IMAGE SEQUENCES AS NECESSARY SUPPLEMENT TO A PATHOLOGICAL VOICE DATA BASE

Matthias Fröhlich, Dirk Michaelis, † Eberhard Kruse
matth@physik3.gwdg.de
Drittes Physikalisches Institut, Georg-August-Universität Göttingen, Bürgerstr. 42-44, D-37073 Göttingen, Germany
† Dept. Phoniatrics and Pedaudiology, Georg-August-Universität Göttingen, Robert-Koch Str. 40, D-37075 Göttingen, Germany

ABSTRACT

Voice corpora of pathological voices are generally based on high-quality audio recordings. These are supplemented with additional information describing the general recording conditions and the individual subjects. While an acoustic analysis cannot be expected to specify certain medical diagnoses it may reflect the glottal phonation conditions. Two examples shall illustrate this point. First, voice quality generally improves for patients undergoing logopedic therapy. These changes are reflected in the acoustic analysis by means of the hoarseness diagram while the diagnosis remains unchanged. Second, a statistically significant differentiation of the acoustic measures is achieved for subjects with laryngeal paralysis only if the phonation mechanism are taken into account. These examples demonstrate the necessity to add visual recordings to an audio data base of pathological voices in order to allow a flexible and comprehensive use of the voice corpus.

1 INTRODUCTION

In general, it may seem sufficient for the potential usefulness of a pathological voice data base to supply high-quality audio recordings. Additionally, an appropriate documentation would contain the description of the recording equipment (e.g., sampling rate, equipment types and specifications) as well as some information on the individual subjects (including at least sex, age, and medical diagnosis). This description of the subjects is important for a later employment of the voice corpus in any context since it serves as the basis for the definition of the specific voice groups which are to be analyzed.

While classifications based on age or sex are straightforward, the ones based on the diagnosed voice pathology poses problems which the experimenter may not even be aware of. An acoustic analysis should be aimed at the description of specific functional and (patho-)physiological conditions during voice generation. However, the labeling of certain organic conditions by pathology names in the past was governed by medical, not acoustical requirements. But if a diagnosis does not correspond to a typical and specific glottal function it cannot be expected to supply an appropriate basis for the definition of a group that is homogeneous in acoustical respect as well. Given the still unresolved debate as to the interpretation of acoustic measures in physiologic or perceptual correlates, an acoustic voice analysis has to be validated using voices with well-defined phonation conditions rather than with specific medical diagnoses.

The assessment of phonation conditions is commonly performed by laryngo-stroboscopic examinations, although high-speed imaging is becoming increasingly important. This visual examination of the vocal fold vibration requires a well-trained and experienced phoniatrician. While for some specific tasks the vibration pattern might be reduced to a few numerical parameters (such as proposed in [1]), a comprehensive parameterization seems effectively impossible due to the complex nature of the vibration. On the other hand, such a parameterization would be needed to offer a flexible use of the speech data. This dilemma may be solved if representative laryngo-stroboscopic image sequences supplement the audio recordings of a voice corpus rather than some numerical, textual, or tabular parameterization of it.

To demonstrate the importance of these considerations, two different applications of a voice corpus are described where they play a crucial role in the interpretation of the results. The first one concerns the monitoring of changes in voice quality for patients undergoing postsurgical logopedic voice rehabilitation. The second one describes the differences in the acoustic analysis results for voice groups showing different kinds of laryngeal paralyses.

2 MATERIAL

2.1 Voice corpus

The pathological voice data base of the Göttingen University hospital from which the analyzed voices are taken comprises 550 recordings (December 1997). Each recording is performed according to the following protocol. First, a phoniatrician examines the patient using laryngo-stroboscopy and diagnoses the current laryngeal health conditions. The laryngo-stroboscopic image sequences are stored on video tape (JVC model HR-S6900EG, Super VHS). After the examination the audio data are recorded in a sound-treated room using a head-mounted microphone (beyer dynamics HEM 191), pre-amplifier (AXR Mic/Dat 2), and DAT recorder (Pioneer D-07, sampling frequency 48kHz). The speech tasks consist of the vowel sequences...
Proceedings of VOICEDATA\textsuperscript{98} and intensity, then at a lower pitch, then at a higher pitch. Next, a standard text (“Nordwind und Sonne”, not used for this study) is read after which the vowel sequence is repeated at comfortable pitch. Before the recording starts the subject is instructed to sustain each vowel for 3 to 5 seconds if possible, without connecting consecutive vowels. The pitches of the different modes are not further specified since for highly disturbed voices a realization of different pitches is often impossible. The recording engineer checks that in the end all isolated vowel tokens have been recorded (i.e., four versions of the seven vowels). If any are missing (e.g., were forgotten, too short in duration, interrupted for some reason) they are repeated until all vowel sets are complete.

2.2 Acoustic analysis

The acoustic analysis is performed by calculating the hoarseness diagram \cite{2-7} for each recording session from the audio data (for a detailed description see \cite{7}). In short, the hoarseness diagram allows a quantitative description of voice quality in two dimensions on the basis of four acoustic features. Jitter, shimmer, and the mean correlation coefficient between successive periods in the time domain contribute in equal parts to the abscissa named the \textit{irregularity component}. The GNE (glottal to noise excitation ratio \cite{8}) is defined as the maximum of the cross-correlation between Hilbert envelopes that are calculated for different frequency channels. A linear function of the GNE defines the ordinate named the \textit{noise component}.

Data pre-processing is performed interactively by marking voice onset and offset using self-developed audio-visual software. The onset and offset segments are excluded from the subsequent automatic and unsupervised analysis. In this way only the “stationary” part of each vowel is analyzed by calculating the hoarseness diagram coordinates on 500ms rectangular windows applying a shift of 250ms. The ensemble of analysis frames is characterized by the mean and standard deviation of each coordinate calculated from all complete frames (i.e., generally excluding the last – usually incomplete – frame of each vowel token). For visualization, an ellipse is displayed in the hoarseness diagram with the center defined by the means and the half-axes defined by the standard deviations of each coordinate. This ellipse characterizes the acoustic voice quality of the subject at the given recording date. If ellipses are averaged for several patients, a “group ellipse” results.

The significance of the differences between voice groups is determined by the two-dimensional Kolmogorov-Smirnov test and by the Wilcoxon two-sample test \cite{9}. Former test is applied to analyze the two-dimensional plane defined by the hoarseness diagram while the latter allows to test one-dimensional differences for the individual coordinate axes. All results are interpreted at a significance level of \( p < 0.001 \).

2.3 Monitoring of voice quality changes

Usually, after a laryngeal surgery a person’s voice is severely disturbed. Its quality normally changes substantially during wound healing. Additionally, a patient undergoing logopedic therapy will often show considerable improvements in voice quality. However, the undifferentiated medical diagnosis does not reflect the current voice condition so that very different acoustic analysis results may be obtained for one given diagnosis. Conversely, a classification based on this diagnosis will not necessarily indicate a typical or specific voice quality. The grouping of several voices with the particular diagnosis may therefore lead to an inhomogeneous group both in patho-physiologic and acoustical respect.

The hoarseness diagram analysis has already been demonstrated to give a meaningful description of acoustic voice quality \cite{2-6}. To support the correspondence between the acoustic analysis results and the phonation conditions, sample images of the vocal fold vibration at different stages of the logopedic therapy are shown for two subjects together with the acoustic analysis results (Figs. 4-7). From our introductory remarks it is obvious that these sample images cannot be regarded as an adequate substitute of the original stroboscopic recordings. However, they may illustrate changes in the phonation conditions in printed media.

The recordings of subject \textit{bw} (male, 55 years) start at the date 3/13/96 after a resection of a second glottal tumor following an earlier partial bilateral chordectomy. During the next half year he underwent a functional voice rehabilitation to stabilize the phonation mechanism. For voice generation he uses both the healthy (left) and the operated (right) vocal fold (“glottic phonation”). Subject \textit{rw} (male, 63 years) underwent logopedic therapy after a complete chordectomy. During phonation only his healthy right vocal fold vibrates while the operated left fold has become stiff. This phonation condition where only the healthy vocal fold vibrates is termed “pseudo-glottic phonation” \cite{5}.

2.4 Paralysis groups

The laryngeal muscles are innervated by different nerves that may be traumatized independently or in combination (see Fig. 1). The recurrent nerve controls only the intrinsic larynx muscles (except for the cricothyroid and the ventricular muscles) so that a lesion will only affect those muscles. On the other hand, a lesion of the vagus nerve will not only paralyze the intrinsic larynx muscles but also body parts below the larynx that are innervated by the vagus nerve. It seems as if in this case the respiratory innervation of the larynx is affected as well. In practice, different typical positions of the paralyzed vocal fold are observable during breathing. These positions depend in a systematic way on the location of the nerve lesion. The differentiation between the recurrent paralysis and the vagus paralysis described in the following thus is achieved primarily
by laryngoscopy, i.e. by the examination especially of the respiratory position of the vocal folds.

The paralysis of the recurrent nerve is the type of laryngeal paralysis most commonly encountered. In general, it leads to a paramedian position of the paralyzed vocal fold. In the laryngoscopic view during breathing the paralyzed fold is found in a position which is similar to the phonation position observed for healthy voices. While this respiratory position deviates clearly from normal it is nevertheless favorable for a vibration during phonation. A relatively stable glottal vibration is often possible due to the entrainment of the paralyzed fold (observable by stroboscopy).

A different paralysis type can result from a lesion of the vagus nerve between the superior laryngeal nerve and the recurrent nerve. Generally, in the early stages of this paralysis the affected vocal fold is found in a more lateral respiratory position compared to the paramedian position commonly observed for a recurrent paresis (as already described by Kirchner [10]). A grouping of laryngeal pareses into a vagus paresis group (vp) and a recurrent paresis group (rp) is therefore possible primarily on the grounds of laryngoscopic examinations. Both groups are analyzed separately for their acoustic voice qualities.

3 RESULTS & DISCUSSION

3.1 Monitoring of voice quality changes

The course of subject bw (Fig. 4) shows a continuous improvement of voice quality during the voice rehabilitation. On the other hand, from a medical point of view the diagnosis “glottic phonation with vibration of the operated vocal fold after tumor resection” remains unchanged. The stroboscopic recordings that are exemplified by the sample images in Fig. 5 show improvements in regularity of the vibration, vibration amplitude, and degree of glottal closure. Thus the acoustic analysis results correspond to the pathophysiologic phonation conditions and reflect the success of the voice rehabilitation.

The same match of the acoustic analysis and the phonation conditions is found for subject rw (Figs. 6, 7). The vibration of the healthy fold in the first recording (1/22/97) is relatively small in amplitude and not very regular. In the stroboscopic recordings of 11/5/97 the vibration has improved considerably. This change in glottal function is reflected by the acoustic analysis. Thus, both case studies support the interpretation that the hoarseness diagram reflects the actual phonation condition whereas an acoustic characterization of the particular diagnosis is not achieved.

3.2 Paralysis groups

The results of the acoustic analyses for the different paralysis groups are shown in Fig. 2 and Table 1. A difference between the recurrent paralysis group rp and the vagus paralysis group vp was expected on the grounds of the different positions of the paralyzed fold and seems to be indicated by the location of the ellipses in Fig. 2. However,
the statistical analysis shows this difference to be insignificant (Table 1). Furthermore, the ellipse of the vagus paresis group is relatively big. A closer look at the vagus paresis group reveals that, unexpectedly, the voices cover a relatively large area in the hoarseness diagram rather than mapping to one well-localized cluster (Fig. 3). Therefore a re-examination of the video-stroboscopic recordings was performed by an experienced phoniatrician (author E.K.) who was unaware of the acoustic analysis results at the time. The stroboscopic video sequences revealed that two subjects (hl, na) were able to achieve a glottal closure during phonation. In spite of the paralysis this closure was almost complete though not very tight (i.e., short in duration). Therefore the statistical analysis was repeated after excluding the two subjects from the vagus paresis group (defining the group vagus paralysis without closure, abbreviated as vp−, see Fig. 2 and Table 1).

While the recurrent paresis group rp shows no significant difference to the complete vagus paresis group vp, the difference between rp and the reduced vagus paresis group vp− is significant. The location of the ellipses in Fig. 2 suggests the irregularity component as main contributor to this difference. This is confirmed by the one-dimensional test of significance where only the difference in the irregularity component is found to be statistically significant (Table 1).

4 CONCLUSION

The interpretability of the acoustic analysis results obtained by means of the hoarseness diagram may be lost if the analysis results are expected to reflect specific medical diagnoses. This is exemplified for two patients undergoing logopedic therapy. In both cases the ellipses in the hoarseness diagram cover a considerable area while the medical diagnosis remains unchanged. The acoustic analysis results correspond to the particular glottal conditions assessed by stroboscopy.

The group analysis of different laryngeal paralysis types (recurrent nerve paralysis vs. vagus nerve paralysis) is found to be insignificant if the grouping is performed on the basis of the diagnosis alone. Once the phonation conditions are taken into account the difference between the group distributions becomes significant.

These examples demonstrate the importance of laryngo-stroboscopic image sequences for the interpretation of acoustic analysis results. Recordings both of the vocal fold vibration and the respiration position are necessary. While the optimum situation would be the simultaneous recording of both vocal fold oscillation by high-speed cameras and the audio signal this seems to be beyond the scope of currently available technology. Therefore the pragmatic solution is to use representative laryngo-stroboscopic video sequences taken before or after the audio recordings as supplement to the audio data. These image sequences seem to be a requirement for a comprehensive applicability of a pathological voice data base.

ACKNOWLEDGMENT

This work is part of a project funded by the Deutsche Forschungsgemeinschaft under Kr 1469/2-2.

REFERENCES


Figure 4: Development of voice quality for subject bw. The ellipse for each recording date is based on 11 analysis frames. The recordings start after a resection of a glottal tumor (3/13/96).

Figure 5: Sample image sequences taken from the stroboscopic recordings of subject bw at different dates. The four images of each row cover one “glottal cycle” of the stroboscopic recordings. Each row refers to a different recording date and corresponds to one ellipse in Fig. 4. Top row: 3/13/96 (no regular stroboscopic image). Middle row: 6/26/96 (broad glottal gap but visible oscillation). Bottom row: 9/4/96 (incomplete closure, good oscillation).

Figure 6: Development of voice quality for subject rw. The ellipse for each recording date is based on 11 analysis frames. The subject was treated for laryngeal cancer and applies glottic phonation without vibration of the operated left vocal fold (“pseudo-glottic phonation”).

Figure 7: Sample image sequences taken from the stroboscopic recordings of subject rw at different dates. The four images of each row cover one “glottal cycle” of the stroboscopic recordings. Each row refers to a different recording date and corresponds to one ellipse in Fig. 6. Top row: 1/22/97 (poor closure, small amplitude of vibration). Bottom row: 1/15/97 (good vibration).