ORIGINAL ARTICLE

Taylor&Francis
healthsciences

Measurement of vocal doses in speech: experimental procedure and signal processing

Jan G. Švec[1], Peter S. Popolo[1,2] and Ingo R. Titze[1,2]

From the [1]National Center for Voice and Speech, the Denver Center for the Performing Arts, 1245 Champa Street, Denver, CO 80204, USA, and [2]Department of Speech Pathology and Audiology, The University of Iowa, 330-WJSHC, Iowa City IA 52242, USA

Received 31 March 2003. Accepted 4 September 2003.

Logoped Phoniatr Vocol 2003; 28: 181–192

An experimental method for quantifying the amount of voicing over time is described in a tutorial manner. A new procedure for obtaining calibrated sound pressure levels (*SPL*) of speech from a head-mounted microphone is offered. An algorithm for voicing detection (k_v) and fundamental frequency (F_0) extraction from an electroglottographic signal is described. The extracted values of *SPL*, F_0, and k_v are used to derive five vocal doses: the time dose (total voicing time), the cycle dose (total number of vocal fold oscillatory cycles), the distance dose (total distance travelled by the vocal folds in an oscillatory path), the energy dissipation dose (total amount of heat energy dissipated in the vocal folds) and the radiated energy dose (total acoustic energy radiated from the mouth). The doses measure the vocal load and can be used for studying the effects of vocal fold tissue exposure to vibration.

Key words: electroglottography, F_0 extraction, sound pressure level, speech processing, vocal dose, vocal loading, voice accumulation, voicing detection, voicing time.

Jan G. Švec, PhD, National Center for Voice and Speech, The Denver Center for the Performing Arts, 1245 Champa Street, Denver, CO 80204, USA. Tel.: +1-303-389-4694. Fax: +1-303-893-6487. E-mail: jsvec@dcpa.org or svecjan@vol.cz

INTRODUCTION

In order to investigate effects of prolonged or excessive vocalization among those who use their voice as an occupational tool, (i.e., teachers, salespeople, phone operators, actors, etc.) it is important to properly quantify the amount of vocalization. Although basic, the problem of proper measurement of the amount of voicing is not a trivial one. There are at least three fundamental factors that can be considered important in prolonged voice use: the duration of voicing, vocal intensity and fundamental frequency. The relative importance of these factors with respect to each other has not, however, been well understood and questions such as: 'Is high-pitched voice potentially more hazardous than low-pitched voice?' or 'Is it more demanding to use soft voice for a long time than a loud voice for a shorter time?' have been difficult to answer.

Exposure of human tissue to factors such as sun radiation or chemicals has traditionally been quantified by a measure called 'dose'. Since prolonged vocal use can be thought of as a problem of exposure of vocal folds to vibration, the term 'vocal dose' has been adopted for measures quantifying the amount of voicing. The theory of the vocal dose measures has been introduced in a previous paper by Titze, Švec and Popolo (1), which provided the definitions of various doses as well as their preliminary normative values for male and female speakers. The present paper complements the previous paper in that it provides a detailed description of the experimental procedures and data-processing algorithms that can be used for experimental vocal dose measurement. This information was not provided in sufficient detail in the previous paper (due to the limited number of pages available). Here, the definitions of the different vocal doses will be reviewed and a method for extracting fundamental frequency (F_0), calibrated sound pressure levels (*SPL*) and unvoiced segments from a speech signal will be offered and used to measure the vocal doses. These descriptions should provide sufficient information for eventual repetition of the experiment and confirmation of the results described in Titze, Švec and Popolo (1) as well as to provide a method that could be adapted and further improved for studying the effects of long-term exposure of vocal folds to vibration.

 ISSN 1401-5439
DOI: 10.1080/14015430310018892

THE DOSE DEFINITIONS

Five vocal doses have been defined, considering various factors that can potentially contribute to voice problems. The simplest vocal dose is the *Time Dose*, D_t, which is identical with the measure known as 'voicing time' or 'vocal accumulation time' (2–7). It accumulates the total time the vocal folds vibrate and is defined as (1):

$$D_t = \int_0^{t_m} k_v \, dt \quad seconds \qquad \text{(eq. 1)}$$

where t_m is the total measurement time and k_v is the voicing unit step function:

$$k_v = \begin{cases} 1 & for\ voicing \\ 0 & for\ non\text{-}voicing \end{cases} \qquad \text{(eq. 2)}$$

It is often useful to relate the time dose D_t to the total measurement time t_m and calculate the *Voicing Ratio*, D_t/t_m, to obtain information on the fraction of total time for which the vocal folds were vibrating. When the voicing ratio is multiplied by 100 one obtains the *Voicing Percentage*.

A second dose measure is the *Cycle Dose*, D_c, which quantifies the total number of oscillatory periods completed by the vocal folds over time. It is defined as

$$D_c = \int_0^{t_m} k_v F_0 \, dt \quad cycles \qquad \text{(eq. 3)}$$

where F_0 is the fundamental frequency of the vocal fold oscillation in Hz. Such a dose was first used by Rantala and Vilkman (8) under the name *Vocal Loading Index* (*VLI*). Since the number of cycles is high (hundreds per second) the *VLI* was adapted to measure the dose in the units of thousands of cycles; then it holds that $VLI = D_c/1000$. The exposure measured by the vocal loading index is higher for subjects with a higher speaking pitch, as equation 3 clearly shows.

In order to account also for amplitude of vibration, the *Distance Dose* D_d was introduced as a third candidate. This dose measures the total distance accumulated by the vocal folds in a cyclic path during vibration. Its definition is (1):

$$D_d = 4 \int_0^{t_m} k_v A F_0 dt \quad meters \qquad \text{(eq. 4)}$$

where A is the amplitude of the vocal folds. Because the vocal folds theoretically travel a distance of four times the amplitude within a cycle (i.e., from neutral position to maximal displacement and back, followed by a similar excursion in the opposite direction), there is the factor 4 in the equation. Since the amplitude of the vocal folds changes with vocal intensity, the distance dose accounts for both the intensity and the fundamental frequency in voicing.

The problem with the distance dose is that the vibration amplitude of the vocal folds is very difficult to measure. To overcome this problem, the amplitude can be approximated from the *SPL* and F_0 using existing normative data. This means that, rather than calculating the distance dose for the particular person, we calculate a distance dose typical for an average person when speaking at the measured *SPL* and F_0. The amplitude A can be approximated using the empirical rules derived in (1):

$$A = 0.05 L_0 [(P_L - P_{th})/P_{th}]^{1/2} \quad m \qquad \text{(eq. 5)}$$

where L_0 is a reference vocal fold length (0.016 m for males and 0.01 m for females), P_L is the lung pressure and P_{th} is the phonation threshold pressure. The empirical rule for the phonation threshold pressure was found by Titze (9):

$$P_{th} = 0.14 + 0.06(F_0/F_{0N})^2 \quad kPa \qquad \text{(eq. 6)}$$

where F_0 is the fundamental frequency and F_{0N} is a nominal (speaking) fundamental frequency (120 Hz for males and 190 Hz for females). The last empirical rule needed for the determination of the distance dose is the rule for the lung pressure, which is, in accordance with Titze and Sundberg (10):

$$P_L = P_{th} + 10^{(SPL-78.5)/27.3} \quad kPa \qquad \text{(eq. 7)}$$

This rule was derived for *SPL* measured at the distance of 50 cm from the mouth.

A fourth dose is the *energy dissipation dose* D_e, which takes into account the factor of thermal agitation of tissue inside the vocal folds and measures the amount of heat produced in the vocal folds during vibration. It can be calculated as (1):

$$D_e = \frac{1}{2} \int_0^{t_m} k_v \eta (A/T)^2 \omega^2 dt \quad joules/m^3 \qquad \text{(eq. 8)}$$

where η is the shear viscosity of the vocal fold tissue (11, 12), T is vertical thickness of the vocal folds and $\omega = 2\pi F_0$ is the angular frequency of the vocal fold vibration. The shear viscosity η and the vertical thickness T can be further approximated from the F_0 of voice using the empirical rules (1):

$$\eta = \left\{ \begin{array}{l} 5.4/F_O \text{ for males} \\ 1.4/F_O \text{ for females} \end{array} \right\} \textit{pascal seconds} \qquad \text{(eq. 9)}$$

$$T = \left\{ \begin{array}{l} \dfrac{0.0158}{1 + 2.15\,(F_O/120)^{1/2}} \text{ for males} \\ \dfrac{0.01063}{1 + 1.69\,(F_O/190)^{1/2}} \text{ for females} \end{array} \right\} \textit{meters} \qquad \text{(eq. 10)}$$

A fifth dose, which is not a measure of exposure to the vocal folds but rather a potential sound exposure to a listener, is the *radiated energy dose* D_r. It quantifies the total energy radiated from the mouth over time. It is calculated as (1):

$$D_r = 4\pi R^2 \int_0^{t_m} k_v 10^{(SPL-120)/10} dt \quad \textit{joules} \qquad \text{(eq. 11)}$$

where R is the distance from the mouth (in our case 0.5 m) at which the sound pressure level (SPL) of voice is registered. To summarize the information contained in equations 1–11, the parameters needed for the calculation of the different doses are listed in Fig. 1.

Using these definitions and empirical rules, all the doses can be derived, for a specified measurement time t_m, by extracting three basic parameters of speech: k_v (voicing-unvoicing parameter), F_0 and SPL. The process of deriving all the necessary parameters for dose calculations is schematically summarized in Fig. 2.

The time, cycle and radiated energy doses are the true doses for the person measured, whereas the distance dose and the dissipated energy dose are approximations based on typical data for male and female vocal fold amplitudes, thicknesses and viscosities.

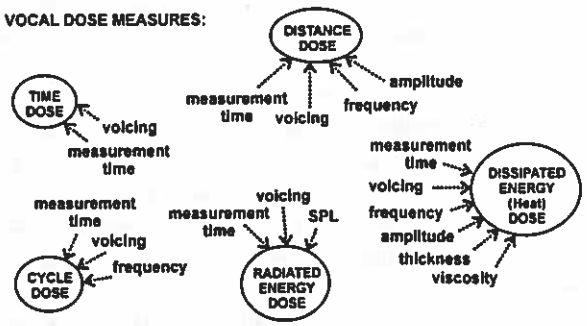

Fig. 1. Parameters needed to measure the five vocal doses.

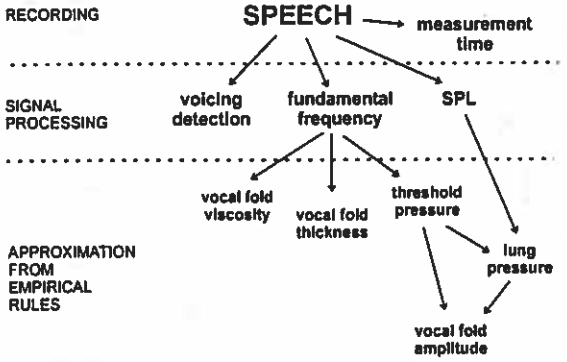

Fig. 2. Schematic summary of the procedure for deriving the speech parameters needed for dose measurement.

EXPERIMENTAL METHOD

There are two basic settings at which the vocal dose measurement can take place—in laboratory experiments and in the occupational workplace. The two conditions require slightly different approaches. Here, a method suitable for a laboratory setting is described, which takes advantage of standard speech instrumentation available in laboratories. This method was used to derive the preliminary normative doses reported in Titze, Švec and Popolo (1). An electroglottographic speech signal was used to determine the individual periods of vocal fold oscillation and to derive the F_0 values and the k_v values of speech. Sound level meter and a head-mounted microphone were used to derive the absolute SPL values of speech.

Equipment set-up

The equipment set-up is shown in Fig. 3. The subject was placed in a sound booth (single wall IAC isolation booth, 2.3 m deep, 2.2 m wide and 2.3 m high) wearing a head-mounted microphone (Shure WH20) with the transducer element off to the side of the mouth and out of the air stream. The mouth-to-microphone distance was about 5 cm, but the distance was not critical as long as it remained constant throughout the recording session, since the microphone signal was ultimately related to SPL at 50 cm through a calibration procedure. The microphone signal was led out of the sound booth where it was amplified with a microphone preamp (Digital Sound Corp., DSC-240). The amplified signal was recorded on the first channel of a digital audio tape (DAT) recorder (Technics SV-DA10) at the sampling rate of 48 kHz.

A sound level meter (Brüel & Kjær 2238, set to linear frequency weighting and fast frequency response) was positioned at the distance of 0.5 m from the mouth. The acoustic signal registered by the sound level meter was led out of the sound booth and

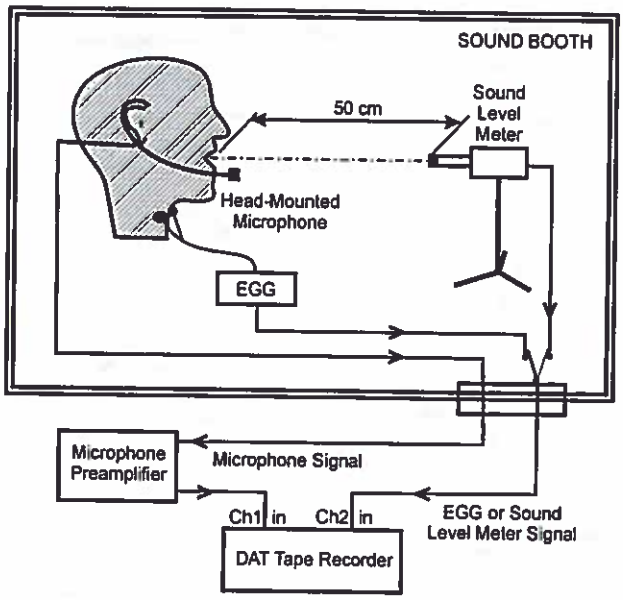

Fig. 3. The experimental set-up.

recorded on the second channel of the DAT recorder during the *SPL* calibration procedure. An electroglottograph (Laryngograph) with electrodes placed on the neck of the subject was used to register the vocal fold contact area. The electroglottographic (EGG) signal was recorded on the second channel of the DAT recorder (replacing the sound level meter signal) during the speech tasks.

Sound level meter issues

A sound level meter with a built-in microphone has traditionally been used for absolute measurement of the sound level of speech at prescribed distance from the mouth (in our case, 50 cm). To ensure correct *SPL* measurement of voice with the sound level meter, the mouth-to-microphone distance shall be constant throughout the measurement and the voice signal shall be minimally contaminated by the acoustic reflections of the room, i.e., the critical distance (reverberation radius) of the room (13) shall be larger than the mouth-to-microphone distance (preferably twice as large). The reverberation radius of our sound booth was calculated to be *c.* 73 cm at 500 Hz, which was not ideal but still acceptable for using the 50-cm reference distance.

Sound level meters are by design, however, rather sensitive to ambient noise that interferes with the voice signal, making the measurement difficult (especially for the soft voice). In our sound booth, the ambient noise level was around 54–55 dB (at linear frequency weighting), which is very close to the normal values of soft speech. A similar or worse situation can be expected in rooms with non-optimal sound isolation.

The disturbing factor of ambient noise can be diminished by using the standard A frequency weighting (14), which attenuates the low-frequency components that are usually strongly present in the ambient noise. Unfortunately, the frequency weighting also influences the frequency spectrum of the speech and changes the resulting sound pressure level of speech. Since the empirical rules for vocal doses are valid only for linearly weighted signals, the use of A frequency weighting is not justifiable if these rules are to be used for dose measures. The C frequency weighting of the sound level meter would be acceptable since it is flat in the spectral frequency range of human voice (less than 1 dB variation in the range 70–5000 Hz (14)), but the C filter is often insufficient in lowering the low frequency ambient noise to levels allowing *SPL* measurement of the softest phonations at the 50 cm distance.

Two-step calibration procedure for the head-mounted microphone

In order to solve the problem of the ambient noise, the head-mounted microphone described above, placed at the distance of about 5 cm from the mouth, was used and its signal level was related to the sound pressure level of the sound level meter at 50 cm distance through a two-step calibration procedure. The use of the head-mounted microphone decreased the noise level by about 10 dB (suppression of the ambient noise by up to 20 dB could theoretically be expected under ideal conditions), which was effectively achieved by picking up the voice signal at a shorter distance than that of the sound level meter. Winholtz and Titze (15) described similar method of relating the sound level meter to the head-mounted microphone signal. The method presented here is different in that it does not require placement of an external sound source in the proximity of the mouth and avoids the process of exponential time averaging of the head-mounted microphone signal used by the above-mentioned authors.

The two steps of the calibration procedure were as follows: 1) A steady calibration tone with the sound level of 94 dB (produced by the calibrator B&K 4231) was registered with the sound level meter. 2) The subject then produced a steady, sustained */a/* which was registered simultaneously with the sound level meter at 50 cm and the head-mounted microphone. All the signals were recorded on a DAT tape (Fig. 4).

To ensure an accurate *SPL* calibration of the head-mounted microphone, the frequency response of the head-mounted microphone shall be flat in the relevant spectral range of the voice (70–5000 Hz as minimum). This is usually best achieved with omnidirectional

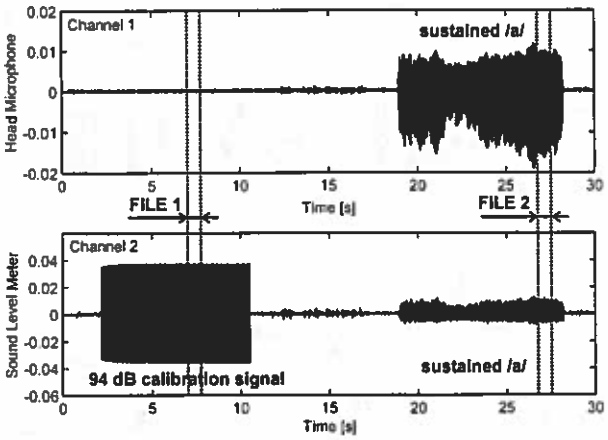

Fig. 4. Two-step calibration procedure for absolute *SPL* measurement with a head-mounted microphone: display of the head-mounted microphone and sound level meter microphone signals recorded simultaneously in the two audio channels. A steady calibration tone (in this case 94 dB *SPL re 20 μPa*) was delivered to the sound level meter, then the subject produced sustained /a/. One second of the calibration signal was selected and saved as the calibration file 1. This file was used to determine the absolute *SPL*s of the digitized sound level meter microphone signal. The most steady part of the sustained /a/ of 1 s duration was selected and saved as the calibration file 2. That file was used to adjust the head-mounted microphone signal to give the same *SPL* as the sound level meter microphone signal.

microphones whereas directional microphones usually suffer from the so-called 'proximity effect' resulting in a boost/attenuation of lower frequencies when the mouth-to-microphone distance is shortened/lengthened (16–18). The frequency response of our Shure WH20 head-mounted microphone was not ideal; it showed a slight drop-off at low frequencies (about 5dB at 100 Hz). This resulted in slightly lower *SPL*s at low-pitched soft voices, but errors for the dose calculations resulting from the use of this microphone are estimated to be only a few percent (soft phonations influence the total doses less than loud phonations) and will be neglected here.

Speech recording

After the two-step *SPL* calibration procedure was finished and the signals recorded, the sound level meter signal on the DAT recorder was replaced by the EGG signal and the speech task was recorded. When recording the speech signals, all the signal controls for the head-mounted microphone were kept exactly at the same levels as they were during the calibration. For the purpose of the present study a segment of Goldilocks passage (19) read by a male subject was chosen, which was used also in our previous study (1). The reader is referred to that study for the results in different male and female subjects. After the reading was finished, the EGG signal on the second channel of the DAT recorder was again replaced by the sound level meter signal and the *SPL* calibration procedure (prolonged steady */a/* vowel and *SPL* calibration tone) was repeated for double-check purposes.

Files for signal processing

The recorded signals were digitally transferred from the digital tape to the computer using the DAT interface of the Computerized Speech Lab (model 4400, Kay Elemetrics) and the CSL software. The original sampling rate of 48 kHz of the DAT recorder was kept also for the computer sound files. From the digitized signals, three separate audio files were created and saved in a standard stereo *wav* format.

File 1: One second of the 94 dB *SPL* calibration signal was selected and saved. Channel 2 contained the calibration signal, whereas channel 1 (the head-mounted microphone channel) contained only noise and was not used for calibration (Fig. 4).

File 2: The most steady part of about 1 s duration was selected from the sustained */a/* phonation and saved. Channel 1 contained the head-mounted microphone signal, channel 2 the sound level meter signal (Fig. 4).

File 3: The speech signal intended for the dose measurement was selected and saved. Channel 1 contained again the head-mounted microphone signal, channel 2 the EGG signal.

The EGG signal from file 3 was analysed to determine the individual cycles of vocal fold vibration. Once the cycles were found, the fundamental frequency of voice was determined for each period. Then, the time indices of the beginning and end of each cycle were applied to the microphone signal of the file 3 and used to calculate the *SPL* for each cycle. In order to obtain the absolute *SPL@50 cm* (in dB re 20 μPa), an automatic calibration was done using the audio files 1 and 2. The details of the processing are now described.

Peak-picking algorithm for the EGG signal of speech

Period extraction from a speech signal is generally problematic, partly because there is not always a uniquely definable period, and partly because there does not seem to exist a universal algorithm which would produce satisfactory results for every subject, every purpose and every recording condition (20). Readers are referred to, for example, Hess (21) for a

detailed overview of frequency extraction algorithms of speech. Since the EGG waveform shape is considerably simpler than the pressure waveform of a microphone signal, the EGG signal was chosen here as the basis for the detection of the voiced/unvoiced segments of speech and for extraction of the speaking fundamental frequency. An original peak-picking algorithm was developed for processing the EGG signal and written as a Matlab script (19). The script was optimized for an EGG signal oriented in accordance with the recommendations of the Voice Committee of the IALP (22), i.e., positive peaks of the EGG curve correspond to the events of maximal vocal fold contact area. The processing phases were as follows:

1) The EGG signal was smoothed by taking a running average of 8 samples (0.17 ms).

2) The value of the baseline noise level was empirically determined from the plot of an unvoiced part of the EGG signal (Fig. 5a and b). The noise threshold was automatically set at 1.2 times the empirically determined noise value. (In cases of signals with EGG waveform artefacts the noise threshold was increased to a larger value, see phase 9.)

3) The maximum extent of the signal envelope of the voiced segments was determined empirically from a plot of the EGG waveform (Fig. 5a and c). The baseline shifts were disregarded in the determination of the value.

4) All the local maxima and minima of the EGG signal were found (Fig. 6a).

5) Global maxima and global minima that were separated by at least 1 ms were selected from the maxima and minima (elimination of frequencies higher than 1000 Hz).

6) A reduced set of maxima and minima was obtained by applying a condition that there is only one maximum between two adjacent minima and only one minimum between two adjacent maxima. In the case of more than one maximum or minimum, the extreme peaks were selected and the others discarded (Fig. 6a). All the following procedures kept this condition of alternating maxima and minima.

7) If a minimum between two maxima was larger than one of the maxima, that minimum was discarded together with the smaller maximum (Fig. 6b).

8) A noise-peaks eliminating procedure was performed: if the value difference between the adjacent maxima and minima was smaller than the measured noise level, the maxima and minima were eliminated by comparing their values to previous valid maxima and minima. The smaller of the two maxima and the larger of the two minima were discarded (Fig. 6c).

9) A threshold criterion for subharmonicity/false harmonic peaks was applied to eliminate any maxima

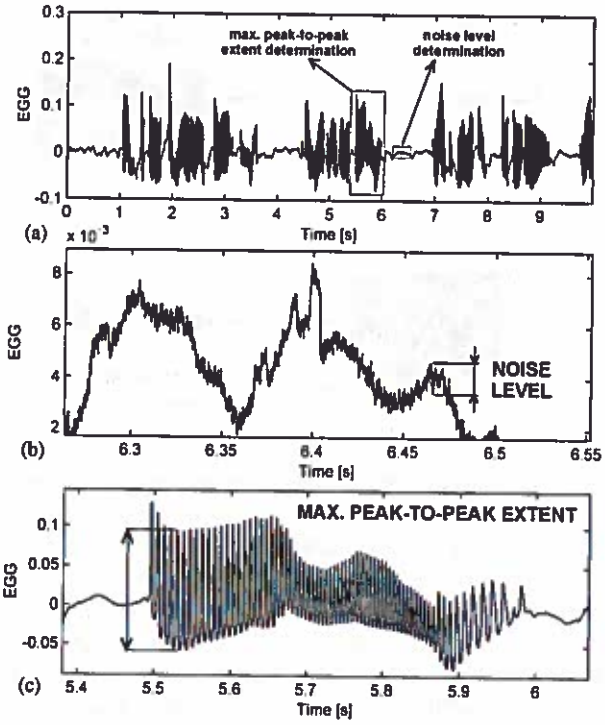

Fig. 5. Empirical determination of the parameters, which were used for the peak-picking algorithm of the EGG signal of speech. (a) The selection of the parts of the signal used for determination of the noise level (at no voice activity) and for the maximal peak-to-peak extent (the portion with greatest EGG amplitude). (b) Detail of the part of the signal with no voice activity (as marked in (a)). The EGG noise level is determined as the peak-to-peak extent of the EGG waveform, disregarding baseline shifts. In this example the noise level is $(4.6 - 3.4) \times 10^{-3} = 0.0012$ (relative units). (c) Detail of the part of the signal with greatest EGG amplitude (as marked in (a)). The maximal peak-to-peak extent is determined by subtracting the minimum value from the maximal value, disregarding baseline shifts. In this example the peak-to-peak extent is $0.1 - (-0.06) = 0.16$ (relative units).

with peak-to-peak values 5.5 times smaller than the previous and following peak-to-peak values (Fig. 6d). The 5.5 value was found by trial and error to work best, but could be modified if needed. In cases with EGG waveform artefacts (for instance those with multiple peaks within a cycle), this condition was bypassed (by changing the value from 5.5 to 100) and the level of noise threshold was increased (to a value which was larger than the peak-to-peak value of the EGG artefacts), which resulted in a smoother F_0 contour.

10) The difference between time indices of the finally selected maxima determined the vibratory

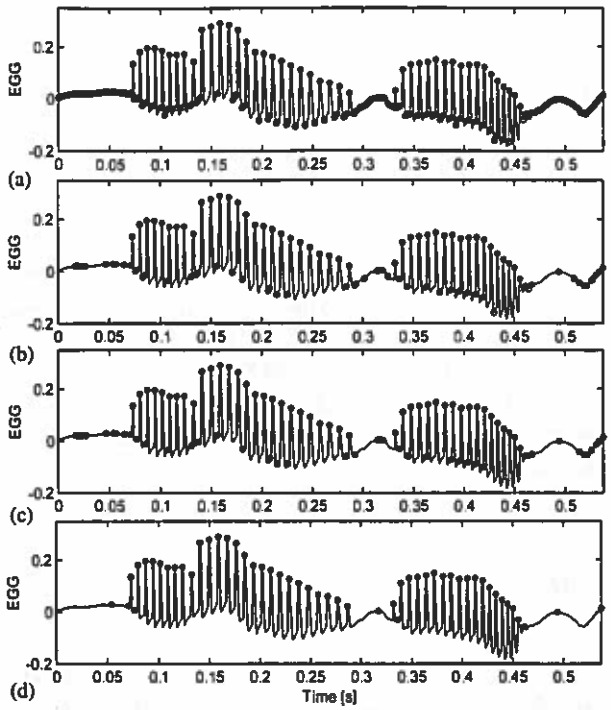

Fig. 6. Determination of the vibratory periods of the vocal folds—the process of consecutive elimination of noise and false peaks in the EGG signal. The empty circles represent the maxima which were eliminated, the black dots represent the remaining maxima. (a) Maxima distant by more 1 ms, situation after phase 6. The empty circles represent the eliminated peaks from all the local maxima found in phase 4. (b) The maxima larger than both the neighbouring minima, situation after phase 7. (c) Maxima with amplitude larger than the noise level, situation after phase 8. (d) Final set of maxima, after phase 9.

periods T, whose values were inverted to obtain cyclic frequency values ($F_0 = 1/T$) in Hz.

11) A high-pass criterion was applied on the frequency values: if the value was below 50 Hz, it was discarded (set equal to zero).

12) A cycle-to-cycle variation check was applied to eliminate irregular peaks not related to vocalization. For that purpose, it appeared useful to distinguish between 'strong' (voicing) and 'weak' (soft phonation and noise) signal. A 100% period-to-period variation was allowed for a 'strong' and 50% for a 'weak' signal. The larger value for the strong signal allowed processing with cycle-to-cycle variations of up to 100% (which might be expected in highly irregular creaky or vocal fry vocalizations with larger vocal fold contact areas). The condition was expressed in equation form as follows:

SET $F_O(j+1) = 0$ IF

$$|F_O(j+1) - F_O(j)| \geq C \cdot MIN[F_O(j), F_O(j+1)]$$

AND

$$F_O(j+1) \neq 0$$

AND

$$|F_O(j+2) - F_O(j+1)| \geq C \cdot MIN[F_O(j+1), F_O(j+2)]$$

where j is the period number, *MIN* is the minimum function, *AND* is the 'logical and' operation and C is the value 1 or 0.5 (corresponding to 100% or 50%) for the strong or weak signals respectively. The signal was automatically recognized as strong/weak if the peak-to-peak amplitude difference in either the cycle of interest or in the preceding cycle was greater/smaller than 30% of the maximal value measured in step 3). This test for three consecutive cycles allowed us not to discard valid pitch jumps exceeding an octave that can occur in certain voices (23).

13) A condition requiring that a valid voicing segment shall contain at least three vibratory cycles of the vocal folds was applied. If there were less than

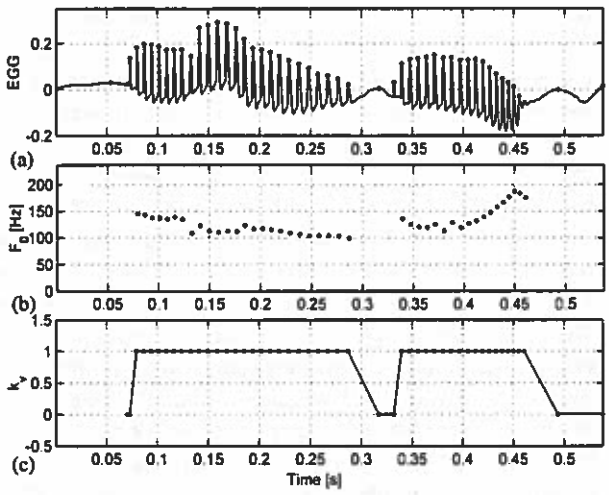

Fig. 7. Results of the F_0 extraction algorithm after phase 15. (a) The EGG segment with the finally selected maxima, defining the period intervals used for F_0 and k_v determination. (b) The final F_0 values for each cycle after the F_0 elimination phases 11–15. (c) The voicing detection values k_v (1 defines voiced, 0 unvoiced period). The F_0 and k_v values are plotted at the end of each cycle. The measured k_v values are marked as dots and for display purposes are interconnected by line, but no actual interpolation was done between the 0 and 1 values. The same holds also for k_v displays in Figs. 8 and 9.

three consecutive non-zero frequency values, they were set equal to zero.

14) The non-zero frequency values were used to make up the final F_0 contour (Fig. 7b). 'Zero' frequencies were not kept as F_0 contour values because an oscillator that is not oscillating does not necessarily have zero frequency.

15) The zero values of F_0 were used to recognize the 'unvoiced' segment of speech, which is the time during which there was no vocal fold vibration. The mathematical coding of the voiced/unvoiced segment was done using the *voicing detection function* k_v (equation 2). The k_v was set to zero when the F_0 was zero; for all other instances, the k_v was set to the value of one. In this way, a k_v value was obtained for each period of vibration (Fig. 7c).

16) The extracted time instances of the beginning and end of each EGG cycle were used as the frame start and stop times in the head-mounted microphone signal to calculate the *SPL* values for every cycle of vocal fold vibration (see further).

Notes: While the EGG signal is easier to use for extraction of the F_0 values than the microphone signal, there are also drawbacks related to using the EGG signal. When processing the speech data from reading the Goldilocks passage (19), large EGG baseline shifts were observed, especially in the female subjects. The large baseline shifts are most likely related to vertical movements of the larynx. As females generally have thinner and shorter vocal folds and thus smaller vocal fold contact area than males, the EGG voicing signal could be slightly weaker so that the baseline shifts become more prominent in the female EGG signal. The greater baseline shifts in females may also be caused by more 'animated' speech than that used by males. The magnitude of the baseline shifts was, in some cases, so great that the EGG signal was 'clipped' as it was being recorded on DAT tape, which made it impossible to perform the F_0 extraction from that particular segment. In other cases the shape of the EGG waveform in each cycle was so distorted (due to the steep rise or decline of the baseline) that there was some ambiguity as to the location and number of peaks in a cycle. These problems, however, affected only small portion of the data (less than about 5%, in all the materials recorded for the experimental dose measures reported in the previous study (1)) and were not considered significant for the overall results.

One consequence of extracting the fundamental frequency and voicing time from the EGG signal is that the voicing detector function can be positive (i.e., $k_v = 1$) even during a time segment where there is little or no energy in the microphone signal; for instance, when the vocal folds are vibrating with the lips completely closed. Thus there are differences between the EGG and microphone waveforms that lead to slightly different results in voicing detection as well as F_0 extraction, which need to be taken into account when calculating vocal doses from different signals (see also, e.g., Howard and Lindsey (24)).

The determination of absolute sound pressure levels

The absolute *SPL* values for the head-mounted microphone signal were obtained in three steps.

1) The root mean square (RMS) amplitude of the 94 dB calibration tone signal p_{cal} from the sound level meter (SLM) microphone (file 1, channel 2) was calculated as

$$RMS(p_{cal}) = \sqrt{\frac{\sum_{k=1}^{K} p_{cal}^2(k)}{K}} \qquad \text{(eq. 12)}$$

where k is the index of the successive sampled values and K is the total number of samples of the calibration signal. (In this case, K equalled about 48000 samples for the signal of 1 s duration). Using the equation

$$SPL_{cal} = 20\log_{10}[g_{SLM} RMS(p_{cal})] \qquad \text{(eq. 13)}$$

the gain factor g_{SIM} was computed to relate the digitized voltage values of the SLM microphone to the SPL_{cal} level actually measured by the SLM. Since SPL_{cal} was 94 dB, the $g_{SLM}$was obtained as:

$$g_{SLM} = \frac{10^{SPL_{cal}/20}}{RMS(p_{cal})} = \frac{10^{94/20}}{RMS(p_{cal})} \qquad \text{(eq. 14)}$$

2) The audio file 2 (i.e., the most steady 1 s segment of simultaneous SLM and head-mounted microphone recording during the production of sustained /a/) was used to obtain the relation between the levels of the head-mounted microphone signal and the SLM signal. The RMS amplitude was calculated from both the signals, and the gain of the head-mounted microphone signal g_{HMM} was obtained as:

$$g_{HMM} = g_{SLM} RMS(p_{SLM})/RMS(p_{HMM}) \qquad \text{(eq. 15)}$$

where p_{HMM} represent all the sample values of the head-mounted signal (file 2, channel 1) and p_{SLM} are all the sample values of the sound level meter signal (file 2, channel 2). The gain factor g_{HMM} adjusts the digitized voltage values p_{HMM} of the head-mounted microphone to the true sound pressure values, so that the absolute *SPL* value (in dB at the 50 cm mouth-to-microphone distance) is obtained from the head-mounted microphone using the equation:

$$SPL = 20\log_{10}[g_{HMM}RMS(p_{HMM})] \qquad \text{(eq. 16)}$$

3) $RMS(p_{HMM})$ values were calculated separately for each period of the vocal fold vibration from the head-mounted microphone signal of the readings of the 'Goldilocks' passage (file 3, channel 1). The periods were defined by the indices obtained from the EGG signal by the peak-picking algorithm. The *SPL* values were calculated for each period of the vocal fold vibration using equation 16. The final results of the analysis are presented in Fig. 8, which shows the F_0 and k_v values derived from the EGG signal and the *SPL@50 cm* derived from the head-mounted microphone signal.

Calculation of the doses

The *SPL* values, F_0 values and k_v values of each cycle were used for calculation of the vocal doses. The *time dose* was obtained by applying equation 1 to the discrete oscillatory cycles:

$$D_t = \int_0^{t_m} k_v\, dt = \sum_{n=1}^{N}[k_v(n)\Delta t(n)] \quad seconds \qquad \text{(eq. 17)}$$

where N is the total number of periods, the symbol $k_v(n)$ denotes the k_v values for each period n (extracted by the peak-picking algorithm) and $\Delta t(n)$ are the durations of the successive periods n in seconds (also extracted by the peak-picking algorithm).

The *cycle dose* D_c was calculated simply by summing all the periods in which the k_v factor was 1, in accordance with the definition (equation 3)

$$D_c = \int_0^{t_m} k_v F_O\, dt = \sum_{n=1}^{N}[k_v(n)F_0(n)\Delta t(n)]$$

$$= \sum_{n=1}^{N} k_v(n) \quad cycles \qquad \text{(eq. 18)}$$

where the inverse relationship $F_O(n) = 1/\Delta t(n)$ was used.

The *distance dose* D_d was calculated as

$$D_d = 4\int_0^{t_m} k_v A F_O dt = 4\sum_{n=1}^{N}[k_v(n)A(n)F_O(n)\Delta t(n)]$$

$$= 4\sum_{n=1}^{N}[k_v(n)A(n)] \quad meters$$

(eq. 19)

where $A(n)$ is amplitude of the vocal folds for each period of the vocal fold vibration, which is obtained from the $SPL(n)$ and $F_0(n)$ by using equations 5–7.

The *energy dissipation dose* D_e was calculated as:

$$D_e = \frac{1}{2}\int_0^{t_m} k_v \eta (A/T)^2 \omega^2 dt$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{k_v(n)\eta(n)[A(n)/T(n)]^2$$

$$[2\pi F_O(n)]^2 \Delta t(n)\} \quad joules/m^3 \qquad \text{(eq. 20)}$$

where the shear viscosity $\eta(n)$ and the vertical thickness $T(n)$ were obtained from the $SPL(n)$ and $F_0(n)$ of voice for each period n using the empirical rules given in equations 9 and 10; and the *radiated energy dose* D_r was calculated as

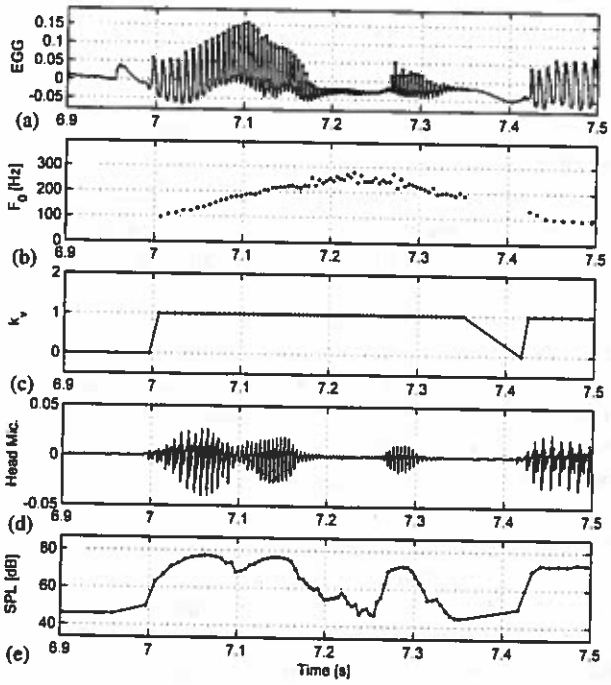

Fig. 8. The results of the F_0 analysis from EGG combined with the absolute *SPL@50 cm* values derived from the head-mounted microphone signal. The F_0 and *SPL* values are obtained for each single extracted period of vocal fold vibration (for unvoiced segments with $k_v = 0$, the F_0 values are not plotted). In (a), notice the segment with alternating stronger and weaker maxima (at *c.* 7.27 s), which are, for the purpose of the dose measurement, identified as separate oscillatory cycles, rather than subharmonic cycles.

$$D_r = 4\pi R^2 \int_0^{t_m} k_v 10^{(SPL-120)/10} dt$$

$$= 4\pi R^2 \sum_{n=1}^{N} [k_v(n) 10^{(SPL(n)-120)/10} \Delta t(n)] \ joules \qquad \text{(eq. 21)}$$

where R is the mouth-to-microphone distance (0.5 m) used for *SPL* measurement.The doses are calculated automatically by the scripts, which are available on the web (19).

RESULTS

The accumulation of the different doses is demonstrated in Fig. 9 on an example of 10 s of speech of a male subject. The extracted F_0, *SPL* and k_v values are shown in Fig. 9a–c. The doses, accumulating over time, are in Fig. 9d–h. Note that the dose values increase during the voiced passages, whereas they stay constant during unvoiced segments as expected.

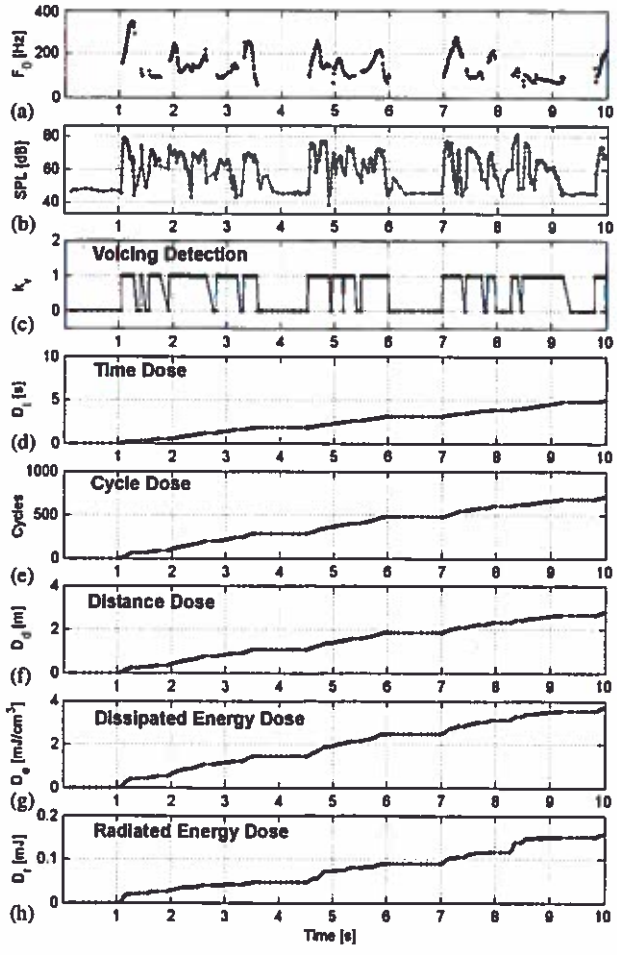

Fig. 9. Accumulation of the five vocal doses over a 10-s segment of speech. The five vocal doses (d–h) are derived from the F_0, *SPL@50 cm* and k_v values shown in (a–c).

The time dose (Fig. 9d) reaches the value of about 5 s after 10 s of speech, which corresponds to the voicing percentage of about 50%. That agrees with the voicing percentages reported by other researchers as being typical for speech (7, 25). The cycle dose (Fig. 9e) reveals the vocal folds accomplished about 700 cycles during the 10 s of speech. Close to 3 meters of accumulated distance is measured by the distance dose (Fig. 9f). The vocal folds dissipated *c.* 3.5 mJ of energy per cubic centimetre of the vocal folds (Fig. 9g) while the energy radiated out of the mouth was about 0.15 mJ (Fig. 9h).

DISCUSSION

The presented method allows quantifying the amount of vocalization over time. The different dose measures allow different vocalizations to be compared and can be used in studying different factors, which can be potentially harmful for the vocal folds if exposed to too much vocalization.

The time dose, although sensitive to neither frequency nor loudness, can be used for quantifying the duration of voicing and the voicing percentages among various vocal activities or occupations (2–7, 26). Furthermore, it can also be used as a normalization factor to obtain doses per second of vocalization (1).

The cycle dose, or the vocal loading index, has been found relevant by Rantala and Vilkman (8) who reported that the teachers with more vocal complaints showed higher values of vocal loading index than the teachers with less vocal complaints. A factor that can contribute to the correlation of the larger number of cycles with the larger number of vocal complaints is the potentially harmful effect of the collisions of the vocal folds. The number of collisions is most likely to be larger with a larger number of oscillations. A more proper quantification of the collisions of the vocal folds would require taking into account the forces resulting from the collision as well as the collision surface of the vocal folds. A dose that would account for all these factors has, however, not been defined yet due to limited knowledge on these factors.

Another interesting dose is the distance dose, because it can be linked to safety limits for hand-transmitted vibration used in industry. Based on exposure criteria for continuous acceleration of tissue, it has been derived that the tissue of the hands shall not accumulate a distance larger than *c.* 520 m per working day; otherwise the dose of vibration is

considered hazardous (1). As shown in Fig. 9f, the vocal fold can travel the distance of 3 meters during only 10 s of speech, which means the distance of 520 m could be achieved by this particular male subject in less than 30 min of speech. This suggests that the limit of 520 m may be too small for the vocal folds. A larger limit than 520 m can be expected also on the basis of the structure of the vocal folds, which is better adapted for vibration than the tissue of hands. Furthermore, the silence periods included in speech could allow for recovery of the tissue traumatized by vibration, and this also could shift the limit of 520 m towards larger values. Nevertheless, similar mechanisms can be potentially involved in the process of damaging the tissue of hands and vocal folds through vibration, which make the distance dose an interesting subject for future studies.

The energy dissipation dose calculates the total amount of heat generated in a unit volume of vocal fold tissues. The increase of the temperature of the vocal folds has been listed as one of the possible factors involved in tissue damage by Titze (27). The present results suggest, however, that the amount of heat generated in the vocal folds is smaller than originally expected. The specific heat capacity of human tissue is 3470 J kg^{-1} $°C^{-1}$ (www.yesican.yorku.ca/home/sh_table.html) and the density is 1020 kg m^{-3}, which means that one would need the energy of *c.* 3.5 J to warm up a cubic centimetre of the vocal folds by 1°C (see, e.g., the calorimetric equation in (1)). As shown in Fig. 9g, the heat generated in the vocal folds over 10 s of speech was approximated to be, in this case, about 3.5 mJ/cm^3, thus only about one thousandth of the above mentioned value of 3.5 J/cm^3. Furthermore, the heat is conducted away from the vocal folds by the phonatory air stream and the circulating blood, making large local increases of tissue temperature unlikely. This underscores the conclusion of the previous study that the heat generated in the vocal folds is of potentially lesser importance for vocal fold damage than factors such as the repetitive acceleration and deceleration of the vocal fold tissue through excessive vibration (1).

The radiated energy dose can potentially be useful for studying the efficiency of the voice production, for instance by relating the energy radiated out the mouth to the total energy dissipated in the vocal folds. Relating the results from Fig. 9g and h and assuming the total volume of the vibrating vocal folds is close to a cubic centimetre, the amount of radiated energy is about 20 times smaller than the amount of dissipated energy in this case. This estimate is, however, only a crude one since a proper quantification of the total dissipated energy would require the knowledge of the length, thickness as well as the depth of the vocal fold vibration. While empirical rules were derived for thickness and length (1), the rule for vibrating depth of the vocal folds has not been available due to lack of experimental data. The importance of the radiated energy dose has been, also due to these factors, slightly limited. The three potentially most useful doses are considered to be, at the present time, the time dose, the cycle dose and the distance dose. But since the internal mechanism of vibrational damage to vocal fold tissue is not sufficiently understood, the other doses should not be rejected.

Overall, proper measurement of the doses depends on the correct extraction of the fundamental frequency, voicing detection and determination of the absolute *SPL* levels at the distance of 50 cm (different distances could also be used but the *SPL* levels then would need to be recalculated for 50 cm, or the empirical rules would need to be modified accordingly). The present paper offers an F_0 extraction and voicing detection algorithm for an EGG signal of speech, which, if the noise value and maximal peak-to-peak extent value are set properly, provides reasonable results for dose calculations even with more complex vibratory patterns, such as subharmonics or pitch jumps. The paper also offers a practical method for deriving absolute *SPL* values from a head-mounted microphone signal.

Since at present time there are no direct methods available for measurement of the vocal fold vibratory amplitude, vocal fold thickness and vocal fold viscosity, which are needed for the determination of the distance and energy dissipation doses (Fig. 1), empirical rules are employed that allow approximating these parameters from F_0 and *SPL* values of speech (Fig. 2). The dispersion of true vocal fold amplitudes around the best fit ((1), Fig. 3) suggests that the difference between the distance dose for an average person and the real person can be up to a factor of two. The accuracy of the energy dissipation dose is only an order of magnitude, especially due to the estimate of the vocal fold viscosity ((1), Fig. 5). More data are needed at this point, but the derived empirical rules allow initial approximation of these two doses, which are otherwise too difficult to be measured directly.

The procedure for vocal dose measurement described here is intended to be ready for use; it does not, however, exclude the possibility of further improvement by, for instance, perfection of the empirical rules when new data become available or by using alternative frequency extraction and voicing detection algorithms.

ACKNOWLEDGEMENTS

The work was supported by the National Institutes of Health, grant number DC RO1 04224-01.

REFERENCES

1. Titze IR, Švec JG, Popolo PS. Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues. J Speech Lang Hear Res 2003; 46: 922–35.
2. Ohlsson A-C, Brink O, Lõfqvist A. A voice accumulation–validation and application. J Speech Hear Res 1989; 32: 451–7.
3. Masuda T, Ikeda Y, Manako H, Komiyama S. Analysis of vocal abuse: Fluctuations in phonation time and intensity in 4 groups of speakers. Acta Otolaryngol (Stockh) 1993; 113: 547–52.
4. Jónsdottir V, Rantala L, Laukkanen A-M, Vilkman E. Effects of sound amplification on teachers' speech while teaching. Logoped Phoniatr Vocol 2001; 26: 118–23.
5. Szabo A, Hammarberg B, Håkansson A, Södersten M. A voice accumulator device: Evaluation based on studio and fields recordings. Logoped Phoniatr Vocol 2001; 26: 102–17.
6. Rantala L, Vilkman E, Bloigu R. Voice changes during work: Subjective complaints and objective measurements for female primary and secondary schoolteachers. J Voice 2002; 16: 344–55.
7. Södersten M, Granqvist S, Hammarberg B, Szabo A. Vocal behavior and vocal loading factors for preschool teachers at work studied with binaural DAT recordings. J Voice 2002; 16: 356–71.
8. Rantala L, Vilkman E. Relationship between subjective voice complaints and acoustic parameters in female teachers' voices. J Voice 1999; 13: 484–95.
9. Titze IR. Phonation threshold pressure: A missing link in glottal aerodynamics. J Acoust Soc Am 1992; 91: 2926–35.
10. Titze IR, Sundberg J. Vocal intensity in speakers and singers. J Acoust Soc Am 1992; 91: 2936–46.
11. Chan RW, Titze IR. Viscosities of implantable biomaterials in vocal fold augmentation surgery. Laryngoscope 1998; 108: 725–31.
12. Chan RW, Titze IR. Viscoelastic shear properties of human vocal fold mucosa: Measurement methology and empirical results. J Acoust Soc Am 1999; 106: 2008–21.
13. Howard DM, Angus JAS. Acoustics and psychoacoustics. 2nd ed. Oxford: Focal Press; 2001.
14. ANSI S1.4-1983, American national standard specification for sound level meters. Acoustical Society of America; 1985.
15. Winholtz WS, Titze IR. Conversion of a head-mounted microphone signal into calibrated SPL units. J Voice 1997; 11: 417–21.
16. Ciletti E. Cardioid-carrying member. Mix [serial online] 2003 March [cited 2003 August 12]. Available from URL: http://www.tangible-technology.com/microphones/proximity/proximity.htm or from URL: http://www.mixmag.com
17. Eargle J. The microphone book. Boston: Focal Press; 2001.
18. Merhaut J. Theory of electroacoustics. New York/London: McGraw-Hill International Book; 1980. Translated from Czech by R. Gerber. (Czech version: Merhaut J. Teoretické základy elektroakustiky. 4 ed. Praha: Academia; 1985).
19. Švec JG, Popolo PS, Titze IR. The Goldilocks passage and scripts for frequency extraction, voicing detection, SPL calculation and vocal dose determination in speech. NCVS Online Technical Memo, April 2003 [cited 2003 August 12; 17 pages]. Available from URL: http://www.ncvs.org/ncvs/library/tech
20. Howard DM. Instrumental voice measurement: Uses and limitations. In: Harris T, Harris S, Rubin JS, Howard DM (eds). The voice clinic handbook. London: Whurr Publishers; 1998. pp. 323–82.
21. Hess W. Pitch determination of speech signals: Algorithms and devices. Berlin-Heidelberg-New York-Tokyo: Springer-Verlag; 1983.
22. Baken RJ. Electroglottography. J Voice 1992; 6: 98–110.
23. Švec JG, Schutte HK, Miller DG. On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges. J Acoust Soc Am 1999; 106: 1523–31.
24. Howard DM, Lindsey GA. Conditioned variation in voicing offsets. IEEE Transactions on Acoustics. Speech and Signal Processing 1988; 36: 406–7.
25. Löfqvist A, Mandersson B. Long-time average spectrum of speech and voice analysis. Folia Phoniatr (Basel) 1987; 39: 221–9.
26. Sala E, Airo E, Olkinuora P, Simberg S, Ström U, Laine A, et al. Vocal loading among day care center teachers. Logoped Phoniatr Vocol 2002; 27: 21–8.
27. Titze IR. Principles of voice production. Englewood Cliffs (NJ): Prentice-Hall: 1994.