Interspeech 2024
1-5 September 2024, Kos, Greece

CreakVC: A Voice Conversion Tool for Modulating Creaky Voice

Harm Lameris, Joakim Gustafson, Éva Székely

Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm
{lameris, jkgu, szekely}@kth.se

Abstract

We introduce a human-in-the-loop one-shot voice conversion tool called CreakVC designed to modulate the level of creaky voice in the converted speech. Creaky voice, often used by speakers to convey sociolinguistic cues, presents challenges to speech processing due to its complex phonation characteristics. The primary goal of CreakVC is to enable in-depth research into how these cues are perceived, using systematic perceptual studies. CreakVC provides access to a diverse range of voice identities exhibiting creaky voice, while maintaining consistency in other parameters. We developed a spectrogram-frame level creak representation using CreaPy and finetuned FreeVC, a one-shot voice conversion tool, by conditioning the speaker embedding and the self-supervised audio representation with the creak representation. An integrated plotting feature allows users to visualize and manipulate portions of speech for precise adjustments of creaky phonation levels. Beyond research, CreakVC has potential applications in voice-interactive systems and multimedia production.

Index Terms: creaky voice, TTS, voice conversion

1. Introduction

In recent studies, spontaneous text-to-speech (TTS) has been used to synthesize stimuli for studies on speech perception (e.g. [1]. TTS built using recordings of spontaneous conversational speech, with added control over prosodic features such as F_0 and speech rate, can be used to synthesize spontaneous speech phenomena not present in the commonly used read-speech corpora, such as hesitations and repetitions. Spontaneous TTS has benefits over both corpus-based studies and stimuli created with actors. It offers greater control over semantic content and prosodic delivery compared to traditional corpus-based approaches. Unlike acted stimuli, spontaneous TTS is data-driven and can more closely represent speech phenomena, because it samples from a distribution of spontaneous speech. In controllable TTS, it is also possible to vary prosodic features independently of each other, which can be a challenging task even for a trained actor.

Historically, the primary limitation of spontaneous TTS has been its limited voice diversity, largely due to the resources required for accurate annotation of spontaneous speech data. This has often confined perception studies to TTS models trained on recordings from a single speaker. An effective way to expand the use cases of spontaneous TTS, and thereby broaden the scope of perception studies, is to apply voice conversion techniques to the synthesized stimuli. Advancements in voice conversion now allow for one-shot transformation of synthetic speech, requiring only about 30 seconds of natural speech from the target speaker. One aspect that is generally not maintained in conversion, however, is the presence of creaky voice quality. Creaky voice, also referred to as vocal fry, is a phonation type which involves lowered subglottal pressure, leading to a low-pitched, crackling sound caused by irregular and slow vibrations of the vocal folds. It fulfills a variety of roles in speech, including phrase boundary marking and turn yielding, it can signal speaker stance as well as a range of emotional states. Creaky voice has seen renewed interest in speech synthesis [2]. Some results indicate that creaky voice is perceived differently for male and female voices, highlighting the need to perform perception studies on a range of voices. In order to generate creaky voice and at the same time maintain the controllability of spontaneous TTS while greatly diversifying the available voices, we combined spontaneous TTS with voice conversion. We modified an open-source one-shot VC tool called FreeVC [3] by conditioning the self-supervised representations and the decoder with a creak embedding.

In our demo, we present a human-in-the-loop voice conversion tool where the user can vary the amount and duration of creaky voice. It has integrated creak detection on the output using CreaPy. The creak probability over time and per word can be plotted, which can then be used to intuitively adjust the supplied creak vector in order to change the amount of creaky phonation present. This can be done in an iterative manner until the desired level and quality of creaky voice is reached. Apart from its primary intended purpose of creating stimuli for researching the perception of creaky voice in different contexts, CreakVC can be used in a variety of other applications such as entertainment and media production for enhancing character voices in films and games, speech therapy tools to assist in vocal training, language learning platforms to demonstrate speech nuances, and voice acting tools to enable a range of expressive possibilities without physical exertion. CreakVC can also enhance robustness for a range of voice-interactive systems, to help ensure that they effectively handle diverse speech patterns.

2. CreakVC Tool

2.1. Data

The clean subsection of the device-recorded CSTR VCTK corpus [4] was used for finetuning FreeVC, featuring 30 speakers and a total of audio length of 9h30 minutes. We reserved two speakers for the validation set. Creaky phonation was annotated at the spectrogram-frame level using CreaPy [5], a creak-detection tool that uses manually selected features including H2-H1, F_0, residual peak prominence, and zero crossing rate to calculate the probability of the presence of creaky voice. The standard configuration of CreaPy was used for all speakers. The annotations consisted of the duration of creaky phonation per the total duration of the spectrogram frame (20 ms).

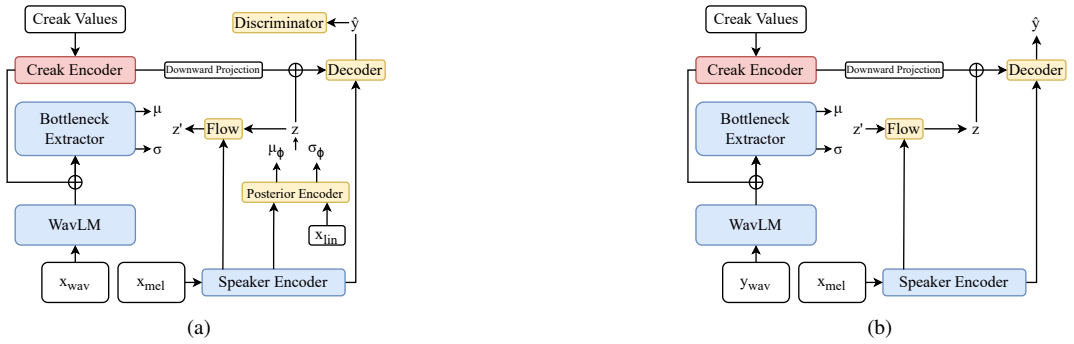

Figure 1: *The architecture of the system during finetuning (a) and inference (b)*

2.2. Architecture

We modified FreeVC, a voice conversion tool that utilizes pre-trained WavLM representations to learn a linguistic content representation and employs a pre-trained LSTM-based speaker encoder to encode speaker information. Functioning as a conditional variational auto-encoder with adversarial training, FreeVC models linguistic content conditioned on a speaker embedding. FreeVC incorporates a prior encoder where WavLM representations are bottlenecked to a lower-dimensional space. This allows it to learn the latent representation of the linguistic content distribution given by $N(z'; \mu_\theta, \sigma_\theta^2)$, while discarding the speaker information and noise. Additionally, FreeVC employs a normalizing flow from the posterior to make the learned prior distribution more complex. At inference, only the representations from the prior encoder are used together with the speaker embedding to generate the output.

We added a creak encoder composed of two affine layers. The first layer projects the creak embeddings into a 1024-dimensional space to align with the outputs of the WavLM model, to which the creak embeddings are added. The second layer performs a down-projection to a 192-dimensional space to match the dimensionality of the inputs to the decoder, and it is added to them.

The architecture during training can be found in Figure 1a, and the process at inference is depicted in Figure 1b. The creak encoder's weights and biases were zero initialized. We fine-tuned the pre-trained FreeVC model with pre-trained speaker encoder for 90k iterations with a batch size of 32.

Figure 2: *A plot of the creak probability (blue) and creak threshold (red) over time of the converted p232_180 from VCTK with zero creak values (a) and with increased creak at the end. (b).*

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).

2.3. Demo

CreakVC includes a human-in-the-loop tool that allows the user to perform voice conversion on both synthetic and natural data and alter the amount of creaky phonation present in the speech that can be run as a Jupyter Notebook. The tool presents an iterative process for the creation of speech samples consisting of voice conversion, to create stimuli for a large number of voices and analysis, to quantify the amount of creaky phonation and change the supplied creak tensor according to these values.

The voice conversion can be performed by indicating the required model paths and supplying the audio from the source- and target speakers as well as a creak tensor. The creak tensor provides the creak conditioning in the model. It is recommended to start with zero values for the complete utterance, as the amount of creak present in the converted speech varies from speaker to speaker. By using the analysis tools and by listening, the values in the creak tensor can be adjusted depending on the requirements.

The analysis consists of running WhisperX [6] to generate TextGrids which, together with the audio, are used to obtain the creak probability from CreaPy. These can then be plotted as shown in Figure 2a. The results can be used to adjust the creak tensor. Figure 2b shows the converted audio after decreasing the creak between 0.4s–1.4s and increasing the creak from 2.0s–2.8s in order to generate positional (end) creak.

It is our hope that CreakVC will contribute to a deeper understanding of the function of creaky voice in speech and the role speaker characteristics might play in its perception.

3. References

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

[2] H. Lameris, M. Wlodarczak, J. Gustafson, and É. Székely, "Neural speech synthesis with controllable creaky voice style," in *ICPhS*, 2023, pp. 3141–3145.

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

[4] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[5] 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.

[6] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Proc. Interspeech*, 2023, pp. 4489–4493.