VoiceQualityGUI: A Tool for Word-Level Voice Quality Modifications

Harm Lameris [1], *Alan Villamil*[2], *Nigel G. Ward*[2]

[1] Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[2] University of Texas at El Paso, USA

lameris@kth.se, avillamilc@miners.utep.edu, nigelward@acm.com

Abstract

Voice quality profoundly affects the perceived pragmatic function and suitability of synthesized speech. However investigation has lagged, due to the lack of tools for exploration. To address this need, we have developed a tool that supports modification of creakiness, breathiness, and nasality down to the word level, using an improved version of the VoiceQualityVC voice conversion system. Users can manipulate these properties and experience how such local modulations alter the perceived conversational meanings.

Index Terms: voice quality, nasality, voice conversion, controllable speech synthesis, pragmatic functions

1. Motivation

Prosody is known to be important in conveying pragmatic intent. While the roles of some prosodic features — notably pitch, speaking rate, and intensity — have been well studied, most have not. In particular, glottal features — such as CPPS, harmonicity, and creakiness — although known to account for a significant amount of variance in pragmatic meaning [1], have been under-studied, as have their perceptual correlates, notably modal, creaky and breathy voice. For example, creaky voice can serve as a turn-taking cue and can be perceived as detached and uninterested, while breathy voice can be perceived as intimate and invested, among many other functions [2]. Nasality has also been shown to have a role in conveying various pragmatic functions.

To date, such findings have relied mostly on unaided observations for hypothesis formation, plus confirmation through controlled stimulus production by humans. Both are time-consuming. We wish to support, in addition, the quick formation and triage of early-stage hypotheses regarding the functions of voice quality features. As these features often bear meaning only in specific temporal configurations with words and other prosodic features, this requires us to support word-level modifications. We accordingly developed a simple tool to support such interactive explorations.

Within this context, our proximate motivation for developing this tool was both the success and the frustrations seen in [3]. In that work, participants used VoiceQualityVC to modify the voice quality of utterances synthesized for a video game to improve the pragmatic suitability. While that effort succeeded in demonstrating the value of control of voice quality in the post-editing of synthesized speech, it required the participants to tediously hand-edit specifications of the value of each property over each time range.

2. VoiceQualityGUI Tool

The graphical user interface (GUI) is shown in Figure 1. The supported workflow is similar to the workflow of [3] and is illustrated in Figure 2.

Before any manipulations happen, a zero-pass of the original audio is performed, that is, it is voice-converted using zero values for both the global pitch settings and voice quality settings. This is intended to provide an initial version of the voice-converted utterance with fairly generic prosody, as a baseline. From this starting point, the user next listens and infers how to adjust the global pitch settings to match the pitch and pitch variation of the original audio.

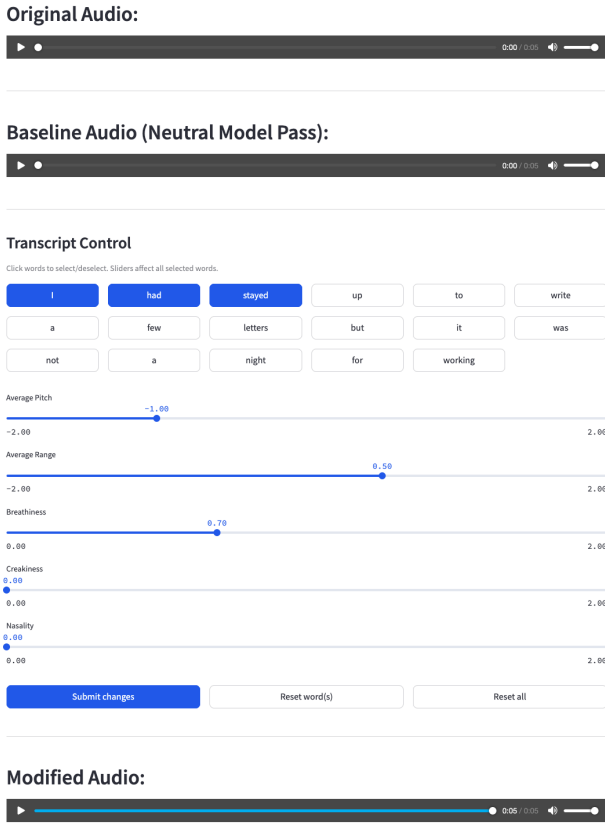

Figure 1: *A screenshot of VoiceQualityGUI. The example shows pitch lowered and pitch variation increase to match the speech of the original audio, and breathiness introduced for the first three words.*

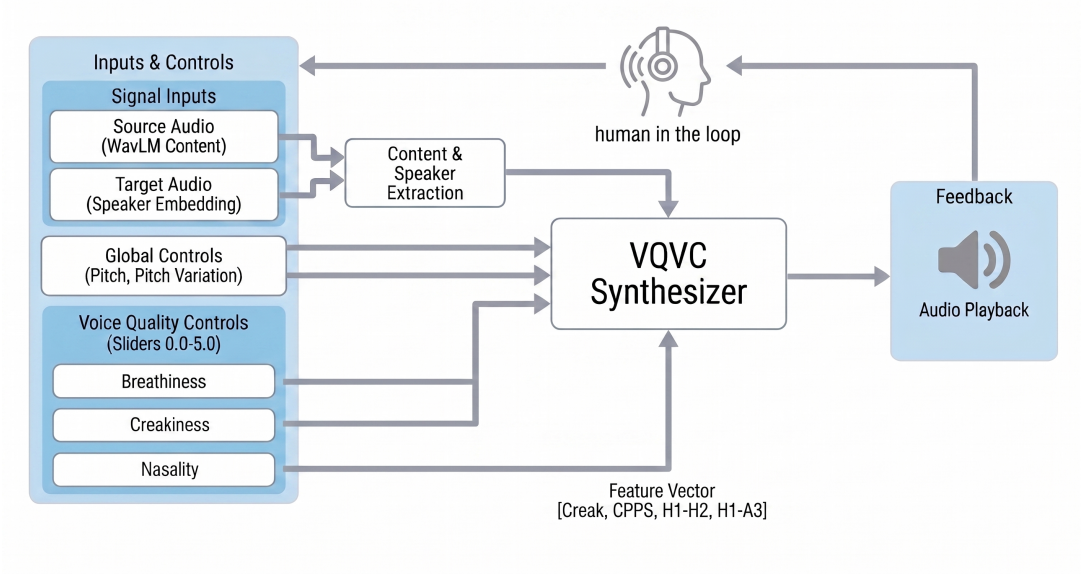

Figure 2: *The components and workflow of VoiceQualityGUI*

After these preparatory coarse adjustments, the user can locally adjust the creakiness, breathiness, and nasality on a word-level basis, to any desired degree of intensity, by selecting one or more words and adjusting the sliders. They then typically listen to the result, subjectively compare it to the original in terms of fidelity to the pragmatic functions conveyed, readjust, and repeat. To support subsequent analyses, the GUI saves the modified audio and the time-aligned adjustments of the intensity of each voice quality feature used to create it.

3. Implementation

This system leverages recent advances in the synthesis of speech with variable voice quality. In particular, it uses a fine-tuned version of VoiceQualityVC [2], a voice conversion tool for the modification of voice quality by directly conditioning on glottal source characteristics. Separate encoders, simple affine transformations, were added for each of four glottal source parameters, namely HNR35, CPPS, H1–H2, and H1–A3. We fine-tuned the open-source FreeVC checkpoint [4] for 45k iterations with a batch size of 32 and a learning rate of 5×10^{-5}, as explained in [2]. Since these parameters are acoustic and do not relate directly to perceptions, these features are exposed to users indirectly, in three combinations, designed to match perceptions of breathiness, creakiness, and nasality, respectively, as detailed in [2].

The data used by the model is the English-language part of the Expressive Speech corpus[1], as it contains several datasets with high expressivity, which we expected to include a wide variety of voice quality-related characteristics. We selected audiofiles which had a prosody score, an automated surrogate of expressivity that has a high correlation with human judgments, of up to 0.78. We annotated these for the four glottal source features at a per-frame level. These specific features were chosen as they correlate with perceptions of creaky and breathy voice quality. While nasality is less clearly related to these features, they were adequate to produce a fair approximation. The total duration of the audio files selected is about 17h20m. Pitch and pitch variation were annotated on a per-utterance basis. All features were z-standardized.

The GUI itself was created using the `streamlit` library. Its main inputs are (1) a path to the audio source file with the utterance that is to be modified, (2) a path to a target file featuring at least 30 seconds of speech of the target speaker, and (3) a time-aligned transcript of the speech in the source file.

[1]Z. Lin, J. Yang, J. Zhao, M. Liu, S. Li, and B. Wang, "Decoding the Ear: A Framework for Objectifying Expressiveness from Human Preference Through Efficient Alignment," arXiv preprint arXiv:2510.20513, 2025

4. Prospects

We have developed the first tool that supports research on voice quality by enabling the creation of directly comparable stimuli varying only in voice quality, while maintaining other aspects of the prosodic delivery, with a high degree of control. By providing a platform for systematic investigation, we hope that this tool will bring us closer to future systems capable of more controllable, or fully automated, intent-driven synthesis.

Acknowledgement: This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-24-1-0281.

5. Generative AI Use Disclosure

Generative AI tools have been used to aid in the creation of the figures, as well as for minor editing of the document.

6. References

[1] N. G. Ward, D. Marco, and O. Fuentes, "Which prosodic features matter most for pragmatics?" in *ICASSP*, 2025.

[2] H. Lameris, J. Gustafson, and É. Székely, "VoiceQualityVC: A voice conversion system for studying the perceptual effects of voice quality in speech," in *Interspeech*, 2025.

[3] H. Lameris and N. G. Ward, "Creakiness, Breathiness, and Nasality Contribute to the Perceived Suitability of Synthesized Speech in a Pragmatically-Rich Domain," in *Speech Synthesis Workshop*, 2025, pp. 89–95.

[4] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *ICASSP*, 2023.