POLITECNICO DI TORINO
Repository ISTITUZIONALE

Subjective and objective speech intelligibility investigations in primary school classrooms

(Article begins on next page)

24 September 2024

Subjective and objective speech intelligibility investigations in primary school classrooms

Arianna Astolfi[a)] and Pasquale Bottalico
Politecnico di Torino, Department of Energetics, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

Giulio Barbato
Politecnico di Torino, Department of Production Systems and Business Economics, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

This work concerns speech intelligibility tests and measurements in three primary schools in Italy, one of which was conducted before and after an acoustical treatment. Speech intelligibility scores (IS) with different reverberation times (RT) and types of noise were obtained using diagnostic rhyme tests on 983 pupils from grades 2–5 (nominally 7–10 year olds), and these scores were then correlated with the Speech Transmission Index (STI). The grade 2 pupils understood fewer words in the lower STI range than the pupils in the higher grades, whereas an IS of ∼97% was achieved by all the grades with a STI of 0.9. In the presence of traffic noise, which resulted the most interfering noise, a decrease in RT from 1.6 to 0.4 s determined an IS increase on equal A-weighted speech-to-noise level difference, S/N(A), which varied from 13% to 6%, over the S/N(A) range of −15 to +6 dB, respectively. In the case of babble noise, whose source was located in the middle of the classroom, the same decrease in reverberation time leads to a negligible variation in IS over a similar S/N(A) range. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3662060]

I. INTRODUCTION

Excessive noise, or the combined effect of excessive noise and reverberation, in a classroom reduces speech intelligibility, which is defined as the percentage of a message understood correctly.[1] A high intelligibility score (IS) is the first requirement for a successful learning environment: If speech sounds are not heard clearly, one cannot speak clearly, and if one does not have good spoken language skills, reading will also suffer; if reading skills are below average, an individual will have difficulty performing academically.[2]

Excessive noise and reverberation are mainly encountered in old school buildings, which are common in Italy: The windows in the classrooms tend to have poor sound insulation, which determines high noise levels due to external noise, and the ceilings are very high, which result in longer reverberation times than modern schools.

Intelligibility rating, on a five-point scale ranging from "Bad" to "Excellent" (Bad, Poor, Fair, Good, Excellent), is related to a range of scores of a number of subjective intelligibility tests,[3] and to a range of values of some objective measures,[1,4] on the basis of some physical properties of the talker–listener transmission path and on speaking and hearing aspects.

To the authors' knowledge, only three studies have dealt with subjective intelligibility tests and measurements carried out in real classrooms for different grades,[5–7] and only one of these concerned primary schools.[7] Numerous studies have instead been made in laboratories.[8–11]

The effect of the speech-to-noise level difference at the listener's position was investigated as one of the main issues in most of these studies, together with the influence of age and hearing disorders of the children. The effect of reverberation was only examined in-field by Bradley[6] in 1986, although it had previously been investigated in many laboratory studies, with some serious limitations, above all regarding the monoaural headphone presentation of the test signals, a problem which has since been overcome.[10,11]

In 1981, Houtgast[5] administered intelligibility tests in classrooms with a variety of road traffic noise conditions and with a reverberation time (RT) in the 0.7–1.5 s range. A Fairbanks rhyme test, composed of meaningful consonant–vowel–consonant phonetically balanced words (CVC-PB words), was conducted on the basis of the teacher's voice with 20 teachers and ∼500 pupils, from 8 to 15 years old. The result was that a relationship was found between the A-weighted speech-to-noise level difference, S/N(A), and IS, which was expressed as articulation loss for consonants,[12] %AL$_{cons}$.

In 1986, Bradley[6] determined the combined effects of S/N(A) and reverberation times, RT, on speech intelligibility for 12–13 year old pupils in their classrooms. The RT range was from 0.39 s to 1.20 s. He used a Fairbanks rhyme test emitted from a small loudspeaker with directionality similar to a human talker. Although the S/N(A) was the main determinant of the intelligibility scores, reverberation time had such a significant effect that the decreased reverberation time was related to increased intelligibility scores.

In a recent work by Bradley and Sato,[7] the mean intelligibility scores were significantly related to S/N(A) and to the grade of primary school pupils. They tested grade 1, 3, and 6 pupils (nominally 6, 8, and 11 years old) in classrooms with similar and very low occupied mid-frequency RT, equal to

a)Author to whom correspondence should be addressed. Electronic mail: arianna.astolfi@polito.it

0.4 s, in the presence of noise produced mainly by pupils in other classrooms. The speech intelligibility scores were obtained using a WIPI (Word Intelligibility by Picture Identification) test in rhyme based on simple phonetically balanced (PB) words to be chosen from a series of six represented by a picture. A female voice emitted the words through a small loudspeaker, with similar directionality to that of a human talker.

After this first extensive survey in primary schools, Yang and Bradley,[10] carried out the same speech intelligibility tests as in the previous study in auralized sound fields with pupils of the same ages, with the aim of investigating the effect of varied room acoustics and S/N(A) on speech intelligibility. The reverberation time was varied from 0.3 to 1.2 s and the spectrum shape of the noise approximated the typical noise of ventilation systems. The IS increased with decreasing RT, for conditions of constant S/N(A), whereas for conditions including realistic increases in speech level with varied reverberation time for constant noise level, the intelligibility scores were nearly maximum for a range of reverberation times.

Neuman and Hochberg,[9] in a test on 25 children, from 5 to 13 years old, also obtained increasing intelligibility scores with increasing age of the listeners and decreasing reverberation times, for constant S/N(A). A nonsense syllable speech test, recorded by a male talker, was presented to the pupils via headphones. Nonreverberant, 0.4 and 0.6 s reverberation time conditions were tested with a very low ambient noise level.

Prodi et al.[11] have recently performed intelligibility tests on 80 pupils aged 8–10 years in auralized classrooms, in which the reverberation time varied from ∼0.5 to 1.8 s and the noise was typical of occupied classrooms. The test material and structure were the same as those used in the present work (see Sec. III A), i.e., a bisyllabic diagnostic rhyme test using pairs of words. The Speech Transmission Index[4] (STI), which combines the speech-to-noise level difference and the room acoustics in a single quantity, was used for the analyses over a range of 0.23–0.72. A linear regression model related IS with STI; no significant difference was observed between the classes, whereas differences were instead observed for the response time.

Hodgson and Nosal[13] used the diffuse field theory to establish the optimal reverberation time in classrooms. They found that the optimal reverberation time is zero when the speech source is closer to the listener than the noise source and nonzero when the noise source is closer than the speech source, and decreases with an increased signal-to noise level difference. Yang and Hodgson[14] carried out speech intelligibility tests with adults that involved auralizing virtual sound fields in the presence of babble noise. The results were in agreement with the earlier work which was based on a model.[13]

From an analysis of the previous studies, which are heterogeneous in terms of subjective tests and type of survey, it emerged that the following issues on the topic of speech intelligibility for pupils in a classroom still need to be tackled:

(a) The effect of different reverberation time and interfering noise should be tested with pupils in their own classrooms;

(b) the effect of the position of the noise sources, with reference to the talker–listener path, should be investigated;

(c) the results of different intelligibility tests and languages and those obtained in-field and in the laboratory, should be compared.

The results of speech intelligibility investigations performed in three primary schools, one of which was conducted both before and after an acoustic treatment, are presented in this work, and, due to the different characteristics of the school buildings and the large size of the sample, the previous issues have been dealt with.

II. CASE STUDIES

The study involved three primary schools in Turin (Italy) located in a residential area far from busy streets. The first two schools, A and B, designed at the end of the 19th century, are characterized by classrooms with high ceilings and large windows, whereas the third school, C, was built in the 1970s, and has lower classroom volumes and clerestory windows. The façades of all the buildings have thick masonry walls and single glazed windows, whereas the internal walls are made of thick masonry in the older schools A and B and of thin masonry in school C.

One classroom was selected in each school as a laboratory in which subjective tests and acoustical measurements were carried out by rotating classes from grades 2–5 (nominally 7–10 years old). All the laboratory classrooms had a parallelepiped shape, whereas the rooms in schools A and B had both vaulted ceilings, whereas school C had a flat ceiling. The length (alongside which the pupils were seated) and the width of the rooms were 7.7 and 6.5 m in school A, 6.8 and 6.7 m in school B, and 6.8 and 6.9 m in school C, respectively. The mean heights were 4.9, 4.3, and 3.2 m, with a volume of 245, 193' and 150 m^3, respectively.

The walls in all of the lab-classrooms were plastered and the floors were covered with ceramic tiles, whereas the ceilings in schools B and C were plastered and those in school A were covered with sound absorption material (acoustical plaster). All the in-field analyses in school B were carried out before, B(a), and after, B(b), an acoustic treatment, which consisted of placing porous sound-absorption material (rock-wool panels) onto the ceiling and the upper part of the back and lateral walls, and plaster board panels onto the lower part of the walls. The lab-classroom height was reduced to 3.8 m and the volume to 173 m^3 after the treatment. A plaster-board panel of ∼7 m^2 was also inserted into the flat absorbing ceiling above the teacher's desk with the aim of increasing the first speech sound reflections to the rear part of the room.

III. EXPERIMENTAL PROCEDURES

A. Speech intelligibility tests

A diagnostic rhyme test (DRT), was used as the speech intelligibility test.[3] It was developed by the Fondazione U. Bordoni of Rome[15] following the rules indicated in Ref. 3, and it consists of 105 bisyllabic word pairs in the Italian

language, given in rhyme, in which the initial consonant is changed. Word pairs were chosen in order to evaluate the six consonant typologies with different phonetic characteristics: Nasals, fricatives, affricates, coronals, fronts, and sonorants. Further, a new category, "filler," was added in order to consider other phonetic features. Some of the items are nonsense words for pupils. A total of 15 tests, each composed of seven 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. Each sentence (carrier phrase and word) was ∼3.5 s long and the next sentence was given to the pupils after 5 s. Each test lasted ∼1 min.

The DRTs were administered, in each school, to the pupils who were sitting in their normal positions in the lab-classroom and who listened to the recorded speech material from a head and torso simulator with directionality similar to a human talker. A total of 7864 tests were administered to 983 pupils aged 7–11, and were evenly distributed among the grades and gender in the four different schools. The native language listeners were 88.1%, 76.1%, 71.4%, and 66.9% in the A, B(a), B(b), and C schools, respectively.

B. Measurement equipment and setup

The equipment used for the measurements is composed of playback and acquisition systems. The tests were recorded by a female talker in an anechoic room (above 250 Hz), so that it was reflection free and with negligible noise. A special sentence, composed of one carrier phrase and a sequence of seven words without pauses, was also 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 recording was made 1 m away from the talker's mouth with an angle of 45° with respect to the front position. The playback system was equalized by an inverse filter, generated with the Aurora 4.2 plug-in. The speech source was a B&K 4128 Head and Torso Simulator (HATS) connected to an amplifier, interfaced to a PC through a TASCAM USB122 sound card.

Typical kinds of classroom noise were presented to each class at different levels. 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 omnidirectional source (B&K mod. 4296). Classroom babble was based on ∼20 pupils chatting, whereas impact noise included trampling and jumping noise and the movement of desks and chairs produced by pupils upstairs.

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

The measurement setup of a typical laboratory classroom is shown in Fig. 1. The setting of school B is shown in particular. The source was located at the teacher's position and oriented toward the pupils' seating area. The loudspeaker, for traffic noise emission, was placed outside the school and oriented toward the lab-classroom. In schools A and C, where the lab-room was on the ground floor, the loudspeaker was located on the street pavement at ∼6 m from the façade, whereas in school B, it was placed on the classroom balcony, ∼2 m from the façade. The omnidirectional source was placed in the center of the classroom at 1.3 m above the floor. One receiver was positioned 1 m from the source's mouth and another six 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, whereas the other receivers were placed at ear height of the seated pupils, 1.1 m above the floor.

C. Test administration and measurement procedure

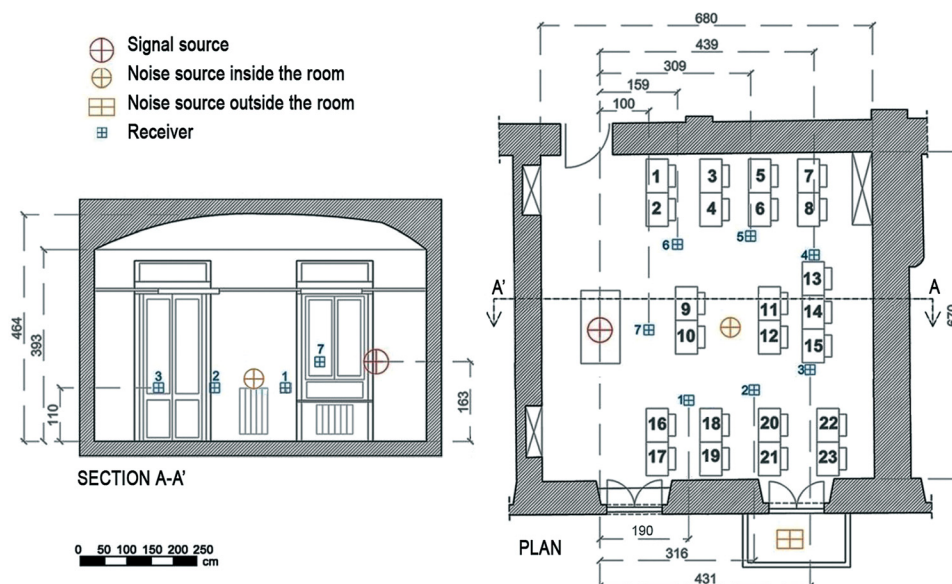Each class spent ∼45 min in the laboratory classroom. After a brief explanation of the experiment and a period in

FIG. 1. (Color online) Measurement setup of a typical laboratory classroom in the schools. The setting of school B is shown in particular.

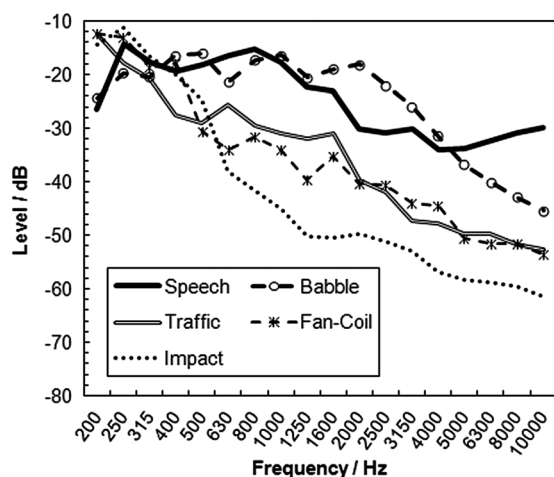Astolfi *et al.*: Intelligibility in primary school classrooms   249

FIG. 2. One-third octave band long-term average spectra of all the noises and the special sentence for the speech signal measurement recorded in school B(b) at 1 m from the B&K 4128 HATS. All the spectra were edited on equal overall sound pressure level and the calibration level is arbitrary.

which the pupils filled in their data sheets, each child was given a parcel of eight sheets, each corresponding to a different test.

With the pupils sitting quiet, the impulse response was measured at the seven points by means of an exponential sweep signal[16] 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. The special sentence for speech level measurement 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, it was checked that the speech level at each measurement location exceeded 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[1] for a single class and it was changed by acting

on the software gain. It was set between 56 and 64 dB(A) for schools A and B(a), between 58 and 61 dB(A) for school B(b) and between 55 and 69 dB(A) for school C.

In order to obtain the exact noise level at each measurement position, the noise sample used for the test was recorded without speech, after the test had been administered. The same recorded sample was used in the tests for each noise. The ambient noise was recorded with the pupils sitting quietly, and there was no significant noise in the classrooms tested in schools A and B, where the level was no higher than 45 dB(A). The situation was different in school C, because of the low sound insulation of the internal partitions. However, the ambient noise level was no higher than 53 dB(A).

As examples, Fig. 2 shows the one-third octave band long-term average spectra of all the noises and the special sentence recorded in school B(b) at 1 m from the B&K 4128 HATS, whereas Fig. 3 shows the relative overall level pattern vs time of the noises and one test signal recorded in dead conditions.

The noise spectra are typical of these kinds of noise, with higher sound energy at medium and high frequencies for babble noise than for impact, fan-coil, and traffic noise. As far as the temporal pattern is concerned, fan-coil noise is a stationary noise, whereas the other three kinds of noise are fluctuating noise: Babble is a fast-fluctuating noise with deep fluctuations, traffic is a slowly fluctuating noise with deep fluctuations and impact noise is fast-fluctuating with shallow fluctuations. The difference between the low and high grade statistical levels (e.g., L_{A10}–L_{A90}) gives an indication of the stationarity of the noise,[17] as the difference is very low for stationary noise, while it becomes higher for noise with deep fluctuation. The statistical level difference is higher for babble and traffic noise, with values of 9.0 and 8.8 dB, respectively, and decreases to 4.8 dB for impact noise and to 1.3 dB for fan-coil noise.

Various speech and noise levels were considered in order to cover an S/N(A) range of −20 to +26 dB. Almost the same S/N(A) range was maintained for each noise in
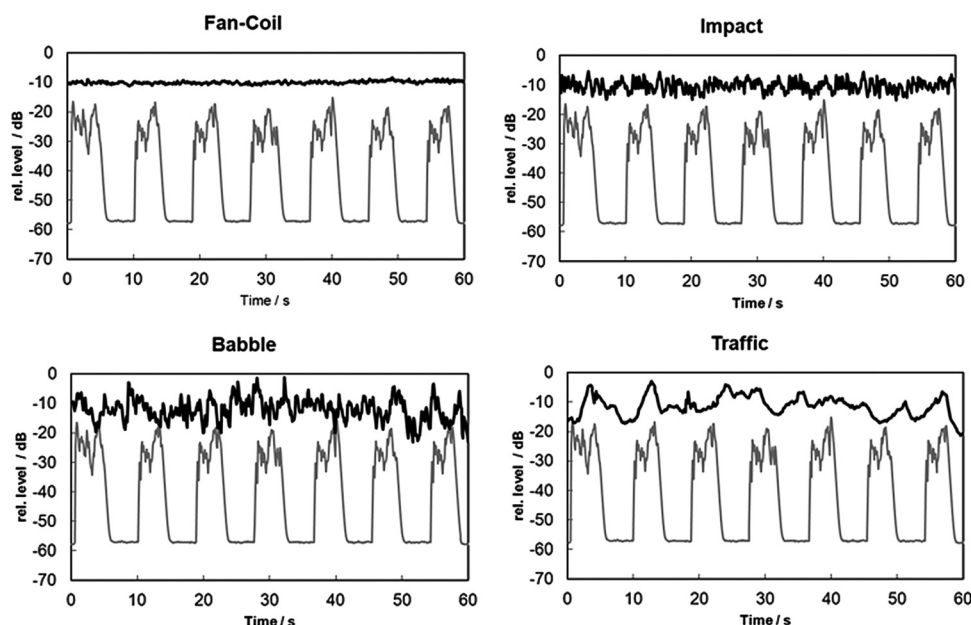

FIG. 3. Relative overall level pattern vs time sampled every 0.1 s of the noises and one test signal recorded in dead conditions. All the noises have the same overall equivalent A-weighted sound pressure level and the test signal has the same level in each subgraph. The black lines represent the noise pattern and the gray lines represent the test signal pattern.

each school, and an overall variation of 6 to 26 dB was determined for ambient noise, -16 to $+18$ dB for traffic noise, -15 to $+20$ dB for babble noise, -15 to $+10$ dB for fan-coil noise, and -20 to $+5$ dB for impact noise.

The STI values were obtained from the squared impulse response and speech-to-noise level difference for each of seven octave bands (center frequencies 125–8000 Hz),[4] using Aurora 4.2.

IV. SPEECH RECOGNITION SCORE MODELS

A. Preliminary analyses

1. Average IS

The speech intelligibility of each pupil was expressed as the percentage of the words understood correctly, with no correction applied for the *a priori* probability of 50% correct responses as a consequence of the two-alternative-choice procedure.[3]

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 correlate the objective parameters to the speech intelligibility scores. The IS of each of these groups was obtained by averaging the answers of the pupils sitting around the same measurement position. The standard deviation of IS of the pupils close to a measurement point follows a similar trend for all the four grades. It increases at the lowest STI values, as expected, as the conditions become more difficult, and varies from $\sim$2.5% for an STI equal to 0.8 to $\sim$8% for an STI equal to 0.1.

There were no significant differences between the non-mother-tongue speaking children and the Italian children, whereas those with hearing or learning disabilities, previously identified by the teachers, were excluded from the analyses. The possible influence of context-sensitive words,[18] i.e., Italian words that vary in orthographic transcription and pronunciation, was also excluded as these words were recognized as consistent words by the pupils of all the grades to the same extent.

2. Room acoustic parameters

The room acoustic parameters[16] were obtained from the impulse response measurements in the occupied laboratory classrooms in which the number of pupils varied between 15 and 20. The average occupied reverberation time, T30, the early decay time (EDT) and the clarity (C50) for combined 500 Hz and 1 kHz octave bands, were determined for each measurement point for at least two different classes in each school.

In order to compare the different schools, the mean spatial values over the six microphone positions representative of the students' seats were calculated, excluding position n. 7, at 1 m from the source mouth. The classrooms were chosen in order to cover a wide range of T30: 0.37 s (s.d. = 0.01) in school B(b), which is representative of the lowest value likely to be found in a classroom; 0.74 s (s.d. = 0.01) in school A and 0.87 s (s.d. = 0.02) in school C,

values that could occur more frequently in classrooms; 1.55 s (s.d. = 0.04) in school B(a), a value that is easily encountered in old Italian school buildings. The EDT values also indicate similar results:[6] 0.30 s (s.d. = 0.04) in B(b), 0.65 s (s.d. = 0.03) in A, 0.85 s (s.d. = 0.02) in C, and 1.54 s (s.d. = 0.04) in B(a).

The clarity shows optimal positive values for speech intelligibility in all but the B(a) school, where negative values are detected for all of the six measurement positions. As expected, the higher the T30 the lower the C50.[6,10] The mean spatial C50 values were the following: 10.67 dB (s.d. = 0.74) for B(b), 4.84 dB (s.d. = 0.62) for A, 1.42 dB (s.d. = 0.61) for C, and -1.68 dB (s.d. = 0.67) for B(a).

The standard deviations of the mean spatial values are equal to or lower than the just noticeable difference (JND) for the correspondent parameters,[16] and therefore demonstrate rather uniform spatial behavior. As far as clarity is concerned, the limited length of the rooms and the reflections from the rear and lateral walls make the values very similar between the desks. No significant differences, equal or lower than the JND of 1 dB, were detected among positions in the schools, even though slightly higher values were measured in the microphone positions correspondent to the first rows of desks (n. 1 and n. 6 in Fig. 1) in A, B(b), and C schools. A maximum difference of $\sim$1 dB between the microphone positions close to the first row of the desks compared to those of the other rows has been noticed in school A, where the classroom, unlike the other laboratory classrooms, has a longer rectangular plan. The same difference of $\sim$1 dB has been found between the microphone positions close to the windows and those close to the absorbent walls in school B(b).

3. Variability in the STI measurements

The STI was obtained from the measurements of the impulse response and the speech signal and noise levels,[4] filtered in octave bands. Since some of the classes involved were very similar, as far as the age and number of pupils are concerned, the classes were divided into homogeneous groups, and the STI values were determined using the audio-records obtained for only one class from each group.

The variability in the STI measurements was evaluated on the basis of this approximation in school B(a), where the impulse responses and speech and noise levels were measured in the lab-classroom for four classes in the school at six out of seven points (excluding the one at 1 m from the source mouth). The repeatability of the impulse response within the same class, the influence of the different grades and numbers of pupils and the vocal effort were tested for each noise. The mean spatial values over the six points of the STI standard deviation for each comparison are lower than the just noticeable difference of 0.05 suggested in literature.[19]

The possibility of different architectural elements, such as the shape of the ceiling or the plan of the room, having an effect on the speech intelligibility scores was excluded since no large variations in room acoustic conditions were detected among the various measurement points for each school.

J. Acoust. Soc. Am., Vol. 131, No. 1, January 2012

Astolfi *et al.*: Intelligibility in primary school classrooms   251

B. Validation methods of the regression models

The speech recognition scores were first examined as a function of STI, as this was the key independent variable in these experiments. The proposed IS vs STI regression model is based on a logarithmic function, according to the following equation:

$$IS = a \ln(STI) + b, \qquad (1)$$

where a and b are the regression coefficients. The logarithmic model has proved to be the best fitting model in relation to the polynomials from grades 1 to 4.

The uncertainty in the adopted model, represented as uncertainty curves, has been analyzed in agreement with the Guide to the expression of Uncertainty in Measurement, GUM,[20] for which the "expanded uncertainty" U associated with an experimental result is obtained by multiplying the "combined standard uncertainty," u_c, by the "coverage factor," k, using the following formula:

$$U = k u_c(y). \qquad (2)$$

The coverage factor is calculated as the Student-t value for a conventional risk of error α of 1% and a number of degrees of freedom, ν, corresponding to $n - 2$, where n is the number of data used. The combined standard uncertainty $u_c(y)$ is the positive square root of the combined variance $u_c^2(y)$, which is given by

$$u_c^2(y) = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)_{x_i}^2 u^2(x_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j), \qquad (3)$$

where y is the estimate of the measurand Y, x_i are the N input estimates, f is the relationship between y and x_i, $u(x_i)$ is a standard uncertainty of the input estimates, and $u(x_i, x_j)$ is the estimated covariance associated with two different inputs, x_i and x_j, taken in pairs. In the case of the model proposed in Eq. (1), y corresponds to IS and x_i are the regression coefficients a and b, so that the combined standard uncertainty is represented as follows:

$$u_c^2(y) = \left(\frac{\partial f}{\partial a}\right)^2 u^2(a) + \left(\frac{\partial f}{\partial b}\right)^2 u^2(b) + 2 \frac{\partial f}{\partial a} \frac{\partial f}{\partial b} \text{cov}(a,b)$$
$$= u^2(a)[\ln(STI)]^2 + u^2(b) + 2 \ln(STI)\text{cov}(a,b). \qquad (4)$$

In order to quantify the goodness of the model compared with others, a robustness coefficient[21] was considered for the regression. The robustness coefficient r is calculated for a and b as the ratio between the values of the a and b coefficients and their uncertainty, according to the following equation:

$$r = \frac{\text{coeff.}}{k s(\text{coeff.})}, \qquad (5)$$

where coeff. is alternatively a or b, $s(\text{coeff.})$ is the a and b standard deviations, and k is the "coverage factor." When r

is higher than 1, the randomness of the regression coefficients can be considered acceptable.

The Normalized Error concept[21] was adopted for the compatibility analysis between two regression models obtained in different conditions. This is useful for comparisons of measurement results produced at the same hierarchical level, i.e., where no value can be taken as the reference value. It is necessary to understand whether the difference in the compared models is due to an effective difference between the evaluated phenomena or to systematic effects, rather than to random effects. The Normalized Error, E_N, is calculated as the ratio between the absolute value of the difference of two states in the evaluated phenomenon and the relative expanded uncertainty of the difference,[20] according to the following formula:

$$E_N = \frac{|\text{coeff}_1 - \text{coeff}_2|}{U} = \frac{|\text{coeff}_1 - \text{coeff}_2|}{k\sqrt{s(\text{coeff}_1)^2 + s(\text{coeff}_2)^2}}, \qquad (6)$$

where coeff_1 and coeff_2 represent the two states in the same regression coefficient (a or b, alternatively) evaluated from data obtained in different conditions.

This analysis can be considered a particular kind of hypothesis test. If the E_N value is higher than unity, the difference between the two values, coeff_1 and coeff_2, is higher than its uncertainty, therefore the difference is not merely due to random effects and the two results can be considered incompatible. On the contrary, if E_N is lower than unity, the difference could be due to random effects and there is no reason to refuse compatibility. Values lower than unity do not mean that real differences or systematic effects are not present, but that random effects cover their presence.

C. Comparison between the schools

Figure 4 shows the average IS scores considering all the intruding noises vs the measured STI for the four schools, together with the logarithmic regressions lines, whereas Table I lists the regression parameters related to the
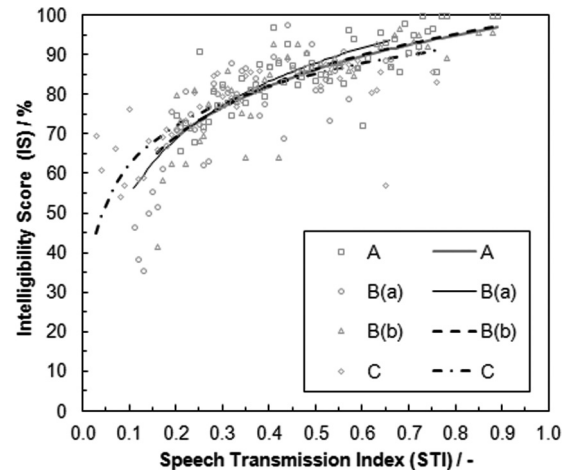

FIG. 4. Logarithmic regression curves of the average speech intelligibility scores, IS, plotted vs Speech Transmission Index, STI, for the A, B(a), B(b), and C schools, considering all the pupils and all the kinds of noise. For a better legibility, punctual values were obtained by averaging the scores over a 0.01 STI interval.

Astolfi *et al.*: Intelligibility in primary school classrooms

TABLE I. Regression parameters related to the regression curves plotted in Figs. 4 and 8.

	Regr. coeff.		s.d.		Degrees of freedom	Cov. fact.	Robustness coeff.		Det. coeff.
	a	b	$s(a)$	$s(b)$	ν	k	r_a	r_b	R^2
School A	18.7	99.1	1.4	1.2	621	2.58	5.2	30.9	0.22
School B(a)	21.0	102.4	1.9	2.1	574	2.58	4.3	18.9	0.18
School B(b)	18.7	99.4	1.4	1.4	584	2.58	5.0	26.8	0.23
School C	14.3	95.1	0.8	1.1	997	2.58	6.8	33.5	0.23
Grade 2	24.5	99.9	1.8	1.8	439	2.59	5.4	21.6	0.31
Grades 3–5	16.8	99.4	1.0	1.0	1342	2.58	6.8	40.4	0.19

curves, including the robustness coefficients, which are all much higher than unity. The expanded uncertainty, U, is lower than 5% over the whole STI interval. The regressions were performed considering the entire data sample, but in order to obtain a better legibility, the represented data were averaged over a 0.01 STI interval and the uncertainty curves were not plotted.

From a visual inspection of Fig. 4, it can be observed that all the models seem similar, but school C has a slightly different slope. The incompatibility of school C was confirmed by applying the Normalized Error method related to the a and b coefficients of the regression models for the schools taken in pairs, the results of which are listed in Table II. The models result to be incompatible, even though only one value of E_N regarding a or b is higher than unity.

The incompatibility of school C could be due to the fact that this is a different type of building from the others, with less sound insulation between the rooms, and the largest STI range during the tests. Because of this, school C was excluded from some of the subsequent analyses.

D. Comparison with other models

Figure 5 shows the common model of the average speech intelligibility scores vs STI that was used in the A, B(a), and B(b) schools, together with the uncertainty curves

TABLE II. Normalized errors, E_N, calculated for regression coefficients a and b related to the trend of speech intelligibility scores, IS, vs Speech Transmission Index, STI, for each school, for each grade and for each noise, taken in pairs. The bold numbers represent higher values than unity, i.e., incompatibilities.

Schools	A B(a)	b B(a)	a B(b)	b B(b)	a C	b C
A	0.38	0.52	0.01	0.06	**1.04**	0.94
B(a)			0.37	0.46	**1.26**	**1.20**
B(b)					**1.03**	0.93
Grades	3	3	4	4	5	5
2	**1.08**	0.30	**1.12**	0.12	**1.48**	0.00
3			0.07	0.44	0.46	0.32
4					0.38	0.12
Noises	Babble	Babble	Fan-coil	Fan-coil	Impact	Impact
Traffic[a]	0.52	**2.98**	**1.53**	**2.70**	0.76	**1.28**
Babble			**1.82**	0.04	**1.08**	**4.05**
Fan-coil					0.55	**3.71**

[a]School C was excluded from the analyses concerning traffic noise as it was incompatible with the other schools.

obtained according to Eq. (2). The expanded uncertainty U is lower than 2% over the whole STI interval. The best-fit regression curve is compared in the figure with another two curves taken from the literature, for adults and pupils, respectively.

The curve for adults is based on a "Modified Rhyme Test" presented via earphones, masked by a random noise, whose spectral shape was that of long-term averaged speech.[3,22] The test consisted of lists of 300 monosyllabic words in six 50-word lists, where each of the 50 words is chosen from six alternatives, administered to eight subjects. The model for primary school pupils is that by Prodi et al.,[11] which uses the same speech material and test procedure as the current study in auralized classrooms, with noise from occupants and babble from an adjacent corridor. The IS from 80 pupils was obtained using a methodology which controls the effect of guessing in the data collection, giving a weight "+1" to a correct word, a weight "−1" to a wrong word and "−0.5" to the option "none of the two."

A subset of the current results was also compared with those from two in-field studies carried out in similar acoustic conditions.[5,7] Figure 6 shows the quadratic regression curves and the uncertainty curves of IS vs S/N(A) in the presence of babble noise in the B(b) school, compared with the quadratic regression curve obtained from the average values in Table IV of Ref. 7, for grade 3 and 6 pupils in the presence of noise from other classrooms. The IS from 840 pupils was obtained
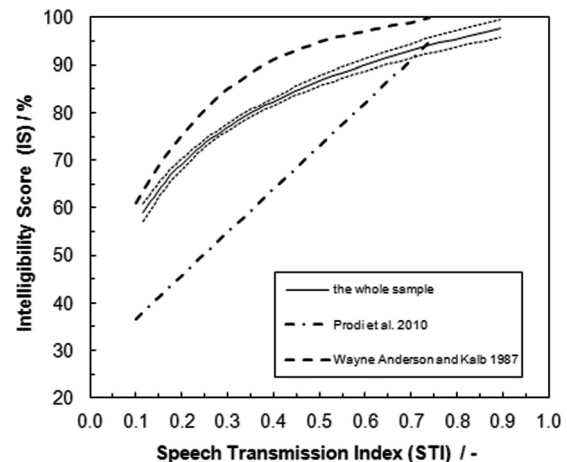

FIG. 5. Best-fit regression curve of the common model of the average speech intelligibility scores, IS, vs Speech Transmission Index, STI, considering schools A, B(a), and B(b) with the uncertainty curves, compared to the regression curves obtained in the laboratories for children (Ref. 1) and for adults (Ref. 22).
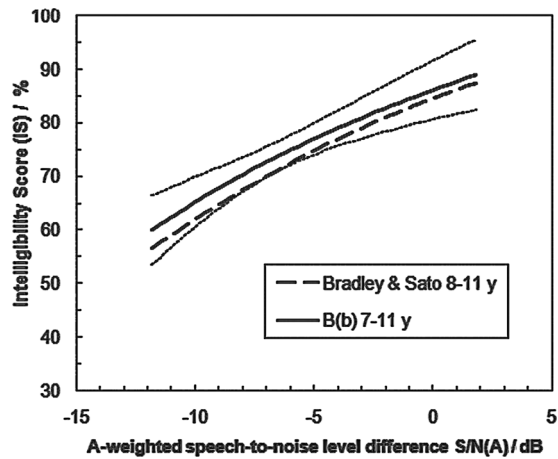
FIG. 6. Best-fit regression and uncertainty bands of the average speech intelligibility scores, IS, vs A-weighted speech-to-noise level difference, S/N(A), with babble noise in school B(b) for children aged between 7 and 11 and regression curves obtained from Bradley and Sato (Ref. 7) for children aged between 8 and 11 with environmental noise.

by means of a WIPI test in rhyme based on 100 PB-simple nouns, in four 25-word lists, where each of the 25 words is chosen by placing a sticker on one of six pictures.

Figure 7 shows the quadratic regression curves and the uncertainty curves of IS vs S/N(A) in the presence of traffic noise in the A, B(a), and B(b) schools, compared with the quadratic regression curve obtained from the values in Table III of Ref. 5, for pupils aged between 8 and 15 in the presence of traffic noise. The IS from 500 pupils was obtained by means of a Fairbanks rhyme test based on 200 meaningful CVC-PB words in four 50-word lists, where each of the 50 words is chosen from four alternatives.

The speech intelligibility regression curve for adults shown in Fig. 5 is higher than the curve obtained in the current study for pupils, for the whole STI range. The two curves show similar IS for lower STI, but adults need lower STI values than pupils to achieve the near-ideal conditions, which corresponds to an IS of 95% of the correct scores.[7]
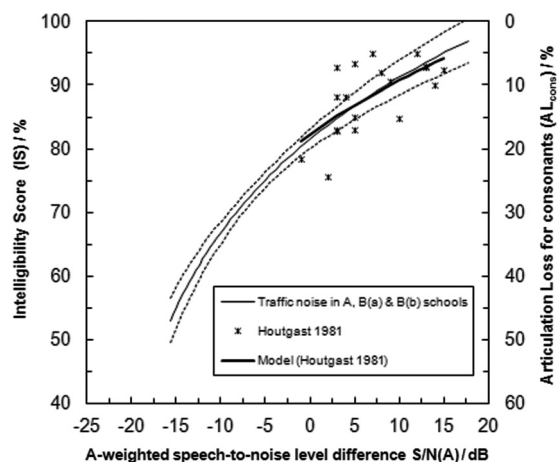
The near-ideal conditions are achievable with an STI of ~0.50 for adults and 0.79 for pupils. This is as might be expected because of the age difference. The curve for pupils presented by Prodi et al.,[11] which is also given in Fig. 5, instead shows a different slope and lower intelligibility scores over the whole STI range, whereas the agreement is very good with the two in-field studies shown in Figs. 6 and 7, whose regression curves fall between the uncertainty curves of the currently compared models.

E. Comparison between the school grades

Table II shows the Normalized Errors calculated for the regression coefficients a and b of the IS vs STI models for each grade, considering all the classes and the intruding noises in schools A, B(a), and B(b). Higher values than 1 indicate incompatibility for grade 2 compared with the other grades, which are instead compatible with each other. Figure 8 shows a plot of the regression curves of IS vs STI for grade 2 and grades 3–5 together and the near-ideal intelligibility conditions corresponding to an IS of 95% of the correct scores,[7] whereas Table I lists the regression parameters related to the curves that include the robustness coefficients, all of which are much higher than unity.

The regression curves indicate that for equal STI values, IS increases from grade 2 to the other grades. The near-ideal conditions are readily achievable with an STI of ~0.82 for grade 2 pupils, and of 0.77 for grade 3–5 pupils, whereas 97% correct scores are achievable with an STI of ~0.9 for all the grades. The difference in IS increases with a deterioration in the acoustic conditions, with a maximum gap between grade 2 and the other grades of 14%, in correspondence to an STI of 0.16.

The gap between grade 2 and the higher grades highlighted in Fig. 8 cannot be imputable to the different ability of the pupils to recognize the simple meaningful bisyllabic words in the administered tests, as, in ideal acoustic conditions, the grade 2 pupils achieved the same score as the higher grades.

FIG. 7. Best-fit regression and uncertainty bands of the average speech intelligibility scores, IS, vs A-weighted speech-to-noise level difference, S/N(A), with traffic noise in the A, B(a), and B(b) schools considering all the pupils, compared with the quadratic regression curve obtained by Houtgast (Ref. 5) for children aged between 8 and 15 with traffic noise.
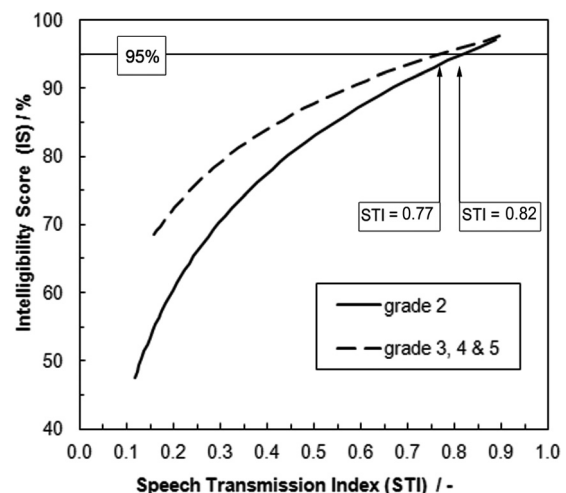

FIG. 8. Regression curves of the average speech intelligibility scores, IS, vs Speech Transmission Index, STI, for grade 2 and grades 3–5 together, considering all the classes and the noises of schools A, B(a), and B(b), and the near-ideal STI value conditions corresponding to an IS of 95% of correct scores.

Astolfi et al.: Intelligibility in primary school classrooms

V. EFFECT OF DIFFERENT KINDS OF NOISE AND REVERBERATION

In order to find a common model of the average IS vs STI for the different types of noise, the compatibility of the results, split into different types of noise, was investigated in the schools. Normalized Errors were calculated for the regression coefficients a and b related to the trend of IS vs STI for each type of noise in the A, B(a), B(b), and C schools, taken in pairs. The only incompatibility was detected for traffic noise in school C, whose E_N values, related to the regression coefficients a and b were 1.04 and 0.64 for school B(a) and 1.56 and 0.96 for school B(b).

Figure 9 shows the regression curves corresponding to the average IS vs STI for the different types of noise, considering the full sample of pupils, with the exception of the results obtained for traffic noise in school C. The E_N values reported in Table II show that all the models are different from each other. A lower slope was shown for fan-coil and impact noise, whereas the highest was for babble noise. Traffic resulted to be the most interfering noise, as it scored lower than all the other types of noise.

The effect of reverberation in the classroom has been highlighted by plotting the speech intelligibility scores vs S/N(A). The analysis was only conducted in school B before and after the acoustical treatment, in order to test the effect of reverberation on speech intelligibility using the same sample of pupils. Traffic and babble noises were chosen for the comparison as representative noises outside and inside classrooms, respectively.

Figure 10 shows the regression curves and uncertainty bands related to the average IS vs S/N(A) plot in schools B(a) and B(b) in the presence of traffic and babble noise. The traffic noise curve is incompatible when analyzed in pairs before and after the intervention. This results from the E_N values of the regression coefficients a and b, which are 0.51 and 1.52. The babble noise curves instead show compatibility, with lower E_N values than unity.

As far as traffic noise is concerned, a decrease in the average reverberation time from 1.6 to 0.4 s after soundproofing determined an IS increase on equal S/N(A), which
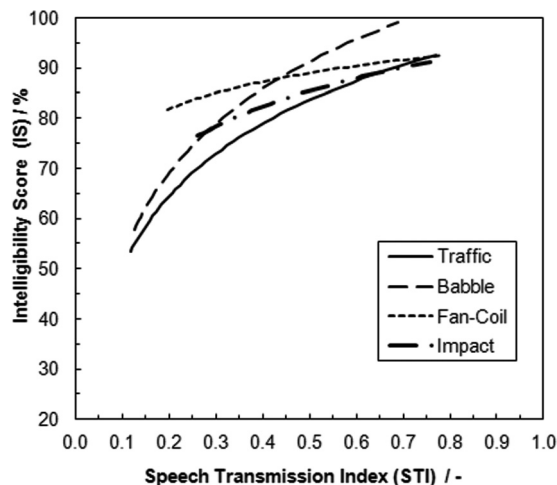
FIG. 9. Regression curves corresponding to the average speech intelligibility scores, IS, vs Speech Transmission Index, STI, for the different types of noise, considering the full sample of pupils, with the exception of the results obtained for traffic noise in school C.
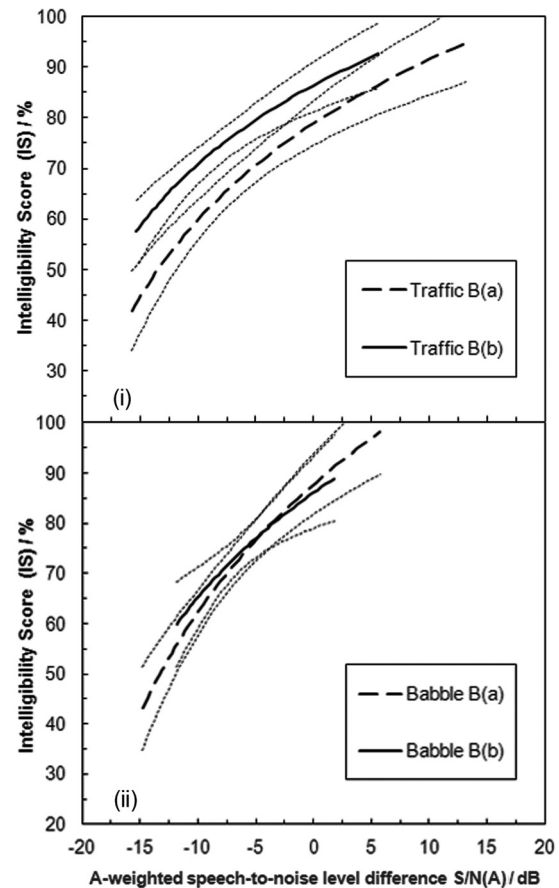
FIG. 10. Regression curves and uncertainty bands related to the average speech intelligibility scores, IS, vs A-weighted speech-to-noise level difference, S/N(A), for the pupils in school B before, B(a), and after, B(b), the acoustic intervention considering traffic noise (i) and babble noise (ii).

varied from 13% to 6%, for the S/N(A) range of −15 to +6 dB, respectively. The same decrease in reverberation time for the babble noise, whose source was located in the middle of the classroom, leads to a negligible variation in IS over the S/N(A) range of −12 to +2 dB.

VI. DISCUSSION

A. Comparison between the different studies

A comparison of the speech intelligibility curves for pupils with the curves from other studies has revealed some similarities, but also some differences. Even though the sample size and the age of the pupils, the language, the test typology, and the word lists were not the same, good agreement was found between IS vs S/N(A) curves when in-field rhyme closed-set tests based on meaningful words were used, with comparable acoustic conditions.[5,7]

On the contrary, the IS vs STI curve of Prodi et al.[11] was not confirmed in the present study, even though the same language, test, age of respondents and acoustic conditions were applied. The curve in Ref. 11 shows a different slope and lower intelligibility scores from the present study, whereas the main differences could be imputable to the sample size and speech intelligibility computational method.

When the IS vs STI curve for the pupils in the present study and the one for adults in Ref. 22 are compared, the

same regression trend is shown, with a higher better performance for adults, as expected. Although both studies are based on a rhyme closed-set test, differences still remain concerning the sample size, language, test typology, and word lists, type of noise, and test setting.

On the basis of this series of comparisons, it seems necessary to investigate, in future works, the systematic effects related to the type of test, language, sample, acoustic conditions and indexes, as well as test-setting, whether in-field or in the laboratory.

B. The influence of different kinds of noise on speech intelligibility

The different IS vs STI regression curves corresponding to the different noises shown in Fig. 9, can mainly be imputable to the different temporal patterns of the noises, as the spectra are taken into account in the STI computation.[4]

The babble noise scores higher due to the fast and deep fluctuations in level, which allow word perception in the temporal gaps. This positive effect is reduced with reverberation, which partially fills the gaps.[22] Traffic noise is characterized by slow and deep fluctuations with high persistent noise levels in correspondence to most of the target words, and this determines the lowest scores. Impact noise is characterized by fast but shallow fluctuations which do not lead to any benefit in speech perception compared with babble noise. The fan-coil model generally shows higher IS than the other models, apart from the babble noise model in the higher STI range, due to the lower levels in correspondence to the target words compared with the other noises.

C. The influence of reverberation on speech intelligibility

The discussion on the influence of reverberation on speech intelligibility is related to the type of noise, whether traffic or babble, and to the position of the noise and the speech sources with respect to the pupils. As far as the first aspect is concerned, the reverberation affects the spectra and the temporal patterns of noises in a different way, and for this reason traffic and babble were considered separately. As far as the second aspect is concerned, Fig. 1 shows that the traffic noise source is far from the pupils, whereas the babble noise source is near and equidistant to most of them. Further, the traffic source is farther or equally distant from most of the pupils than the speech source, whereas the babble source is closer than the speech source for all of them.

The lowering of RT from 1.6 s to 0.4 s involved a 10% quite constant IS increase over the S/N(A) range of −15 to +6 dB in the case of traffic noise in this study. No significant improvement in IS has followed the same lowering of RT over a comparable S/N(A) range in the case of babble noise. As clearly described in Ref. 13, these results can be explained physically in an approximate way as follows.

In the case of traffic noise, for pupils at the same distance from the noise and the speech source, the lowering of RT equally decreases the speech and the noise levels. This results in a constant S/N(A) and leads to a lower IS in the case of higher RT, which is only due to reverberation. This behavior was also confirmed in a study in the laboratory with pupils by Yang and Bradley.[10]

For pupils farther from the traffic noise source than the speech source, the noise levels decrease more with reverberation than the speech levels, and this results in increased S/N(A), which leads to a higher IS. In both conditions, reducing RT leads to an IS increase, which in turn points out an optimal reverberation time of zero, a result that is in agreement with Hodgson and Nosal.[13]

In the case of babble noise, the pupils are closer to the noise source than the speech source and the lowering in RT decreases the speech levels and, to a lesser extent, the noise levels for which the contribution of the direct sound dominates. This results in an S/N(A) and IS decrease. On the other hand, for equal S/N(A), the lowering of RT does not produce any significant IS changes: It seems that the RT decrease, whose effect is more on speech than on the babble signal, does not affect intelligibility to any great extent, probably due to the temporal pattern of babble noise, which allows word perception in the temporal gaps either with high or low reverberation.

Although similar findings have been encountered in the cited works, the optimal RT in a classroom still needs further investigations, above all through acoustic simulations and then through experimental studies performed with children, in the laboratory and in-field, considering a wide speech and noise source position scenario.

VII. CONCLUSIONS

This study provides data that describe the ability of primary school pupils to understand speech with noise and reverberation in real classrooms. Four laboratory classrooms in three schools in Italy, with different reverberation times and realistic traffic, speech babble, fan-coil, and impact noise, were used for the speech intelligibility tests on 983 pupils from grades 2–5 (nominally 7–10 years old).

The intelligibility scores were then correlated with the Speech Transmission Index corresponding to seven microphone positions evenly distributed in the lab-classrooms.

The following main findings emerged from the experimental investigations:

(a) Two logarithmic regression curves of IS vs STI for grade 2 and grades 3–5 together were obtained; an IS of 95% is readily achievable with an STI of ∼0.82 for grade 2 pupils and 0.77 for grade 3–5 pupils. A maximum gap between grade 2 and the other grades of 14% corresponds to a STI of 0.16.

(b) different types of noise result in significantly different IS vs STI trends: Traffic noise in particular is the most interfering noise;

(c) in the case of traffic noise source outside the building, a lower RT from 1.6 s to 0.4 s results in a 10% quite constant IS increase over the S/N(A) range of −15 to +6 dB. In the case of the babble noise, whose source was located in the middle of the classroom, the same decrease in reverberation time leads to a negligible variation in IS over a similar S/N(A) range.

Although some of the main open questions regarding speech intelligibility in classrooms have been addressed in this study, further research is still needed to extensively investigate the following topics: The difference between laboratory and in-field experiments, the influence of different types of intelligibility test and procedures, and the determination of the optimal reverberation time on the basis of the type of noise source and its position in the classroom.

ACKNOWLEDGMENTS

[1]ISO9921, *Ergonomics—Assessment of Speech Communication* (International Organization for Standardization, Genève, 2003).

[2]C. Flexer, "The impact of classroom acoustics: Listening, learning, and literacy," Semin. Hear. **25** (2), 131–140 (2004).

[3]ISO TR4870, *Acoustics—the Construction and Calibration of Speech Intelligibility Tests* (International Organization for Standardization, Genève, 1991).

[4]EN 60268-16, *Objective Rating of Speech Intelligibility by Speech Transmission Index* (European Committee for Standardization, Brussels, 2003).

[5]T. Houtgast, "The effect of ambient noise on speech intelligibility in classrooms," Appl. Acoust. **14**, 15–25 (1981).

[6]J. S. Bradley, "Speech intelligibility studies in classrooms," J. Acoust. Soc. Am. **80**(3), 846–854 (1986).

[7]J. S. Bradley and H. Sato, "The intelligibility of speech in elementary school classrooms," J. Acoust. Soc. Am. **123**(4), 2078–2086 (2008).

[8]L. L. Elliott, "Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability," J. Acoust. Soc. Am. **66**(3), 651–653 (1979).

[9]A. Neuman and I. Hochberg, "Children's perception of speech in reverberation," J. Acoust. Soc. Am. **73**, 2145–2149 (1983).

[10]W. Yang and J. S. Bradley, "Effects of room acoustics on the intelligibility of speech in classrooms for young children," J. Acoust. Soc. Am. **125**(2), 922–933 (2009).

[11]N. Prodi, C. Visentin, and A. Farnetani, "Intelligibility, listening difficulty and listening efficiency in auralized classrooms," J. Acoust. Soc. Am. **128**(1), 172–181 (2010).

[12]V. M. A. Peutz, "Articulation loss of consonants as a criterion for speech transmission in a room," J. Aud. Eng. Soc. **19**(11), 915–919 (1971).

[13]M. Hodgson and E. M. Nosal, "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms," J. Acoust. Soc. Am. **111**(2), 931–939 (2002).

[14]W. Yang and M. Hodgson, "Auralization study of optimum reverberation times for speech intelligibility for normal and hearing-impaired listeners in classrooms with diffuse sound fields," J. Acoust. Soc. Am. **120**(2), 801–807 (2006).

[15]P. Bonaventura, A. Paoloni, F. Canavesio, and P. Usai, "Realizzazione di un test di intelligibilità per la lingua italiana (Development of a diagnostic intelligibility test in the Italian language)," Rapporto Tecnico 3C1286, Fondazione Ugo Bordoni, Roma (1986).

[16]ISO 3382-1, *Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces* (International Organization for Standardization, Genève, 2009).

[17]I. D. Griffiths and F. J. Langdon, "Subjective response to road traffic noise," J. Sound Vib. **8**, 16–32 (1968).

[18]L. Barca, A. W. Ellis, and C. Burani, "Context-sensitive rules and word naming in Italian children," Read. Writ. **20**, 495–509 (2007).

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

[20]JCGM100, *Evaluation of measurement data—guide to the expression of uncertainty in measurement* (Joint Committee for Guides in Metrology, 2008), http://www.iso.org/sites/JCGM/GUM-JCGM100.htm (last viewed May 5 2011).

[21]ISO/IEC Guide 43-1, *Proficiency Testing by Interlaboratory Comparisons. Part1: Development and Operation of Proficiency Testing Schemes* (International Organization for Standardization, Genève, 1997).

[22]B. Wayne Anderson and J. T. Kalb, "English verification of the STI method for estimating speech intelligibility of a communications channel," J. Acoust. Soc. Am. **81**(6), 1982–1985 (1987).

[23]E. L. J. George, J. M. Festen, and T. Houtgast, "The combined effects of reverberation and nonstationary noise on sentence intelligibility," J. Acoust. Soc. Am. **124**(2), 1269–1277 (2008).