Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

VoiceQualityVC: A Voice Conversion System for Studying the Perceptual Effects of Voice Quality in Speech

Harm Lameris[1], *Joakim Gustafson*[1], *Éva Székely*[1]

[1]Department of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

{lameris, jkgu, szekely}@kth.se

Abstract

Voice quality is an often overlooked aspect of speech with many communicative functions. Voice quality conveys both paralinguistic and pragmatic information, such as signalling speaker stance and aids in grounding. In this paper, we present VoiceQualityVC, a tool that can manipulate the voice quality of both natural and synthesized speech using voice quality features including CPPS, H1–H2, and H1–A3. VoiceQualityVC is a research tool for perceptual experiments into voice quality and UX experiments for voice design. We perform an objective evaluation demonstrating the control of these features as well as subjective listening tests of the paralinguistic attributes of intimacy, valence, and investment. In these listening tests breathy voice was rated as more intimate and more invested than modal voice and creaky voice was rated as less intimate and less positive. The code and models can be found at https://github.com/Hfkml/VQVC.

Index Terms: Voice conversion, Voice quality, Pragmatics, Paralinguistics

1. Introduction

The improved capabilities of voice assistants and voice-based chatbots have seen these tools become commonplace in daily life. People are increasingly using chatbots in social settings, for tasks and situations where the chatbot plays an advisory role [1], and for creative tasks [2]. For smooth, voice-based interaction, these situations demand subtle prosodic changes that are key in accurately conveying the pragmatic implication in a manner that reflects human speech [3], and understanding the underlying phenomena within these interactions is crucial.

Non-lexical aspects of speech, particularly paralinguistic cues, play a key role in communication. These cues signal relationships between speakers and listeners, convey stance on prior statements, and indicate levels of engagement. Such paralinguistic information is often expressed through variations in voice quality—i.e., changes in glottal phonation—including modal, creaky, breathy, tense, and harsh voice [4].

Perceptual studies paint a complex one-to-many mapping between voice quality and affect. A study using several synthesized voice qualities including creaky and breathy voice suggests breathy voice is perceived as more intimate and creaky voice being perceived as bored [5]. Breathy voice plays a role in grounding [6], and several studies echo the perception of breathy voice as intimate as well as indicating care [7, 8]. The perception of creaky voice is especially complex, and sometimes contradictory across groups. It can be evoke strong negative attributes, such as in [9] where creaky voice is judged to be *vain* and *uninterested*, while others in the same study regard creaky voice from female speakers to be *sophisticated* and *cool*. Other studies show creaky voice is used to distance the speaker from the lexical content in face-saving contexts [10] and to signal a detached stance regarding a previous utterance [11]. In parenthetical comments, however, it signals off-the-cuff remarks and inner thoughts [11], thus signalling a certain intimacy between the listener and the speaker.

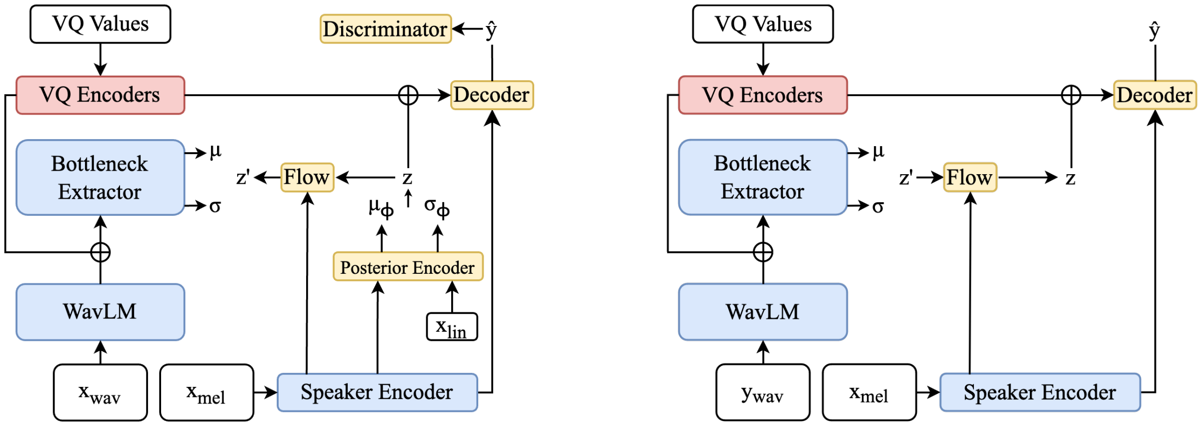
Recent work has focussed on the synthesis of different voice qualities, either by conditioning on voice quality features for creaky voice [12, 13, 14], or through learned perceptual ratings of voice quality ratings [15]. We present an open-source framework for the manipulation of voice quality by using five voice quality features: creakiness, cepstral peak prominence smoothed (CPPS), H1–H2, pitch (F_0), and H1–A3. The tool can play a role in filling the research gaps that persist in terms of the perceptual impact of voice quality by enabling the creation of near-identical stimuli with systematic variation of linguistic, prosodic, and extralinguistic characteristics. By gradually manipulating the voice quality features, the tool could also be useful for UX experiments for voice assistants or social robots by manipulating the target voice in an interpretable manner until the desired voice quality is achieved.

For this paper, we created stimuli in modal, creaky, and breathy voice qualities for a study on the perceptual differences between these voice qualities. Objective evaluation shows that the voice quality features were accurately reproduced. In a subjective listening experiment, we asked participants to rate the dimensions of intimacy, valence, and investment. Breathy voice was rated as more intimate, and more invested than modal voice, while creaky voice was rated less intimate and less positive.

2. Background

Voice quality refers to quasi-permanent characteristics of a speaker's voice [16]. We use the definition of voice quality in the narrow sense [17], that is, deriving solely from laryngeal activity (e.g. breathy, modal, creaky, harsh), and not the general acoustic colouring of a speaker's voice or the definition of voice quality used to describe pathological voices as measured by the GRBAS scale. In English, voice quality conveys a range of emotions, attitudes, and social cues. It carries paralinguistic information, such as emotional expression, as well as sociolinguistic cues about the speaker's identity and role. It also provides pragmatic signals related to intention, emphasis, stance, and certainty, which are key elements in subtle communication [6, 18]. The realization of voice quality differs between languages and between speakers. In [19], it is noted that voice quality is not absolute. What constitutes breathy for one speaker might constitute modal phonation for another. Specifically we focus on modal, creaky, and breathy voice, as these account for upwards of 90% of English speech [18].

10.21437/Interspeech.2025-902



(a)

(b)

Figure 1: *The architecture of the system during finetuning (a) and inference (b)*

2.1. Modal voice

Modal voice, as the name suggests, refers to the phonation type in which the constriction of the glottis is in between creaky and breathy phonation [19]. In English, modal voice is seen as unmarked, and it is by far the most used voice-quality [18]. Vowels in modal phonations have a measurable pitch, and human hearing is attuned to subtle pitch changes in modal voice [19].

2.2. Creaky voice

Creaky voice, also referred to as vocal fry or glottal fry, is a voice quality that uses more constriction of the glottis compared to modal voice [19]. Creaky voice tends to have low, often irregular F_0, a lower Harmonics to Noise Ratio (HNR), and smaller harmonic differences compared to modal voice [20]. In English, creaky voice be a prosodic cue used to mark phrase-boundaries [21] and turn yields [22].

2.3. Breathy voice

Breathy voice is a voice quality that is produced with low laryngeal effort and with only a slight amount of frication [16]. Acoustic measures of breathiness generally correlate only moderately with the perception of creaky voice including HNR and Cepstral Peak Prominence Smoothed (CPPS), i.e. the amplitude of the first rahmonic (N.B. energy peak im the cepstral domain) relative to the regression line over the log power cepstrum of a signal, indicative of its harmonic structure [6]. Yet CPPS is often used in the detection and analysis of breathy voice. Other measures for breathy voice include H1–H2 and H1–A3, which are both positively correlated with breathiness [19].

3. Method

3.1. Data

We used the publicly available LibriTTS-R [23], a restored version of LibriTTS for training. The dataset was annotated for five voice quality features: creakiness, Cepstral Peak Prominence smoothed (CPPS), H1–H2, pitch, and H1–A3. The voice quality features were chosen based on Esposito[1]who suggests that these features were the most successful at distinguishing breathy and modal phonation.

For creakiness, we used frame-level annotations of the dataset for creaky voice from CreaPy [24]. CreaPy is a threshold-based creak-detection tool that uses features including H2–H1 (inverse H1–H2), F_0, and residual peak prominence to annotate segments of speech for creaky voice. CreaPy computes the creak probability by thresholding these features optimized for the recall of creaky phonation. More details about the creakiness annotation may be found in [13].

We extracted utterance-level CPPS over voiced segments of the speech using Praat. CPPS is used as a measure of dysphonia in pathological voices as well as in phonetics [25] and correlates with both breathy and creaky voice [6]. The CPPS values were standardized per speaker.

Parselmouth [26], a python wrapper for Praat was used to annotate the frame-level H1–H2 (dB), also referred to as L1-L2, a spectral measurement obtained by subtracting the amplitude of the second harmonic (H2) from the amplitude first harmonic (H1) of the speech signal. H1–H2 is long-established as a voice quality feature to distinguish breathy [27], as well as creaky phonation from modal phonation [28]. Physiologically, lower values of H1–H2 correspond to a lower glottal Open Quotient (OQ), more constriction, and increased medial vocal fold thickness [29]. As H1–H2 is sensitive to recording conditions, the values were standardized per speaker.

Pitch (Hz) was extracted using the Wavelet Prosody Toolkit [30], which applies continuous wavelet transforms to capture pitch contours over time. For each utterance, pitch was computed at the frame level and then averaged to produce a single pitch value per utterance. These values were subsequently standardized across the entire dataset to allow for comparisons between speakers.

Parselmouth was used to extract frame-level H1–A3 (dB), a spectral measure calculated by subtracting the amplitude of the third formant (A3) from that of the first harmonic (H1). Physiologically, H1–A3 corresponds to the speed of the vocal fold closing phase [31], and it is frequently used to distinguish between breathy and modal voice qualities.

[1] C. M. Esposito, "The effects of linguistic experience on the perception of phonation," Ph.D. dissertation, University of California, Los Angeles , 2006.

3.2. Architecture

We present VoiceQualityVC, an architecture based on [13], in turn based on [32]. It consists of a conditional variational autoencoder-based (CVAE) voice conversion tool that is trained using adversarial training to directly output the converted waveforms. VoiceQualityVC uses a prior encoder that embeds audio using WavLM representations with a bottleneck extractor to represent the prosody and lexico-semantic information. These are then passed through a normalizing flow to learn a more complex distribtution. During training, a posterior encoder learns the posterior distribution using an RNN-based speaker encoder for the speaker information as wel as a linear spectrogram. During inference, however, this information is extracted from the speaker encoder and the learned prior distribution. An overview of the architecture can be found in Figure 1, with more details available in [32]. We add separate voice quality encoders, each consisting of an affine layer with the same dimensionality as the WavLM representation to encode the values for each voice quality feature at the frame-level. The encoded features are in turn added as conditioning to the WavLM embeddings before being passed through the bottleneck extractor. A downward projection of the same voice quality features are used as general conditioning before the decoder.

3.3. Training Procedure

We finetuned the weights from the open-source pre-trained FreeVC model on the complete LibriTTS-R dataset using the annotated voice-quality features. The pre-split training and validation set from LibriTTS-R were used. The voice quality encoders were zero-initialized, and the complete model was finetuned for 7k iterations on 4 24 GB NVIDIA GeForce RTX 3090 GPUs using a batch size of 32. The standard configurations from FreeVC without the spectrogram distortion-based data augmentation were used. The model has a total of 39,351,872 parameters, an increase of 12,160 parameters over FreeVC.

4. Experiments

4.1. Objective Evaluation

For the objective evaluation, we measured the values of the voice quality features to ascertain that the modifications behave as expected. The objective evaluation, thus, serves as a sanity check to establish objective grounds for the feature modification. While the voice quality features serve as correlates for voice quality, it cannot be ascertained whether the produced speech can be classified as creaky or breathy, as these features operate in concordance.

To perform the objective evaluation, we synthesized a single utterance 100 times and individually modified each of the voice quality features for the values [-3..3] in increments of 0.5. We used four different voices, a low male, a high male, a low female, and a high female voice as target voices to ensure the robustness of the model. As the features were standardized, the input corresponds to standard deviations from the mean. The specific feature values were then measured using CreaPy for creakiness, parselmouth for H1–H2, H1–A3, and pitch, and Praat for CPPS.

4.2. Subjective Evaluation

For the subjective evaluation, we synthesized 10 sentences using the conversational Chris voice on ElevenLabs using the multilingual v2 model. The voice was chosen for its conversational properties and since the synthesis quality was sufficient for voice conversion. The sentences were intended to be neutral for the following dimensions: intimacy, valence, and investment. An example sentence is: *I think you should know, I saw your name in the performance review summary*. The high male voice from the objective evaluation was used as the target speaker for the subjective evaluation using VoiceQualityVC. Each sentence was converted into three voice qualities: modal voice, creaky voice and breathy voice. Modal voice was achieved by listening for areas with creak, and modifying the feature values to have less creakiness and a higher CPPS. Creaky voice was achieved by using a high value for creakiness, generally 5 std. above the mean, a lowered CPPS, and a lower H1–H2 and H1–A3, generally -1 st.d. from the mean. Breathy voice was achieved with lowered creak (-1 st.d), lowered CPPS (-1 st.d), and high H1–H2 and H1–A3 (2 st.d.). All converted stimuli were manually examined and the feature values were changed until the intended phonation type was achieved. Separate subjective listening experiments were devised for each of the dimensions, and participants rated either modal and creaky voice or modal and breathy voice.

We recruited 25 participants for each test using Prolific[2]. These participants were paid £12.00 per hour, and took an average of 7.35 minutes to complete the listening test. For the *intimacy* dimension, participants were asked: rate how close the relationship between the speaker and the listener is based on the way the sample is said on a scale from 1 (Not at all close) to 7 (Very close), with the additional anchor of 4 (Somewhat close). For *valence*, participants were asked: rate the attitude of the speaker towards what he is saying on a scale from 1 (Very negative) to 7 (Very positive), with the additional anchor of 4 (Neutral). For *investment*, participants were asked: Rate how much the speaker cares about what he is saying on a scale from 1 (Does not care at all) to 7 (Cares a lot). Participants additionally received the anchor 4 (Cares somewhat).

5. Results

Table 1: *The results of the subjective evaluation.* ***Bold*** *indicates a significantly higher rating.*

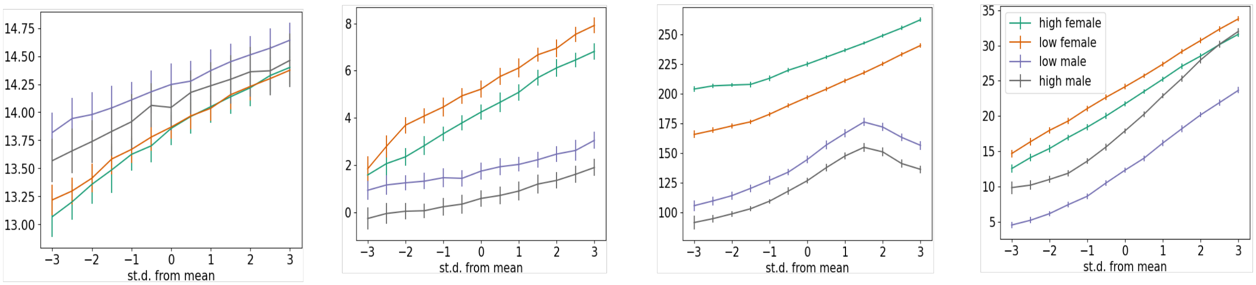
	Breathy	Modal	Creaky	Modal
Intimacy	**4.93±1.39**	4.65±1.44	4.42±1.54	**4.65±1.57**
Valence	4.23±1.55	4.20±1.44	4.01±1.29	**4.27±1.30**
Investment	**5.03±1.49**	4.71±1.55	4.72±1.35	4.85±1.35

5.1. Objective Evaluation

Figure 2 contains the results of the objective evaluation for each feature. Figure 2b and 2d show predictable control over the full range of of [-3, 3] standard deviations from the mean for both the male and the female voices. H1-H2 and H1-A3 are highly speaker and recording setup dependent, and therefore show different patterns, especially between the male and the female voices.

The results for CPPS and pitch may be viewed in Figure 2a and 2c. Although CPPS demonstrates control over the full range for both the male and the female voices, there are some irregularities for, especially for the high male voice. These irregularities are most likely caused by the fact that CPPS is generally measured over sustained vowels, whereas the measurements in

[2] `https://app.prolific.com`



(a) *CPPS (dB)*

(b) *H1–H2 (dB)*

(c) F_0 *(Hz)*

(d) *H1–A3 (dB)*

Figure 2: *The effect of input feature modification on the corresponding speech property in the output speech*

Figure 2a are over all voiced segments. For pitch, there is predictable control between [-3, 1.5] for the male voices and for [-1.5, 3] for the female voices. This can be explained by the fact that the pitch inputs are normalized over the complete corpus. Simply put, the male voices cannot extrapolate beyond 1.5 standard deviations since there are no male voices in that pitch range in the corpus, and vice-versa for the female voices.

5.2. Subjective evaluation

The mean ratings for each voice quality for each dimension are listed in Table 1. The results of a relative t-test subjective evaluation show that breathy voice was rated significantly more intimate ($p = .01$) and more invested ($p < .01$) than modal voice, while creaky voice was rated less intimate ($p = .04$) and less positive ($p < .01$) than modal voice. No significant differences were found between breathy and modal voice for valence and no significant differences were found between creaky and modal voice for investment.

6. Discussion

The results show that breathy voice was rated as more intimate and as indicating greater investment from the speaker. This is in line with the literature on breathy voice, e.g. [5, 7, 8] who found that breathy voice suggests greater intimacy and care. A suggested reason for this effect is given in [8], who claim that breathy voice indicates closer physical proximity, suggesting greater intimacy between speaker and listener. Similarly, breathy voice has been associated with self-directed speech, as well as for establishing common ground [6]. No significant difference was found between breathy voice and modal voice in terms of valence. This contrasts the finding by [5] in which breathy voice is associated with sadness and other attributes with mildly negative valence.

Creaky voice was rated as less intimate and less positive than modal voice, consistent with previous research. For instance, [5] suggest that creaky voice is perceived as conveying less interest and is often associated with boredom, echoing earlier claims by [16]. The lower intimacy ratings also align with findings by Butler[3] and [11], who both link creaky voice to a sense of detachment. However, no significant difference was found for investment. Interestingly, the reduced perception of intimacy appears to contrast with the function of parenthetical creak described by [11], in which creaky voice is used for off-the-record remarks and inner thoughts—contexts that typically suggest a more intimate speaker–listener relationship.

Our findings also replicate earlier results from synthesized speech studies. [33] reported that creaky voice generated via text-to-speech was rated as less positive than modal voice, which they attributed to the particular quality of creak in their corpus. The present study extends this line of work using VoiceQualityVC, a system capable of generating creaky voice from a more diverse corpus spanning multiple speakers, thus offering broader generalizability.

Despite VoiceQualityVC's ability to reliably manipulate voice quality features, it is not without limitations. The system exhibits inconsistencies in CPPS values, which may stem from the fact that CPPS is typically measured on sustained vowels. In this study, CPPS was averaged over voiced segments at the utterance level, which likely introduced minor irregularities. Additionally, the pitch values did not generalize between genders, possibly due to dataset-level standardization, which included both male and female speakers. As VoiceQualityVC models acoustic correlates of voice quality rather than voice quality itself, more phonetic analysis is recommended to ensure the naturalness of these modifications. Although all authors have received training in phonetics, systematic evaluation by phoneticians could help to ensure the validity of the conversions.

7. Conclusion

In this paper, we presented VoiceQualityVC, a voice conversion tool that explicitly models the voice quality features of creakiness, CPPS, H1–H2, H1–A3 as well as pitch. The tool can be used to produce utterances in both modal and non-modal voice qualities including breathy and creaky. An objective evaluation showed that the tool has intuitive control over the voice quality features. A subjective evaluation consisting of a listening experiment examined the perceptual differences between breathy and modal voice and creaky and modal voice. The results indicate that breathy voice was rated as more intimate and more invested than modal voice, while creaky voice was rated as less intimate and less positive than modal voice. These results are largely in line with the perception of breathy and creaky voice for natural stimuli and show the viability of using VoiceQualityVC for the creation of stimuli for perceptual experiments regarding voice quality. Future work could focus on implementing VoiceQualityVC into dialogue systems in order to enhance pragmatic competence by using breathy or creaky voice when prosodically appropriate.

[3] E. Butler, "The use of creaky voice in mitigating face-threatening acts," *Student Research Submissions*, vol. 164, University of Mary Washington, Fredericksburg, VA, 2017.

8. Acknowledgments

This research is supported by the Swedish Research Council project Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

9. References

[1] J. Wester, S. de Jong, H. Pohl, and N. van Berkel, "Exploring people's perceptions of LLM-generated advice," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100072, 2024.

[2] M. Skjuve, P. B. Brandtzæg, and A. Følstad, "Why do people use ChatGPT? exploring user motivations for generative conversational ai," *First Monday*, vol. 29, no. 1, 2024.

[3] S. Herment and L. Leonarduzzi, "The pragmatic functions of prosody in English cleft sentences," in *Speech Prosody 2012*, 2012.

[4] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *Proc. ICPhS*, 2003, pp. 2417–2420.

[5] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[6] N. Ward, A. Kirkland, M. Włodarczak, and É. Székely, "Two pragmatic functions of breathy voice in american english conversation," in *11th International Conference on Speech Prosody, Lisbon, Portugal, May 23-26, 2022*. International Speech Communication Association, 2022, pp. 82–86.

[7] N. Audibert, V. Aubergé, and A. Rilliard, "When is the emotional information? a gating experiment for gradient and contours cues," in *Proc. ICPhS*, 2007, pp. 6–10.

[8] L. Tsvetanova, V. Aubergé, and Y. Sasa, "Multimodal breathiness in interaction: From breathy voice quality to global breathy "body behavior quality"," in *Proceedings of the Proc. of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots—VIHAR*, 2017.

[9] C. Ligon, C. Rountrey, N. V. Rank, M. Hull, and A. Khidr, "Perceived desirability of vocal fry among female speech communication disorders graduate students," *Journal of Voice*, vol. 33, no. 5, pp. 805–e21, 2019.

[10] E. Butler, "The use of creaky voice in mitigating face threatening acts," in *Student Research Submissions (University of Mary Washington, Fredericksburg,VA),Vol.164*, 2017.

[11] S. Lee, "Creaky voice as a phonational device marking parenthetical segments in talk," *Journal of Sociolinguistics*, vol. 19, no. 3, pp. 275–302, 2015.

[12] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous TTS with neural HMMs," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] H. Lameris, J. Gustafson, and É. Székely, "CreakVC: A voice conversion tool for modulating creaky voice," in *Proc. Interspeech*, 2024, pp. 1005–1006.

[14] H. Lameris, M. Włodarczak, J. Gustafson, and É. Székely, "Neural speech synthesis with controllable creaky voice style," in *International Congress of Phonetic Sciences (ICPhS)*, 2023, pp. 3141–3145.

[15] F. Rautenberg, M. Kuhlmann, F. Seebauer, J. Wiechmann, P. Wagner, and R. Haeb-Umbach, "Speech synthesis along perceptual voice quality dimensions," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[16] J. D. M. Laver, "Voice quality and indexical information," *International Journal of Language & Communication Disorders*, vol. 3, no. 1, pp. 43–54, 1968.

[17] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[18] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," in *Annual meeting of the Berkeley linguistics society*, 2011, pp. 427–448.

[19] P. A. Keating and C. Esposito, "Linguistic voice quality," *UCLA Working Papers in Phonetics*, vol. 105, no. 105, pp. 85–91, 2007.

[20] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice." in *ICPhS*, 2015, pp. 2–7.

[21] L. Davidson, "The effects of pitch, gender, and prosodic context on the identification of creaky voice," *Phonetica*, vol. 76, no. 4, pp. 235–262, 2019.

[22] M. Włodarczak and M. Heldner, "Contribution of voice quality to prediction of turn-taking events," in *Proc. Speech Prosody*, 2022, pp. 485–489.

[23] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech*, 2023, pp. 5496–5500.

[24] M. Paierl, T. Röck, S. Wepner, A. Kelterer, and B. Schuppler, "Creapy: A python-based tool for the detection of creak in conversational speech," in *Proc. ICPhS*, 2023, pp. 1716–1720.

[25] Y. Maryn and D. Weenink, "Objective dysphonia measures in the program praat: smoothed cepstral peak prominence and acoustic voice quality index," *J. Voice*, vol. 29, no. 1, pp. 35–43, 2015.

[26] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[27] E. Fischer-Jørgensen, "Phonetic analysis of breathy (murmured) vowels in gujarati," *Annual Report of the Institute of Phonetics University of Copenhagen*, vol. 2, pp. 35–85, 1967.

[28] L. Davidson, "Perceptual coherence of creaky voice qualities," in *Proceedings of the 19th International Congress of Phonetic Sciences. Canberra, Australia: Australasian Speech Science and Technology Association Inc*, 2019, pp. 147–151.

[29] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2625–2632, 2012.

[30] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Comput. Speech Lang.*, vol. 45, pp. 123–136, 2017.

[31] C. Menezes, K. Maekawa, and H. Kawahara, "Perception of voice quality in paralinguistic information types," in *Proceedings of the 20th General meeting of the Phonetic Society of Japan, Special issue of the 80th Anniversary. Tokyo, Japan*, 2006, pp. 153–158.

[32] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[33] H. Lameris, É. Székely, and J. Gustafson, "The role of creaky voice in turn taking and the perception of speaker stance: Experiments using controllable TTS," in *Proceedings of LREC-COLING*, 2024, pp. 16 058–16 065.