

Enhancing Acoustic Voice Quality Through Real-time Auditory Feedback Modulation

Isabel S. Schiller^{a,*}, Karolin Krüger^b, Patricia Weede^b, Melf T. Sopha^a,
Gerhard Schmidt^b

^aTeaching and Research Area Work and Engineering Psychology Institute of Psychology
RWTH Aachen University Kackertstrasse 9 52072 Aachen Germany

^bDigital Signal Processing and System Theory Department of Electrical and Information
Engineering Kiel University Kaiserstrasse 2 24143 Germany

Abstract

Auditory feedback modulation (AFM) – altering how speakers hear their own voice during phonation – is a well-established method for investigating vocal motor control. Healthy speakers typically respond to AFM with compensatory adjustments in the opposite direction, such as lowering pitch in response to upward pitch shifts. Whether hoarseness-induced auditory feedback elicits comparable compensatory behaviour – specifically, vocal adjustments that enhance acoustic voice quality – remains unknown. To address this gap, the present study pursued two aims: (1) to introduce and evaluate a real-time voice resynthesis system, *VQ-Synth*, designed to induce the percept of hoarseness in otherwise healthy voices, and (2) to test whether hoarseness-induced auditory feedback leads to compensatory improvements in acoustic voice quality, measured by smoothed cepstral peak prominence (CPPS). In Study 1, participants rated recordings of their own voice processed with four different resynthesis methods. Overall, the *anti-peak-window* method, which inserted noise between the pitch-related amplitude peaks, produced the strongest percept of dysphonia. In Study 2, this method was applied in an AFM experiment to modulate voice quality in real-time. Participants sustained the vowel /a:/ 140 times. In the AFM group, auditory feedback was modulated according to a phase design (baseline, ramp, hold, after), whereas the control group received unaltered feedback throughout. CPPS in the AFM

*Corresponding author

Email address: isabel.schiller@psych.rwth-aachen.de (Isabel S. Schiller)

group increased significantly from baseline through ramp and hold and remained elevated in the after phase, while there was no CPPS increase in the control group. These results indicate that the improvement in acoustic voice quality was not a practice effect but indeed driven by hoarseness-induced AFM. Future work will explore *VQ-Synth*'s potential in connected speech and its application as a therapeutic tool for individuals with dysphonia.

Keywords: VQ-Synth, Voice Perception, Voice Quality, Auditory Feedback Modulation, Motor Learning

Introduction

When we speak, we rely on auditory and somatosensory feedback to compare our intended vocal output with how we actually perceive our voice.^{1,2} If our brain detects a mismatch, corrective motor commands are sent to the phonatory or articulatory muscles to compensate for it – a process governed by vocal motor control. A common method for investigating speakers' vocal motor control involves modulating their auditory feedback, and studying the resulting phonatory responses (see Alves et al.³, for a scoping review). In auditory feedback modulation (AFM) paradigms, participants are asked to phonate into a microphone and receive an acoustically perturbed real-time feedback of their voice through headphones. Healthy speakers typically compensate for this perturbation, modifying their voice in the opposite direction.³

The pitch-shift reflex (PSR) is the most extensively studied example of such compensatory responses, elicited by upward or downward pitch shifts in the auditory feedback.⁴⁻⁸ When a speaker's pitch is shifted upward, they typically compensate by lowering their pitch, and vice versa. Such compensatory responses may temporarily persist even after auditory feedback is restored to normal (post-modulation effect).^{5,6} Similar responses have been observed for vocal loudness, with speakers speaking more softly when receiving louder feedback of their own voice through headphones.^{8,9} In PSR research, the magnitude of pitch shifts applied often varies between 50 and 100 cents, with 100 cents corresponding to one semitone.^{4,8,10} Larger pitch shifts often increase response magnitude, but this relationship is not strictly linear. For example, Arbeiter et al.¹¹ applied shifts of 700 cents and observed responses of only about 11 cents, whereas other studies using shifts of just 50 cents reported responses ranging between 13 and 20 cents.^{8,10}

Compensatory responses to AFM can be explained within the framework of the Directions Into Velocities of Articulators (DIVA) model by Guenther et al.^{1,2,12}. The DIVA model describes speech production as involving the interaction of auditory and somatosensory feedback mechanisms with feedforward control. The feedforward system generates motor commands that control articulatory movements, while the feedback system monitors the resulting output. The feedback system itself is further divided into auditory and somatosensory components. During speech production, motor commands from the feedforward system are transmitted to the articulatory and phonatory musculature, and the resulting auditory and somatosensory feedback is compared against predicted feedback. If a mismatch is detected, corrective motor commands are sent to the relevant muscles. Such corrective adjustments manifest as compensatory responses – that is, error corrections aimed at achieving the phonatory or articulatory target. While this has repeatedly been demonstrated in pitch- or amplitude shift paradigms, the effects of perceiving a degraded voice quality via auditory feedback are still unknown.

Part of the reason might be that, in contrast to the unidimensional parameters of pitch and amplitude, voice quality is a multidimensional construct that includes a broader range of auditory features.¹³ Kreiman et al.¹⁴ describe voice quality as the auditory colouring of a person’s voice, shaped by different configurations of the larynx and vocal tract, which allows the speaker to be distinguished from others. Voice quality is primarily an auditory-perceptual phenomenon.¹³ Healthy voices are generally perceived as clear and resonant, whereas disordered voices are characterised by hoarseness (dysphonia).¹⁵ More specifically, the quality of a dysphonic voice can be described and assessed in terms of roughness, breathiness, asthenia, and strain¹⁶. In line with the predictions of the DIVA model^{1,2,12}, we propose that degrading a speaker’s voice in the auditory feedback could elicit compensatory adjustments that promote more efficient and physiological voice use, manifested as improved voice quality. If confirmed, this finding could have significant implications for voice training and therapy.

Voice disorders affect a substantial portion of the population, particularly professional voice users (e.g., teachers, singers, lawyers), with prevalence reaching 44%.¹⁷ In a study involving over 14,000 US teachers in the Miami region, Rosow et al.¹⁸ assessed the economic impact of voice disorders using a self-administered survey. The study found that voice-related absenteeism cost society an estimated \$1 billion, and the financial burden of

presenteeism (the impact of voice disorders on work productivity) amounted to approximately \$300 million. Beyond the personal and economic effects, voice disorders can also negatively influence listeners, leading to reduced cognitive performance and greater listening effort.^{19–22} One of the most common types of voice disorders, observed in approximately 30% of voice patients, is primary muscle tension dysphonia (MTD).²³ In primary MTD, the vocal folds are structurally normal, but their vibration is disrupted by excessive, imbalanced muscular activity.²⁴ Traditionally, primary MTD is treated with voice therapy,¹⁵ which includes exercises aimed at reducing tension and improving vocal function. If tailored to support these therapy goals, AFM could provide a remote option, allowing patients to practice and enhance their voice quality beyond traditional clinical settings.

However, because of its multidimensional nature, generating an authentic hoarse voice quality for real-time AFM is challenging. Previous attempts to “hoarsify” a voice – by altering jitter, shimmer, spectral decay, or adding noise – have, to the best of our knowledge, all been implemented offline.^{25–27} In the absence of a real-time solution, we developed *VQ-Synth*, a voice resynthesis system whose primary function is to induce or amplify hoarseness in speaker’s auditory feedback. Implemented in the Kiel Real-Time Application Toolkit (KiRAT; <https://kirat.de/>), *VQ-Synth* has evolved from its original offline version²⁷ to offer real-time functionality, enabling direct perturbation of voice quality with a minimal hardware processing delay of 3.5 ms and an overall system delay of approximately 16 ms. In terms of hardware, *VQ-Synth* consists of a microphone to capture the speaker’s voice, a low-latency audio interface to route the voice signal to a computer or laptop for processing, headphones to deliver the modified feedback to the participant, and a screen to present instructions (see supplementary Figure S.1). For signal processing (see supplementary Figure S.2), linear prediction (LP) is used to obtain the error signal (which is the difference of the short term predicted and original speech signal), from which pitch related amplitude peaks are detected; additional noise is then generated based on this information and added back to the error signal at various signal-to-noise ratios (SNRs). Here, the ratio is defined based on the original speech signal in contrast to the noise that has been intentionally added. Before being played back to the participant via headphones, the recombined signal can be further modified by adjusting the spectral decay per octave (SDO) to enhance the perceptual dampening of the output signal.

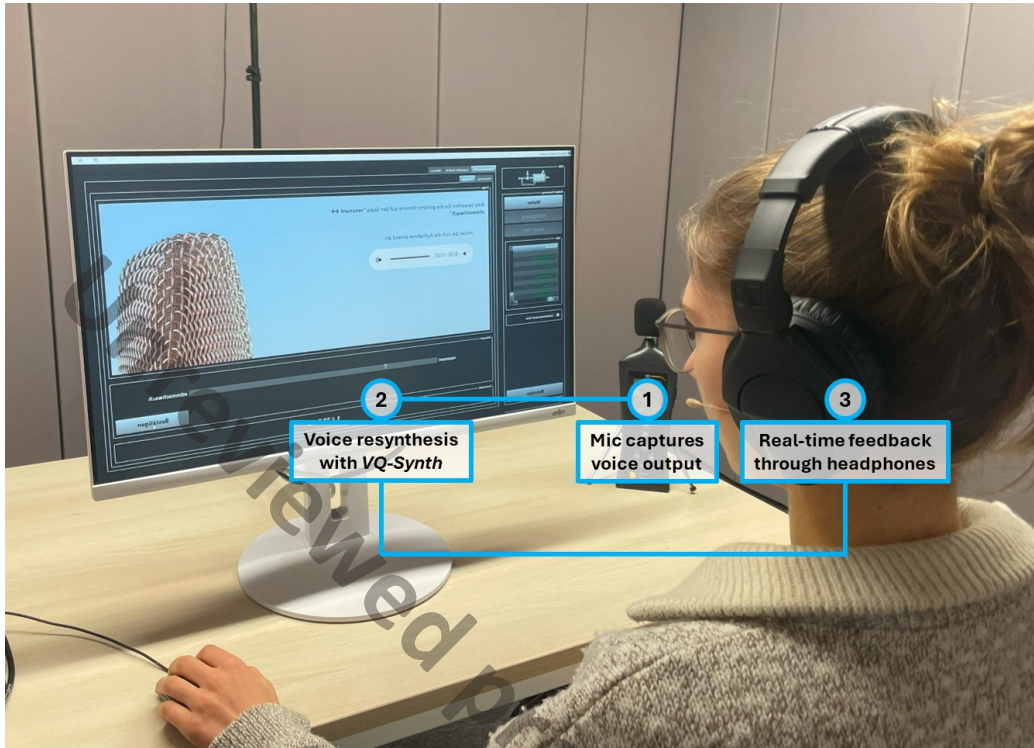


Figure S.1. Illustration of the *VQ-Synth* system in use. Boxes indicate the key processing steps: voice capture, resynthesis and manipulation to induce hoarseness, and real-time playback through headphones.

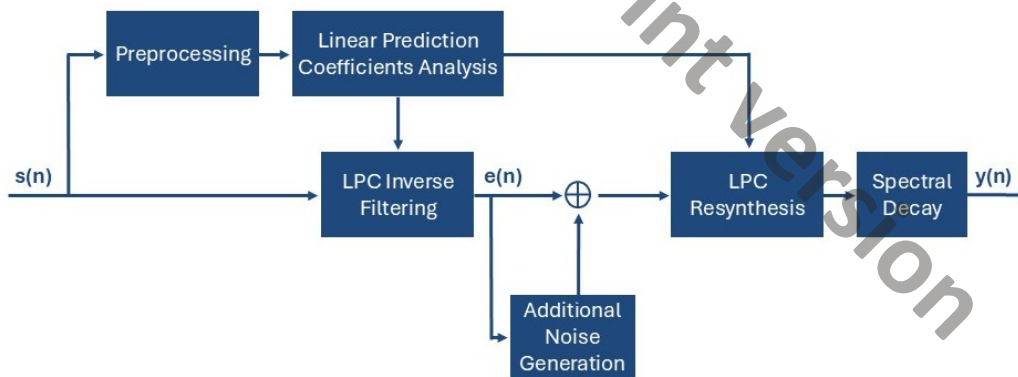


Figure S.2. Block diagram of the *VQ-Synth* system. The incoming voice signal $s(n)$ undergoes preprocessing and linear prediction analysis to extract the error signal $e(n)$. Additional noise is generated and added to $e(n)$ at various signal-to-noise ratios. The signal is then resynthesized and modified via spectral decay before being presented as $y(n)$.

When analysing phonatory responses to AFM, a speaker's voice quality can be characterised using a range of acoustic voice parameters (e.g., jitter,

shimmer, harmonics-to-noise ratio)²⁸, with cepstral peak prominence (CPP) currently regarded as especially reliable.^{29–32} Derived from the cepstrum (the inverse Fourier transform of the logarithm of a signal’s spectrum), CPP measures the strength of harmonic structure relative to noise, with lower values indicating reduced periodicity and higher values reflecting a clearer, more harmonic voice quality. CPP is typically extended to smoothed cepstral peak prominence (CPPS), which averages cepstral magnitudes across time and frequencies to minimise the effects of short-term fluctuations.³⁰ Supporting its value as a measure of voice quality, CPPS has been found to show strong correlations with auditory perceptual measures of hoarseness, including breathiness, roughness, and strain.³² To distinguish between healthy and dysphonic voices based on CPPS measures from sustained vowels calculated with Praat³³, Murton et al.³¹ reported a cutoff of 14.5 dB CPPS. Importantly, when analysing and interpreting CPP or CPPS, it should be considered that the values can be influenced by multiple factors; they are generally higher for vowels than for continuous speech,^{31,34} higher in male than female voices,^{31,35} higher for louder versus softer voice SPL,³⁶ and may also vary depending on the software algorithms used.^{31,35}

Despite CPPS is now a well-established measure of acoustic voice quality, a gap remains in evaluating CPPS changes in response to AFM. We are aware of only one study that has investigated this issue.³⁷ Specifically, Schenck et al.³⁷ explored the impact of loudness and pitch shifts on healthy participants’ changes in acoustic voice quality. Their findings showed no significant effect of pitch shifts, and while loudness shifts produced a relative increase in CPPS, the direction of the voice adaptation varied across participants and could not be tied to upward or downward shifts. In light of these mixed findings, and given our assumption that phonatory compensation primarily relates to the specific voice characteristic being modified, we sought to move beyond pitch and loudness. Thus, we investigated the impact of hoarseness-induced auditory feedback on CPPS.

Research intent

In a series of two studies, our aim was to carry out an auditory-perceptual evaluation of the *VQ-Synth* system and then apply it in the context of AFM. Specifically, Study 1 evaluated *VQ-Synth*’s ability to induce perceived hoarseness in healthy speakers listening to pre-recorded, modified vowels from their own voice, while Study 2 examined the effect of real-time, hoarseness-induced auditory feedback on acoustic voice quality, as measured by CPPS.

In Study 1, participants first recorded a sustained vowel /a:/, which was subsequently manipulated using four resynthesis methods that differed in how noise was added to the signal. In a listening task, participants then repeatedly rated the perceived voice quality of their unaltered and manipulated vowel samples on visual analogue scales. We hypothesised that the manipulated samples would be judged as more impaired, breathy, strained, asthenic, and hoarse than unaltered controls (H1.1). As a side effect of voice manipulation, we also anticipated the manipulated samples to be judged as less natural than the controls (H1.2), as has previously been observed.^{26,27} Furthermore, we predicted differences between the resynthesis methods in their ability to produce a dysphonic yet natural-sounding voice percept (H1.3). The most suitable resynthesis method was selected for real-time AFM in Study 2.

Study 2 employed a real-time AFM paradigm in form of a phase design. Participants produced 140 repetitions of sustained /a:/-vowels, while their auditory feedback was either unaltered or manipulated. In the AFM group, feedback progressed in a fixed order through four phases: baseline (unaltered), ramp (stepwise modulation increase), hold (maximum modulation), and after (unaltered). In the control group, feedback remained unchanged throughout. For the AFM group, we hypothesised that hoarseness-induced auditory feedback would significantly increase CPPS compared to baseline (H2.1). We also expected CPPS to be higher in the hold phase compared to the ramp phase (H2.2). Finally, we anticipated a post-modulation effect, with higher CPPS in the after phase compared to baseline (H2.3). For the control group, we did not expect CPPS to vary across the trials corresponding to the different phases.

Study 1

This study focused on an auditory-perceptual evaluation of the *VQ-Synth* system, assessing its effectiveness in inducing a dysphonic voice percept. To this end, voice resynthesis was performed offline, and participants listened to and rated both unaltered and modified recordings of their own voice.

Method

Participants

Study 1 included 36 participants (25 ♀, 11 ♂), aged between 19 and 42 years ($M = 24$, $SD = 6$), mainly psychology students at RWTH Aachen

University. All met the following inclusion criteria: ≤ 1 voice symptom (self-administered vocal health questionnaire; supplementary material T.1), no self-reported cold in the past 14 days, an Acoustic Voice Quality Index (AVQI) score below 3.05 (German cut-off, indicating healthy voice quality³⁸), normal hearing (≤ 20 dB HL between 500 Hz and 4 kHz; pure-tone audiometry screening, Auritech ear 3.0), C2-level German proficiency (self-report), and normal or corrected-to-normal vision (self-report). Of an initial sample of 56 participants, 20 were excluded for symptoms of dysphonia ($n = 10$), hearing loss ($n = 5$), technical problems ($n = 3$), or participant attrition ($n = 2$). Ethical approval was obtained from the Ethics Committee of the Faculty of Arts and Humanities at RWTH Aachen University (ref. 2023_18_FB7_RWTH Aachen). Participants provided informed consent prior to the study and were reimbursed with €10 or one course credit.

Task

Participants' task was to listen to pre-recorded samples of their own sustained /a:/-vowel (unaltered or manipulated according to four different resynthesis methods described below) and provide voice quality ratings after each sample. Using visual analogue scales, they rated each sample across 6 dimensions: vocal health (0 = *impaired*, 100 = *healthy*), naturalness (0 = *synthetic*, 100 = *natural*), hoarseness (0 = *hoarse*, 100 = *clear*), strain (0 = *strained*, 100 = *effortless*), asthenia (0 = *weak*, 100 = *resonant*), and breathiness (0 = *breathy*, 100 = *notbreathy*). Each sample could be replayed by the participant as many times as they wished.

Technical set-up

VQ-Synth was used to generate 24 resynthesis conditions from four methods, each differing in how noise was integrated with the error signal. Based on the temporal structure of the vocal fold cycle, we applied four methods to introduce noise at specific areas order aligned to specific points. Noise regions were shaped and positioned using a Blackmann window, with adjustments in height, width, and alignment to the pitch peaks in the error signal. We chose this approach to simulate the physiological mechanisms typically underlying hoarseness, such as incomplete glottal closure or irregular vocal fold vibration, which cause audible turbulent airflow and air leakage.¹⁵ The *anti-peak-window* method added noise between peaks; the *peak-window* method added it around peaks; the *peak-window+noise* method combined constant average and peak-related noise; and the *short-term-envelope* method corre-

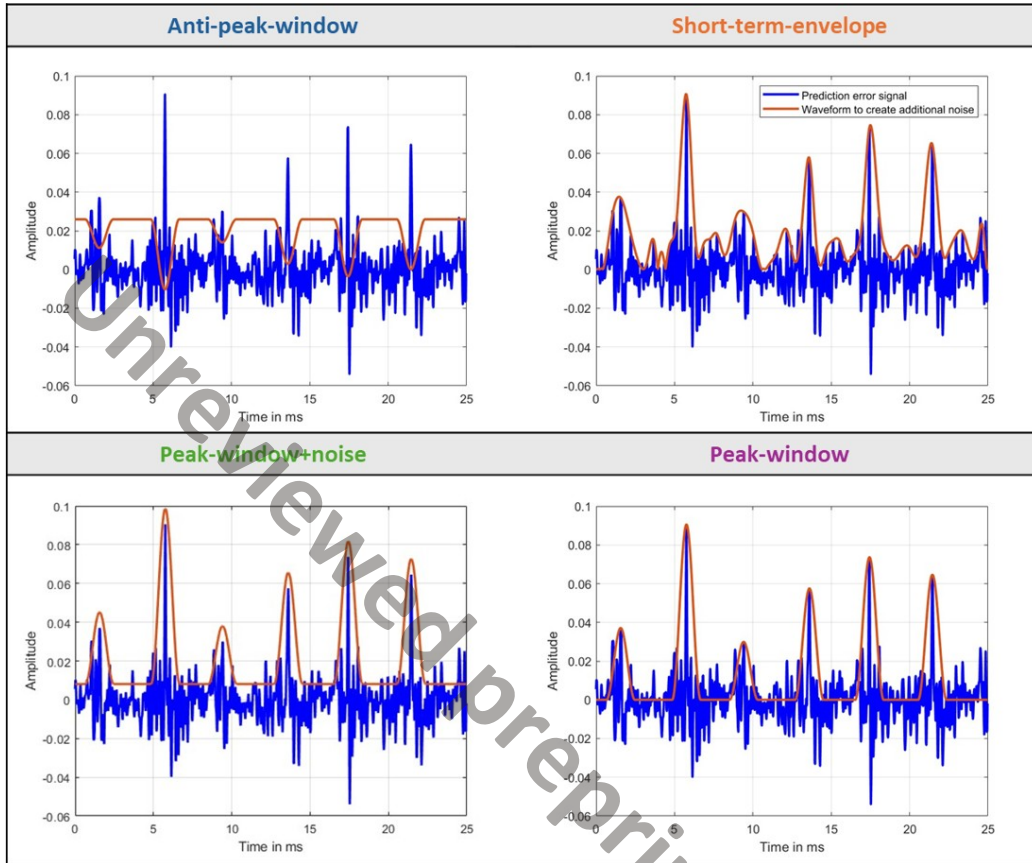


Figure 1: Comparison of the four resynthesis methods for noise addition. The LP error signal is shown in blue; orange lines indicate waveforms multiplied by noise to modulate the voice signal.

lated noise with signal amplitude. For a direct comparison of these methods, see Figure 1. For each method, the six conditions were generated by combining different SNRs (12–26 dB) and SDOs (0 or 2 dB, starting at 1000 Hz).

Prior to the experiment, participants’ voice was recorded using a DPA 4066 headset microphone (2.5 cm from the mouth), and processed via an RME Fireface UC sound card. Voice resynthesis was performed in KiRAT. The unaltered and modified vowel samples were then embedded within the listening task. The presentation level was set to 65 dBA, with stimuli presented via headphones (Sennheiser HD 280 Pro).

Procedure

The experiment took place in a soundproof booth and lasted about 45 minutes. After an introduction and eligibility screening, participants were fitted with the headset microphone and headphones, then trained to produce sustained /a:/ vowels at a constant pitch and amplitude. During the training phase, the experimenter – trained by a speech-language pathologist – first demonstrated a stable vowel production as a model. Participants then practiced several productions, aiming for approximately 4-second vowels with minimal variation in loudness and pitch. Amplitude was monitored using a sound level meter (PCE-322A) positioned in front of the participant, and practice continued until they could consistently maintain a sound pressure level of 65 dBA (± 2 dBA) at a mouth–microphone distance of 30 cm. Pitch stability was assessed by the experimenter, who provided verbal feedback until the vowel productions were perceived as stable.

This training was followed by a voice recording of both a sustained vowel /a:/ and a short phonetically balanced text. Both recordings were used to determine the AVQI score as part of the inclusion criteria. The sustained vowel was additionally used to generate the unaltered and manipulated voice samples for the subsequent listening task. Voice resynthesis occurred automatically in the background using *VQ-Synth*.

The main task, which involved listening to and rating the samples, consisted of 27 randomised trials (preceded by five practice trials), where participants rated each voice sample following on-screen instructions. Out of the 27 trials, three consisted of unaltered voice samples, while the remaining 24 were manipulated according to the resynthesis conditions described above.

Statistical analysis

Data analysis was conducted in RStudio (4.4.2; R Core Team³⁹). We calculated a composite score from the voice-quality ratings to determine the resynthesis method that most effectively induced dysphonia. Each rating dimension contributed equally, resulting in a composite score between 0 and 100. For calculating the composite score, ratings on the visual analogue scales were recoded so that the more impaired, hoarse, strained, asthenic, breathy, and natural a sample was rated, the closer its score was to 100. Thus, higher composite scores indicate samples reflecting better resynthesis performance (i.e., a more dysphonic percept).

To assess the effect of the resynthesis methods, we fitted linear mixed-effects models (LMMs) using the *lme4* package,⁴⁰ with *method* as a fixed

factor and random intercepts for condition (nested in method) and participant. Additional random factors (AVQI, trial, age) did not improve model fit. Diagnostic checks on the LMM confirmed that all model assumptions were satisfied. Post-hoc analyses were conducted using the emmeans package,⁴¹ with Tukey’s HSD adjustment for multiple comparisons.

Results

Figure 2 displays participants’ ratings of the unaltered (control) samples and those modified according to the four resynthesis methods. As expected, unaltered samples were rated as relatively healthy, natural, clear, effortless, resonant, and not breathy (means: 69–78; SDs: 22–25). In contrast, altered voices received ratings clustered nearer the centre of the scale (means: 39–58; SDs: 28–29). Notably, for the *anti-peak-window* method, ratings shifted further toward the impaired end of each scale – the direction targeted by our manipulations.

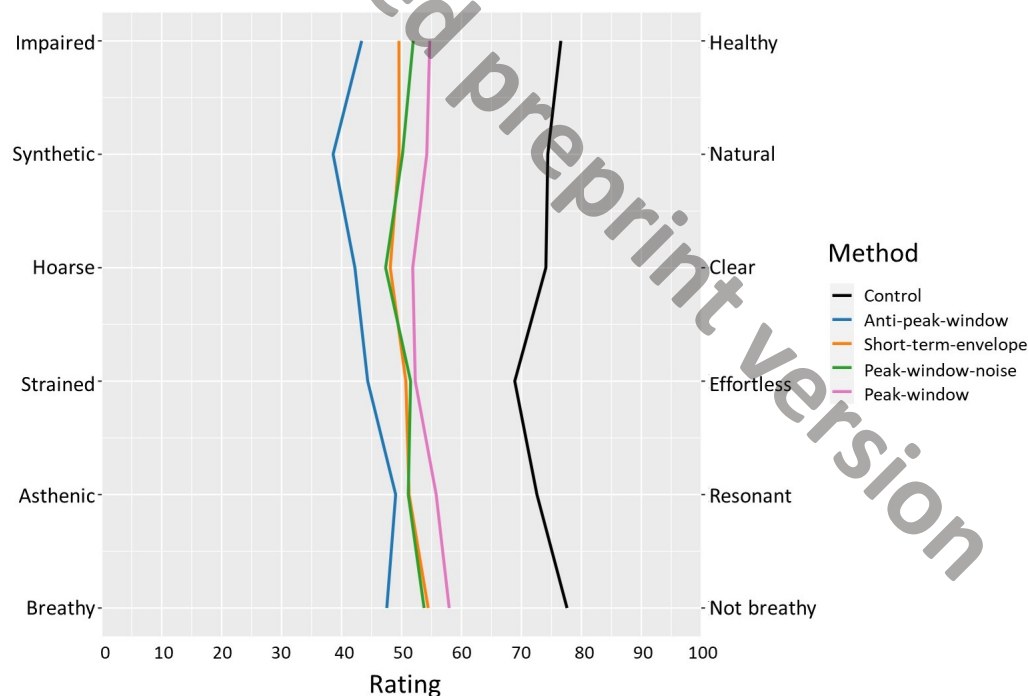


Figure 2: Semantic differential of participants’ auditory-perceptual ratings for unaltered (control) and manipulated voice samples (*anti-peak-window*, *short-term-envelope*, *peak-window+noise*, *peak-window*).

This is also reflected in the composite scores, where the highest scores were achieved with the *anti-peak-window* method ($M = 52, SD = 14$), followed by the *short-term-envelope* ($M = 49, SD = 15$), *peak-window+noise* ($M = 49, SD = 15$), and *peak-window* methods ($M = 47, SD = 16$). As expected, composite scores were lowest for the control ($M = 34, SD = 16$), reflecting normal unimpaired voice quality. The LMM revealed a significant effect of resynthesis method on composite scores ($\chi^2(4) = 23.89, p < .001, \eta_p^2 = 0.59$). Post-hoc analyses showed that all four methods outperformed the control (all p 's $< .05$) with no significant differences between them ($p > .20$).

In summary, Study 1 showed that all resynthesis methods induced the desired dysphonic-voice percept. Auditory-perceptual ratings indicated the *anti-peak-window* method was most effective. Although its advantage was not statistically significant, we selected this method for Study 2 because it was descriptively the most successful in inducing perceived dysphonia.

Study 2

This study examined the effect of hoarseness-induced auditory feedback on healthy speakers' voice quality. In an auditory feedback task, participants repeatedly sustained vowels while feedback followed a phased design (baseline, ramp, hold, after). To account for practice effects, a control group completed the same task with consistently unaltered feedback.

Method

Participants

A new sample of participants was recruited for Study 2, using the same eligibility criteria as in Study 1. The AFM group consisted of 34 participants (22 female, 12 male; age 19–35 years, $M = 24, SD = 3$), after excluding 13 of the initial 47 due to dysphonia ($n = 12$) or technical issues ($n = 1$). The control group included 22 participants (16 female, 6 male; age 19–57 years, $M = 24, SD = 8$) after excluding 12 of 34 due to dysphonia ($n = 7$), technical issues ($n = 4$), or hearing loss ($n = 1$). Again, participants were primarily psychology students who received €10 or a course credit. Informed consent was obtained prior to the study. The study was approved by the Ethics Committee of the Faculty of Arts and Humanities at RWTH Aachen University (AFM group: ref. 2023_18_FB7_RWTH Aachen; control group: ref. 2024_11_FB7_RWTH Aachen).

Task

Participants were asked to sustain the vowel /a:/ for 4 seconds each for a total of 140 trials (plus five practice). For the AFM group, the experiment followed a phase design, similar to Stepp et al.⁴². The first 20 trials involved no feedback modulation (baseline), the subsequent 50 trials involved a stepwise increase of modulation intensity, followed by 50 trials at maximum modulation intensity (hold), and finally another phase of 20 trials with no feedback modulation (after). For the control group, the task was identical, except that auditory feedback was kept unaltered for the entire 140 trials.

Technical set-up

During the experiment, participants' voice was recorded using a DPA 4066 headset microphone (2.5 cm mouth-microphone distance), and processed via an RME Fireface UC sound card. Auditory feedback was presented via closed headphones (Sennheiser HD 280 Pro). Prior to the study, headphone output was calibrated to approximately match participants' own perceived vocal intensity. Calibration was performed using a dummy head (HMS II.3) inside a soundproof booth, which recorded the headphone signal. This was compared to a speaker's voice level (in dBA) measured near the ears with a class 2 sound level meter (PCE-322A), while the speaker, positioned outside the booth, repeatedly produced sustained /a:/ vowels. Headphone gain was adjusted until the levels were matched.

For the AFM group, auditory feedback was modulated using the *anti-peak-window* method (added noise between pitch related amplitude peaks) in *VQ-Synth*. Following the unaltered baseline, the ramp phase involved a stepwise decrease in SNR from +17 dB to +13 dB, with SDO held constant at 2 dB. During the hold phase, maximum modulation was applied at +12 dB SNR and 2 dB SDO, after which the alteration returned to zero in the after phase. In the ramp and hold phases, AFM began quasi-randomly between 1.0 and 1.3 s after phonation onset, preceded by a 100 ms fade-in, and was maintained until the end of each trial. For the control group, the auditory feedback was unaltered throughout all trials. The experiment was implemented in KiRAT allowing real-time modulation.

Procedure

The experiment was conducted in a soundproof booth (studiobox premium) and lasted approximately one hour. After eligibility screening, participants were fitted with the headset microphone and headphones, trained

to produce /a:/ at a constant pitch and amplitude, and recorded samples for AVQI calculation. They then performed the main auditory feedback task while seated in front of a computer screen that guided them through the procedure. Voice input was recorded and digitised at 44.1 kHz with a 24 bit resolution. Participants remained blind to the specific study aim until debriefing, being told only that the goal was to assess voice changes over time.

Analysis

To assess the effect of hoarseness-induced auditory feedback on CPPS, we extracted CPPS values from participants’ voice recordings using Parselmouth, a Python interface to Praat.^{33,43} For trials with unaltered auditory feedback (baseline and after trials in the AFM group, and all trials in the control group), CPPS was calculated from 2-s steady state mid-vowel portions. In the AFM group’s ramp and hold phases, CPPS was calculated from a 300-ms window, following Schenck et al.³⁷. This analysis window was chosen to span 100–400 ms after modulation onset, based on PSR research indicating that phonatory responses do not occur within the first 100 ms.^{8,10,44}

Statistical analysis was conducted in RStudio (v4.4.2; R Core Team³⁹). Using the lme4 package,⁴⁰ three LMMs were fitted: one assessing the effect of *phase* on CPPS in the AFM group, one in the control group, and a third evaluating the *phase-by-group* interaction on standardised CPPS (z-scored calculated within each group, $M = 0$, $SD = 1$) across all data to ensure group differences reflect true effects. In addition to *phase*, we included *gender* as a fixed effect to control for male–female differences. All participants self-identified as male or female; none reported a non-binary gender. Random effects – participant ID, AVQI score, trial, RMS input (root-mean-square amplitude of participants’ voice input), f_0 input (mean f_0 in participants’ voice input), jitter input (refers to the degree of frequency fluctuations in the voice input), and shimmer input (refers to the degree of amplitude fluctuations in the voice input) – were sequentially added to the LMMs and only retained if they significantly improved the model fit, which was assessed through likelihood ratio tests. The final LMMs were specified as:

$$\begin{aligned} \text{CPPS}_{\text{AFM}} &\sim \text{Phase} + \text{Gender} + (1 \mid \text{Trial}) + (1 \mid \text{RMS}_{\text{input}}) + (1 \mid f_{0_input}) + (1 \mid \text{Jitter}_{\text{input}}) + (1 \mid \text{ID}) \\ \text{CPPS}_{\text{control}} &\sim \text{Phase} + \text{Gender} + (1 \mid \text{Trial}) + (1 \mid \text{RMS}_{\text{input}}) + (1 \mid f_{0_input}) + (1 \mid \text{Jitter}_{\text{input}}) + (1 \mid \text{ID}) \\ \text{CPPS}_{\text{standardised}} &\sim \text{Phase} * \text{Group} + \text{Gender} + (1 \mid \text{RMS}_{\text{input}}) + (1 \mid f_{0_input}) + (1 \mid \text{Jitter}_{\text{input}}) + (1 \mid \text{ID}) \end{aligned}$$

Diagnostic checks indicated that model assumptions were met by all three LMMs. Post-hoc analyses were conducted as in Study 1.

Table 1: Results from the LMM modelling the effect of phase (baseline, ramp, hold, after) on CPPS estimates in Study 2’s AFM group

Effect	Estimate	SE	df	t	p
Intercept	14.44	0.31	35	46.97	< 0.001
Ramp	0.34	0.06	160	5.61	< 0.001
Hold	0.58	0.06	160	9.46	< 0.001
After	0.56	0.07	172	7.59	< 0.001
Gender(m)	2.60	0.51	32	5.12	< 0.001

Note. Intercept = estimated CPPS for women in the baseline phase; Ramp, Hold, and After = phase effects relative to baseline; Gender(m) = relative CPPS increase for male compared to female participants; Estimate = model coefficient; SE = standard error; df = degrees of freedom; t = t-statistic, p = p-value.

Results

The LMM modelling the data from the AFM group revealed a significant effect of *phase* (baseline [trials 1–20], ramp [trials 21–70], hold trials [71–120], after [trials 121–140]) on CPPS, $\chi^2(3) = 100.21$, $p < .001$, $\eta_p^2 = .37$ (see Table 1 for a summary of the model output). Assessing this further, pairwise comparisons showed that CPPS significantly increased from baseline to ramp ($\Delta = 0.34$, $SE = 0.06$, $p < .001$, $d = 0.29$), from baseline to hold ($\Delta = 0.58$, $SE = 0.06$, $p < .001$, $d = 0.49$), and from ramp to hold ($\Delta = 0.23$, $SE = 0.05$, $p < .001$, $d = 0.20$). Moreover, CPPS remained elevated in the after phase, with a significant difference compared to baseline ($\Delta = 0.56$, $SE = 0.07$, $p < .001$, $d = 0.48$) and no significant difference compared to hold ($p = .997$). Notably, CPPS varied with *gender*, $\chi^2(1) = 26$, $p < .001$, $\eta_p^2 = .45$, with higher values in males compared to females ($\Delta = 2.60$, $SE = 0.51$, $p < .001$, $d = 2.20$). Descriptive CPPS values per *phase* were: baseline $M = 15.39$, $SD = 2.70$; ramp $M = 15.71$, $SD = 2.92$; hold $M = 15.93$, $SD = 3.03$; after $M = 15.91$, $SD = 2.94$. The CPPS trajectory is visualised in Figure 3, showing mean CPPS across all 140 trials with a LOESS curve (locally weighted polynomial regression) and shaded 95% confidence interval.

In the control group, where auditory feedback remained unaltered throughout the experiment, mean CPPS per *phase* were relatively stable across baseline ($M = 16.08$, $SD = 2.70$), ramp ($M = 16.02$, $SD = 2.92$), and hold ($M = 16.13$, $SD = 3.03$), and slightly decreased in the after phase ($M = 15.92$, $SD = 2.94$). The LMM revealed a significant but small effect of

phase on CPPS, $\chi^2(3) = 8.08$, $p = .044$, $\eta_p^2 = .05$, with pairwise comparisons indicating that CPPS decreased from *hold* to *after* ($\Delta = -0.26$, $SE = 0.09$, $p = .027$, $d = 0.21$), while all other comparisons were not statistically significant ($p \geq .19$). Again, CPPS varied with *gender*, $\chi^2(1) = 5.7$, $p = .017$, $\eta_p^2 = .23$, with higher values in males compared to females (see Table 2).

Finally, our LMM based on standardised CPPS from both groups (AFM and control) revealed a small but significant *phase-by-group* interaction, $F(3, 6647) = 15.74$, $p < .001$, $\eta_p^2 = .007$ (see Table 3). Figure 4 illustrates this interaction, demonstrating the distinct trajectories of standardised CPPS as a function of *group* and *phase*: in the AFM group, CPPS increased from baseline to ramp, peaked during hold, and remained high in after, whereas in the control group, CPPS showed no increase during ramp and hold and even decreased in after. Reflecting the results of the group-separated LMMs, standardised CPPS was higher in males than females, $F(1, 51.5) = 28.69$, $p < .001$, $\eta_p^2 = .36$.

Table 2: Results from the LMM modelling the effect of phase (baseline, ramp, hold, after) on CPPS estimates in Study 2's control group

Effect	Estimate	SE	df	<i>t</i>	<i>p</i>
Intercept	15.12	0.62	20	24.55	< 0.001
Ramp	-0.02	0.09	143	-0.27	0.784
Hold	0.05	0.09	160	0.54	0.588
After	-0.21	0.11	155	-1.87	0.064
Gender(m)	2.81	1.18	19	2.39	0.027

Note. Intercept = estimated CPPS for women in the baseline phase; Ramp, Hold, and After = phase effects relative to baseline; Gender(m) = relative CPPS increase for male compared to female participants; Estimate = model coefficient; SE = standard error; df = degrees of freedom; *t* = t-statistic, *p* = p-value.

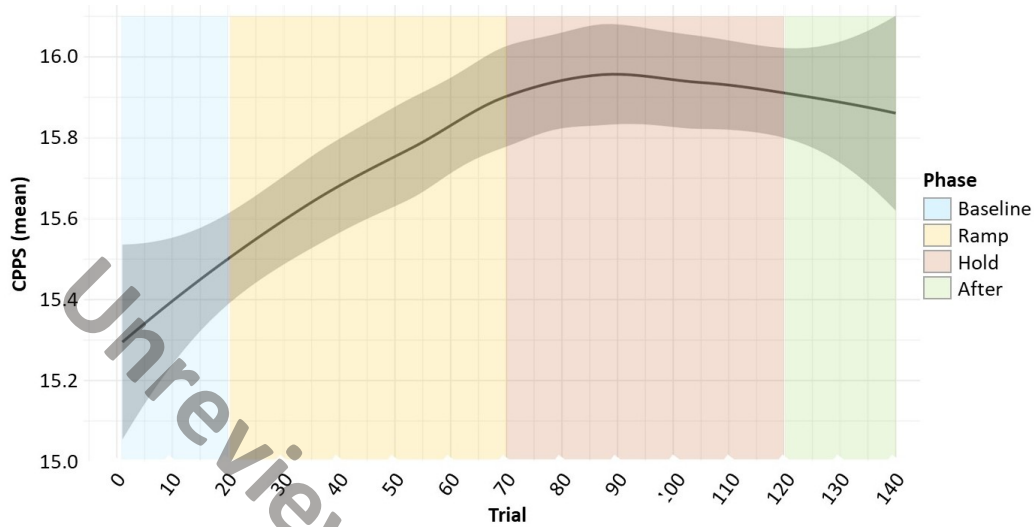


Figure 3: Changes in CPPS across experimental phases in Study 2’s AFM group. In the baseline phase, auditory feedback was unaltered; the ramp phase involved a stepwise increase of hoarseness; in the hold phase, the maximum level of hoarseness was applied; and in the after phase, feedback returned to its unaltered state.

Note. The LOESS line shows the smoothed CPPS trajectory across all trials. The early rise (first 20 trials) may reflect both a practice effect and smoothing over multiple phases, which averages phase-specific changes.

Table 3: Fixed effects from LMM fitted by REML, showing the effect of phase and study on standardised CPPS estimates

Effect	Estimate	SE	df	t	p
Intercept	-0.56	0.14	53	-3.92	< 0.001
Group (Control)	0.20	0.20	54	1.00	0.321
Phase (Ramp)	0.15	0.02	7081	6.05	< 0.001
Phase (Hold)	0.26	0.03	7046	10.50	< 0.001
Phase (After)	0.26	0.03	6525	8.53	< 0.001
Gender (Male)	1.09	0.20	51	5.36	< 0.001
Group:Phase (Control:Ramp)	-0.16	0.04	7154	-3.93	< 0.001
Group:Phase (Control:Hold)	-0.23	0.04	6368	-5.68	< 0.001
Group:Phase (Control:After)	-0.31	0.05	6750	-6.31	< 0.001

Note. The estimates refer to standardised CPPS (z-scores). The intercept represents the reference group at baseline for female participants. Other terms indicate differences between *group*, *phase*, or *gender*, as well as the *group-by-phase* interaction. Estimate = model coefficient; SE = standard error; df = degrees of freedom; t = t-statistic; p = p-value.

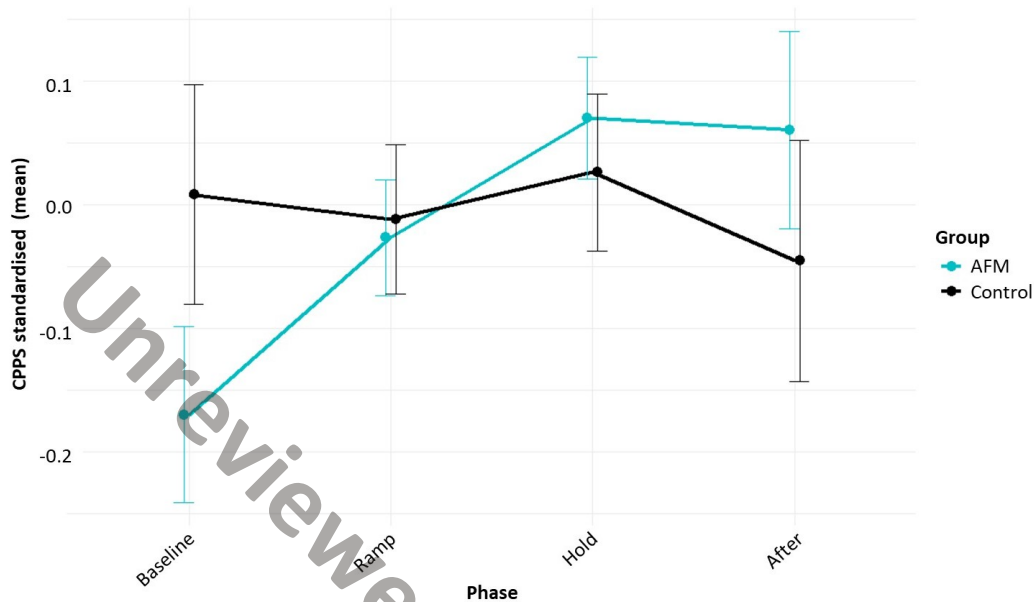


Figure 4: Interaction between group (AFM and control) and experimental phase (baseline, ramp, hold, after), based on standardised CPPS values (z-scores).

Note. Considering that CPPS was z-scored within each group, the *group-by-phase* interaction reflects differences in CPPS trajectories across phases, not baseline levels.

Discussion

While previous research has shown that pitch- and loudness-shifted AFM can elicit compensatory vocal responses in healthy speakers, it has remained unclear whether hoarseness alterations produce similar effects. We have addressed this gap by developing, evaluating, and testing a novel voice resynthesis system, *VQ-Synth*, designed to induce hoarseness in speakers' auditory feedback. Our findings suggest that *VQ-Synth* can induce the percept of a dysphonic voice and may also elicit compensatory vocal responses that enhance acoustic voice quality.

In Study 1, we compared four resynthesis methods (*anti-peak-window*, *short-term-envelope*, *peak-window+noise*, and *peak-window*) in terms of their effectiveness at inducing a dysphonic voice percept. Auditory-perceptual ratings revealed that, relative to the control voice samples, all voice manipulation methods increased perceived voice impairment, hoarseness, strain,

asthenia, and breathiness. This supports our first hypothesis (H1.1) and indicates that *VQ-Synth* meets its design goal. At the same time, consistent with our second hypothesis (H1.2), manipulated samples were judged as less natural than unaltered ones. Our third hypothesis (H1.3), which predicted differences in the methods' effectiveness at inducing perceived dysphonia, was not confirmed. The *anti-peak-window* method produced the highest composite score, indicating it was most effectively in generating perceived hoarseness, but differences remained descriptive.

The samples manipulated with *VQ-Synth* were rated more dysphonic but also less natural compared to control samples, which aligns with earlier work by Ruinskiy et al.²⁶ and Schiller et al.²⁷. Both of these studies aimed to increase perceived hoarseness in prerecorded healthy – and, in the case of Schiller et al.,²⁷ also impaired – voices, with perceptual effects evaluated through listening experiments. In both studies, modified voices were rated as hoarser but less natural, especially with stronger manipulations. Notably, because *VQ-Synth* is designed for real-time AFM rather than offline manipulations, slight losses in naturalness may be less problematic, assuming that the altered feedback is integrated into ongoing self-monitoring rather than perceived as an external stimulus. This is supported by compensatory responses in pitch-shift experiments⁴⁻⁸, which show that speakers automatically correct perceived deviations in their own voice.

In Study 2, we assessed the effect of hoarseness-induced auditory feedback on participants' acoustic voice quality, measured via CPPS. The *anti-peak-window* method was used in a phased design. For the AFM group – but not the control group – we observed a significant increase in CPPS during the ramp and hold phases compared to baseline. This confirms H2.1, indicating that hoarseness-induced auditory feedback improved acoustic voice quality. Consistent with H2.2, we also found significantly higher CPPS in the hold phase compared to the ramp phase, reflecting stronger effects at higher modulation intensity. Finally, as predicted by H2.3, the effect of hoarseness-induced AFM even persisted into the after phase, reflecting a post-modulation effect.

These findings can be interpreted in light of the DIVA model^{1,2,12} and previous pitch- and amplitude-shift experiments.^{4,7-9} Consistent with the DIVA model, the CPPS increase in response to hoarsified AFM suggest that participants detected a mismatch between expected and perceived voice quality which triggered updated motor commands to the laryngeal musculature to reduce the discrepancy. These vocal motor adaptations may have involved

physiological processes such as adjusting glottal closure to reduce phonation noise, fine-tuning vocal fold tension, and regulating subglottal pressure – processes that are often impaired in voice disorders (see e.g., Reiter et al.¹⁵).

A comparison between the AFM and control groups confirmed that the CPPS increase observed in the ramp and hold phases was not due to practice effects. In fact, the control group, which received unaltered auditory feedback throughout the experiment, showed no CPPS improvement; their CPPS even declined in trials corresponding to the after phase, suggesting the onset of vocal fatigue.⁴⁵ In light of this, it is all the more notable that the AFM group exhibited a post-modulation effect, with CPPS remaining elevated in the after phase. Similar post-modulation effects have previously been observed in pitch-shift experiments.^{5,6} Taken together, it appears that hoarseness-induced auditory feedback not only improved acoustic voice quality but also counteracted the vocal fatigue that tends to emerge toward the end of a demanding vocal task. This is crucial: if hoarseness-induced auditory feedback enhances voice quality in healthy speakers, those with functional voice disorders such as MTD could exhibit even greater improvements due to their greater capacity for change, a possibility that remains to be investigated.

Despite these promising findings, several limitations of the present study should be acknowledged. In Study 1, some uncertainty remains regarding the extent to which the different voice resynthesis methods produced an authentic dysphonic voice percept. This is reflected in participants' ratings for the altered voice qualities, which tended to cluster around the midpoint of the visual analog scales – particularly for the *short-term-envelope*, *peak-window+noise*, and *peak-window* methods. Two factors may have contributed to this pattern. First, the participants were inexperienced raters, which, despite receiving brief definitions of each perceptual dimension, may have caused uncertainty in how they rated the voice samples along these dimensions. Second, the varying modulation intensities within each method may have promoted this central clustering, as the results were presented within methods but averaged across the SNR and SDO conditions. In Study 2, while the significant CPPS increase of 0.6 dB from baseline to hold represents a large and statistically significant effect, its perceptual relevance remains unclear. Further research is needed both to determine which resynthesis settings might elicit even larger CPPS increases, and to identify the threshold at which such increases are subjectively perceived. For context, Hofman et al.⁴⁶ reported that, in 22 patients with various voice disorders,

average CPPS measured from sustained vowels increased from approximately 14.5 dB at baseline to 16.4 dB following voice therapy. As a last limitation, we acknowledge that our Study 2 analysis does not account for potential variation in individual response patterns to AFM. This would have required a data stratification into different response types (compensatory, following, and nil responses), which was beyond the scope of the present study. While our choice was to explore hoarseness-induced AFM effects at a more global level, delving deeper into individual response patterns is planned for future works.

Our next step will be to evaluate *VQ-Synth*'s performance in individuals with dysphonia. Specifically, we aim to repeat Study 2 with participants with MTD, a population that has not yet been investigated in this context. Previous research using pitch-shifted AFM in individuals with MTD yielded mixed results, with a study by Sonj et al.⁴⁸ showing compensatory responses similar to healthy talkers, with compensation magnitude being correlated with psychological measures of depression, while a study by Stepp et al.⁴² suggested that phonatory responses may be less consistent than in healthy speakers. In their study, only four of nine dysphonic participants compensated for pitch shifts, while the remaining five followed the shift. Generalisation of these results is limited by the small sample size and the fact that pitch rather than voice quality was altered.

In our future research, we will also validate the system's ability to hoarsify voice in connected speech and consider analysing additional acoustic parameters (e.g., jitter, shimmer, or HNR; see e.g., Warhurst et al.²⁸) to evaluate the speaker's phonatory responses to hoarseness-induced auditory feedback. Currently, *VQ-Synth* employs traditional signal processing for voice manipulation rather than neural networks, because (1) it enables precise control of acoustic features and (2) there is no sufficiently large dysphonic-voice training dataset. Nevertheless, we have begun exploring neural network-based AFM approaches to induce hoarseness,⁴⁷ which, while still in an early stage, may ultimately prove more effective. Ultimately, we envision *VQ-Synth* enhancing traditional voice therapy and aiding professional voice users, with future research needed to explore if this potential can be realised.

Conclusion

In two studies, we evaluated the performance of the voice-resynthesis system *VQ-Synth* in inducing perceived hoarseness in healthy voices and

examined its effects within the context of auditory feedback modulation. Our findings demonstrated that *VQ-Synth* achieves its design goal: especially under the *anti-peak-window* method – where noise was added between amplitude-related pitch peaks – voice samples were judged as significantly more dysphonic than unaltered control samples. A key advantage of *VQ-Synth* over previous systems is its ability to alter voice quality in real-time. When applied in an AFM experiment, feedback modulation using *VQ-Synth* resulted in compensatory responses towards improved acoustic voice quality, assessed via CPPS. Future research will focus on refining *VQ-Synth* to enhance response magnitude, validate its functionality in the context of connected speech, and evaluate its potential to improve voice quality in individuals with dysphonia.

Acknowledgements

We thank F. Schellong, M. Kretz, C. Lei, and L. Willms for their support in CPPS extraction. We further thank C. Baur, E. Dürrwächter, M. Heinrichs, L. Köchling, L. Kruse, and M. Tenberg for their support in data collection. Finally, we gratefully acknowledge Dr. Catherine Madill for reviewing the manuscript prior to submission and for providing valuable feedback.

Data Availability Statement

The dataset supporting this study is available on the Open Science Framework (OSF) through a private, view-only link for peer review. The dataset will be made publicly available with a DOI upon article acceptance.

Author Contribution

I. S. Schiller: Conceptualisation, Methodology, Validation, Formal Analysis, Data Curation, Writing-original draft preparation, Visualisation, Supervision, Project Administration, Funding Acquisition; K. Krüger: Methodology, Software, Writing - Review & Editing, Visualisation; P. Weede: Methodology, Software und Writing - Review and Editing; M. T. Sopha: Investigation, Methodology, Writing - Review & Editing; G. Schmidt: Methodology, Software, Writing - Review & Editing

Funding Sources

I. S. Schiller's contribution to this study was supported by the Excellence Strategy of the Federal Government and the Länder (StUpPD_445-23) and by a grant from the HEAD-Genuit-Foundation (P-16/10-W), which was awarded to Prof. Dr. Sabine J. Schlittmeier, head of the Teaching and Research Area for Work and Engineering Psychology.

Declaration of generative AI in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

- [1] Guenther FH. Cortical interactions underlying the production of speech sounds. *J Commun Disord.* 2006;39(5):350-365. <https://doi.org/10.1016/j.jcomdis.2006.06.013>
- [2] Guenther FH, Hickok G. Neural models of motor speech control. In: Hickok G, Small SL, eds. *Neurobiology of Language*. Academic Press; 2016:725-740. <https://doi.org/10.1016/B978-0-12-407794-2.00058-4>
- [3] Alves M, Mancini PC, Teixeira LC. Modifications of auditory feedback and its effects on the voice of adult subjects: a scoping review. *CoDAS.* 2023;35:e202220202. <https://doi.org/10.1590/2317-1782/20232022202en>
- [4] Jones JA, Munhall KG. Perceptual calibration of F0 production: evidence from feedback perturbation. *J Acoust Soc Am.* 2000;108(3):1246-1251. <https://doi.org/10.1121/1.1288414>
- [5] Ning LH. Sensorimotor adaptation and aftereffect to frequency-altered feedback in Mandarin-speaking vocalists and non-vocalists. *Concentric.* 2020;46(2):125-147. <https://doi.org/10.1075/consl.00015.nin>
- [6] Feng Y, Xiao Y, Yan Y, Max L. Adaptation in Mandarin tone production with pitch-shifted auditory feedback: influence of tonal contrast requirements. *Lang Cogn Neurosci.* 2018;33(6):734-749. <https://doi.org/10.1080/23273798.2017.1421317>
- [7] Scheerer NE, Jones JA. Detecting our own vocal errors: an event-related study of the thresholds for perceiving and compensating for vocal pitch errors. *Neuropsychologia.* 2018;114:158-167. <https://doi.org/10.1016/j.neuropsychologia.2017.12.007>
- [8] Larson CR, Sun J, Hain TC. Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F0 and amplitude control. *J Acoust Soc Am.* 2007;121(5):2862-2872. <https://doi.org/10.1121/1.2715657>
- [9] Bauer JJ, Mittal J, Larson CR, Hain TC. Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude. *J Acoust Soc Am.* 2006;119(4):2363-2371. <https://doi.org/10.1121/1.2173513>

- [10] Liu P, Chen Z, Jones JA, Huang D, Liu H. Auditory feedback control of vocal pitch during sustained vocalization: a cross-sectional study of adult aging. *PLoS One*. 2011;6(7):e22791. <https://doi.org/10.1371/journal.pone.0022791>
- [11] Arbeiter M, Petermann S, Hoppe U, et al. Analysis of the auditory feedback and phonation in normal voices. *Ann Otol Rhinol Laryngol*. 2018;127(2):89-98. <https://doi.org/10.1177/0003489417744567>
- [12] Guenther FH. A neural network model of speech acquisition and motor equivalent speech production. *Biol Cybern*. 1995;72(1):43-53. <https://doi.org/10.1007/BF00206237>
- [13] Kreiman J, Vanlancker-Sidtis D, Gerratt BR, et al. Defining and measuring voice quality. In: *Proceedings of From Sound To Sense*. 2003:115-120. Accessed October 14, 2025. https://www.isca-archive.org/voqual_2003/kreiman03b_voqual.pdf
- [14] Kreiman J, Sidtis D. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell; 2011.
- [15] Reiter R, Hoffmann TK, Pickhard A, Brosch S. Hoarseness-causes and treatments. *Dtsch Arztebl Int*. 2015;112(19):329-337. <https://doi.org/10.3238/arztebl.2015.0329>
- [16] Dejonckere PH, Obbens C, De Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr Logop*. 1993;45(2):76-83. <https://doi.org/10.1159/000266220>
- [17] Oliveira P, Ribeiro VV, Constantini AC, et al. Prevalence of work-related voice disorders in voice professionals: systematic review and meta-analysis. *J Voice*. 2024;39(1):84-104. <https://doi.org/10.1016/j.jvoice.2022.07.030>
- [18] Rosow DE, Szczupak M, Saint-Victor S, et al. The economic impact of vocal attrition in public school teachers in Miami-Dade County. *Laryngoscope*. 2016;126(3):665-671. <https://doi.org/10.1002/lary.25513>
- [19] Lyberg-Åhlander V, Haake M, Brännström J, Schötz S, Sahlén B. Does the speaker's voice quality influence children's performance on a language comprehension test? *Int J Speech Lang Pathol*. 2015;17(1):63-73. <https://doi.org/10.3109/17549507.2014.898098>

- [20] Brännström KJ, Holm L, Lyberg-Åhlander V, et al. Children's subjective ratings and opinions of typical and dysphonic voice after performing a language comprehension task in background noise. *J Voice*. 2015;29(5):624-630. <https://doi.org/10.1016/j.jvoice.2014.11.003>
- [21] Schiller IS, Breuer C, Aspöck L, et al. A lecturer's voice quality and its effect on memory, listening effort, and perception in a VR environment. *Sci Rep*. 2024;14(1):12407. <https://doi.org/10.1038/s41598-024-63097-6>
- [22] Schiller IS, Aspöck L, Schlittmeier SJ. The impact of a speaker's voice quality on auditory perception and cognition: a behavioral and subjective approach. *Front Psychol*. 2023;14:1243249. <https://doi.org/10.3389/fpsyg.2023.1243249>
- [23] Van Houtte E, Van Lierde K, D'haeseleer E, Claeys S. The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia. *Laryngoscope*. 2010;120(2):306-312. <https://doi.org/10.1002/lary.20696>
- [24] Desjardins M, Apfelbach C, Rubino M, Verdolini Abbott K. Integrative review and framework of suggested mechanisms in primary muscle tension dysphonia. *J Speech Lang Hear Res*. 2022;65(5):1867-1893. https://doi.org/10.1044/2022_JSLHR-21-00575
- [25] Böhm T, Audibert N, Shattuck-Hufnagel S, Németh G, Aubergé V. Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *J Acoust Soc Am*. 2008;123(5):3886.
- [26] Ruinskiy D, Lavner Y. Stochastic models of pitch jitter and amplitude shimmer for voice modification. In: *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*. 2008:489-493. <https://doi.org/10.1109/EEEI.2008.4736577>
- [27] Schiller IS, Schnapka A, Eggert C, Birkholz P, Stone S. VQ-Synth: development and perceptual evaluation of a system for voice quality modification. In: Astolfi A, ed. *Proceedings of the 10th Convention of the European Acoustics Association, Forum Acusticum 2023*. European Acoustics Association; 2023:3533-3540. <https://doi.org/10.61782/fa.2023.0250>
- [28] Warhurst S, Madill C, McCabe P, Heard R, Yiu E. The vocal clarity of female speech-language pathology students: an exploratory study. *J Voice*. 2012;26(1):63-68. <https://doi.org/10.1016/j.jvoice.2010.10.008>

- [29] Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *Am J Speech Lang Pathol*. 2018;27(3):887-905. https://doi.org/10.1044/2018_AJSLP-17-0009
- [30] Ponce JRC, Montelongo LAG, Silva JEJ, no González JLT. Smoothed cepstral peak prominence: a comparison between dysphonic and non-dysphonic Mexican adults employing the Praat software. *Cureus*. 2024;16(10):e72292. <https://doi.org/10.7759/cureus.72292>
- [31] Murton O, Hillman R, Mehta D. Cepstral peak prominence values for clinical voice evaluation. *Am J Speech Lang Pathol*. 2020;29(3):1596-1607. https://doi.org/10.1044/2020_AJSLP-20-00001
- [32] Antonetti AES, Siqueira LTD, Gobbo MPDA, Brasolotto AG, Silverio KCA. Relationship of cepstral peak prominence-smoothed and long-term average spectrum with auditory-perceptual analysis. *Appl Sci*. 2020;10(23):8598. <https://doi.org/10.3390/app10238598>
- [33] Boersma P, Weenink D. Praat: doing phonetics by computer [computer software]. 2023. <http://www.praat.org/>
- [34] Madill C, Chacon A, Kirby E, Novakovic D, Nguyen DD. Active ingredients of voice therapy for muscle tension voice disorders: a retrospective data audit. *J Clin Med*. 2021;10(18):4135. <https://doi.org/10.3390/jcm10184135>
- [35] Saggio G, Costantini G. Worldwide healthy adult voice baseline parameters: a comprehensive review. *J Voice*. 2022;36(5):637-649. <https://doi.org/10.1016/j.jvoice.2020.08.028>
- [36] Brockmann-Bauser M, Van Stan JH, Sampaio MC, Bohlender JE, Hillman RE, Mehta DD. Effects of vocal intensity and fundamental frequency on cepstral peak prominence in patients with voice disorders and vocally healthy controls. *J Voice*. 2021;35(3):411-417. <https://doi.org/10.1016/j.jvoice.2019.11.015>
- [37] Schenck A, Hilger AI, Levant S, Kim JH, Lester-Smith RA, Larson C. The effect of pitch and loudness auditory feedback perturbations

- on vocal quality during sustained phonation. *J Voice*. 2023;37(1):37-47. <https://doi.org/10.1016/j.jvoice.2020.11.001>
- [38] Maryn Y, De Bodt M, Barsties B, Roy N. The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Otorhinolaryngol*. 2014;271(6):1609-1619. <https://doi.org/10.1007/s00405-013-2730-7>
- [39] R Core Team. R: A Language and Environment for Statistical Computing [computer software]. R Foundation for Statistical Computing; 2024. <https://www.R-project.org/>
- [40] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48. <https://doi.org/10.18637/jss.v067.i01>
- [41] Lenth RV. Least-squares means: the R package lsmeans. *J Stat Softw*. 2016;69(1):1-33. <https://doi.org/10.18637/jss.v069.i01>
- [42] Stepp CE, Lester-Smith RA, Abur D, et al. Evidence for auditory-motor impairment in individuals with hyperfunctional voice disorders. *J Speech Lang Hear Res*. 2017;60(6):1545-1550. https://doi.org/10.1044/2017_JSLHR-S-16-028
- [43] Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: a Python interface to Praat. *J Phon*. 2018;71:1-15. <https://doi.org/10.1016/j.wocn.2018.07.001>
- [44] Naunheim ML, Yung KC, Schneider SL, et al. Vocal motor control and central auditory impairments in unilateral vocal fold paralysis. *Laryngoscope*. 2019;129(9):2112-2117. <https://doi.org/10.1002/lary.27680>
- [45] Nanjundeswaran C, G, Konstanty K, Keinath C, Huber JE. Respiratory responses to vocal demand tasks: a scoping review. *J Voice*. Published online October 10, 2024. <https://doi.org/10.1016/j.jvoice.2024.10.031>
- [46] Hofman EC, Dassie-Leite AP, Martins PN, Pereira EC. Acoustic measurements of CPPS and AVQI pre and post speech therapy. *CoDAS*. 2023;35:e20220136. <https://doi.org/10.1590/2317-1782/20232022136en>

- [47] Eikens H, Krüger K, Röhrdanz F, Schmidt G, Schiller I. Neural-network based auditory feedback modulation in real-time. In: *2025 IEEE International Professional Communication Conference (ProComm)*. 2025:337-339. <https://doi.org/10.1109/ProComm64814.2025.00071>
- [48] Sonj SS, Torabinezhad F, Saffarian A, Abolghasemi J, Behroozmand R. Psychological correlates of auditory-motor integration in primary muscle tension dysphonia: a preliminary study. *J Voice*. Published online July 30, 2025. <https://doi.org/10.1016/j.jvoice.2025.07.030>

Unreviewed preprint version