10.2478/msr-2013-0012

MEASUREMENT SCIENCE REVIEW, Volume 13, No. 2, 2013

Management of Truncated Data in Speech Transmission Evaluation for Pupils in Classrooms

G. Genta[1], A. Astolfi[2], P. Bottalico[2], G. Barbato[1], R. Levi[1]

[1]Politecnico di Torino, Department of Management and Production Engineering, Corso Duca degli Abruzzi 24, 10129 Torino, Italy, gianfranco.genta@polito.it

[2]Politecnico di Torino, Department of Energy, Corso Duca degli Abruzzi 24, 10129 Torino, Italy, arianna.astolfi@polito.it

Speech intelligibility is a subjective performance index defined as the percentage of a message understood correctly. Often the results of speech intelligibility tests would suggest that conditions are acceptable, with Intelligibility Score (IS) of the order of 90% or more, while speech transmission performance may not be satisfactory. Subjective ratings of the Listening Easiness Score (LES), based on a discrete questionnaire, provide an alternative approach. A total of 239 primary school pupils, aged 7 to 11, evenly distributed among the grades, participated in the survey. The objective indicator Speech Transmission Index (STI) was also measured for each test setting in seven different positions in the laboratory classroom used for the test. Both IS and LES are inherently bounded, and their data distributions exhibit a significant accumulation of scores in the upper and lower parts. The resulting truncation problem has been addressed with a method based on the normal probability plot, enabling identification of mathematical models relating IS and LES to STI, as well as the estimation of related uncertainties. IS and LES exhibit substantially similar metrological capabilities, as, for both, model relative uncertainty does not exceed 4% and uncertainties in prediction of new observations are about twice as large.

Keywords: Speech intelligibility, listening easiness, data accumulation, truncated data, uncertainty

1. INTRODUCTION

IMPORTANT MEASURES at the boundaries between physics and psychology need to be addressed frequently; specific attention shall then be paid to particular characteristics that cannot be managed with the usual methods normally applied in physical measurements. According to an axiomatic approach, only quantities evaluated on interval or ratio scales [1] can be added, allowing consequently the use of common methods like the evaluation of average and standard deviation. Subjective evaluations involve generally ordinal scales, therefore average and standard deviation should be axiomatically considered as forbidden operations. However, the central limit theorem may frequently be relied upon to produce a nearly normal distribution of results, allowing, therefore, calculation of average and standard deviation; in fact in the common practice, starting from school mark evaluation, average is of general use, and frequently works adequately, while strict application of axioms may lead to complications which may well be dispensed with. A reasonable way between the strict application of axioms and the blind use of inadequate generic methods shall be found in [2].

In this work the actual case study of speech transmission evaluation in classrooms allows to deal with the above matter, where application of common statistics may not be made in a straightforward way to the type of subjective measurement scale considered. Moreover, current speech transmission scales may not have an extension adequate for managing the variability of responses in all the observed acoustic conditions, so that a significant accumulation of subjective scores in the upper and lower part of the scales occurs [3]. The resulting truncation problem [4] should be addressed in order to obtain meaningful relationships between subjective evaluations and speech transmission measures.

Speech, a major mean of communication between people, involves three sequential components, namely speaker (or talker), transmission channel and listener. The transmission channel between speaker's mouth and listener's ear affects the deterioration of the speech signal: important influences are ambient noise, reverberation, echoes, limitation in the frequency response and non-linearities [5].

The quality of speech communication is usually expressed in terms of speech intelligibility, quantified as the intelligibility score (IS), i.e., the percentage of a message understood correctly. It is evaluated by a performance test, i.e., the ability of a group of listeners to understand speech by a talker in different acoustic conditions. The intelligibility scores are not only connected with the physical situation, but also with the organization of the performance test (e.g., number and type of word pairs proposed). It frequently happens, as is also evidenced in the case examined, that over a certain level of goodness of the transmission channel most of the subjective answers are crowded on the 100% value of IS. The same happens for bad conditions of the transmission channel, where the IS scores plummet down to 0%.

The complexity of such evaluation is evident, underlining the importance of a measurement method based upon objective data pertaining to signal, noise, reverberation, etc., in order to take into account the condition of speech transmission. The most frequently used method evaluates the Speech Transmission Index, STI [6]-[8], which is based on the concept of the modulation transfer function. The latter quantifies, for a talker-to-listener speech transmission path, the reductions in the intensity modulations of a speech signal when sounded in a room or through a communication channel. STI, whose unit upper bound refers to optimal transmission conditions, is not affected by the scale truncation of IS. A difference should be underlined, namely

that IS is a direct evaluation of intelligibility, while STI refers to the capability of the transmission channel. Other methods, directly related to intelligibility, were therefore studied in order to avoid the problem of saturation of IS.

Sato et al. [9] proposed, as an alternative approach, the subjective rating of the "listening difficulty" of speech recognition. It is based upon a discrete subjective scale, from 1 to 4 [10]-[12], and calculated as the percentage of the sum of the difficulty responses "2", "3" to "4" (i.e., except "1" – not difficult). Listening difficulty ratings result in an ordinal scale from 0% to 100%, but, again, the values of 0% (perfect speech transmission performance) and 100% (worst performance) do not correspond to physical limits, therefore they also may suffer the same problem of saturation evidenced for IS. It should be noted that listening difficulty ratings decrease for conditions with improved speech transmission, contrary to IS.

Listening difficulty can be turned into its contrary "listening easiness" and its computational method changes in order to make it increasing with improved speech transmission conditions, as shown below in the paper.

To sum up, three methods have been proposed by standards or scientific literature in order to evaluate speech transmission: the Speech Intelligibility Score (IS) that comes from a performance test, the listening difficulty score that comes from a subjective personal evaluation, more dependent on the typical vagaries of human judgment, and the Speech Transmission Index (STI) that is an objective method. Advantages and disadvantages of the three methods may be summarized as follows: IS and listening difficulty are aimed at a direct evaluation of the effects of speech transmission conditions, but require the answers of a jury of adequate size. STI, based on instrumental measurement, is generally applicable, but evaluates the channel transmission characteristics and not directly the intelligibility performances, requiring therefore the use of specific correlation functions. A rough relationship, valid only for adults, is given in the ISO Standard 9921 [5] by a conventional five level scale: "Bad" speech intelligibility entails an STI lower than 0.30, "Poor" between 0.30-0.45, "Fair" between 0.45-0.60, "Good" between 0.60-0.75, while an "Excellent" rating corresponds to an STI exceeding 0.75.

An analysis of such main metrological characteristics as resolution and reproducibility of the three methods shows that the five level scale does not properly represent the real capability of IS and listening difficulty. Some researchers therefore identify the above functions with mathematical models, which should also include the relevant model uncertainty.

Relationships between IS and STI have been recently given for pupils of age ranging from 7 to 11 years ([3], [13], [14]). The present work is based on the previous research described in [3], where the influence of different room acoustics and types of noise were experimentally investigated in four primary schools.

The case study concerns a primary school where speech intelligibility score and the listening easiness score were obtained in a laboratory classroom. Pupils of different classes, from grade 2 to grade 5, were tested there with different noises and various speech and noise levels in order to cover a wide range of A-weighted speech-to-noise level differences (S/N(A)). STI was also measured for each test setting in seven different positions in the classroom, and the results correlated with speech intelligibility scores and listening easiness scores.

2. CASE STUDY

The study involved a primary school in Turin (Italy) located in a quiet residential area. The school, designed at the end of the nineteenth century, is characterized by classrooms with high ceilings and large windows. In the classroom (4.9 m height and a volume of 245 m^3) selected as a laboratory, speech intelligibility and listening easiness tests and acoustical measurements were carried out by rotating classes from grades 2, 3, 4 and 5 (nominally 7, 8, 9 and 10 year olds). The walls were plastered and the floor was covered with ceramic tiles, while the ceiling was covered with acoustical plaster. The number of pupils in the lab-classroom during the tests ranged from 15 to 20, and the average occupied reverberation time, for combined 500 Hz and 1 kHz octave bands and over microphone positions, was 0.74 s (st. dev. 0.01) [3].

A. Speech intelligibility and listening easiness tests

A Diagnostic Rhyme Test, DRT, used as the speech intelligibility test ([15], [16]), consists of 105 bisyllabic word pairs in the Italian language, given in rhyme, in which the initial consonant is changed in order to evaluate different phonetic characteristics. Some of the items are nonsense words for the pupils. A total of 15 tests, each composed of 7 word pairs in rhyme, one for each phonetic category, were obtained from the word list. Each word was presented in a carrier phrase randomly chosen from a set of eight. The pupils heard one word at a time and marked the answering sheet by indicating which of the two words they thought was correct and, immediately after marking the word, they rated the listening easiness on a 5-point scale labeled with the descriptors: "very difficult", "difficult", "fairly easy", "easy" and "definitely easy".

B. Measurement set-up and equipment

Test sentences were recorded by a female talker in an anechoic room (above 250 Hz), reflection free and with negligible noise. The DRTs were administered to the pupils sitting in their normal positions in the lab-classroom, listening to the recorded sentences from the B&K 4128 Head and Torso Simulator (HATS) located at the teacher's desk.

Typical kinds of classroom noise were presented to each class at different levels during the tests. A typical traffic noise sample, recorded next to a busy street, was reproduced using a digital audio player and a loudspeaker (B&K mod. 4224). Classroom babble, fan-coil and impact noise were recorded in a dead and occupied room and reproduced by means of an omni-directional source (B&K mod. 4296).

The measurement set-up of the laboratory classroom is shown in Fig.1. The HATS was located at the teacher's position and oriented towards the pupils' seating area. The

loudspeaker, for traffic noise emission, was placed outside the school and oriented towards the lab-classroom. The omni-directional source was placed in the center of the classroom at 1.3 m above the floor.

The acquisition system consisted of 7 omni-directional microphones (ECM 8000) connected through an amplifier to 7 sound card inputs (Echo Audiofire 8), linked to a PC.

A receiver was positioned 1 m away from the source's mouth and six others were positioned at representative students' seats, uniformly distributed over the seating area. The receiver in front of the source's mouth was placed at mouth height, 1.5 m above the floor, while the other receivers were placed at ear height of the seated pupils, 1.1 m above the floor.

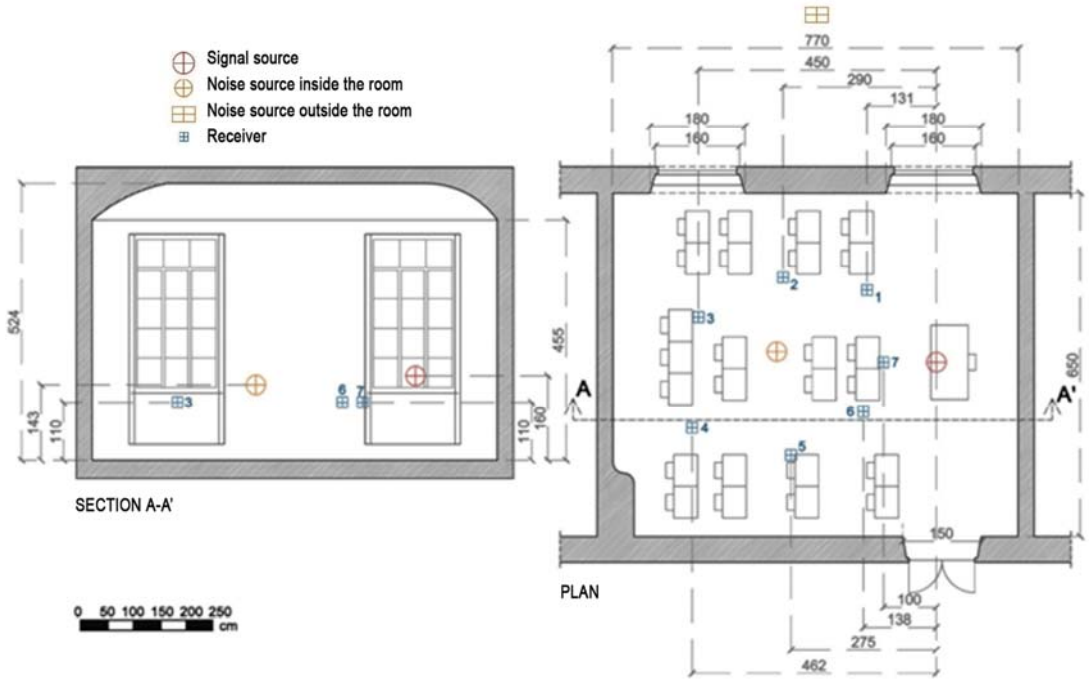

Fig.1. Laboratory classroom and position of measurement setup.

C. Test administration and measurement procedure

Each class spent approximately 45 minutes in the laboratory classroom. After a brief explanation of the experiment and a period in which the pupils filled in their data sheets, each child was given a set of eight sheets, each corresponding to a different test. Test administration and measurement of STI were carried out in the same work session.

Distortions in the time domain (e.g., echoes and reverberation), noise interference and non-linear distortions may degrade the fluctuating speech signal and reduce the intelligibility. This is modeled in the STI measurement which determines the degree to which the intensity envelope of the speech signal is affected by a transmission channel. By using *AURORA* 4.2 [17] the impulse response was acquired as well as the speech and the noise levels, the modulation transfer function derived from which the STI was subsequently calculated.

With the pupils sitting quietly, the impulse response was measured at the seven points by means of an exponential sweep signal [18] emitted by the HATS. The eight intelligibility tests were then administered with the different noises in a random order among the classes in order to prevent effects, such as tiredness or a decrease in concentration, from affecting the same noise.

A special sentence, composed of one carrier phrase and a sequence of seven words without pauses, was edited and recorded for the speech level measurement in order to have a continuous speech sample. The overall A-weighted level difference between the special sentence and each single test without pauses was under 1 dB.

The special sentence was emitted by the HATS and recorded for each class at the end of the session with the pupils sitting quietly. In order to minimize the influence of noise on the signal recording, speech level at each measurement location was checked to exceed the noise level by at least 6-10 dB for each octave-band from 125 Hz to 8 kHz.

As a general rule, all the tests were administered with the same vocal effort [5] for a single class in the range between 56-64 dB(A), which was set by acting on the software gain.

In order to obtain the correct noise level at each measurement position, the noise sample used for the test was recorded without speech, after the test had been administered. The ambient noise was recorded with the pupils sitting quietly, and there was no significant noise in the classroom, the level not exceeding 45 dB(A).

Various speech and noise levels were considered in order to cover a wide S/N(A) range. Almost the same S/N(A) range was maintained for each noise, and an overall variation of 9 to 21 dB was determined for ambient noise, -10 to 18 dB for traffic noise, -9 to 20 dB for babble noise, -6 to 10 dB for fan-coil noise and -15 to 2 dB for impact noise.

3. DATA EXPLORATION AND ANALYSES

An exploratory data analysis was preliminary performed ([19], [20]). The eight tests were administered to 239 pupils aged 7 to 11, evenly distributed among the grades for a grand total of 1912 useful tests. The native language listeners were 88%, and 51% were male. More details about the selection of the sample are given in [3].

The speech intelligibility score (IS) of each pupil was expressed as the percentage of the words understood correctly, with no correction applied for the *a priori* probability of 50 per cent correct responses as a consequence of the two-alternative-choice procedure ([13], [15]).

For each test seven pairs of words were used, hence the possible answers are integers from 0 (none correct) to 7 (all correct), and therefore resolution is 1/8 (12.5%). Clearly, when the transmission channel is worse than a threshold low level the answer is always 0, while when it is better than a given high level the answer is always 7. Beyond that range of transmission, the method lacks sensitivity, leading therefore to an accumulation of answers on 0 and/or 7 scores.

To address the problem of the lack of sensitivity beyond boundaries frequently too close, the listening difficulty score was proposed by Sato et al. [9] as an alternative approach to IS. In this work, its contrary, the Listening Easiness Score, LES, was determined. For each word, it is based on a five-point discrete scale, ranging from 0 to 100%: the lowest corresponds to “extremely difficult”, 25% to “difficult”, 50% to “fairly easy”, 75% to “easy” and 100% to “definitely easy”. LES is calculated as the average of the scores obtained for the seven words pertaining to each test. It should be noted that LES increases for conditions with improved speech transmission, coherently to IS.

The elementary variation of ± 25% on each single evaluation gives, on the average of seven words, a LES resolution of one seventh of 25% (3.6%). In the case of questionnaires with one or more missing answers the real resolution can change, but for evaluating the method it is sufficient to consider the nominal resolution, that for LES is really better than for IS, and this helps to overcome the problem of the limited range. Moreover, LES being influenced by personal preferences, the boundaries at 0% and 100% are not as strictly determined as for IS, but are quite variable among different jury members, and this too helps to get a larger range of LES *vs.* transmission conditions. Nevertheless, also for LES an accumulation, mainly on upper value of 100%, is frequently present.

The pupils’ seating area in each classroom was divided into seven approximately equal areas around each measurement point. Each area included at least two pupils’ positions in order to relate the objective parameters to the speech intelligibility scores. The IS and LES for each pupil was associated with the STI value measured by the closest microphone.

A rigorous evaluation of STI resolution is difficult and hardly meaningful if based only on instrumental reading resolutions, because of the involvement of human related aspects. It therefore makes sense to accept as resolution the Just Noticeable Difference (JND) of 5% indicated in [22].

A. Truncation problems

The IS and LES were examined as a function of STI, as this was the key independent variable in the experiments. Both IS and LES are inherently bounded. As far as IS is concerned, when acoustic conditions are particularly favorable and over a certain upper level, all the answers (7 in this study) are given correctly and a 100% score is regularly reached, the test being unable to detect any further variation. The same happens also in poor acoustic conditions, worse than a certain lower level, below which no correct results can be obtained and therefore a 0% score is always recorded. These effects are also revealed for LES, as already underlined elsewhere [10], and clearly shown in Fig.2., where the histogram of LES is compared with the normal distribution of data obtained using only data of the central part of the scale, excluding 0% and 100%. Clearly the frequency density for 0% and 100% compensates the presence of measurands in the tail zone.

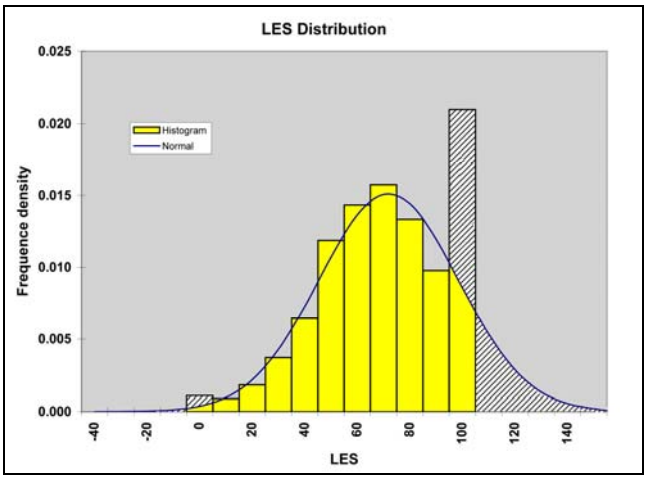

Fig.2. Histogram of LES (in percent) with superimposed normal distribution, fitted using only data of the central part of the scale (excluding 0% and 100%). The upper part exhibits a compensation between the empty tail and the accumulation on 100% level. On the lower part accumulation and compensation are also present, albeit less evident owing to average shift.

Accumulation effect is also evident for data distributions given as normal probability plot (NPP) in Fig.3(a). and 3(b)., for IS and LES, respectively, exhibiting significant accumulation of scores at the upper bound. In the case of LES, accumulation of scores at the lower bound is also noticed.

These situations are avoided in metrological practice, e.g., by selecting instruments with adequate coverage of all the measurand range; in some instances, however, accumulations have to be managed, as occurred elsewhere [23].

Average and standard deviation may not be resorted to, owing to results falling at both bounds ([1], [2]). Fig.3. points out that distortion problems, evidenced by the amount of asymmetry, are more significant for IS than for LES and it also confirms that IS presents a resolution lower than LES.

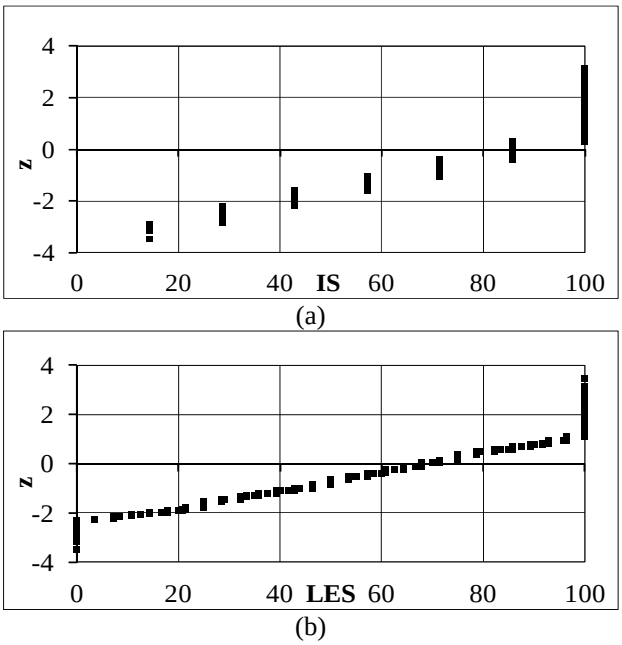

Fig.3. Normal probability plot of IS (a) and LES (b), both in percent, for the complete data set. Standard normal order statistic medians *z* shown in ordinate, [21].

B. Management of truncated data

A severe truncation problem was detected, both for intelligibility score (IS) and listening easiness score (LES), particularly felt for higher values of STI. To address this problem, data were divided into 10 groups, identified by their decile based on STI, each group consisting of about 190 entries. STI ranges pertaining to the first eight groups are by and large comparable to the just noticeable difference (JND) of the index corresponding to 0.05 [19], while the last two groups exhibit a larger scatter, as shown in Table 1.

Table 1. Descriptive statistics relevant to the 10 groups identified by decile based on STI (adimensional).

Group	STI (central value)	STI (range)	N
1	0.237	0.056	195
2	0.287	0.040	191
3	0.323	0.031	188
4	0.357	0.033	195
5	0.409	0.063	197
6	0.462	0.041	182
7	0.494	0.019	207
8	0.542	0.076	176
9	0.648	0.122	205
10	0.804	0.184	176

Normal probability plots (NPP) of both IS and LES are shown in Fig.4(a). and 4(b). for group No. 5, and in Fig.5(a). and 5(b). for group No. 10. For increasing STI values the saturation problem is seen to become more severe. Similar probability plots were drawn for all the groups in the course of the analysis. If in the NPP the extreme scores (i.e., 0 and 100%) are excluded, paths can reasonably be approximated with straight lines, as can be argued from Fig.3.

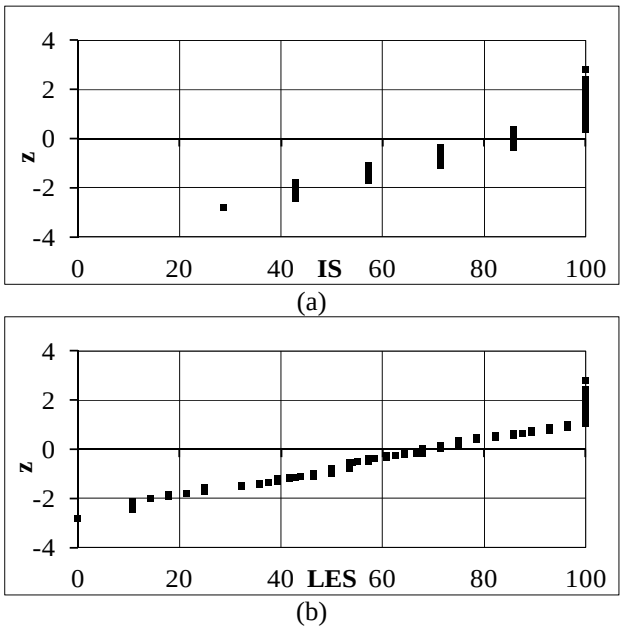

Fig.4. Normal probability plot of IS (a) and LES (b), both in percent, for group No. 5, i.e., for STI ranging from 0.377 to 0.440.

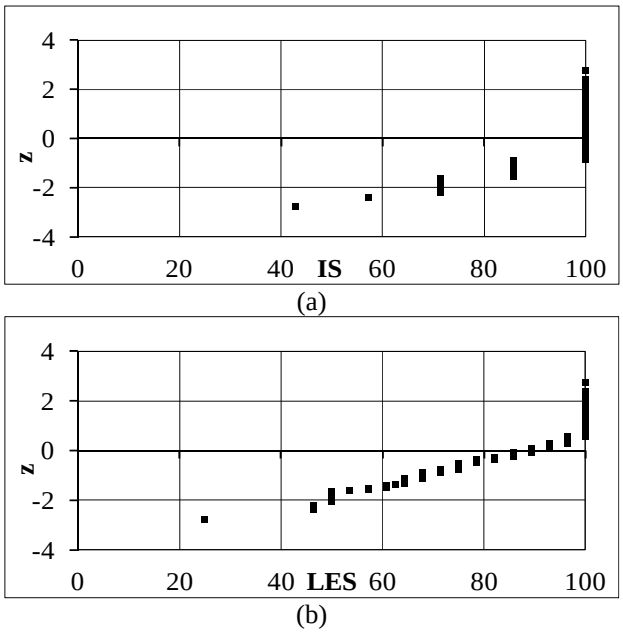

Fig.5. Normal probability plot of IS (a) and LES (b), both in percent, for group No. 10, i.e., for STI ranging from 0.712 to 0.896.

The hypothesis that measurands are distributed normally, but the limited range of the measurement methods adopted accumulates on the bounds all the results of the tails, can therefore be considered.

For the evaluation of statistical parameters, the simple elimination of the extreme data certainly produces erroneously mean values and smaller values of standard deviations.

However, the statistical parameters of LES and IS can be directly evaluated from the rectilinear part of NPP by taking into account the standardizing transformation:

$$z = \frac{x - m}{s} \tag{2}$$

where x is the value of IS or LES, m is the average value, s is the standard deviation and z is the relevant standard normal variable.

The least squares line derived from the data subset obtained excluding the extreme scores is the following:

$$z = a + bx \tag{3}$$

where a and b are the regression coefficients.

Mean m and standard deviation s may be calculated as follows:

$$\begin{aligned} z = 0 &\Rightarrow x = -\frac{a}{b} = m \\ z = 1 &\Rightarrow x = \frac{1-a}{b} = m + s \Rightarrow s = \frac{1}{b} \end{aligned} \tag{4}$$

Statistical parameters for the ten groups are given in Table 2.

Table 2. Mean value m, standard deviation s and expanded uncertainty of the mean U_m for each group of IS and LES derived with least squares applied to the rectilinear part of the NPP.

Group	STI	IS/%			LES/%		
		m	s	U_m	m	s	U_m
1	0.240	74.0	22.0	0.73	55.4	25.6	0.31
2	0.281	74.5	22.2	0.75	52.8	27.3	0.23
3	0.323	80.0	22.0	1.1	59.2	24.5	0.31
4	0.357	82.8	20.1	1.1	62.0	22.1	0.25
5	0.405	85.5	20.7	1.0	69.7	25.8	0.35
6	0.469	87.4	17.2	1.1	70.3	22.6	0.34
7	0.493	89.0	21.4	1.1	74.8	19.9	0.28
8	0.537	94.7	23.5	1.7	73.5	22.2	0.31
9	0.662	99.5	24.1	2.3	81.1	25.3	0.34
10	0.741	114.4	24.1	4.6	88.4	20.5	0.39

The uncertainty of the mean values, evaluated as prescribed by the *Guide to the Expression of Uncertainty in Measurement* (GUM) [24], is generally around 0.3% for LES, while for IS it is about 1% and up to 5% for the highest STI intervals (groups 8, 9 and 10), where the intersection of the rectilinear part of the NPP with abscissa axis is significantly higher than 85.7%, i.e., the maximum IS available before 100%. This involves extrapolation and consequently a much higher uncertainty.

In order to show that theoretical axioms ([1], [2]) correctly forbid estimates of IS and LES by the usual computation of average and standard deviation from all data, Table 3 reports the differences between mean values obtained considering the complete data set and those derived with least squares from the rectilinear part of NPP. Conventional mean values are lower than those derived with least squares from the rectilinear part of NPP, and absolute differences tend to increase for increasing STI values.

Table 3. Mean value m and standard deviation s obtained with conventional calculations, expanded uncertainty of the mean U_m and absolute differences of mean values with reference to Table 2. Over a STI of about 0.5, the differences exceed the relevant expanded uncertainties, showing that conventional calculations do not give acceptable approximations.

Group	STI	Conventional calculations						Absolute diff. /%	
		IS/%			LES/%				
		m	s	U_m	m	s	U_m	IS	LES
1	0.240	72.8	19.4	3.1	55.2	24.4	3.6	1.2	0.2
2	0.281	73.3	19.6	3.2	52.5	26.0	3.9	1.2	0.3
3	0.323	78.0	18.2	3.2	58.9	23.6	3.5	2.0	0.3
4	0.357	81.1	17.3	2.8	61.7	21.3	3.1	1.7	0.3
5	0.405	83.0	16.5	2.9	68.3	23.1	3.6	2.5	1.4
6	0.469	85.6	13.7	2.5	69.4	20.9	3.3	1.8	0.9
7	0.493	85.4	16.1	2.9	74.0	18.3	2.7	3.6	0.8
8	0.537	87.8	15.2	3.5	72.2	20.4	3.3	6.9	1.3
9	0.662	90.3	14.4	3.3	77.7	20.9	3.5	9.2	3.4
10	0.741	96.2	8.9	3.6	84.7	15.6	3.0	18.2	3.7

Over a STI of about 0.5, the differences exceed the relevant expanded uncertainties, therefore, conventional calculations do not give acceptable approximations. Furthermore, standard deviations calculated using the conventional methods are smaller for higher values of STI as a consequence of truncation, while standard deviations derived from the NPP regression are substantially more uniform over the ten groups.

C. Subjective parameters as functions of STI

Using mean values derived from the NPP regressions, the following linear regressions between STI, IS and LES are identified:

$$IS = 73 \cdot STI + 55 \tag{5}$$

$$LES = 68 \cdot STI + 38 \tag{6}$$

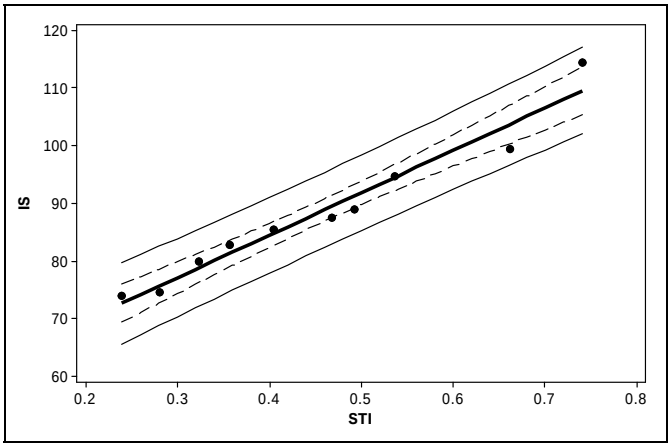

Fig.6. Regression of intelligibility score IS (in percent) vs. speech transmission index (STI) with 95% confidence bands for observations (solid) and for regression line (dashed). Since the upper limit of IS is 100%, values of STI nominally exceeding 0.62 in the present case entail saturation, in substantial agreement with other findings, [14].

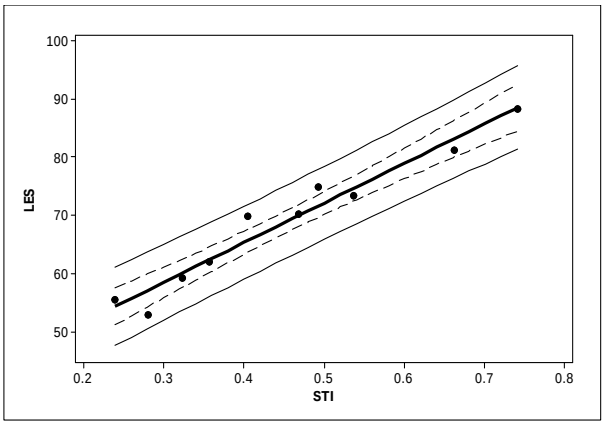

Fig.7. Regression of listening easiness score LES (in percent) vs. speech transmission index (STI) with 95% confidence bands for observations (solid) and for regression line (dashed).

Some fitting models, e.g., logarithmic, proved to be adequate elsewhere [3], however, in the case at hand the linear models have proved to be the best ones.

Figs.6 and 7. show mean values with regression lines for IS and LES versus STI. Possible bias of the models is evidenced in the plots by 95% confidence bands for regression lines (dashed), while the external band (solid) combines the effect of the reproducibility of observation, in the averaging conditions adopted. Both models explain over 95% of variation, thus proving adequate for estimating IS and LES in terms of STI. For both regression models the uncertainty involved is about 4%, while for observations it is about 8%.

4. Discussion and conclusions

Three different methods were considered for evaluation of speech transmission characteristics in a comprehensive series of classroom tests. Two methods are direct evaluations: Speech Intelligibility Score, IS, described in ISO 9921 [5], is a performance test, and Listening Easiness Score, LES, described in literature [10], is based on a discrete scale questionnaire. The third one, Speech Transmission Index, STI, described in EN 60268-16 [8], is aimed at the evaluation of the characteristics of a transmission channel by instrumental measurements. A metrological analysis of IS and LES showed different performances and measurement capabilities. Removal of the IS and LES non-normality characteristics catered for evaluation of such terms as averages, standard deviations and linear relationships. Had efficient estimation of IS and LES characteristics been the main goal, a suitably smaller range would have been selected.

The metrological characteristics for IS and LES not only concern the methods themselves, but are also related to their implementation. However, the considerations given below, specific for the test schemes adopted in the present case study, may also be adapted to other cases, e.g., when the number of target words increases (improvement of resolution) or when the type of speech material has changed (variation of difficulty).

Given the main purpose of STI, identification of mathematical models allowing evaluation of IS and LES in terms of STI was deemed important, as well as analysis of performances and characteristics of IS and LES. IS resolution was found to be remarkably inferior to that of LES and both methods exhibited a problem of measurement range, the scale being truncated at 0% and 100%, not covering the span of the effects of acoustic conditions of the transmission channel. This entails an abrupt change of sensitivity at the boundaries, underlining that the corresponding measurement scale of both methods does not fulfill the conditions of an interval or ratio scale over the whole range of possible values of the measurand; therefore conventional measures of average and standard deviation become misleading.

An alternative method based on NPP was therefore developed to fill the gap, enabling identification of mathematical models relating IS and LES to STI, as well as estimation of parameters and related uncertainty over a reasonably broad range. Statistical inferences may thus be drawn on IS and LES in terms of STI at specified confidence levels.

Comparison of performances of IS and LES shows that these methods exhibit substantially similar metrological capabilities, as model relative uncertainty does not exceed 4% for both IS and LES, uncertainties in prediction of new observations being about twice as large for both. This is due to the fact that, while the reproducibility of single points for LES is lower than for IS (see standard deviations in Table 2.), the scatter of LES points about the regression line exceeds that pertaining to IS, as shown in Figs.6. and 7., explaining substantially equal overall uncertainty for IS and LES.

Dependence of the measurement capabilities of IS and LES on their practical implementation suggests, as future work, to further investigate, from the metrological point of view, alternative implementations of the methods and their resolution, reproducibility and uncertainties.

Not strictly metrological, but substantial factors to be taken into account when considering an improvement of the method, are the effects on jury responses and the costs involved. A change based for instance on increasing the number of target words can produce a lack of attention of the listeners [14] and higher costs involved in terms of the time consumed. Therefore, metrological and practical conditions deserve careful harmonization.

References

[1] Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103 (2684), 677-680.

[2] Barbato, G., Farné, S., Genta, G. (2008). Management of subjective evaluations represented by ordinal scales. In *12th IMEKO TC1 & TC7 Joint Symposium on Man, Science & Measurement*, 3-5 September 2008. IMEKO, 89-94.

[3] Astolfi, A., Bottalico, P., Barbato, G. (2012). Subjective and objective speech intelligibility investigations in primary school classrooms. *Journal of Acoustical Society of America*, 131 (1), 247-257.

[4] Klein, J.P., Moeschberger, M.L. (2003). S*urvival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer.

[5] International Organization for Standardization (2003). *Ergonomics-Assessment of speech communication*. ISO 9921. Genève.

[6] Houtgast, T., Steeneken, H.J.M. (1971). Evaluation of speech transmission channels by using artificial signals. *Acustica*, 25, 355-367.

[7] Steeneken, H.J.M., Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of Acoustical Society of America*, 67, 318-326.

[8] European Committee for Standardization (2011). *Objective rating of speech intelligibility by speech transmission index*. EN 60268-16. Brussels.

[9] Sato, H., Yoshino, H., Nagatomo, M. (1998). Relationship between speech transmission index and easiness of speech perception in reverberatory fields. *Journal of Acoustical Society of America*, 103 (5), 2999.

[10] Sato, H., Morimoto, M., Sato, H., Wada, M. (2008). Relationship between listening difficulty and acoustical objective measures in reverberant sound fields. *Journal of Acoustical Society of America*, 123 (4), 2087-2093.

[11] Morimoto, M., Sato, H., Kobayashi, M. (2004). Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces. *Journal of Acoustical Society of America*, 116 (3), 1607-1613.

[12] Sato, H., Bradley, J.S., Morimoto, M. (2005). Using listening difficulty ratings of conditions for speech communication in rooms. *Journal of Acoustical Society of America*, 117 (3), 1157-1167.

[13] Prodi, N., Visentin, C., Farnetani, A. (2010). Intelligibility, listening difficulty and listening efficiency in auralized classrooms. *Journal of Acoustical Society of America*, 128 (1), 172-181.

[14] Prodi, N., Visentin, C., Feletti, A. (2013). On the perception of speech in primary school classrooms: Ranking of noise interference and of age influence. *Journal of Acoustical Society of America*, 133 (1), 255-268.

[15] International Organization for Standardization (1991). *Acoustics – The construction and calibration of speech intelligibility tests*. ISO TR 4870. Genève.

[16] Bonaventura, P., Paoloni, A., Canavesio, F., Usai, P. (1986). *Development of a diagnostic intelligibility test in the Italian language*. Roma: Fondazione Ugo Bordoni. (Technical Report 3C1286).

[17] Aurora site (last viewed February 24, 2013). http://pcfarina.eng.unipr.it/Aurora_XP/STI.htm.

[18] International Organization for Standardization (2009). *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces*. ISO 3382-1. Genève.

[19] Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

[20] Hofmann, D., Linss, G. (2003). Challenges and chances of internet metrology. *Measurement Science Review*, 3 (1), 1-17.

[21] NIST/SEMATECH. *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/, February 2013.

[22] Bork, I. (2000). A comparison of room simulation software: The 2nd round robin on room acoustic computer simulation. *Acustica*, 86, 943-956.

[23] Johnson, L.G. (1964). *Theory and Technique of Variation Research*. Amsterdam: Elsevier.

[24] Joint Committee for Guides in Metrology (2008). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement (GUM)*. JCGM 100:2008. Sèvres.

Received October 10, 2012.
Accepted April 2, 2013.