Data Augmentation for Diverse Voice Conversion in Noisy Environments

Avani Tanna, Michael Saxon, Amr El Abbadi, William Yang Wang

University of California, Santa Barbara

Avani@ucsb.edu, saxon@ucsb.edu, amr@cs.ucsb.edu, william@cs.ucsb.edu

Abstract

Voice conversion (VC) models have demonstrated impressive few-shot conversion quality on the clean, native speech populations they’re trained on. However, when trained on accents, background noise conditions, or microphone characteristics differ from training, quality voice conversion is not guaranteed. These problems are often left unexamined in VC research, giving rise to frustration in users trying to use pretrained VC models on their own data. We are interested in accent-preserving voice conversion for name pronunciation from self-recorded examples, a domain in which all three of the aforementioned conditions are present, and posit that demonstrating higher performance in this domain correlates with creating VC models that are more usable by otherwise frustrated users. We demonstrate that existing SOTA encoder-decoder VC models can be made robust to these variations and endowed with natural denoising capabilities using more diverse data and simple data augmentation techniques in pretraining.

Index Terms: voice conversion, robustness, data augmentation

1. Introduction

Voice conversion (VC) is the task of generating utterances in a target speaker’s voice that carry the content and prosody from a source utterance from a different speaker [1], preserving the content of the source utterance while reproducing the characteristics and style of the target speaker. VC was originally conceived of as a data-efficient way to add styles and personalities to 90s-era text-to-speech systems, to improve the quality of decoded speech in telephony, and as a way to preserve speaker individuality under speech translation [2]. Many of these problems have been solved with other techniques, and VC has come to be treated as a novelty task for demonstrating innovations in data-efficient and few-shot generative modeling [3], and real-world use-cases for VC technologies are no longer centered. However, niche real-world applications for VC still exist. One example is multicultural name pronunciation (e.g., in graduation ceremonies) where the desired standard is for one’s own name to be read aloud as one pronounces it themselves. Typically, ceremony organizers solicit self-recordings of awardees voices, typically produced on their own cell phones. We propose treating these recordings as source utterances for conversion into the organizer’s target voice. Under these conditions, represented phonemes differ considerably from those present in the VC model training speech distribution and utterances have inconsistent microphone characteristics and environmental noise. Unfortunately, we find that existing pretrained few-shot VC models, such as AutoVC [3] and FragmentVC [4] perform poorly in these conditions. Is it possible to adapt these SOTA VC models to be performant in this setting? Yes, we find.

1.1. Related Work & Existing Problems

Encoder-decoder voice conversion models such as AdaIN-VC [5], AutoVC [3], and AutoPST [6] have recently set the standard for high-quality many-many voice conversion. A current state-of-the-art model in this vein, FragmentVC [4], achieves high-quality parallel-data-free VC results using the pretrained Wav2Vec [7] speech encoder to characterize the source speech. It is challenging to document how to record a quality source utterance that will work well with a VC model produced on an unnatural distribution. We found that OOD sampling rate, volume levels, microphone quality, and clarity of speech due to distance all had significant impacts on output quality for the aforementioned models. This leads to significant frustration for users trying to use open-source VC models on their own data, evidenced by a litany of GitHub issues. Furthermore, these models are trained on speech from native English speakers (UK/US accents), which also was problematic for our target task. We chose FragmentVC [4] for adaptation because of its use of the pretrained Wav2Vec encoder, which we believed would be better able to handle the diverse set of source utterance phonemes.

1.2. Contributions

We find that yes, this adaptation is easy. Simply by pretraining on a diverse sample of accented English speech from CommonVoice [8] under a variety of input noise conditions, FragmentVC [4] is able to simultaneously convert and denoise diverse speech, producing clean output speech in the target speaker’s voice, while preserving the accent and content of the source utterance. We open-source our noising data-augmentation scripts and release our noise-robust FragmentVC checkpoint.

Figure 1: Speakerwise mean quality Likert score histograms for baselines and our robust models CV-finetune and CV-VCTK.

![Figure 1: Speakerwise mean quality Likert score histograms for baselines and our robust models CV-finetune and CV-VCTK.](image-url)

---

2. https://github.com/avanitanna/RobustFragmentVC#checkpoints

---

2024
We add a noising module to augment data during training with randomized effects including volume level changes, adding background noise, simulating room reverberation effects and simulating compression artifacts from telephony. These effects are randomly sampled as data is loaded for feature extraction. We train FragmentVC from scratch using a more comprehensive dataset—a combination of CommonVoice and VCTK corpus (checkpoint CV-VCTK)—and perform finetuning of the existing FragmentVC using only CommonVoice (CV-finetune). Both checkpoints are trained for 1 million steps. FragmentVC was optimized with AdamW optimizer and lr 1e-5.

To assess the robustness of the model, we use a test suite consisting of name clips with multilingual speakers with varying accents. We ensure that the files represent real-world challenges with audio data and train FragmentVC to better adapt to these challenges as well as accent variations. We make a more useful checkpoint for real-world, arbitrary, multi-lingual users which can enable them to use such VC models more easily with real-world data.

4. Conclusion

We demonstrate denoising capabilities in FragmentVC, providing a denoising objective is used at train time. We address difficulties faced by users in replicating these models and their performance with their own data. We consider real-world challenges with audio data and train FragmentVC to better adapt to these challenges as well as accent variations. We make a more useful checkpoint for real-world, arbitrary, multi-lingual users which can enable them to use such VC models more easily with real-world data.

5. References


