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. 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. 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. Krenzli et al evaluated and compared the voice quality of 27 patients undergoing radiotherapy and 30 patients undergoing laser cordectomy. Diplophonia occurred in 18.5% of the patients after radiotherapy, and in 23.3% of the patients after laser
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. 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. 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. 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, and Tsukahara et al reported another case in which a direct pull of the lateral cricoarytenoid muscle resolved high-pitched diplophonia. Kimura et al evaluated effects of arytenoid adduction on diplophonia in 6 patients via analysis of high-speed videolaryngoscopy. Diplophonia disappeared in all cases after treatment. Molteni et al assessed auto-crosslinked hyaluronan gel injections in a group of 40 patients. 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. 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. 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. In a recent objective approach, the simultaneous existence of two fundamental frequencies is detected in sustained vowels. Also, the degree of subharmonics is used as an indicator for the presence of diplophonia. Additionally, high-speed videolaryngoscopy and (multiline) kymographic imaging were used in the past to investigate diplophonia.

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. 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. 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. Also De Bodt et al reported moderate interlistener correlations, and poor to fair correlations were reported by Wuyts et al. 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-rest kappa statistics ranging from 0.56 to 0.64 (moderate to good). The same authors reported later a correlation between grade G and self-perceived voice quality of only 0.32. 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. 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. Helou et al investigated the influence of listener experience on reliability. 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. 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. 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. 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). Sakakibara et al showed imperfect agreement among 6 listeners. Hammarberg et al as well as Iwarsson et al reported that reliability of diplophonia ratings is high. An overview regarding the issues related to auditory/perceptual judgment of voice quality is given in Barsties and De Bodt. 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.

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 diplophonia. 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. 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.

**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.

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, \( (f_1 - f_2) \).”

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).

\[
DR = \frac{T_D}{T_V} \times 100\%
\]
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 infrequently. 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, i.e., 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, i.e., 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.
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 $D_{Re}$ 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 $D_{Re}$ 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 $D_{Re}$ with respect to the reference labels. The estimated $D_{Re}$ 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 $D_{Re}$ above 9.04% are predicted to be frequently diplophonic, whereas subjects with a $D_{Re}$ 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.

$^1$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 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. 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 naive and expert listeners. One may hypothesize that (i) distinct clusters of diplophonia rates are observed above and
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 diplophonied did not diplophonied during the recording of the standard text reading, whereas 14 diplophonied unfrequently ($DR_m < 8.39\%$) and another 14 diplophonied frequently ($DR_m > 8.39\%$). Thus, approximately only two-thirds of the subjects who were clinically rated as diplophonied diplophonied during read speech. No subject diplophonied 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.

Diplophoniedations 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 diplophonied speech fragments. Nevertheless, the performance of the proposed method for detecting diplophonieda automatically from running speech is promising.

REFERENCES


