Detection of Diplophonation in Audio Recordings of German Standard Text Readings

*Philipp Aichinger, and †Jean Schoentgen, *Vienna, Austria, and †Brussels, Belgium

Summary: Objectives. Diplophonia is a common symptom of voice disorder that is in need of objectification. We investigated whether diplophonia can be detected from audio recordings of text readings by means of dedicated audio signal processing, ie, a descendant of a formerly published "Diplophonia Diagram."

Study design. Diagnostic study.

Methods. Forty subjects were included who had been clinically rated in the past as diplophonic. For each subject, the audio signal of the German standard text "Der Nordwind und die Sonne" was recorded. First, subject groups regarding the frequency of occurrence of diplophonic episodes were established via manual labeling of audio recordings. Reference boundaries of diplophonic time intervals and the boundaries of voiced time intervals were manually obtained. Each time interval was labeled as diplophonic or nondiplophonic, as well as voiced or unvoiced. The diplophonia rate was defined as the total duration of diplophonation among the total duration of voiced phonation. Based on the diplophonia rate obtained from manual annotations, subjects were distinguished who were (1) frequently diplophonic, (2) unfrequently diplophonic, and (3) nondiplophonic during the reading of the standard text.

Second, the grouping was predicted automatically via audio signal processing, and the performance of automatic prediction was evaluated. The audio recordings were analyzed with a purpose-built audio signal processor that estimated the diplophonia rate automatically. Two cut-off threshold classifiers were trained to detect automatically (1) frequently diplophonic, and (2) nondiplophonic subjects. In addition, multinomial logistic regression was performed to enable automatic 3-way classification.

Results. Among all subjects, 14 were frequently diplophonic during the reading of the text, 14 were unfrequently diplophonic, and the remaining 12 were nondiplophonic. In automated detection of frequently diplophonic subjects, a sensitivity of 71% and a specificity of 88% were obtained. The sensitivity and specificity regarding automated detection of nondiplophonic subjects were 68% and 92%. In 3-way classification, 62.5% of the subjects were classified into the correct group.

Conclusions. Only two-thirds of the subjects who had been labeled as diplophonic on the base of auditory impression during clinical anamnesis diplophonated during the reading of a standard text. This demonstrates that the ecological validity of audio recordings of standard text readings is limited. Subject groups regarding the frequency of occurrence of diplophonic episodes were established and audio signal processing enabled automated classification. The observed performance of automated classification was promising and may be relevant to future clinical and scientific work. Possible applications include objective clinical voice assessment for diagnostic purposes and feedback based training of clinical raters.

Key Words: diplophonia–detection–signal processing–audio recordings–standard text readings–running speech.

INTRODUCTION

Diplophonia is a common sign or symptom of disordered voices. It is characterized on the auditory level by the simultaneous presence of two pitches in the voice sound.[1] Diplophonia is observed in a wide variety of clinical diagnoses, including paresis, edema, dysfunctional dysphonia, cyst, polyp, laryngitis, sulcus, scar, benign tumor, and bamboo nodes.[2] Diplophonia may be caused by (i) tension imbalances of the vocal folds, typically observed in paresis and dysfunctional dysphonia, (ii) imbalances of mass, typically observed in edema, cyst, polyp, laryngitis, or benign tumor, (iii) imbalances of stiffness, typically observed in sulcus, scar, or bamboo nodes. Any of these imbalances or combinations of these may cause a desynchronization of the vocal folds that may result in the simultaneous vibration of distinct glottal oscillators at different frequencies, and in two simultaneously perceived pitches.

The detection of diplophonia is clinically relevant, because it aids the indication, selection, evaluation, and optimization of clinical treatment in the role of a functional outcome variable. Diplophonia has been assessed in past clinical studies dealing with early glottic cancer, paralyses, mutational dysphonia, and other etiologies. Krengli et al evaluated and compared the voice quality of 27 patients undergoing radiotherapy and 30 patients undergoing laser cordectomy.[3] Diplophonia occurred in 18.5% of the patients after radiotherapy, and in 23.3% of the patients after laser

Accepted for publication June 27, 2018.
Declarations of interests: none.
From the *Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Vienna, Austria; and the †Department of Bio-, Electro-and Mechanical Systems (BEAMS), Faculty of Applied Sciences, Université libre de Bruxelles, Brussels, Belgium.
Address correspondence and reprint requests to Philipp Aichinger, Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Waehringer Guertel 18-20, 1090 Vienna, Austria. E-mail: philipp.aichinger@meduniwien.ac.at
Journal of Voice, Vol. 33, No. 6, pp. 949.e1–949.e10
0892-1997

https://doi.org/10.1016/j.jvoice.2018.06.009

cordectomy. Bibby et al assessed diplophonia before and after radiotherapy on a 6-point Likert scale, in which 1 and 6 reflect normal voice and severe impairment respectively.[4] The observed diplophonia ratings before and after treatment ranged from 1 to 3 and their means were 1.2 and 1.1 respectively. Bertino et al compared cordectomy with and without subsequent reconstruction in 14 patients.[5] Three of 7 patients were diplophonic after cordectomy with reconstruction, whereas 2 of 7 were diplophonic after cordectomy without reconstruction. Nishiyama et al evaluated in 8 patients autologous transplantation of fascia graft into the vocal fold for the treatment of glottal insufficiency.[6] Diplophonia was reported in 3 patients before treatment, which in all three cases disappeared after treatment. Iwamura and Kurita reported elimination of diplophonia in 29 out of 31 patients undergoing a pull of the lateral cricoarytenoid muscle,[7] and Tsukahara et al reported another case in which a direct pull of the lateral cricoarytenoid muscle resolved high-pitched diplophonia.[8] Kimura et al evaluated effects of arytenoid adduction on diplophonia in 6 patients via analysis of high-speed videolaryngoscopy.[9] Diplophonia disappeared in all cases after treatment. Molteni et al assessed auto-crosslinked hyaluronan gel injections in a group of 40 patients.[10] Before injection, 23.7% of the patients were diplophonic. Three and twelve months after injection, 7.9% and 25.9% of the patients were diplophonic. Voice therapy for mutational dysphonia was assessed by detecting diplophonia in a group of 15 patients by Lim et al.[11] Nine patients were diplophonic before voice therapy, and none afterward. Kocak et al reported diplophonia to evaluate effects of Isshiki type II anterior commissure relaxation laryngoplasty in a group of 21 patients who believed that their high-pitched voices conflicted with their body image and/or gender identity.[12] Diplophonia was observed in 11 of these cases before treatment, and in 5 afterward.

Current approaches to the clinical detection of diplophonia are subjective or objective. In the most common subjective approach, diplophonia is detected by auditory judgement, aiming at the perception of two simultaneous pitches.[1,13] In a recent objective approach, the simultaneous existence of two fundamental frequencies is detected in sustained vowels.[2] Also, the degree of subharmonics is used as an indicator for the presence of diplophonia.[14] Additionally, high-speed videolaryngoscopy and (multiline) kymographic imaging were used in the past to investigate diplophonia.[15–31]

Problems with regard to existing approaches are the following. First, subjective detection may suffer from intercenter and intrarater variability. Intercenter variability impedes comparisons between clinical studies conducted at different centers. Intrarater variability impedes pre–post treatment comparisons. Second, a limitation of an existing objective approach, ie, a formerly published "Diplophonia Diagram," is that it was only tested on sustained vowels in the past, and that before analysis the audio signals were required to be manually segmented into intervals of homogeneous voice quality.

The so-called "Diplophonia Diagram" is a precursor of the proposed approach.[2] It was tested in the past with audio recordings of 185 phonations that were split up manually into 285 analysis intervals of homogeneous voice quality. The test signals were obtained during high-speed videolaryngoscopy from 28 diplophonic, 22 nondiplophonic dysphonic, and 30 euphonic subjects. Diplophonic subjects were found to diplophonate in 28.4% of the total phonation time. Nondiplophonic dysphonic and euphonic subjects diplophonated in 4.3% and 7.0% of the total phonation time. The diplophonia diagram only worked with sustained vowels because it relied on the manual presegmentation of audio recordings into intervals of homogeneous voice quality. An accuracy of 94.2% was achieved in a 2-way classification task (diplophonic versus nondiplophonic). In the present study we test a different procedure with audio recordings of standard text readings. The procedure uses automatic temporal segmentation, which enables analysis of standard text readings.

Issues related to subjective/perceptual analysis of disordered voices were reported in the past, and one may conclude that room for improvement of subjective/perceptual analysis exists. Investigations included experiments on roughness ratings, the grade, roughness, breathiness, asthenia, strain (GRBAS) scale, the roughness, breathiness, hoarseness (RBH) scale, consensus auditory perceptual evaluation-voice (CAPE-V), and ratings regarding diplophonia. Regarding roughness ratings, Gerratt et al showed limited reliability and drifting over time, especially if no anchor stimuli are available to the listeners.[32] Dejonckere et al reported a moderate correlation between judges for the overall grade of severity of 0.7, and a lower interjudge correlation for asthenia and strain.[33] Also De Bodt et al reported moderate interlistener correlations,[34] and poor to fair correlations were reported by Wuyts et al.[35] Webb et al reported inter-rater kappa statistics ranging from 0.22 to 0.45 (fair to moderate), inter-rater kappa statistics ranging from 0.42 to 0.52 (moderate), and test-retest kappa statistics ranging from 0.56 to 0.64 (moderate to good).[36] The same authors reported later a correlation between grade G and self-perceived voice quality of only 0.32.[37] Sellars et al reported inter- and intrarater reliability rates of 64.7% and 69.6% respectively. Karnell et al and Zraick et al assessed both GRBAS and CAPE-V.[38,39] The former reported good reliability of grade G as well as the overall severity of CAPE-V, reflected by Spearman correlation coefficients exceeding 0.8. The latter showed that intrarater reliability coefficients range from good (0.82 for breathiness B and 0.76 for the overall severity of CAPE-V) to fair or poor (0.35 for the strain S and 0.28 for the pitch of CAPE-V). Studies dealing with the assessment of CAPE-V exclusively were the following. Krival et al reported very good intrarater reliability in terms of intraclass correlation coefficients exceeding 90%, but coefficients regarding interrater reliability ranging from 80% for overall severity to 0% for strain.[40] Helou et al investigated the influence of listener experience on reliability.[41] Reported intraclass correlation coefficients regarding overall severity ratings were 0.91 (intrarater reliability) and 0.72 (inter-rater reliability) for experienced listeners. Coefficients were smaller for inexperienced listeners (0.838 and 0.528).

Solomon et al compared CAPE-V ratings made in a clinical setting, ie, with raters' knowledge of patient's identity and clinical status, against those made under randomized and blinded laboratory conditions.[42] Intraclass correlations across raters were moderate in the laboratory setting for overall severity (0.645). Correlations of clinical ratings with laboratory ratings ranged from 0.526 to 0.792. Klechner et al reported intraclass correlation coefficients for inter-rater reliability that ranged from 71% for breathiness to 35% for strain.[43] Mean coefficients for intrarater reliability ranged from 87% for overall severity to 63% for strain. Regarding RBH, moderate inter-rater agreement was reported by Koreman et al[44]. Regarding the judgment of diplophonia, Aichinger et al reported moderate inter-rater reliability (Cohen's Kappa = 0.67), and good intrarater reliability (0.86 and 0.88).[45] Sakakibara et al showed imperfect agreement among 6 listeners.[46] Hammarberg et al as well as Iwarsson et al reported that reliability of diplophonia ratings is high.[47,48] An overview regarding the issues related to auditory-perceptual judgment of voice quality is given in Barsties and De Bodt.[49] In addition, several types of cognitive biases were reported in the past and may be considered in future discussions on issues of auditory/perceptual analysis.[50]

The objectives of the study are the following. A development of the diplophonia diagram is tested that is able to analyze running speech via automated temporal segmentation. First, a feature called "diplophonia rate" is proposed that quantifies the frequency of occurrence of diplophonia in recordings. Second, the "diplophonia rate" obtained via manual annotations of audio recordings is used to distinguish groups of speakers according to the frequency of occurrence of diplophonation. Finally, because manual annotations are tedious to obtain, it is investigated if a computerized estimate of the diplophonia rate enables (1) the automatic detection of frequently diplophonic subjects, (2) the automatic detection of nondiplophonic subjects, and (3) an automatic 3-way classification into frequently diplophonic, unfrequently diplophonic and nondiplophonic speakers via multinomial logistic regression.

MATERIAL AND METHODS

Data collection

Forty subjects who were clinically rated as diplophonic in the past are included in the study.[51,52] The subjects were recruited among the outpatients of the Medical University of Vienna, Austria, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics. Each subject was rated by an experienced clinician as diplophonic if he/she diplophonated in any utterance during clinical examination, ie, during the conversation between the clinician and the subject, during stroboscopic examination, during the recording of the voice range profile, when sustaining vowels, or when reading the standard text. For each subject, an audio recording of the German standard text "Der Nordwind und die Sonne" has been obtained. The study has been approved by the ethical review committee of the Medical University of Vienna, Austria. The observed diagnoses are summarized in Table 1.

TABLE 1.
Summary of Diagnoses of Included Subjects

Diagnosis	Number of Subjects
Bamboo nodes	1
Benign tumor	1
Cyst	4
Dysfunction	5
Edema	6
Laryngitis	2
Nodules	1
Paresis	13
Polyp	3
Scar	1
Sulcus	2
Unknown	1

Manual labeling

To obtain the reference segmentation and labeling, the audio recordings of the diplophonic subjects are annotated manually by the first author with regard to phonatory intervals during which diplophonia and voicing are observed. Figure 1 shows an example. Interval boundaries are positioned using Praat, and intervals are labeled according to the presence (1) or absence (0) of diplophonia, as well as voicing.[53]

Figure 2 illustrates the beating phenomenon. The upper plot shows a waveform of a sine with a frequency of 138 Hz and the middle plot a sine with a frequency of 117 Hz. The sines have frequencies that are typical of fundamental frequencies observed in diplophonic voices. The bottom plot shows the sum of the two sines and a sine of 21 Hz, ie, the difference between the frequencies of the sines. One sees a beating in the sum, which is characterized by, according to the American National Standard for Acoustical Terminology, "[p]eriodic variations that result from the superposition of two simple harmonic quantities of different frequencies, f_1 and f_2. Beats involve the periodic increase and decrease of the amplitude of the sum at the beat frequency, $(f_1 - f_2)$."[54] The sines amplify or attenuate each other according to their relative phase. When both are in phase, they interfere constructively, and when they are out of phase they interfere destructively. The result is a periodic increase and decrease of the amplitude of their sum. The dashed 21 Hz sine approximates the peak amplitudes of the sum. Similar patterns are observed in diplophonic voices.

A "diplophonia rate" (DR) is proposed, which is the ratio of the total duration of diplophonation T_D and the total duration of voiced phonation T_V, in percent (Equation 1).

$$DR = \frac{T_D}{T_V} \cdot 100\% \tag{1}$$

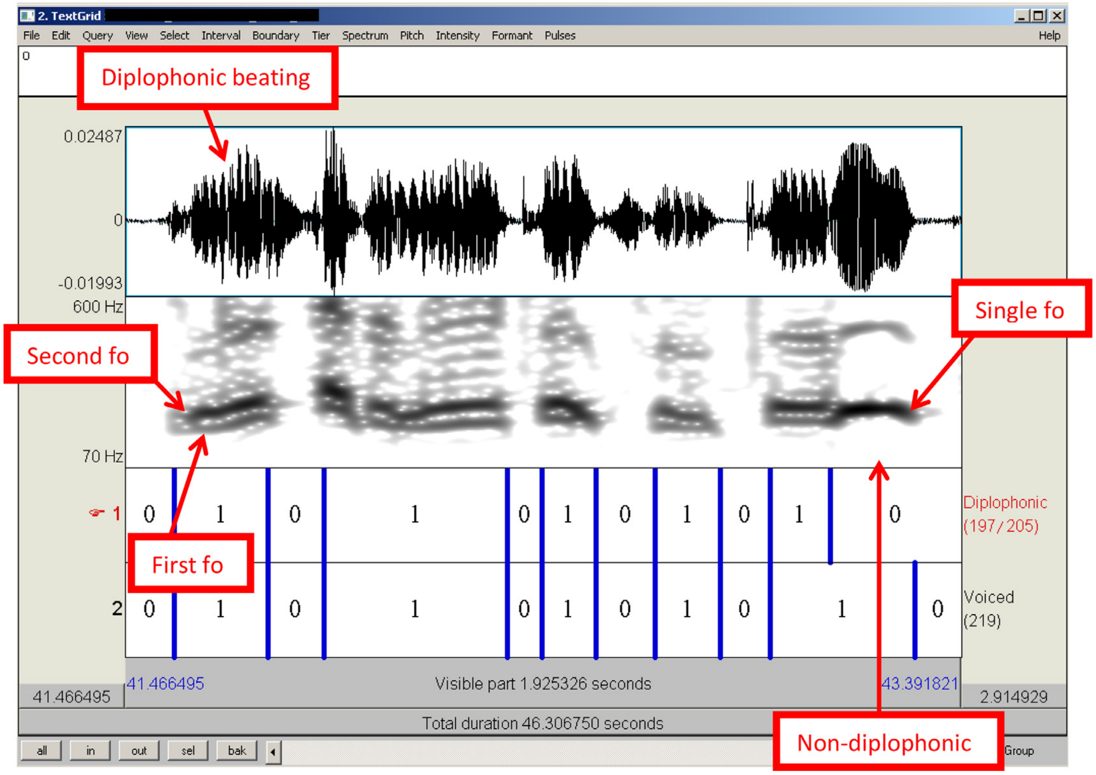

FIGURE 1. Example of a recording of a frequently diplophonic subject, annotated with Praat. The signal corresponds to the text fragment "… da musste der Nordwind zugeben, … ." The top plot shows the waveform, and the plot below the spectrogram. Annotations regarding the presence of diplophonia and voicing are underneath. Diplophonic and voiced intervals are labeled 1, whereas nondiplophonic and unvoiced segments are labeled 0. During diplophonation, beating is observed in the waveform, and two fundamental frequencies are observed in the spectrogram. Only one fundamental frequency is observed in a nondiplophonic voiced interval.

The diplophonia rate obtained via manual annotation is denoted as DR_m.

Subjects who never diplophonate during the recorded readings are assigned to a nondiplophonic group. For the remaining subjects, the median of the DR_m is obtained and used as a separating threshold to establish a distinction between subjects who diplophonate frequently and unfrequently. These subject group assignments based on manual annotations provide the reference labels.

Automated analysis

All audio recordings are analyzed with a purpose built software audio signal processor that estimates the number of fundamental frequencies that co-exist. The processing involves (1) spectral peak picking, (2) multiple Viterbi tracking of the peak positions, which carries out automatic temporal segmentation and thus enables the analysis of recordings of connected speech, (3) candidate waveform synthesis, and (4) candidate selection. In spectral peak picking, the frequencies that correspond to the maxima of the audio recordings' spectrograms are identified. These frequencies are fundamental frequency candidates that are extracted frame by frame. They include the fundamental frequencies, their partials/harmonics, and the frequencies of the combination tones, ie , the beats. Multiple Viterbi tracking is applied to connect peak positions (frequencies) over time. Thus, tracking identifies those frequencies that belong together over subsequent frames. The temporal evolution of the tracks is represented via a hidden Markov model, which is a probabilistic model of the voice onset and offset probabilities, as well as the tracks' temporal evolution, ie, the rise and fall of the frequencies in time. Next, for each candidate track, a candidate waveform is obtained via a filter that only lets through the fundamental together with the corresponding harmonics. Candidate waveforms are added together and compared to the audio recording by sample wise subtraction. All possible candidate combinations are evaluated, and the one combination is retained that minimizes the sample wise difference between the candidates' sum and the audio recording. This combination gives the final estimate of the fundamental frequencies. Additional explanations regarding the method are given elsewhere.[2,55]

Signal fragments during which two fundamental frequencies are detected are labeled as "diplophonic," whereas other fragments are labeled as "nondiplophonic." Fragments without a detected fundamental frequency are labeled as "unvoiced," whereas the remaining fragments are labeled as "voiced" (either diplophonic or monophonic). The automatically estimated diplophonia rate is denoted as DR_e, (Equation 1). The group for each subject is predicted via statistical analysis, using DR_e as predictor. These group assignments provide the subjects' predicted labels.

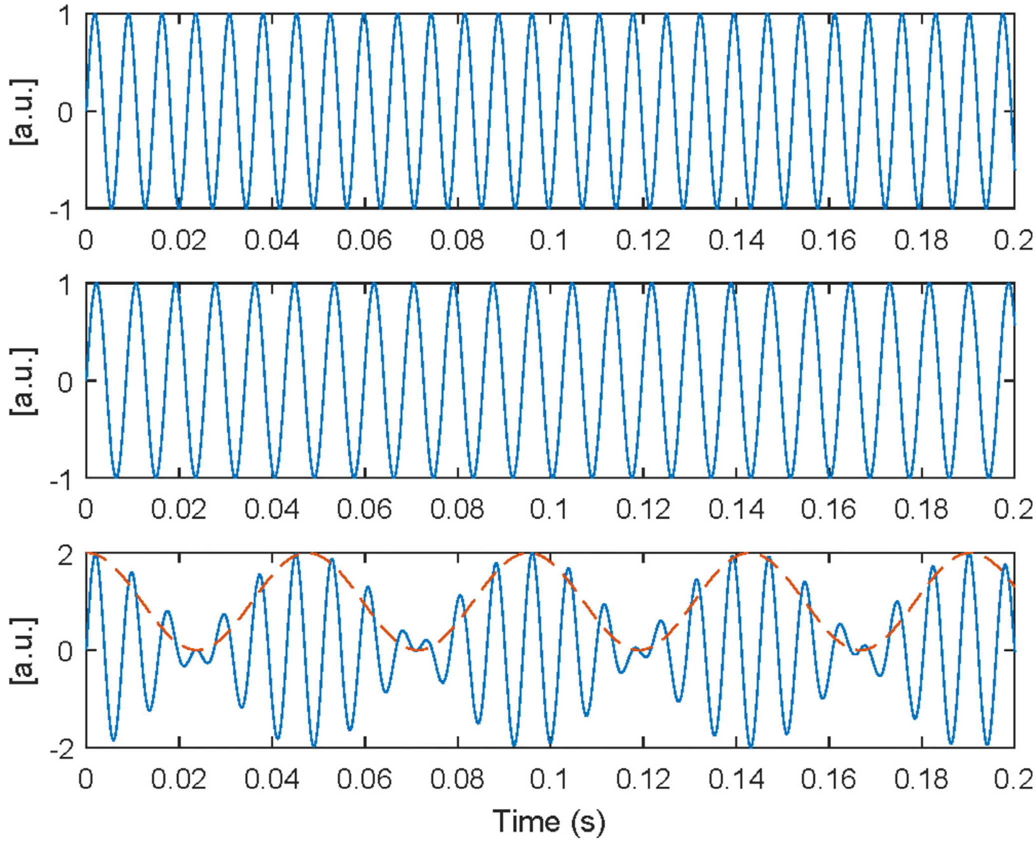

FIGURE 2. Didactic illustration of the beating phenomenon. The upper and middle plots show the waveforms of two sines $s_1 = \sin(2\pi f_1 \bullet t)$ and $s_2 = \sin(2\pi f_2 \bullet t)$, with $f_1 = 138$ Hz and f_2=117 Hz. The bottom plot shows the sum $s_1 + s_2$, and a cosine $c = 1 + \cos(2\pi(f_1 - f_2))$ (dashed line), which approximates the amplitude of the sum $s_1 + s_2$.

Statistics

First, descriptive statistics regarding the number of frequently and unfrequently diplophonating subjects, as well as nondiplophonic subjects are reported. Second, the automatically obtained diplophonia rate DR_e is compared to manually obtained groups. An analysis of variance (Anova) and three post-hoc t tests for pairwise group comparisons are carried out. Third, two DR_e cut-off threshold classifiers are proposed by means of receiver operating characteristic curves. The first serves to detect frequently diplophonic subjects, and the second to detect nondiplophonic subjects. Finally, multinomial logistic regression is used for a 3-way classification, which is reported in a cross-table.

RESULTS

Figure 3 shows the distribution of the manually obtained rates DR_m for each subject who diplophonates during the recorded standard text reading (n = 28). The DR_m ratio ranges from close to 0.1% to almost 100%. No subject is included who diplophonates permanently in the strict sense (DR_m = 100%). The median of the DR_m is 8.39% and separates the subjects into an unfrequently diplophonic (n = 14) and a frequently diplophonic (n = 14) group. These group assignments constitute the subjects' reference labels.

Figure 4 shows the boxplots of DR_e with respect to the reference labels. The estimated DR_e increases across groups from nondiplophonic to unfrequently diplophonic to frequently diplophonic. Anova confirms the statistical significance of the increase ($P = 0.005$). Three-fold Bonferroni corrected post-hoc testing reports a significant difference between the means of the nondiplophonic and frequently diplophonic groups ($P = 0.036$).

Figures 5 and 6 show the receiver operating characteristic curves regarding the automatic detection of (1) frequently diplophonic subjects, and (2) nondiplophonic subjects. Subjects with DR_e above 9.04% are predicted to be frequently diplophonic, whereas subjects with a DR_e below 5.91% are predicted to be nondiplophonic. The threshold of 9.04% is close to the median used in the categorization based on manual annotation (8.39%).

Regarding the detection of frequently diplophonic subjects an area under the curve of 0.77, a sensitivity of 71%, a specificity of 88%, and an accuracy of 83% are observed. Regarding the detection of nondiplophonic subjects an area under the curve of 0.84, a sensitivity of 68%, a specificity of 92%, and an accuracy of 73% are observed[1].

[1]Detection of nondiplophonic subjects is equivalent to the detection of diplophonic subjects, if frequently and unfrequently diplophonic subjects are pooled into one positive group. In this option, the sensitivity and the specificity are swapped, ie, a sensitivity of 92% and a specificity of 68% are observed.

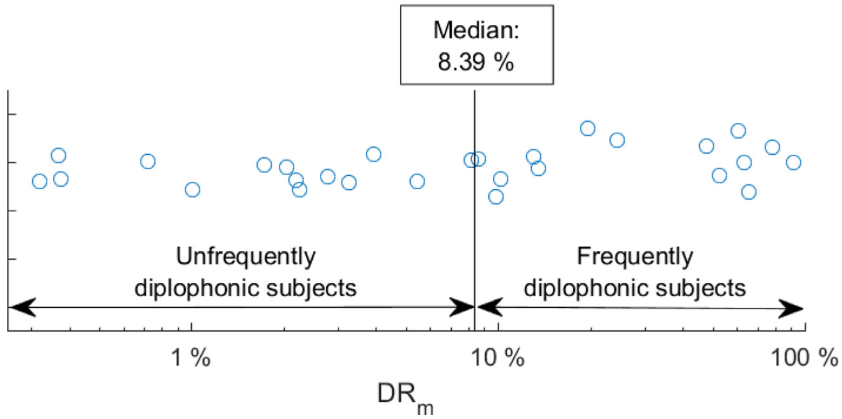

FIGURE 3. Diplophonia rates DR_m, obtained via manual annotation of audio recordings. Each circle represents one subject. The subjects are separated into a frequently diplophonic (n = 14) and an unfrequently diplophonic (n = 14) group using the median of 8.39% as the threshold. Subjects who do not diplophonate during the recorded standard text readings are not shown (n = 12). These group assignments constitute the subjects' reference labels.

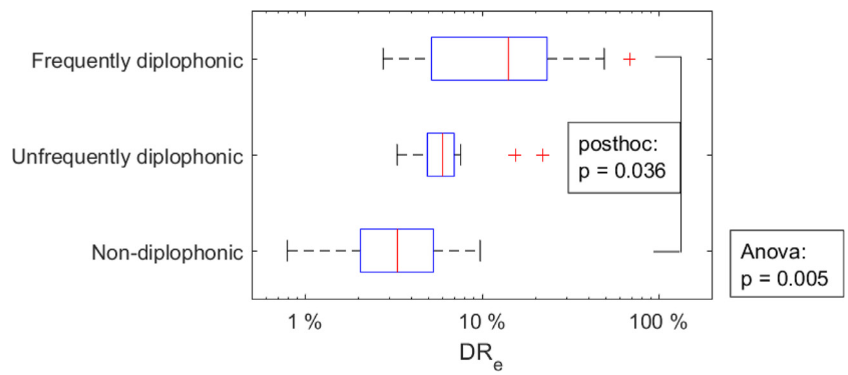

FIGURE 4. Boxplots of automatically estimated diplophonia rates DR_e, which increase across groups. An analysis of variance (Anova) confirms the statistical significance of the increase. Threefold Bonferroni corrected post-hoc testing reports a significant difference between the means of the nondiplophonic and the frequently diplophonic groups.

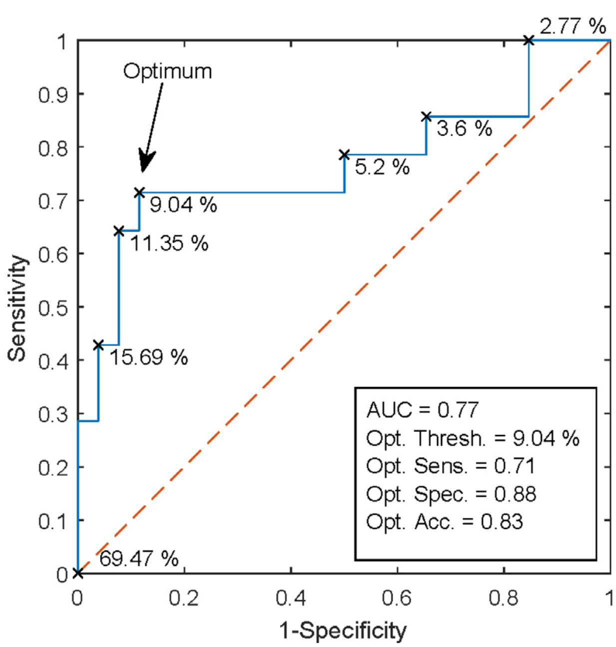

FIGURE 5. Receiver operating characteristic curve regarding the detection of frequently diplophonic subjects. Subjects are predicted to be frequently diplophonic, if the estimated diplophonia rate exceeds 9.04%. The sensitivity and specificity are 71% and 88%.

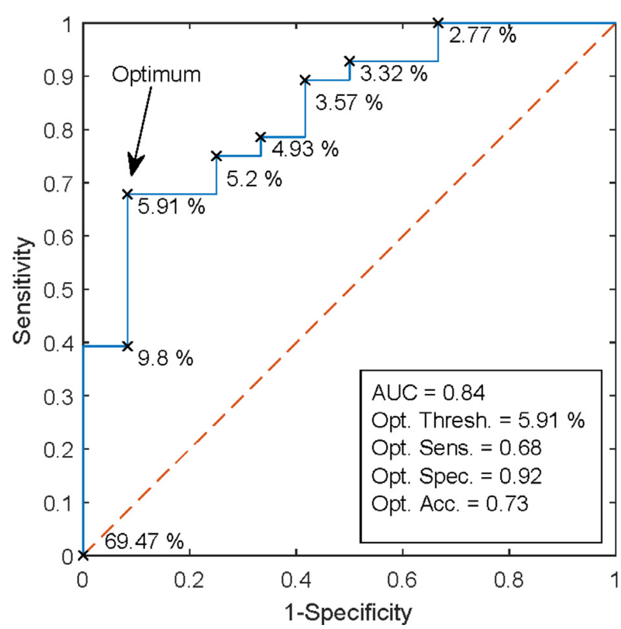

FIGURE 6. Receiver operating characteristic curve regarding the detection of nondiplophonic subjects. Subjects are predicted to be nondiplophonic, if the estimated diplophonia rate is below 5.91%. The sensitivity and specificity are 68% and 92%.

TABLE 2.
Cross-Table Reporting the Numbers of Subjects Assigned Via 3-Way Classification by Means of Multinomial Logistic Regression. Manually Annotated and Automatically Assigned Subject Groups are Shown in the Rows and Columns. The Number of Correct Classifications are on the Diagonal (highlighted). 25 (9+9+7) Out of 40 Subjects are Assigned to the Correct Group Automatically, ie, 62.5% of the Subjects

		Predicted labels (based on $\boldsymbol{DR_e}$)		
		Frequently diplophonic	Unfrequently diplophonic	Non-diplophonic
Reference labels (based on $\boldsymbol{DR_m}$)	Frequently diplophonic	9	2	3
	Unfrequently diplophonic	2	9	3
	Non-diplophonic	0	5	7

Table 2 reports the numbers of subjects regarding a multinomial logistic regression 3-way classification, using DR_e as the predictor. Subject groups based on manual annotations are shown row-wise (reference labels), and groups predicted automatically via DR_e are shown columnwise (predicted labels). The number of correct classifications are shown on the diagonal. 25 (9+9+7) out of 40 subjects are assigned to the correct group automatically, ie, 62.5% of the subjects.

DISCUSSION

The diplophonia rate quantifies the frequency of occurrence of diplophonic episodes in human voices. DR_m is obtained via manual annotation of the recordings and used to obtain the reference labels for each subject (1) frequently diplophonic, (2) unfrequently diplophonic, and (3) nondiplophonic. The first two relate to the groups referred to as "permanent" and "intermittent" diplophonia.[1] To the best of our knowledge, we provide the first descriptive statistics regarding the frequency of occurrence of diplophonation in standard text readings by subjects clinically labeled as diplophonic. The automated prediction of these labels via an estimation of the DR_e appears to be possible.

An added value of the proposed method is that it is applicable to recordings of read speech, which enables the analysis of patients' voices in a standardized setting. We had observed previously in other settings that changes in the recording conditions and instructions given by the examiner may influence the occurrence of diplophonia, which had raised the issues of limited repeatability, reliability, and objectivity. However, we observe that not all subjects clinically labeled as diplophonic diplophonate during the reading of a standard text, which now raises the issue of limited ecological validity of standard text readings. In recordings of standard texts read by subjects who were clinically rated as diplophonic, approximately only two-thirds of the subjects diplophonate. Patients were initially labeled as diplophonic if they diplophonated in any utterance during clinical examination, ie, during the conversation between the clinician and the subject, during stroboscopic examination, during the recording of the voice range profile, when sustaining vowels, or when reading. One may argue that the analysis of standard text readings better relates to the (self-) perceived voice handicap than sustained vowels, and that a gold standard may be ambulatory monitoring. In principle, our tool may also be used in ambulatory monitoring.

A mean DR_m of 28.4% was obtained in a previous study of vowels sustained by diplophonic subjects, which is higher than the median DR_m of 8.39% observed in the present study. We interpret this difference as an effect of a deliberate instigation of the patients during the recordings of the previous study. In particular, patients were instructed to adjust pitch and loudness of their sustained phonation to elicit diplophonia. The increase of DR that may be attributed to deliberate instigation can be interpreted as an observer-induced bias. Thus, analyses of standard text readings may be more objective than the past analyses of sustained vowels. Also, analyses of standard text readings may relate better to voice handicap.

The threshold that is found to detect frequently diplophonic recordings ($DR_e > 9.04\%$) is close to the median used in the categorization based on manual annotation ($DR_m > 8.39\%$). This agreement justifies speculations regarding the importance of a threshold close to 9%. The possibility exists that this threshold is an auditorily salient one, and clinically relevant. Such a threshold could be built into future tools for automatic categorization of frequently and unfrequently diplophonic speakers. Additional work that is needed to confirm this threshold includes the analysis of the distribution of diplophonia rates in clinical practice, and experiments with naïve and expert listeners. One may hypothesize that (i) distinct clusters of diplophonia rates are observed above and below 9% on the DR scale if more patients are included, and (ii) naïve listeners may only be able to detect diplophonia auditorily if it occurs frequently. If this would hold true, mild cases of diplophonia, ie, unfrequent diplophonia, may remain unrevealed in subjects with voice problems for a long time if no clinical voice specialist is sought out timely.

The largest accuracy observed in the current study is 83% for the distinction between frequently diplophonic subjects and others. This accuracy is for two reasons smaller than the one observed in a previous study (94.2%). First, connected speech is more variable than sustained vowels, which makes the analysis more difficult. Second, the manual pre-segmentation of audio recordings that was done in the past favored the correct classification of the signals by the Diplophonia Diagram. The automatic temporal segmentation in the present study enables automatic analysis of connected speech but decreases the accuracy.

Limitations of our approach exist, which motivate further research. First, we assumed that the percept "pitch" is equivalent to the fundamental frequency of the speech signal. This assumption is necessary to enable automated detection, but is not true strictly speaking. It is known from psychoacoustics, that the perception of pitch is not exclusively linked to the frequencies of prominent peaks in the magnitude spectrum.[56,57] This partly explains differences observed between the manually obtained DR_m and the automatically obtained DR_e. Second, another difficulty is the correct manual labeling of the audio recordings. In particular, some speech fragments are borderline diplophonic due to a slight phase desynchronization of the vocal folds. These fragments are ambiguous with regard to perceived diplophonia, which shows that the labeling of diplophonic speech fragments based on perception and spectral analysis has limitations. The same is true for the labeling of voicing in highly irregular voices, ie, the voiced/unvoiced status may be unclear. Finally, it has been decided to define 3 groups regarding the frequency of occurrence of diplophonation, which is a more or less arbitrary choice. In particular, we see in Figure 3 that DR_m may be distributed uniformly across its whole range. Thus, no distinct clusters are observed in the data, which would motivate the use of a finer grained scale. However, in preliminary experimentation, we also used terciles instead of the median in order to distinguish 4 groups regarding the frequency of occurrence of diplophonia. Classification performance and statistical significance decreased a lot when more than 3 groups were used.

Investigation of features of vocal fold vibration was beyond the scope of this study, but had laid the foundations of our approach in the past.[29,51,55] In particular, a distinction between left−right asymmetry, and anterior−posterior asymmetry is well established for diplophonic voices. In left−right asymmetry, the left and the right vocal folds vibrate at different frequencies. In anterior−posterior asymmetry, the anterior and the posterior parts of the vocal folds vibrate at different frequencies. While video analyses are more cause-oriented, the advantage of audio analyses is their noninvasiveness, which enables faster and cheaper completion, as well as a less obstructive examination.

The expected clinical utility of the proposed tool is here discussed with reference to suggestions for future work. First, the proposed tool can be further validated by pre-/post-treatment comparisons. Ratings and features of diplophonia should be compared prior to and following voice therapy or medical/surgical management of dysphonia. Second, the proposed tool can be used for the training of raters. Thus, clinicians can use the results provided by the tool as feedback to their own perceptual ratings. Agreement between their ratings with the automatically obtained diplophonia rate may hereby be improved. As a benefit, reliability of perceptual ratings may hereby be improved. Thus, clinical judgments of diplophonia can be validated with the proposed tool in the future. In particular, the algorithm may be implemented in a software that runs on standard personal computers, or mobile devices, such as smart phones. Users may install the software on their own devices and use it preferably with high-quality external microphones and in acoustically treated rooms.

Other suggestions for future work with a view toward the proposed tool's clinical and research utility are the following. First, one might wish to investigate whether the frequency of occurrence of diplophonia may be obtained at a finer grained scale, ie, whether more than the 3 groups are distinguishable. From a statistical point of view the realization of this wish also depends on the amount of available data, which may be increased in the framework of a multicenter study. In addition, we observed different types of diplophonia that are distinguishable auditorily. In particular, a rumbling type of diplophonia caught our attention. It may be due to a slow beating caused by two close fundamental frequencies. The distinction of different types of diplophonia may be included in future voice assessment procedures that are more accurate than existing ones. Second, the frequencies of occurrence of diplophonia may be reported for other types of vocal utterances that are used for voice assessment in clinical practice, eg, for spontaneous/conversational speech, singing, including application in ambulatory monitoring. Third, the proposed tool may facilitate the investigation (i) whether other acoustic features besides doubled fundamental frequencies may contribute to clinical ratings of diplophonia, and (ii) whether doubled fundamental frequencies occur in voice signals that are not perceived as diplophonic, and what the reasons may be. Features of vocal fold vibration may contribute to acoustic features that factor into the model, and may also be investigated in the future using laryngeal high-speed videos and/or (multiline) kymograms. Finally, mutual training may not only involve the training of raters based on feedback obtained from the tool, but also training of the tool based on feedback obtained from raters.

CONCLUSION

The ecological validity of recordings of the readings of a standard text was found to be limited with regard to diplophonia. Twelve of the patients who were clinically labeled as diplophonic did not diplophonate during the recording of the standard text reading, whereas 14 diplophonated unfrequently (DR_m < 8.39%) and another 14 diplophonated frequently (DR_m > 8.39%). Thus, approximately only two-thirds of the subjects who were clinically rated as diplophonic diplophonate during read speech. No subject diplophonated permanently in the strict sense (DR_m = 100%). However, read speech is ecologically more valid than sustained vowels, and read speech may thus be preferred for voice assessment. Also, the examiner's influence on recordings of read speech may be smaller than on recordings of sustained vowels.

Diplophonations could be detected with the described automated analysis in read speech. Perfect accuracy cannot be achieved, given the residual uncertainties involved in manual reference labeling of diplophonic speech fragments. Nevertheless, the performance of the proposed method for detecting diplophonia automatically from running speech is promising.

REFERENCES

1. Dejonckere P, Lebacq J. An analysis of the diplophonia phenomenon. *Speech Commun*. 1983;2:47–56.
2. Aichinger P, Roesner I, Schneider-Stickler B, et al. Towards objective voice assessment: the diplophonia diagram. *J Voice*. 2017;31. 253. e17−253.e26.
3. Krengli M, Policarpo M, Manfredda I, et al. Voice quality after treatment for T1a glottic carcinoma. *Acta Oncol*. 2004;43:284–289.
4. Bibby JRL, Cotton SM, Perry A, et al. Voice outcomes after radiotherapy treatment for early glottic cancer: assessment using multidimensional tools. *Head Neck*. 2008;30:600–610.
5. Bertino G, Bellomo A, Ferrero FE, et al. Acoustic analysis of voice quality with or without false vocal fold displacement after cordectomy. *J Voice*. 2001;15:131–140.
6. Nishiyama K, Hirose H, Iguchi Y, et al. Autologous transplantation of fascia into the vocal fold as a treatment for recurrent nerve paralysis. *The Laryngosc*. 2002;112:1420–1425.
7. Iwamura S, Kurita N. A newer arytenoid adduction technique for one-vocal-fold paralysis: a direct pull of the lateral cricoarytenoid muscle. (In Japanese). *Head Neck Surg. (Tokyo)*. 1996;6:1–10.
8. Tsukahara K, Tokashiki R, Hiramatsu H. A case of high-pitched diplophonia that resolved after a direct pull of the lateral cricoarytenoid muscle. *Acta Oto-Laryngol*. 2005;125:331–333.
9. Kimura M, Imagawa H, Nito T, et al. Arytenoid adduction for correcting vocal fold asymmetry: high-speed imaging. *Ann Otol Rhinol Laryngol*. 2010;119:439–446.
10. Molteni G, Bergamini G, Ricci-Maccarini A, et al. Auto-crosslinked hyaluronan gel injections in phonosurgery. *Otolaryngol Head Neck Surg*. 2010;142:547–553.
11. Lim J-Y, Lim SE, Choi SH, et al. Clinical characteristics and voice analysis of patients with mutational dysphonia: clinical significance of diplophonia and closed quotients. *J Voice*. 2007;21:12–19.
12. Kocak I, Dogan M, Tadihan E, et al. Window anterior commissure relaxation laryngoplasty in the management of high-pitched voice disorders. *Arch Otolaryngol Head Neck Surg*. 2008;134:1263–1269.
13. Dejonckere P, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Oto-Rhino-Laryngol*. 2001;258:77–82.
14. Deliyski D. Acoustic model and evaluation of pathological voice production. *Third Eur Conf Speech Commun Technol.*. 1993;3:1969–1972.
15. Drioli C, Foresti GL, Scienze V. Enhanced video kymographic data analysis based on vocal folds dynamics modeling. In: *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*. 2172–2176.
16. Echternach M, Döllinger M. Vocal fold vibrations at high soprano fundamental frequencies. *J Acoust Soc Am*. 2013;133:EL82–EL87.
17. Isshiki N, Tanabe M, Ishizaka K. Clinical significance of asymmetrical vocal cord tension. *Ann Otol Rhinol Laryngol*. 1977;86:58–66.
18. Sakakibara KI, Imagawa H, Kimura M, et al. Modal analysis of vocal fold vibrations using laryngotopography. In: *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*. 917–920.
19. Tralie CJ and Perea JA. (Quasi)Periodicity quantification in video data, using topology. *SIAM J on Imaging Sci.* 2018;11:1049−1077.
20. Unger J, Meyer T, Herbst CT. Phonovibrographic wavegrams: visualizing vocal fold kinematics. *J Acoust Soc Am*. 2013;133:1055–1064.
21. Unger J, Schuster M, Hecker DJ, et al. A generalized procedure for analyzing sustained and dynamic vocal fold vibrations from laryngeal high-speed videos using phonovibrograms. *Artif Intell Med*. 2016;66:15–28.
22. Wang J-S, Olszewski E, Devine EE, et al. Extension and application of high-speed digital imaging analysis via spatiotemporal correlation and eigenmode analysis of vocal fold vibration before and after polyp excision. *Ann Otol Rhinol Laryngol*. 2016;125:660–666.
23. Woo P, Baxter P. Flexible fiber-optic high-speed imaging of vocal fold vibration: a preliminary report. *J Voice*. 2017;31:175–181.
24. Yan Y, Ahmad K, Kunduk M, et al. Analysis of Vocal-fold Vibrations from High-Speed Laryngeal Images Using a Hilbert Transform-Based Methodology. *J Voice*. 2005;19:161–175.
25. Yan Y, Chen X, Ahmad K, et al. High-speed laryngeal imaging analysis of vocal fold dynamics. *International Conference on Voice Physiology and Biomechanics*. 4.
26. Zhang Y, Bieging E, Tsui H, et al. Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging. *J Voice*. 2010;24:21–29.
27. Tigges M, Mergell P, Herzel H, et al. Observation and modelling of glottal biphonation. Acust Acta Acus*t*. 1997;83:707–714.
28. Neubauer J, Mergell P, Eysholdt U, et al. Spatio-temporal analysis of irregular vocal fold oscillations: biphonation due to desynchronization of spatial modes. *J Acoust Soc Am*. 2001;110:3179–3192.
29. Aichinger P, Roesner I, Leonhard M, et al. Comparison of an audio-based and a video-based approach for detecting diplophonia. *Biomed Signal Process Control*. 2017;31:576–585.
30. Aichinger P, Roesner I, Schneider-Stickler B, et al. Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation. In: *Proceedings of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. 81–84.
31. Aichinger P, Schneider-Stickler B, Bigenzahn W, et al. Double pitch marks in diplophonic voice. In: *Proceeedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7437–7441.
32. Gerratt BR, Kreiman J, Antonanzas-Barroso N, et al. Comparing internal and external standards in voice quality judgments. *J Speech Hearing Res*. 1993;36:14–20.
33. Dejonckere PH, Obbens C, de Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr. (Basel)*. 1993;45:76–83.
34. De Bodt MS, Wuyts FL, Van De Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
35. Wuyts FL, De Bodt MS, Van De Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517.
36. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Oto-Rhino-Laryngol*. 2004;261:429–434.
37. Webb AL, Carding PN, Deary IJ, et al. Optimising outcome assessment of voice interventions, I: reliability and validity of three self-reported scales. *J Laryngol Otol*. 2007;121:763–767.

38. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576–590.
39. Zraick R, Kempster G, Connor N, et al. Establishing validity of the consensus auditory perceptual evaluation of voice (CAPE-V). *Am J Speech-Lang Pathol*. 2011;20:14–22.
40. Krival K, Kelchner L, Weinrich B, et al. Vibratory source, vocal quality and fundamental frequency following pediatric laryngotracheal reconstruction. *Int J Pediatr Otorhinolaryngol*. 2007;71:1261–1269.
41. Helou L, Solomon N, Henry L, et al. The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V) ratings of post-thyroidectomy voice. *Am J Speech-Lang Pathol*. 2010;19:248–258.
42. Solomon N, Helou LB, Stojadinovic A. Clinical versus laboratory ratings of voice using the CAPE-V. *J Voice*. 2011;25:e7–14.
43. Kelchner L, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
44. Koreman J, Pützer M, Just M. Correlates of varying vocal fold adduction deficiencies in perception and production: methodological and practical considerations. *Folia Phoniatr Logopaedica*. 2004;56:305–320.
45. Aichinger P, Hagmüller M, Roesner I, et al. Diplophonia disturbs jitter and shimmer measurement. *Folia Phoniatr Logopaedica*. 2016;68:22–28.
46. Sakakibara KI, Imagawa H, Yokonishi H, et al. Physiological observations and synthesis of subharmonic voices. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. 1079–1085.
47. Hammarberg B, Fritzell B, Gaufin J, et al. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol*. 1980;90:441–451.
48. Iwarsson J, Petersen NR. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *J Voice*. 2012;26:304–312.
49. Barsties B, De Bodt M. Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx*. 2015;42:183–188.
50. “List of cognitive biases,” *Wikipedia*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/List_of_cognitive_biases. Accessed July 15, 2018.
51. Aichinger P. *Diplophonic Voice—Definitions, Models, and Detection*. Austria: Graz University of Technology; 2015. PhD dissertation.
52. Aichinger P, Roesner I, Leonhard M, et al. A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 10, 767–770.
53. Boersma P, Weenink D. Praat: doing phonetics by computer; 2014. [Online]. Available:; 2014. www.praat.org. Accessed July 14, 2014.
54. “American National Standard—Acoustical Terminology ANSI S1.1-1994 (ASA 111-1994). Revision of ANSI S1.1-1960 (R 1976),” vol. 1994. New York, 2005.
55. Aichinger P, Hagmuller M, Schneider-Stickler B, et al. Tracking of multiple fundamental frequencies in diplophonic voices. *IEEE/ACM Trans Audio, Speech, Lang Process*. 2018;26:330–341.
56. Terhardt E. Pitch, consonance, and harmony. *J Acoust Soc Am*. 1974;55:1061–1069.
57. Parncutt R. *Harmony: A Psychoacoustical Approach*. Berlin: Springer; 1989.