Duration of voicing and silence periods of continuous speech in different acoustic environments

Duration of voicing and silence periods of continuous speech in different acoustic environments

Arianna Astolfi[a)]
Politecnico di Torino, Department of Energy, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

Alessio Carullo
Politecnico di Torino, Department of Electronics and Telecommunications, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

Lorenzo Pavese and Giuseppina Emma Puglisi
Politecnico di Torino, Department of Energy, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

(Received 25 July 2014; revised 19 December 2014; accepted 31 December 2014)

This work deals with the duration of voicing and silence periods of continuous speech in rooms with very different reverberation times (RTs). Measurements were conducted using the Ambulatory Phonation Monitoring (APM) 3200 (Kaypentax, Montvale, NJ) and Voice-Care devices (developed at the Politecnico di Torino, Italy), both of which have a contact microphone placed on the base of the neck to detect skin vibrations during phonation. Six university professors and 22 university students made short laboratory monologs in which they explained something that they knew well to a listener 6 m away. Seven students also described a map with the intention of correctly explaining directions to a listener who drew the path on a blank chart. Longer speech samples were made by 25 primary school teachers in classrooms. A tendency to increase the voicing periods as the RT increased was on average observed for the university professors, the school teachers, and the university students who described a map. These students also showed longer silence periods than the students who made short monologs. The recognized trends concerned voice professionals or subjects who were highly motivated to make themselves understood in a perturbed speaking situation. Nonparametric statistical tests, which were applied to detect the differences in distributions of voicing and silence periods, have basically supported the findings. © *2015 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4906259]

[NX] Pages: 565–579

I. INTRODUCTION

Speakers continually adapt the acoustic-characteristics of their speech in response to a difficult communication context in order to improve speech intelligibility. The type of interlocutor and the environment are the main factors of influence.[1,2] Among the strategies adopted to counteract challenging acoustic conditions, a slower speech rate increases the reception of phonetic information and decreases the cognitive effort of the listener. A speech rate reduction is mainly obtained by inserting more frequent and longer pauses in the speech stream and, to a lesser extent, by speech segment lengthening.[2]

A speaking rate decrease is typical of the so-called "clear speech."[1,3,4] Clear speech can be produced in response to instructions to speak clearly,[5] as in the case of clear, read speech, but also spontaneously in order to adapt to a perturbed communication situation or to help a listener with reduced comprehension capability.[2,6] Clear speech elicited through instructions shows more extreme changes in speech characteristics than speech produced in spontaneous interactions, when actual challenging conditions are experienced.[1] Moreover, the characteristics of spontaneous clear speech vary according to the needs of the interlocutor and, even in a perturbed communication situation, the proportion of clarification strategies diminishes compared to clear, read speech.[1] It also appears that the strategies used by talkers vary according to the adverse listening conditions. There are likely to be individual differences in the strategies used by talkers to clarify their speech in different conditions, as well as in the degree of success they have in achieving effective communication.[1,3]

Both the number and duration of pauses increase in clear speech, compared to conversational speech.[5] "Pauses" can be defined as any period of silence of at least 10 ms, even though the threshold commonly used to define a pause in natural speech is 250 ms. Shorter pauses of 10 ms can be seen as the result of the speaker's attempt to enunciate both word-final and word-initial consonants as clearly as possible, while the longer pauses serve to mark syntactic boundaries or phrases.[1]

Speakers also lengthen individual words in clear speech. Picheny *et al.*,[5] who studied phrases containing a substantial number of monosyllabic words spoken under instruction to produce clear speech, found the average syllable length to be in the 520–590 ms range, that is, almost double that of conversational speech, which was found, in turn, to be in the 250–330 ms range. These values are higher than the average syllable length range for natural speech, which varies from

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

170 ms to 220 ms, as the average syllable length measured for phrases containing a substantial number of monosyllabic words is usually longer than that measured for polysyllabic words.

Speech prosody under instructions to provide clear speech has much in common with Lombard Speech (LS), i.e., speech produced in the presence of noise, and has long been studied; it typically exhibits evidence of word lengthening and the insertion of more and longer pauses.[7,8]

Very few studies have dealt with the influence of room acoustics on speech production and, in particular, on variations in the speaking rate. Black[9] investigated the effect of size and reverberation time (RT) on vocal intensity and speech duration. His study was based on a group of 184 male speakers, who, with a microphone placed 33 cm from their mouths, individually read 12 five-syllable phrases in 8 rooms (23 subjects per room). The rooms were different in volume (4 and 45 m^3) and shape (drum and rectangular) and had two different RTs of 0.8–1.0 s and 0.2–0.3 s. Each subject was instructed to read naturally with the aim of making himself understood by the listener, who was positioned at a distance of 2.5 m. The speech rate was found to be slower in large rooms than in small ones, and in large rooms, the rate was slower in live rooms than in dead ones.

Pelegrín-García *et al.*[10] investigated the effect of the acoustical environment on natural English speech, evoked by means of a map task,[11] conducted with 13 male, nonnative speakers at doubled communication distances (from 1.5 m to 12 m) in the absence of background noise. They considered very different environments, including an anechoic room and a reverberation room with average RTs of 0.04 s and 5.38 s at 500 Hz and 1 kHz ($T_{30,0.5-1\,\mathrm{kHz}}$), respectively. In the case of a communication distance of 6 m, which is typical of a lecturing situation, the phonation time ratio, which is the ratio between the phonation time and the running speech time, was 0.70 for the anechoic room and 0.72 for the reverberation room. The standard deviation of the intersubject variation was estimated as 0.059 for both rooms.

In addition to environmental factors, phonological and phonetic factors can also influence the duration of vocalic segments,[12] but these factors are beyond the scope of the present study. Differences in languages,[13] linguistic issues,[12] and extralinguistic factors, such as the speaker's mood and physical state,[14] can also affect this duration.

Dauer[15] found that the number of syllables per second in continuous natural speech varies greatly from language to language and is ~4.5 syllables/s for English and ~7.3 syllables/s for Italian. The average length of a syllable is ~220 ms in English and ~130 ms in Italian. Another study by Klatt[12] reported that the normal range of conversational speaking rates in English varies from ~4 to 7 syllables/s, which corresponds to an average syllable length in the 140–250 ms range. Excluding pauses of >200 ms, Klatt[12] stated that pauses constitute ~20% of the time during fluent reading, and a good deal more, ~50% of the time, in conversation.

From the linguistic point of view, lexical items that contain more information tend to be longer. Similarly, referring expressions that introduce new information into a discourse are longer than their anaphors. Moreover, words are given longer and more intelligible pronunciation when they occur in contexts that do not predict them and shorter or less intelligible pronunciations when they can be predicted from the context. A word appears to be more susceptible to degradation when it can be identified from the context, whether linguistic or extralinguistic, e.g., shortening and a loss of intelligibility accompany second co-referential mentions in extended discourse, or reference to objects visible to the speaker and listener, or even when informal or close relationships exist between the speaker and listener.[11]

Anger, fear, and sorrow situations tend to produce differences in the temporal characteristics of speech.[14] Some syllables are produced with increased intensity or emphasis; the duration of words uttered in anger is usually longer, but this effect is not so obvious and is not consistent for all voices.

Klatt[12] observed durational patterns in English sentences and argued that considerable interspeaker and intraspeaker variability exists. He stated, in particular, that interspeaker differences may be greater than the differences that can be attributed to contextual constraints. Interspeaker variability was confirmed in the study by Cristal and House,[13] in which it was found that natural reading rates varied sufficiently to allow a separation to be made between a fast group and a slow group on the basis of the total time that elapsed for two specific scripts. These groups on average lasted 77.9 s and 103.8 s, respectively. The average slow reader was 33% slower than the average fast reader. The mean pauses were 574 ms for the fast readers and 728 ms for the slow readers, and the ratio of speech-to-elapsed time ratio was on average 82.5 for the former and 76.4 for the latter. This increase was attributed to the introduction of new pauses (54%), the increased duration of existing pauses (27%), and the increased duration of speech segments (19%).

Klatt[12] also established a just-noticeable difference (JND) for a voice segmental duration of ~25 ms. From the perceptual point of view, systematic changes of about less than one JND are considerably less important than changes that exceed one JND. Since the JND for duration approximately follows Weber's law, this constraint could be reformulated so that only changes of ~20% or more could be used as primary perceptual cues. A minimum JND of 25 ms has been found, but this JND systematically increases by as much as a factor of 4 in certain sentence positions.

From the medical point of view, the duration of voicing and silence periods can be related to vocal fatigue and vocal recovery, respectively. According to Hunter and Titze,[16] vocal overuse is the cause of physiological vocal fatigue, which can be broken down into laryngeal muscle fatigue and laryngeal tissue fatigue. The former results in soreness, discomfort, and/or muscle tension in the neck region, while the latter likely stems from changes or damage to the vocal fold lamina propria caused by vibration exposure, and results in pain or a scratchy voice and/or increased voice breaks, instability, and the inability to produce a soft voice.

Although the primary aim of evaluating vocal fatigue is the quantification of the voicing time, equal importance should be given to the recovery time (or silence time), which can be broken down into long- and short-term recovery.[16,17] Subjective ratings seem to be better at quantifying the effect of long-term recovery than objective metrics, as pointed out by Hunter and Titze,[16] who, by means of perceptual ratings, quantified a full long-term recovery time of 12–18 h after 2 h of oral reading. They hypothesized that there is continual damage of the laryngeal tissue with daily use of the voice, and that the healing mechanism is in a state of constant repair. They also stated that the recovery time was similar to the trajectory of a healing dermal wound. As far as short recovery time is concerned, the minimum period of silence necessary for tissues to experience any degree of recovery is not known, and further investigations are required.

Titze *et al.*[17] began with an investigation of the distribution of voicing and silence periods for teachers at work and those not at work using the National Center for Voice and Speech (NCVS) Voice Dosimeter, a device for the long-term monitoring of vocal parameters, which is based on the vibration of vocal folds sensed at the base of the neck of each subject using a small accelerometer. They measured the occurrences of voicing and silence periods, taking into account the typical frame lengths of the speech rhythms and pauses. This led to the adoption of a scale with a bin duration of half a decade of logarithmic time, that is, voicing periods ranging from $(0.0316 \div 0.10)$ s for the shortest period to $(31.6 \div 31.6)$ s for the longest. Silence was considered for periods of up to several hours. The voice accumulation of each period was then obtained, in seconds, by multiplying the number of occurrences by the corresponding duration. The greatest accumulation of voicing periods per hour was found in the $(0.316 \div 1.0)$ s range for the two-week monitoring of 31 subjects. This included voicing periods at the word and sentence level, and those of silence in the $(3 \div 10)$ s range, with pauses between sentences. On the basis of this analysis, Titze *et al.*[17] suggested that the greatest accumulation of voicing periods might be related directly to vocal fatigue, while the greatest accumulation of silence periods could be related to short-term vocal recovery. Further analyses are needed to associate the accumulation intervals to uncomfortable speaking, as in the case of LS or in the presence of reverberation.

The present study has the aim of investigating the influence of different acoustic environments on the duration of voicing and silence frames in continuous speech. In particular, it has been supposed that the length of voicing periods can increase under more reverberant conditions, with a consequent increase in vocal fatigue. Data were obtained from the long-term monitoring of voices using a contact microphone placed on the base of the neck, near the larynx, which sensed the skin vibrations caused by vocal fold vibrations. Different communication conditions were analyzed, including short monologs spoken by voice professionals and nonprofessionals in front of a listener in a laboratory, and long in-field speech samples, involving teachers during primary school lessons.

II. LONG-TERM VOICE MONITORING

A. Voice monitoring devices

Two portable devices for voice monitoring were used in this study, the characteristics of which are summarized in Table I: The commercial Ambulatory Phonation Monitoring (APM) device,[18,19] model 3200, by Kaypentax (Montvale, NJ), and the new Voice-Care device,[20] which has recently been developed at the Politecnico di Torino.

The long-term monitoring of the voice was carried using a contact microphone placed at the jugular notch, which measured the skin vibrations at the base of the neck that occur during phonation, and an acquisition device that processed the signal at each designated time interval to estimate the vocal parameters.

The devices provide an estimation of the sound pressure levels (SPLs) of the speaker's voice at a fixed distance from the speaker's mouth after a calibration against a reference microphone.[20–22] The phonation time percentage, $D_{t\%}$, is obtained through the procedure described hereafter, which allows the voiced and unvoiced frames to be separated. The fundamental frequency, F_0, is extracted from the voiced frames, with a specific routine that is based on an autocorrelation algorithm.

The APM 3200 consists of a data-logger, connected to a small accelerometer sensor, which was glued to the talker's jugular notch, or fixed using hypoallergenic tape. The interval over which the average vocal parameter value was computed and stored in the memory was 50 ms. Vocal parameters can be downloaded to a personal computer (PC) via a serial port connection.

Voice-Care consists of a data-logger that is based on a low-cost micro-controller board, connected to an electret condenser microphone (ECM), which is used as a contact microphone and is held in place at the jugular notch of the person being monitored by means of surgical tape. The acquired samples are stored in a micro-secure digital-card and then transferred to a PC, where they are processed by subdividing the data stream into frames of 30 ms, which correspond to the inter-syllabic pauses, in order to estimate the vocal parameters.[23] A comparison between APM 3200 and Voice-Care, conducted on the monitoring of the vocal activity of the same female professor during two different university lessons, has proved to be very satisfactory.[19]

Collecting data by measuring skin vibrations through a contact microphone offers many advantages over the use of an air microphone. Besides the reduced size and light weight of the contact microphone (which allows a person to wear it all day), an operational battery life of >10 h, and the possibility of collecting objective vocal data in a person's natural environment, there is the further advantage of its capability to minimize background-noise effects.[18,20,22] Accelerometers and ECMs, in fact, collect data through vibrations rather than from air pressure, hence, the effect of background noise becomes negligible. However, since an ECM is not completely insensitive to air pressure, dedicated experiments were carried out to investigate this aspect. It was found that acoustic noise has a negligible effect

TABLE I. Main characteristics of the Ambulatory Phonation Monitoring (APM) device, Model 3200, and Voice-Care.

Name	Sensor	Bandwidth	Frame length	Estimated parameters
APM 3200	Accelerometer BU7135 (Knowles Corp., Itasca)	2 Hz ÷ 3 kHz Flatness: ±1.5 dB (50÷1000 Hz)	50 ms	SPL, F_0, $D_{t\%}$, and vocal doses as defined in Titze *et al.* (Ref. 26)
Voice-Care	ECM AE38 [Alan Electronics GmbH (Dreieich, Germany)]	10 Hz ÷ 4 kHz	30 ms	SPL, F_0, $D_{t\%}$

on the measurement of the SPL parameter, provided its level does not exceed ~100 dB on the ECM surface.

B. Processing of the acquired data

1. Discrimination between silence and speech and calculation of the phonation time percentage

Of all the information provided by the devices, only the detection of the presence or absence of voiced excitation, i.e., voiced-silence discrimination, has been of interest for the aims of this work. Voiced–voiceless segments could not be discriminated since the contact microphones only detected vocal fold vibrations. Voiceless sounds do not usually make throat walls vibrate.[24]

In order to discriminate between silence and speech, the following procedure was implemented. The voltage signals that were acquired at the contact-microphone chain output were grouped for each designated time interval, and the corresponding root-mean-square (rms) values were calculated (30 ms for Voice-Care and 50 ms for APM 3200). An rms voltage value, which acted as a discrimination threshold that divided the voiced from the silence periods, was manually chosen for each speech, and a separation between speech and silence intervals was allowed.[18,20,24] A further check was applied to critical cases. A histogram of the voltage values (or of the logarithm of the voltage values) was built with two significant maxima, the first for the noise floor value and the second for the voiced region.[25] A minimum exists somewhere in between, since the transitional frames of voiced and noise excitation occur less frequently. This minimum was assumed as the correct threshold for speech–silence discrimination.[24]

The phonation time percentage, $D_{t\%}$, was then calculated as the percentage of the total period spent voicing over the total monitoring time. The occurrence distributions of the voicing and silence segments of different durations were then obtained through a finer-tuned analysis of the available data.[17,19]

2. Analysis of loud speech

Starting from quite similar values of $D_{t\%}$ in normal and "exaggerated" speech, a speech characterized by a higher voice level compared to normal speech, as observed by Titze *et al.*,[26] a specific method for the detection and analysis of loud speech, i.e., speech produced with a high vocal effort, has been proposed. This method can be applied efficiently to in-field speech samples, e.g., to teachers during lessons, as it is supposed that teachers raise their voices in classrooms to catch the attention of pupils, and at times speak with a louder voice. The algorithm is able to detect voicing and silence duration within loud speech intervals and to identify whether specific voice frames are typical of this louder voice.

Loud speech has arbitrarily been identified as the speech level that is exceeded for 10% of the phonation time, i.e., the $L_{v,10}$ speech percentile level. Assuming a typical $D_{t\%}$ value for this type of speech, the proposed algorithm automatically selects loud speech and silence time windows of variable widths with the same $D_{t\%}$. The number of voiced (V) and silence (S) intervals varies, for each selected window, according to the requirement of equal $D_{t\%}$, e.g., in the case of a $D_{t\%}$ of 25%, the windows can be VSSS or VSSVSSSS or VSSVSSVSSSSS, etc. Only windows for which all the voice levels are equal to or higher than $L_{v,10}$ have been considered.

C. Subjects and communication scenarios

Two communication scenarios were considered in this work. These scenarios included short monologs in laboratories and long in-field speeches during primary school lessons. Voice professionals and non-professionals were involved in the tests, the former being university professors and primary school teachers and the latter university students.

1. Laboratory monitoring

Laboratory monitoring was carried out in the semi-anechoic and reverberant rooms of the National Institute of Metrological Research (INRiM) in Turin (Italy), and in the anechoic, semi-reverberant and reverberant rooms of the London South Bank University (LSBU). APM 3200 and Voice-Care were both used in Turin, while only Voice-Care was used in London.

Twenty-two university students, aged 20–30 yr, were monitored in Turin, whereas six middle-aged university professors were involved in the London tests. All the students were native Italian speakers, while four of the professors were native English speakers and two of them spoke English very well since they had been living in England for several years. The speakers were asked to make a continuous 5 min-long free speech, with the aim of transmitting information on something they knew well (e.g., the research topic they dealt with, a recipe, the rules of a game, the path from their house to the workplace, etc.), while standing 6 m away from a young female listener who sat in front of them.

TABLE II. Number of investigated subjects in the LSBU and INRiM laboratory settings divided according to age, gender, and voice professionals. The numbers in brackets are related to the subjects that were monitored twice, but not added to the overall sample.

		LSBU				INRiM				
		Voice-Care		APM 3200		Voice-Care		Voice-Care Map		
		M	F	M	F	M	F	M	F	Overall
Subjects		4	2	3	6	8	5	(4)	(3)	28
Age	20–30	—	—	3	6	8	5	(4)	(3)	22
	31–40	1	—	—	—	—	—	—	—	1
	41–70	3	2	—	—	—	—	—	—	5
Voice professionals		4	2	—	—	—	—	—	—	6

The decision to make the subjects speak freely about a topic they knew well was related to the fact that this was considered the best way of making speakers express themselves in a normal speech manner. Reading or acting would have implied an inflection or an unnatural rhythm, so the vocal parameters would probably not only have been influenced by the room acoustics,[11] but also by the inflection and/or rhythm.

Seven university students performed the experiment twice in Turin. They were also asked to describe a map in order to evoke another form of natural speech in a very specific mode of communication.[10] The map contained 12 landmarks (e.g., "school bus," "shop," and "yacht club"), starting and ending point marks, and a path connecting these two points. Following the same procedure reported in Ref. 11, the speakers were instructed to describe the route from the start to the end points, indicating the landmarks along the path (e.g., "go to the west until you find the yacht club"), while trying to maintain visual contact with the listener. The speakers had the objective of making the listener draw the path correctly on a blank map containing all the items, except the path and the ending mark. Cardinal points and a 2.5 cm background square grid were provided on the map to facilitate speaker-to-listener communication. Two maps were provided, one for each room, each sized 29.7 cm × 42.0 cm. The maps were printed on fabric and laid over a sound absorbing panel hung on a music stand in front of the speaker's eyes at a distance of 1.5 m slightly to the left so that the listener's view was not perturbed.

Table II shows the characteristics of the subjects and the monitored samples in the LSBU and INRiM laboratory settings, while the volume of the rooms, the mid-frequency RT, $T_{\text{mean 0.5–2 kHz}}$, and the A-weighted equivalent background noise level, $L_{Aeq,bn}$, are shown in Table III.

The RT in the empty rooms, T_{30}, at INRiM was measured in the one-third octave bands with a center frequency of 100 Hz–8 kHz, applying the integrated impulse response method using a sine sweep excitation signal.[27] The equipment consisted of an omnidirectional source B&K mod. 4296 (B&K, Nærum, Denmark) connected to an amplifier, interfaced to a notebook PC through a sound card TASCAM US-144 (TEAC America, Inc., Montebello, CA), and a 1/2 in. microphone Schoeps CMC5-U (Schoeps GmbH, Karlsruhe, Germany). DIRAC 5 measurement software was used to generate the excitation signal and to process the recorded signal in order to obtain the impulse response.[28] The results measured for the two source and five microphone positions were combined for the room as a whole to obtain spatial average values. In the case of LSBU, the one-third octave band RT, T_{30}, was measured using a sound analyzer Nor140 (Norsonic AS, Tranby, Norway) that generated pink noise, which was emitted from a hemi-dodecahedron loudspeaker Nor275 placed on the floor. The results measured for the 2 source and 27 (9 points and 3 heights) microphone positions in the semi-reverberant and reverberant rooms were combined for the room as a whole to obtain spatial average values. Average results were found in the anechoic chamber for two source and six microphone positions.

2. In-field monitoring

The in-field experiments involved 23 female and 2 male primary school teachers, who were monitored in 6 schools in Italy using the APM 3200 device. A total of 42 working-day speech samples of 4 h each were considered. The subjects were extracted from the full sample of 40 primary school teachers, monitored by Bottalico and Astolfi,[29] for whom vocal doses and parameters had been obtained and subjective

TABLE III. Physical volume, mid-frequency reverberation time, and A-weighted equivalent background noise level in the LSBU and INRiM laboratory settings, and the number of subjects who were monitored with two different devices in the different rooms.

		LSBU			INRiM	
Room		Anechoic	Semi-reverberant	Reverberant	Semi-anechoic	Reverberant
Volume (m^3)		102	203	203	384	294
$T_{\text{mean 0.5–2 kHz}}$ (s) (standard deviation)		0.05 (0.01)	1.73 (0.03)	3.51 (0.18)	0.11 (0.01)	7.38 (1.61)
$L_{A,eq,bn}$ (dB)		25.9	35.0	38.7	24.5	30.3
Device	APM 3200	—	—	—	9	9
	Voice-Care	4	4	4	13	13

TABLE IV. Number of investigated subjects and speech monitorings collected in the primary schools, divided according to age and gender, for the A and B school groups.

		Group A				Group B					
		M		F		M		F		Overall	
	Age	Subjects	Monitorings	Subjects	Monitorings	Subjects	Monitorings	Subjects	Monitorings	Subjects	Monitorings
Teachers	31–40	—	—	6	10	—	—	5	9	11	19
	41–70	—	—	6	8	2	4	6	11	14	23

impressions had been collected. The selection was based on the homogeneity of the speech task of the teachers. This task was only characterized by whole working days of traditional teaching in the typical school classrooms. Even though the sample was reduced, it was similar to the one surveyed by Titze *et al.*[17] for the same type of study.

The selected teachers were divided into two groups of three schools, A and B, where A grouped the older buildings and B grouped the newer schools. Higher average values of the mid-frequency RT were measured in the classrooms in the older-building group, while the RTs were lower in the newer group and in agreement with the optimal range for a speaker in a classroom as estimated by Bottalico and Astolfi.[29]

The RT was measured in occupied classrooms applying the backward integration technique to the impulse response obtained using a balloon-pop as the impulse source.[27] The ambient noise level was also monitored in the classrooms during plenary lessons where usually one person (pupil or teacher) spoke at a time. The calibrated sound level meter B&K 2250 (B&K, Nærum, Denmark) was placed close to the teacher's desk at a height of 1.5 m from the ground.

Table IV shows the number of investigated subjects and speech monitorings collected in the primary schools, subdivided according to age and gender, for the A and B school groups. Table V shows the physical volume, the average values and the standard deviations of the mid-frequency RTs and the background noise levels in the classrooms, estimated as an A-weighted percentile level, $L_{A,90}$, related to the ambient noise recordings. A significant difference between the two groups was detected for the RT (p-value <0.01).

III. RESULTS

A. Occurrences of voicing and silence periods in different speech scenarios

In a first phase, the results of each monitored scenario were reported as ensemble averages of histograms of voicing and silence occurrences for specific durations, which were multiples of the processed data sampling period (30 ms in the case of Voice-Care and 50 ms in the case of APM 3200), and comparisons were made between the highest occurrences of voicing and silence periods that had been detected from the average distributions.

TABLE V. Physical volume, average values and standard deviation of the mid-frequency reverberation time and background noise level, estimated as the A-weighted percentile level, $L_{A,90}$, in the primary school classrooms for the A and B school groups.

	Group A	Group B
Volume (m^3)	≈240	≈160
$T_{\text{mean } 0.5\text{–}2 \text{ kHz}}$ (s) (standard deviation)	1.15 (0.20)	0.81 (0.11)
$L_{A,90}$ (dB) (standard deviation)	51.5 (8.3)	51.0 (6.2)

In a second phase, a statistical analysis was conducted to test the difference between two or more distributions in the different room settings. Three different nonparametric tests were applied, depending on whether the samples were considered to be independent or dependent.

1. Highest occurrences of voicing and silence periods

Figure 1 shows the average occurrences and standard deviations for voicing and silence periods obtained over 5 min of continuous free speech by the university professors monitored with Voice-Care in the anechoic, reverberant, and semi-reverberant rooms of LSBU. The results show the highest occurrences of silence for the shorter periods with a peak at 90 ms in all three rooms, while the highest occurrence of voicing periods increases with an increase in the RT in all the rooms, that is, 90 ms in the anechoic room, 120 ms in the semi-reverberant room, and 150 ms in the reverberant room.

Figure 2 shows the average occurrences and standard deviations for voicing and silence periods obtained over 5 min of continuous free speech by university students monitored with Voice-Care in the semi-anechoic and reverberant rooms of INRiM. The results show that the highest occurrences of silence and voicing periods are 60 ms and 90 ms, respectively, for both rooms. Since the Voice-Care monitoring sessions were carried out on different days, a reproducibility check was performed, which showed the same results when different groups of students were investigated separately. Figure 3 shows the same results as Fig. 2, but obtained with APM 3200. In this case, the highest occurrences of silence and voicing periods are 50 ms and 100 ms, respectively, for both rooms. No changes were observed in speech duration in the rooms for the two different speech samples. The results obtained with Voice-Care are in perfect agreement with those obtained with APM 3200 for this experiment, the only difference being imputable to the different frame lengths of the processed data.

Figure 4 shows the average values and standard deviations of the occurrences of voicing and silence periods related to the speech samples involving the description of a map by the university students monitored using Voice-Care in the semi-anechoic and reverberant rooms of INRiM. The results show that the highest occurrence of silence periods is 60 ms in both rooms, while the highest occurrence of voicing

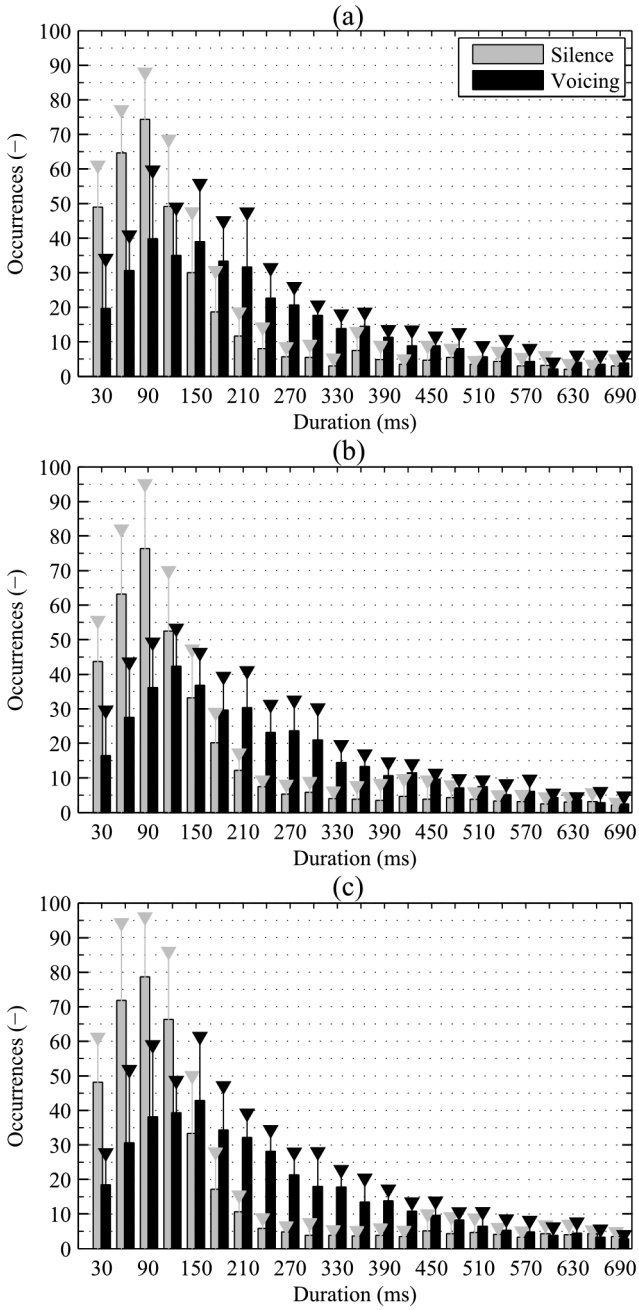

FIG. 1. Ensemble averages (six samples) and standard deviations of histograms for voice and silence occurrences for multiple durations of 30 ms, related to 5 min of continuous free speech made by university professors monitored using Voice-Care in the (a) anechoic, (b) semi-reverberant, and (c) reverberant rooms of the LSBU.

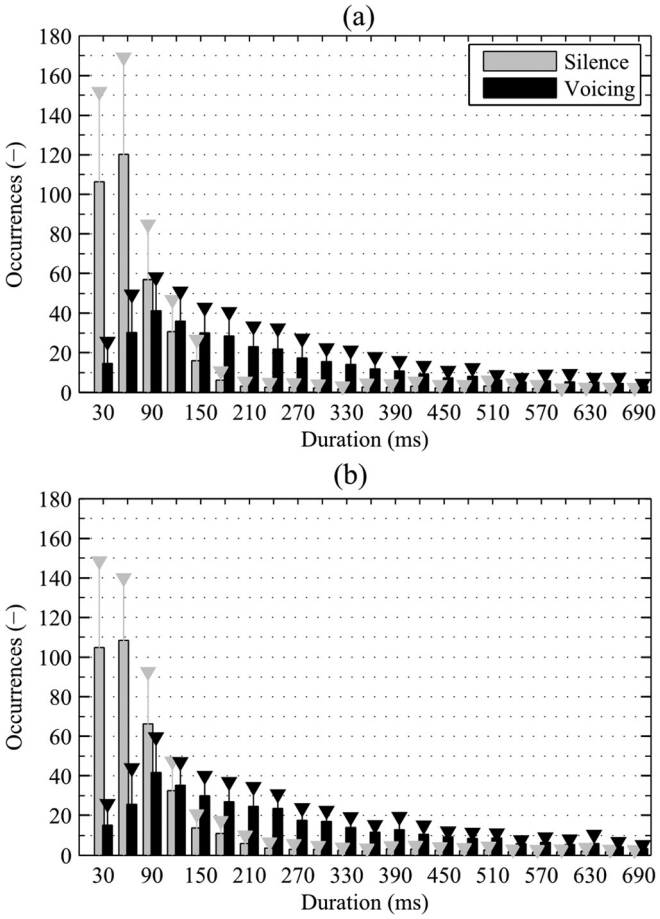

FIG. 2. Ensemble averages (thirteen samples) and standard deviations of histograms for voice and silence occurrences for multiple durations of 30 ms, related to 5 min of continuous free speech made by university students monitored using Voice-Care in the (a) semi-anechoic and (b) reverberant rooms of INRiM in Turin.

periods increases with an increase in the RT in the rooms, that is, 90 ms in the semi-anechoic room and 120 ms in the reverberant room. From a comparison of voice occurrences for the case of free speech (Fig. 2) and when the university students were asked to describe a map (Fig. 4), it can be seen that higher values occur for longer voicing periods in the reverberant room. Moreover, even though the highest occurrence of silence periods does not change from that of free speech, higher occurrences of longer silence periods are shown, in general, when the speakers describe a map.

Figure 5 shows the average occurrences and standard deviations for voicing and silence periods for >4 h of speech for the primary school teachers from the A and B school groups, respectively, monitored with APM 3200. The results show that the highest occurrence of silence periods is 50 ms for both groups. Instead, the highest occurrence of voicing periods is 50 ms for the teachers in group B, who spoke in classrooms with a shorter RT, and 100 ms for the teachers in group A, who spoke in classrooms with longer RTs. The background noise levels in the two school groups were not significantly different and it has, thus, been supposed that the change in the vocal behavior only depended on the different RTs.

2. Statistical analysis

Statistical analyses were carried out with the IBM SPSS statistics package (version 21.0, Armonk, NY). The outcomes of two conditions were initially compared using the Mann-Whitney U test.[30,31] This is a nonparametric test that allows two independent distributions to be compared without making the assumption that data have a pre-specified distribution. The main requirement of the test is that the observations must be independent, which means that there must be no relationship between the observations in each group or between the groups themselves. In the case of a comparison

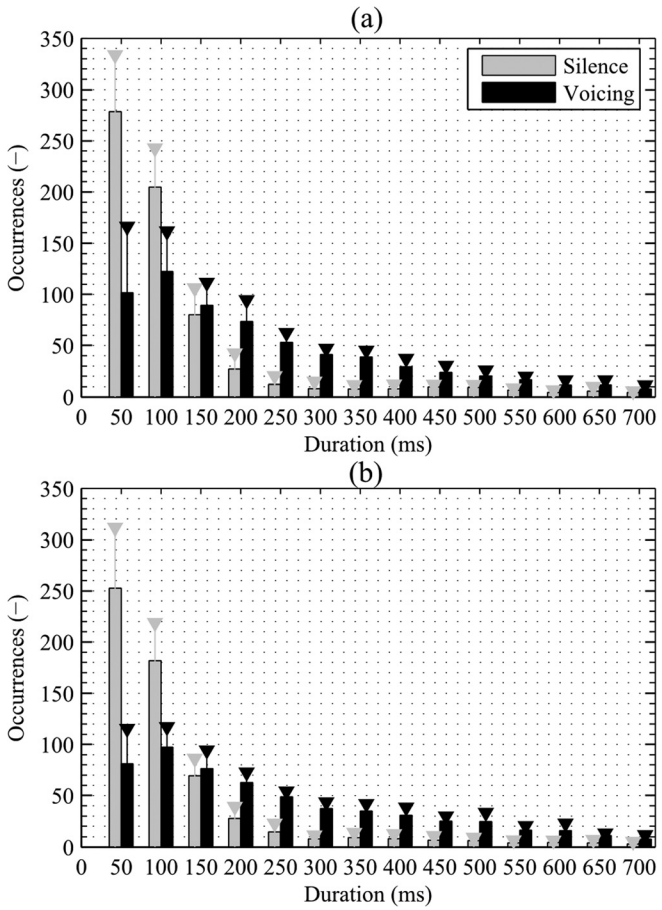

FIG. 3. Ensemble averages (nine samples) and standard deviations of histograms for voice and silence occurrences for multiple durations of 50 ms, related to 5 min of continuous free speech made by university students monitored using APM 3200 in the (a) semi-anechoic and (b) reverberant rooms of INRiM in Turin.

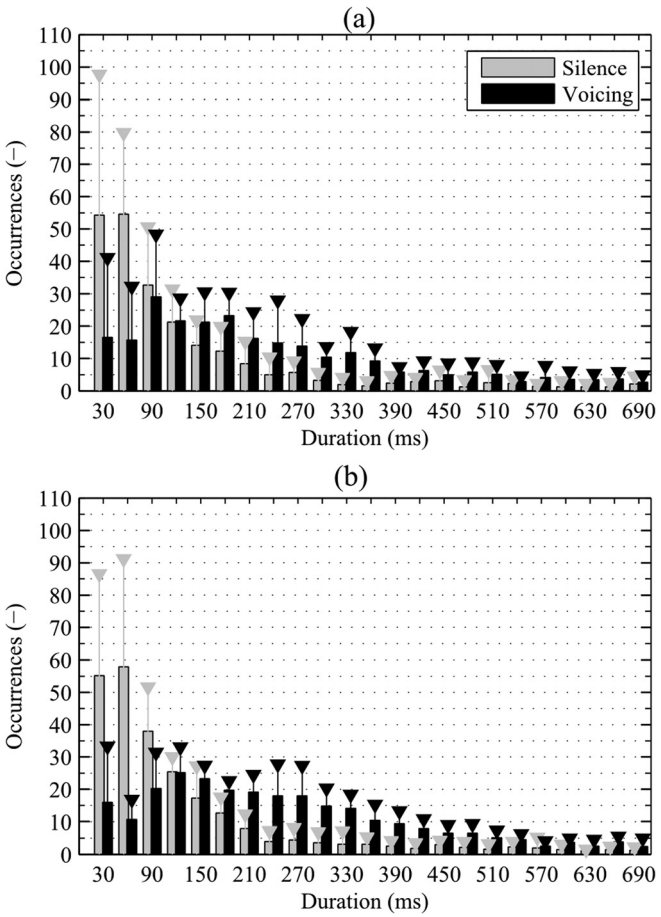

FIG. 4. Ensemble averages (seven samples) and standard deviations of histograms for voice and silence occurrences for multiple durations of 30 ms, related to speech samples in which a map was described by university students monitored using Voice-Care in the (a) semi-anechoic and (b) reverberant rooms of INRiM in Turin.

of distributions of voicing (or silence) occurrences related to the same subject in two different rooms, the samples can be considered independent as long as the speech made by the subject was different in the two rooms. A paired comparison was not possible in this condition, and it was assumed that the endogenous factors of the subject did not influence the differences between the distributions.[30]

The Kruskal–Wallis test was applied when three independent distributions had to be compared. This test extends the Mann–Whitney U test to more than two groups. In the case of a significant difference between groups, the Kruskal–Wallis test does not identify the different samples; hence, the Mann–Whitney U test can be applied after analyzing three samples in pairs.

As a first analysis, the difference between the average distributions of voice (and silence) occurrences of the university professors in the anechoic, semi-reverberant, and reverberant rooms of the LSBU was assessed with the Kruskal–Wallis test, while the Mann–Whitney U test was used to compare the average distributions concerning the university students in the semi-anechoic and reverberant rooms at INRiM.

Statistical significant differences were not found (two-tailed p-value <0.05) between the rooms for either the voice or the silence average distributions in the two laboratory settings. P-values of 0.868 and 0.857 were obtained for the comparison of the average voice and silence distributions, respectively, in the three rooms at LSBU, and p-values of 0.498 and 0.288, respectively, were found for the two rooms at INRiM for the case of university students who made a free speech monitored with Voice-Care, while values of 0.097 and 0.895 were found, respectively, for the students monitored with APM. P-values of 0.136 and 0.962, respectively, were found for the students who described a map monitored with Voice-Care.

Even though the average distributions did not show any significant differences between the rooms, a large inter-speaker variability characterized the monitorings as shown by the high standard deviations of the average occurrences highlighted in Figs. 1–4. In order to better investigate this variability, the same statistical tests were applied to each subject.

Table VI shows the two-tailed p-values of the significance of the differences (p-value < 0.05) related to the voice and silence distributions of each university professor in the anechoic, semi-reverberant, and reverberant rooms of the LSBU according to the Kruskal–Wallis test, and of each

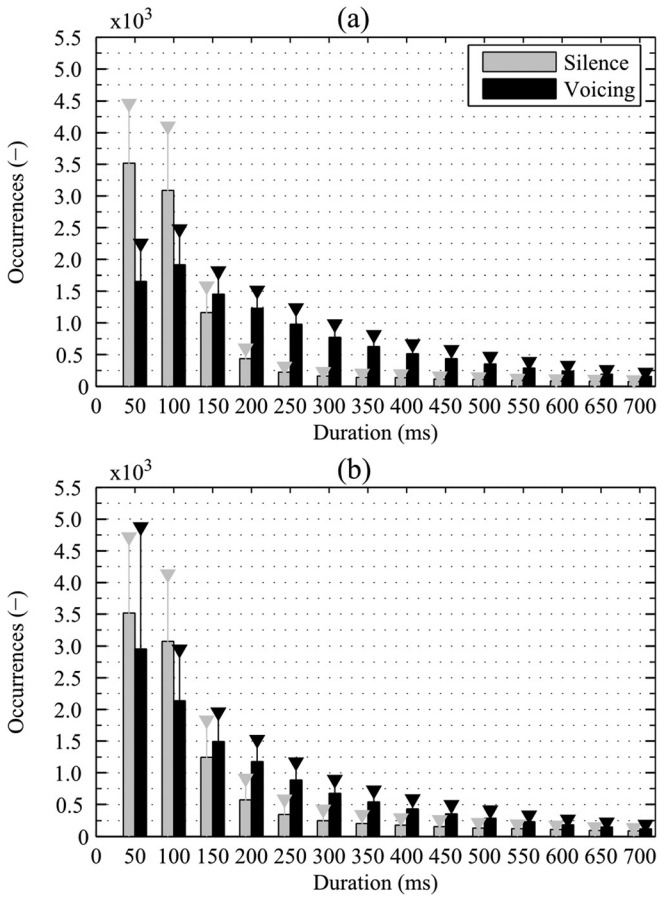

FIG. 5. Ensemble averages and standard deviations of histograms for voice and silence occurrences for multiple durations of 50 ms, related to 4 h of speech at work made by the primary school teachers of groups (a) A, 18 samples, and (b) B, 23 samples, monitored using APM 3200.

university student in the semi-anechoic and reverberant rooms at INRiM according to the Mann–Whitney U test.

For the LSBU setting, three and one out of six subjects showed a significant difference in voice and silence distributions between the rooms, respectively. A Mann–Whitney U test was carried out on these four subjects with the rooms in pairs, and the result was that both the voice and silence distributions of all but one subject differed significantly between the anechoic and semi-reverberant rooms and between the anechoic and reverberant rooms, while no difference was detected between the semi-reverberant and reverberant rooms. A significant difference was only found for voice distribution between the semi-reverberant and reverberant rooms for the SLM01 subject.

As far as the INRiM setting is concerned, 4 out of 13 university students who made a free speech and were monitored with Voice-Care showed significant differences in both voice and silence distributions between the rooms; while four and six out of nine voice and silence distributions were different in the case of students monitored with APM, respectively. Significant differences in voice and silence distributions for the group of students who described a map were found for four and two out of seven students, respectively.

After the previous analysis, based on a comparison of independent samples referring to the same subject, the Wilcoxon signed-rank test[30] was applied to the laboratory data following a different approach: The monitorings in the two rooms of the same subject were considered dependent and a test based on paired samples was applied. The Wilcoxon signed-rank test is a nonparametric test that establishes the significance of the difference between the distributions of two non-independent samples. It requires two related samples or repeated measurements on a single sample, taken in pairs, without any specific assumptions on the distributions. In order to apply the test, the medians and the Kurtosis and Skewness coefficients of the voice distributions and, similarly, those of the silence distributions were calculated in two different rooms for each subject involved in the study, and a pair was thus obtained for each subject. The Wilcoxon signed-rank was then applied to all the paired lists of medians and Kurtosis and Skewness coefficients related to the voice and silence distributions of the same group of subjects who spoke in two different rooms. When comparisons between three rooms were conducted, paired tests between rooms were carried out.

According to the Wilcoxon signed-rank test, significant differences between the rooms were only found for the silence distributions of the students who made a free speech at INRiM, monitored with Voice-Care (p-value $= 0.025$ for the median comparison) and with APM (p-value $= 0.038$ and 0.028 for the Kurtosis and Skewness coefficient comparison, respectively).

The main drawback of the Wilcoxon signed-rank test is related to the presence of ties, i.e., subjects that have the same score in both conditions. In this case, the test discards the individual from the analysis and, thus, reduces the sample size. When the medians were compared in this work, many ties occurred and the sample was, therefore, reduced and the reliability of the test undermined. For this reason, the results of the Wilcoxon signed-rank test have not been considered in the subsequent discussion.

The Mann–Whitney U test was also applied to compare the two series of medians and the Kurtosis and Skewness coefficients, related to the distribution of the two different groups of university students monitored with Voice-Care who made a free speech (13 people) and described a map (7 people) at INRiM. Significant differences were found between the medians (p-value $= 0.037$), as well as between the Kurtosis (p-value $= 0.002$) and Skewness (p-value $= 0.001$) coefficients of the silence distributions in the semi-anechoic room. Significant differences were also found between the Kurtosis (p-value $= 0.008$) and Skewness (p-value $= 0.005$) coefficients of the silence distributions in the reverberant room. No differences were found for the voice distributions.

The Mann–Whitney U test was applied with the same method to compare the two series of medians and the Kurtosis and Skewness coefficients related to the in-field distributions of the two different groups of teachers in schools A and B (18 and 23 samples, respectively). Significant differences were found between the medians (p-value $= 0.013$) and between the Kurtosis (p-value $= 0.013$) and Skewness (p-value $= 0.008$) coefficients of the voice distributions, while no differences were found for the silence

TABLE VI. Two-tailed p-values of the significance of the difference in the distributions of voice and silence occurrences for each subject in the anechoic, semi-reverberant, and reverberant rooms of the LSBU according to the Kruskal–Wallis (K-W) test, and in the semi-anechoic and reverberant rooms of INRiM, according to the Mann–Whitney U (M-W U) test. Values lower than a significance level of 0.05 are reported in bold. The phonation time percentage, $D_{t\%}$, is also shown for each individual, as well as the average values and standard deviations related to the different rooms, devices, and types of speech. M and F stand for male and female, respectively, while SL stands for second language.

			K-W test p-value		$D_{t\%}$ in LSBU Rooms		
Subject	Device	Speech	Voice	Silence	Anechoic	Semi-reverberant	Reverberant
F01	VC	Free	0.703	0.160	52.0	49.0	51.1
M01	VC	Free	**0.001**	0.597	47.8	56.7	55.0
M02	VC	Free	0.698	**0.001**	46.5	44.8	43.7
M03	VC	Free	0.349	0.069	53.9	55.9	59.4
SLF01	VC	Free	**0.005**	0.096	51.2	51.1	54.6
SLM01	VC	Free	**0.000**	0.749	40.4	39.1	48.0
Average					*48.6*	*49.4*	*52.0*
Standard deviation					*4.9*	*6.7*	*5.6*
			M-W U test p-value		$D_{t\%}$ in INRiM rooms		
			Voice	Silence	Semi-anechoic		Reverberant
F02	VC	Free	**0.000**	0.499	68.0		69.8
F03	VC	Free	0.070	**0.050**	71.7		68.3
F04	VC	Free	0.154	0.346	64.9		72.6
F05	VC	Free	0.462	0.468	74.5		68.1
F06	VC	Free	**0.000**	**0.005**	57.1		68.5
M04	VC	Free	**0.019**	**0.001**	63.5		58.1
M05	VC	Free	**0.000**	0.150	72.2		76.7
M06	VC	Free	0.297	0.521	63.6		64.8
M07	VC	Free	0.165	0.389	71.6		70.2
M08	VC	Free	0.290	0.571	66.1		68.7
M09	VC	Free	0.868	**0.009**	61.5		40.4
M10	VC	Free	0.143	0.315	59.5		53.4
M11	VC	Free	0.117	0.594	66.9		65.2
Average					*66.2*		*65.0*
Standard deviation					*5.3*		*9.5*
F04	VC	Map	**0.044**	0.063	48.8		55.4
F05	VC	Map	0.762	0.895	55.9		49.7
F06	VC	Map	**0.007**	0.707	53.7		52.4
M08	VC	Map	0.230	**0.000**	34.4		67.6
M09	VC	Map	**0.000**	0.119	31.5		32.0
M10	VC	Map	0.126	**0.000**	59.5		43.4
M11	VC	Map	**0.005**	0.688	49.9		40.8
Average					*47.7*		*48.8*
Standard deviation					*10.7*		*11.4*
F07	APM	Free	**0.023**	**0.017**	67.9		72.8
F08	APM	Free	**0.015**	**0.000**	61.8		72.8
F09	APM	Free	0.347	0.377	73.5		73.7
F10	APM	Free	0.101	**0.000**	63.6		60.9
F11	APM	Free	**0.001**	**0.036**	67.9		55.8
F12	APM	Free	0.948	0.515	63.5		66.5
M12	APM	Free	0.117	0.268	66.8		68.2
M13	APM	Free	**0.000**	**0.031**	48.6		60.8
M14	APM	Free	0.738	**0.003**	61.2		57.6
Average					*63.9*		*65.4*
Standard deviation					*6.9*		*6.9*

distributions (p-value = 0.248 for the median comparison, and p-value = 0.067 and 0.060 for the Kurtosis and Skewness coefficient comparison, respectively).

Other comparisons related to different laboratories and different devices would be meaningless due to the differences in acoustic conditions of the premises where the

monitorings took place and the different data sampling periods of the Voice-Care and the APM.

B. Phonation time percentage

Table VI shows the phonation time percentage, $D_{t\%}$, for the investigated subjects in the LSBU and INRiM laboratory settings. The average $D_{t\%}$ values and standard deviations in the anechoic, semi-reverberant, and reverberant rooms at LSBU were 48.6 (4.9), 49.4 (6.7), and 52.0 (5.6), respectively. In the semi-anechoic and reverberant rooms at INRiM, the average $D_{t\%}$ values were 63.9 (6.9) and 65.4 (6.9), respectively, with APM 3200, and 66.2 (5.3) and 65.0 (9.5), respectively, for the case of free speech with Voice-Care. The average $D_{t\%}$ values in the case of describing a map with Voice-Care were 47.7 (10.7) and 48.8 (11.6), respectively. No differences were found when the Wilcoxon signed-rank test was applied to the $D_{t\%}$ values shown in Table VI when analyzed in pairs.

As far as the experiment that involved the two primary school groups is concerned, an average $D_{t\%}$ value of 24.1% (7.4) was obtained for group A for teachers who spoke in classrooms with longer RTs, and 21.2% (7.7) for group B for teachers who spoke in classrooms with shorter RTs.

Considerable differences have been detected in $D_{t\%}$ for higher values than ~45% in the case of laboratory monitorings and between 20% and 25% for the in-field measurements. These differences are due to the differences in the speaking task: One was a short-term monolog in the laboratory, without any long hesitation periods, while the other was a 4-h long in-field monitoring with longer pauses, which are typical of teaching activities.[17] A particular remark should be made concerning the laboratory task of describing a map, which, in comparison to the free speech task, shows a $D_{t\%}$ reduction from ~65% to a little below 50%. This can be explained considering that the former, owing to the difficulties involved in finding the best path from the start to the end point in the map, involves many more hesitation pauses than the latter for which the speakers were instructed to speak freely on their own topic.

C. Occurrences of voicing and silence periods in loud speech

Figure 6 shows the ensemble averages and standard deviations of the histograms for voice and silence occurrences for multiple durations of 50 ms related to loud speech made by the primary school teachers in groups A and B, monitored with APM 3200, who spoke in classrooms characterized by a higher and lower RT, respectively.

The primary school teachers showed $D_{t\%}$ values of between 20% and 25% for the A and B school groups, but, in order to compare the results, the same $D_{t\%}$ value of 25% was assumed for the loud speech intervals for both groups. The algorithm proposed in this work was able to automatically select loud speech and silence time windows of variable widths for which $D_{t\%}$ was 25%.

The highest occurrence of silence periods for both groups of teachers was 150 ms and the highest occurrence of voicing periods was 50 ms, thus, showing that the RT did not influence changes in voice duration in the case of loud speech. The occurrences of voice and silence in longer periods of 100 ms and 300 ms, respectively, also appear for group A, thus, indicating a slower rate in more reverberant rooms. Nevertheless, the main difference between the schools is the number of average occurrences, which is much lower in school group A than in school group B. This different behavior in the way of speaking can be explained by considering the lack of "support" to the speech from the room acoustics in the classrooms of group B, characterized by a lower RT, which can result in higher occurrences toward higher speech levels in group B than in group A.

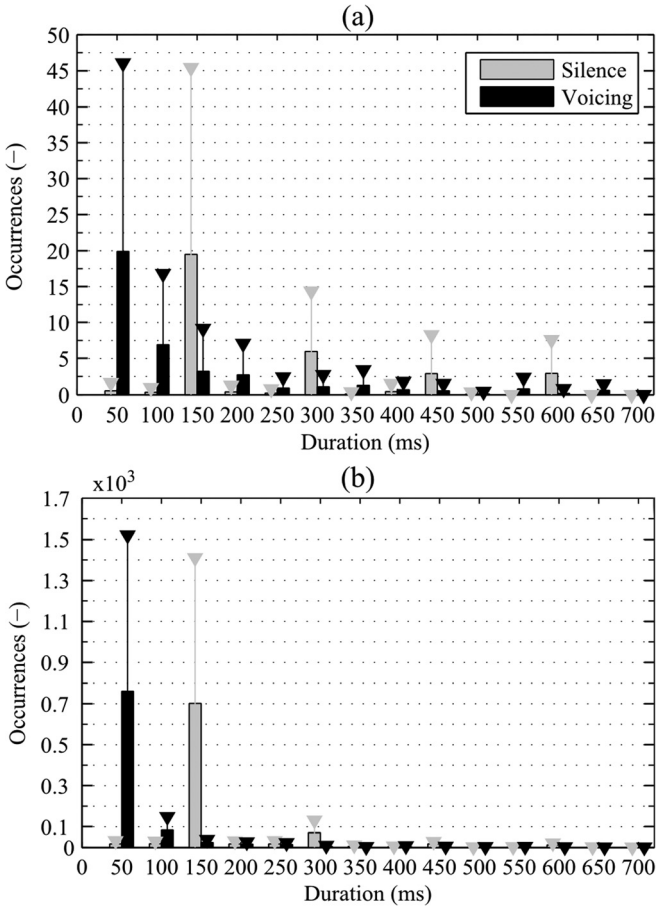

FIG. 6. Ensemble averages and standard deviations of histograms for voice and silence occurrences for multiple durations of 50 ms, related to loud speech at work made by the primary school teachers of groups (a) A, 18 samples, and (b) B, 23 samples, monitored using APM 3200.

IV. DISCUSSION

A. Occurrences of voicing and silence periods for voice professionals and non-voice professionals

In the case of the LSBU voice professionals who made short monologs in the laboratory under very different acoustic conditions, the results, on average, showed a tendency to increase the voicing occurrence for longer periods as the RT in the room where the speech was made increased. The same tendency was found in the case of non-voice professionals, i.e., the university students at INRiM when the speech task was that of clearly describing a map and the speakers were highly motivated to make themselves understood.

Coherent results were also found for the primary school teachers, another category of voice-professionals, in the case of long-term monitoring in classrooms with different

acoustic conditions. On the other hand, in the case of free speech made by non-voice professionals at INRiM, no differences were found in the monitoring of voicing occurrences in very different acoustics conditions.

As far as silence periods are concerned, no differences have been found for the highest occurrence between rooms, but higher occurrences for longer periods were found overall for the voice professionals at the LSBU and for the non-voice professionals at INRiM for the speech task of describing a map in all the acoustic conditions, compared to the non-voice professionals at INRiM in the case of free speech. This result is also supported by the lower $D_{t\%}$ values that were obtained at LSBU with voice professionals and at INRiM with non-voice professionals who described a map, compared to the case of INRiM with non-voice professionals speaking freely. This suggests a slower speech rate for voice professionals and non-voice professionals only in the case of a specific speech task more focused on the listener's needs.

The trends described concerning voicing and silence occurrences for voice professionals and non-voice professionals have been detected from the ensemble averages of distributions. As these were made over a small population, they are affected by a rather large variability as can be seen from the high relative standard deviations. However, the lack of consistent samples has been compensated by the fact that similar results have been obtained in similar experiments with different subjects, even when different monitoring devices were used.

The statistical analysis concerning the comparison of these distributions has basically supported these results. A higher number of significant differences in voice distributions and a lower number of significant differences in silence distributions between rooms were detected overall in the voice professional category at LSBU and in the non-voice professional category at INRiM for the speech task of describing a map, compared to the case of non-voice professionals at INRiM who made a free speech. Moreover, the differences in voicing occurrences at LSBU were mostly between the anechoic and the semi-reverberant and reverberant rooms. In the case of the in-field monitoring of the voice professionals, that is, the primary school teachers, significant differences were only found in the voicing occurrences (and not in the silence occurrences) between classrooms with different RTs.

The finding concerning the occurrences of silence periods between speech contents has also been supported by the statistical analysis. In the case of different speech tasks, significant differences in silence distributions were found between the non-voice professionals at INRiM for the speech task of describing a map and the free speech in all the acoustic conditions.

Durational patterns can be influenced considerably by interspeaker and intraspeaker variability,[12,13] and various effects can influence their duration. One of these effects could be a longer sound tail, which can be compared to noise when it is considered as a challenging listening condition. Speech produced in the presence of noise, i.e., LS, usually exhibits evidence of the lengthening of words and the insertion of more and longer pauses.[7,8] A speaking rate decrease is typical of "clear speech," which is produced in order to adapt to a perturbed communication situation.[2] Clear speech is more intelligible than conversational speech in a variety of difficult listening situations and, in the case of voice professionals whose aim is to be understood by one or more interlocutors, speaking clearly could be a form of natural adaptive behavior.[1,2]

The same can be said for non-voice professionals when their speech task is similar to those of voice-professionals, that is, to make themselves clearly understood by the listeners. In the case of free speech produced by the university students, other factors could have occurred that influenced the results, such as an informal or close relationship between the speaker and the listener, which surely occurred during the experiments in some cases.[11] Shorter and less intelligible words are pronounced when they are predictable from the context, or when they transfer information already known by the listener. Speech produced during interaction between two speakers is, in fact, oriented toward the listener's needs, but when communication occurs efficiently for some reasons, even in the case of a communication barrier, the degree of clarification decreases.[1]

B. Phonation time percentage in different room acoustic conditions

The phonation time percentage values for the voice professionals in the LSBU rooms and in the A and B primary school groups with slightly higher values in more reverberant rooms than in dead rooms, although not significantly different, support the tendency to increase the voice period duration as the RT increases. This behavior is observable from the occurrence distributions in Figs. 5(a) and 5(b) in which slightly higher occurrences of longer voicing periods, together with slightly lower occurrences of silence periods, are shown in classrooms with a longer RT compared to classrooms with a shorter RT. The same behavior is still observable in the case of non-voice professionals for the task of describing a map when Figs. 4(b) and 4(a) are compared, as this particular task had the aim of making the listener draw the path correctly and, hence, of being clearly understood.

These results are in agreement with those obtained by Pelegrín-García *et al.*,[10] who found an average $D_{t\%}$ of 70 for an anechoic room and 72 for a reverberation room with a standard deviation of 5.9 for both rooms for 13 speakers who produced natural speech during the description of a map. Further evidence of the lengthening of the speech segments in more reverberant environments was given by Black.[9] He investigated the effect of RT on speech duration in large rooms with different RTs on 23 subjects who were instructed to read naturally with the aim of making themselves understood by the experimenter. The speech rate was found to be slower in live rooms than in dead ones. The mean duration of phrases was 1.74 s in rooms with a RT of between 0.8 s and 1 s, and 1.53 s in rooms with a RT of between 0.2 s and 0.3 s (the *t*-test showed that these mean durations were significantly different at a confidence level of 99%).

However, for the case of continuous free speech made by non-voice professionals, the almost perfect overlapping of the voicing and silence occurrence patterns in the different rooms has not been fully proved by the average $D_{t\%}$ values, which are slightly different from room to room. When the Voice-Care monitoring at INRiM is considered, the average $D_{t\%}$ value is slightly higher in the semi-anechoic room than in the reverberant one. When the ungrouped Voice-Care data obtained during free speech in the reverberant room is analyzed, an outlier (M09 in Table VI) can be detected whose phonation time percentage is much shorter than the average value. After the removal of this value, the average $D_{t\%}$ in the reverberant room is 67.0% (6.1), a value that is closer to that of 66.2% (5.3) obtained in the semi-anechoic room with a lower standard deviation, thus, supporting the hypothesis of unchanged voice behavior of the speakers. For the case of the non-voice professional speakers, monitored with APM 3200 in the same rooms, a slight increase in the average $D_{t\%}$ is shown in the reverberant room compared to the semi-anechoic room. Again, in this case, an outlier (M13) with a much shorter phonation time percentage than the average percentage for the semi-anechoic room has been detected. After the removal of this value, the average $D_{t\%}$ becomes 65.8% (4.1), a value that is in perfect agreement with the value of 65.4% (6.9) obtained in the reverberant room.

C. Vocal fatigue and recovery

According to Hunter and Titze,[16] the knowledge of the distribution of voicing and silence periods during long-term speech activity associated to the perceptual rating of the talkers allows one to determine which of these periods affects vocal fatigue and vocal recovery, respectively, and these results could be of interest for health-care providers.

Titze *et al.*[17] obtained average values of the occurrences and accumulations of voicing and silence periods per hour, over two weeks, monitoring 31 teachers who spoke with an NCVS voice dosimeter attached to their neck. Accumulation was obtained for each period by multiplying the number of occurrences by the corresponding duration. The data were acquired from an accelerometer that was placed at the base of the subject's neck and were then processed in 30 ms intervals. The authors grouped the occurrences of voicing and silence periods in bin durations of half a decade of logarithmic time in the 0.0316 s–31.6 s range for voicing and up to 10^3 s for silence. The first shortest bin, $(0.0316 \div 0.10)$ s, included voicing and silence periods below and up to the phonemic segmental level, the second bin, $(0.10 \div 0.316)$ s, contained all the occurrences of voicing and silence periods at the phonemic and syllabic level, the third bin, $(0.316 \div 1.0)$ s, included voicing and silence periods at the word and sentence level, the fourth bin, $(1.0 \div 3.16)$ s, grouped all-voiced sentences and pauses between sentences, the fifth bin, $(3.16 \div 10)$ s, included sustained phonations and pauses between sentences, the sixth bin, $(10 \div 31.6)$ s, included rare long phonations and silences in a dialog,[17,32] etc.

Two occurrence peaks were found in the voicing periods in the work by Titze *et al.*[17] below and up to the phonemic segmental level, i.e., bin $(0.0316 \div 0.10)$ s, and at the phonemic and syllabic level, i.e., bin $(0.10 \div 0.316)$ s. The occurrence peak for silence was found in the period below and up to the phonemic segmental level, i.e., bin $(0.0316 \div 0.10)$ s. The greatest accumulation of voicing was found for the word and sentence level, i.e., bin $(0.316 \div 1.0)$ s, while the greatest accumulation of silence was found for the pauses between sentence periods, i.e., bin $(3.16 \div 10)$ s.

The same results were found for the two groups of primary school teachers investigated in this work by clustering data in five bins of half a decade of logarithmic time in the 0.0316 s–10 s range, as shown in Fig. 7. No significant difference was found between group A (older school buildings with higher RT) and group B (newer schools with lower RT).

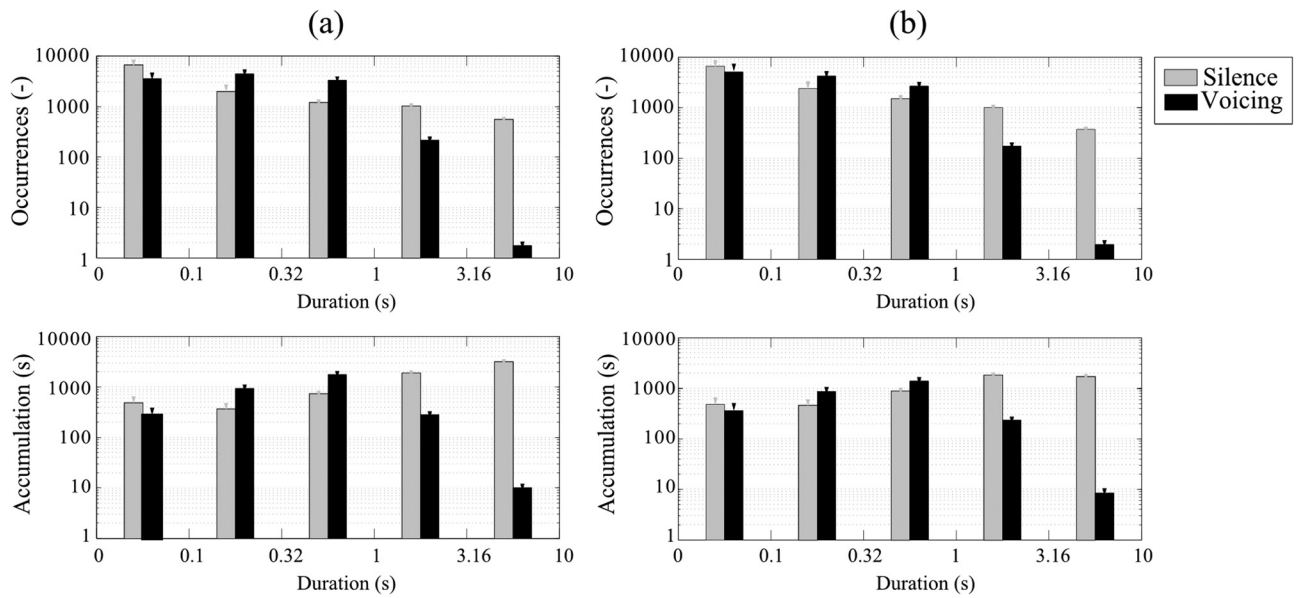

FIG. 7. Ensemble averages and standard deviations of histograms for voice and silence occurrences and accumulations for specific durations in logarithm bins, related to 4 h of speech at work made by the primary school teachers of groups (a) A, 18 samples, and (b) B, 23 samples, monitored using APM 3200.

Bottalico and Astolfi[29] did not find any significant difference between two school groups concerning vocal dose values and vocal parameters, while a significant difference was found in the subjective average scores that teachers assigned to a number of aspects in the classroom. The following aspects were covered: the influence of acoustics on teaching; noise intensity and noise disturbance, i.e., the intensity of the average noise in the classroom and the effect of the disturbance on lessons and practical lessons; noise intensity, noise disturbance, and the frequency of occurrence of different sources perceived by the teachers in the classrooms; reverberation, i.e., reverberation of the sounds and of the teachers' and students' voices; speech comprehension, i.e., how well the teacher comprehended the words spoken by the pupils during traditional lessons; teacher's vocal effort, i.e., the perceived vocal effort of the teacher; acoustical quality satisfaction, i.e., satisfaction of the classroom acoustics.

Significantly worse scores were achieved in group A where the classrooms were more reverberant, than in group B where the classrooms had optimal RT values. This result was also supported by a series of physical problems that were perceived by the teachers at the end of each traditional lesson: 35.2% reported sore throats, 35.2% aphonia, 40.7% raucousness, 18.5% neck stiffness, 11.1% headaches, and 5.6% general illnesses.

Only when the occurrences of the voicing periods are clustered into multiple intervals of 50 ms and represented on a linear scale, as shown in Fig. 5, does the greatest occurrence in the longer period in group A (100 ms), compared to group B (50 ms), support the difference in subjective scores. The hypothesis that the length of the voicing periods can increase due to the longer sound tail with a consequent increase in the vocal fatigue, could only be pointed out with a more fine-tuned analysis of the voicing and silence segments.

V. CONCLUSIONS

Variations in duration of voice and silence periods have been investigated in this work, which is related to continuous speech produced by voice-professionals, i.e., university professors and primary school teachers, and non-voice professionals, i.e., university students, in rooms with very different RTs. The laboratory experiments were held in anechoic, semi-reverberant, and reverberant rooms and involved six middle-aged university professors who made short free monologs to a young listener seated 6 m in front of them. The monologs entailed explaining something they knew well. Longer in-classroom speeches of 4 h each were made by 25 primary school teachers in real communication scenarios.

Twenty-two university students made short free monologs in front of a young listener in both semi-anechoic and reverberant rooms, and seven of them also described a map with the intention of correctly explaining directions to a listener who drew the path on a blank chart.

Measurements were carried out using two devices for the long-term monitoring of vocal parameters, APM 3200 by Kaypentax and Voice-Care, a new device that has recently been developed by the authors. The devices include a contact microphone that is placed at the jugular notch in order to detect the skin vibrations that occur at the base of the neck during phonation.

Ensemble averages of histograms of voice and silence occurrences for multiple durations of the frame length of the processed data (50 ms in the case of APM 3200 and 30 ms in the case of Voice-Care) have been obtained for each monitored scenario.

Although the findings are based on average occurrences with a large uncertainty that is influenced by a very high interspeaker and intraspeaker variability, a tendency to increase the occurrence of longer voicing periods was observed for increasing reverberation. This tendency was only found for the voice-professionals and the non-voice professionals who described a map. These subjects were highly motivated to make themselves understood in the presence of a challenging environmental condition. As far as silence is concerned, higher occurrences of longer periods characterized these two focused speaker categories than for the non-voice professionals who produced free speech. This finding is also in agreement with the lower average phonation time percentage that was observed.

The results have been obtained from different, but homogeneous, speech samples and the reproducibility of some of these samples has also been checked as they were acquired in different monitoring sessions. Even though not completely exhaustive, since it was based on a small number of people, the statistical analysis has basically supported the recognized trends. These trends are in agreement with the literature findings related to speech in adverse communication conditions, although, in the literature, they were essentially oriented to the case of speech in a noisy environment, i.e., LS. Lengthened words and pauses are typical of "clear speech," which is produced spontaneously when high intelligibility is required in a perturbed communication situation. Excessive reverberation can be considered an example of a perturbed speaking situation the same way as speaking in a noisy environment. In the case of voice professionals, whose occupation requires them to be intelligibly understood, speaking clearly can be a natural adaptive action. The same is true of non-voice professionals when they are given a speaking task that is focused on the listener's needs.

Finally, a specific method for the detection and analysis of loud speech, i.e., speech produced with a high vocal effort, has been proposed. RT does not influence changes in voice duration in the case of teachers speaking loudly, even though a slower speaking rate appears in more reverberant rooms. The proposed method should be considered as a preliminary attempt to investigate whether specific voicing periods are used by teachers at work when they change intensity to maintain interest. Further research is planned in order to detect the changes in loud speech in different room acoustic conditions and to investigate vocal fatigue since a specific metric to show vocal impairment has not yet been identified.

ACKNOWLEDGMENTS

This work has been funded by the Italian National Institute for Occupational Safety. The kind cooperation of

the professors from the London South Bank University, the students from the Politecnico di Torino, the primary school teachers from the D. Muratori and L. Fontana schools in Turin, and the E. De Amicis, A. Gramsci, and A. Mei schools in Beinasco have made this work possible. Last, particular thanks are extended to the INRiM researchers who made their laboratories available for the experiments.

[1]V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," J. Acoust. Soc. Am. **130**(4), 2139–2152 (2011).
[2]M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," Comput. Speech Lang. **28**, 543–571 (2014).
[3]J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**(1), 362–378 (2004).
[4]R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," J. Speech Hear. Res. **39**, 494–509 (1996).
[5]M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446 (1986).
[6]K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise ad reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**(3), 1581–1592 (1994).
[7]W. V. Summers, D. P. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928 (1988).
[8]J. C. Junqua, "The Lombard reflex and its role on human listener and automatic speech recognizers," J. Acoust. Soc. Am. **93**, 510–524 (1993).
[9]J. Black, "The effect of room characteristics upon vocal intensity and rate," J. Acoust. Soc. Am. **22**, 174–176 (1950).
[10]D. Pelegrín-García, B. Smits, J. Brunskog, and C. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," J. Acoust. Soc. Am. **129**(4), 1981–1990 (2011).
[11]A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," Lang. Speech **34**, 351–366 (1991).
[12]D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**(5), 1208–1221 (1976).
[13]T. H. Crystal and A. S. House, "Segmental durations in connected speech signals: Preliminary results," J. Acoust. Soc. Am. **72**(3), 705–716 (1982).
[14]C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," J. Acoust. Soc. Am. **52**(4), 1238–1250 (1972).
[15]R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," J. Phon. **11**, 51–62 (1983).
[16]E. Hunter and I. R. Titze, "Quantifying vocal fatigue recovery: Dynamic vocal recovery trajectories after a vocal loading exercise," Ann. Otol. Rhinol. Laryngol. **118**, 449–460 (2009).
[17]I. R. Titze, E. J. Hunter, and J. G. Švec, "Voicing and silence periods in daily and weekly vocalizations of teachers," J. Acoust. Soc. Am. **121**(1), 469–478 (2007).
[18]H. A. Cheyne, H. M. Hanson, R. P. Genereux, K. N. Stevens, and R. E. Hillman, "Development and testing of a portable vocal accumulator," J. Speech Lang. Hear. Res. **46**(6), 1457–1467 (2003).
[19]A. Astolfi, A. Carullo, A. Vallan, and L. Pavese, "Influence of classroom acoustics on the vocal behavior of teachers," in *Proceedings of 21st International Congress on Acoustics*, Montreal (June 2–7, 2013), Vol. 19.
[20]A. Carullo, A. Vallan, and A. Astolfi, "Design issues for a portable vocal analyzer," IEEE T. Instrum. Meas. **62**(5), 1084–1093 (2013).
[21]R. E. Hillman, J. T. Heaton, A. Masaki, S. M. Zeitels, and H. A. Cheyne, "Ambulatory monitoring of disordered voices," Ann. Otol. Rhinol. Laryngol. **115**(11), 795–801 (2006).
[22]P. S. Popolo, J. G. Švec, and I. R. Titze, "Adaptation of a pocket PC for use as a wearable voice dosimeter," J. Speech Lang. Hear. Res. **48**, 780–791 (2005).
[23]P. S. Popolo, M. K. Rogge, J. G. Švec, and I. R. Titze, "Technical considerations in the design of a wearable voice dosimeter," The National Center for Voice and Speech Online Technical Memo, No. 5, 2002, version 1.1, available at http://www.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo05.pdf (Last viewed June 26, 2014).
[24]W. Hess, *Pitch Determination of Speech Signals. Algorithms and Devices* (Springer, Berlin, 1983), pp. 133–144.
[25]B. S. Atal and L. R. Miner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust., Speech, Signal Process. ASSP **24**(3), 201–212 (1976).
[26]I. R. Titze, J. G. Švec, and P. S. Popolo, "Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues," J. Speech Lang. Hear. Res. **46**, 919–932 (2003).
[27]ISO 3382-2:2008, Acoustics—Measurement of room acoustic parameters—Part 2: Reverberation time in ordinary rooms (International Organization for Standardization, Geneva, Switzerland, 2008).
[28]Acoustics Engineering, "Measuring impulse responses using Dirac," Technical Report, Acoustics Engineering (2007), Technical Note 001, available at http://www.acoustics-engineering.com/support/technotes.htm (Last viewed May 19, 2014).
[29]P. Bottalico and A. Astolfi, "Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms," J. Acoust. Soc. Am. **131**(4), 2817–2827 (2012).
[30]S. Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1988), pp. 1–399.
[31]R. Bergman, J. Ludbrook, and W. P. J. M. Spooren, "Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages," Am. Stat. **54**(1), 72–77 (2000).
[32]B. Zellner, "Pauses and the temporal structure of speech," in *Fundamentals of Speech Synthesis and Speech Recognition*, edited by E. Keller (Wiley, Chichester, 1994), pp. 41–62.