The Role of Creaky Voice in Turn Taking and the Perception of Speaker Stance: Experiments Using Controllable TTS

Harm Lameris, Éva Székely, Joakim Gustafson
Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
{lameris, szekely, jkgu}@kth.se

Abstract

Recent advancements in spontaneous text-to-speech (TTS) have enabled the realistic synthesis of creaky voice, a voice quality known for its diverse pragmatic and paralinguistic functions. In this study, we used synthesized creaky voice in perceptual tests, to explore how listeners without formal training perceive two distinct types of creaky voice. We annotated a spontaneous speech corpus using creaky voice detection tools and modified a neural TTS engine with a creaky phonation embedding to control the presence of creaky phonation in the synthesized speech. We performed an objective analysis using a creak detection tool which revealed significant differences in creaky phonation levels between the two creaky voice types and modal voice. Two subjective listening experiments were performed to investigate the effect of creaky voice on perceived certainty, valence, sarcasm, and turn finality. Participants rated non-positional creak as less certain, less positive, and more indicative of turn finality, while positional creak was rated significantly more turn final compared to modal phonation.

Keywords: speech synthesis, voice quality, creaky voice, speech perception

1. Introduction

Advancements in generative modeling have significantly improved text-to-speech (TTS) technologies, allowing for the use of spontaneous data that reflects the intricacies of genuine human conversations.(Székely et al., 2019b). Spontaneous TTS models effectively reproduce characteristics absent in read speech, such as fillers (Székely et al., 2019a), and they extend the prosodic spectrum beyond the uniform patterns found in scripted speech corpora (Ben-David and Shechtman, 2021). A recent area of focus in spontaneous speech synthesis is the realistic generation of non-modal voice qualities, such as creaky voice (Lameris et al., 2023b,c). Voice quality, including phonation types such as creaky, breathy or tense voice, is used to convey pragmatic and paralinguistic information (Campbell and Mokhtari, 2003). Creaky voice can signal the end of a turn, and it serves to distance the speaker from the content of their message (Lee, 2015). This characteristic makes it a valuable tool for expressing notions such as certainty in TTS-based dialogue systems, with the potential to enhance the expressiveness and nuance of an interaction. To harness voice quality within communication technologies, a deeper understanding of the multifaceted relationship between voice qualities and their communicative roles is needed. This study, thus, investigates the effect of creaky voice on listeners' perception of speaker stance and turn taking behaviour.

2. Background

2.1. Properties of Creaky Voice

Creaky voice refers to several phonation types that generally exhibit a low rate of glottal fold vibration (F_0) that is often irregular (Laver, 1980) and is articulated with a constricted glottis, as well as low glottal airflow (Keating et al., 2015). Vocal fry, a term that is often used interchangeably with creaky voice (Dallaston and Docherty, 2020), displays similar characteristics, but has a regular F_0 (Keating et al., 2015).

The utility of creaky voice is diverse, encompassing a range of pragmatic, socio- and paralinguistic functions. Lee (2015) identifies five types of creak based on the function the creak serves. Among these, positional creak is the most commonly observed, serving as a prosodic cue to mark phrase finality, extending over multiple syllables or even words (Davidson, 2019). Acoustically, positional creak shares similarities with non-constricted creak, characterized by highly aperiodic and irregular vibrations (Keating et al., 2015; Lameris et al., 2023c). Positional creaky voice appears to play a role in turn taking. In Laver (1976), the use of creaky voice is claimed to signal a turn yield for speakers of Received Pronunciation when accompanied by a low falling intonation. Włodarczak and Heldner (2022) found that the speech before non-overlapping speaker changes had less modal phonation than those ending in turn holds.

In contrast, non-positional creak, which lacks positional constraints, exhibits acoustic properties akin to vocal fry (Lameris et al., 2023c; Keating et al., 2015). Non-positional creak serves various functions, including indicating parenthetical comments and humor, sometimes extending beyond textu-

This research was supported by the Swedish Research Council projects Connected (VR-2019-05003), Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ally parenthetical phrases (Lee, 2015). Additionally, creaky voice can signal the speaker's stance towards a preceding utterance. According to Lee (2015), the use of creaky voice in stance-taking often conveys a sense of detachment. This detachment is also evident when creaky voice serves as a hedging device in face-saving contexts (Butler, 2017), as it allows the speaker to distance themselves from the lexical content. The amount of creaky phonation present affects the nature of this perception. When a full utterance is uttered in a creaky voice, it is argued to convey a sense of "bored resignation" (Laver, 1980). Finally, Fónagy (1981) suggests that creaky voice is present in expressions of sarcasm or irony.

Several studies have been performed regarding the perceptual judgments related to creaky voice. The results, however, paint a complex picture with little consensus, especially regarding female creaky voice (e.g. Wolk et al. (2012)). In Yuasa (2010), the results suggest female creaky voice is perceived as *professional* and associated with young urban women. In Anderson et al. (2014) and Taylor et al. (2022), on the other hand, a negative impact of female creaky voice in terms of perceived intelligence and labour market prospects is described that appears to be less marked for male creaky voice (Greer and Winters, 2015). The results in Ligon et al. (2019) echo this complexity with creaky voice being rated as *vain* and *disinterested*, although some participants rated female creaky voice as *sophisticated* and *cool*.

2.2. Synthesis of Creaky Voice

In an early TTS study on the role of creaky voice quality in the perception of emotion and attitude, Gobl and Ní Chasaide (2003) found that creaky voice stimuli received low scores for *interest* and *happiness*. Several TTS papers have investigated the synthesis of creaky voice in the HMM era of speech synthesis (Raitio et al., 2013; Csapó and Németh, 2013b,a) to enhance the diversity and expressivity of synthesized speech. Recently, neural creaky voice synthesis has been achieved in two ways. In Lameris et al. (2023b) creaky voice was modelled implicitly using pitch, as a result of prosodic modification. Lameris et al. (2023c) trained a TTS system on a corpus with automatic creak annotations, and achieved natural-sounding creaky phonation.

While Lameris et al. (2023b,c) used experts to rate the presence of creaky voice, this paper examines the perception of positional and parenthetical (non-positional) creaky voice by listeners without formal training, and finds that presence of creak influences the raters' perception of certainty, valence, and turn finality. Samples can be found at: `www.speech.kth.se/tts-demos/lrec-creak/`.

3. Method

3.1. Data

We obtained our spontaneous speech corpus from a publicly available English-language multimodal multi-party dataset described in Kontogiorgos et al. (2018) named *AptSpeech*. This dataset consists of 15 multi-party interactions between a mediator, who is kept the same in all interactions, and two unique participants per session, in which the participants were tasked with designing an apartment on a large touchscreen with a GUI. The data was collected with the intended purpose of developing a social robot that could be used in collaborative tasks similar to the task presented to the participants.

Each interaction consists of four distinct parts. The mediator first made small talk with the participants about their experiences in shared living situations. The mediator then instructed the participants about the setup of the experiment highlighting the following aspects: The participants were to design an apartment where they would hypothetically live together for three months while being filmed for a reality television series. They had 70,000 crowns to purchase items to decorate the apartment, and the mediator would give advice as an interior decorator. Additionally, the mediator explained the operation of the GUI. During the extent of the experiment, the mediator gave advice concerning aspects related to interior decorating. Lastly, the mediator engaged in self-directed speech, for example when adjusting settings on the GUI. While the setup was identical in each of the sessions and the moderator was provided with a general outline and topics to cover, the interactions were unscripted. This structure results in spontaneous yet pre-planned extemporaneous conversations for each interaction with greater inter-session variation for the small talk, advice, and self-directed parts, and less variation for the instructions.

For this paper, the speech data described in Lameris et al. (2023a) were used. This consists of the mediator's speech data that were extracted and segmented into breath groups, sections of speech between two breaths, of 1-11 seconds. Those breath groups were transcribed using automatic speech recognition software. These transcriptions were manually corrected and annotated for filled pauses, audible breaths, turn-internal pauses and turn endings. The annotations were performed to be able to synthesize these speech phenomena as to influence the speaking style at inference, taking inspiration from Gustafson et al. (2021). To diversify the control of speaking styles and increase the stability of the synthesis, we supplemented the corpus with audio of the mediator reading 1129 of the CMU Arctic sentences

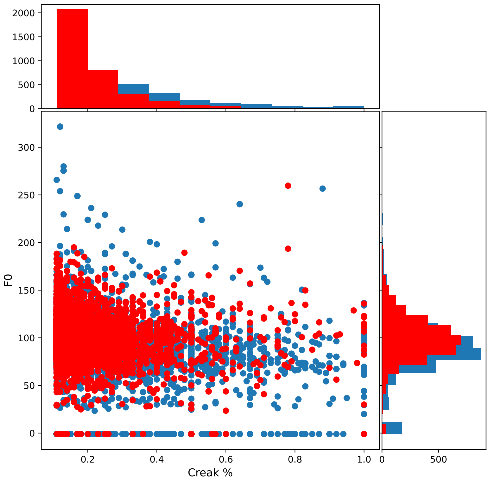

Figure 1: The per-word creak percentage and F_0 of the read-speech (red) and spontaneous-speech (blue) corpora.

(Kominek and Black, 2004) and 1132 sentences read from online newspaper texts. The complete corpus used has a duration of approximately 8 hours: 2h 26min of reading and 5h 40min of spontaneous speech.

3.2. Data Annotation

In previous studies on the role of creaky voice on turn taking, the voice quality has usually been manually annotated Ogden (2001). The corpus in the current study was automatically annotated for the use of creaky voice using the per-word percentage of creaky voice (creak percentage), which includes voiceless segments similar to Lameris et al. (2023c). Although creaky voice exclusively appears in voiced segments, the annotation of creak over voiced segments leads to a complex distribution. Per-word creak percentage is used in order to create a more learnable distribution. We extracted the durations of creaky voice using two publicly available python-based tools: DeepFry (Chernyak et al., 2022) and CreaPy (Paierl et al., 2023). DeepFry is a deep-learning based multi-head classifier trained to classify creaky voice, voicing, and pitch. We used the pre-trained model that was trained on both the nuclear and pre-nuclear datasets from the paper. CreaPy uses manually selected features to detect creaky voice including H2-H1, F_0, residual peak prominence, and zero crossing rate, optimized for recall.

We extracted word-level alignments using the Montreal Forced Aligner (McAuliffe et al., 2017) and calculated the creak percentage for each word for each detection method using Formula 1 in Appendix B. We aggregated the annotations from each creak detection method using the heuristic in Table 2 in Appendix B. Figure 1 shows the distribution of creak percentage for both the spontaneous and read speech compared to the measured F_0 values, and Figure 2 shows the creak percentage with the distance of the word to the sentence boundary.

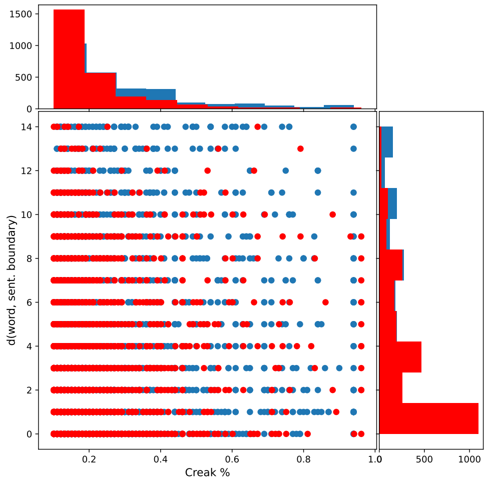

Figure 2: The per-word creak percentage and distance to the sentence boundary for the read-speech (red) and spontaneous-speech (blue) corpora.

3.3. Architecture

A modified version of the PyTorch implementation of Tacotron2 was used[1] (Shen et al., 2018). We added an 8-dimensional speaker-like embedding, similar to Valle et al. (2020), which is appended to each utterance's encoded text and passed on to the attention and decoder blocks of the model. This embedding serves to indicate whether the speech is spoken in a read style or a spontaneous style that constitute the two styles present in the corpus. Additionally we used a creak embedding consisting of the word-level creak percentage copied for each phone in the word, which is appended to the speaker-like embedding.

This model was initialised on a pre-trained read speech model and trained for 70k iterations with only the speaker-like embedding as published in Székely et al. (2023). The model was finetuned for a further 30k iterations including the creak embedding as well. The model's dimensions were extended for additional embeddings between these steps as per Székely et al. (2023). The model was trained on 4 NVIDIA GeForce RTX 3090 12 GB GPUs with batch size 28. 5% of the data was withheld as a validation set. We used a HiFi-GAN vocoder (Kong et al., 2020) finetuned on the same corpus for 383k iterations on top of the pre-trained model[2]. At inference, denoiser strength was set at 0.04. An overview of the architecture can be found in Figure 3.

[1] https://github.com/NVIDIA/tacotron2

[2] https://github.com/jik876/hifi-gan

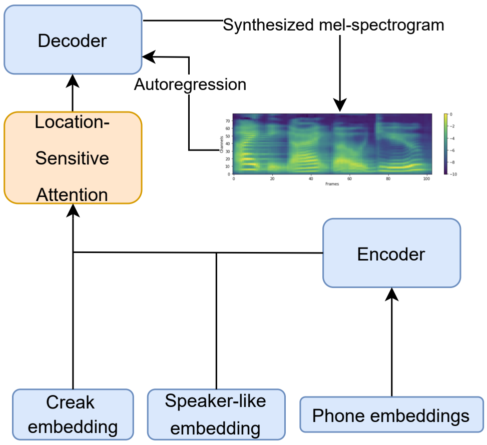

Figure 3: The model architecture.

4. Experiments

4.1. Experimental Setup

We conducted an objective analysis and two subjective listening experiments. The objective analysis aimed to quantify the creakiness present in the stimuli from the subjective experiments, while the subjective listening experiments were conducted to assess the participants' perception of creaky voice. The first subjective experiment focused on non-positional creaky voice, while the second explored positional creaky voice. In all experiments, we utilized the TTS engine described in section 3.3 to synthesize stimuli with the following phonation types: without creak (modal phonation), with non-positional creak, and with positional creak. More information about the synthesis can be found in Appendix A.

For the subjective listening experiments, participants were asked to listen to the generated stimuli and rate aspects of the perceived stance of the speaker as well as turn finality, phrased in the following manner: "How certain does the speaker sound" (certainty), "How positive does the speaker sound?" (valence), "How sarcastic does the speaker sound?" (sarcasm), "How likely is it the speaker has finished speaking?" (turn finality). Participants were informed that the speaker is responding to someone in a conversation about activities. The ratings were performed on a seven-point scale, with 1 being the lowest rating (i.e. very uncertain, very negative, not at all sarcastic, and very unlikely to have finished speaking) and 7 being the highest rating (i.e. very certain, very positive, very sarcastic, and very likely to have finished speaking).

4.2. Objective Evaluation

In the objective evaluation, we obtained the creak percentage for the stimuli that were used in the subjective listening tests using CreaPy (Paierl et al., 2023) to measure the extent that creaky phonation was present. It should be noted that this was calculated over the full stimulus including voiceless segments. The all-gender model and the off-the-shelf settings for creak-probability threshold and zero-crossing rate threshold were used.

4.3. Non-Positional Creak Experiment

In the non-positional creak experiment, 20 creaky and 20 non-creaky stimuli were synthesized, consisting of two phrases. The first phrase served an introductory purpose and was followed after 200ms with a phrase that in the creaky version was synthesized with non-positional creak to simulate parenthetical creak. The second phrase was ambiguous in semantic meaning as to whether the speaker was reacting sarcastically or genuinely. The non-creaky stimuli were chosen to approximately match the prosody of the creaky version. The following is an example stimulus with the potential creaky phonation in italics: (1) A trip to the desert, (2) *isn't endless sand and scorching sun a cool experience*? We recruited 25 native speakers of English living in majority English-speaking countries on Prolific[3], who were paid £4.00 and completed the experiment in an average of 16 minutes and 4 seconds.

4.4. Positional Creak Experiment

In the positional creak experiment, 20 creaky and 20 non-creaky stimuli were synthesized consisting of the same introductory phrase as in 4.3, followed after 200ms by a phrase that introduced semantic ambiguity regarding its intended purpose as a turn-yielding or a turn-holding cue. For the creaky stimuli, the final noun phrase was synthesized with positional creak, as indicated in italics. An example of a stimulus is: (1) A trip to the desert, (2) it's a way to experience *nature's beauty*.

We recruited 25 participants on Prolific using the same criteria as in 4.3 and completed the experiment in an average of 14 minutes and 54 seconds.

5. Results

Creak type	Creak percentage
No creak	0.04±0.03
Positional creak	0.09±0.06
Non-positional creak	0.13±0.08

Table 1: The per-utterance creak percentage and confidence interval for the stimuli, as obtained using CreaPy (Paierl et al., 2023)

[3] https://www.prolific.com

5.1. Objective Evaluation

Table 1 shows the objective evaluation results. A one-way ANOVA with post-hoc Tukey, showed a significant difference between positional and non-positional creak (p=.04), positional creak and no creak (p<.01), and non-positional creak and no creak (p=.03). The means and 95% confidence intervals are shown in the table.

5.2. Non-Positional Creak Experiment

The median ratings for each category can be found in Table 3 in Appendix C. A Wilcoxon signed-rank test showed a significant difference for certainty (p=.04) where the creaky stimuli were rated as less certain, valence (p=.0002) where the creaky stimuli were rated as more negative, and turn finality (p=.008) where the creaky stimuli were rated as more turn final.

5.3. Positional Creak Experiment

The median ratings for each category can be found in Table 4 in Appendix C. A Wilcoxon signed-rank test showed significance exclusively for turn finality (p<.0001) where the creaky stimuli were rated as more turn final.

6. Discussion

Our results indicate that participants without formal training in phonetics could perceive synthesized creaky phonation, and had a differing perception of non-positional creak, positional creak, and modal phonation. This suggests that there are clear perceptual differences between these types of phonation, highlighting the viability of using synthesized creak for perceptual studies about speaker stance.

The perception of non-positional creak as less certain and less positive than modal voice is in line with findings about the perception of creak as *bored resignation* in Laver (1980) as well as its association with boredom in Gobl and Ní Chasaide (2003). A possible reason for the perception of non-positional creak as less certain than modal voice is the usage of creaky voice as a hedge, in addition to its function of distancing the speaker from the speech (Butler, 2017; Lee, 2015). The finding of decreased valence also echoes Gobl and Ní Chasaide (2003). The result that creaky voice suggests less certainty seems at odds with studies by Ward (2006) and Lefkowitz and Sicoli (2007), which have described creaky voice as conveying an "authoritative" stance, often associated with greater certainty. A plausible explanation might be the context and purpose behind the speaker's use of creaky voice in the dataset, particularly when it's deployed in self-directed speech. In such instances, creaky voice could be interpreted as conveying uncertainty. Although non-positional creaky phonation was rated as less certain and less positive, it should be noted that the median certainty and valence ratings for creaky voice were still positive. This indicates that, although creaky voice decreases certainty and valence, it does not automatically imply uncertainty or negative valence.

Although there are some associations with creaky voice with rhetorical devices such as sarcasm, irony (Fónagy, 1981), and derision (Lee, 2015), our study's findings reveal no significant difference in how participants rated sarcasm between modal and creaky phonation. This finding suggests that, although creaky phonation often accompanies sarcastic remarks, it may not serve as the primary or exclusive cue for sarcasm perception. These results are in line with Yang (2021), who found the presence of prosodic correlates of creak, such as high jitter and a low harmonic-to-noise ratio in sarcastic utterances, but did not find a significant difference in the rating of sarcasm between creaky and modal voice.

Furthermore, we demonstrate that both positional and non-positional creak are perceived as indicators of turn completion. This observation extends beyond the commonly discussed phrase-final role of positional creaky voice (Lee, 2015) and its association with turn-taking mechanisms (Włodarczak and Heldner, 2022). The ambiguity surrounding the exact function of creak in signaling turn yields in English (Cutler and Pearson, 2018) becomes even more pronounced when considering non-positional creak. However, our findings, which resonate with those reported for German and Swedish by Włodarczak et al. (2023), propose that creaky phonation is perceived as turn-final whether occurring in the last voiced interval or spanning broader temporal domains. This is also in line with studies on creak as a turn-yielding cue in Finnish (Ogden, 2001).

7. Conclusion

In this paper, we investigated non-experts' perception of positional and non-positional creaky voice using neural speech synthesis. We annotated a conversational speech corpus for the presence of creak using two automatic creak detection tools. The presence of creak was objectively analyzed, showing three distinct categories for modal voice, non-positional creak, and positional creak. In two subjective listening experiments, non-positional creak was rated as less certain, less positive, and more turn final than identical stimuli synthesized with modal voice. Positional creak at the end of an utterance was rated as more turn final than modal phonation. Future work includes investigating the link between sarcasm, irony and creak, and how it interacts with other indexations associated with creaky voice. More research is required investigating sociolinguistic aspects of the perception of creaky voice, especially pertaining to gender.

8. Bibliographical References

Rindy C Anderson, Casey A Klofstad, William J Mayew, and Mohan Venkatachalam. 2014. Vocal fry may undermine the success of young women in the labor market. *PLoS one*, 9(5):1–8.

Avrech Ben-David and Slava Shechtman. 2021. Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis. In *Proc. SSW*.

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

Nick Campbell and Parham Mokhtari. 2003. Voice quality: the 4th prosodic dimension. In *Proc. ICPhS*, pages 2417–2420.

B.R. Chernyak, T.B. Simon, Y. Segal, J. Steffman, E. Chodroff, J.S. Cole, and J. Keshet. 2022. DeepFry: Identifying vocal fry using deep neural networks. In *Proc. Interspeech*, pages 3578–3582.

Tamás Gábor Csapó and Géza Németh. 2013a. Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):209–220.

Tamás Gábor Csapó and Géza Németh. 2013b. A novel irregular voice model for HMM-based speech synthesis. In *Proc. SSW*, pages 229–234.

Anne Cutler and Mark Pearson. 2018. On the analysis of prosodic turn-taking cues. In *Intonation in discourse*, pages 139–156. Routledge.

Katherine Dallaston and Gerard Docherty. 2020. The quantitative prevalence of creaky voice (vocal fry) in varieties of english: A systematic review of the literature. *PLOS ONE*, 15(3):1–18.

Lisa Davidson. 2019. The effects of pitch, gender, and prosodic context on the identification of creaky voice. *Phonetica*, 76(4):235–262.

Ivan Fónagy. 1981. Emotions, voice and music. *Research aspects on singing*, 33:51–79.

Christer Gobl and Ailbhe Ní Chasaide. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1-2):189–212.

Sarah DF Greer and Stephen J Winters. 2015. The perception of coolness: Differences in evaluating voice quality in male and female speakers. In *Proc. ICPhS*.

Joakim Gustafson, Jonas Beskow, and Éva Székely. 2021. Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. *Proc. SSW*, pages 48–53.

Patricia A Keating, Marc Garellek, and Jody Kreiman. 2015. Acoustic properties of different kinds of creaky voice. In *Proc. ICPhS*, pages 2–7.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. NeurIPS*, 33:17022–17033.

Harm Lameris, Joakim Gustafson, and É Székely. 2023a. Beyond style: Synthesizing speech with pragmatic functions. In *Proc. Interspeech*.

Harm Lameris, Shivam Mehta, Gustav Eje Henter, Joakim Gustafson, and Éva Székely. 2023b. Prosody-controllable spontaneous tts with neural HMMs. In *Proc. ICASSP*. IEEE.

Harm Lameris, Marcin Wlodarczak, Joakim Gustafson, and Éva Székely. 2023c. Neural speech synthesis with controllable creaky voice style. In *International Congress of Phonetic Sciences (ICPhS)*, pages 3141–3145.

John Laver. 1976. Language and nonverbal communication. *Handbook of perception*, 7:345–361.

John Laver. 1980. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186.

Sinae Lee. 2015. Creaky voice as a phonational device marking parenthetical segments in talk. *Journal of Sociolinguistics*, 19(3):275–302.

Daniel Lefkowitz and Mark Sicoli. 2007. Creaky voice: Constructions of gender and authority in american english conversation. In *106th annual meeting of the American Anthropological Association*.

Claire Ligon, Carrie Rountrey, Noopur Vaidya Rank, Michael Hull, and Aliaa Khidr. 2019. Perceived desirability of vocal fry among female speech communication disorders graduate students. *Journal of Voice*, 33(5):805–e21.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech*, volume 2017, pages 498–502.

Richard Ogden. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1):139–152.

Michael Paierl, Thomas Röck, Saskia Wepner, Anneliese Kelterer, and Barbara Schuppler. 2023. Creapy: A python-based tool for the detection of creak in conversational speech. In *Proc. ICPhS*.

Tuomo Raitio, John Kane, Thomas Drugman, and Christer Gobl. 2013. HMM-based synthesis of creaky voice. In *Proc. Interspeech*, pages 2316–2320.

J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, et al. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*, pages 4779–4783.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019a. How to train your fillers: uh and um in spontaneous speech synthesis. In *Proc. SSW*.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019b. Spontaneous conversational speech synthesis from found data. In *Proc. Interspeech*, pages 4435–4439. ISCA.

Bryn Taylor, Karen Wheeler-Hegland, and Kenneth J Logan. 2022. Impact of vocal fry and speaker gender on listener perceptions of speaker personal attributes. *Journal of Voice*.

Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *Proc. ICASSP*, pages 6189–6193.

Nigel Ward. 2006. Non-lexical conversational sounds in american english. *Pragmatics & Cognition*, 14(1):129–182.

Marcin Włodarczak and Mattias Heldner. 2022. Contribution of voice quality to prediction of turn-taking events. In *Proc. Speech Prosody*, pages 485–489.

Marcin Włodarczak, Mattias Heldner, Anna Bruggeman, and Petra Wagner. 2023. Voice quality dynamics of turn-taking events in Swedish and German. In *Proc. ICPhS*, pages 3477–3481.

Lesley Wolk, Nassima B Abdelli-Beruh, and Dianne Slavin. 2012. Habitual use of vocal fry in young adult female speakers. *Journal of Voice*, 26(3):e111–e116.

Seung-yun Yang. 2021. Listener's ratings and acoustic analyses of voice qualities associated with english and korean sarcastic utterances. *Speech Communication*, 129:1–6.

Ikuko Patricia Yuasa. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women? *American Speech*, 85(3):315–337.

9. Language Resource References

Kominek, John and Black, Alan W. 2004. *The CMU Arctic speech databases*.

Kontogiorgos, D. and Avramova, V. and Alexanderson, S. and Jonell, P. and Oertel, C. and Beskow, J. and Skantze, G. and Gustafson, J. 2018. *A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction*.

Székely, Éva and Wang, Siyang and Gustafson, Joakim. 2023. *So-to-Speak: an exploratory platform for investigating the interplay between style and prosody in TTS*. International Speech Communication Association.

A. Creak Values at Synthesis

Although the creak percentage that was supplied during training as the word-level creak embedding contained values between 0 and 1, these values were extrapolated between -1 and 3 at synthesis. The different types of creak were achieved in the following manner: the non-creaky phrases were synthesized with a creak percentage of -1 for each phone, the phrases with non-positional creak were synthesized with a creak percentage of 2 for each phone, and the phrases positional creak were synthesized with a creak percentage of 0 until the desired creak location and a creak percentage of 3 for the remaining phones. The speaker-like embeddings were set with a 0.1x weight on read speech, and a 0.9x weight on the spontaneous corpus.

B. Formulae and Heuristics

$$\text{creak percentage} = \frac{\text{total creak duration}}{\text{total duration}} \tag{1}$$

Creak Annotation	Chosen Value
DeepFry>0 & CreaPy>0	Highest value
DeepFry>0 or CreaPy>0	Highest if value>0.1 else 0
DeepFry=0 & CreaPy=0	0

Table 2: The heuristic used in the creak annotation

C. Detailed Results

	Non-positional creak	No creak
Certainty	5	**5**
Valence	4	**4**
Sarcasm	4	4
Turn finality	**6**	6

Table 3: The medians for the non-positional creak experiment. Bold indicates a significantly higher rating.

	Positional creak	No creak
Certainty	5	5
Valence	5	4
Sarcasm	2	2
Turn finality	**6**	6

Table 4: The medians for the positional creak experiment. Bold indicates a significantly higher score.