.71 to 0.29 whereas for GMM from 0.85 to 0.61. The drop of the coefficient states that SA and SL keep on having a growing and related tendency but weaker. The matrix points out as GMM and KM calculate SA differently. The acoustic treatments affect deeply the anti-correlation between A_{occ} and SA, which is completely lost for GMM and strengthen for KM. Concerning SL, despite its weak correlation with A_{occ}, it is worth noting how the regression slope changes for both GMM and KM as seen in Fig. 3.15. More in general, A_{occ} loses its anti-correlation with almost all parameters except for SA calculated via KM.

Summary

The present chapter shows some applications of the proposed method in university lecture halls. An acoustic discomfort in educational spaces causes different drawbacks, like the Lombard effect, listening and vocal efforts of students and teachers, respectively. In such spaces, where communication is the main activity, the ability to identify the sound contribution of each source is needed. The analyses conducted are focused on the detection of the student activity (SA), i.e., the chatting made among students during lectures. SA can be deemed as a metric to assess the focus extent of students. The chapter is basically divided into two parts.

In the first part of the chapter, the methodology to detect SA is discussed. Based on the literature, different techniques are compared: two visual methods, i.e., the conventional praxis, called PL, and a conditioned GMM, called PD; two blind methods, i.e., the Gaussian Mixture Model, called GMM, and the K-means clustering, called KM. Results show similarities between PL and KM and between PD and GMM. The main reason lies in the difference between hard and soft clustering performed by the algorithms. The first assigns a single data point to only and only one cluster. The second could assign a single data point to more than one cluster based on probabilities weights. KM is a hard algorithm. PD and GMM are both soft clustering algorithms that perform a Gaussian mixture fitting with different boundary conditions. PD performs the fitting according to constraints decided by the operator, and GMM does it via the Expectation-Maximization algorithm. PL is not an iterative algorithm but describes the acoustic scene through the equivalent continuous level and the 90th acoustical percentile of the entire SPL population collected during the monitoring. Hence, it is highlighted how PL relies on strong assumptions, not always fulfilled. Visible peaks of the occurrence curves and flex points of the cumulative curves are detected by both GMM and KM. Blind methods prove to be methodologically more robust and consistent with the features of the curves obtained by the measurement.

The second part of the chapter analyzes the behavioral changes of students after acoustical renovation works in two of the lecture halls shown in the first part. After a comprehensive overview of the treatments, both active and passive, the SA is analyzed only through blind methods, i.e., GMM and KM. However, the comparison with the conventional praxis, the PL, is kept. Results show a decrease in the regression curves concerning the correlation between SA and the other metrics, i.e., the teachers' speech levels and the occupancy during lectures. Spectral analyses show the reliability of the methods to reconstruct the spectra of each source measured simultaneously. Further discussions prove how SA is a reliable and measurable

metric to assess the comfort and the behavior of the students and their relationship with the acoustical context.

Chapter 4

Applications in offices

Keywords: *speech intelligibility, background noise, mechanical noise, human noise.*

The method proposed in this work has been deeply studied and tested in offices. The acoustic environment of workspaces represents a complicated and sensitive context to address. Workplaces are one of the most lived-in spaces by people. Offices can have different shapes, volumes, acoustical properties, and intricate characteristics to monitor. This complexity deals with the need of an appropriate acoustic comfort experienced by workers. Noises can affect deeply the individual perception basing on personal factors, tasks to accomplish, and the nature of the noise itself [46, 15, 82]. In such a challenging context, the acoustic comfort results to be a function of the task to perform, the architectural setting of workstations, and the balance needed between the ease of communication and concentration [113].

The short-term memory is strictly bound up with attention. This was proved in several psychological cognitive experiments that showed how irrelevant sounds disrupted the ability to accomplish tasks by participants. Besides different kinds of sounds, also speech signals were used [77, 47]. Thus, the understanding of colleagues' speech, when not involved in the conversation, is one of the most distracting noises in offices [64, 15].

The acoustic comfort in workspaces raised the attention in last years because of the increase of open-plan offices. Here, different activities are carried out by workers simultaneously. However, activities could be acoustically contradictory, e.g. speech communication and the need of quiet to focus on individual works. Further, a certain extent of speech privacy has to be provided per each employee. Thus, the design of open-plan offices involves accurate considerations concerning the layout of workstations and mutual arrangement of teams or workgroups. Acoustical performance in such complex spaces regards sound absorption, background noises,

height of screens and placement of furnitures and desks considering their mutual distances.

The reverberation time is commonly used as the main indicator of the acoustical properties of spaces. In some cases it could result as relevant to outline the characteristics of an office. In other cases it is not enough to describe the performance of rooms designed for different and simultaneous acoustical purposes. Spatial decay of SPLs, STI, and background noise levels must be considered as metrics for a more complete evaluation.

Two ISO standards are focused on the measurement, characterization, and design of open-plan offices: the ISO 3382-3 and the ISO 22955 [73, 71]. The ISO 3382-3 defines the metrics needed to accurately describe the performance of offices. Besides the spatial decay rate of speech $D_{2,S}$ and the A-weighted SPL of speech at a distance of 4 m $L_{p,A,S,4m}$, most of the metrics are STI-related. Thus, it is confirmed that the intelligibility represents one of the most effective workers' performance metrics [64]. Figure 4.1 shows to what extent the STI affects the performance of cognitive tasks according to four different models. The x-axis shows the STI and the y-axis the percentage of performance's decrease [52]. The relationship describes the change in performance, not the magnitude. Different tasks could be more or less affected by the irrelevant speech, e.g. intensive concentration tasks are more influenced by intelligibility than routine tasks. The negative effects of speech on work performance begin to vanish when the STI is about below 0.50 for most of the models. Therefore, the distraction distance metric r_D described in the ISO has been set at the distance where STI reaches 0.50. The negative effects of speech on work performance disappear when the STI is below 0.20. Therefore, the privacy distance metric r_P described in the ISO has been set at the distance where STI reaches 0.20.

The strong relationship between workers' performance and well-being and the intelligibility highlights the importance of the perceived noise levels at workstations. Background noise assumes a key-role, indeed. According to the considerations mentioned above, the amount of noise experienced by employees during working hours represents a balancing factor for the design of workspaces. Low SNRs disrupt the speech signal decreasing the intelligibility. Thus, theoretically speaking, high levels of noise are useful to prevent the distraction due to the irrelevant speech. At the same time, quiet places are needed to boost the focus abilities. As a consequence, noise should be high enough to disrupt the intelligibility but low enough to allow the concentration [130, 40]. However, the ISO 3382-3 states that the STI should be evaluated ignoring the noise produced by the human activity. This assumption leads to underestimate the real acoustic environment experienced in the office [55]. Moreover, it implies that the worst distracting scenario is represented by a single

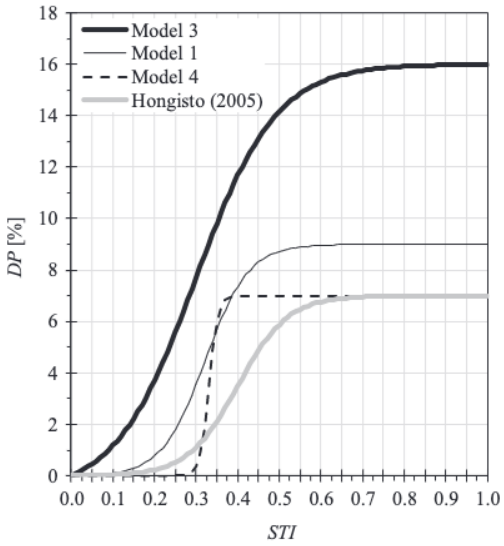


Fig. 4.1 Comparison of models about the relationship between the speech transmission index (STI) and the decrease in performance (DP). From "The relation between the intelligibility of irrelevant speech and cognitive performance—A revised model based on laboratory studies", by A. Haapakangas, V. Hongisto, A. Liebl [52].

talker and not by a multi-talker context [153, 151]. The ISO 22955 focuses on containing speech propagation [71]. Its purpose is to limit the disturbance among adjacent workstations but also to optimize the comfort within short-distance conversations. The most important contribution of the standard is the assessing noise levels at workstations depending on the activity carried out within the space. Also, Annex D provides surveys to subjectively rate the annoyance of noises in the office. Table 4.1 shows the kinds of activities considered and the respective A-weighted target values.

The combined approach of the two ISO standards deeply affect the acoustical design of workspaces, not only the open-plan type [54]. The contribution of the human activity broadens the analysis towards the assessment of a dynamic context where persons are considered no longer as passive receivers but also as active sound sources [119, 120]. The STI can be corrected in post-processing with more accurate background noise values. This means that the measurement of noise is crucial and can be performed in an accurate way [123, 44]. Different criteria, like the *liveliness* and percentile levels differences, were proposed to assess the impact of persons in the acoustical environment and the corresponding comfort and productivity [141, 121].

This chapter shows two real-world applications in two different offices of the method under study. The main focus of the discussions concerns the separation, identification, and measurement of the main kinds of sound sources in active offices. Based on the requirements described in ISO 22955 to measure the workstation noise

Table 4.1 ISO 22955:2021: target values for workspaces and workstations depending on the activity conducted.

Activities in its own space	
Activity	Target values (dBA)
Activity mainly focusing on outside of the room communication	55
Activity mainly based on collaboration between people at the nearest workstation	52
Activity mainly based on a small amount of collaborative work	48
Activity can involve receiving public	55
Workstation noise levels assumed for different types of activity	
Receiver space type	Target values (dBA)
Informal meetings (open plan)	48
Outside of the room communication (phone)	48
Collaborative	45
Non-collaborative	42
Focused phone	42
Focused individual work	40

levels $L_{Aeq,T}$, SPLs monitoring long at least 4 hours and with an occupancy of at least 80% were performed in two offices. In this context, the analyses concerned the most extensive generalization as possible of the method. The aim is to separate the human contribution from the other sources, e.g. mechanical noises or traffic. The ability of performing this kind of separation would represent a step forward for the technical praxis used in measurements and designs of workplaces at the time of writing. The first case study involved the unsupervised analysis of long-term monitoring carried out via GMM and KM, called Algorithm 1 and Algorithm 2. The results were compared with the common metrics used by scholars and standards, i.e. the continuous equivalent L_{eq} and the statistical levels L_{90}, L_{50}, L_{10}. Preliminary considerations about the features – means and standard deviations – were conducted according to literature. Further remarks about the reconstructed spectra and the influence of a realistic measurement on the STI were made in [34]. The second case study shows the same analysis, improved in some steps, to assess its reliability in a different environment. Besides the clustering analysis presented in [34], a dual analysis was carried out through a deep clustering approach. Thus, both machine and deep learning approaches were used to conduct cluster analyses on two different datasets of the same event. In this case, the sound level meter recorded both SPLs and the digital audio of the whole working day in one office. Then, SPLs were processed via GMM and KM, whereas 1-second length spectrograms of the digital audio were used as input of a variational autoencoder (VAE). A semi-supervised analysis of the VAE's latent space was used to assess the ability of a neural network to recognize patterns between the two different active sound sources in the office. Discussions concerned the spectra reconstructed via GMM and KM, hints about the

Table 4.2 Reverberation time T_{30} measured in the office under study. Results are shown in octave bands from 125 up to 4000 Hz.

	Frequency octave band (Hz)					
	125	250	500	1000	2000	4000
T_{30}	0.41	0.45	0.40	0.47	0.49	0.53

influence of the room's frequency response on the measurements, and the use of the VAE as a validation tool for the proposed unsupervised machine learning method.

4.1 Active sources in active office

4.1.1 Case study

To assess the proposed methods, an office with four workstations was chosen as case study. The room is placed in a building outside the city, far away from road traffic too. Thus, the noise expected during working hours is made only by the "inner" sources: the human activity, and the mechanical noise due to the HVAC system and the electronic equipments, e.g. computer fans, printers, etc. During the inspection conducted to set the sound level meter, the room was measured according to ISO 3382-2 to evaluate its T_{30} [72]. The office has a treated ceiling and, consequently, a low reverberation time. Hence, the room can be considered as a "dead" environment.

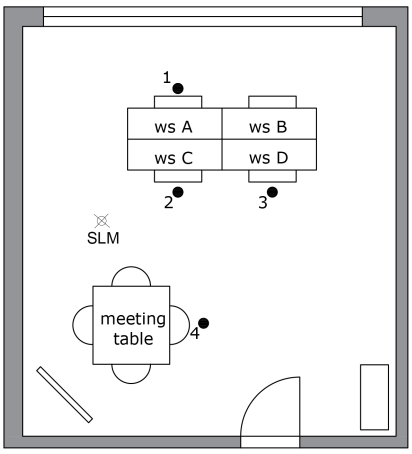


Fig. 4.2 Plan of the office under study. Numbers from 1 up to 4 indicate the measurement positions according to ISO 3382-2. SLM indicates the position of the sound level meter used for the long-term monitoring. From "Unsupervised analysis of background noise sources in active offices" by De Salvio et al [34].

The office under study is made by 4 workstations and a meeting table. Despite the small size of the space, the meeting table is far enough from the workstations. Figure 4.2 shows the plan of the office. The workstations are indicated with the letters A, B, C, and D. Numbers from 1 up to 4 show the positions used for the measurements. Table 4.2 shows the measured T_{30}. The SLM point shows the position of the sound level meter used for the long-term monitoring of the activities.

4.1.2 Long-term monitoring of the activities

An entire working day was monitored through a sound level meter. According to ISO 22955, the recording was long enough to obtain a significant SPL statistical population [71]. Short-time equivalent levels were acquired with a 100 ms interval time. Octave-band filtered SPLs were obtained from 125 up to 4000 Hz. The dataset obtained has been processed via GMM and KM, according to the procedure described in Chapter 2. Briefly, an outline of the three-step method is described:

- The first step concerns the preparation of the candidate models. Here, candidates were evaluated via GMM and KM from a number of clusters K from 2 up to 6.
- The model selection metrics evaluate the "best" among the candidate models: Silhouette coefficient (SC), Davies-Bouldin index (DB), Calinski-Harabasz coefficient (CH), and Gap statistic (GS). The best model is chosen according to the majority rule, i.e. the most frequent result obtained among the metrics.
- The features, means and standard deviations, are collected per each sound source described by the model selected. Temporal and metric features – i.e. standard deviation (s.d.) and the average intracluster distance (AICD), respectively – were used to label the sources as human or mechanical.

In offices, the human sound source can be deemed totally described by the speech, the most relevant human activity. Thus, previous studies about the speech s.d. can be useful to set the threshold to identify the human clusters. Concerning this, Bottalico and Astolfi studied vocal doses of elementary male and female teachers. They found an uncertainty of the SPL mean of about 4 dB [13]. Olsen measured a s.d. of the speech in the range of 4-6 dB [107]. Iannace et al. measured the mechanical system noise within an open-plan office in three operating conditions: two different speeds and the background noise with the HVAC system off. The s.d. in the first two cases were about 1 dB, in the third it was of about 4 dB [68]. Leonard and Chilton reported the measured ambient noise levels of previous studies in open-plan offices. It is

Table 4.3 Model selection step for the office under study. Results are shown for both GMM and KM algorithms. Each metric has been evaluated per each octave band from 125 up to 4000 Hz.

Gaussian Mixture Model						
Metric	Frequency octave band (Hz)					
	125	250	500	1000	2000	4000
Silhouette	2	2	2	2	2	2
Davies-Bouldin	2	2	2	2	2	2
Calinski-Harabasz	3	6	6	4	4	6
Gap Statistic	2	2	2	2	6	2
Best	**2**	**2**	**2**	**2**	**2**	**2**
K-means clustering						
Metric	Frequency octave band (Hz)					
	125	250	500	1000	2000	4000
Silhouette	2	2	2	2	2	2
Davies-Bouldin	6	3	4	4	6	2
Calinski-Harabasz	6	6	6	6	6	6
Gap Statistic	2	2	2	2	2	2
Best	**2**	**2**	**2**	**2**	**2**	**2**

shown how the difference between minima and maxima SPLs spans between 5 and and 11 dB [114]. Based on this literature, it is possible to set a preliminary threshold of about 5 dB. A cluster obtained via GMM is labelled as human if the s.d. is greater than 5 dB, otherwise it is assumed as mechanical.

Concerning the KM, there is no literature about the AICD of acoustical measurements. Thus, the classification follows the one made via GMM. Qualitatively, low values of AICDs are associated to mechanical clusters, otherwise to the human ones.

Further considerations can be made about the proportions of data contained in the clusters with respect to the total dataset size. Thus, the cluster mixing proportions for GMM and the percentage of data over the total size for KM were collected. These parameters can be deemed as temporal information about the activity of each source during the monitoring. To preliminary assess the reliability of the results and compare with previous studies, the equivalent levels L_{eq} and the statistical levels L_{90}, L_{50}, L_{10} were collected [121].

Table 4.3 shows the results of the candidate model selection step for GMM and KM, respectively. Each metric has been evaluated per each octave band. The selected model are shown in the last row named "Best". Results show that $K = 2$ is the best number of clusters in the collected data for both algorithms. This is consistent with

Table 4.4 Results of the third step for both algorithms, GMM and KM. The final outcomes associated to mechanical or human sources are shown. They are obtained running both algorithms with $K = 2$, the optimal number of clusters found in the second step through the model selection metrics (see Table 4.3). For GMM, s.d. and the mixing proportions of the Gaussian curves are shown in brackets. For KM, the AICD and the size of each cluster expressed as percentage on the whole population are shown in brackets. Lastly, the equivalent levels L_{eq} and the statistical levels L_{90}, L_{50}, L_{10} are shown for comparison. All values are presented in dB per each octave band. From "Unsupervised analysis of background noise sources in active offices" by De Salvio et al [34].

Source type	Frequency octave band (Hz)					
	125	250	500	1000	2000	4000
GMM – K=2						
Mech. (L_B)	32.5 (2.7 – 0.73)	30.0 (3.1 – 0.67)	28.1 (3.9 – 0.65)	22.2 (2.3 – 0.53)	18.6 (1.3 – 0.44)	21.6 (0.8 – 0.61)
Human (L_S)	41.6 (7.0 – 0.27)	41.3 (7.7 – 0.33)	40.7 (9.1 – 0.35)	32.5 (7.8 – 0.47)	28.0 (6.9 – 0.56)	27.5 (5.4 – 0.39)
KM – K=2						
Mech. (L_B)	32.7 (2.9 – 82%)	30.6 (3.4 – 79%)	28.6 (4.1 – 77%)	23.4 (3.2 – 75%)	20.3 (2.8 – 74%)	22.2 (1.7 – 84%)
Human (L_S)	45.5 (5.3 – 18%)	45.8 (5.6 – 21%)	45.8 (6.8 – 23%)	37.8 (6.4 – 25%)	33.7 (5.7 – 26%)	32.4 (4.6 – 16%)
L_{10}	42.9	45.4	45.5	37.7	34.0	30.0
L_{50}	33.3	31.7	30.3	24.6	21.3	22.1
L_{90}	29.5	26.7	23.7	19.9	17.5	20.7
L_{eq}	42.6	44.1	46.2	40.1	34.6	30.3

the expectations about the number of active sound sources in the office, i.e. human and mechanical.

Table 4.4 shows the results of the models obtained via GMM and KM with $K = 2$. Results are associated with each cluster, human or mechanical. Spectra have been reconstructed from 125 up to 4000 Hz. In brackets, s.d. and mixing proportions are shown for GMM, AICD and the percentage of the cluster size over the whole dataset is shown for KM. Besides the clustering results, the equivalent levels L_{eq} and the statistical levels L_{90}, L_{50}, L_{10} are shown.

4.1.3 Statistical insights about the active sound sources

The labelling step is strongly based on the SPLs variation within the same cluster. Concerning the GMM the s.d. is used to evaluate the size of clusters, whereas the AICD is used for KM. The intuition is the same for both algorithms. The larger the cluster, the more random the source is. High SPLs variations mean that the source varies in time, like the human activity. By contrast, low temporal SPLs variations suggest a mechanical cycle. This is confirmed by Table 4.4, where the tendencies of s.d. and AICDs are proportional. Consistent results have been obtained concerning the mixing proportions in GMM and the size of clusters in percentage in KM, where the larger clusters regard the mechanical noise, as expected. The only exception is represented by the 2000 Hz octave band for GMM. Here, the human cluster has a mixing proportion equal to 0.56. This could be due to the low contribution of the mechanical noise and the significant speech energy in that octave band.

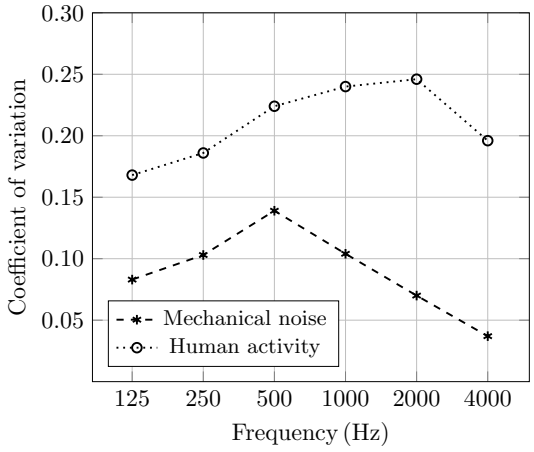


Fig. 4.3 Coefficient of variation of mechanical and human sources. From "Unsupervised analysis of background noise sources in active offices" by De Salvio et al [34].

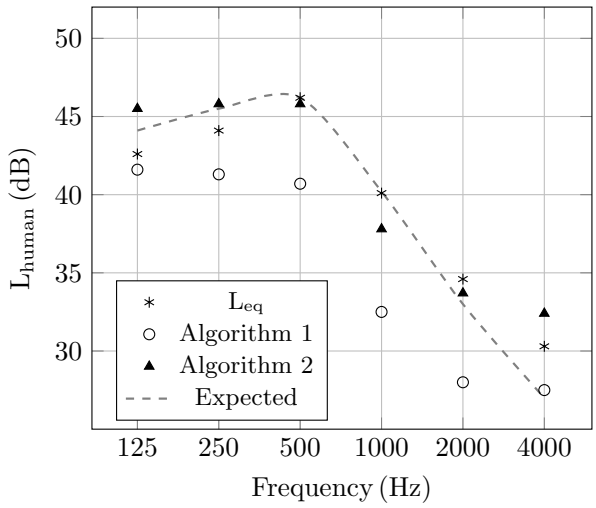
As seen in Section 1.1.3, GMM and KM can be related. In fact, GMM can be described as a generalization of KM for small variances [94]. This is confirmed by the results in Table 4.4. Mechanical clusters have similar means, especially in low frequencies where HVAC systems have higher power. Thus, the higher the variances, the larger the differences between GMM and KM are. As a consequence, the data homoscedasticity is not fulfilled.

To deepen the understanding of the dispersion of data in clusters, the coefficient of variation (CV) has been investigated. CV is also called "relative standard deviation" and is defined as the ratio between the s.d. and the mean of a population. Figure 4.3 shows the CV of both sources, human and mechanical, per each octave band obtained via GMM. The tendencies seem to be proportionally equal for both sources up to 500 Hz. Then, the behavior changes increasing the frequency. The spread of the human clusters increases up to the 2 kHz octave band. In this range the human activity randomness grows showing a higher dynamical behavior. This is consistent with the expectations since from 500 up to 2000 Hz most of the formants of the speech occur; hence, the biggest part of the energy of the speech. In general, the trend of CV confirms the assumption of identifying the human activity with the larger clusters.

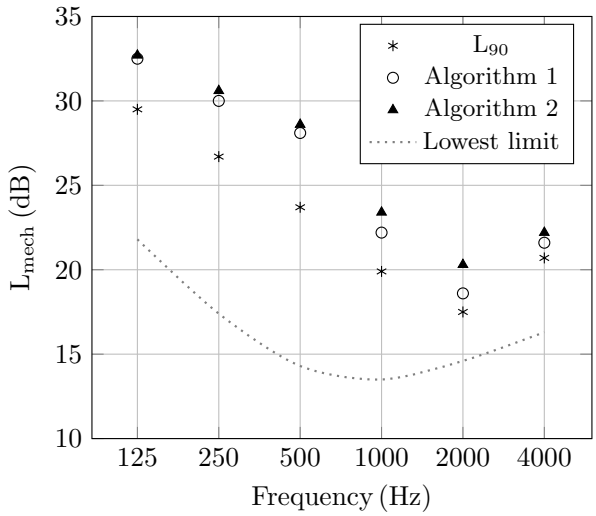
4.1.4 Spectral insights about the active sound sources

Further investigations to confirm the assumptions made to conduct the unsupervised analysis concerns the spectral matching. Since there is no a reference spectrum for HVAC noises, the analysis is focused on the speech cluster. The sound power levels of the speech shown in ISO 3382-3 were contextualized in the office under study. Values were averaged between male and female speakers and for a normal voice effort. Thus, the SPLs of the speech were calculated using the diffuse field hypotheses and the reverberation time in-situ measurements [73, 63]. The diffuse field can be assumed because all the distances among the workstations are greater than the critical radius. A talking time of about the 20% of the whole monitoring time was considered.

The background noise levels measured in the office never exceed 45 dB. Thus, the Lombard effect is not triggered and it is possible to use a constant value of speech power level [114]. Previous studies showed how the speech spectrum changes in noisy environments, especially in lower bands [87, 124]. This is observable in the results with an increase of 6 and 3 dB in the 125 and 250 Hz octave bands, respectively. The spectra obtained via GMM and KM are compared with spectral equivalent levels L_{eq}. Figure 4.4a shows the reconstructed and the expected spectra.



(a) Speech spectral matching.



(b) Inferred values of mechanical noise.

Fig. 4.4 Reconstructed spectra via GMM, KM, L_{eq}, and L_{90}. Figure 4.4a shows the spectral matching between the expected spectra calculated in the room under the assumptions of a diffuse sound field (dashed line) and the human clusters calculated. Figure 4.4b shows the inferred values of mechanical noise per each method. The dotted line shows the lowest detectable limit of the equipment used. From "Unsupervised analysis of background noise sources in active offices" by De Salvio et al [34].

The first is indicated with markers, the second with a dashed line. Algorithm 1 and 2 refer to GMM and KM, respectively.

The qualitative results confirm the assumptions: the human cluster is comparable with the speech, the main detectable human activity. However, differences have been achieved among methods. GMM gives back lower values with respect to the other techniques. The gap could be associated with the different way of the algorithms to cluster data. GMM performs a soft clustering process unlike KM which performs a hard clustering. Data points in GMM could belong to more than one cluster with an assigned probability. This does not happen in KM where each data point belongs to one and only one cluster [129]. Thus, the overlap area between two Gaussian curves obtained via GMM could affect the corresponding means. On the other hand, hard clustering seems to obtain results similar to the expected and L_{eq} tendencies. However, it should be noted that L_{eq} levels are strongly affected by low SNRs. This is confirmed by the flat trend of the speech spectra in low frequencies. The high contribution due to the HVAC system influences the speech energy that in 125 and 250 Hz octave bands is represented only by the formants. Nevertheless, the flat tendency at the low frequencies falls within the uncertainties mentioned above. Moreover, as also seen in Figure 4.3, the coherent tendencies of CV, noted for both sources, confirm the challenge to separate the human and the mechanical contributions when both have high energies.

Further proof about the reliability of the proposed method can be found in the size of clusters. The average percentage over all the octave band is of about 21% of the monitoring time. This can be assumed, in a first approximation, as the percentage of speech occurrence in the office during work hours. Thus, KM gives back tendencies more similar to the energy model, since it is near the expected curve of the speech for the 20% of the whole monitoring time. Short gaps of the mechanical noise especially in the frequencies where the most of the energy lies, i.e. from 125 up to 500 Hz, obtained via GMM and KM confirm that the heteroscedasticity condition is fulfilled. High differences between measured spectra and L_{90} seem to highlight how the common praxis to assume the mechanical noise as a percentage of the exceed time is not robust.

An unforeseen tendency can be noted in the 4 kHz octave band in both spectra, mechanical and speech. In fact, a strong decrease of these values is expected for both sources. The dotted line represents the lowest detectable limit of the sound level meter used. Considering the quiet environment of the office, the growth of the levels in the 4000 Hz band of the mechanical spectrum, as well as the small decrease of this octave band in the speech spectrum, can be attributed to the intrinsic error of the instrument.

4.1.5 Influence of background noise on STI evaluation

Raytracing simulations of the office allowed to assess a more realistic scenario of the intelligibility among workstations. The 3D model was created using SketchUp and imported into ODEON Room Acoustic software. The geometry and the modelling pipeline follows the recommendations of the state-of-the-art [145]. All the surfaces were modelled up to the size of 0.35 m and the sound absorption coefficients were supplied by the scientific literature [29, 145, 31]. Since the software describes the sound through rays, the wave nature of the phenomenon is ensured by the introduction of scattering coefficients [122]. The layers were managed dividing the elements with high absorption and scattering coefficients from reflective and smooth elements. The model was deemed calibrated when the differences of the measured and simulated T_{20} lie within the range of the just noticeable difference (JND).

The intelligibility was evaluated through the STI and corrected with different values of background noise. The first was STI_{∞}, i.e. without any contribution from the background noise. Then, the mechanical contribution was added, and finally the sum of both sources was considered. Figure 4.5 shows a gray scale variations of the STI matrix with the different contributions explained above. The scale varies from 0.5 (in black), i.e. the minimum STI detected, up to 1 (in white), the maximum intelligibility achievable when source and receiver are in the same workstation. Algorithm 1 and 2 refer to GMM and KM, respectively.

Focusing on the rows of the matrix, slight differences are noticeable between the first two matrices, i.e. the STI_{∞} and the contribution of the mechanical noise. Only in the third matrix, when both the types of noise are considered, the shades are darker. Results highlight the importance of a detailed analysis of the background noise to evaluate the intelligibility in a more realistic scenario, avoiding the overestimation of this fundamental metric. Measuring the privacy condition considering only the mechanical noise, as currently required by the standards, is not sufficient to assess the effective privacy environment of the office, which is significantly affected by the social context [117].

When the active noise masking is used, the speech privacy can be assumed as quite constant over the working areas. In other cases, like the one under study, the privacy fluctuates dynamically in time. The results obtained with the procedure presented in this work allow to assess different scenarios, thus broadening the characterization of privacy criteria of ISO 3382-3. Further analyses could be done with these unsupervised Algorithms with longer monitoring times, in order to investigate the existence of more significant correlations with percentile levels [121].

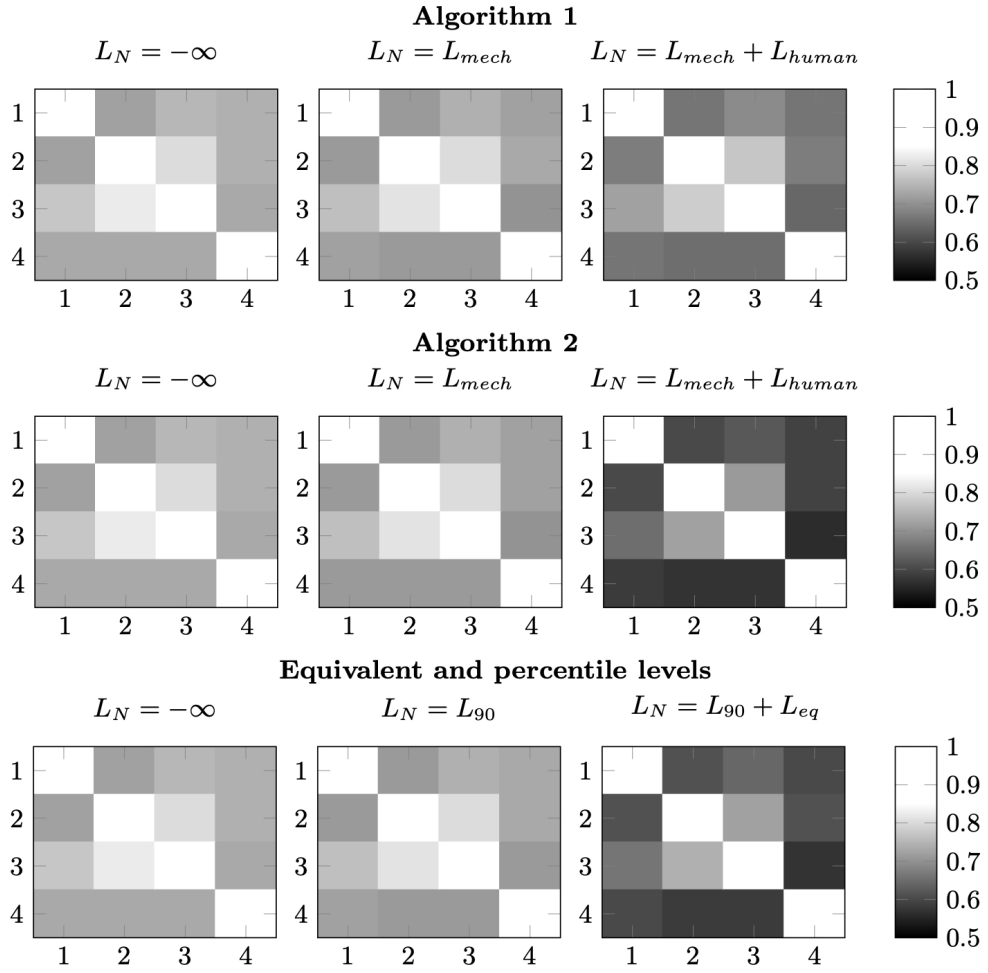


Fig. 4.5 Matrices of the STI values among the workstations in the office under study. The source has been set at the "normal" speech level. On each row, going from the left to right, the STI is presented first without background nose (indicated as $L_N = -\infty$), then corrected with the background noise levels obtained through the unsupervised analysis. First adding the mechanical contribution only (indicated as $L_N = L_{mech}$), then summing up the human contribution as well (indicated as $L_N = L_{mech} + L_{human}$). The sidebar on the right represents the legend of the STI values. On the axis of the matrix are reported the source-receiver positions (1 – 4) corresponding to the three workstations and the meeting table (see Figure 4.2). From "Unsupervised analysis of background noise sources in active offices" by De Salvio et al [34].

4.2 Offices short survey

The proposed method has been tested in three offices to broaden its application and collect more data. The activities carried out are different in each case study and have been identified according to ISO 22955 [71] (see Table 4.1). Here, a brief description of the three offices is presented:

> **Office A** - Open plan with 8-12 workers. The amount of employees can vary during the day. This is a sales office. Hence, the activity is mainly focused on outside of the room communication.
>
> **Office B** - Open plan with 10-12 workers. The amount of employees can vary during the day. This is a design office. Hence, the activity can be more or less collaborative.
>
> **Office C** - Small office with a maximum of 2 workers. The amount of employees can vary during the day. The activity carried out has a small amount of collaborative work.

This study was conducted after COVID19. Thus, besides the occupancy's variations, all offices were treated with screens 120 cm height.

Activity in two different days within three different offices was measured through a sound level meter. Sound pressure levels (SPLs) were recorded each 0.1 seconds to reach a high resolution monitoring. Such a short acquisition time allows to record SPLs even in the pauses among syllables of the speech [146]. About 430k samples for each day were collected in octave bands from 125 to 4000 Hz besides the global A-weighted average levels. The arrays obtained by the time series represent the database for the application of the two algorithms, GMM and KM.

The procedure used in the present study follows the analysis described so far. Thus, in a nutshell, the optimal number of cluster is obtained according to the majority rule among the model selection metrics: DB, CH, SC, GS. After the best model is picked, the next step labels the sound source as mechanical or human. The means and the standard deviations for GMM and the centroid and the average intra-cluster distance (AICD) for KM represent the feature used to assign the labels. The logic behind the labelling is based on two assumptions. The first concerns the sound pressure levels of the sources. The mechanical noise should be lower than the speech, indeed. Thus, the lower mean and the lower centroid will be assigned to the mechanical sources. Higher values will be assigned to the human noise. The second assumption regards the variance of the SPLs source. A mechanical process should measure similar SPLs because of the mechanical cycles, while the speech

does not follow always the same rhythm. Thus, lower s.d. and lower AICD will be assigned to the mechanical sources. Higher values will be assigned to the human noise. Preliminary analyses found a value of s.d. equal to 5 dB as a good threshold to separate mechanical from human sources [89].

Numerical models allow to deepen the context analyzing the acoustic properties of the spaces under study. In this work, one or two metrics were used to describe the spaces. For all the offices the reverberation time T_{60} is considered. For the biggest offices A and B, the spatial decay of the A-weighted level of speech doubling the distance from the source $D_{2,S}$ was also taken into account. Following, a brief summary of the simulated parameters:

Office A - $T_{60} = 0.7$ s; $D_{2,S} = 2.4$ dB;

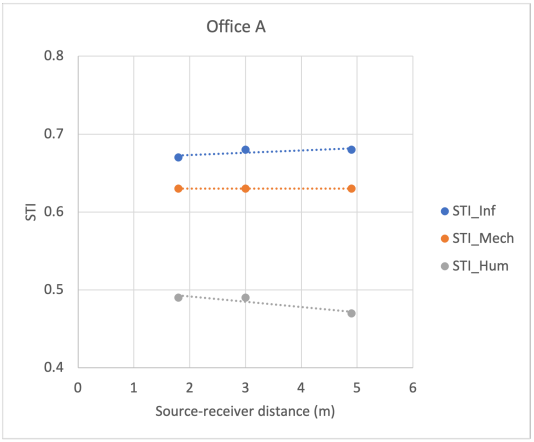
Office B - $T_{60} = 0.7$ s; $D_{2,S} = 3.0$ dB;

Office C - $T_{60} = 1.3$ s.

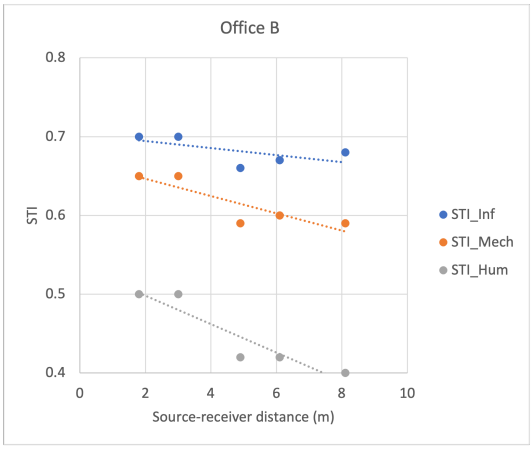
Concerning the model selection, the majority rule gave back an optimal number of clusters equal to 2 for both GMM and KM. This means that we can consider the sound context within the measured office made by two main sound sources: mechanical (air systems, electronic devices, ...) and human noise (speech).

Table 4.5 shows the results of the clustering analysis carried out over all offices. Values are shown for the octave bands from 125 to 4000 Hz and the global A-weighted average levels. Differences between the two algorithms are deemed as a consequence of the heteroscedasticity and the respective soft/hard clustering assignment, as already discussed in the previous case studies. It is worth noting that human noise obtained via KM has similar values to the A-weighted average levels. Differences are less than 1 dB for offices A and B. Office C shows differences of 1.3 and 2.2 dB. It can be correlated to the type of activity carried out in this office. The small amount of collaborative work can affect the occurrence curve of the human cluster. It is noticeable by the higher AICD measured in this office with respect to A and B.

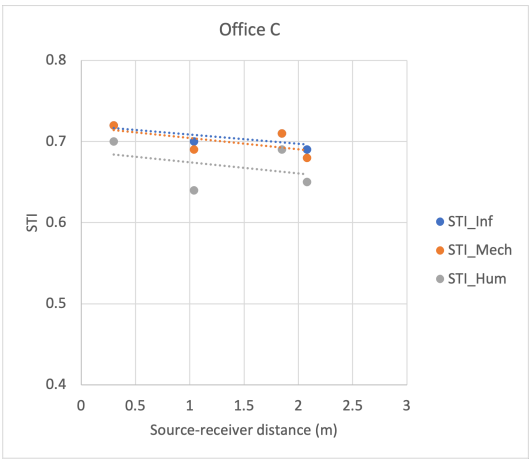
Spectral tendencies confirm the type of sources as shown in previous section. Figure 4.6 shows the decay of the STI with respect to the distance. Three curves are shown: the STI_{Inf} without background noise, the STI_{Mech} considering only the mechanical noise obtained by the clustering, and the STI_{Hum} considering all the sound sources inside the offices. The plots show how the ability of separate sound sources provides deeper insights of the spaces under study.



(a) **Office A**



(b) **Office B**



(c) **Office C**

Fig. 4.6 Relationship between STI and distance. Blue lines show the STI evaluated without background noise. Orange and gray lines show the STI considering the HVAC and HVAC combined with the human noise contributions, respectively. From "Assessing human activity noise in workspaces using machine learning and numerical models" by De Salvio et al [36].

Table 4.5 Results of the clustering carried out over long-term monitoring of the three offices. The office, the correspondent measurement day, the algorithm and the kind of source are shown. Measured SPLs are shown for each octave band from 125 to 4000 Hz, besides the A-weighted values. Moreover, the correspondent A-weighted continuous-equivalent level $L_{Aeq,T}$ measured through the sound level meter is shown. From "Assessing human activity noise in workspaces using machine learning and numerical models" by De Salvio et al [36].

				ML-based methods							Traditional method
Office	Day	Algorithm	Source	Frequency octave band						A-weighted	$L_{Aeq,T}$
				125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz		
A	1	GMM	Mechanical	47.7 (1.7)	43.8 (1.4)	37.0 (1.2)	30.3 (1.3)	24.8 (2.3)	19.8 (3.7)	39.8 (1.0)	56.5
			Human	53.2 (4.3)	52.3 (5.3)	52.6 (8.5)	42.5 (8.6)	37.6 (8.5)	32.6 (8.4)	51.9 (7.7)	
		KM	Mechanical	47.8 (1.8)	44.3 (1.9)	38.3 (3.0)	32.5 (3.5)	27.6 (4.2)	22.3 (4.7)	41.3 (2.9)	
			Human	55.5 (3.1)	55.1 (3.9)	57.2 (5.9)	49.6 (6.4)	43.6 (6.4)	37.8 (6.0)	57.0 (5.4)	
A	2	GMM	Mechanical	47.2 (1.7)	43.8 (1.7)	37.7 (1.5)	31.7 (1.8)	26.6 (2.5)	22.3 (3.6)	40.2 (1.4)	55.4
			Human	52.9 (4.3)	53.0 (5.6)	51.8 (8.0)	43.4 (8.0)	38.2 (7.7)	32.8 (7.4)	51.8 (7.1)	
		KM	Mechanical	47.4 (1.8)	44.3 (2.1)	38.9 (3.0)	33.3 (3.3)	28.4 (3.7)	23.3 (3.9)	41.5 (2.8)	
			Human	55.3 (3.1)	55.8 (4.1)	56.4 (5.6)	49.2 (6.0)	43.3 (5.7)	36.9 (5.4)	56.3 (5.1)	
B	1	GMM	Mechanical	40.9 (2.2)	38.7 (2.4)	34.4 (2.2)	33.7 (2.1)	29.8 (2.3)	25.2 (3.1)	38.7 (2.1)	55.0
			Human	49.6 (5.1)	49.3 (6.4)	49.3 (8.3)	44.3 (7.4)	40.2 (7.0)	35.1 (6.8)	50.6 (7.1)	
		KM	Mechanical	41.9 (2.7)	40.1 (3.3)	36.4 (3.9)	35.2 (3.2)	31.7 (3.5)	26.5 (3.6)	40.7 (3.5)	
			Human	52.3 (3.6)	53.0 (4.5)	54.1 (5.8)	49.6 (5.7)	45.2 (5.2)	39.1 (5.0)	55.0 (5.1)	
B	2	GMM	Mechanical	41.7 (1.9)	38.2 (1.9)	33.3 (1.7)	32.7 (1.6)	28.7 (1.9)	23.8 (2.6)	37.9 (1.7)	54.5
			Human	49.1 (5.0)	48.2 (6.5)	47.9 (8.8)	43.2 (7.7)	39.3 (7.4)	33.6 (6.8)	49.5 (7.5)	
		KM	Mechanical	42.3 (2.3)	39.4 (2.9)	35.0 (1.1)	34.0 (3.0)	30.5 (3.2)	25.2 (3.4)	39.5 (3.1)	
			Human	51.8 (3.6)	52.3 (4.8)	53.6 (6.1)	49.3 (6.1)	44.9 (5.7)	37.9 (5.1)	54.6 (5.5)	
C	1	GMM	Mechanical	35.4 (3.2)	35.7 (3.2)	33.8 (4.8)	29.0 (4.2)	22.4 (3.7)	14.2 (1.1)	34.7 (3.5)	53.2
			Human	46.1 (7.4)	46.2 (7.5)	45.4 (8.8)	39.6 (8.1)	35.9 (8.3)	29.5 (8.5)	46.3 (7.9)	
		KM	Mechanical	36.2 (3.6)	36.5 (3.6)	34.2 (4.5)	30.3 (4.4)	25.9 (4.9)	20.6 (4.7)	36.6 (4.3)	
			Human	50.3 (5.3)	50.5 (5.4)	49.8 (6.4)	40.5 (5.7)	41.2 (5.7)	35.7 (5.9)	51.0 (5.7)	
C	2	GMM	Mechanical	34.7 (2.4)	33.5 (2.2)	31.4 (3.8)	27.7 (4.3)	22.0 (4.5)	16.9 (3.0)	33.5 (3.5)	51.7
			Human	43.8 (7.1)	44.1 (7.2)	45.5 (8.8)	38.6 (8.4)	34.7 (8.7)	29.6 (8.2)	45.7 (8.1)	
		KM	Mechanical	35.3 (3.0)	34.5 (3.0)	32.3 (4.2)	28.2 (4.2)	23.6 (4.8)	19.2 (4.3)	34.7 (4.1)	
			Human	48.3 (4.9)	48.8 (5.3)	50.0 (6.4)	43.0 (6.1)	39.7 (6.2)	34.8 (6.0)	50.4 (5.9)	

4.2.1 Overlapping areas

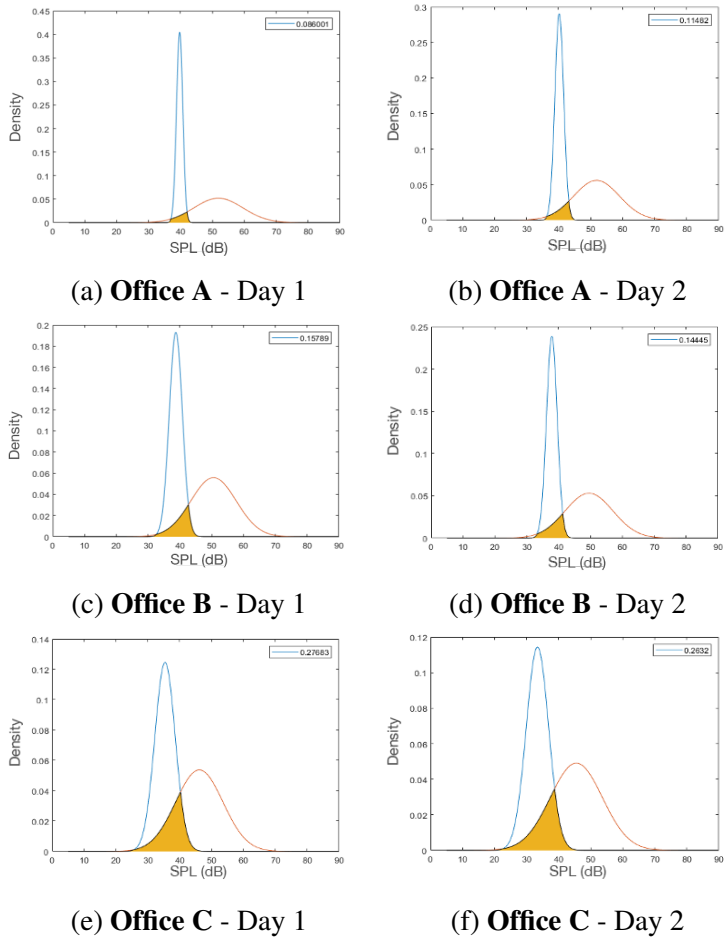
A soft clustering algorithm allows to evaluate data with a more realistic assignment, closer to the human perception [56]. GMM assigns data points to each cluster with a probability weight. Thus, each data point can belong to more than one cluster. As a consequence, GMM can have overlapped areas among the components, i.e. the Gaussian curves. Under the assumption that the mechanical component does not change during long-term monitoring, it may be deduced that the overlapped area depends on the speech component. Thus, it depends on the extent of the collaboration among workers.

On the basis of these considerations, the overlapping value (OvA) of the two components is proposed as a metric to assess the amount of collaboration among employees according to the ISO 22955. Measuring the overlapping areas between clusters is an important issue in the machine learning field. Hence, several algorithms were proposed [105, 137]. In the present analysis, the OvA value lies in the range [0,1]. OvA is equal to 0 when the two Gaussians do not have any overlapping and is equal to 1 when the two components are the same, i.e. totally overlapped.

Table 4.6 Results of the overlapping areas OvA for each combination of office and day. Values are shown for data population obtained per each octave band from 125 to 4000 Hz and the overall A-weighted average level. From "Assessing human activity noise in workspaces using machine learning and numerical models" by De Salvio et al [36].

Office - Day	**Frequency octave band**						
	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	$L_{A,eq}$
A - 1	0.307	0.154	0.067	0.119	0.179	0.256	0.086
A - 2	0.286	0.166	0.098	0.164	0.212	0.310	0.115
B - 1	0.208	0.194	0.122	0.203	0.216	0.291	0.158
B - 2	0.244	0.187	0.113	0.172	0.186	0.256	0.144
C - 1	0.277	0.288	0.365	0.365	0.232	0.082	0.276
C - 2	0.272	0.208	0.233	0.357	0.309	0.218	0.263

Table 4.6 shows the results obtained by the evaluation of the OvA. The values achieved for the average A-weighted levels show small differences between the days of each office but large differences among the case studies. Both results seem reasonable. The activity inside the space can vary day by day but globally, the fluctuations remain in small intervals. At the same time, the offices show different average OvAs. This result may be influenced by the number of workers in the office



(a) **Office A** - Day 1 (b) **Office A** - Day 2

(c) **Office B** - Day 1 (d) **Office B** - Day 2

(e) **Office C** - Day 1 (f) **Office C** - Day 2

Fig. 4.7 Overlapping areas per each office and each day. Blue Gaussian curves represent mechanical sources, red Gaussian curves represent human sources. The overlapping areas are highlighted in orange. From "Assessing human activity noise in workspaces using machine learning and numerical models" by De Salvio et al [36].

besides the collaboration among them. Offices A and B, both containing an average of 10 people, show a similar scale of values. This does not happen in office C that contains 2 people at most.

Figure 4.7 shows the plots of the components obtained via GMM. The considerations made above can be visualized here. The overlapping areas are highlighted in orange. Blue and red lines indicate respectively the mechanical and the human sources.

Preliminary results show how the OvA can be considered as a promising feature to measure the amount of activity within the offices. Increasing the database among further difference offices and corresponding correlation analyses will provide more

robust insights about this feature. The calculation of OvA over normalized probability density functions allows exploratory comparisons among different kinds of offices.

4.3 Qualitative validation through dual analysis

As seen in previous applications, the 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 [152]. However, the $L_{A,eq}$ does not show any detail about the acoustic scene [51]. 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 [109, 99, 100]. 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.

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 levels of abstraction [84]. Inspired by the cocktail party effect, i.e. the ability of humans to focus the auditory attention to one speaker filtering other stimuli [19], 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 [143]. 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 [57]. One of the most popular category to perform this kind of analysis is represented by the autoencoders. These kinds of networks perform 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 [101]. Applications of autoencoders in acoustics have been in the field of speech enhancement and clustering of geophysical data [92, 76, 108].

In this work, variational inferences were used to perform a deep clustering analysis. A variational autoencoder is a deep generative model that forces the latent code of autoencoders to follow a predefined distribution [101]. It has the same architecture of autoencoders, high-dimensional data are encoded into a low-dimensional latent space [79]. The ability of parametrizing data through a probability

distributions gained broad attention in the deep learning community. Variational autoencoders have been successfully applied to speech enhancement, blind source separation, and sound source localization in reverberant spaces [88, 103, 7].

The present section, based on the methods shown in previous sections, proposes a dual analysis of the same phenomenon. 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.

4.3.1 Case study

A small office with 3 workers place in 3 different workstations was selected as the case study for the dual analysis. It is worth noting that measurements were made during the COVID-19 emergency. Thus, persons wore face masks during the day. According to ISO 22955, the activity carried out among workers is collaborative (see Table 4.1).

The whole working day was monitored through a sound level meter placed similarly far from each workstation. Figure 4.8 shows the plan of the office and the placement of workstations and sound level meter.

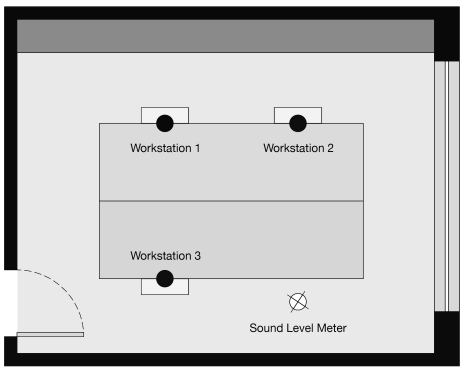


Fig. 4.8 Plan of the office under study. The arrangement of workstations and the placement of the sound level meter are shown. From “Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis” by De Salvio et al [33].

The same event was recorded to collect two different datasets:

1. Octave-band filtered SPLs from 125 up to 4000 Hz obtained with an interval time equal to 0.1 s;

2. Digital audio recording obtained with a sample rate equal to 51.2 kHz and a depth equal to 32 bit.

The two datasets represent the raw data used to conduct the two analyses. The sound level meter collected about 6 hours of working activity in the office. Figure 4.9 shows a 10-minute length example of time history that provides the two different databases. 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 on the bottom are exploited for the deep learning process.

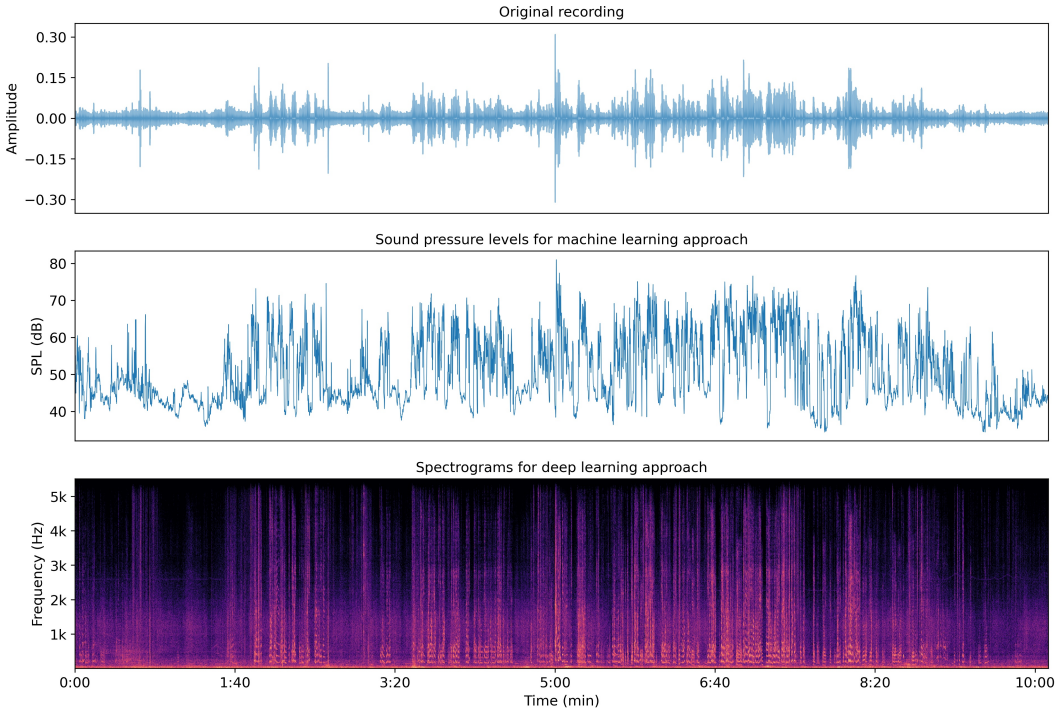


Fig. 4.9 Example of the raw data used in this study. On the top, a sample of 10 minutes recording. In the middle, the sound pressure levels obtained in the same sample. This constitutes one of the databases for the machine learning approach. On the bottom, the spectrograms obtained by the same sample used for the deep learning approach. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

The air system was turned off during the measurement and the window is exposed towards a highly busy road. Thus, the sound environment can be described as created by two kinds of sound sources: the traffic and the speech. The room has a volume of about 60 m^3 with no acoustic treatments and can be considered as a "live" environment. To better understand the acoustic context in which the working activities were conducted, the reverberation time T_{20} and the façade insulation $D_{2m,nT}$ were measured according to the precision method described in the ISO 3382-2 and

the global method of the ISO 16283-3 [72, 70], respectively. Figure 4.10 shows the measurements' results. Solid and dashed lines show the T_{20} and $D_{2m,nT}$ tendencies in octave bands from 125 up to 4000 Hz, respectively.

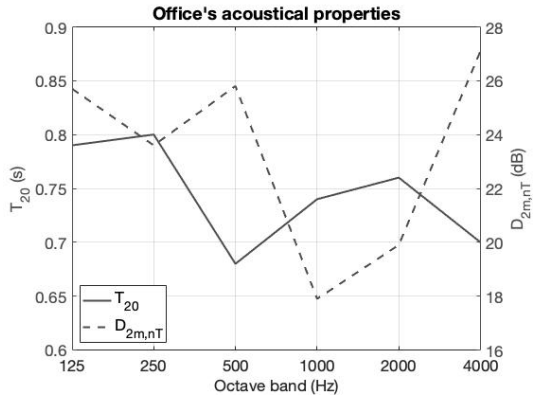


Fig. 4.10 Acoustical properties of the office under study. The reverberation time T_{20} is shown on the left axis, the façade insulation $D_{2m,nT}$ on the right axis. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

The office has a reverberation time averaged in 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 drop in the 500 Hz band maybe due to two steel closets. The façade has an average insulation of about 22 dB on the mid frequencies of 500 and 1000 Hz. The drop of the $D_{2m,nT}$ in the 1 kHz band is due to the glass coincidence effect of the window.

4.3.2 The dual analysis

Machine learning approach

The machine learning approach consists of the same analysis carried out in previous case studies, regardless whether it involved classrooms or offices. Details about the method are referred to the previous sections. Here, just a brief summary of the process is shown to recall the main steps:

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 source labelling according to statistical or distance metrics.

Step 2, the selection step, evaluated candidate models with a number of clusters from 2 up to 10 components. Spectral data were analyzed in the octave band from 125 up to 4000 Hz. The same analysis has been carried out for GMM and KM.

Labelling the sound sources, i.e. linking the spectra to the corresponding clusters obtained, is basically depending on the temporal characteristics of the sound source. Dense clusters represent continuous noises, while spread data refer to a random source. The machine learning approach being 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 s.d. equal or greater than 5 dB refers to a speech source. Lower values of 5 dB describe a mechanical or more in general a more or less steady source. Comparing preliminary studies, this value is considered to be a good threshold to separate continuous sound sources from human-related noises [107, 89]. 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 s.d., lower values are associated to continuous noises, otherwise to human noises.

Deep learning approach

The digital audio recording has been divided in 1-second length samples to obtain the dataset for the analysis through the VAE. Spectrograms of each segment were used as input for the network. A pre-processing flow has been carried out before feeding the encoder, briefly described below:

- The audio has been resampled at 11025 Hz to make the input comparable to the octave band range (125-4000 Hz) used in the machine learning approach. Moreover, observing the spectrograms, no useful information was 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 inferred that in each FFT only one sound source is detected.
- MinMAX normalization has been applied to each spectrogram to have all the amplitude values in the [0,1] range.

Overall, the dataset contained about 23k samples.

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,

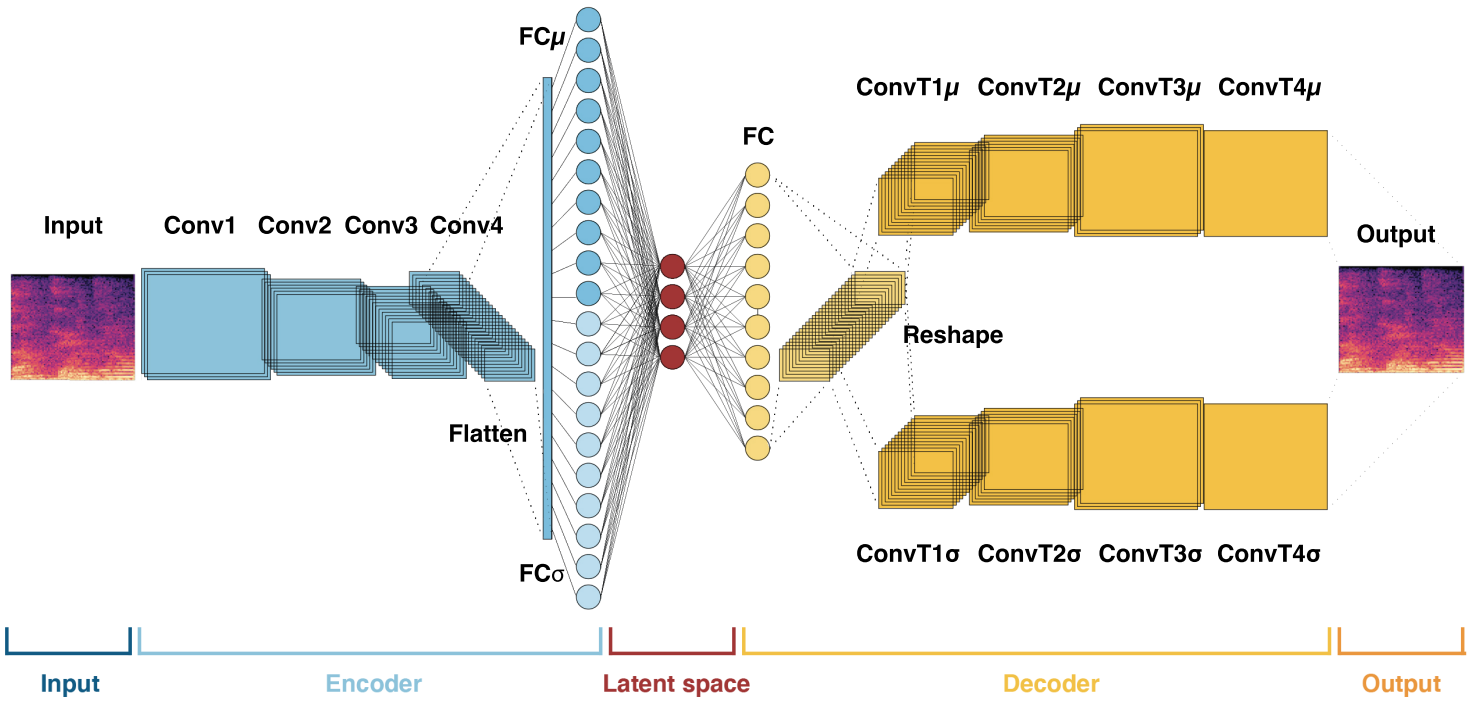


Fig. 4.11 Architecture of the VAE. The encoder is constituted by four convolutional layers, shown in light blue. The latent space is shown in red and the decoder is represented by the yellow blocks. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

speech, and unclassified sounds. The latter category was useful to label all the samples where the main listened sound 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 how the VAE learns to recognize the labels. At the end of the labelling 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 spectrograms' input size is 128×87. The encoder is made by four strided convolutional layers (stride $= 2$). Then, a flatten layer links the convolutional layers to the fully connected layers. A VAE maps the input to a multivariate latent distribution. The distribution used in the present analysis is the Gaussian distribution. Thus, each input is mapped through means and variances. As a consequence, the fully connected layer of the encoder is doubled. Here, the inputs are processed into the 30-dimensional latent space. The Pyro library was used to perform the stochastic variational inference. To sample inputs from the latent space according a Normal distribution, Pyro requires means and variances. Thus, the decoder is made by four strided transposed convolutional layers (stride $= 2$) for both parameters [7]. Then, spectrograms are reconstructed reshaping the output of mean and variances obtained by the two sections of the decoder. Non-linearities are activated through *ReLU* functions for all layers except for the output parameters. The decoder is parametrized according 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 4.11 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 4.7. 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.

4.3.3 Clustering results

First, the results of the clustering analysis carried out via GMM and KM are shown. Table 4.8 shows the results of model selection. Silhouette (SC), Davies-Bouldin

Table 4.7 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. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

	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]

Table 4.8 Model selection step for 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 most likely number of clusters used to run both GMM and KM algorithms. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

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

(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 4.12 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 selected references from standards. Blue lines show the results for GMM, red lines for KM. In the relative analyses plots, yellow lines refer to the reference spectra. To compare the reconstructed with references, each measured spectrum is shifted by setting the 1 kHz octave band to 0 dB. Table 4.9 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 case study described in Section 4.1.4, low frequencies seem to be easier separated for both algorithms. 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 traffic spectrum shown in EN 1793-3 [48]. 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 trend of the results. The shape of the traffic spectra seem to be 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 [24].

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 [73]. This is the reference for the speech source; the related spectra obtained via clustering have similar tendencies as shown on the bottom of Figure 4.12. 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, as seen in the above-mentioned previous section, 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 [124, 87]. 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 neither the same as the reference, nor omnidirectional. Moreover, at low frequencies, modal effects could have affected the results since the sound level meter was used only in one position.

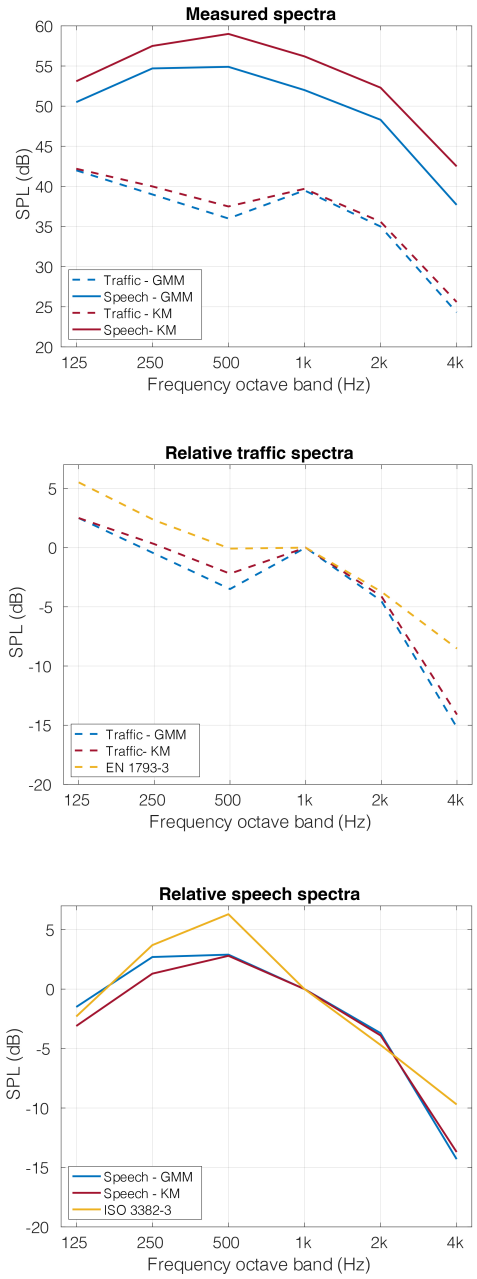


Fig. 4.12 Results of the clustering analyses. On the top: reconstruction of the spectra from 125 up to 4000 Hz. Blue and red lines represent the spectra reconstructed respectively via GMM and KM. Dashed and solid lines represent respectively the traffic and the speech spectra. In the middle and on the 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 [48, 73]. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

Table 4.9 SPLs of each sound source obtained via GMM and KM. Standard deviations s.d. for GMM and average intra-cluster distance AICD for KM are reported. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

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
s.d.	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
s.d.	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

Overall, concerning the relationship between the acoustical properties of the space and the spectra obtained, the tendencies of measurements' results in Figure 4.10 may bring preliminary insights about the comparison of measured and reference spectra shown in Figure 4.12. Traffic and speech spectra seem to be related to the tendency of the T_{20}. The drop in the 500 Hz octave band is visible in both sources, indeed. 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 combination of both T_{20} and $D_{2m,nT}$ seems to affect the energy of both sources in the 4 kHz octave band. Thus, a preliminary analysis of the room's acoustics seems to give further reliability to the results obtained through the machine learning approach. The drop in high frequencies may be explained considering that these can be strongly affected by their interactions with surfaces and furnitures in the room.

Further considerations can be made regarding the size of the clusters. This is described by the s.d. and the AICD; both are shown in brackets in Table 4.9. The physical meaning associated to s.d. 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 s.d. for continuous sources because the corresponding Gaussian curve will be narrow. On the contrary, a human-related noise produce higher s.d.. 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 s.d. lie in the range 3.0 - 4.3 dB for each octave band. Previous case studies in several offices showed mechanical s.d. 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.

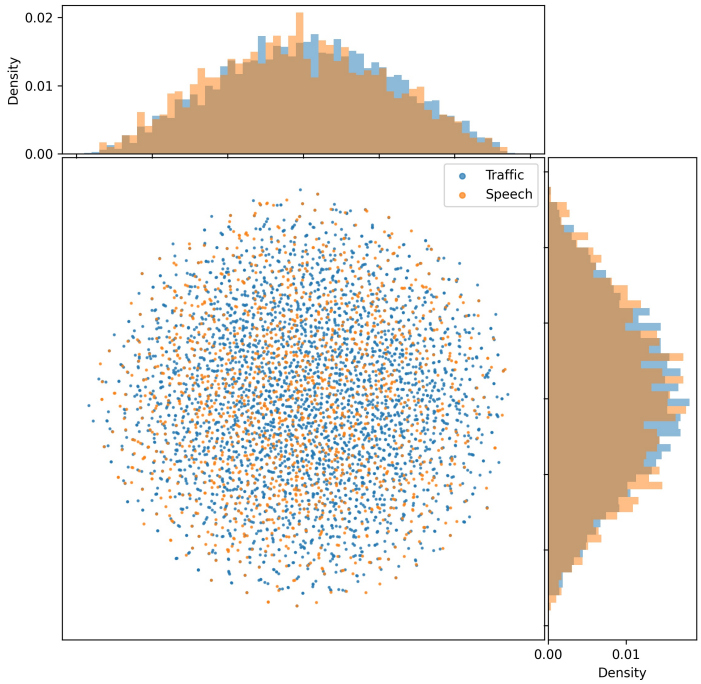
The absolute spectra shown on the top of Figure 4.12 point out differences between SPLs of the two methods that can be related to the homoscedasticity of data, i.e. constant variances of data. This is shown in Table 4.9. SPLs are the same for GMM and KM when s.d. and AICD are almost equal, e.g. in the 125 and 1000 Hz octave bands of the traffic source. This result confirms that AICD can be deemed as a reliable metric to assess the shape of the cluster.

4.3.4 Deep clustering results

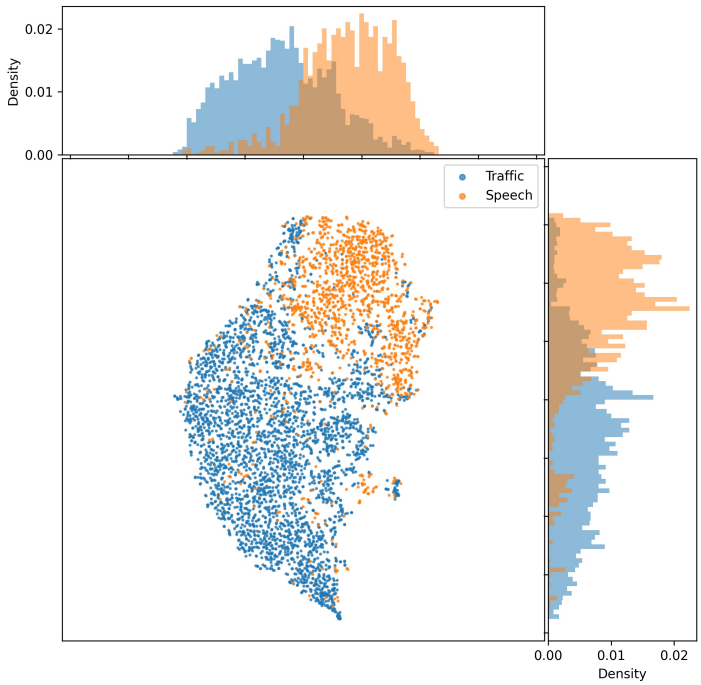
Latent representation of the working day

The clustering analysis carried out through the machine learning approach is totally unsupervised. Thus, considerations about its results have been based on assumptions and spectral matching. The discussions of these evidences depend on the operators' knowledge. Hence, it is useful to find an objective method to either confirm or not the quality of using the proposed method via 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 to. 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 analysis it was made at the end of the process, in the deep learning analysis data were previously labelled. Thus, the latent space of the VAE aims to be a qualitative tool to assess the machine learning approach.

Figures 4.13a and 4.13b show the latent distributions of the untrained and trained network, respectively. Because the dimension of the latent space is equal to 30, a 2D t-stochastic neighbor embedding (t-SNE) visualization was used [140]. This is a dimensionality reduction technique commonly used to visualize high-dimensional data. The t-SNE algorithm evaluates the similarity between pairwise instances in both high and low dimensional space. Then, through a cost function, the similarities are optimized. Figures 4.13a and 4.13b are obtained with a perplexity equal to 30, which resulted to be a stable value for the configuration obtained.



(a) Untrained network



(b) 400 epochs training

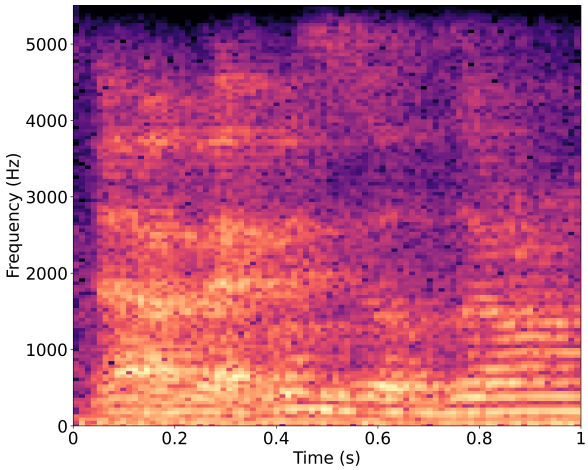
Fig. 4.13 Latent space of the untrained 4.13a and trained 4.13b VAE. Histograms show the x- and y- axes projections of the density distributions of the data. Blue and orange dots and histograms represent respectively the traffic and the speech data. From “Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis” by De Salvio et al [33].

Data in the latent space are represented based on their categorical label, i.e. blue dots refer to traffic spectrograms, orange dots to the speech. The untrained latent space in Figure 4.13a shows a circular distribution of data since it is perfectly described by a Gaussian distribution [30]. However, there is no categorical separation among data, i.e. blue and orange dots are mixed up. Figure 4.13b 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 overlap or not. Hence, histograms of the trained network prove that the two clusters are close but do not overlap. 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.

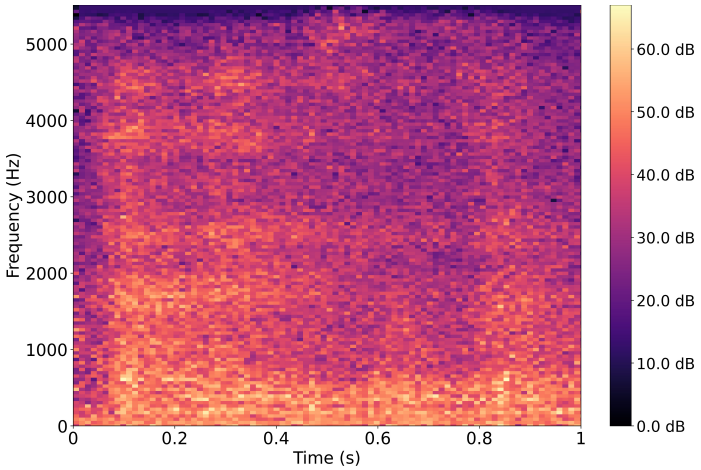
Measuring through deep clustering

The aim of the proposed method is focused on measuring different sound sources in real-world contexts. Through a deep learning approach this is possible reconstructing the audio samples. Frames of each class can be selected from the latent space and post-processed through octave-band filters 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 4.14a and 4.14b. The reconstruction is blurred and this is common in VAEs [103]. 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 in 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, it is easy to deduce that 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 pre-processing

step itself adds further uncertainty to the results. VAEs can identify underlying



(a) Untrained network



(b) 400 epochs training

Fig. 4.14 Example of original 4.14a and reconstructed 4.14b magnitude spectrograms obtained through the VAE. From "Blind source separation by long-term monitoring: a variational autoencoder to validate the clustering analysis" by De Salvio et al [33].

structures of data. With respect to standard autoencoders, they push the latent code to follow a predefined distribution [101]. 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. This represents additional confirmation to the considerations about the use of Gaussian distributions for long-term monitoring data made in Chapter 2.

Based on all the results and considerations made in the present chapter, clustering techniques, GMM and KM, 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 SPL occurrences. The second factor concerns the need of a deep learning approach for recording audio in work contexts. This can lead to privacy issues, one of the most important aspects on the application of big data approaches in real contexts [78]. On the contrary, clustering techniques provide simple and smooth applications on measuring sound environments. It is worth to recall that GMM can be considered as a generalization of KM. Thus, the GMM can be deemed as the most reliable method to perform blind source separation on sound level meter data for these reasons: the better performance in the method's flow, e.g., the results concerning the optimal number of clusters; the ability to explore more in details the results obtained, e.g., the OvA application; and the convenience of using an underlying well-known probabilistic model and the resulting features to use.

Summary

The present chapter shows some applications of the proposed method in offices. The nature of a sound source inside workspaces can affect in different ways the employees' productivity. The most annoying sound for a worker is represented by irrelevant speech, i.e., colleagues' conversations. The understanding of others' speech deeply affects the cognitive tasks of a worker. Thus, intelligibility represents the most delicate issue in offices and depends mainly on two factors: the acoustical properties of the space and the background noise. High reverberation and low signal-to-noise ratios drop the understanding of the speech. Thus, mechanical continuous noises, like HVAC or air change systems, increase the background noise disrupting intelligibility. It follows that the ability to separate the noise contribution of each kind of source inside offices is essential. The chapter is divided into three parts.

The first part investigates the proposed method in a small office with four workstations. The monitoring lasted for the entire day. The aim is to separate the contribution of the speeches and the mechanical noises, e.g., HVAC and computer fans. The analysis follows the same methodology of classroom applications, i.e., besides the Gaussian Mixture Model (GMM), the separation is carried out via the K-means clustering (KM) and the conventional praxis. Results show the ability of both GMM and KM to identify two different spectra. A preliminary spectral matching through standards proves the reliability of the reconstructions. Discussions about the statistical insights of the active sources explore the possible features of the clusters

obtained to label the sound sources more efficiently. Means and standard deviations for GMM, centroids and average intra-clusters distances for KM represent the basic couple of parameters to describe the nature of a sound source. Then, the coefficient of variation, i.e., the relative standard deviation, shows further insights to distinguish the different sources. Further remarks through STI evaluation matrices show how the intelligibility assessment changes taking into account each sound source separately.

The second part of the chapter applies the proposed method in three different offices. In these cases, the monitoring lasted for two days. Evaluations of the similarities of results between the two days in the same office show the repeatability of the method. Then, a study proposes a further parameter to measure the amount of collaboration among colleagues according to ISO 22955: the overlapping area. Preliminary outcomes seem to be consistent with the kinds of activities carried out in the offices and the ISO categories.

The third and last part of the chapter shows a comprehensive study of the validation of the unsupervised proposed method. Both machine and deep learning approaches are used to assess the sound context of the same office. The machine learning approach follows the same method of the previous studies reconstructing two different spectra: speech and traffic. The deep learning approach exploits the audio recording of the entire day as the database to train and test a variational autoencoder (VAE). The whole database has been segmented into frames 1 second long and manually labelled as speech or traffic. Thus, the semi-supervised analysis via the VAE provides the lack of labels of the unsupervised analysis via the GMM and KM. The comparison between the machine and the deep learning approaches is made through the latent space of the VAE. Here, it is possible to check how the encoder successfully separated the two clusters. In this case study, the encoder maps the input to a multivariate latent Gaussian distribution. Thus, the use of a Gaussian distribution to obtain the latent space of the VAE and recalling that the GMM can be considered as a generalization of KM, it is possible to validate the proposed method and state that the GMM seems to be the best algorithm to perform a source separation through a sound level meter.

Conclusions

Summary

In recent years, the increasing attention on ML applications strongly influenced many research fields, acoustics included. The ability to exploit data-driven methods has led to advances in manifold acoustics topics, like sound source separation and localization in free-field and reverberant environments, signal processing, underwater acoustics, and scene classification. Despite the limitations due to the need for large amounts of data, ML techniques provide encouraging results in many scenarios. The whole work of this thesis lies in this context and aims to propose a method to measure the noise contributions of different coexisting sound sources in real-world environments. The purpose is to provide an unsupervised training-free workflow to deepen the analysis of complex scenes.

Nowadays, the technical praxis handled by acousticians underlies strong assumptions to measure different portions of energy due to noise components. A statistical approach over long-term monitoring provides tools to analyze different acoustic contexts. A representative sample of case studies shows preliminary results of the applications of the proposed methods in two active environments: university lecture halls and offices.

An exhaustive summary of the main topics provides a suitable starting point to understand the context at the beginning of the manuscript. The motivations and objectives of the present work frame the current state-of-the-art and point out the lack that needs to be compensated. A comprehensive theoretical description, but limited to the focus of the work, gives the proper awareness about the mathematical and statistical properties of the tools used to build the method. Links and relationships between each metric and algorithm, either machine or deep learning, have particular attention. These represent a fundamental step allowing the comparison of the results and performances of each algorithm.

Details and motivations describe the reasons underlying each step of the proposed method, spanning from the data acquisition to the labelling step. The framework's

setting allows technicians and researchers to use the procedure in any scenario measurable through a sound level meter. The basic knowledge of the problem to address leads the operator to set the workflow properly.

The first applications of the proposed method are shown in university lecture halls. Here, the measurement of the student activity during lectures represents the main goal. The quantification of the chatting among students permits the evaluation of the extent to which students stay focused on teachers' speech. Thus, a statistical approach, rather than the conventional one, can bring further insights into the sound context. The cumulative distribution function of SPLs shows how the conventional energy-based approach works. It assesses the acoustic scene without considering how the distribution is shaped by the monitored activity. Percentiles and equivalent levels do not correspond to any feature of the statistical population. On the contrary, peaks and points of inflections of both probability and cumulative curves seem to provide a more consistent readability of the acoustic environment. In the specific case of university lecture halls, it has been shown how an objective measurement carried out through the proposed method describes a change in subjective behaviors by students. This result represents the ability of a statistical approach to conduct analyses between the objective – room criteria and energetic noise levels – and subjective – surveys and soundscape analyses – evaluations.

The detailed refinement of the method concerned the application of the procedure in offices. Unlike the learning context, the monitoring of active workplaces does not involve the measurement of a specific metric but the assessment of the entire acoustic scene. Thus, a model selection step ensures a preliminary overview of the number of type sources. Controlling each noise contribution allows technicians to evaluate the acoustic comfort in offices through intelligibility. Both the statistical features of each cluster and the basic knowledge of the problem permit the inference of the nature of the measured sound source. The broadening to other case studies pointed out the chance of creating additional metrics to assess the extent of collaborative work according to ISO 22955. However, this is possible only through a fuzzy algorithm. Thus, the overlapping areas between Gaussian curves represent a promising way to classify the dynamic cooperation among colleagues.

The deep learning analysis, exploiting the audio recording, represents an important tool to assess the assumption underlying the proposed method, especially in the labelling step. A latent representation of an entire working day obtained through a variational autoencoder confirms the ability of a Gaussian parametrization to recognize different kinds of sound sources in long-term monitoring. The underlying gaussianity among unsupervised algorithms and variational autoencoder validates the hypotheses partially confirmed by spectral matching.

The variational autoencoder performs a model-based deep clustering. Thus, it is possible to generate from the latent space, according to the parametrized model, the original measured samples and reconstruct the sound level meter. However, the latent representation is obtained through the dimensionality reduction of the inputs, i.e., loss of information. As a result, the decoder adds uncertainties to the reconstructions reproducing noisy spectrograms making the variational autoencoder incapable of being used for accurate measurements. The K-means clustering is a heuristic algorithm and tends to find clusters of comparable spatial extent. It can be compared to the Gaussian Mixture Model only if the homoscedasticity is fulfilled and performs a hard subdivision. As a result, K-means does not provide features such as the overlapping area. All these cons do not affect the Gaussian Mixture Model, which resulted to be the most flexible and appropriate algorithm to analyze sound level meter long-term monitoring. It provides features to univocally label the sound sources and the dynamics of the acoustical context. It is model-based, i.e., the measurements are described through statistical models. Moreover, using SPLs as input, no pre- or post-processing is required on the data.

Resulting remarks on sound level meter measurements

Noise descriptors used in the conventional approach are represented by the equivalent sound level L_{eq}, and statistical levels, i.e., percentiles. The latter indicate the sound pressure levels exceeded for a certain percentage of time indicated by their subscripts. The most used are L_{10}, L_{50}, and L_{90}. Since L_{eq} is defined in energy and statistical levels are numerical, their relationship is not straightforward. Moreover, the physical meaning of statistical levels is not always accurate, except for L_{90}, which usually referred to the background noise without the investigated source. However, some rules of thumb allow technicians to assess noise environments through these parameters. In little fluctuating contexts, L_{eq} will be close to L_{50}. On the contrary, with high fluctuations, L_{eq} will be closer to L_{10} and will exceed L_{90} by 10 or more dB. Generally, L_{eq} is between L_{50} and L_{10} with the latter about 3 dB above L_{eq}.

All these considerations about the sound environment are general and lack in details. A statistical approach would dig deeper into data describing the whole phenomenon that shaped the occurrences obtained by means of long-term monitoring. To visualize better the difference between the two approaches, 6 different cases of ideal mixtures with 2 components have been created and shown in Figure 4.15. All the plots show the means of each component – i.e., the corresponding sources' SPLs according to the statistical approach – the L_{eq} and the 10, 50, and 90

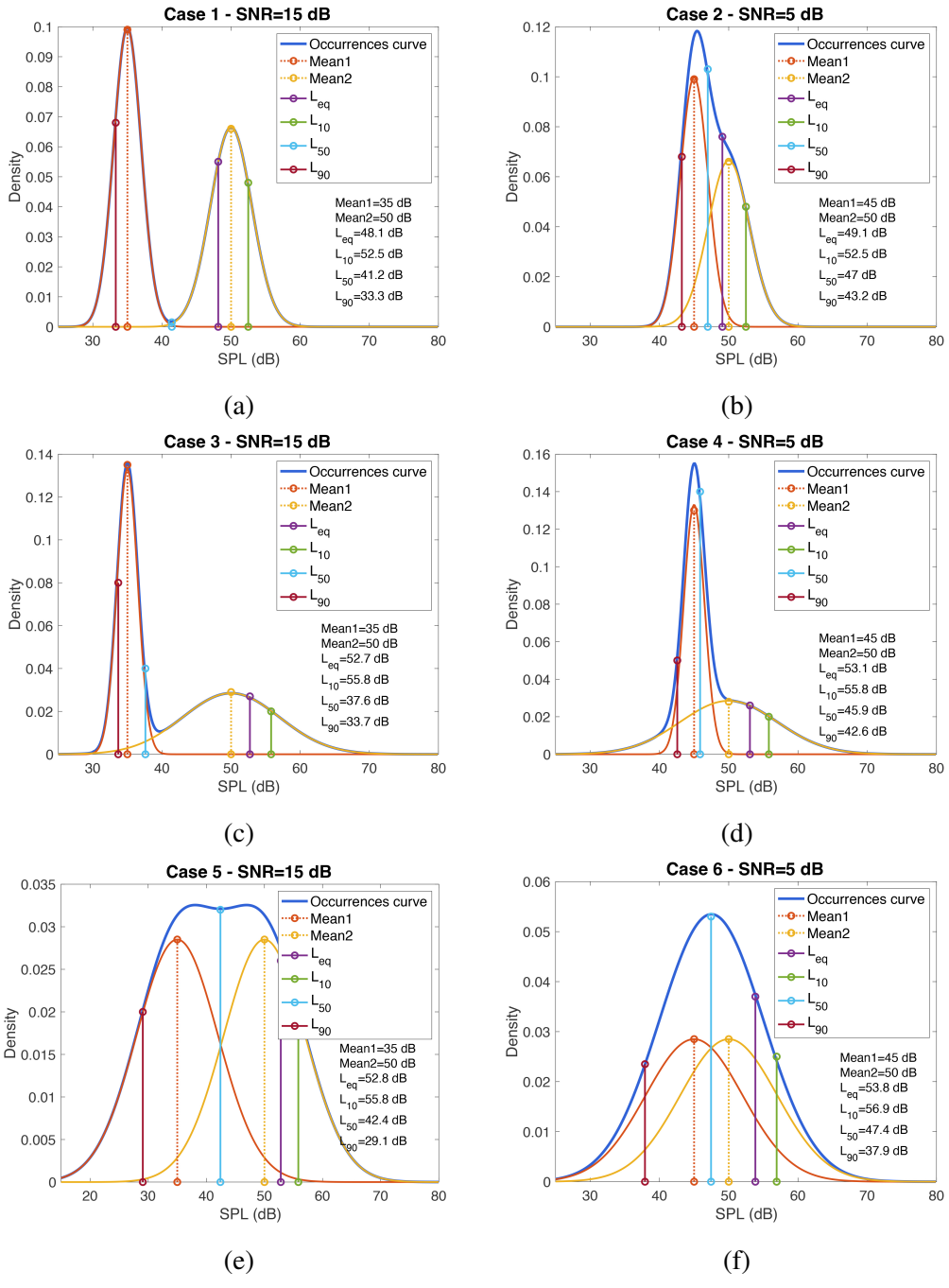


Fig. 4.15 Synthetic cases of Gaussian mixtures with different signal-to-noise ratios (SNR) and standard deviations. Means of each component, and the corresponding L_{eq}, besides the 10, 50, 90 statistical levels of each distribution are shown.

statistical levels of the corresponding synthetic distributions. The cases represent ideal situations with the mixing proportions of each Gaussian component equal to 0.5. Thus for simplification, the fluctuations of a source are considered only through the component's standard deviation.

Cases 1, 3, and 5 show a signal-to-noise ratio (SNR) between the two sources equal to 15 dB. Cases 2, 4, and 6 show an SNR equal to 5 dB. Narrow and large standard deviations (s.d.) have been used in different cases to simulate more steady and random sources, respectively. According to the literature and the case studies presented in this work, L_{eq} and L_{90} are assumed, in the conventional method, to have the same physical meaning as Mean2 and Mean1. In all cases, it is possible to notice how L_{eq} is more or less close to Mean2, i.e., the SPL of the highest sound source. L_{eq} is higher than Mean2 in all the cases except cases 1 and 2, where the s.d. of both sources are low. L_{90} is always lower than Mean1. Differences, as seen for L_{eq} and Mean2, are less noticeable when the s.d. is low. More than the SNR, the s.d. seems to affect the results, especially for L_{50}.

In summary, synthetic ideal distributions show how neither L_{eq} nor L_{90} is able to adequately measure the SPL of a sound source. L_{eq} and statistical levels result to being useful in describing the extent of noise fluctuations and depicting a general overview of the sound environment. However, they do not seem accurate enough to measure a sound source in a mixture. The combination of the conventional approach and the proposed one shows how few features would bring a lot of information to technicians to analyze a sound context.

Outlook and future work

At the end of the present dissertation, many open issues remain to be debated and investigated. The use of SPLs in the present work is strictly connected to the need of acousticians to measure each noise source in complex scenarios. Standards and requirements rely on clear thresholds, intervals, or single values to satisfy. Thus, the proposed method aims to obtain a source separation and the interpretation of real contexts only through SPLs. Thus, the phased data are lost but they are not used in standard measurements. As seen in the chapter concerning the application of the method in offices, this could mean preserve privacy. However, information is lost and it represents one of the limits of the method, indeed. Other drawbacks and limits concern the possible future scenarios of the acoustic monitoring. The use of short-term equivalent levels as input data of the method allows fast and efficient calculations preserving huge storage spaces with respect to audio data. Hence, a noise

monitoring station could implement in the software the computation of mixtures or wireless transfers very easily. At the same time, the lost of raw audio signals makes some possible applications, like the implementation in devices that exploit speech recognition or similar technologies, less versatile.

Another important step is the need to increase the number of case studies. A larger database of applications of the proposed method can provide further and more accurate insights into the physical meaning of the results. The outcomes presented here can constitute a preliminary analysis of the method's assessment. Once the method has been used in different environments for a number of case studies that is large enough, the statistical behavior of the results can be traced. The features represent the framework to assess the reliability of the procedure. Thus, the analysis of each characteristic useful to identify, separate, and label a sound source paves the way for the accurate estimation of the acoustical context.

First, the standard deviation ranges play a key role in understanding the randomness of a sound source. Cross-checks between s.d. in their spectral tendencies and several case studies could outline a robust way to label different kinds of sources. Monitoring over an extended period of time in active contexts gathers all interferences, masking effects, and other interactions between the sources' energy. Considering how complex the interactions of sound waves with the environment can be, the coefficient of variation constitutes a valuable feature to compare similar sources in different contexts. The relative standard deviation tendency could be the most accurate feature to verify similarities among different occurrence curves, especially for the tightest clusters, i.e., mechanical sources.

Second, the overlapping areas could constitute an objective metric to evaluate the dynamic of complex scenarios. In the present work, it has been introduced in office applications. However, it could become useful in each context where the sound sources have significant differences in temporal evolution.

Moreover, a quantitative validation experiment would provide preliminary insights about the influence of reverberation on the occurrences curve. Differences between long-term monitoring in anechoic and reverberant environments could point out to what extent SPLs vary in different contexts.

Lastly, the physical interpretation of the mean needs to be deepened. It is the most important feature since it represents the SPL of the sound source. Here lies the main difference between the conventional and the proposed approach. Sound power levels measured in controlled environments – reverberation or anechoic chambers – identify a sound source univocally. Measuring SPLs (sound pressure levels) means considering complex interactions between the emitted sound and the surrounding environment, even at the closest distances. The same source can emit different SPLs

in the free field or in enclosed spaces. Dealing with the measurement of sound sources in realistic and complex scenarios means associating a specific SPL to a single source. Even in a long-term measurement of a sound source much higher than the background noise, the L_{eq} would not describe that source accurately. The conventional approach is strongly influenced by the highest level. Thus, each outlier, i.e., high SPLs, would affect the result. The proposed approach avoids the influence of outliers because only the most probable SPL is considered to be the representative metric of the source.

The present work aims to provide an additional method to acousticians for measuring sound sources in complex acoustic scenarios. The ability to identify, separate, and label different coexisting sound sources would bring more accurate diagnoses and design proposals improving the acoustic comfort in manifold contexts.

References

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

[2] Anand, S., Bottalico, P., and Gray, C. (2019). Vocal fatigue in prospective vocal professionals. *Journal of Voice*, page in press.

[3] ANSI 3.5 (1997). Methods for the calculation of the speech intelligibility index (SII).

[4] Astolfi, A., Parati, L., D'Orazio, D., and Garai, M. (2019). *The new Italian standard UNI 11532 on acoustics for schools*. Universitätsbibliothek der RWTH Aachen.

[5] Astolfi, A. and Pellerey, F. (2008). Subjective and objective assessment of acoustical and overall environmental quality in secondary school classrooms. *The Journal of the Acoustical Society of America*, 123(1):163–173.

[6] Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

[7] 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.

[8] 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.

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

[10] Bistafa, S. R. and Bradley, J. S. (2000). Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. *The Journal of the Acoustical Society of America*, 107(2):861–875.

[11] Bistafa, S. R. and Bradley, J. S. (2001). Predicting speech metrics in a simulated classroom with varied sound absorption. *The Journal of the Acoustical Society of America*, 109(4):1474–1482.

[12] Bottalico, P. (2018). Lombard effect, ambient noise, and willingness to spend time and money in a restaurant. *The Journal of the Acoustical Society of America*, 144(3):EL209–EL214.

[13] Bottalico, P. and Astolfi, A. (2012). Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms. *The Journal of the Acoustical Society of America*, 131(4):2817–2827.

[14] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

[15] Braat-Eggen, E., vd Poll, M. K., Hornikx, M., and Kohlrausch, A. (2019). Auditory distraction in open-plan study environments: Effects of background speech and reverberation time on a collaboration task. *Applied Acoustics*, 154:148–160.

[16] Bradley, J. (1985). *Uniform derivation of optimum conditions for speech in rooms (BRN 239)*.

[17] Bradley, J. S. (1986a). Predictors of speech intelligibility in rooms. *The Journal of the Acoustical Society of America*, 80(3):837–845.

[18] Bradley, J. S. (1986b). Speech intelligibility studies in classrooms. *The Journal of the Acoustical Society of America*, 80(3):846–854.

[19] 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.

[20] Brunskog, J., Gade, A. C., Bellester, G. P., and Calbo, L. R. (2009). Increase in voice level and speaker comfort in lecture rooms. *The Journal of the Acoustical Society of America*, 125(4):2072–2082.

[21] Building Bulletin 93 (2003). Acoustic design of schools - A design guide (BB 93).

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

[23] Calosso, G., Puglisi, G. E., Astolfi, A., Castellana, A., Carullo, A., and Pellerey, F. (2017). A one-school year longitudinal study of secondary school teachers' voice parameters and the influence of classroom acoustics. *The Journal of the Acoustical Society of America*, 142(2):1055–1066.

[24] 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.

[25] Choi, Y.-J. (2016). Effect of occupancy on acoustical conditions in university classrooms. *Applied Acoustics*, 114:36–43.

[26] Choi, Y.-J. (2018). Effects of the distribution of occupants in partially occupied classrooms. *Applied Acoustics*, 140:1–12.

[27] Choi, Y.-J. (2020a). Evaluation of acoustical conditions for speech communication in active university classrooms. *Applied Acoustics*, 159:107089.

[28] Choi, Y.-J. (2020b). The intelligibility of speech in university classrooms during lectures. *Applied Acoustics*, 162:107211.

[29] Christensen, C. (2011). Odeon room acoustics program ver. 11. *User Manual: Industrial, Auditorium and Combined Editions,(Odeon A/S, Lyngby, 2011)*.

[30] Connor, M., Canal, G., and Rozell, C. (2021). Variational autoencoder with learned latent structure. In *International Conference on Artificial Intelligence and Statistics*, pages 2359–2367. PMLR.

[31] Cox, T. and d'Antonio, P. (2016). *Acoustic absorbers and diffusers: theory, design and application*. Crc Press.

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

[33] De Salvio, D., Bianco, M. J., Gerstoft, P., D'Orazio, D., and Garai, M. (2023a). 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.

[34] 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.

[35] De Salvio, D. and D'Orazio, D. (2022). Effectiveness of acoustic treatments and pa redesign by means of student activity and speech levels. *Applied Acoustics*, 194:108783.

[36] De Salvio, D., Fratoni, G., D'Orazio, D., and Garai, M. (2023b). Assessing human activity noise in workspaces using machine learning and numerical models. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 265, pages 5259–5269. Institute of Noise Control Engineering.

[37] 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*, volume 253, pages 4117–4126. Institute of Noise Control Engineering.

[38] 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.

[39] Deutsche Institut für Normung (2016). Hörsamkeit in räumen—anforderungen empfehlungen und hinweise für die planung (DIN 18041).

[40] Di Blasio, S., Shtrepi, L., Puglisi, G. E., and Astolfi, A. (2019). A cross-sectional survey on the impact of irrelevant speech noise on annoyance, mental health and well-being, performance and occupants' behavior in shared and open-plan offices. *International journal of environmental research and public health*, 16(2):280.

[41] D'Orazio, D., De Cesaris, S., Guidorzi, P., Barbaresi, L., Garai, M., and Magalotti, R. (2016). Room acoustic measurements using a high spl dodecahedron. In *Audio Engineering Society Convention 140*. Audio Engineering Society.

[42] Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

[43] 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.

[44] D'Orazio, D., Rossi, E., and Garai, M. (2018). Comparison of different in situ measurements techniques of intelligibility in an open-plan office. *Building Acoustics*, 25(2):111–122.

[45] Egli, M., Roper, T., Feurer, I., and Thompson, T. (1999). Architectural acoustics in residences for adults with mental retardation and its reltion to perceived homelikeness. *American journal on mental retardation*, 104(1):53–66.

[46] 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.

[47] Ellermeier, W. and Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology*, 35(1):10–16.

[48] 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).

[49] Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

[50] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

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

[52] 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.

[53] Harold, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321.

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

[55] Harvie-Clark, J., Larrieu, F., and Opsanger, C. (2019). Iso 3382-3: Necessary but not sufficient. a new approach to acoustic design for activity-based-working offices. *Proceedings of the 23rd International Congress on Acoustics*.

[56] Hasegawa, Y. and Ryherd, E. (2020). Clustering acoustical measurement data in pediatric hospital units. *The Journal of the Acoustical Society of America*, 148(1):265–277.

[57] 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)*, pages 31–35. IEEE.

[58] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[59] Hodgson, M. (2002). Rating, ranking, and understanding acoustical quality in university classrooms. *The Journal of the Acoustical Society of America*, 112(2):568–575.

[60] Hodgson, M. (2004). Case-study evaluations of the acoustical designs of renovated university classrooms. *Applied Acoustics*, 65(1):69–89.

[61] Hodgson, M. and Nosal, E.-M. (2002). Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms. *The Journal of the Acoustical Society of America*, 111(2):931–939.

[62] 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.

[63] Hodgson, M., Steininger, G., and Razavi, Z. (2007). Measurement and prediction of speech and noise levels and the lombard effect in eating establishments. *The Journal of the Acoustical Society of America*, 121(4):2023–2033.

[64] Hongisto, V. (2005). A model predicting the effect of speech of varying intelligibility on work performance. *Indoor air*, 15(6):458–468.

[65] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[66] Houtgast, T. (1981). The effect of ambient noise on speech intelligibility in classrooms. *Applied Acoustics*, 14(1):15–25.

[67] Houtgast, T. and Steeneken, H. J. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica united with Acustica*, 28(1):66–73.

[68] Iannace, G., Ciaburro, G., and Trematerra, A. (2018). Heating, ventilation, and air conditioning (hvac) noise detection in open-plan offices using recursive partitioning. *Buildings*, 8(12):169.

[69] IEC 60268 (2020). Sound System Equipment – part 16: Objective rating of speech intelligibility by speech transmission index. (International Electrotechnical Commission, Geneva, Switzerland).

[70] 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).

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

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

[73] ISO 3382 - 3 (2012). Acoustics - Measurement of room acoustic parameters – part 3: Open-plan offices. (International Organization for Standardization, Geneva, Switzerland).

[74] ISO 3741 (2010). Acoustics – Determination of sound power levels and sound energy levels of noise sources using sound pressure – Precision methods for reverberation test rooms. (International Organization for Standardization, Geneva, Switzerland).

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

[76] 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.

[77] Jones, D. (1999). The cognitive psychology of auditory distraction: The 1997 bps broadbent lecture. *British Journal of Psychology*, 90(2):167–187.

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

[79] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

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

[81] Klatte, M., Hellbrück, J., Seidel, J., and Leistner, P. (2010). Effects of classroom acoustics on performance and well-being in elementary school children: A field study. *Environment and Behavior*, 42(5):659–692.

[82] 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*, pages 673–675.

[83] Larsen, J. B. and Blair, J. C. (2008). The effect of classroom amplification on the signal-to-noise ratio in classrooms while class is in session. *Language, Speech, and Hearing Services in Schools*, 39(4):451–460.

[84] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[85] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[86] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[87] 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.

[88] 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)*, pages 101–105. IEEE.

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

[90] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

[91] Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Mal. de L'Oreille et du Larynx*, pages 101–119.

[92] Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Interspeech*, volume 2013, pages 436–440.

[93] Lundquist, P., Holmberg, K., Landstrom, U., et al. (2000). Annoyance and effects on work from environmental noise at school. *Noise and Health*, 2(8):39.

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

[95] Markides, A. (1986). Speech levels and speech-to-noise ratios. *British Journal of Audiology*, 20(2):115–120.

[96] McCulloch, W. S. (1943). A logical calculus of ideas imminent in nervous activity. *Biol Math. Biophys*.

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

[98] McSporran, E., Butterworth, Y., and Rowson, V. J. (1997). Sound field amplification and listening behaviour in the classroom. *British Educational Research Journal*, 23(1):81–96.

[99] 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.

[100] 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.

[101] 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.

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

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

[104] Nijs, L., Van Berlo, D., and de Vries, D. (2001). The development of architectural guidelines for the acoustical quality in rooms for mentally challenged people. *Int Congr. Acoustics*.

[105] Nowakowska, E., Koronacki, J., and Lipovetsky, S. (2014). Tractable measure of component overlap for gaussian mixture models. *arXiv preprint arXiv:1407.7172*.

[106] Oberdorster, M. and Tiesler, G. (2006). Acoustic ergonomics of school: Research report fb 1071. *Federal Institute for Occupational Safety and Health, Dortmund, Germany*.

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

[108] 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.

[109] 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.

[110] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

[111] Peng, J., Lau, S.-K., and Zhao, Y. (2020). Comparative study of acoustical indices and speech perception of students in two primary school classrooms with an acoustical treatment. *Applied Acoustics*, 164:107297.

[112] Peng, J., Zhang, H., and Wang, D. (2018). Measurement and analysis of teaching and background noise level in classrooms of chinese elementary schools. *Applied Acoustics*, 131:1–4.

[113] Perrin Jegen, N. and Chevret, P. (2017). Effect of noise on comfort in open-plan offices: application of an assessment questionnaire. *Ergonomics*, 60(1):6–17.

[114] Peter, L. and Anthony, C. (2019). The lombard effect in open plan offices. In *Proceedings of the Institute of Acoustics*, volume 41, pages 216–226. Institute of Acoustics.

[115] Prodi, N., Visentin, C., Borella, E., Mammarella, I., and Di Domenico, A. (2021). Using speech comprehension to qualify communication in classrooms: Influence of listening condition, task complexity and students' age and linguistic abilities. *Applied Acoustics*, 182:108239.

[116] Puglisi, G. E., Warzybok, A., Astolfi, A., and Kollmeier, B. (2021). Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios. *Building and Environment*, page 108137.

[117] Rasmussen, B. and Carrascal García, T. (2019). Acoustic regulations for offices-comparison between selected countries in europe. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 259, pages 8141–8150. Institute of Noise Control Engineering.

[118] Reich, R. and Bradley, J. (1998). Optimizing classroom acoustics using computer model studies. *Canadian Acoustics*, 26:15–21.

[119] Renz, T., Leistner, P., and Liebl, A. (2018a). Auditory distraction by speech: Can a babble masker restore working memory performance and subjective perception to baseline? *Applied Acoustics*, 137:151–160.

[120] Renz, T., Leistner, P., and Liebl, A. (2018b). Auditory distraction by speech: Sound masking with speech-shaped stationary noise outperforms- 5 db per octave shaped noise. *The Journal of the Acoustical Society of America*, 143(3):EL212–EL217.

[121] Renz, T., Leistner, P., and Liebl, A. (2019). Use of energy-equivalent sound pressure levels and percentile level differences to assess the impact of speech on cognitive performance and annoyance perception. *Applied Acoustics*, 153:71–77.

[122] Rindel, J. H. (2000). The use of computer modeling in room acoustics. *Journal of vibroengineering*, 3(4):219–224.

[123] Rindel, J. H. (2018). Open plan office acoustics–a multidimensional optimization problem. *Proceedings of DAGA2018, Munich, Deutsche Gesellschaft für Akustik*.

[124] 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*, volume 2012, pages 954–966. Institute of Noise Control Engineering.

[125] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

[126] 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.

[127] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[128] Sato, H. and Bradley, J. S. (2008). Evaluation of acoustical conditions for speech communication in working elementary school classrooms. *The Journal of the Acoustical Society of America*, 123(4):2064–2077.

[129] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681.

[130] Schlittmeier, S. J. and Liebl, A. (2015). The effects of intelligible irrelevant background speech in offices–cognitive disturbance, annoyance, and solutions. *Facilities*.

[131] Selim, S. Z. and Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, (1):81–87.

[132] Shield, B., Conetta, R., Dockrell, J., Connolly, D., Cox, T., and Mydlarz, C. (2015). A survey of acoustic conditions and noise levels in secondary school classrooms in england. *The Journal of the Acoustical Society of America*, 137(1):177–188.

[133] Shield, B. and Dockrell, J. E. (2004). External and internal noise surveys of london primary schools. *The Journal of the Acoustical Society of America*, 115(2):730–738.

[134] Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

[135] Steeneken, H. J. M. and Houtgast, T. (1982). Some applications of the speech transmission index (sti) in auditoria. *Acta Acustica united with Acustica*, 51(4):229–234.

[136] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928.

[137] Sun, H. and Wang, S. (2011). Measuring the component overlapping in the gaussian mixture model. *Data mining and knowledge discovery*, 23(3):479–502.

[138] 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.

[139] UNI (2020). Caratteristiche acustiche interne di ambienti confinati – metodi di progettazione e tecniche di valutazione – parte 2: Settore scolastico (UNI 11532-2). Technical report, Ente Nazionale Italiano di Unificazione.

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

[141] Vellenga, S., Bouwhuis, T., and Höngens, T. (2017). Proposed method for measuring 'liveliness' in open plan offices. In *Proceedings of the 24th International Congress on Sound and Vibration, London, UK*, pages 23–27.

[142] Vér, I. L. and Beranek, L. L. (2005). *Noise and vibration control engineering: principles and applications*. John Wiley & Sons.

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

[144] Visentin, C., Prodi, N., Cappelletti, F., Torresin, S., and Gasparella, A. (2018). Using listening effort assessment in the acoustical design of rooms for speech. *Building and Environment*, 136:38–53.

[145] Vorländer, M. and Summers, J. E. (2008). Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and acoustic virtual reality. *Acoustical Society of America Journal*, 123(6):4028.

[146] Wang, D. and Narayanan, S. S. (2007). Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2190–2201.

[147] 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.

[148] Whitlock, J. and Dodd, G. (2006). Classroom acoustics—controlling the cafe effect... is the lombard effect the key. *Proceedings of ACOUSTICS, Christchurch, New Zealand*, pages 20–22.

[149] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

[150] Wooldridge, M. (2020). *The road to conscious machines: The story of AI*. Penguin UK.

[151] Yadav, M. and Cabrera, D. (2019). Two simultaneous talkers distract more than one in simulated multi-talker environments, regardless of overall sound levels typical of open-plan offices. *Applied Acoustics*, 148:46–54.

[152] 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.

[153] Yadav, M., Kim, J., Cabrera, D., and De Dear, R. (2017). Auditory distraction in open-plan office environments: The effect of multi-talker acoustics. *Applied Acoustics*, 126:68–80.