Vocal effort with changing talker-to-listener distance in different acoustic environments

David Pelegrín-García,[a] Bertrand Smits, Jonas Brunskog, and Cheol-Ho Jeong
Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark

(Received 11 February 2010; revised 16 October 2010; accepted 14 January 2011)

Talkers adjust their vocal effort to communicate at different distances, aiming to compensate for the sound propagation losses. The present paper studies the influence of four acoustically different rooms on the speech produced by 13 male talkers addressing a listener at four distances. Talkers raised their vocal intensity by between 1.3 and 2.2 dB per double distance to the listener and lowered it as a linear function of the quantity "room gain" at a rate of −3.6 dB/dB. There were also significant variations in the mean fundamental frequency, both across distance (3.8 Hz per double distance) and among environments (4.3 Hz), and in the long-term standard deviation of the fundamental frequency among rooms (4 Hz). In the most uncomfortable rooms to speak in, talkers prolonged the voiced segments of the speech they produced, either as a side-effect of increased vocal intensity or in order to compensate for a decrease in speech intelligibility. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3552881]

PACS number(s): 43.55.Hy, 43.70.Mn [NX] Pages: 1981–1990

I. INTRODUCTION

In face-to-face communication, a talker makes a decision about the desired vocal output based on the given communication scenario. Some factors affecting this decision are the intention of the talker (dialog, discipline, rebuke…), the distance between talker and listener, and special requirements of the listener, due to hearing impairment or language disorders. Once the decision is made, the talker starts to speak and uses a series of feedback mechanisms (auditory, tactile, proprioceptive, and internal) to grant that the actual vocal output matches the desired vocal output.[1]

Speaking in various rooms leads to different experiences or sensations for a talker, due to changes in auditory feedback. The vocal effort required for communicating with a listener at different distances changes with room acoustic conditions, as does also the feeling of vocal comfort. One should differentiate between the concepts of vocal effort and vocal comfort. Vocal effort, according to Traunmüller and Eriksson,[2] is a physiological magnitude different from vocal intensity, which accounts for the changes in voice production required for the communication at different distances. This definition of vocal effort can be extended to also include the changes in voice production induced by noise or the physical environment. These changes include vocal intensity, fundamental frequency (F0), vowel duration, and the spectral distribution of speech. Vocal comfort, according to Titze,[3] is a psychological magnitude determined by those aspects that reduce the vocal effort. Vocal comfort reflects the self-perception of the vocal effort by the feedback mechanisms listed above.

The maximization of vocal comfort should be a priority in situations of very high vocal demands, which are hazardous for the vocal health, such as teaching environments. A recent study revealed that around 13% of teachers suffer from voice problems.[4] Indeed, the prevalence of voice problems among teachers is much higher than it should be, compared to their representation in overall population.[5–7] Vilkman[8] points out "bad classroom acoustics" as one of the hazards for voice health from the testimonies of teachers who had suffered from voice disorders. These disorders are related, in many cases, to the intensive use of the voice as an occupational tool.

To characterize the amount of voice use, and to estimate the risk of suffering from voice problems, Titze *et al.*[9] introduced a set of measures of the accumulated exposure of vocal fold vibration, called vocal doses. The vocal doses are calculated from the phonation time, F0, and the vocal fold vibration amplitude. In the present work, the variations of vocal intensity (as a rough estimate of the vocal fold vibration amplitude), F0, and the phonation time are reported without going further into a detailed risk analysis, leaving this task to future studies and more advanced analytical models. As in the study by Rantala *et al.*,[10] both the mean and the standard deviation of F0 are measured as indicators of vocal effort.

Although bad classroom acoustics might be hazardous for voice health, only a few works have attempted to relate classroom acoustics to voice production. Hodgson *et al.*[11] suggested a simple empirical prediction model to calculate average voice levels used by teachers in university lecture rooms, depending on individual factors, acoustical characteristics of the room, and student activity noise. Brunskog *et al.*[12] found that the average vocal intensity used by teachers in different classrooms is closely related to the amplification of the room on the talker's perceived own voice (defined as "room gain"). From this study, it appears that teachers speak louder in rooms with a low room gain and softer in rooms with a high room gain, at a rate of −13.5 dB/dB (decibels of voice level per decibels of room gain).[13] However, none of the two previous studies took into account the distance

[a] Author to whom correspondence should be addressed. Electronic mail: dpg@elektro.dtu.dk

between teachers and students, which could explain by itself some of the changes in voice level. From a different perspective, Kob *et al*.[14] found that teachers with voice disorders were more affected by unfavorable classroom acoustics than their healthy colleagues.

In a more general communication context, several investigations have analyzed the vocal intensity used by a talker to address a listener located at different distances. One general finding is that the vocal intensity is approximately proportional to the logarithm of the distance. The slope of this relationship is in this paper referred to as the compensation rate (in decibels/double distance), meaning the variation in voice level (in decibels) each time that the distance to the listener is doubled (double distance). Warren[15] found compensation rates of 6 dB/dd when talkers produced a sustained vocalization (/a/) addressing listeners at different distances, suggesting that talkers had a tacit knowledge of the attenuation of sound with the distance. However, a sound attenuation of 6 dB/dd is only found in free-field or very close to the source. Warren did not provide information on the experimental acoustic surroundings. Michael *et al*.[16] showed that the speech material (natural speech or bare vocalizations) influenced the compensation rates and found lower values than Warren, 2.4 dB/dd for vocalizations and 1.3 dB/dd for natural speech. Healey *et al*.[17] obtained compensation rates in a range between 4.5 and 5 dB/dd when the task was to read a text aloud to a listener at different distances. Liénard and Di Benedetto[18] found an average compensation rate of 2.6 dB/dd in a distance range from 0.4 to 6 m using vocalizations. Traunmüller and Eriksson[2] carried out their experiments with distances ranging from 0.3 to 187.5 m to elicit larger changes in vocal effort, finding a compensation rate of 3.7 dB/dd with spoken sentences. In general, there is a substantial disagreement among the results of different studies.

Each of the previous experiments analyzing voice production with different communication distances was carried out in only one acoustic environment. Michael *et al*.[16] pointed out that unexplained differences among experimental results might be ascribed to the effect of different acoustic environments, because the attenuation of sound pressure level (SPL) with distance depends on the room acoustic conditions. Zahorik and Kelly[19] investigated how talkers varied their vocal intensity to compensate for the attenuation of sound with distance in two acoustically different environments (one indoor and one outdoor), when they were instructed to provide a constant SPL at the listener position. When uttering a sustained /a/, the talkers provided an almost uniform SPL at each of the listener positions, which indicated that talkers had a sophisticated knowledge of physical sound propagation properties. The measured compensation rates laid between 1.8 dB/dd for an indoor environment and 6.4 dB/dd for an outdoor environment.

In addition, some of the studies investigated further indicators of vocal effort at different communication distances. Liénard and Di Benedetto[18] also found a positive correlation between vocal intensity and F0 and significant spectral changes in vowels. Traunmüller and Eriksson[2] observed that the duration of vocalic segments increased with communication distance, and thus, with vocal effort.

In summary, there have been many studies reporting vocal intensity at different communication distances, as well as other descriptors of vocal effort: F0 and vowel duration. Only one study[19] analyzed the additional effect of the acoustic environment on the vocal intensity, although the instruction—*provide a constant SPL at the listener position*—and the speech material—vocalizations—were not representative of a normal communication scenario. The aim of the present study is to analyze the effect of the acoustical environment on the natural speech produced by talkers at different communication distances in the absence of background noise, reporting the parameters which might be relevant for the vocal comfort and for assessing the risks for vocal health.

II. EXPERIMENTAL METHOD

The speech from 13 talkers speaking to one listener at four different distances in four different rooms was recorded. The speech signals were processed to calculate measures of vocal intensity, F0, and the relative duration of the phonated segments.

A. Subjects

Thirteen male talkers participated in the experiment as talkers. Two of the talkers were acting as listeners and experimenters at different times. All 13 subjects had ages between 23 and 40 yr and had neither hearing and visual impairments nor vocal disorder. None of the subjects were native English speaker, but nevertheless all of them used English as the spoken language during the tests.

B. Instruction

Before the start of the tests, the listener/experimenter explained the instructions verbally to each talker at a close distance. The talkers were given a map that contained roughly a dozen of labeled items (e.g., "diamond mine," "fast flowing river," and "desert"), starting and ending point marks, and a path connecting these two points. They were instructed to describe the route between the starting point and the finish point, indicating the items along the path (e.g., "go to the west until you find the harbor"), while trying to enable eye-contact with the talker. There were 16 maps in total, and a different map was used at each condition. The order of the maps was randomized differently for each subject. These maps have been used extensively in previous research to obtain a dialog-based speech corpus.[20] The object of using maps was evoking natural speech from the talkers in a very specific context and mode of communication. An alternative method for obtaining natural speech could have been instructing talkers to speak freely. However, there would have been different modes of communication and contexts among subjects, which would have introduced higher variability in the data.

After explaining the task to the talker, the listener stood at different positions and indicated the talker non-verbally when to start talking. The listener gave no feedback to the talker, either verbally or non-verbally, about the voice level perceived at his position.

TABLE I. Physical volume, reverberation time, room gain, STI (mouth-to-ears), and A-weighted background noise level measured in the four environments: anechoic chamber, lecture hall, corridor, and reverberation room.

	V [m^3]	T_{30} [s]	G_{RG} [dB]	STI	$L_{N,Aeq}$ [dB]
Anechoic room	1000	0.04	0.01	1.00	<20
Lecture hall	1174	1.88	0.16	0.93	28.2
Corridor	410	2.34	0.65	0.83	37.7
Reverberation room	500	5.38	0.77	0.67	20.6

At the end of the experiment, the subjects were asked about the experience of talking in the different rooms and they could answer openly.

C. Conditions

For each subject, the experiment was performed in a total of 16 different conditions, resulting from the combination of four distances (1.5, 3, 6, and 12 m) and four different environments (an anechoic chamber, a lecture hall, a long, narrow corridor, and a reverberation room). The environments were chosen so as to represent a wide range of room acoustic conditions, while being large enough to allow distances between talker and listener of up to 12 m. However, not all of these rooms were representative of everyday environments. The order of the rooms was randomized for each subject, but the distances from talker-to-listener were always chosen from closest to furthest. Talker and listener stood further than 1 m from the walls and faced each other.

The volume V, reverberation time T_{30}, room gain G_{RG}, speech transmission index (STI) between talker's mouth and ears, and A-weighted background noise levels $L_{N,\mathrm{Aeq}}$, measured in the rooms are shown in Table I.

1. Reverberation time

The reverberation time T_{30} was measured according to ISO-3382,[21] using a dodecahedron loudspeaker as an omnidirectional sound source and a 1/2 in. microphone, Brüel & Kjær (B&K) type 4192 (Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark). The measurements were carried out with DIRAC,[22] using an exponential sweep as the excitation signal. The T_{30} obtained from the impulse response using Schroeder's method[23] and averaging the measurements in the 500 Hz and 1 kHz one-octave bands is shown in Table I.

2. Room gain

The room gain G_{RG} was measured with the method proposed by Pelegrin-Garcia[13] in the empty rooms, using a Head and Torso Simulator (HATS) B&K type 4128 with left ear simulator B&K type 4159 and right ear simulator B&K type 4158. The software measurement DIRAC was used to generate an exponential sweep as an excitation signal and extract the impulse responses from the received signals on the microphones at the ears of the HATS. The HATS was placed at the talker position, with the mouth at a height of 1.6 m and more than 1 m away from reflecting surfaces. The G_{RG} values reported for each room correspond to the average of the values at the two ears and three different repetitions and are shown on Table I. No filtering was applied to the impulse response to calculate G_{RG}.

3. STI

The STI was derived with the AURORA software suite[24] from the same mouth-to-ears impulse responses used for the G_{RG} measurements and ignoring the effect of background noise. The values resulting from averaging three repetitions and the two channels (left and right) at each environment are shown on Table I. One should note that the STI parameter was not originally intended to explain the transmission of speech between the mouth and the ears of a talker, as in this case, but to characterize the transmission channel between talker and listener. The STI values presented here are used only as rough indicators of the perceived degradation in one's own voice due to reverberation and ignoring completely the bone-conducted component of one's own voice.

4. Background noise level

The A-weighted, 20-s equivalent background noise levels ($L_{N,\mathrm{Aeq}}$) were measured in the empty rooms using a sound level meter, B&K type 2250. The results from averaging the measurements across four positions in each room are shown in Table I. Possible noise sources contributing to the reported levels are ventilation systems, traffic, and the activity in neighboring areas. All the measured background noise levels were below 45 dB(A) so, according to Lazarus,[25] the produced voice levels were not affected by the noise.

5. Speech sound level

The *speech sound level*[26] S is defined as the difference between the SPL L_p produced by a source with human voice radiation characteristics at a certain position and the level L_{ref} produced by the same source at 10 m in free-field, averaged over all directions in space,

$$S = L_p - L_{\mathrm{ref}}. \quad (1)$$

A directive loudspeaker JBL Control One (JBL Professional, Northridge CA) was used as the sound source and was placed at the talker position, with the edge of the low frequency driver at a height of 165 cm above the floor and pointing toward the listener. The SPL L_p produced by the loudspeaker reproducing pink noise was analyzed in one-octave bands with a sound level meter, B&K type 2250, at the listener position for each of the four distances in each room.

The reference SPL L_{ref} was calculated as the average of 13 measurements in an anechoic chamber with a distance of 10 m between the sound level meter and the loudspeaker. For each measurement, the loudspeaker was turned at steps of 15° from 0° to 180° and reproduced the same pink noise signal with the same gain settings as used for the measurement of L_p.

The resulting S, as a function of distance, averaged across the one-octave mid-frequency bands of 500 Hz and 1 kHz, is presented in Fig. 1.

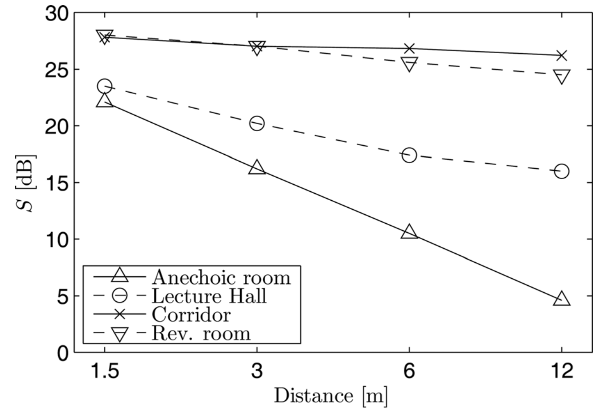

FIG. 1. Speech sound level S as a function of distance.

D. Processing of the voice recordings

The acoustic speech signal was picked up with a DPA 4066 headworn microphone (DPA Microphones A/S, Allerød, Denmark), placed on the talker's cheek at a distance of 6 cm from the lips' edge. The signal was recorded with a Sound Devices 722 digital recorder (Sound Devices, LLC, Reedsburg, WI) in 24 bits/44.1 kHz pulsed-code modulation (PCM) format and later processed with MATLAB. The length of the recordings varied between 1 and 2 min, depending on the map and the talker.

1. Voice power level

Vocal intensity is related to the strength of the speech sounds. There are many ways to represent this magnitude, e.g., on-axis SPL at different distances in free-field, sound power level (L_W), or vibration amplitude of the vocal folds. Among these parameters, the sound power level appears to be the most appropriate one to characterize the total sound radiation from a source. Indeed, it is possible to determine the sound power level if the on-axis SPL in free-field conditions and the directivity of the speaker are known. Following the works of Hodgson *et al.*[11] and Brunskog *et al.*,[12] the sound power level was chosen as the main index of vocal intensity and is also referred to as voice power level.

To determine the voice power level of the recordings, the equivalent SPL in the one-octave bands between 125 Hz and 4 kHz was first calculated. A correction factor due to the increase of SPL at the headworn microphone in the different rooms was applied (see values in Table II). The correction factor was measured by analyzing the SPL produced by the HATS, reproducing pink noise with a constant sound power level in the different rooms, at the headworn microphone, which was placed on the HATS. The SPL readings from the anechoic chamber were subtracted to the readings in each room. The difference between the corrected SPL at the headworn microphone and the voice power level was determined by performing sound power measurements in a reverberation room in a similar way as described by Brunskog *et al.*[12] However, instead of using a dummy head (as in Brunskog *et al.*), the speech of six different talkers, one by one, was recorded simultaneously using a headworn microphone DPA 4066 and a 1/2 in. microphone, B&K type 4192, positioned in the far field, where the sound field is assumed to be diffuse. The difference between the mean corrected SPL measured at the headworn microphone and the voice power level as a function of frequency is shown in Fig. 2.

TABLE II. Increase of SPL (in decibels) at the headworn microphone due to sound reflections (used as correction factor), measured with a dummy head. The reference situation is the measurement of SPL in anechoic conditions. Abbreviations are used instead of the complete name of the rooms: LH for the lecture hall, COR for the corridor, and REV for the reverberation room.

	Frequency (Hz)					
Room	125	250	500	1000	2000	4000
LH	0.27	0.05	0.12	0.22	0.07	0.15
COR	0.58	0.32	0.46	0.54	0.59	0.69
REV	0.30	0.18	0.38	0.49	0.43	0.51

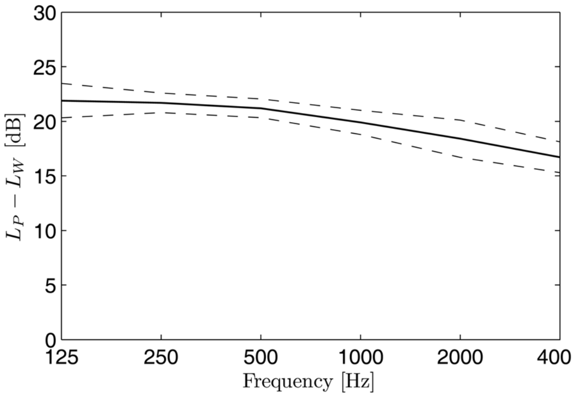

FIG. 2. Difference between the SPL measured at the headworn microphone, corrected for the increase in SPL due to sound reflections, and L_W. Bold line: mean value. Dashed lines: one standard deviation above and below the mean value.

2. Fundamental frequency

F0 was extracted from the recordings with the application WAVESURFER[27] using the entropic signal processing system method at intervals of 10 ms. Taking a sequence with the F0 values of the voiced segments (the only segments for which the algorithm gave an estimation of F0), the mean (noted as $\bar{F}_0$) and the standard deviation (noted as σ_{F_0}) were calculated.

3. Phonation time ratio (PTR)

Due to the large variations in the length of speech material among subjects and conditions, the absolute phonation time is not reported, but the ratio of the phonation time t_P to the total duration of running speech t_S in each recording, referred to as PTR. The calculation procedure is shown in Fig. 3. First, the original speech signal [Fig. 3(a)] is processed to obtain the running speech signal [Fig. 3(b)]. Then, this signal is split into N non-overlapping frames or segments of a duration $t_F = 10$ ms [Fig. 3(c)]. In the ith frame, the logical variable k_i ($k_i = 0$ if the segment is unvoiced;

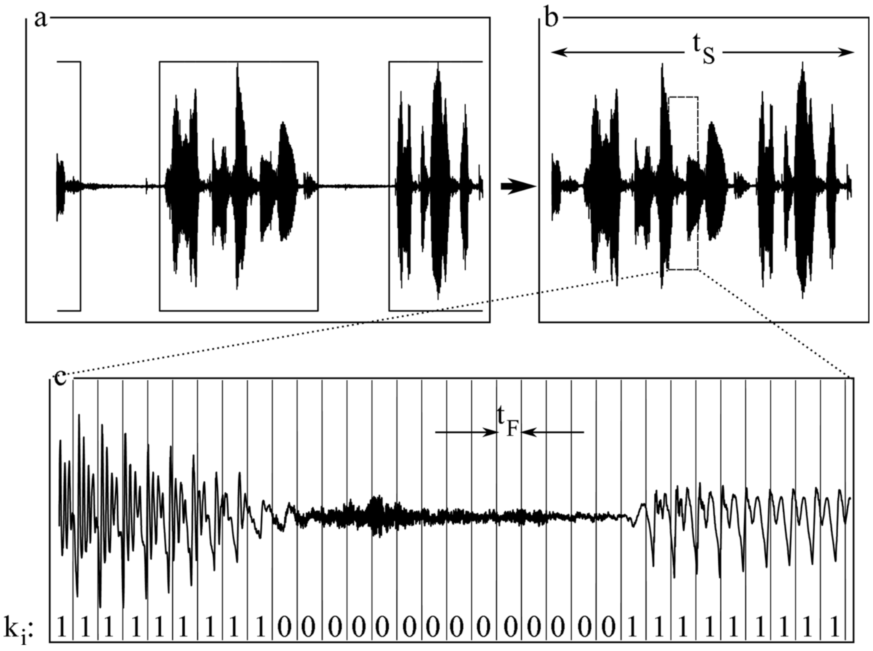

FIG. 3. Post-processing of the recordings and computation of the PTR. (a) Original speech signal. (b) Running speech signal of duration t_S, obtained from the original signal by removing 200 ms-long frames with very low energy. (c) Calculation of the phonation time by splitting the running speech signal in frames of length $t_F = 10$ ms, determining whether each segment i is phonated ($k_i = 1$) or not ($k_i = 0$) and adding up the time of all phonated segments.

$k_i = 1$ if it is voiced) is determined with WAVESURFER. The total duration t_P of phonated segments is $t_F \times \sum_{i=1}^{N} k_i$. Thus,

$$\mathrm{PTR} = \frac{\text{Phonation time}}{\text{Running speech time}} = \frac{t_F \sum\limits_{i=1}^{N} k_i}{t_s}, \quad N = \left\lfloor \frac{t_s}{t_F} \right\rfloor. \tag{2}$$

The floor operator $\lfloor \cdot \rfloor$ results in the closest integer not larger than the operand.

E. Statistical method

For each parameter (L_W, $\bar{F}_0$, σ_{F_0}, and PTR), a linear mixed model[28] was built from a total of 208 observations (13 subjects × 4 distances × 4 rooms), using the `lem4` method in the library `lme4`[29] of the statistical software R.[30] The "full model" included the logarithm of the distance as a covariate and the acoustic environment (or room) as a factor and the interaction between the distance and the room. In the present paper, the mixed model for a response variable y which depends on the ith subject, the jth distance d_j, and the kth room is presented in the form

$$y_{ijk} = a_k + \alpha_i + (b_k + \beta_i) \times \log_2(d_j/1.5) + \varepsilon_{ijk}. \tag{3}$$

The fixed-effects are written on roman characters (a_k and b_k) and the random effects are written on greek characters (α_i, β_i, and ε_{ijk}). The random effects are stochastic variables normally distributed with zero mean. The distance dependence is contained in the parameters b_k and β_i (fixed slope and random slope, respectively). On the fixed part, the subscript k indicates an interaction between room and distance. If there is no interaction, b_k becomes a constant b. The presence of β_i indicates that the dependence of the response variable y on the distance d is different for each subject. The intercept ($a_k + \alpha_i$) adjusts the overall value of y, and it has a fixed part a_k and a random part α_i. The fixed intercept contains the effect of the room k on the response variable. The random part is also referred to as intersubject variability. The residual or unexplained variation ε_{ijk} is also regarded as a random effect. The standard deviations of the random effects α_i, β_i, and ε_{ijk} are notated as σ_α, σ_β, and σ_ε, respectively.

The actual models were built as simplifications of the "full model." First, the significance of the interaction (room-dependent slope b_k) was tested by means of likelihood ratio tests (using the function `anova` in R), comparing the outcomes of the full model and a reduced model without the interaction (constant slope b). If the full model was significantly better than the reduced model, the first one was kept. Otherwise, the reduced model was used. Another test for the suitability of random slopes was made by comparing the full model to another one with fixed slopes by means of a likelihood ratio test. In the same way, if the model with random slopes was significantly better than the one with fixed slopes, the first one was chosen. The suitability of including the basic variables (room and distance) was assessed by comparing the chosen model from the previous tests to a reduced version that only contained one variable (room or distance) with likelihood ratio tests. However, all the parameters showed dependence on the room and the distance. The models did not include a random effect for the room due to the subject.

The p-values for the overall models were calculated by means of likelihood ratio tests comparing the fit of the chosen model to the fit of a reduced model which only contained the random intercept due to the effect of the subject (and no dependence on room or distance). The p-values associated to each predictor and the standard deviations of the random effects were obtained with the function `pvals.fnc` (…, `withMCMC = T`) of the library `languageR` (Ref. 31) in R, which makes use of the Markov Chain Monte Carlo (MCMC) sampling method.

TABLE III. Fixed and random effects included in the mixed models. The fixed-effects are characterized for the intercepts a and slopes b, whereas the random effects have zero mean and only their standard deviation is shown. Abbreviations are used instead of the complete name of the rooms: ACH for the anechoic room, LH for the lecture hall, COR for the corridor, and REV for the reverberation room. Note that the b values for, $\bar{F}_0$ σ_{F_0}, and PTR are independent of the room.

	Fixed-effects								Random effects		
	a_k (Intercept)				b_k (Slope)				Intercept	Slope	Residual
Parameter	ACH	LH	COR	REV	ACH	LH	COR	REV	σ_α	σ_β	σ_ϵ
L_w [dB]	56.8	56.0	54.8	56.2	2.2	2.0	1.9	1.3	2.74	0.76	1.33
$\bar{F}_0[Hz]$	123.6	120.1	119.8	119.3	3.8				16.3	2.95	3.6
σ_{F_0} [Hz]	23.2	22.0	20.6	19.2	0.63				5.22	1.29	2.77
PTR	0.65	0.55	0.56	0.67	0.026				0.059	—	0.062

The choice of mixed models has the following basis: a considerable amount of the variance in the observations is due to the intersubject differences (which could be revealed with an analysis of variance table), so the subject is regarded as a random effect. Conceptually, it is similar to applying a normalization for each subject or regarding the subject as a factor in traditional statistical modeling.

III. RESULTS AND ANALYSIS

The measurements of L_W, $\bar{F}_0$, σ_{F_0}, and PTR were used to build four different linear mixed models according to Eq. (3). The coefficients for the intercepts and slopes corresponding to the fixed-effects of the models, together with the standard deviations of the random effects, are presented in Table III. The statistical significance (p-value) of the fixed-effects and interactions included in each model, along with the overall significance levels, is shown in Table IV.

A. Voice power level

The measured L_W, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 4. In the same figure, the lines show the fixed-effects part of the empirical model described in Eq. (3) and Table III. L_W depends almost linearly on the logarithm of the distance (with slopes between 1.3 and 2.2 dB per doubling distance) and changed significantly among rooms (intercepts between 54.8 and 56.8 dB). At each distance, the highest L_W was always measured in the anechoic room. A significant interaction was found between the room and the logarithm of the distance, because the variation of L_W with distance in the reverberation room (1.3 dB per doubling distance) was lower than the variation in the other rooms (1.9 to 2.2 dB per doubling distance). The standard deviation of the intersubject variation was estimated to be 2.7 dB, whereas the individual differences in the variation of L_W with distance had a standard deviation of 0.76 dB per doubling distance.

TABLE IV. Statistical significance and p-values of the fixed-effects and interactions considered in the empirical models and overall significance of the models. NS: Non-significant.

	Main effects		Interaction	
	log (distance)	Room	Room × log (distance)	Overall
L_W	<0.001	<0.001	0.009	<0.001
$\bar{F}_0$	<0.001	<0.001	NS	<0.001
σ_{F0}	0.10	<0.001	NS	<0.001
PTR	<0.001	<0.001	NS	<0.001

B. Fundamental frequency

Figure 5 shows the subject-averaged measured $\bar{F}_0$ (data points) and the corresponding empirical model (lines) described in Eq. (3) and Table III, for the different distances and rooms. $\bar{F}_0$ changed significantly among rooms (intercepts between 119.3 and 123.6 Hz) and had an almost linear dependence on the logarithm of the distance, with a slope of 3.8 Hz per doubling distance, identical for all the rooms. However, by visual inspection of Fig. 5, in the anechoic and reverberant rooms, there was less variation between the distances of 1.5 and 3 m than at further distances. $\bar{F}_0$ in the anechoic room was about 4 Hz higher than in the other rooms for all distances. The standard deviation of the intersubject variation was estimated in 16.3 Hz, whereas the individual differences in the variation of $\bar{F}_0$ with distance had a standard deviation of 2.95 Hz per doubling distance.

The measured σ_{F_0}, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 6. The lines in the figure show the fixed-effects part of the empirical model described in Eq. (3) and Table III. σ_{F_0} changed significantly among rooms (intercepts between 19.2

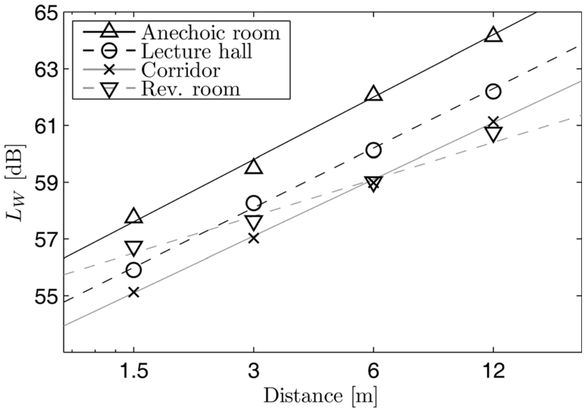

FIG. 4. Average voice power level used by the talkers at different distances to the listener. The lines show the predictions of the empirical model. The different slopes of the lines show an interaction between the room and the distance.

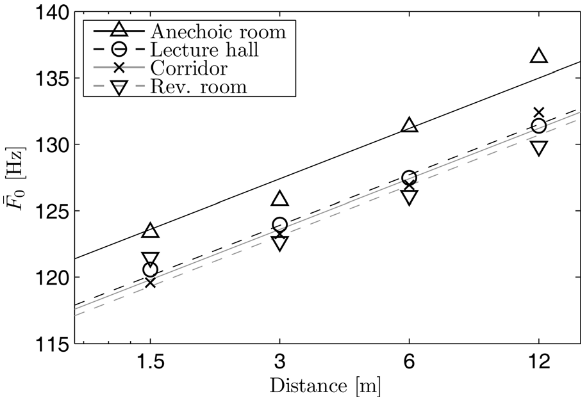

FIG. 5. Average mean fundamental frequency used by talkers at different distances to the listener. The lines show the predictions of the empirical model.

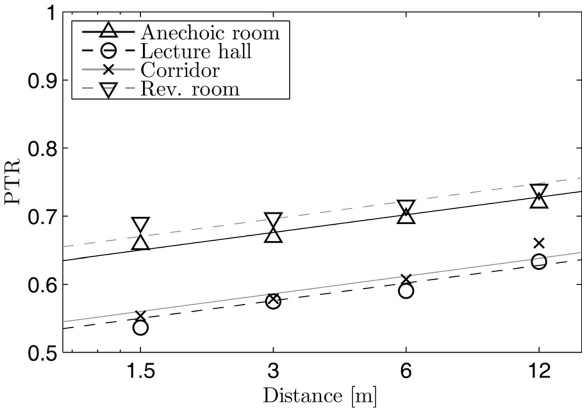

FIG. 7. Average PTR (relative appearance of voiced segments in running speech) used by talkers at different distances to the listener. The lines show the predictions of the empirical model.

and 23.2 Hz) and had a weak linear dependence on the logarithm of the distance, with a slope of 0.63 Hz per doubling distance, equal among the rooms. The standard deviation of the intersubject variation was estimated in 5.22 Hz, whereas the individual differences in the variation of σ_{F_0} with distance had a standard deviation of 1.29 Hz per doubling distance. The latter value is larger than the fixed-effect slope (0.63 Hz) which means that, for a number of subjects, σ_{F_0} decreased with distance. This is the reason for the low statistical significance of the σ_{F_0} dependence with the logarithm of the distance shown on Table IV. Therefore, the amount of σ_{F_0} change as a function of distance was mainly an individual factor.

C. PTR

The measured PTR, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 7. In the same figure, the lines show the fixed-effects part of the empirical model described in Eq. (3) and Table III. PTR had a weak linear dependence on the logarithm of the distance (with a slope of 0.026 per doubling distance, equal for all rooms) and changed significantly among rooms, especially between two groups: one formed by the anechoic room and the reverberation room (intercepts 0.65 and 0.67) and a second group formed by the lecture hall and the corridor (intercepts 0.55 and 0.56). The standard deviation of the intersubject variation was estimated in 0.059. The change in PTR with distance was not significantly different among subjects, so the model does not include a random slope.

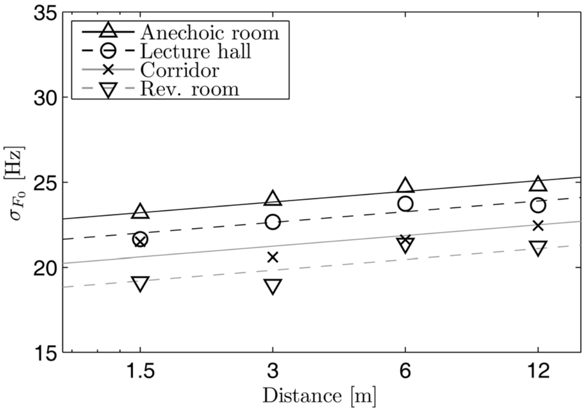

FIG. 6. Average long-term standard deviation of the fundamental frequency used by talkers at different distances to the listener. The lines show the predictions of the empirical model.

D. Subjective impressions

The talkers expressed their opinions verbally about the experience of talking in the different rooms. One general comment was that the anechoic chamber was an unnatural place to speak in, due to the lack of sound reflections, and that they felt moved to raise their vocal intensity to make themselves heard at the listener location, and for this reason, it was not a comfortable environment for talking. The reverberation room was very unpleasant for speaking, due to the excessive reverberation. Talkers admitted that they had to modify their speech strategy to compensate for the poor acoustic conditions. A few of the subjects preferred overall the corridor, due to the sensation of support or being helped by the room to reach longer distances without having to increase their voice level too much, although they pointed out some acoustical deficiencies like a noticeable echo. Most of the subjects preferred the lecture hall for speaking. However, they admitted that it was demanding to talk at the longest distance (12 m). Many subjects commented that the acoustic conditions of the experimental rooms were not the desirable ones in rooms for speech.

IV. DISCUSSION

Figures 4 to 7 show the variation of the measured parameters (L_W, $\bar{F}_0$, σ_{F_0}, and PTR) with distance and across environments. As all of the measured parameters indeed have variation with distance and acoustic environment, they are potential indicators of vocal effort.

The measurements shown in Fig. 4 reveal that the average variations of L_W when the distance increases from 1.5 to 12 m are in the range between 3.9 dB in the reverberation room and 6.6 dB in the anechoic room. These variations are mainly the consequence of a conscious decision of the talker to raise the voice level as a response to a change in communication distance. However, the fact that the compensation

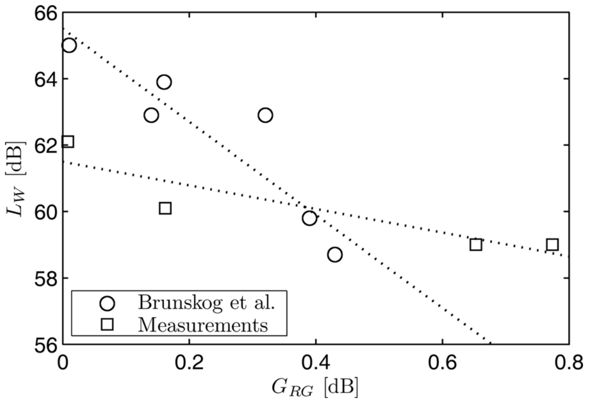

FIG. 8. Average L_w at 6 m vs room gain G_{RG}, as compared to the results of Brunskog *et al.*

rates differ among rooms shows the influence of auditory feedback in voice level adjustment. Furthermore, the effect of room on L_W varies between 2 dB at 1.5 m and 3.3 dB at 12 m. These values are smaller but comparable to the effect of distance on L_W. Thus, the perception of one's own voice via reflections in the room boundaries is important for voice level regulation, together with the direct air transmission and the bone-conducted components, as Siegel and Pick[32] stated.

Brunskog *et al.* used G_{RG} as a metric to quantify the importance of the reflected sound from one's own voice. This measure is indeed a measure of sidetone (one's own voice reaching the ears) amplification. Taking the subject-averaged L_W values measured at 6 m, a distance which is representative of a lecturing scenario, the least squares regression model using G_{RG} as a predictor is

$$L_{W,6} = 61.5 - 3.56 \times G_{RG}. \tag{4}$$

The R^2 for this regression model is 0.82, whereas the p-value is 0.09. The L_W values, with the regression line (4), are compared to the results of Brunskog *et al.*[12,13] in Fig. 8. The slope of the regression line in the current measurements is much lower than the slope obtained by Brunskog *et al.* (−3.6 dB/dB vs −13.5 dB/dB). The difference between slopes might be explained by the fact that the distance was not taken into account by Brunskog *et al.* In their study, the rooms with high G_{RG} values were small rooms where the listeners stood close to the talker whereas the rooms with low room gain were larger and the listeners stood far from the talker. Thus, there is an unwanted correlation between the room gain and the distance, due to the experimental design, but which is found in typical real rooms. The model from Brunskog *et al.* predicts L_W in a general situation with varying distance to the listeners, but the model (4) accounts for the variation due exclusively to changes in auditory feedback.

As in some studies of sidetone amplification,[33] L_W decreases with increasing sidetone amplification (estimated by G_{RG}). However, there are two differences between these studies and the present study. One is the range of L_W variation and the second is the magnitude of the effect. In the present study, talkers raised L_W by 3.2 dB on average while speaking in the anechoic room at a distance of 12 m, compared to the reverberant room. In other studies of voice production with altered sidetone, variations in voice level of up to 20 dB were reported. In these studies, the sidetone was altered by inducing temporary hearing loss on the subjects, thus decreasing all components of sidetone (direct, reflected, and bone-conducted sound) or attenuating the airborne sound while bone conduction is preserved. The significantly different ranges of voice level variation obtained in the previous studies (up to 20 dB) and in this study (approximately 3.2 dB by the effect of room) might be due to the fact that only the reflected component was changed in this study, while the direct and bone-conducted components of the talker's own voice were kept unchanged. Therefore, the overall sidetone variations were much smaller than in the other studies. The magnitude of the effect on traditional sidetone compensation was in the range between −0.25 and −0.57 dB/dB, whereas in the present study the magnitude of the effect was −3.6 dB/dB, as can be seen in Eq. (4). These differences could be explained by two alternative hypotheses. The first is that the changes in L_W are purely due to the Lombard effect and that the room reflections alter the loudness of one's own voice to a greater extent than indicated by the single figure G_{RG}. The second is that there are additional psychological attributes related to room perception affecting the voice regulation at a cognitive level, through internal feedback mechanisms.

The measured compensation rates for L_W due to changes in distance between talker and listener were between 1.3 dB/dd in the reverberation room and 2.2 dB/dd in the anechoic chamber. These compensation rates are much lower than the ones obtained by Warren,[15] Healey *et al.*,[17] and Traunmüller and Eriksson.[2] However, they are closer to other studies[16,18] and especially close to the 1.8 dB/dd measured indoor by Zahorik and Kelly.[19] Differences from the previous studies might arise from the selection of subjects or different instruction. In the present study, there were significant differences in vocal behavior among subjects, indicated by the random slope effect in Table III, which predicts a standard deviation of 0.76 dB/dd over the fixed slopes 1.3 to 2.2 dB/dd. In any case, the individual compensation rates were not as large as 6 dB/dd.[15,19] In addition, natural speech was evoked in the present experiment by means of the map task, which resulted

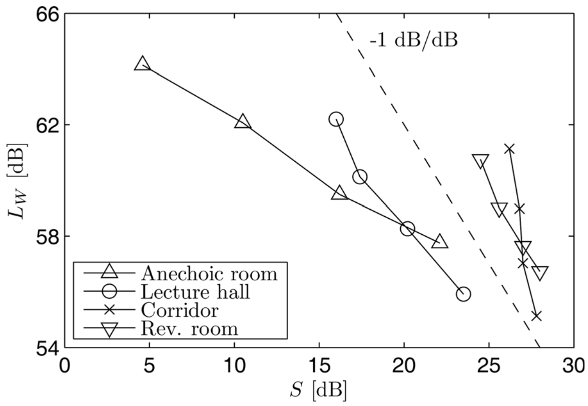

FIG. 9. Voice power level vs speech sound level S at the listener's position. The dashed line has a slope of −1 dB/dB. If the L_w values laid in a line with the same slope, talkers would be providing a constant SPL at the listener position.

in lower compensation rates than would be obtained by using short vocalizations, as Michael *et al.*[16] stated.

Figure 9 shows the relationship between the L_W produced by the talkers and the sound speech level S at the listener position, which is an alternative representation of the data in Fig. 4. The dashed line in Fig. 9 represents the theoretical L_W values that would keep the SPL constant at the listener position. According to Zahorik and Kelly,[19] if talkers accurately compensated for the sound propagation losses—providing an almost constant average SPL at the listener position—the expected L_W would lay exactly on top of a line with the same slope as the dashed line, meaning that a talker would lower L_W by 1 dB whenever S increases by 1 dB. The L_W data points in Fig. 9 follow approximately straight lines with different slopes for each room: -0.4 dB/dB in the anechoic chamber, -0.8 dB/dB in the lecture hall, -1.1 dB in the reverberation room, and -3.8 dB/dB in the corridor. In the lecture hall and the reverberation room, talkers approximately compensated for sound propagation losses. However, there was an undercompensation in the anechoic chamber, meaning that the SPL produced at the listener position decreased with distance, and an overcompensation in the corridor, where the SPL increased with the distance. Undercompensation appears to take place in rooms with big differences of S between short and long distances, i.e., rooms with dominating direct sound. Overcompensation takes place in rooms where differences in S at short and long distances were small, i.e., rooms with strong reverberant field. Undercompensation and overcompensation were present because the talkers were not explicitly asked to compensate for sound propagation losses, and many of the talkers were not used to talk in the environments of the study. It is presumed that talkers would be able to compensate for sound propagation losses with an explicit instruction and training to get acquainted with the acoustical properties of each room.

Compensation rates have a meaning when the distance between talker and listener is well defined, such as in a face-to-face conversation. In the case of a distributed audience, as in the usual teaching context, the situation is more complex and it is not clear what is the distance estimation of the talker. In that case, according to Brunskog *et al.*,[12,13] talkers apparently adjust their voice levels guided by the room gain or degree of amplification provided by the room at their ears (Fig. 8).

The changes in $\bar{F}_0$ were similar to those in L_W, as both parameters increased linearly with the logarithm of the distance, and it was in the anechoic room where the highest $\bar{F}_0$ were obtained at each distance. Table III shows that $\bar{F}_0$ changed 3.8 Hz by doubling the distance and was 4 Hz higher in the anechoic room than in the other rooms. In simplified terms, the extra vocal effort demanded to speak in the anechoic room is comparable to the effect of doubling the distance to the listener in other rooms. However, the changes among other rooms (maximum of 0.8 Hz) were not as important so as to ascribe a significant effect to the room. It seems more likely that the unfamiliarity of talkers with the anechoic room accentuated some changes in speech production too much, which are not observed in everyday rooms. Nevertheless, $\bar{F}_0$ is an important measure of vocal effort to show that, at long communication distances, the number of vocal fold vibrations (or collisions) increases, which leads to higher vocal doses that might eventually result in vocal fold trauma.

The talkers had the general remark that the anechoic room and the reverberation room were the most uncomfortable environments to speak in. Both environments were the two most extreme rooms in terms of T_{30}, STI, and G_{RG}, as shown in Table I. The anechoic chamber demanded an increased vocal effort due to lacking support, with a G_{RG} value of 0.01 dB. On the other hand, it was very unpleasant and stressing to speak in the reverberation room, which could be explained by the remarkably lower STI value (only 0.67) corresponding to the transmission between mouth and ears. Talkers' comments suggest that there is a compromise between STI and G_{RG}, in order for rooms to be comfortable. The poor vocal comfort rating for the reverberation room cannot be explained by the measured L_W or $\bar{F}_0$, as the L_W and $\bar{F}_0$ in this room were not higher than the values measured in the lecture hall and the corridor, the most preferred rooms. This observation supports the idea that the concepts of vocal effort and comfort are not exactly opposite.

As shown in Fig. 6 and Table III, the model predicted significant differences in σ_{F_0} among the environments for all distances. The highest σ_{F_0} was found in the anechoic room, followed by those in the lecture hall, the corridor, and the reverberation room, in reverse order to the reverberation times: the reverberation room, the corridor, the lecture hall, and the anechoic chamber (in decreasing order), or in the same order as the STI. According to this observation, speech produced in acoustically live rooms is more monotonous (meaning low variability in F0) than in acoustically dry rooms. The extreme values of σ_{F_0} were obtained in the least preferred rooms. The highest σ_{F_0} in the anechoic room might be an indication of increased vocal demands (increased L_W and $\bar{F}_0$), whereas the low σ_{F_0} in the reverberant room might be an observable feature of the speech produced under low STI conditions. However, this assertion needs to be proved in a broader range of acoustic conditions.

In Fig. 7, the average PTR was remarkably different between two groups of environments and correlated well with the subjective impressions of talkers regarding vocal comfort. The highest PTR values were measured in the most uncomfortable rooms (0.67 in the reverberation room and 0.65 in the anechoic room), whereas the PTR in the other two rooms was significantly lower (0.55 in the lecture hall and 0.56 in the corridor). The increased voice levels or vocal efforts explain the high values obtained for the anechoic chamber, as Lienard and Di Benedetto[18] also reported. However, the high PTR obtained in the reverberation room might be due to the adaptation of the talker to the environment. It seems that talkers tried to improve the speech intelligibility in such a reverberant environment by separating the consonant segments of their speech, resulting in longer vocalic segments.

V. CONCLUSIONS

The present paper studies the changes in different speech parameters (voice power level, fundamental frequency, PTR) describing vocal effort when talkers addressed a single listener at different distances under various room acoustic conditions in the absence of background noise. The main conclusions are as follows:

(1) The decision of using a certain voice level depends on the visually perceived distance to the listener and varies between 1.3 and 2.2 dB per double distance to the listener.

(2) The room acoustic conditions modify the auditory feedback of the talker's own voice, inducing significant changes in voice level with an approximately linear dependence on the amplification of the room to one's own voice, given by the magnitude "room gain," at a rate of -3.6 dB/dB.

(3) The mean fundamental frequency increases with distance at a rate of 3.8 Hz per double distance to the listener and is 4 Hz higher in anechoic conditions.

(4) A room that provides vocal comfort requires a compromise between room gain and STI, supporting the voice from a talker but not degrading the perceived speech quality.

(5) The standard deviation of the fundamental frequency and the relative duration of voiced segments in a running speech signal might be symptomatic indicators of vocal comfort in a room.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to all the participants in the measurements, to one anonymous reviewer that enabled significant improvements in this work, and to Finn Jacobsen and Torben Poulsen for their positive feedback toward improving the quality of the paper. This research has been partially funded by the Swedish company AFA Försäkring as a part of the project "Speaker's comfort and voice health in classrooms."

[1] L. Raphael, G. Borden, and K. Harris, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, 5th ed. (Lippincott Williams & Wilkins, Baltimore, 2007), pp. 167–176.

[2] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," J. Acoust. Soc. Am. **107**, 3438–3451 (2000).

[3] I. Titze, *Principles of Voice Production* (National Center for Voice and Speech, Iowa City, 2000), pp. 49–51.

[4] V. Lyberg-Åhlander, R. Rydell, and A. Löfqvist, "Speaker's comfort in teaching environments: Voice problems in Swedish teaching staff," J. Voice. (2010) Available online March 26, 2010.

[5] J. Preciado-López, C. Perez-Fernández, M. Calzada-Uriondo, and P. Preciado-Ruiz, "Epidemiological study of voice disorders among teaching professionals of La Rioja, Spain," J. Voice **22**, 489–508 (2008).

[6] N. Roy, R. Merrill, S. Thibeault, R. Parsa, S. Gray, and E. Smith, "Prevalence of voice disorders in teachers and the general population," J. Speech Lang. Hear. Res. **47**, 281–293 (2004).

[7] A. Russell, J. Oates, and K. M. Greenwood, "Prevalence of voice problems in teachers," J. Voice **12**, 467–479 (1998).

[8] E. Vilkman, "Voice problems at work: A challenge for occupational safety and health arrangement," Folia Phoniatr. Logop. **52**, 120–125 (2000).

[9] I. R. Titze, J. G. Svec, and P. S. Popolo, "Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues," J. Speech Lang. Hear. Res. **46**, 919–932 (2003).

[10] L. Rantala, E. Vilkman, and R. Bloigu, "Voice changes during work: Subjective complaints and objective measurements for female primary and secondary school teachers," J. Voice **16**, 344–355 (2002).

[11] M. Hodgson, R. Rempel, and S. Kennedy, "Measurement and prediction of typical speech and background noise levels in university classrooms during lectures," J. Acoust. Soc. Am. **105**, 226–235 (1999).

[12] J. Brunskog, A. Gade, G. Payá-Ballester, and L. Reig-Calbo, "Increase in voice level and speaker comfort in lecture rooms," J. Acoust. Soc. Am. **125**, 2072–2082 (2009).

[13] D. Pelegrín-García, "Comment on "Increase in voice level and speaker comfort in lecture rooms" [J. Acoust. Soc. Am. **125**, 2072–2082 (2009)] (L)," J. Acoust. Soc. Am. 129 (2011).

[14] M. Kob, G. Behler, A. Kamprolf, O. Goldschmidt, and C. Neuschaefer-Rube, "Experimental investigations of the influence of room acoustics on the teachers voice," Acoust. Sci. & Tech. **29**, 86–94 (2008).

[15] R. Warren, "Vocal compensation for change in distance," in *Proceedings of the 6th International Congress of Acoustics* (International Commission for Acoustics, Tokyo, 1968), pp. 61–64.

[16] D. Michael, G. Siegel, and H. Pick, Jr., "Effects of distance on vocal intensity," J. Speech Hear. Res. **38**, 1176–1183 (1995).

[17] E. C. Healey, R. Jones, and R. Berky, "Effects of perceived listeners on speakers' vocal intensity," J. Voice **11**, 67–73 (1997).

[18] J. S. Liénard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422 (1999).

[19] P. Zahorik and J. W. Kelly, "Accurate vocal compensation for sound intensity loss with increasing distance in natural environments," J. Acoust. Soc. Am. **122**, EL144–EL150 (2007).

[20] 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).

[21] International Organization for Standardization, *ISO-3382:2009, Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces* (ISO, Geneva, Switzerland, 2009).

[22] 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 October 16, 2010).

[23] M. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am. **37**, 409–412 (1965).

[24] A. Farina, "Aurora plug-ins," available at http://www.aurora-plugins.com (Last viewed October 12, 2010).

[25] H. Lazarus, "Prediction of verbal communication in noise—A review: Part 1," Appl. Acoust. **19**, 439–463 (1986).

[26] M. Barron, *Auditorium Acoustics and Architectural Design* (Taylor & Francis, London, 1993), pp. 223–240.

[27] K. Sjölander and J. Beskow, "WAVESURFER," Stockholm: Centre for Speech Technology (CTT) at KTH, available at http://sourceforge.net/projects/wavesurfer/ (Last viewed October 16, 2010), (2000).

[28] D. Hedeker, "Generalized linear mixed models," in *Encyclopedia of Statistics in Behavioral Science*, edited by B. Everitt and D. Howell, 2nd ed. (Wiley, New York, 2005).

[29] D. Bates and M. Maechler, *lme4: Linear mixed-effects models using S4 classes* (2010), available at http://CRAN.R-project.org/package=lme4 (Last viewed October 16, 2010), R package version 0.999375–33.

[30] R Development Core Team, *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna, Austria (2010), available at http://www.R-project.org (Last viewed October 16, 2010), ISBN 3-900051-07-0.

[31] R. H. Baayen, *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"* (2009), available at http://CRAN.R-project.org/package=languageR (Last viewed October 16, 2010), R package version 0.955.

[32] G. Siegel and H. Pick, Jr., "Auditory feedback in the regulation of voice," J. Acoust. Soc. Am. **56**, 1618–1624 (1974).

[33] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," J. Speech Lang. Hear. Res. **14**, 677–709 (1971).