NEURAL SPEECH SYNTHESIS WITH CONTROLLABLE CREAKY VOICE STYLE

Harm Lameris[1], Marcin Włodarczak[2], Joakim Gustafson[1], Éva Székely[1]

[1]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Linguistics, Stockholm University, Stockholm, Sweden
lameris@kth.se, wlodarczak@ling.su.se, jkgu@kth.se, szekely@kth.se

ABSTRACT

The use of creaky voice, or vocal fry in speech has been extensively studied for its linguistic, paralinguistic, and sociolinguistic functions. However, much of the existing research on this topic is fragmented and often contradictory. In order to gain a deeper understanding of the communicative functions of creaky voice, we propose the use of comparative perceptual studies with natural sounding speech synthesis. We present a neural speech synthesizer that produces highly natural-sounding synthetic speech with controllable creaky voice styles. In a subjective listening experiment, speech experts were able to identify the presence and intensity of creaky voice produced by the synthesizer. Our results suggest that neural speech synthesis can be a valuable tool in furthering our understanding of the communicative functions of creaky voice.

Keywords: creaky voice, vocal fry, speech synthesis, TTS, spontaneous speech

1. INTRODUCTION

Creaky voice, also known as vocal fry, has been the subject of numerous linguistic, extralinguistic, and sociolinguistic studies in recent years. However, the diverse uses of creaky voice have often led to inconclusive results. Phonetically, creaky voice can refer to several types of voice production, including a low, irregular pitch with a low spectral-to-noise ratio [1]. It has a weaker first harmonic, as well as low harmonic differences, particularly between the first and second harmonic. There has been an increase in both intra-speaker usage of creaky voice in Finnish [2] and inter-speaker use in American English and Hiberno English, particularly among young women [3, 4, 5]. In terms of linguistic function, creaky voice varies widely across languages and can serve pragmatic functions such as turn-yielding in Finnish [6], signalling a phrase boundary for English [7], as well as semantic functions like providing an acoustic cue for tone in Vietnamese [8]. Creaky voice can be used to convey a wide range of attitudes and affective states [7, 9] including boredom [10], hesitation [11], sarcasm and disgust [12], and can signal romantic intentions [13]. However, not all of these results appear to be simple one-to-one mappings, with [9] pointing out that more research is needed to elucidate the relationship between creaky voice and affective state.

The sociolinguistic aspects of creaky voice for English also paint a complex and often contradictory picture. Extensive use of creaky voice carries strong social signals [14], and creaky phonation can be distinguished equally well for male and female voices by naive listeners. Despite the absence of gender disparity in the identification of creak, there appears to be a gender difference in the perception of it [3]. While earlier research indicates that the production of creaky voice by American women was perceived as "professional" and "urban-oriented" [15], more recently the perception of creaky voice has become overwhelmingly negative. In [3] vocal fry is identified as potentially harmful to women's career prospects.

The increased frequency of the occurrence of creaky voice in English, the large number of creaky voice functions, and lack of unequivocal results with regards to the sociolinguistic aspects of creaky voice highlight big gaps in the current research. This is exacerbated by the absence of a method to reliably and systematically produce creaky synthetic speech. We therefore propose a method that controls the synthesis of creaky voice ito facilitate systematic research into the sociolinguistic aspects of creaky voice. We modified a neural speech synthesizer [16], which is itself a modification of [17], to explicitly model the (non)presence of creaky voice. A corpus that extensively contains creak was analyzed for frame-level creaky phonation using DeepFry [18]. The frame-level creakiness annotations were aligned with word-level transcripts from the Montreal Forced Aligner (MFA) [19]. Two types of creak, stylistic creak and end-of-turn creak were chosen for a phonetic analysis, as these types commonly occurred in the corpus. In the analysis,

the creaky voice from the corpus closely resembled the synthesized creak for both types. Experts were asked to examine the word-level location of creaky phonation, and rate its intensity in a subjective listening experiment. While [16] implicitly model creaky voice in speech synhtesis, this is the first paper to explicitly model creaky voice using neural speech synthesis.

2. METHOD

2.1. Data

We used two speech corpora to train our models: RyanSpeech [20] and TSGD [21]. The RyanSpeech corpus is a scripted conversational corpus that was used for base training the speech synthesis model. It consists of 10 hours divided into 11,279 utterance spoken by a male speaker of American English speaking in a conversational style. The speaker reads aloud real-world conversational settings, task-oriented dialogues, supplemented by a selection of LibriTTS excerpts. This corpus contains only minute levels of creaky phonation.

Secondly, we used the audio recordings of the TSGD corpus [21], which were denoised using [22]. The corpus features a male speaker of Hiberno English and consists of 25 unscripted monologues of spontaneous speech of approximately 10 minutes each. The monologues are delivered in a colloquial style, and according to our observations the speaker makes extensive use of creaky phonation: as a stylistic feature to convey e.g. disinterest, as a phrase-final marker, and as a speech planning tool to conserve breath at the end of a respiratory cycle. The audio was transcribed and processed identically to [23, 24] by segmenting the corpus into breath groups, i.e. speech segments occurring between two breath events. The breath groups were combined into overlapping breath group bigrams to create audio files of up to 11 seconds long [25].

2.2. Creak detection

DeepFry, a neural network-based identification method for creaky voice was used to obtain the values to quantify creakiness [18]. It was chosen due to its substantially higher recall and comparable precision compared to other methods. Utterances which could not be parsed by DeepFry were excluded from further analysis. After the creak locations and durations were obtained, we used MFA to retrieve word duration alignments, which were used to calculate the percentage of creaky phonation per word, a measure we refer to as *creak value*. The *creak values* were used as additional input for the TTS model during training.

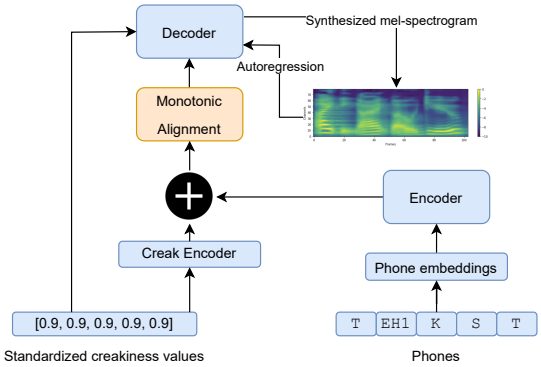


Figure 1: The model architecture

2.3. Model architecture

For the creaky voice synthesis, we modified an existing prosody-controllable sequence-to-sequence neural TTS architecture presented in [16], itself a modification of the TTS engine proposed by [17]. This architecture features an encoder-decoder model based on Tacotron 2 [26], while using a left-to-right no-skip neural Hidden Markov Model (HMM) to generate the alignments for the phone durations. The system by [16] allows for control of prosodic features such as f0 and speech rate on an utterance level. In order to achieve fine-grained control of voice quality rather than prosodic features, we modified the prosodic feature encoder from [16] to take word-level creak values as input, which we call *creak encoder*. The word-level creak values between 0 and 1 were standardized using the corpus-level mean and standard deviation, in order to create a more learnable distribution, and were copied for each phone. Copying the creakiness value for each phone instead of only the voiced phones creates a more general conditioning that allows the deep learning alogrithm to capture the general patterns of creaky phonation. In order to match the embedding size of the phone embeddings, the standardized values were passed through a feed-forward layer into a 512-dimensional control space. We used a two-step conditioning of the phone embeddings. First, we added the creak value projections to the output of the encoder. Secondly, we used a skip connection that appends the standardized creak values to each dimension of the decoder input that defines the final states of the HMM for more robust control (Figure 1).

2.4. Experimental setup

For the experiments, we trained a speech synthesizer using the speech corpora with creak annotations. As the TSGD corpus is too small for training a stable voice, we base-trained a voice on the RyanSpeech corpus for 28k iterations to increase the synthesis

stability before transfer learning on the TSGD voice for 4k iterations to learn the voice quality and creaky voice. Training on RyanSpeech does not change the voice quality after finetuning on TSGD. To analyze the synthesized creaky phonation, we performed two experiments: an acoustic analysis in which the creaky phonation present in the corpus and the creaky phonation produced by the synthesizer were compared, and a subjective listening test for which phoneticians and speech technologists were asked to mark the location and intensity of creaky phonation. To facilitate the comparison of the location of the creaky phonation for each condition, we divided each sentence in quarters based on the word count. This division was not visible to the participants. We compared the ratings for the general intensity of the creaky phonation by comparing the distribution of the intensity of the creak for the full utterances per condition. The location of the creaky phonation was compared by analyzing the creak ratings for the words in each quarter.

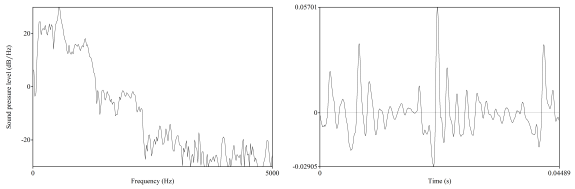
3. EXPERIMENTS

3.1. Acoustic analysis of synthesized creak

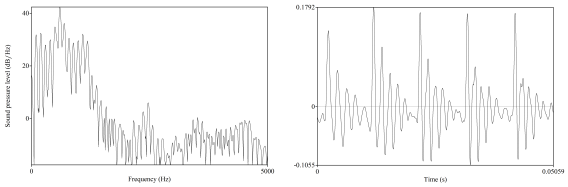
In order to compare the creaky phonation present in the corpus with the synthesized creaky phonation, two realizations of the /ɒ/ vowel preceding /t/ from the corpus with their synthesized equivalents were examined, one with end-of-phrase creak as voice quality and one with stylistic creak. The /ɒ/ vowel preceding /t/ was chosen, as it was a common combination in creaky segments. We selected a representative sample for each of the creak types and synthesized a matching sample for each creak type.

Figure 2 shows spectra and waveforms for each of the natural and synthesized vowels. Figure 2a and Figure 2c show the natural and synthesized end-of-phrase creak. The synthesized creak shows many similarities with the natural creak in the spectrum, both having weak harmonic structure and little energy for the higher frequencies. The waveform for the natural and synthesized end-of-phrase creak show high aperiodicity, which caused pitch tracking to fail for both the natural and synthesized speech.

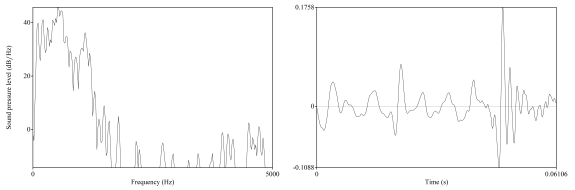
For Figure 2b and Figure 2d, showing stylistic creak on the other hand, the spectra show a clear harmonic structure with weak fundamental (i.e. low L1-L2). Additionally, it has more energy in the higher frequencies compared to the aperiodic creak. In terms of the waveform, both the natural and synthesized creak are periodic, albeit in a more regular manner for the natural creak. Pitch tracking retrieved the correct values for both examples.



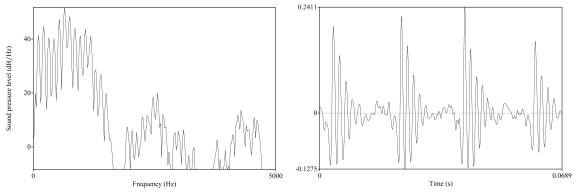
(a) The spectrum and aperiodic waveform of the natural /ɒ/ with end-of-phrase creak from [21].



(b) The spectrum and periodic waveform of the natural /ɒ/ with stylistic creak from [21].



(c) The spectrum and aperiodic waveform of synthesized /ɒ/ with end-of-phrase creaky phonation.



(d) The spectrum and periodic waveform of synthesized /ɒ/ with stylistic creak.

Figure 2: Spectra and waveforms for the natural and synthesized creaky phonation.

3.2. Subjective listening test

To examine the perception of synthesized creaky voice, 23 participants with expertise in phonetics or speech technology were presented iteratively with 36 stimuli. These stimuli comprised 12 sentences that were synthesized in three different voice qualities: no creak, i.e. modal phonation, stylistic creak with creaky phonation present throughout the utterance, and end-of-turn creak with creaky phonation present at the end of the utterance. The participants were provided with word-level transcripts for each stimulus and were asked to rate the presence of creaky phonation using a scale of 0—no creak, 1—creak, or 2—intense creak. Additionally, the participants rated the naturalness of the creaky phonation and could provide comments at the end of the experiment.

Figure 3 shows an overview of the ratings given by experts per creak condition per utterance quarter.

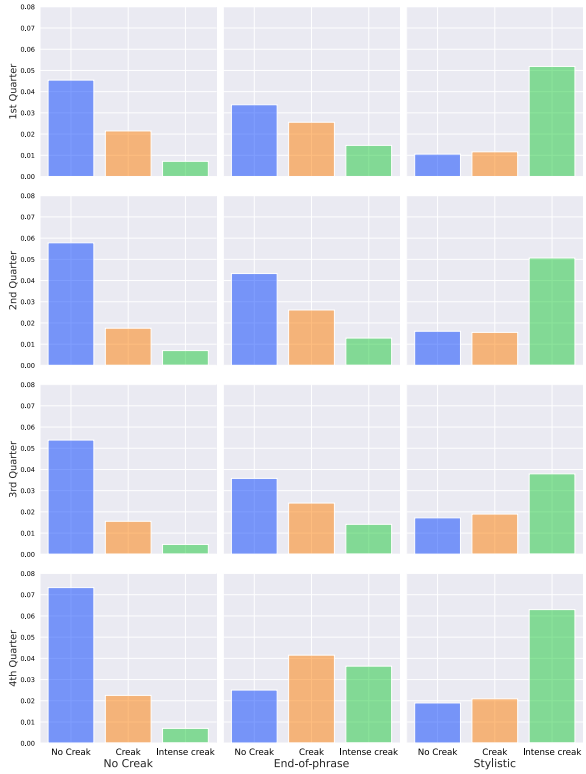


Figure 3: The distribution of the experts' ratings per condition and per quarter as percentage of the complete set of responses.

A Kruskal-Wallis H test showed significant results between the conditions for both the full utterance and for each quarter (all $p < 0.0001$). A post-hoc Dunn test with Bonferroni adjustment showed significant differences between every combination of conditions for both the full utterance and for each quarter ($p < 0.0001$). *Stylistic creak* was universally rated the strongest for each category with a median of 2 for the full utterance and each quarter ($p < 0.0001$). This was followed by *end-of-phrase creak* which had a median of 1 for the full utterance, the first, third, and fourth quarter, and a median of 0 for the second quarter, but had higher creak ratings than the no creak condition for each category ($p < 0.0001$). *No creak* had the lowest creak rating, with a median of 0 for full utterance and each quarter.

4. DISCUSSION

The acoustic analysis showed the synthesis of two types of creak: an end-of-phrase creak and a stylistic creak. The end-of-phrase creak was highly aperiodic with slow and irregular vibration. This type of creak is similar to non-constricted creak, as described by Keating [1], which also exhibits low and irregular f0, as well as slow irregular vibrations. It also showed many similarities to creaky phonation present in the corpus. The stylistic creak is more periodic, albeit with a very low f0. This creak is similar to vocal fry, as described in [1].

The ratings suggest that participants were generally able to distinguish *no creak*, *end-of-phrase creak*, and *stylistic creak*. Participants rated utterances with no creak as not creaky. Stylistic creak was rated as more intense and more present throughout the utterance than end-of-phrase creak.

Twenty-two of the experts rated the creaky phonation as natural or very natural for most utterances. Six participants noted occasional difficulty to distinguish creaky from modal phonation for stretches of speech with a high speech rate, partly due to vocoder artefacts that persist due to the nature of the data. Three participants highlighted the role of prosody in the rating of the utterances; participants had more difficulty distinguishing the phonation type for higher-pitched utterances. Two participants mentioned the influence of sociolinguistic and paralinguistic aspects in their perception of creaky voice.

5. CONCLUSION

In this paper, we present a novel approach to investigating the perception and communicative functions of creaky voice by utilizing neural speech synthesis. Our method opens up the possibility to systematically test the perception of creaky voice, and provides the opportunity to clarify, disambiguate, and build upon existing knowledge. We created a speech corpus with automatic annotation of creaky voice and modified a state-of-the-art speech synthesis architecture to explicitly model creaky voice. The synthesizer was trained on the speech corpus containing creak annotations. A phonetic analysis demonstrated that the two types of creaky phonation generated by the synthesizer closely resembled their natural counterparts found in the corpus. In a subjective listening test, experts successfully distinguished between the two creaky phonations and rated their creakiness as natural or very natural. This work reinforces the potential of neural speech synthesis as a valuable tool for advancing our understanding of the communicative functions of creaky voice and its sociolinguistic implications.

ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council projects Connected (VR-2019-05003), STANCE (VR-2020-02396), *Prosodic functions of voice quality dynamics* (VR-2019-02932), and the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

6. REFERENCES

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

[2] M. O'Dell, T. Nieminen, and L. Mustanoja, "Creak rate variation in individual speakers of Finnish," in *Proceedings of Nordic Prosody 2022*, in press.

[3] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatachalam, "Vocal fry may undermine the success of young women in the labor market," *PLoS one*, vol. 9, no. 5, pp. 1–8, 2014.

[4] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111–e116, 2012.

[5] R. Hickey, "Irish English in the anglophone world," *World Englishes*, vol. 36, no. 2, pp. 161–175, 2017.

[6] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[7] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech Language*, vol. 28, no. 5, pp. 1233–1253, 2014.

[8] J. Edmondson and N. V. Loi, "Tones and voice quality in modern northern Vietnamese: instrumental case studies." *Mon-Khmer Studies*, vol. 28, no. 35, pp. 1–18, 1997.

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

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

[11] M. Horne, "Creaky fillers and speaker attitude: data from Swedish," *Studies in Pragmatics*, pp. 277–288, 2009.

[12] K. Fischer and O. Niebuhr, "The effects of the online visualization of acoustic-prosodic features of speech on speakers' productions," in *Studientexte zur Sprachkommunikation: elektronische Sprachsignalverarbeitung*. TUD Press, 2022, pp. 180–187.

[13] J. Jelle, "A study in Scarlett: Creaky voice and romantic intention in Spike Jonze's Her," *Leviathan: Interdisciplinary Journal in English*, no. 1, pp. 35–44, 2017.

[14] 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.

[15] I. P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[16] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous TTS with neural HMMs," *Proc. ICASSP*, 2023.

[17] S. Mehta, É. Székely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *Proc. ICASSP*. IEEE, 2022, pp. 7457–7461.

[18] B. R. Chernyak, T. Ben Simon, Y. Segal, J. Steffman, E. Chodroff, J. Cole, and J. Keshet, "DeepFry: Identifying Vocal Fry Using Deep Neural Networks," in *Proc. Interspeech 2022*, 2022, pp. 3578–3582.

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[20] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," in *Proc. Interspeech*, 2021, pp. 2751–2755.

[21] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. IVA*, 2018, pp. 93–98.

[22] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, pp. 3291–3295.

[23] J. Gustafson, J. Beskow, and É. Székely, "Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," in *Proc. SSW*, pp. 48–53.

[24] É. Székely, G. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. SSW*, 2019, pp. 245–250.

[25] É. Székely, G. Henter J., Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *Proc. ICASSP*, 2020, pp. 7649–7653.

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.