A Real-time Objective Speech Analysis Tool for Analysis of Impaired Speech

Karolin Krüger[1], Patricia Weede[1], Gerhard Schmidt[1]

[1] *Digital Signal Processing and System Theory, Department of Electrical and Information Engingeering,*
Kiel University, 24143 Kiel, Germany, Email: {kkru, pp, gus}@tf.uni-kiel.de

Introduction

Speech is a fundamental element of communication, playing a key role in personal relationships, social life, and professional success. Therefore, it is integral to human identity [1]. A person's voice can be affected by neurological or organic conditions, or due to functional voice disorders or high vocal demands leading to occupational dysphonia. Common symptoms include changes in voice quality, hoarseness, reduced vocal stamina, frequent throat clearing, and a sensation of effort while speaking [2]. Speech therapy can often help to reduce these limitations. Thereby, the focus is mainly on the patient's personal experience and difficulties, meaning a formal diagnosis is not always necessary. Improvements in voice quality are often assessed subjectively by the speech therapist or by using offline analysis tools to provide spectral analysis and further parameters like for example jitter or shimmer [2, 3].

Therefore, our goal is to develop an objective speech analysis tool for the analysis of impaired speech, enabling a reproducible, therapist-independent, comparable speech analysis to accompany traditional speech therapy. In addition, we want to allow further insights into changes in speech by creating real-time feedback tests, that could interact with the patient's prerequisites and show how their ability to adapt and react to auditory feedback is impaired or can even be improved. To do so, we implemented a modular and flexible tool in our real-time framework Kiel Real-time Application Toolkit (KiRAT). It enables feature extraction of basic voice parameters (voice activity, loudness, pitch, shimmer, jitter, ...), survey functions to access self-assessment questionnaires, as well as an interactive loudness feedback test. Therefore, we conducted tests with healthy participants to confirm the hypothesis, that auditory feedback during lower loudness phases leads to an increase in the voice level.

Background and Motivation

Voice and speech disorders can be characterized by many different symptoms. Impaired phonation can lead to lowered and unstable voice intensity, rough and breathy sound, as well as a reduced pitch range. Also changes in rhythm and articulation can occur, such as changes in speed and pronunciation. The voice can sound monotonous because of a loss of modulation ability and changes in fluency can occur, as well as a monoloudness [4, 5, 6].

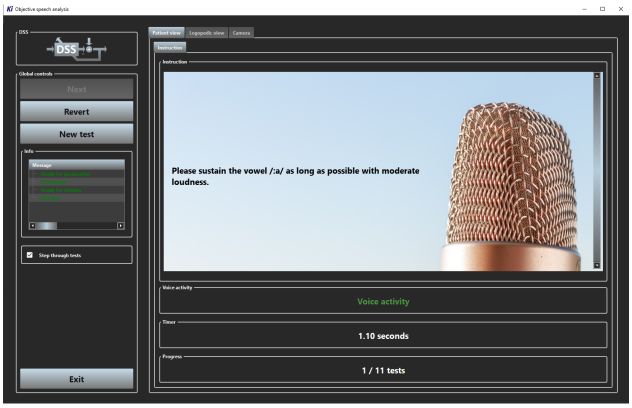
To improve intelligibility, one training concept, Lee Silverman Voice Treatment (LSVT), focuses on training the loudness resulting in an additional increase in speech intelligibility [7]. We wanted to create an adaptive test and training based on this approach and the Lombard effect. Lombard speech is current in noisy environments. It is characterized by an increased vocal intensity level, pitch, and vowel duration, a shift in formant frequency, and spectral tilting, and a reduced speaking rate, leading to more intelligible speech. Since *Al-Fwaress* [6] investigated that the Lombard effect is also present in patients with Parkinson's disease, we assume that as long as there is no organic cause, this is also applicable to various speech disorders. Due to changed speech-motor control in dysarthric patients, the auditory and somatosensory feedback can be altered, resulting in a dysfunctional feedback loop [8]. By inducing an additional feedback source, patients could be stimulated and trained to learn again how to produce speech that is loud and intelligible.

To analyze how participants react to instant auditory feedback depending on their loudness, a loudness feedback test was designed. In the future, this test could either function to extract parameters such as the reaction time, how long the correction lasts, or how many corrections are necessary. Additionally, it could also function as training for a constant loudness level. The basic principle is to define a loudness threshold and whenever the smoothed loudness over time falls below it, noise is activated depending on the actual loudness value to provide feedback and the reflex to use Lombard speech.

Methods

The speech analysis process is based on a state machine that enables the pre-, running, and post-processing phases to enable new settings for each test, real-time processing, and analysis of the extracted features. In general, one loop of processing can be redone several times based on the number of speech tasks one has defined via a configuration beforehand. In addition, the advanced user chooses instructions, descriptions for each trial, and test categories enabling different setting groups designed for different voice tasks. For the whole test procedure, global parameters such as countdowns before speech tasks, report settings, and storage settings can be defined. In addition, survey options can be enabled to get information about how the participants rate their speech. To use the analysis afterward and to interpret the results, a report is done including options like plotting, showing patient data, and additional feature and task descriptions.

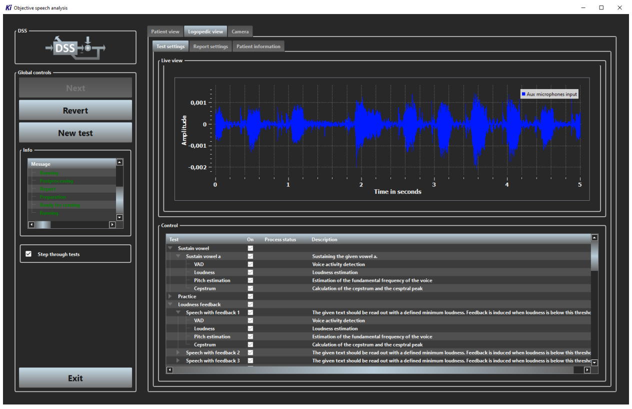
The graphical user interface (GUI) is subdivided into a patient and a logopedic view. In both views control buttons for starting, reversing, and restarting the whole test, as well as exiting are visible. In the patient view, shown in Figure 1, the upper part of the screen shows automatic instruction, countdown, and speech tasks, enabling self-explaining usability. The lower part enables real-time feedback as in a voice activity display, a timer that shows the time since the actual voice task started, and the progress in terms of how many trials from the whole tests are done.



Please sustain the vowel /:a/ as long as possible with moderate loudness.

Figure 1: GUI of KiRAT in patient view during the first trial and active speech.

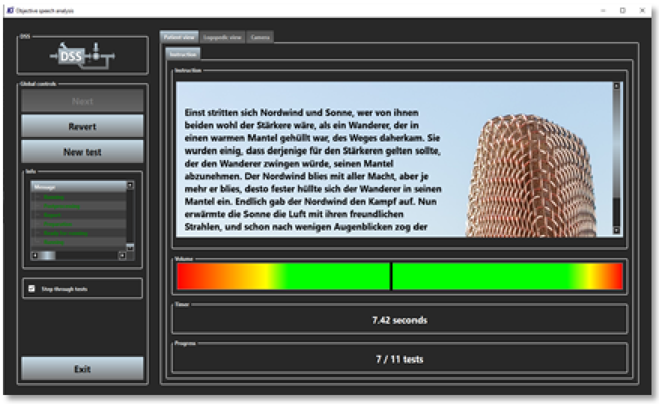
The logopedic view, shown in Figure 2, allows the user to enable and disable trials and to select the features to be extracted from the voice for each trial. During recording, the speech signal is displayed in real-time. For the report, an additional tab allows choices such as location, format, and additional report features such as patient information, feature, and test descriptions.



Test	On	Process status	Description
Sustain vowel	☑		
Sustain vowel a	☑		Sustaining the given vowel a.
VAD	☑		Voice activity detection
Loudness	☑		Loudness estimation
Pitch estimation	☑		Estimation of the fundamental frequency of the voice
Cepstrum	☑		Calculation of the cepstrum and the cesptral peak
Practice	☑		
Loudness feedback	☑		
Speech with feedback 1	☑		The given text should be read out with a defined minimum loudness. Feedback is induced when loudness is below this thresho
VAD	☑		Voice activity detection
Loudness	☑		Loudness estimation
Pitch estimation	☑		Estimation of the fundamental frequency of the voice
Cepstrum	☑		Calculation of the cepstrum and the cesptral peak
Speech with feedback 2	☑		The given text should be read out with a defined minimum loudness. Feedback is induced when loudness is below this thresho
Speech with feedback 3	☑		The given text should be read out with a defined minimum loudness. Feedback is induced when loudness is below this thresho

Figure 2: GUI of KiRAT in logopedic view during a speech trial and active speech.

Within this framework the proposed adaptive loudness feedback method was implemented. The feedback, which is given when the detected loudness falls below the target loudness, can be white noise or a file-based and therefore freely selectable noise type. These settings and others, such as an input loudness smoothing parameter or target loudness threshold, can be selected from a configuration file. Each trial starts with a countdown for the participants to prepare and estimate the background noise. Therefore, the first step in estimating the noise floor is to calculate the root mean squared value of the microphone input signal. During speech activity, if the loudness threshold is not reached, the loudness feedback level is calculated as the difference between the actual smoothed loudness value and the threshold, and an additional tuning value for personal adjustment of feedback strength. Smoothing parameters, also defined in the configuration, are used to slowly ramp up during the start of feedback. The up and down ramps are defined independently of each other to ensure that the feedback starts slowly enough not to unsettle the participants, but quickly enough to let them know that the target loudness has been reached. In addition to the desired target loudness, a maximum loudness threshold is defined to ensure that no excessively loud audio signals are output. The user can choose to receive visual feedback in addition to auditory. If this is chosen, an upper and lower threshold must be defined for the displayed area, as shown in Figure 3. The green area is then defined by the selected thresholds for auditory feedback and maximum loudness. If the noise type is to be different from white noise, the configuration allows the audio file to be added.



Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der

Figure 3: GUI during the adaptive loudness test with visual feedback using a German phonetically balanced text.

Pilot Study

A small pilot study was conducted with four participants to test the functionality and basic principle of the test. Table 1 shows the demographics of the participants who reported no hearing or speech problems.

Table 1: Demographic data of participants in the pilot study.

Participant	Gender	Age
1	Female	29
2	Female	24
3	Male	27
4	Male	30

The study was carried out using the German equivalent of the rainbow passage. The phonetically balanced text ”Nordwind und Sonne” was used in all three sections of the study. The first section was designed to provide a baseline so that the standard deviation (SD) of loudness and pitch could be calculated over five trials to capture normal variation. The second section was designed to compare the effects of visual and audiovisual feedback. For this purpose

- one trial with no feedback,
- one trial with visual feedback only,
- six trials with visual and auditory feedback,
- and two blind trials

were conducted. The six audiovisual feedback trials consisted of three trials with white noise of increasing feedback intensity and three trials with rain noise of equally increasing intensity (small, medium, and big). The blind trials contained no auditory feedback, although the participants were unaware of this. In the third section, only the effect of the auditory feedback was investigated. Therefore, the visual feedback was neglected while the experimental procedure was carried out as in the second section. The introduction of visual feedback as the first component of feedback was chosen to clarify the general principle of the test and thus ensure that the concept was understood.

Results and Discussion

In the pilot study, loudness and pitch were analyzed to investigate patient responses to the Lombard effect. The baseline for the pilot study, i.e. the first section, did not result in highly varying SD for the different participants. On average, Table 2 shows a SD of 1.5 dB for loudness and 2.8 Hz for pitch.

Table 2: Mean and standard deviation (SD) of loudness and pitch in section one of the pilot study.

Partici - pant	Mean loudness [dB]	SD loudness [dB]	Mean pitch [Hz]	SD pitch [Hz]
1	72.8	1.3	185.9	3.8
2	63.2	1.7	199.5	2.3
3	66.3	2.0	87.7	2.3
4	70.6	0.8	150.4	2.6

In the second section, the effect of audiovisual feedback was analyzed. Figure 4 and 5 show the effects of all visual and audiovisual modes as well as the two blind trials for loudness and pitch. Therefore, the results of all participants and the mean across all participants (blue) are shown. A comparison of the standard deviation shows that the visual feedback leads to an increase in loudness of 2.0 dB on average and therefore above the normal variation and that all audiovisual modes lead to a further increase (as seen by the dotted blue line). At the lowest intensity, only a minimal increase above the visual feedback is achieved for both noise modes, while at the medium and high intensities respectively, an average of 4.4 dB is achieved for both the white noise and the rain noise. This increase continues through the blind trials, although the second trial shows a smaller increase in loudness than the first blind trial. A very similar picture emerges for pitch, with an average effect of 5.3 Hz for the visual feedback. The maximum average increases in fundamental frequency are again observed for the medium intensity of the white noise and the high intensity of the rain noise at 10.1 and 11.2 Hz respectively.

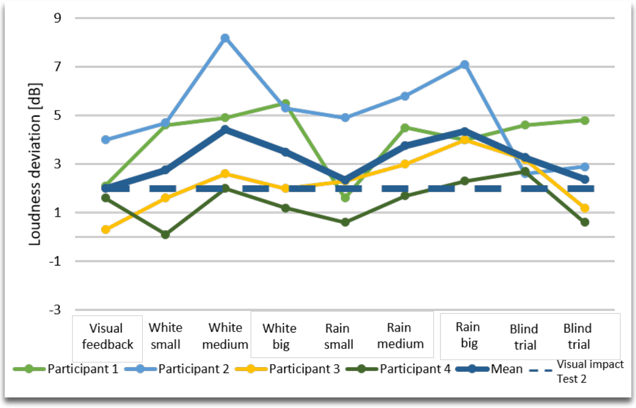


Figure 4: Mean deviation of loudness in the second section in contrast to the mean resulting from the baseline test depending on the feedback modes.

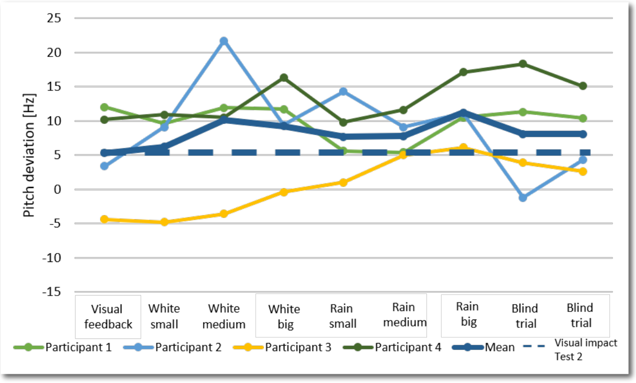


Figure 5: Mean deviation of pitch in the second section in contrast to the mean resulting from the baseline test depending on the feedback modes.

In the third section, in contrast to section 2, only the auditory feedback is analyzed. Figure 6 and 7 show the auditory modes as well as the visual effect from the second trial (black dotted line) in comparison. For both loudness and pitch, there is less of an effect. For loudness, the maximum effect is reached at the highest intensities of both noise modes, while for pitch, the maximum effect is reached at the medium modes. The mean maximum loudness feedback is consistent with the deviation due to visual feedback from the second trial. For pitch, the maximum average effect is 5.9 Hz, slightly higher than the effect of the visual feedback.

The significantly smaller effect in the third section can be attributed to a few limitations: Firstly, the test group was relatively small, and secondly, all participants had no voice impairment, i.e. they speak at an appropriate loudness and are aware of this. To investigate an effect, the target loudness had to be set relatively and unnaturally high during the test. In contrast, dispho-

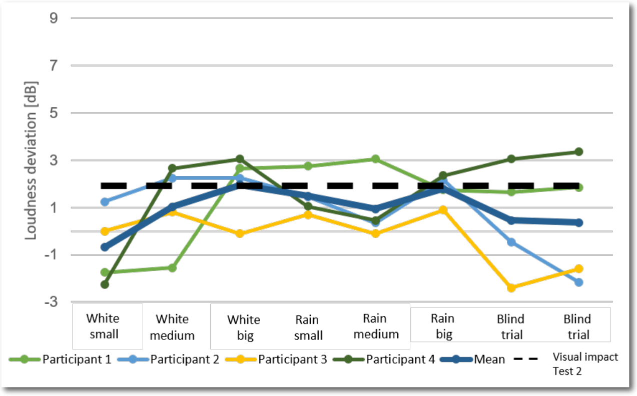


Figure 6: Mean deviation of loudness in the third section in contrast to the mean resulting from the baseline test depending on the feedback modes.

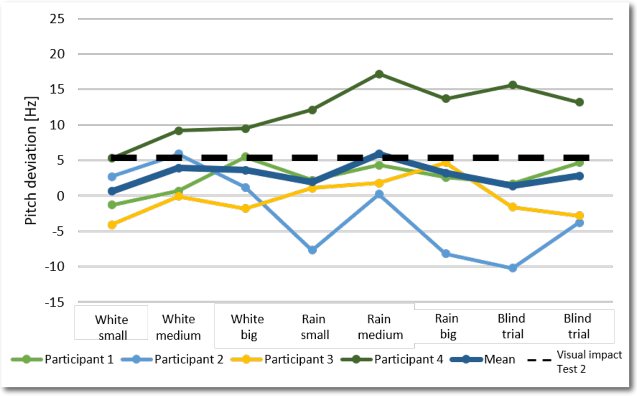


Figure 7: Mean deviation of pitch in the third section in contrast to the mean resulting from the baseline test depending on the feedback modes.

nic participants are aware of their tendency not to speak out loudly enough, which could lead to different behaviors and results. In addition, a headband microphone should have been used. In the conducted study, participants were asked to maintain their position, which seems difficult, especially when the reflex to increase loudness is engaged. This would not influence the results of the pitch estimation but perhaps the loudness results and therefore should be changed in the future. Another important aspect is that a pre-test for the calculation of the individual target loudness should be introduced. Particularly in the third section of the study, most participants had little auditory feedback. This is because the participants had learned to adjust their average loudness in the second section. Also, unlike visual feedback, auditory feedback should lead to an increase in loudness above the lower target threshold. On the other hand, visual feedback depicted a target range so all healthy participants automatically tried to adjust their loudness to reach the middle of the range of the visualization bar. This resulted in different target loudnesses and therefore different results. In addition, fatigue effects as well as the explanation of the experiment and the order of the feedback modes could be adjusted in the future to further analyze whether they could have influenced the results. Also, the after-effects should be analyzed in more detail to predict whether this training can help patients re-adjust their speaking loudness in general and maintain it in everyday life.

Summary and Outlook

In this paper, we presented an objective real-time speech analysis tool with the ability to use adaptive testing and training methods. A loudness feedback test was created using the Lombard effect to induce increased loudness and speech intelligibility. This test or training allowed the intensity, feedback noise type, and target threshold to be personalized for each participant and speech task. In a small pilot study, audiovisual feedback was shown to increase loudness and pitch, depending on the strength of the feedback. The use of auditory feedback alone resulted in a smaller effect on speaker loudness. However, this may be due to the test design and the resulting limitations. With this knowledge, and assuming that people with Parkinson's disease behave in the same way as they do in terms of the Lombard effect, a more detailed test scenario should be designed and tested with a larger group of participants and dysphonic speakers.

References

[1] Podesva, R.J., Callier, P. Voice Quality and Identity. *Annual Review of Applied Linguistics.* 2015; 35:173-194. doi:10.1017/S0267190514000270

[2] Desjardins, M., Halstead, L., Cooke, M., Shaw Bonilha, H. A Systematic Review of Voice Therapy: What "Effectiveness" Really Implies. *Journal of Voice.* 2017; 31(3). doi:10.1016/j.jvoice.2016.10.002

[3] Dhamyal, H., Singh, R. Objective Measurements of Voice Quality. *ArXiv.* 2024; URL: `https://arxiv.org/abs/2410.09578`.

[4] Hernandez, A., Yeo, E.J., Kim, S., Chung, M. Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics. *Proc. Interspeech 2020*, 2897-2901. doi:10.21437/Interspeech.2020-2354

[5] Galaz, Z., et al. Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease. *Computer Methods and Programs in Biomedicine.* 2016; 127:301-317. doi:10.1016/j.cmpb.2015.12.011.

[6] Al-Fwaress, F.S.D. The Lombard Effect on Speech Clarity in Patients with Parkinson Disease. Doctoral dissertation, University of Cincinnati, OhioLINK Electronic Theses and Dissertations Center. 2008; URL: `http://rave.ohiolink.edu/etdc/view?acc_num=ucin1208661860`.

[7] Wight, S., Miller, N. Lee Silverman Voice Treatment for People with Parkinson's: Audit of Outcomes in a Routine Clinic. *Int J Lang Commun Disord.* 2015; 50(2):215-225. doi:10.1111/1460-6984.12132

[8] Kröger, B.J. Modeling Speech Processing in Case of Neurogenic Speech and Language Disorders: Neural Dysfunctions, Brain Lesions, and Speech Behavior. *Frontiers in Language Sciences.* 2023; doi:10.3389/flang.2023.1100774