DOI: 10.61782/fa.2023.0250

forum acusticum 2023

VQ-SYNTH: DEVELOPMENT AND PERCEPTUAL EVALUATION OF A SYSTEM FOR VOICE QUALITY MODIFICATION

Isabel S. Schiller[1*] **Alexander Schnapka**[1] **Christina Eggert**[2]
Peter Birkholz[2] **Simon Stone**[2]
[1] Work and Engineering Psychology, RWTH Aachen University, Germany
[2] Institute of Acoustics and Speech Communication, TU Dresden, Germany

ABSTRACT

This paper introduces *VQ-Synth*, a prototype system for the voice-quality modification, designed to increase breathiness in sustained vowels. The system is currently operating offline but shall be used for real-time auditory feedback alteration in the near future. Here, we describe *VQ-Synth*'s architecture and operating principles and evaluate its efficacy through two listening experiments. In experiment 1, we examined the impact of different *VQ-Synth* settings on listeners' breathiness and naturalness perception of an unknown speaker's normal and hoarse voice recordings. In experiment 2, an extension of experiment 1, we tested the same but based on listeners' own-voice recordings. Both experiments confirmed our hypothesis that, with stronger voice manipulation, vowels were perceived as increasingly breathy. At the same time, perceived naturalness declined. We conclude that *VQ-Synth* meets its intended purpose to increase perceived breathiness in vowels – not only based on unknown voices of varying quality but also on listeners' own voices. Further investigation into optimal parameter settings is needed, especially in regards to naturalness. With refinements and an implemented real-time mode, *VQ-Synth* may soon be used to study vocal motor control in healthy speakers and those with voice disorders.

Keywords: *voice quality resynthesis, auditory feedback alteration, vocoder, auditory-perceptual evaluation*

[*] ***Corresponding author***: *isabel.schiller@psych.rwth-aachen.de*

1. INTRODUCTION

Functional voice disorders (FVDs) are characterized by a dysfunction of the voice system resulting in impaired voice quality (i.e., hoarse voice), without any identifiable organic or neurological causes [1], [2]. The etiology of FVDs is still not fully understood and current voice therapy approaches focus primarily on symptom management [2]–[4]. In the study of FVDs and the search for treatment methods, recent studies have highlighted the potential of auditory feedback alteration [4], [5]. Such experiments involve speakers phonating into a microphone while receiving acoustically perturbated real-time feedback of their own voice through headphones. This feedback triggers automatic phonatory (vocal) responses that can provide insights into the underlying neural mechanisms of vocal control.

In persons without voice disorders, auditory feedback alteration usually generates compensatory responses in the opposite direction of the perturbation (e.g., upward pitch shifts prompt participants to lower their voice). Intriguingly, patients with FVDs have also shown following responses and overriding compensatory responses [4], [5]. More research is needed to analyze, classify, and detect the reasons behind these different response patterns. Previous studies on patients with FVD have only altered the parameter of fundamental frequency f_0, thus pitch, to study how this affects phonation. But FVDs are characterized by impaired voice quality (e.g., breathiness, roughness, and tenseness) which is independent of pitch [8]. Future experiments would benefit from a vocoder that specifically alters voice quality, for example, to enhance the patient's awareness of particular voice features they need to adjust for a more physiological phonatory behavior.

To our knowledge, Perrotin and Mcloughlin [9], [10] are the only researchers who developed a vocoder for real-time voice quality manipulation. *GFM-Voc* modifies voice

quality in terms of perceived vocal force and tenseness, using iterative adaptive inverse filtering (IAIF) for source-filter separation. IAIF may provide a more accurate estimation of the glottal volume flow compared to basic inverse filtering, because it repeatedly applies linear prediction and inverse filtering [9], [11]. Since glottal and vocal tract configurations can be extracted independently from one another, modifications become more precise. *GFM-Voc* alters voice quality by manipulating the vocalic formants and the spectral shape of the glottal flow. However, to our knowledge, *GFM-Voc* has not undergone a thorough auditory-perceptual evaluation to ensure that listeners actually perceive the voice quality shifts as intended and natural sounding. It is also not clear whether *GFM-Voc* can deal with hoarse input voices.

Inspired by Perrotin and Mcloughlin's [9] work, we have developed a similar voice-quality resynthesis system, namely *VQ-Synth*. The system has been designed to increase vocal breathiness (i.e., the auditory impression arising when a high amount of unused air escapes from the vocal folds during phonation, resulting in a whispery sound) in sustained vowel recordings. Currently, *VQ-Synth* is restricted to offline voice-quality manipulation. However, the underlying algorithms only result in short delays, and a real-time mode is currently in development.

The goal of this paper is to describe *VQ-Synth*'s architecture and operating principles, and present the results from two listening experiments in which we evaluated the system's performance on vowels recorded in a healthy (modal) and hoarse voice quality by an unfamiliar speaker (Exp. 1), and on vowels vocalized by the listeners themselves (Exp. 2).

2. VQ-SYNTH ARCHITECTURE

VQ-Synth is a voice modification system implemented in MATLAB that uses IAIF for source-filter decomposition. To increase the degree of perceived breathiness, two parameters are configurated: spectral decay per octave (SDO) and signal-to-noise ratio (SNR), both measured in dB. The lower the SDO values are set, the steeper the spectral decay of the manipulated glottal excitation signal. As we are presently limited to increasing the spectral slope, resulting in a faster spectral decay towards higher frequencies, the SDO value is interpreted as a requested difference in the spectral slope after the manipulation. If, for example, an SDO value of –6 dB is specified, and the original input signal features a spectral slope of –4 dB/oct, the output speech signal will have a spectral slope of –10 dB/oct. SNR refers to the level of simulated aspiration noise: The lower the SNR, the higher the level of additive, temporally shaped white noise and thus the perceived aspiration. The system currently processes input speech files offline and as a single chunk. Modification of running speech is not yet possible.

The current processing pipeline is as follows: The original speech signal is divided into a source and a filter signal with the IAIF implementation from COVAREP 1.4.2. (https://github.com/covarep/covarep/releases/tag/v1.4.2). The parameters for the algorithm are currently fixed to the COVAREP authors' recommended values: the order of the LPC analysis of the vocal tract is the sampling rate of the input signal in kHz plus 3, the LPC order for the glottal source is 3, the leaky integration coefficient is 0.99, and the high-pass filter is applied three times. Next, the spectral slope of the glottal flow signal is manipulated according to the requested SDO value. First, the current spectral slope is determined by means of peak detection in the magnitude spectrum of the glottal signal and a subsequent linear fit through the detected peaks. Since only a few (in fact, theoretically only two) true peaks are needed to calculate the spectral slope, the peak detection algorithm is parametrized to yield high true positives and low false positives at the cost of a few negligible false negatives (see Table 1).

Table 1. Parameters for the peak picking algorithm (MATLAB function *findpeaks*).

Analyzed frequency range	$(0.9 \cdot f_0, 2000\ \text{Hz})$
Minimum peak prominence	30
Minimum peak distance	$0.9 \cdot f_0$

Next, the spectral slope of the original flow signal is decreased (stronger amplitude decay per octave) by low-pass filtering of the signal with a frequency-sampled FIR filter of order 512. The filter is designed on-the-fly based on the specified desired delta of the spectral slope using an arbitrary response magnitude filter specification object (*fdesign.arbmag* in MATLAB, single-band design, 100 frequency taps between f_0 and half the sampling frequency). After filtering the glottal flow signal, the spectral slope is determined again (as described above). If the desired spectral slope is not successfully set (which may happen due to inaccurate peak detection), an error is raised.

After the manipulation of the spectral slope, aspiration noise is added. The noise is sampled from a Gaussian process with zero mean, a standard deviation of $\frac{1}{\sqrt{12}}$ and limited to the interval [–1, 1]. The resulting broadband noise signal is then filtered using an FIR low-pass filter to enforce a typical spectral slope of –9.4 dB/kHz [12]. The noise signal is also

gated (i.e., multiplied) by the glottal flow signal so that the added noise amplitude is modulated by the flow amplitude [13]. Finally, the modulated noise signal is added to the flow signal according to the specified SNR. The thusly modified glottal flow signal is then filtered using the original filter coefficients identified in the initial source-filter-separation step to create the final utterance containing the manipulated breathiness. Figure 1 depicts the processing pipeline.

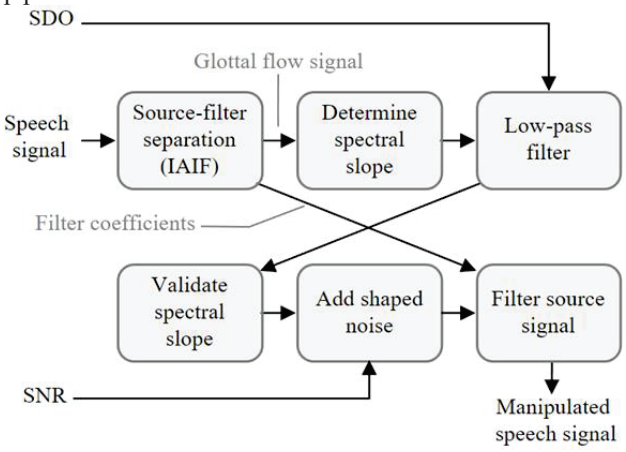

Figure 1. Overview of the processing pipeline.

3. EXPERIMENT 1

3.1 Method

For an auditory-perceptual evaluation of *VQ-Synth*, we designed and conducted a listening test that took place in a soundproof booth at the Institute of Psychology, RWTH Aachen University. The goal was to determine perceived breathiness and naturalness in modal and hoarse voice samples, manipulated with a pre-defined set of resynthesis configurations. We also included hoarse voice samples to evaluate the system's ability to handle dysphonic voices, which will be relevant in future voice therapy applications.

3.1.1 Participants

We tested 31 participants (21 female, 10 male) between 18 and 32 years old (M = 23, SD = 4). Inclusion criteria were self-reported normal hearing, normal or corrected-to-normal vision, and an advanced level of German (B2 level). Most participants were psychology students who received study credits as compensation.

3.1.2 Material

A computer-based listening task was designed in PsychoPy (Version 2022.1.1; [14]). The task was to listen to samples of the sustained vowel /aː/, and rate perceived breathiness and naturalness. Visual analogue scales were used for this purpose (Figure 2). The breathiness scale endpoints were "not at all breathy" and "maximally breathy", and the naturalness scale endpoints were "synthetic" and "natural".

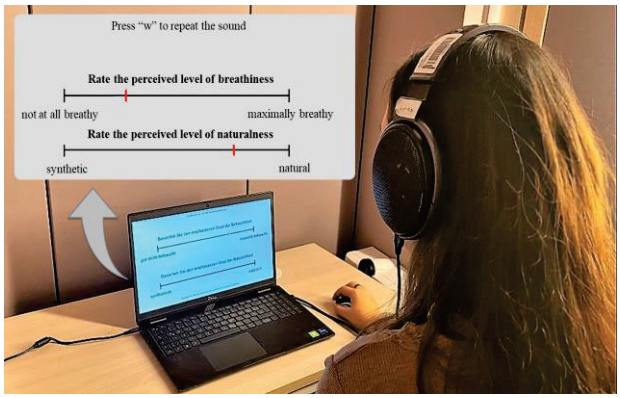

Figure 2. Visual analogue scale for breathiness and naturalness ratings of the vowel samples.

Recordings were made by a 34-year-old female voice expert, who sustained /aː/ in her modal voice and while imitating a hoarse voice. The vowel /ɛː/ was also recorded for applying it in the practice block of the listening task. Recordings were digitized at a sampling frequency of 44.1 kHz and a 16-bit resolution. Using Praat (Version 6.1.47; [15]), we cut the vowel productions to a length of 3 s and included a fade-in and fade-out of 150 ms.

In preparation for the listening task, voice recordings were resynthesized by means of the following *VQ-Synth* settings: 5 SDO levels (–6, –7, –8, –9, –10 dB per octave)[1] x 5 SNR levels (+30, +25, +20, +15, +10 dB). Thus, in total, the listening tasks consisted of 52 test trials, including the two unmanipulated control conditions (modal, hoarse) and the resynthesized samples. Figure 3 depicts spectrograms of the controls and the most extreme resynthesis conditions.

[1] The –6 dB SDO with +30 dB SNR condition sounded almost identical to the original recordings. This was because a simple passthrough of source-filter separation and recombination without any manipulation resulted in a positive spectral slope delta of about +6 dB, caused by the selected separation method (IAIF).

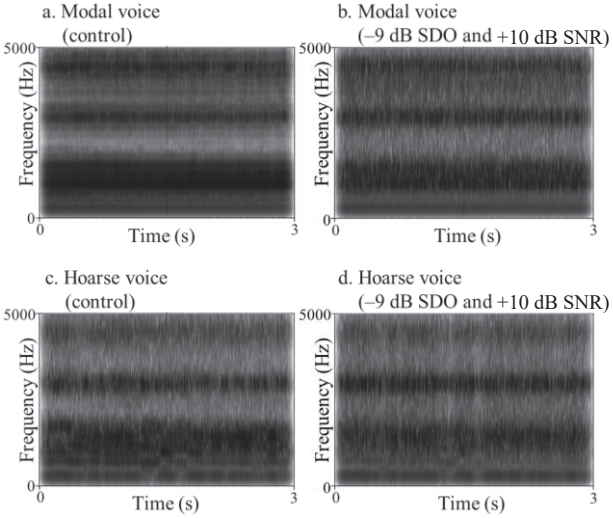

Figure 3. Spectrograms of the sustained vowel /a:/ in the modal and hoarse control condition and the most extreme SDO and SNR manipulations.

3.1.3 Procedure

First, participants were given task instructions that included definitions of breathiness and naturalness. Breathiness was described as an auditory impression that results when there is an excessive escape of air during phonation, which decreases the voice's tonicity. Naturalness was defined as how closely the sustained vowel resembled a human-produced sound compared to a synthetic computer voice.

The experiment lasted about 30 minutes and participants were seated in front of a computer (Dell Latitude 3590) equipped with headphones (HD 650, Sennheiser electronic GmbH) and a mouse to indicate their responses. The calibrated presentation level was 60 dB (A). The experiment started with a practice block to familiarize participants with the stimuli and their task, followed by a test block containing randomized unmanipulated and resynthesized samples of the vowel /a:/ in modal and hoarse voice. Each trial started with automatic playback of the sample, which could be repeated as many times as needed before the participant registered their response.

To analyze the data, we coded breathiness ratings using a numerical scale ranging from 0 ("not at all breathy") to 100 ("maximally breathy"), and naturalness ratings using the same scale (0 = "synthetic"; 100 = "natural".). Data were analyzed with R Studio (Version 2022.02.3), using repeated measures correlations with the package rmCorr [16]. Specifically, we calculated the correlation between SDO and perceived breathiness and naturalness, and SNR and perceived breathiness and naturalness. We then calculated pairwise comparisons to identify voice samples in which the subjectively perceived breathiness was significantly higher than in the controls, without a significant reduction in perceived naturalness.

3.2 Results

Figure 4 presents the descriptive results for the breathiness and naturalness ratings as a function of SDO variations (4a.) and SNR variations (4b.). We found significant negative correlations between SDO and perceived breathiness, $r_{rm}(123) = -.75$, $p < .001$, and between SNR and perceived breathiness, $r_{rm}(123) = -.89$, $p < .001$. In other words, the steeper the spectral decay per octave and the poorer the signal-to-noise ratio, the breathier the talker's voice was perceived. Regarding naturalness, the opposite was found. The steeper the SDO or the lower the SNR, the less natural the voice was perceived, $r_{rm}(123) = .61$, $p < .001$, and $r_{rm}(123) = .66$, $p < .001$, respectively.

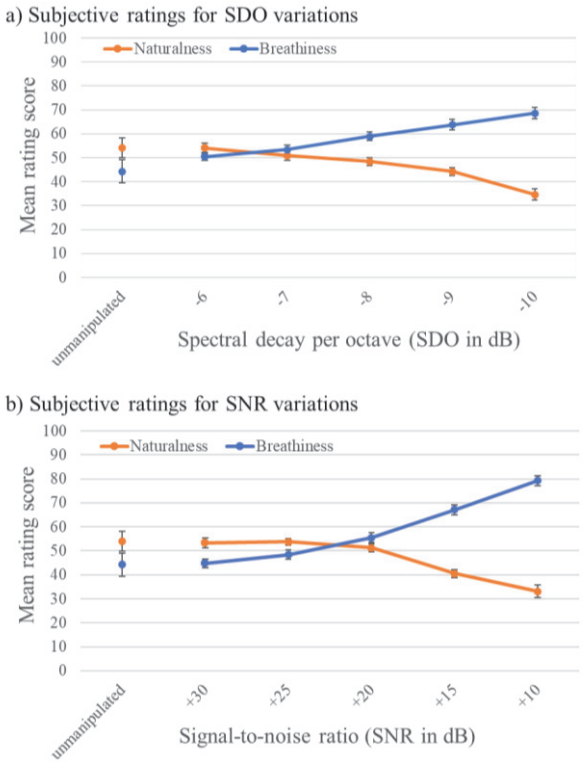

Figure 4. Mean breathiness and naturalness ratings in Exp. 1 as a function of SDO (a) and SNR (b). Error bars represent standard errors.

In practice, *VQ-Synth* could be successfully applied to resynthesize not only the modal but also the hoarse input voice. Nevertheless, identifying *VQ-Synth* parameter settings that significantly increased perceived breathiness without significantly reducing naturalness was more successful with respect to the modal input voice. Results from directed pairwise comparisons are shown in Table 2.

3.3 Discussion

In Exp. 1, we assessed the effect of different *VQ-Synth* resynthesis configurations on listeners' perception of breathiness and naturalness, based on recordings of the vowel /a:/, produced by a female voice expert in a modal and imitated hoarse voice. Results indicate that the system achieves its pre-determined objective (i.e., to increase perceived breathiness in sustained vowels) through the applied parameter settings. Importantly, we were able to show that assumptions underlying the system, i.e., the source-filter-separation, still hold for a hoarse input voice. Still, regarding the modal input voice, the system offers greater flexibility to increase breathiness – at least based on the input voices. It appears that stronger modifications are necessary to increase breathiness in an already hoarse voice. Interestingly, relatively low ratings of naturalness were observed, even for the unmanipulated control samples. Especially for these latter samples, we had expected high naturalness rating scores, but instead, we found scores as low as 50-60. This finding might relate to how naturalness was defined prior to the listening task and/or the nature of the samples. That is, in real life, connected speech does not contain sustained vowels of several seconds. Another possible explanation for the relatively low ratings of naturalness in this study is that participants might have avoided extreme ratings on the visual analogue scale. This phenomenon is known as central tendency bias.

Table 2. Parameter settings resulting in significantly increased breathiness without decreasing naturalness.

Input voice quality	SDO (in dB)	SNR (in dB)	Comparison with breathiness ratings of the control conditions, using *t*-tests with Bonferroni-Holm corrected *p*-values and Cohen's effect size (*d*)
modal	–9	+25	$t(30) = 4.13$, $p = .001$, $d = 0.83$
	–8	+20	$t(30) = 4.50$, $p < .001$, $d = 0.85$
	–6	+15	$t(30) = 5.53$, $p < .001$, $d = 0.87$
	–7	+15	$t(30) = 7.01$, $p < .001$, $d = 1.64$
	–9	+20	$t(30) = 7.08$, $p < .001$, $d = 1.59$
	–10	+25	$t(30) = 7.48$, $p < .001$, $d = 1.47$
	–8	+15	$t(30) = 8.25$, $p < .001$, $d = 2.05$
	–6	+10	$t(30) = 8.35$, $p < .001$, $d = 2.03$
	–10	+20	$t(30) = 9.37$, $p < .001$, $d = 1.91$
hoarse	–8	+10	$t(30) = 3.34$, $p = .027$, $d = 3.76$

4. EXPERIMENT 2

4.1 Method

As a next step toward real-time auditory feedback alteration, we conducted a second listening experiment, approved by the ethics committee of the Faculty of Arts and Humanities (ref. 2022_14_FB7_RWTH Aachen). The goal was to assess *VQ-Synth*'s performance on each participant's own voice vowel productions. Based on manipulated and unmanipulated samples of /a:/, /i:/, and /u:/, participants rated breathiness and naturalness.

4.1.1 Participants

In total, we tested 77 participants, mostly consisting of psychology students who received study credits for their participation. Three data sets were excluded due to technical issues ($n = 2$) or erroneous log files ($n = 1$). The final data set includes 74 participants (52 female, 22 male), aged 18-37 years old ($M = 22$, $SD = 3$). None of the participants had previously participated in Exp. 1. Inclusion criteria were self-reported normal hearing, normal or corrected-to-normal vision, and German skills on B2 level.

4.1.2 Material

The listening experiment was programmed in PsychoPy (Version 2022.1.1; [14]). As in Exp. 1, the task consisted of listening to and rating perceived breathiness and naturalness in vowel samples (set-up depicted in Figure 1). This time, however, ratings were based on each participant's own voice. Thus, prior to the experiment, participants recorded four vowels (/a:/, /i:/, /u:/, and /ɛ:/ for practice purposes), which were then manipulated according to the same *VQ-Synth* configurations as in Exp. 1. These samples, along with the unmanipulated samples, were then presented to them in the listening experiment (processed in the same manner as in Exp. 1). As more vowels were included, the task consisted of 78 (in addition to the practice block on /ɛ:/): three unmanipulated samples and 75 manipulated samples, based on the combinations of 3 vowels (/a:/, /i:/, /u:/) x 5 SDO levels (–6, –7, –8, –9, –10 dB per octave) x 5 SNR levels (+30, +25, +20, +15, +10 dB SNR).

4.1.3 Procedure

The experiment lasted about 45 min. The instructions and definitions provided to the participants were similar to Exp 1., but contained additional information regarding the naturalness rating, and vowel recordings. Since naturalness was rated surprisingly low in Exp. 1, we now explicitly instructed the participants to rate this dimension solely based on whether the voice resembled more to a human (themselves) or a computer, and to neglect the fact that sustained vowels usually do not occur in everyday speech.
First, participants were asked to sustain each vowel for 5 s in front of a microphone (Neumann KM 184), aiming to maintain stable amplitude and pitch at about 65 dB (A) SPL, measured with a sound level meter (Nor116, Norsonic) at a distance of 15 cm. The recordings were repeated until the desired stability was achieved. The vowel samples were then processed using the *VQ-Synth* script in Matlab R2018a (version 9.4), which took about three min.
The subsequent listening task was similar to Exp. 1. However, we included three vowels, thus, participants were randomly presented with three test blocks, each containing all unmanipulated and manipulated samples of the respective vowel (/a:/, /i:/, /u:/) in random order.
Rating scores were coded from 0 to 100 as in Exp. 1. Again, data were analyzed concerning how different *VQ-Synth* parameter settings would relate to participants' breathiness and naturalness ratings, for which we calculated repeated measure correlations [16] in R Studio (Version 2022.02.3). To investigate whether *VQ-Synth*'s performance varied with vowels, we calculated repeated measures ANOVA.

4.2 Results

Figure 5 shows the rating results for the unmanipulated samples and those manipulated by varying SDO (5a) and SNR (5b). This time, unmanipulated samples were rated with a naturalness of 70-80 thus, notably higher.

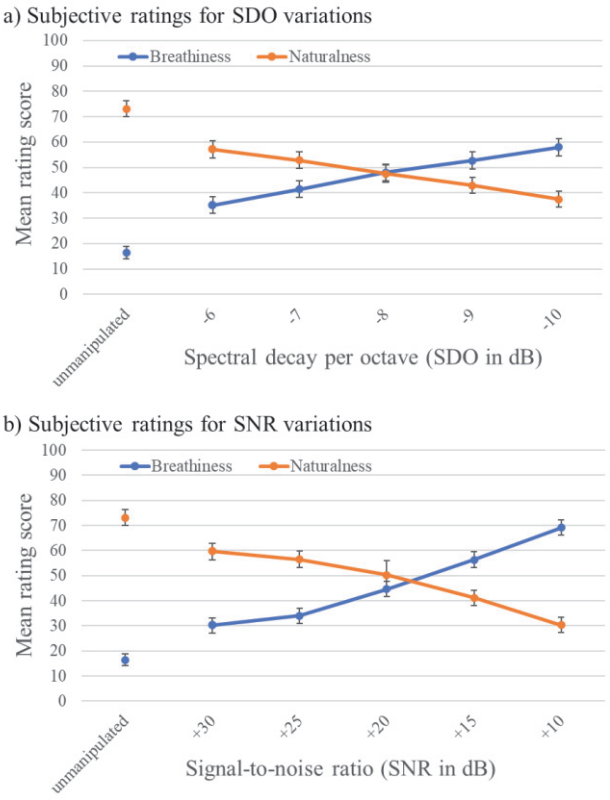

Figure 5. Mean breathiness and naturalness ratings in Exp. 2 as a function of SDO (a) and SNR (b). Error bars represent standard errors.

As in Exp 1., we found significant negative correlations between perceived breathiness and SDO, $r_{rm}(295) = –0.77$, $p < .001$, as well as SNR, $r_{rm}(295) = –.87$, $p < .001$, indicating that steeper SDO and lower SNR resulted in a breathier sound. Again, naturalness ratings significantly dropped with steeper SDO and lower SNR, $r_{rm}(295) = 0.73$, $p < .001$, and $r_{rm}(295) = 0.74$, $p < .001$, respectively.
Descriptive results of the breathiness and naturalness ratings for each vowel are presented in Table 3. Repeated measures ANOVA revealed that the main effect of vowel was significant with respect to both breathiness ratings, $F(2, 146) = 9.506$, $p < .001$, $\eta_p^2 = .12$, and naturalness ratings, $F(2, 146) = 6.86$, $p = .001$, $\eta_p^2 = .09$, indicating that

perception varied significantly across /a:/, /i:/, and /u:/. Post-hoc analysis using appropriate corrections for multiple comparisons showed that /a:/ was perceived significantly more breathy than /u:/ ($p < .001$) and /i:/ ($p = .007$), while the ratings for /u:/ and /i:/ did not significantly differ from each other ($p = .50$). Moreover, /a:/ and /u:/ were perceived as significantly more natural than /i:/ ($p = .01$, and $p = .003$, respectively), but there was no significant difference in naturalness ratings between /a:/ and /u:/ ($p = .9$).

Table 3. Breathiness and naturalness ratings of the manipulated samples, according to each vowel.

Vowel	Breathiness Mean (*SE*)	Naturalness Mean (*SE*)
/a:/	51.50 (1.63)	49.17 (1.59)
/i:/	45.86 (1.80)	43.57 (1.56)
/u:/	43.80 (1.96)	49.99 (2.00)

4.3 Discussion

In Exp. 2, we investigated the influence of different *VQ-Synth* resynthesis configurations on listeners' breathiness and naturalness perception of their own voice. We found that stronger manipulations correlated with increased perceived breathiness, but also with decreased naturalness. This finding is consistent with the results of Exp. 1, indicating that while the system meets its intended purpose, further technical refinements are necessary to ensure that listeners perceive the modifications as more natural. Importantly, we were able to validate the functionality of *VQ-Synth* for various input voices, male and female, and show that the system is capable of adjusting own voice perception as desired. This is not immediately apparent because multiple factors can affect self-perception such as voice attributes, personal traits, and emotional tone [17].

Furthermore, we were able to demonstrate that *VQ-Synth* can successfully be applied to different vowels, not only /a:/. Nevertheless, manipulated samples of the vowel /a:/ were perceived as most breathy and natural, compared to /i:/, and /u:/, suggesting that the system performs best for this vowel. This could be because the vowel /a:/ is characterized by a largely open vocal tract, allowing for the greatest independence between source and filter. In the case of /i:/ and /u:/, the acoustic load on the source and therefore the source-filter interaction is greater, resulting in poorer separability and, therefore, presumably poorer manipulation performance. With respect to the idea that *VQ-Synth* should one day be applied for auditory feedback alteration in voice therapy, it is necessary to further test the systems' performance on different speech sounds.

5. GENERAL DISCUSSION

In this paper, we presented *VQ-Synth*, a prototype voice resynthesis system for increasing perceived breathiness in sustained vowels. Two auditory-perceptual studies were conducted to evaluate the system's performance on resynthesizing pre-recorded voice samples. In Exp. 1, listeners rated breathiness and naturalness in unmanipulated and manipulated samples of the vowel /a:/, produced in a modal and hoarse input voice. In Exp. 2, listeners evaluated these perceptual dimensions with respect to unmanipulated and manipulated recordings of their own voice, based on the vowels /a:/, /i:/, and /u:/. Findings suggest that the system meets its designed purpose to increase perceived breathiness. However, the stronger the manipulations, the more synthetical the voice samples were perceived, which must be addressed in future works.

Our results represent an important step toward using *VQ-Synth* in the context of auditory feedback alteration, but several limitations should be acknowledged. Interindividual differences in the concepts of breathiness and naturalness might have confounded the results, although we carefully defined these concepts to each participant before the task. Thus, there is a degree of uncertainty about the agreement between the results and participant perception. Adding more diverse perceptual rating dimensions to the task might have reduced this possible bias. Moreover, *VQ-Synth* does not yet operate in real-time. Thus, it remains to be tested whether the system performs in the same manner when a person's auditory feedback is manipulated during speech production rather than offline.

Our vision is that, in the future, *VQ-Synth* will allow researchers to gain a deeper knowledge of FVDs. Our approach is unique in that we manipulate perceived voice quality rather than pitch [4]–[7], as impaired voice quality is a key symptom of FVDs [1]. The phonatory reactions triggered by auditory feedback alteration may provide information about how auditory and kinaesthetic information is processed and integrated in FVD patients. This information could also be used to adapt the system as a tool for voice therapy to promote more physiological voice use. The next steps include refining the technology for more natural-sounding manipulated voices conducting another listening experiment to compare listeners' perception of manipulated vowel samples versus unmanipulated vowel samples of real breathy voices, and enabling real-time functionality through frame-wise processing. Other source-filter separation algorithms will be compared, and a more user-friendly interface will be implemented to improve scalability.

6. CONCLUSION

To conclude, the first prototype of the *VQ-Synth* successfully demonstrated its ability to increase pereived breathiness (as a facet of impaired voice quality) in modal and hoarse voice recordings, male and female speakers, and different vowel sounds. Importantly, the indended voice-quality modification was not only achieved when listeners evaluated unfamiliar voice samples but also when they evaluated manipulated samples of their own voice. Moving forward, we aim to modify *VQ-Synth* to be able to conduct real-time auditory feedback alterations, which could greatly contribute to understanding functional voice disorders and lead to recommendations for new voice therapy approaches.

ACKNOWLEDGMENTS

Isabel Schiller's contribution to this study was also funded by a grant from the HEAD-Genuit-Foundation (P-16/10-W) awarded to Prof. Sabine Schlittmeier. We thank the Institute of Hearing Technology and Acoustics, RWTH Aachen University, especially Lukas Aspöck, for the technical support and the calibration. We acknowledge the contribution of Jian Pan, who programmed the experiment. Thanks to Tom Jungbauer for the help in data collection, and to Mai Ly Tenberg for the help in preparing this paper.

REFERENCES

[1] M. Andrea, Ó. Dias, M. Andrea, and M. L. Figueira: "Functional voice disorders: the importance of the psychologist in clinical voice assessment," *J. Voice*, vol. 31, no. 4, pp. 507.e13–507.e22, 2017.

[2] P. Carding, M. Bos-Clark, S. Fu, P. Gillivan-Murphy, S. M. Jones, and C. Walton: "Evaluating the efficacy of voice therapy for functional, organic and neurological voice disorders," *Clin. Otolaryngol.*, vol. 42, no. 2, pp. 201–217, 2017.

[3] J. Ruotsalainen, J. Sellman, L. Lehto, and J. Verbeek: "Systematic review of the treatment of functional dysphonia and prevention of voice disorders," *Otolaryngol. Neck Surg.*, vol. 138, no. 5, pp. 557–565, 2008.

[4] C. E. Stepp, R. A. Lester-Smith, D. Abur, A. Daliri, J. Pieter Noordzij, and A. A. Lupiani: "Evidence for auditory-motor impairment in individuals with hyperfunctional voice disorders," *J. Speech Lang. Hear. Res.*, vol. 60, no. 6, pp. 1545–1550, 2017.

[5] A. Ziethe *et al.*: "Control of fundamental frequency in dysphonic patients during phonation and speech," *J. Voice*, vol. 33, no. 6, pp. 851–859. 2019.

[6] N. E. Scheerer and J. A. Jones: "Detecting our own vocal errors: an event-related study of the thresholds for perceiving and compensating for vocal pitch errors," *Neuropsychologia*, vol. 114, pp. 158–167, 2018.

[7] C. R. Larson, K. W. Altman, H. Liu, and T. C. Hain: "Interactions between auditory and somato-sensory feedback for voice F0 control," *Exp. Brain Res.*, vol. 187, no. 4, pp. 613–621, 2008.

[8] S. N. Awan: "Cepstral analysis of voice," *The SAGE encyclopedia of human communication sciences and disorders*. Sage Publications, pp. 327–330, 2019.

[9] O. Perrotin and I. Mcloughlin, "GFM-Voc: A real-time voice quality modification system," in *Proc. of* the *20th Annual Conference of the International Speech Communication Association – Interspeech 2019*, (Graz, Austria), pp. 3685–3686, 2019.

[10] O. Perrotin and I. V. Mcloughlin, "GFM-Voc: A tool for analysis and modification of the glottis signal," presented at the *12th International Conference on Voice Physiology and Biomechanics – ICVPB2020*, (Grenoble, France), 2020.

[11] P. Mokhtari and H. Ando, "Iterative Optimal Preemphasis for Improved Glottal-Flow Estimation by Iterative Adaptive Inverse Filtering," *Proc. of* the *20th Annual Conference of the International Speech Communication Association – Interspeech 2017,* (Stockholm, Sweden), pp. 1044–1048, 2017

[12] R. Hillman, E. Oesterle, and Feth, L.: "Characteristics of the glottal turbulent noise source," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 691–694, 1983.

[13] D. J. Hermes: "Synthesis of breathy vowels: some research methods," *Speech Commun.*, vol. 10, no. 5–6, pp. 497–502, 1991.

[14] J. Peirce *et al.*: "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, 2019.

[15] P. Boersma and D. Weenink: "Praat doing phonetics by computer (Version 6.1.47)," [Software], 2021. Available: http://www.praat.org/

[16] J. Z. Bakdash and L. R. Marusich: "Repeated measures correlation," *Front. Psychol.*, vol. 8, Art. no. 456, 2017.

[17] H. J. Chong, J. H. Choi, and S. S. Lee: "Does the perception of own voice affect our behavior?," *J. Voice*, in press, 2022.