H. Eikens*, A. Culcu, K. Hollborn, I. S. Schiller, and G. Schmidt

Targeted Cepstral Jitter Modeling Using Lightweight Neural Networks for Real-time Auditory Feedback Modulation

https://doi.org/10.1515/

Abstract: Real-time auditory feedback modulation requires signal representations that allow perceptually effective voice-quality control under strict latency constraints. This work investigates a cepstrum-based approach to modulating perceived vocal roughness, a central aspect of hoarseness, in sustained vowels. The proposed method introduces frame-wise cepstral peak shifts to impose jitter-like temporal instability and uses a lightweight gated recurrent unit (GRU) to generate short-term modulation trajectories. The model is motivated by an analysis of imitated hoarse vowel recordings showing a characteristic jitter distribution and a pronounced negative autocorrelation at lag 1. Analytical evaluation shows that the proposed system reproduces these global short-term jitter properties well and induces peak-based cepstral changes consistent with imitated hoarse speech. A listening study further shows that the modulation increases perceived hoarseness and roughness, but also reduces naturalness and increases technical artifacts relative to unmodified speech. The results suggest that the neural model captures the intended local jitter dynamics, while the main limitations arise from the cepstral real-time representation and reconstruction constraints.

Keywords: auditory feedback modulation, hoarseness, cepstrum, jitter, GRU, real-time speech processing

1 Introduction

Auditory feedback modulation (AFM) provides a means to influence voice production by altering the speaker's own auditory feedback signal during phonation [1]. While many AFM approaches focus on pitch or loudness perturbations, voice-quality modulation is of particular interest for applications in voice training and therapy [2–4]. In this context, hoarseness is a relevant modulation target because it is associated with irregular phonation and impaired voice quality, while hoarseness-related feedback manipulations may elicit compensatory vocal adjustments toward more stable phonation [2, 3, 5, 6].

This paper evaluates a real-time cepstrum-based method for targeted modulation of hoarseness-related roughness in sustained vowels. The approach extends an existing AFM framework based on cepstral-domain processing [7]. The central idea is to impose structured temporal instability on the excitation-related cepstral peak by means of frame-wise jitter modulation, generated by a lightweight GRU under real-time constraints. Imitated hoarse phonation was used as reference, since naturally hoarse paired recordings from the same speakers are difficult to obtain in sufficient quantity.

2 Data-Driven Motivation

The chosen approach is based on an analysis of sustained vowel recordings in modal and imitated hoarse phonation. In the hoarse recordings, frame-wise variations of the first cepstral peak reveal two consistent properties. First, the distribution of relative jitter values is sharply centered around zero with symmetric, rapidly decaying tails, as shown in Figure 1. Second, the temporal structure is dominated by a pronounced negative autocorrelation at lag 1, i.e., a negative statistical dependence between consecutive frames. A value of about -0.2 indicates a weak-to-moderate tendency for a positive deviation in one frame to be followed by a negative deviation in the next frame, and vice versa. This effect decays rapidly toward zero for higher lags, as shown in Figure 2. Together, these observations indicate short-term corrective dynamics rather than long-range sequential structure.

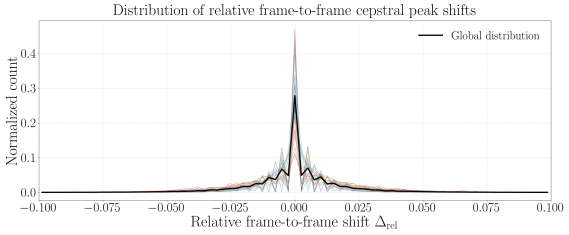


Fig. 1: Distribution of relative jitter values in hoarse sustained-vowel recordings, showing a sharp central peak near zero with symmetric, rapidly decaying tails.

***Corresponding author: H. Eikens, A. Culcu, K. Hollborn, G. Schmidt,** Christian-Albrecht-Universität zu Kiel, Kiel, Germany, e-mail: heei@tf.uni-kiel.de
I. S. Schiller, RWTH Aachen University, Aachen, Germany, e-mail: Isabel.Schiller@psych.rwth-aachen.de

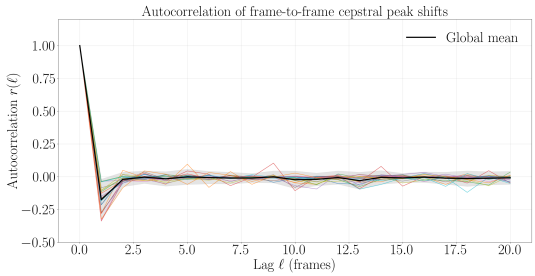


Fig. 2: Autocorrelation of frame-wise cepstral-peak jitter in hoarse sustained-vowel recordings, showing a pronounced lag 1 component of about -0.2 with rapid decay toward zero for higher lags.

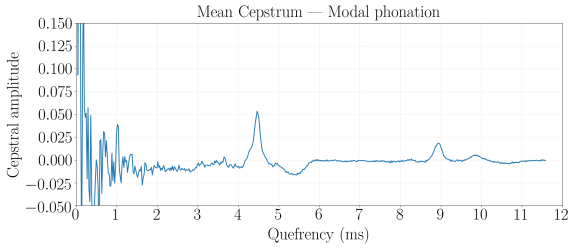


Fig. 3: Example of an excitation-related cepstrum obtained after LPC analysis. The dominant peak corresponds to the fundamental period and serves as the target for jitter-based modulation.

3 Method

The proposed method operates on the excitation-related cepstrum obtained from linear predictive coding (LPC) analysis and modifies the first cepstral peak to introduce jitter-like temporal instability. An example of such an excitation-related cepstrum is shown in Figure 3. The network is trained on normalized frame-wise relative jitter sequences extracted from sustained vowel recordings (/a, e, i, o, u/) in imitated hoarse phonation, produced by 31 speakers (11 female, 20 male) with no diagnosed voice disorders, corresponding to approximately 216 minutes of annotated speech material recorded at two sites under identical conditions. No speaker-specific or vowel-specific conditioning is applied; the model is trained to capture the shared temporal structure of the jitter process across speakers and vowels. The generated jitter values are integrated into a block-wise real-time cepstral peak-shifting pipeline. Speech is processed at 44.1 kHz. To satisfy the overall latency requirement of approximately 20 ms, a frame length of 1024 samples, a frame shift of 256 samples, and an FFT order of 1024 are chosen. For each frame, the first excitation-related cepstral peak is detected and its position serves as an estimate of the local pitch period. The network output is a normalized shift value that is applied to this detected peak position, introducing frame-wise jitter-like instability into the excitation signal. To predict the next shift value, the model uses a short history of previous frames. A systematic hyperparameter search over sequence lengths of $L \in \{3, 5, 10, 20, 30\}$ previous frames showed that longer histories did not yield consistent improvements; the final model therefore uses the last $L = 10$ frames of the normalized jitter sequence. The resulting compact architecture is

$$\mathrm{GRU}(64) \rightarrow \mathrm{GRU}(32) \rightarrow \mathrm{Dense}(32) \rightarrow \mathrm{Dense}(1).$$

The network is trained using the Huber loss, which combines the robustness of mean absolute error for large deviations with the smoothness of mean squared error near zero and is well suited to the sharply peaked jitter distribution observed in the data (see Figures 1 and 2).

Let $\Delta_{\mathrm{norm}}(n)$ denote the normalized frame-wise cepstral peak shift at frame n, and let $\hat{\Delta}_{\mathrm{norm}}(n)$ denote the corresponding network prediction. The residual between target and prediction is

$$e_{\mathrm{gru}}(n) = \Delta_{\mathrm{norm}}(n) - \hat{\Delta}_{\mathrm{norm}}(n). \tag{1}$$

To restore variability not captured by the deterministic GRU prediction, a stochastic term parameterized by the training residuals is added:

$$\tilde{\Delta}_{\mathrm{norm}}(n) = \hat{\Delta}_{\mathrm{norm}}(n) + \alpha\, \epsilon(n), \tag{2}$$

where $\epsilon(n) \sim \mathcal{N}(0, \sigma^2_{\mathrm{res}})$, σ_{res} is estimated from $e_{\mathrm{gru}}(n)$ on the training data, and $\alpha \geq 0$ controls the strength of the injected variability. This preserves the learned short-term temporal trend while better matching the spread of the target jitter process.

4 Analytical Evaluation

The analytical evaluation examined whether the modulated cepstrum ($\alpha = 1$) reproduces the short-term jitter properties previously observed in imitated hoarse vowels. As shown in Figure 4, the distribution of frame-wise cepstral peak shifts in the modulated signals remains sharply centered around zero, exhibits a symmetric shape with rapidly decaying tails, and matches the overall spread of the target hoarse distribution well. At the same time, the modulated signals appear to shift more regularly from frame to frame than the imitated hoarse reference, which is reflected in a more pronounced isolated peak at zero. Likewise, the autocorrelation function in Figure 5 reproduces the characteristic negative lag-1 component of approximately -0.2. As in the imitated hoarse recordings, the autocorrelation of the modulated signals remains close to zero for lags greater than 1 and fluctuates only within a narrow range of about ± 0.05, indicating that the dominant temporal dependency is restricted to very short time scales.

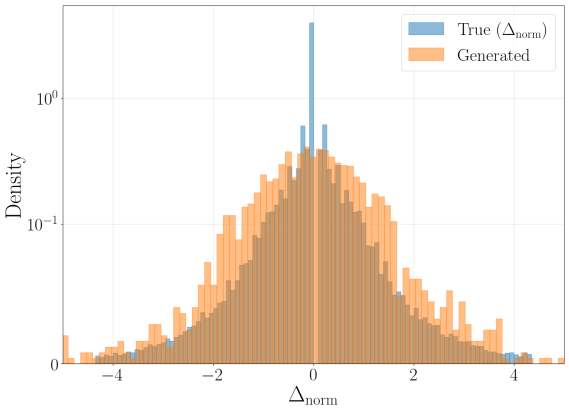


Fig. 4: Distribution of frame-wise cepstral peak shifts in the modulated signals compared with the target hoarse distribution.

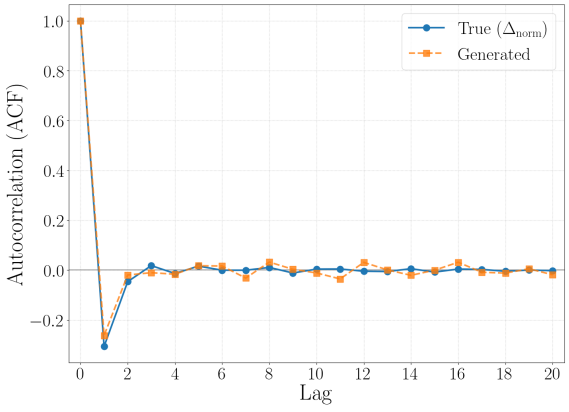


Fig. 5: Autocorrelation of frame-wise cepstral peak shifts in the modulated signals, showing the characteristic negative lag-1 component and values close to zero for higher lags.

Beyond these aggregated jitter statistics, peak-based cepstral measures also move in the expected direction. As shown in Figures 6 and 7, the speaker-normalized peak height decreases from a median of about 1.0 in modal speech to about 0.55 in the modulated condition, close to the imitated hoarse reference at about 0.5. Conversely, the normalized peak width increases from about 0.8 in modal speech to about 1.5 in the modulated condition, essentially matching the imitated hoarse reference. Thus, the modulation reproduces the characteristic hoarseness-related tendency toward a lower and broader main cepstral peak, while the resulting distributions remain somewhat more compact than in the imitated hoarse condition.

Limitations become visible in the time-resolved cepstral trajectories. While imitated hoarse vowels exhibit segmented and nonstationary temporal organization, including relatively stable intervals, more unstable regions, and abrupt local disruptions, the modulated signals tend to show more uniform and temporally homogeneous fluctuations. The target jitter process is therefore reproduced well on average, but the current cepstral modulation framework does not yet reproduce the more irregular longer-term behavior observed in imitated hoarseness.

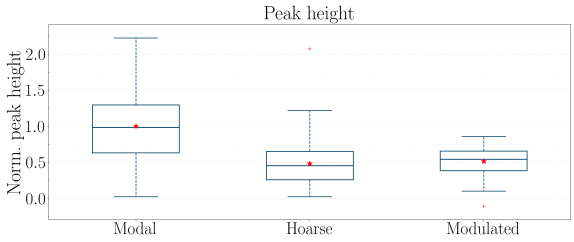


Fig. 6: Speaker-normalized peak height for modal, imitated hoarse, and modulated speech.

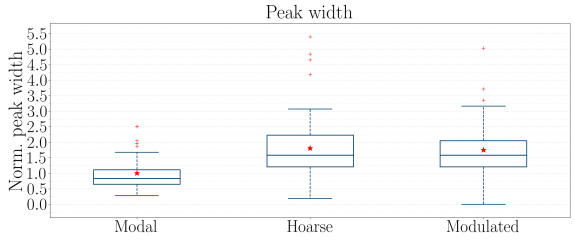


Fig. 7: Speaker-normalized peak width for modal, imitated hoarse, and modulated speech.

5 Perceptual Evaluation

A controlled within-subject listening study with 20 native speakers was conducted to assess whether the analytically matched jitter properties translate into perceptually convincing voice-quality changes [8]. Listeners rated randomized sustained vowel stimuli on 7-point Likert scales with respect to naturalness, technical artifacts, hoarseness, roughness, and breathiness under blinded conditions. The key conditions reported here are unmodified modal speech, imitated hoarse reference speech, and modulated speech ($\alpha = 1$).

The proposed modulation increased perceived hoarseness and roughness relative to modal input speech. Median hoarseness ratings increased from about 2 for modal recordings to about 3 for modulated recordings, while imitated hoarse recordings reached about 5.5. A similar pattern was observed for roughness, with median ratings increasing from about 2.5 to 3.5, compared with about 4.5 for imitated hoarse speech. Breathiness was also affected, increasing from about 2 to 3, whereas imitated hoarse recordings reached about 4.5. Overall, the perceptual effect moved the modified signals in the intended direction, but did not reach the severity of imitated hoarse reference speech.

Perceptual quality constitutes a further limitation. While the modulated signals remained interpretable as voice-quality variations rather than complete signal breakdown, listeners also reported reduced naturalness and increased technical artifacts. Technical-artifact ratings increased from about 2 for unmodified recordings to about 4.5 for modulated speech, indicating that perceptual quality degradation remains a substantial limitation of the current approach.

6 Discussion

The data analysis identified short-term corrective dynamics as the dominant structure of hoarse jitter, motivating a compact sequence model. The results confirm that the network captures this short-term temporal dependency and yields trajectories with realistic global distributional behavior. The main limitations arise instead from the representation and system constraints imposed by real-time cepstral processing. Several factors contribute to these limitations. FiFirst, the fixed frame size and latency requirement restrict the usable cepstral range, particularly for low-pitched voices, where realistic peak displacements approach the boundary of the accessible region. Second, the method acts mainly on the first excitation-related cepstral peak and cannot represent more complex harmonic instability patterns. Third, the excitation cepstrum provides a meaningful target mainly in voiced, sufficiently periodic regions, limiting applicability in unvoiced or transitional segments. Finally, the reconstruction stage introduces technical artifacts that reduce signal quality even when the intended modulation is preserved. In addition, the modulation may produce cepstral configurations that are not fully physically plausible or reconstructable, further contributing to audible artifacts. From a method-evaluation perspective, the key finding is therefore not that cepstral jitter modulation fails, but that its success is bounded.

7 Conclusion

This paper presented a real-time method for targeted cepstral jitter modeling using a lightweight GRU within an auditory feedback modulation framework. Motivated by data analysis of imitated hoarse sustained vowels, the method was shown to reproduce the global short-term jitter properties in the cepstral domain and to increase perceived roughness and hoarseness. At the same time, the results revealed fundamental limits. Although the sequence model is adequate for generating local jitter-like dynamics, the overall approach remains constrained by the cepstral representation, real-time processing requirements, and reconstruction-related artifacts. In view of these limitations, and given the comparatively high methodological effort required to achieve only a moderate modulation effect, the present cepstral real-time formulation does not appear to be the most suitable approach for introducing hoarseness in AFM paradigms. Future work should therefore investigate alternative signal representations or hybrid approaches that preserve real-time feasibility while providing better access to structured temporal variation and broader voice-quality characteristics [8].

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. I. S. Schiller's contribution to this study was supported by the Excellence Strategy of the Federal Government and the Länder (StUpPD445-23). Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Guenther FH, Hickok G. Role of the auditory system in speech production. Handbook of Clinical Neurology 2015;129:161–175.

[2] Schiller IS, et al. VQ-Synth: Development and perceptual evaluation of a system for voice quality modification. Proceedings of the 10th Forum Acusticum 2023; 2023:3533–3540.

[3] Schiller IS, et al. Enhancing acoustic voice quality through real-time auditory feedback modulation. Journal of Voice 2026.

[4] Schenck A, et al. The effect of pitch and loudness auditory feedback perturbations on vocal quality during sustained phonation. Journal of Voice 2023;37(1):37–47.

[5] Yanagihara N. Significance of harmonic changes and noise components in hoarseness. Journal of Speech and Hearing Research 1967;10(3):531–541.

[6] Childers DG, Skinner DP, Kemerait RC. The cepstrum: A guide to processing. Proceedings of the IEEE 2005;65(10):1428–1443.

[7] Eikens H. Design and Implementation of a Neural Network Based Auditory Feedback Evaluation System in Real-Time. MA thesis. Christian-Albrecht-Universität Kiel; 2025.

[8] Culcu A. Extension of a Real-time System for Auditory Feedback Modulation. MA thesis. Christian-Albrecht-Universität Kiel; 2026.