

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Measuring the speech level and the student activity in lecture halls: Visual- vs blind-segmentation methods

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

Published Version:

D'Orazio, D., De Salvio, D., Anderlucci, L., Garai, M. (2020). Measuring the speech level and the student activity in lecture halls: Visual- vs blind-segmentation methods. *APPLIED ACOUSTICS*, 169, 1-8 [10.1016/j.apacoust.2020.107448].

Availability:

This version is available at: <https://hdl.handle.net/11585/763146> since: 2020-09-15

Published:

DOI: <http://doi.org/10.1016/j.apacoust.2020.107448>

Terms of use:

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

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

(Article begins on next page)

Measuring the speech level and the student activity in lecture halls: visual- Vs blind-segmentation methods

Dario D'Orazio^{a,*}, Domenico De Salvio^a, Laura Anderlucci^b, Massimo Garai^a

^a*Department of Industrial Engineering, University of Bologna, Bologna, Italy*

^b*Department of Statistical Sciences "Paolo Fortunati" University of Bologna, Bologna, Italy*

Abstract

The background noise has a fundamental role in oral communication, since the higher the speech level with respect to the background noise the greater the intelligibility. In occupied lecture halls the main contribution to background noise is related to the human noise, which is called by scholars student activity. Scholars proposed methods to measure both student activity and speech level through short-time sound level meter measurements during lessons. However, a comparison of their relative effectiveness on a relevant set of data in different situations is still lacking. In this study, basing on recordings of university lessons performed with public address system, student activity and speech level values were extracted using different methods. Various scenarios of university lectures were recorded: frontal lessons, media-aided lectures, open discussions. Visual-segmentation and blind-segmentation procedures were compared for each case. Results show the benefits of blind-segmentation methods, which seem to be reliable and affordable methods for this kind of analyses.

1. Introduction

Intelligibility can be measured as the degree of preservation of the vocal information carried by modulation frequencies [1]. A high reverberation of the room and a low value of signal-to-noise ratio (SNR) are detrimental in the preservation of the vocal information [2, 3]. The acoustic characteristics of the room can also influence the speaker behaviour; in fact, with a high reverberation time a lecturer reads or speaks slower than in *dead rooms* [4]. Objective parameters linked to the room acoustics criteria can describe the change of the sound power emitted by the teachers in their speeches [5]. In classroom acoustics, the SNR is defined as the difference between the speech level (SL) – which is the level of the speech signal at the listener position – and the background noise. In occupied conditions, during the lecture, the background noise may be due to various contributions: the systems equipment noise (HVAC systems), the external traffic and the human noise, called student activity (SA). In university classrooms, this latter factor – the human noise – is generally higher than the other two, due to the large number of listeners. It was proved that the human noise depends on the occupancy degree and its value may increase in presence of PA [6, 7].

The speech structure is a quasi-continuous signal with short breaks due to the division of the single words or sentences and longer breaks due to the teacher writing on the blackboard [8]. During the pauses the main sound source is the background noise. However, a long break leads to an increase of the student noise and thus, for an accurate evaluation it is more interesting to study the short breaks between sentences. Taking advantage of these breaks, several methods have been proposed in the literature in order to distinguish the noise sources during lectures.

Early approaches analysed manually the short-time sound level recordings [9, 10]. The acquisition of time histories allowed an easier measure of both SA and SL during the lessons. The equivalent sound pressure level $L_{A,eq}$ and the 90-th percentile $L_{A,90}$ can be assumed as the SL value and the SA value, respectively [11, 8].

*Corresponding author

Basing on the assumption of Gaussian distribution of the statistical occurrences of the short-time levels for both SA and SL, if the recordings were long enough, Hodgson used a Peak-detection algorithm to extract SA and SL values from time histories of 15min-long recordings [12]. This approach was used in various contexts such as in auditoria to evaluate the noise due to the audience during musical performances or in the speaker recognition [13, 14].

More recently, improvements were obtained with blind techniques. Because this kind of measurements does not need an operator, it could be used in permanent setups of data acquisition. *K*-means clustering was used by Brill et al., in order to identify the lecture time in classrooms monitoring [15].

SA and SL are not independent values but they can be correlated by the *Lombard effect* [16] and the *Cafè effect* [17], which may affect both the speaker and the students [18]. SA is related to the language as well: SA increases in case of non-native listeners, so a higher SNR value is recommended in this case [19].

Furthermore, the PA system may influence the SA values. It is a common solution that is surely useful in practice but that still lacks the enhancing of the listening environment for the students [17]. Other studies did not find a high correlation between their results and the *Lombard effect* when the lecturers used a sound reinforcement system, differently from the case when the PA system is not used [18]. Many studies did not take into account the PA system in measured classrooms. They use a direct observation of the sound phenomena during the recording of lessons instead of statistical methods, so they can associate each peak of speech signal to a source [20]. Beside this, the high occupation keeps the reverberation time low, due to the high acoustic absorption of students. Thus, it can be assumed that the intelligibility, in occupied university halls, can depend on SNR only.

In the present work, active lecture halls were recorded. The aim is comparing four techniques to detect the students activity levels and the teachers' speech level, investigating their strong and weak points. In Section 2 a theoretical overview of each method is provided, while in Section 3 the method is described; the obtained results are shown and discussed in Sections 4 and 5, respectively.

2. Theoretical background

2.1. Visual segmentation: Percentile levels and peak detection

Early methods extract speech and student activity levels directly from sound level meter. SA and SL are extracted from time windows [10, 20], both the equivalent and percentile levels are extracted from whole recordings [8, 11]. The difference of the two levels, respectively SL and SA, is considered as an estimation of SNR basing on the percentage time of teacher's speech [21], where the $L_{A,eq}$ is assumed as the SL of the teacher and the $L_{A,90}$ as the SA due to the students (see Figure 1).

The peak detection (PD) technique is based on the assumption of Gaussian distribution of occurrences over 15 minutes of data-collecting. Depending on noise conditions, the statistical distribution of occurrences of sound levels may be fitted with two or more Gaussian curves, allowing to distinguish the noise sources and their levels [12]. In case of HVAC switched off and negligible traffic noise, the number of Gaussian curves may be fixed as two. Multi-peak analysis and curve fitting can return a significant Gaussian regression of the data. An asymmetrical curve, the occurrence density, can be fitted into two symmetrical normal-distribution curves with the maximum values in the neighbourhood of the measured peaks.

Percentile levels and peak detection techniques are linked. The same dataset can be seen in two different ways: through the cumulative distribution or through the occurrence curve (see respectively Fig 1 and Fig 2). In fact, evaluating a percentile level means doing a backward integration of the occurrence curve (see Figure ??) covering the percentage of its area until the percentile required. For example, determining the acoustic percentile level L_{90} corresponds to the backward integral until the 10th percentile of the occurrence curve.

In statistics, the rank r of a percentile q of N observations is defined as:

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

Consequently, the value of the *acoustic* percentile level L_q of a certain dataset is equal to the rank r of $100-q$. For a large number of observations, if the density occurrence is represented by $f(x)$, the value q can

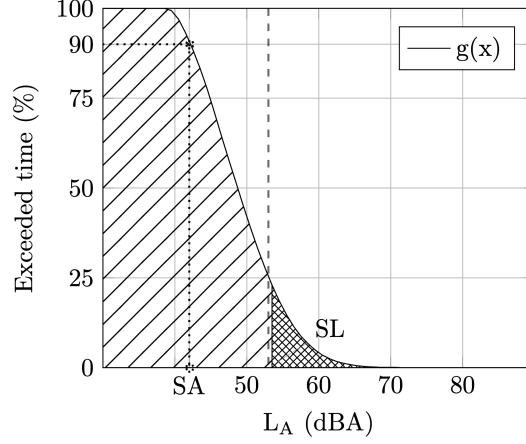


Figure 1: Percentile levels technique: the continuous line represents the cumulative distribution of the SPL values recorded by a sound level meter. The example highlights the L_{90} percentile level, named as SA, with the dotted line and the equivalent level $L_{A,eq}$, named as SL, with the dashed line.

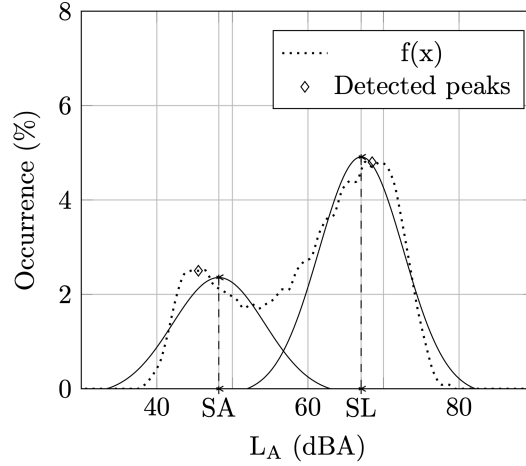


Figure 2: Peak detection technique: the dotted line represents the distribution of the occurrences $f(x)$. This density function was fitted with two Gaussian curves after a peak detection (peaks are indicated with \diamond). The means of Gaussian curves (*) correspond to two different sound sources levels.

be expressed as:

$$q = P(x > L_q) = \int_{r(100-q)}^{\infty} f(x) dx. \quad (2)$$

Taking the complement to 1 of the cumulative function:

$$g(x) = \int_{\infty}^x f(\xi)(-d\xi) = \int_x^{\infty} f(\xi) d\xi = 1 - \int_{-\infty}^x f(\xi) d\xi. \quad (3)$$

from the properties of the probability density function. This approach was used by Hodgson et al. [12] to identify the student activity during lessons.

2.2. Blind segmentation

Running iterative algorithms allows to arrange the recorded values avoiding the visual detection needed by the methods seen in the previous section. The following statistical techniques are used in unsupervised

machine learning to detect patterns and correlations among data as well [22].

2.2.1. Model-based clustering: Gaussian Mixture Models (GMM)

Model-based clustering assumes that the data are generated by a probabilistic model and tries to recover the original model from the data. The model estimated from the data via the Maximum Likelihood (ML) method then defines clusters and assigns points to the mixture component with the highest a posteriori probability of belonging to; in fact, each component probability distribution describes the shape of a cluster. In this context, like in the peak detection technique, Gaussian probability distributions have been assumed for both SA and SL [18].

Let X be a set of independent observations, x_1, \dots, x_n , drawn from a mixture of Gaussian distributions; the density $p(x_i)$ can be written in the form

$$p(x_i; \psi) = \sum_{k=1}^K \pi_k \phi(x_i; \theta_k) \quad i = 1, \dots, n, \quad \psi = \{\theta, \pi\} \quad (4)$$

where the $\phi(x_i, \theta)$ s are the Gaussian densities with parameter vector $\theta_k = \{\mu_k, \sigma_k^2; k = 1, \dots, K\}$ and π_k are the so called *mixing proportions*, non-negative quantities that sum to one; that is, $0 \leq \pi_k \leq 1 \quad (k = 1, \dots, K)$ and $\sum_{k=1}^K \pi_k = 1$ [23]. The likelihood function for a mixture model with K univariate Normal components is:

$$\mathcal{L}(\psi|x) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \phi(x_i|\theta_k) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}}. \quad (5)$$

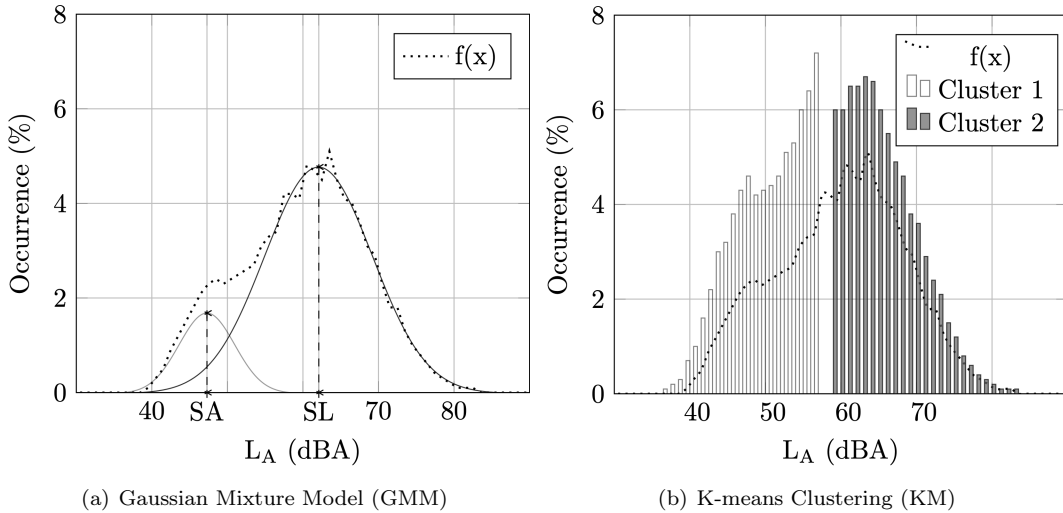


Figure 3: Gaussian Mixture Model technique (left figure): the dotted line represents the distribution of the occurrences of the recorded SPL while the continuous lines represent the Gaussian curves found. K -means clustering technique (right figure): the continuous line represents the distribution of the occurrences. Two clusters of the recorded dataset are highlighted with bars.

Figure 3(a) shows an example of the GMM method where the recorded SPL values are separated in two different Gaussian curves.

2.2.2. Distance-based clustering: K -means algorithm (KM)

Distance-based clustering algorithms manage data by optimizing a distance metric. K -means algorithm finds a partition such that the squared Euclidean distance between the empirical mean of a cluster and the points in such cluster is minimized.

Let $X = \{x_i\}$, $i=1, \dots, n$ be the set of n points to be clustered into a set of K clusters, $C = \{c_k; k = 1, \dots, K\}$. Let μ_k be the mean of cluster c_k . The squared Euclidean distances between μ_k and the points in cluster c_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2. \quad (6)$$

The goal of K -means is to minimize the sum of the squared Euclidean distances over all K clusters; therefore, the objective function to be minimized is the following:

$$J(C) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2. \quad (7)$$

Minimizing this function is known to be an NP-hard problem, hence K -means, which is a greedy algorithm [24], can only converge to a local minimum. A classical estimation algorithm for minimizing $J(C)$ consists of two steps sequentially iterated until convergence [25]. In the first step, for fixed μ_k the best partition C is found by assigning each point to the nearest cluster center. Then in the second step, for fixed C , the centroids μ_k ($k = 1, \dots, K$) are computed. To the best of authors' knowledge in this field, K -means have only been used to recognize the occupied or unoccupied state of a classroom [15]. In this work it is used to identify the speech level of the teacher and the student activity as well. An example is given in Figure 3(b) where the occurrences of the recorded short-time A-weighted equivalent levels are divided in two clusters.

3. Method

Table 1: General data of the lecture halls and ISO 3382-1 measurements results [31], 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 subscripts “M” indicate a value averaged over all the receivers in the octave bands of $500 \div 1000$ Hz.

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

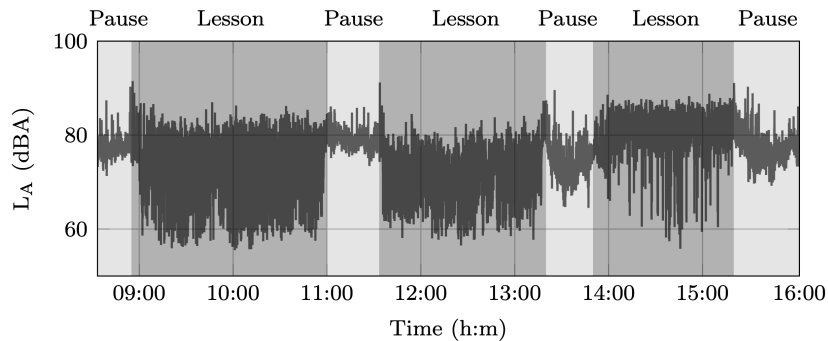


Figure 4: Example of a temporal history extracted from a measurement campaign. The two shades of grey show the parts of the temporal history meant as pause (light grey) or lesson (dark grey).

3.1. General method

Short-time sound levels values (with an integration time of 100 ms) were recorded during lessons in university lecture halls. Lessons were done in the same building, the Faculty of Humanities and Philosophy of the University of Bologna, in three large lecture halls. Students and teachers varied for each lesson. Data were acquired during the entire lesson activity, collecting several hours of measurements, and then post-processed. Basing on these data, SA and SL values were extracted using the four methods outlined in Section 2: percentile levels (PL), peak detection (PD), Gaussian mixture model (GMM) and K -means clustering (KM). PL were extracted using the post-processing commercial software 01dB dBTrait [26]. Supervised PD was made through *OriginPro* software [27]. For what concerns GMM, the maximum likelihood estimates of the parameters was derived via the EM algorithm [28], by using the `Mclust` function from the homonym R package [29]. The algorithm to perform KM was the `kmeans` function of the `stats` R library [30].

3.2. In situ measurements

The lecture rooms were chosen for their high occupancy and the variability of the lessons given in them. Despite the fact that these halls were designed specifically for this use, an acoustic discomfort was complained by lecturers and students because of the excessive reverberation. Hall I and Hall II are historical rooms with an amphitheater geometry; they have plastered walls and wooden seats and benches. Hall III has a quite regular shoe-box shape, 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.

A preliminary measurement campaign was carried out in the halls under study aimed at qualifying their room criteria in an unoccupied state, using procedures and equipment according to ISO 3382 [31] standard. Monaural impulses responses were acquired with an ESS signal, length 512 K and sampled at 48 kHz, using a high-SPL dodecahedron as omnidirectional sound source [32]. The variable occupancy by the students influence the total absorption area in the halls [33]. Based on values measured in unoccupied condition, the reverberation times in occupied condition were evaluated using the equation [34]:

$$T_{occ} = \frac{T_{unocc}}{1 + \frac{T_{unocc}CN\Delta A_{1p}}{0.16V}} \quad (\text{s}) \quad (8)$$

where N is the maximum occupancy of the hall and C is the percentage of occupancy ($C=1$ means a full occupied hall, $C=0.8$ an occupancy of 80%). ΔA_{1p} is the increase of the equivalent absorption area due to one person in m^2 . Its values are taken from the datasets of German acoustic regulation for classrooms [34]. The main geometrical data and the measured reverberation times in unoccupied condition are reported in Table 1.

3.3. Measurement setup

Two sound level meters were placed in the middle of the audience area in each hall, on two different sides, at a height of 1.2 m, maintaining a distance of at least 1 m from any surrounding surfaces.

An operator attended the recordings during lessons inside the halls. He reported the activities done during the lesson to analyse potential peculiarities in the recorded sample data. He took note of any unexpected sound phenomena also, so as to delete the corresponding peaks in post-processing analysis. The whole lessons were analysed removing such intervals from the time series and focusing on the lesson activity (see Figure 4). Supervising the recorded lessons helps to understand the differences between the considered methods. The size of the lecture halls and the PA support make the movements of the teachers insignificant, as the SL source position is always the same and does not affect the recordings.

4. Results

Twelve lessons of about 90 minutes each were recorded. All of them have a similar number of samples (52000 on average). About half of the lecturers were male and the other half female. For each statistical population (i.e. a single lesson), SA and SL values were extracted using the four techniques above mentioned

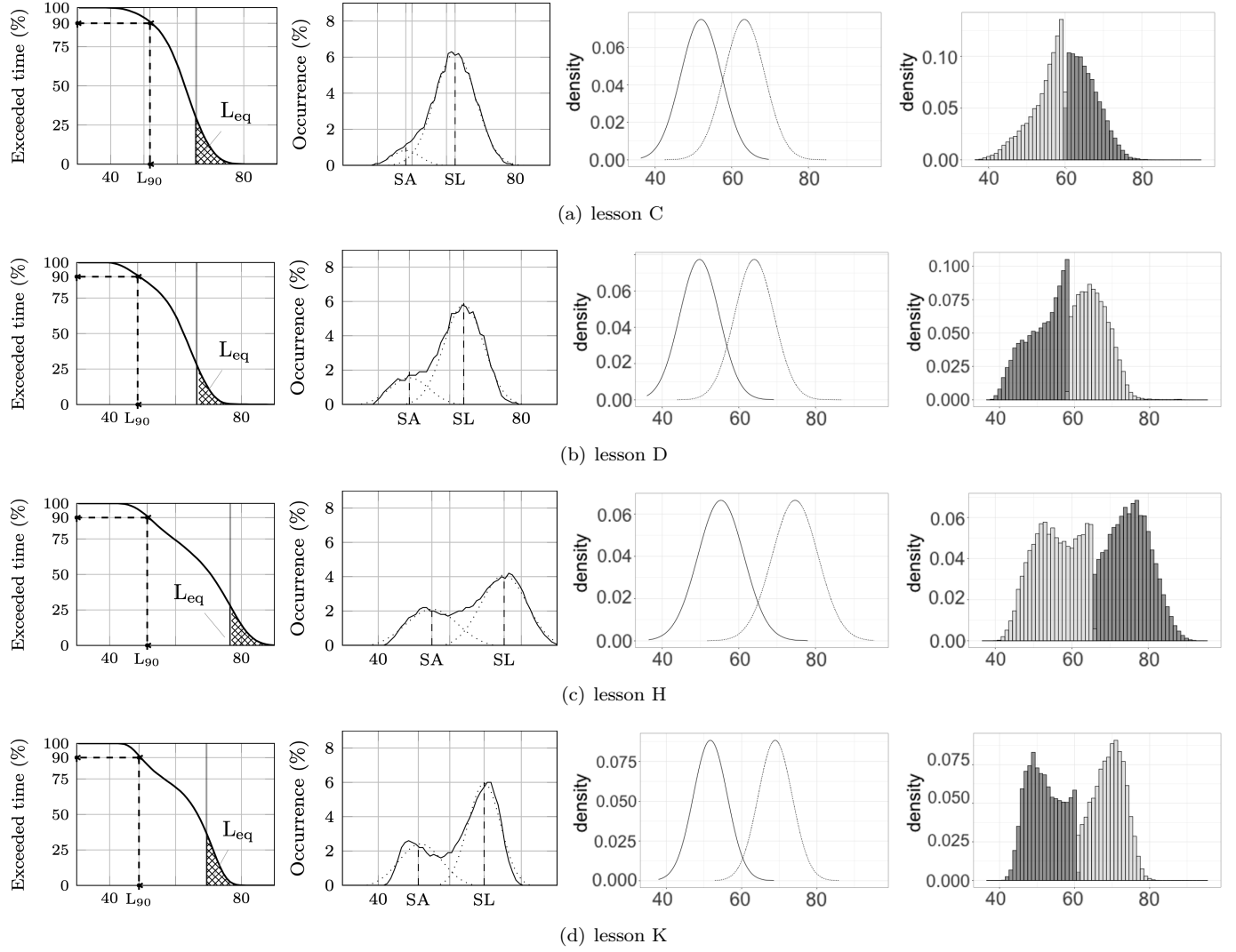


Figure 5: Graphs of the statistical analysis of lessons C, D, H and K for each technique. From left to right: PL, PD, GMM and KM. The x-axes values, for each graph, are in dBA.

Table 2: 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.

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
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
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
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
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
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
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
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
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
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
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
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
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
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

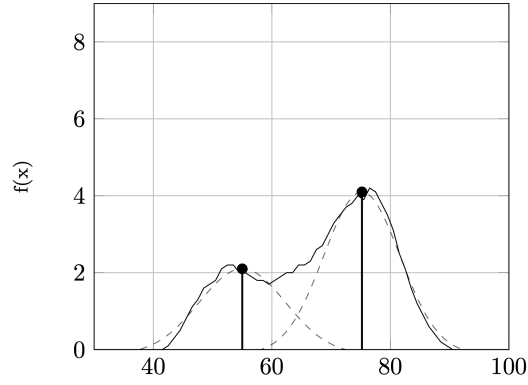
for each sound level meter (Figure 5). For each lesson and technique mean and standard deviation values are shown in Table 2.

Almost all the measured lessons were done in a traditional way with the teacher speaking from the desk. In spite of the fact that some lessons were conducted in different ways, they were kept in the dataset for two reasons: firstly, they are representative of a different use of these spaces; secondly, they allow for a wider analysis of the pros and cons of the investigated techniques.

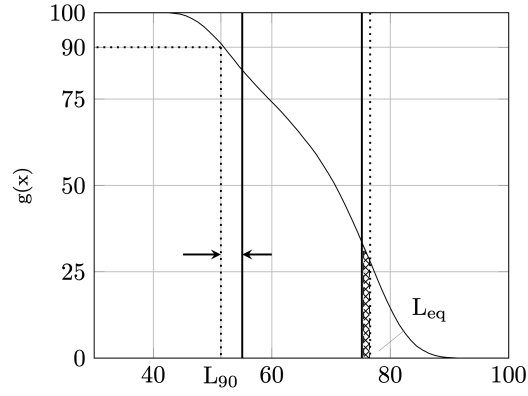
The lesson C (see Figure 5(a)) was a meeting for internship, so more than one teacher talked and students were very active in answering. The measured SA involves the intentional speaking and the non-intentional speaking. Only this latter part is affected by Lombard effect, as will be discussed in Section 5.2. Moreover, it could be assumed that intentional speaking of students is not overlapped to the speech level: teacher’s speaking 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. Methods like KM seem to be more able in the speakers detection, however if the students and the teacher speak at the same time, the higher level overcomes the lower one. Moreover, the interaction between students and teacher may influence the method performance, as it may result in a wider SL curve. GMM is able to account for an increased standard deviation, while PD seems less able to identify this aspect.

The lesson D (see Figure 5(b)) had a long time of media streaming which are transmitted by the PA as the voice of the teacher. This brings to a detected SL with less pauses since its signal is more continuous in time. With reference to the values of table 2, PD and GMM return similar values of SA, whereas KM returns a higher value and PL a lower one. The reasons of these latter differences will be discussed in the next section.

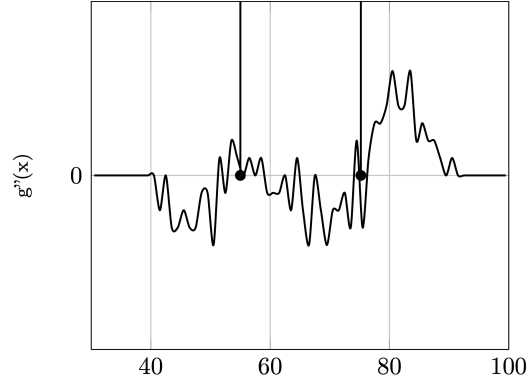
The lesson H (see Figure 5(c)) was an interactive lecture between teacher and students. This means that students paid more attention, indeed the SNRs increase. In this case, PL return a SA value higher than the other methods. As already mentioned, and as it will be discussed in the next section, this depends on the fixed threshold of the 90th percentile used for extracting SA value in PL method. When the student-talking is high, as the case of lesson D, this threshold seems to underestimate the SA; when the students are more silent, PL method seems to overestimate the SA.



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



(b) Cumulative function $g(x)$



(c) $g''(x)$

Figure 6: Relationship between occurrence curve (on the top), cumulative curve (in the middle) and its numerical second derivative (on the bottom). The analysis of the zeros of the numerical second derivative reduces the error among the techniques. The dashed lines in the cumulative curve graph indicate the L_{90} and the L_{eq} levels while the solid vertical lines indicate the corresponding inflection points. The arrows and the dashed area show the differences between the values. The x-axes values, for each graph, are in dBA.

5. Discussions

5.1. Differences between techniques

SA is due to the contribution of several sound sources (the talkers) which vary in time and in space, but they can be treated as homogeneously distributed if they are integrated over the whole lesson time. As a consequence, SA values should be quite homogeneous over the audience. This should be confirmed by low differences between the two sound level meters. Indeed, PD, GMM and KM return low standard deviation values of SA, whereas PL return higher standard deviations. Moreover, teacher’s voice may be heard louder or higher depending on the PA coverage. As a consequence, SL values could show higher differences between the two sound level meters. As it could be expected, due to the differences of the PA coverage, the standard deviations of SL can be comparable only for the lessons carried out in the same lecture hall. As a matter of fact, the standard deviation values for methods PD, GMM and KM are comparable for the lessons done in the same lecture hall. Indeed, they are in the range 4-4.9 dB for Hall I (lesson A-D), in the range 1-1.8 dB for Hall II (lesson E-I), and in the range 1.3-2.4 dB for Hall III (lesson J-L). Instead, PL method does not follow this trend, returning a larger spread of value in the same hall.

Nevertheless, the PL technique had the largest use by scholars [10, 20, 8, 11]. The largest spread among results can be exemplified as follows. In Figure 6(a) the occurrence curve $f(x)$ is plotted highlighting the maxima of the two Gaussians of Figure 6(a). As shown in Figure 6(b) there is a bias, in this case, on SA corresponding to the distance between the 90-th percentile of the $g(x)$ function (dashed line) and the maximum point (i.e. the mode) of the Gaussian on the left of Figure 6(a). Instead, the bias on SL is the area under the $g(x)$ curve highlighted in Figure 6(b). It should be noted that the local maxima highlighted in Figure 6(a) match with the inflection points of the $g(x)$ function in Figure 6(b). It can be confirmed by the numerical second derivative of $g(x)$ function, being the inflection points the zeros of $g''(x)$ function in Figure 6(c).

Furthermore, the results of GMM and K -means are slightly different; this can be due to the fact that the assumption of homoscedasticity is not always fulfilled or to the initialization that may have led the two algorithms to local maximum solutions. As instance of the first sentence, the distance between methods seems to increase when the variances are, respectively, high for speech and low for student noise, e.g. in case of the lessons G and I. Although usually no probability assumption is usually mentioned, K -means 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, \dots, x_n are independently drawn from $\mathcal{N}(\mu_{x_i \in c_k}; \sigma^2)$, $i = 1, \dots, n$, where $\mu_{x_i \in c_k}$, $k = 1, \dots, K$ are parameters giving the cluster memberships of the x_i .

Finally, it should be noticed that KM technique seems to give SL values barely larger than GMM technique. This can be due to the different behaviour of the two sound sources. Within each short-time integration, the student noise can be assumed as a continuous signal due to the high number of simultaneous talkers. Instead, the teacher voice is a non-continuous signal thus, in each integration window, if the teacher and the students talk together, the corresponding sound level is clustered in the group of the teacher. This occurs even if the difference between the two sources is lower than 10 dB. This effect may increase the resulting SL.

5.2. Signal-to-Noise Ratio and Lombard effect

As it was shown in the mean values of Table 2, averaged values of measured SNR are close to +18 dB for PL method, +16 dB for PD and +15 dB for GMM and KM. This value agrees with the previous Shield’s paper [11], but is higher than other measurements done without PA. Moreover, the SNR values seem to be mildly dependent on SA values. Table 3 compares the results of the present work with the ones of previous scholars [9, 10, 12, 39, 20, 8, 11, 15, 18, 35].

Figure 7 plots the measured values from both sound level meters of Table 2, placing on the x-axis the student activity and on the y-axis the speech level and plotting the regression lines which fit the results of each method. Fitting curves of the blind-segmentation methods (GMM and KM) show similar behaviour. The little offset can be due to the reasons explained in the last paragraph. According to this result, there is a sort of self-matching – from the listener point of view – between the noise produced by the student activity and the speech level. As it could be expected, SA values are more diffuse (having small s.d.) whereas SL

Table 3: Comparison of measurement condition and results among the present study and the previous 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.

Ref.	Grade	Hall type	Rooms/Less.	Pos.	Method	P.A.	SNR (dB)	s.d. (dB)	W (ms)
[9]	High school	Classrooms	10/–	1	–	–	9.5	4.6	–
[10]	Elementary	Classrooms	12/–	1	PL	–	-4.5	–	–
[12]	University	Classrooms	11/18	3	PD	No	7.9	3.1	200
[39]	Elementary	Lecture halls	27/27	4	PD	No	11.1	2.5	200
[20]	Elementary	Classrooms	4/–	9	PL	Yes	13.0	–	850
[8]	Elementary	Classrooms	–/54	1	PL	No	2.0	–	50
[11]	Secondary	Classrooms	80/274	1	PL	No	–	–	–
[15]	University	Classrooms	110/–	–	KM	No	14.8	4.6	–
[18]	Elementary	Classrooms	46/59	1 ÷ 2	GM	–	–	–	–
[35]	University	Classrooms	11/15	–	PD	Yes	11.6	–	200
					No	No	9.2	–	200
					PL	No	7.7	2.4	200
					PD	Yes	18.1	3.2	
					GM	Yes	16.3	2.5	
					KM	Yes	15.3	3.1	
					KM	Yes	15.4	3.2	
Present work	University	Lecture halls	3/12	2	PL	Yes	18.1	3.2	100
					PD	Yes	16.3	2.5	
					GM	Yes	15.3	3.1	
					KM	Yes	15.4	3.2	

strictly depend on the PA. KM and GMM have a similar behaviour instead PL and PD which have different biases and slopes. Each student ‘sets’ his/her own speech level in order to not disturb the listening process. It should be considered that a SNR=15 dB is the threshold level below which the background noise influences the intelligibility. Below this value, the modulation functions are penalized and, consequently, the STI value decreases [36]. This self matching may be considered as an “inverse” Lombard effect [37]. Being the slope values less than one, it means that this inverse effect is more evident at low SA values and at high SA values.

The results of the present study can be related to the behaviour of the student population. In the present case, the students are quieter at the beginning of the lesson and after the break, while are noisier at the end of each semi-lesson. They are interested and prone to listen to the lesson, as the attendance is often not mandatory in university courses. The averaged values of SNR decrease in case of secondary or elementary school, but it can be due to non-acoustic reasons. Indeed, apart university lecture rooms, the attendance is mandatory and the lessons are planned for the whole day in the same classroom. This may influence the listening effort, increasing the student activity [38].

6. Conclusions

An in-depth analysis was conducted in three university lecture halls in order to evaluate the noise which affects the speech intelligibility during a lecture. Since the HVAC system was switched off, the analysis is focused on the noise due to the student activity (SA) and the useful signal of speech level (SL). Four techniques were used to extract SA and SL values: visual-segmentation methods such as Percentile levels (PL) and Peak Detection (PD); blind-segmentation methods such as Gaussian Mixture Model (GMM) and K-means clustering (KM).

Based on recordings of twelve lessons, the study shows a mutual comparison among the four methods. PL method returns a larger spread values compared to the other techniques. As discussed, the student activity

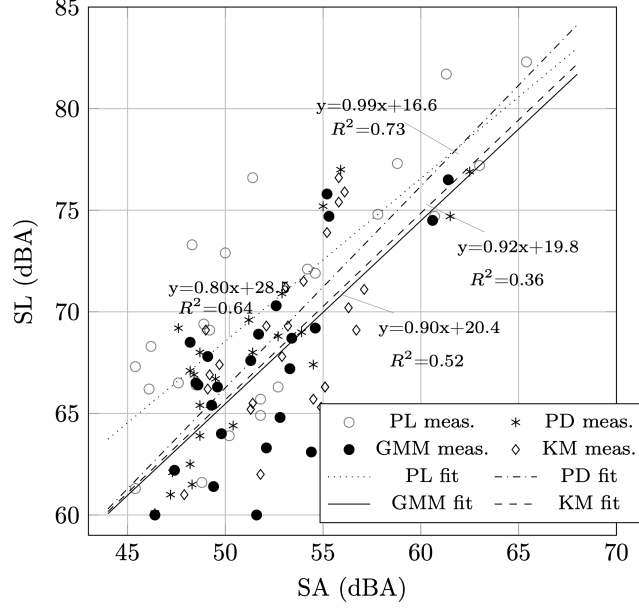


Figure 7: Measured SA and SL values for both sound level meters. Fitting curves for each method are also shown.

in university lecture halls may vary depending on the kind of lesson. The fixed threshold of 90th percentile can overestimate or underestimate this behaviour. Moreover, the acoustic coverage of PA system seems to spread the SL values extracted by PL, more than the other methods.

PD seems to return different values from other methods when more than one speech sources coexist, such as the case of meetings or multi-talker lessons. While this case is negligible in primary or secondary schools, this can happen in university lecture halls.

Blind-segmentation methods seem to be the most reliable techniques to make a segregation of the sound sources, but they need statistical post-processing. Due to the temporal properties of teacher's voice, KM seems to return SL values barely larger than GMM, but as a matter of fact this difference could be negligible. Both blind-segmentation methods return comparable values of SA .

Furthermore, the paper proposes some considerations which can be useful to improve the PL technique, which is the most and often the only one used by acoustic consultants. The comparison between cumulative and occurrence curves shows the correspondence between their shapes basing on the recorded data. It has been shown that, in order to reduce the bias between different techniques, it is possible to refine the analysis in a simple way. Being the cumulative curve shaped by the inflection points, the study of its second derivative can bring the exact thresholds of the accumulation function, from which to extract SA and SL values.

The present study returns also some findings on signal-to-noise ratio that enrich the currently limited literature on student activity in lecture halls with PA. When the teachers use the PA support during the lectures, the measurement results pointed out an inverse Lombard effect, which switches the SA to lower values maximizing the intelligibility. University students seem to automatically set their noise contribution in order to not influence the intelligibility of teacher's voice. Indeed the difference between SL and SA – which is an estimate of signal-to-noise ratio when the HVAC systems are switched off – is around 15 dB over a wide range of SL values. This “inverse effect” seems to be “saturated” when the SL is loud.

References

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

- [2] Houtgast, T., Steeneken, H. J., and Plomp, R. (1980). Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. *Acta Acust. united Acust.*, 46(1), 60-72.
- [3] Secchi, S., Astolfi, A., Calosso, G., Casini, D., Cellai, G., Scamoni, F., Scrosati, C. and Shtrepi, L. (2017). Effect of outdoor noise and façade sound insulation on indoor acoustic environment of Italian schools. *Appl. Acoust.*, 126, 120-130.
- [4] Black, J. W. (1950). The effect of room characteristics upon vocal intensity and rate. *J. Acoust. Soc. Am.*, 22(2), 174-176.
- [5] Brunskog, J., Gade, A. C., Bellester, G. P., and Calbo, L. R. (2009). Increase in voice level and speaker comfort in lecture rooms. *J. Acoust. Soc. Am.*, 125(4), 2072-2082.
- [6] Reich, R., and Bradley, J. (1998). Optimizing classroom acoustics using computer model studies. *Can. Acoust.*, 26(4), 15-21.
- [7] Hodgson, M. (1999). Experimental investigation of the acoustical characteristics of university classrooms. *J. Acoust. Soc. Am.*, 106(4), 1810-1819.
- [8] Bottalico, P., and Astolfi, A. (2012). Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms. *J. Acoust. Soc. Am.*, 131(4), 2817-2827.
- [9] Houtgast, T. (1981). The effect of ambient noise on speech intelligibility in classrooms. *Appl. Acoust.*, 14(1), 15-25.
- [10] Markides, A. (1986). Speech levels and speech-to-noise ratios. *Br. J. Audiol.*, 20(2), 115-120.
- [11] 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. *J. Acoust. Soc. Am.*, 137(1), 177-188.
- [12] Hodgson, M., Rempel, R., and Kennedy, S. (1999). Measurement and prediction of typical speech and background-noise levels in university classrooms during lectures. *J. Acoust. Soc. Am.*, 105(1), 226-233.
- [13] Jeong, C. H., Marie, P., Brunskog, J., and Møller Petersen, C. (2012). Audience noise in concert halls during musical performances. *J. Acoust. Soc. Am.*, 131(4), 2753-2761.
- [14] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal process.*, 10(1-3), 19-41.
- [15] Brill, L. C., and Wang, L. M. (2016). Comparison of occupied and unoccupied noise levels in K-12 classrooms. *J. Acoust. Soc. Am.*, 139(4), 1979-1979.
- [16] Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Mal. de L'Oreille et du Larynx*, 37, 101-119.
- [17] Whitlock, J., and Dodd, G. (2006). Classroom acoustics—controlling the cafe effect... is the Lombard effect the key. *Proceedings of Acoustics, Christchurch, New Zealand*, 20-22.
- [18] Peng, J., Zhang, H., and Wang, D. (2018). Measurement and analysis of teaching and background noise level in classrooms of Chinese elementary schools. *Appl. Acoust.*, 131, 1-4.
- [19] Bradley, J. S. (2002). Acoustical design of rooms for speech. NRC-IRC.
- [20] 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. *Lang. Speech Hear. Ser.*.
- [21] Titze, I. R., Hunter, E. J., and Švec, J. G. (2007). Voicing and silence periods in daily and weekly vocalizations of teachers. *J. Acoust. Soc. Am.*, 121(1), 469-478.

- [22] 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. *J. Acoust. Soc. Am.*, 146(5), 3590-3628.
- [23] McLachlan, G. J., and Peel, D. (2004). *Finite mixture models*. John Wiley and Sons.
- [24] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.*, 31(8), 651-666.
- [25] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE T. Inform. Theory*, 28(2), 129-137.
- [26] dBTrait 5.5. 01dB ACOEM group, 2015.
- [27] OriginPro 2017. OriginLab Corporation, 2016.
- [28] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B.*, 39(1), 1-22.
- [29] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.*, 8(1), 205-233
- [30] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation; 2017.
- [31] ISO 3382-2:2008, Acoustics - Measurement of room acoustic parameters - Part 2: Reverberation time in ordinary rooms. Geneva: International Organization for Standardization.
- [32] D’Orazio, D., De Cesaris, S., Guidorzi, P., Barbaresi, L., Garai, M., and Magalotti, R. (2016, May). Room acoustic measurements using a high-SPL dodecahedron. 140th Audio Engineering Society International Convention 2016, AES 2016.
- [33] Choi, Y. J. (2018). Effects of the distribution of occupants in partially occupied classrooms. *Appl. Acoust.*, 140, 1-12.
- [34] DIN 18041:2016, Acoustic quality in rooms - Specifications and instructions for the room acoustic design, DIN-Normenausschuss Bauwesen (in German).
- [35] Choi, Y. J. (2020). Evaluation of acoustical conditions for speech communication in active university classrooms. *Appl. Acoust.*, 159, 107089.
- [36] IEC 60268-16:2011, Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index. Geneva: International Electro-technical Commission.
- [37] Bottalico, P. (2018). Lombard effect, ambient noise, and willingness to spend time and money in a restaurant. *J. Acoust. Soc. Am.*, 144(3), EL209-EL214.
- [38] Lundquist, P., Holmberg, K., and Landstrom, U. (2000). Annoyance and effects on work from environmental noise at school. *Noise Health*, 2(8), 39.
- [39] Sato, H., and Bradley, J. S. (2008). Evaluation of acoustical conditions for speech communication in working elementary school classrooms. *J. Acoust. Soc. Am.*, 123(4), 2064-2077.