INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

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 source or target speech 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 themself. 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.

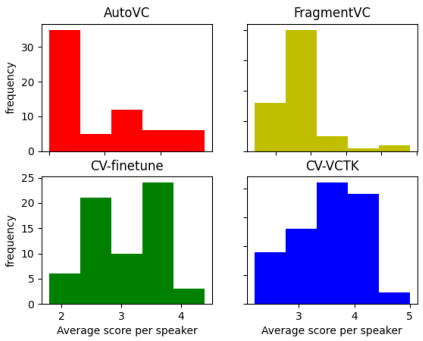


Figure 1: *Speakerwise mean quality Likert score histograms for baselines and our robust models CV-finetune and CV-VCTK.*

1.1. Related Work & Existing Problems

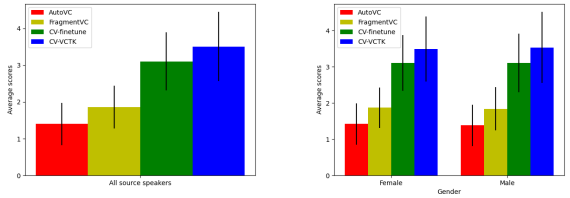
Encoder-decoder voice conversion models such as AdaIN-VC [5], AutoVC [3], and AutoPST [6] have recently set the standard for high-quality many-many voice conversion. A current state-of-the-art model in this vein, FragmentVC [4], achieves high-quality parallel-data-free VC results using the pretrained 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[1]. 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.

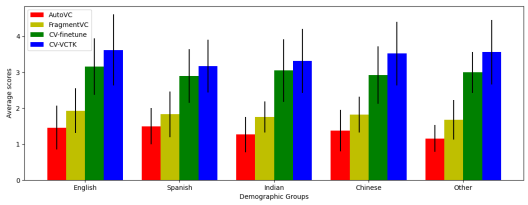
1.2. Contributions

We find that **yes, this adaptation is easy**. Simply by pretraining on a diverse sample of accented English speech from Common-Voice [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.[2]

[1] github.com/auspicious3000/autovc/issues/108, /28, /19, /14
[2] https://github.com/avanitanna/RobustFragmentVC#checkpoints



(a) *All speakers* (b) *Speakers grouped by gender*



(c) *Speakers grouped by demographic groups*

Figure 2: *Average scores and standard deviations representing model performance of all 4 models across all annotators.*

2. Method

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. The recordings come from students who self-recorded their full names. The recordings are of varying lengths and differ in audio quality. We ensure that the files represent real-world challenges with audio data (background noise, volume levels, microphone quality, etc.) by further noising them. Our code is available on GitHub[3] and the checkpoints are provided there as well[2].

3. Results

Annotator	A1	A2	A3	A4	A5
A1	1	0.6	0.45	0.56	0.59
A2	0.6	1	0.51	0.41	0.7
A3	0.45	0.51	1	0.3	0.45
A4	0.56	0.41	0.3	1	0.38
A5	0.59	0.7	0.45	0.38	1

Table 1: *Inter-rater agreement based on average Pearson Correlation (PCC) scores across all models.*

We present opinion scores from 5 human annotators. In this test, each subject was asked to listen to converted utterances (with the same target speaker) of 64 testing pairs for each of the 4 models - AutoVC, vanilla FragmentVC, CV-finetune FragmentVC, and CV-VCTK-scratch FragmentVC. The subjects were asked to score the models (1 to 5) based on their confidence in how the utterances sounded (1: doesn't sound like speech; 2: sounds like speech, weird noise and incomprehensible; 3: some comprehensible bits, can't fully parse name; 4: can hear what the name is, still some noise or quality issues; 5: clearly can hear the name, sounds clean). The test suite consists of real student name utterances; we do not release the test suite for privacy reasons. Figure 1 shows Likert score histograms for baselines and our robust models. Figure 2 shows the average scores and standard deviation (rounded to 2 decimal places) representing how well the each model performs across all annotators for all source speakers, source speakers grouped by gender (Male or Female voice), and source speakers grouped by self-reported first-name demographic group affiliation (English, Spanish, Indian, Chinese, or Other). We can clearly see that CV-VCTK has better scores than the rest, followed by CV-finetune. We report inter-rater agreement scores in Table 1.

[3]https://github.com/avanitanna/RobustFragmentVC

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 [2] for real-world, arbitrary, multi-lingual users which can enable them to use such VC models more easily with real-world data.

5. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] E. Moulines and Y. Sagisaka, "Editorial," *Speech Communication*, vol. 16, no. 2, pp. 125–126, 1995, voice Conversion: State of the Art and Perspectives. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167639395900543

[3] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[4] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5939–5943.

[5] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," 2019. [Online]. Available: https://arxiv.org/abs/1904.05742

[6] K. Qian, Y. Zhang, S. Chang, J. Xiong, C. Gan, D. Cox, and M. Hasegawa-Johnson, "Global prosody style transfer without text transcriptions," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8650–8660.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.