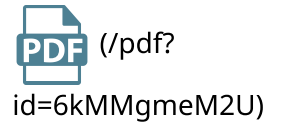


[← Go to ICML 2024 Conference homepage \(/group?id=ICML.cc/2024/Conference\)](#)

SelfVC: Voice Conversion With Iterative Refinement using Self Transformations



Paarth Neekhara (/profile?id=~Paarth_Neekhara1), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), Rafael Valle (/profile?id=~Rafael_Valle1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shlomo Dubnov (/profile?id=~Shlomo_Dubnov1), Farinaz Koushanfar (/profile?id=~Farinaz_Koushanfar1), Julian McAuley (/profile?id=~Julian_McAuley1)

Published: 01 May 2024, Last Modified: 01 May 2024 ICML 2024 Conference, Senior Area Chairs, Area Chairs, Reviewers, Publication Chairs, Authors Revisions (/revisions?id=6kMMgmeM2U) BibTeX CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Verify Author List: I have double-checked the author list and understand that additions and removals will not be allowed after the submission deadline.

Keywords: speech, voice conversion, representation learning, speech synthesis, self refinement

Abstract:

We propose SelfVC, a training strategy to iteratively improve a voice conversion model with self-synthesized examples. Previous efforts on voice conversion focus on factorizing speech into explicitly disentangled representations that separately encode speaker characteristics and linguistic content. However, disentangling speech representations to capture such attributes using task-specific loss terms can lead to information loss. In this work, instead of explicitly disentangling attributes with loss terms, we present a framework to train a controllable voice conversion model on entangled speech representations derived from self-supervised learning (SSL) and speaker verification models. First, we develop techniques to derive prosodic information from the audio signal and SSL representations to train predictive submodules in the synthesis model. Next, we propose a training strategy to iteratively improve the synthesis model for voice conversion, by creating a challenging training objective using self-synthesized examples. We demonstrate that incorporating such self-synthesized examples during training improves the speaker similarity of generated speech as compared to a baseline voice conversion model trained solely on heuristically perturbed inputs. Our framework is trained without any text and is applicable to a range of tasks such as zero-shot voice conversion, voice conversion across different languages, and controllable speech synthesis with pitch and pace modifications. We conduct extensive comparisons against prior work and find that SelfVC achieves state-of-the-art results in zero-shot voice conversion on metrics evaluating naturalness, speaker similarity, and intelligibility of synthesized audio.

Primary Area: General Machine Learning (active learning, clustering, online learning, ranking, reinforcement learning, supervised, semi- and self-supervised learning, time series analysis, etc.)

Position Paper Track: No

Paper Checklist Guidelines: I certify that all co-authors of this work have read and commit to adhering to the Paper Checklist Guidelines, Call for Papers and Publication Ethics.

Submission Number: 2905

Filter by reply type

Filter by author

Search keywords...

Sort: Newest First



Everyone
 Program Chairs
 Submission2905 Authors
 Submission2905...
 12 / 14 replies shown

Submission2905 Area...
 Submission2905...
 Submission2905...
 Submission2905...

Submission2905...
 Submission2905...
 ✕

Add: **Withdrawal**

Official Review of Submission2905 by Reviewer SW7S

Official Review Reviewer SW7S 15 Apr 2024, 04:59 (modified: 15 Apr 2024, 04:59)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SW7S

Revisions (/revisions?id=frgDLU97fj)

Summary:

the authors suggest a tts framework where speech is decomposed into content, duration, pitch and speaker identity. the VC is achieved by extracting the above features, and replacing the speaker identify vector with that of the target speaker. to better improve speaker similarity, this papers draws inspiration from previous methods that perturb the source recording (using formant shift and pitch modulation), and suggest to incorporate intermediate outputs from the model with randomly selected speaker embeddings.

Strengths And Weaknesses:

strengths:

1. the paper is very well-written and easy to follow.
2. method - the authors propose and easy improvement over existing methods by using intermediate outputs from the model. this modifications improves speaker similarity while preserving all other metrics.
3. experimental results - the authors provide extensive experimental results and evaluate their approach in different settings: reconstruction, speaker similarity, and cross-lingual VC. evaluations include both objective metrics and subjective metrics. the relevant baselines were also evaluated against the proposed method.

weaknesses:

1. results - while the objective speaker similarity results (SV-EER & SV-sim) are substantially better, the difference between Baseline-Heuristic and SelfVC falls inside the confidence interval for Sim-MOS and Naturalness-MOS. PER/CER seem to be comparable to Baseline-Heuristic.
2. PER metric for cross-lingual setting - it seems that because the ASR model was trained on English-only the PER metric for all methods (except for (Libri + CSS10)) are comparable (23.3%-23.7%), while the result for (Libri + CSS10) is still in the high ranges for PER. i would suggest the authors re-evaluate using a multi-lingual ASR model for a more informative comparison.
3. baselines were trained on different datasets?

Questions:

1. "Conversely, models that have the ability to explicitly control prosody lack the ability to use SSL, making it extremely hard to support multiple languages" -- why? i believe that [1] also used prosody controls together with (quantized) SSL representations.
2. "Training such systems requires transcribed speech data and the synthesis is limited to the language the model is trained on" -- can the authors please provide a citation or further explanation?

Limitations:



i have not found a limitations section in the manuscript, would be happy to edit in case i missed it.

Ethics Flag: No

Soundness: 3: good

Presentation: 4: excellent**Contribution:** 3: good**Rating:** 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.**Confidence:** 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.**Code Of Conduct:** Yes

Official Review of Submission2905 by Reviewer WvW2

Official Review  Reviewer WvW2  14 Mar 2024, 11:17 (modified: 21 Mar 2024, 05:17) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WvW2 Revisions (/revisions?id=iBo2OaHrBx)**Summary:**

This paper proposes SelfVC, a voice conversion system that uses mismatched speaker and content embeddings as a data augmentation method to disentangle entangled representations. SelfVC produces Mel Spectrograms from input pitch, content embedding, and speaker embedding. Experiments demonstrate small improvements relative to baseline models on similarity, intelligibility, and naturalness metrics.

Strengths And Weaknesses:

The primary contribution (the self-transformation strategy) is indeed powerful, but it is not novel. It was first proposed in (Wang et al, 2022) as "AIC loss".

The contributions state that the proposed system is "a controllable synthesizer that can either mimic the prosody of a source utterance or adapt the prosody of the target speaker". I would expect to see objective prosody reconstruction metrics to support the claim of accurate prosody transfer/reconstruction/control. Other models have shown high accuracy on this task. So it cannot be stated as a contribution unless you're demonstrating some relative improvement. Isolated GPE of your model is not sufficient for that.

What you call your SSL representation is not dissimilar to the latent ASR models often called phonetic posteriorgrams. PPG models offer pitch controllability, duration control, voice conversion, and more without requiring a transcript (Kovela, 2023). Interpretable PPGs further allow pronunciation control (Churchwell, 2024). How is an SSL-based system more "simple" and "efficient"? How have you demonstrated that to be the case? Have you demonstrated that PPGs cannot scale, or extend to other languages?

I'm not entirely convinced that the baseline models represent the state-of-the-art in VC. What about DiffHier-VC or NaturalSpeech 2? PYin does not take as input a spectrogram; Figure 2 is mildly misleading in that regard

Wang, Yunyun, et al. "Controllable speech representation learning via voice conversion and aic loss." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022. Kovela, Sudheer, et al. "Any-to-Any Voice Conversion with F0 and Timbre Disentanglement and Novel Timbre Conditioning." International Conference on Acoustics, Speech and Signal Processing, 2023. Churchwell, Cameron, Max Morrison, and Bryan Pardo. "High-Fidelity Neural Phonetic Posteriorgrams." ICASSP XAI-SA Workshop, 2024.

Questions:

How does deriving durations from the SSL tokens compare to traditional methods of forced phoneme alignment? Have you tried forced phoneme alignment (e.g., MFA)? How does your pitch representation represent unvoiced regions? How does this impact the user's ability to control the signal, or perform transfer?

Limitations:

A limitation statement regarding the ethical implications of (especially zero-shot) voice conversion should be added to the manuscript.

Ethics Flag: No**Soundness:** 2: fair

Presentation: 2: fair

Contribution: 1: poor

Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

28 Mar 2024, 18:20 (modified: 29 Mar 2024, 05:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=h9LBjgfT37)

Rebuttal:

We thank the reviewer for your valuable feedback and suggestions.

Differences with AIC Loss Our approach is fundamentally different from AIC loss (Wang et al, 2022) in two aspects:

1. AIC loss is formulated as a loss in the feature embedding space, calculated using the L1 distance between the content embedding of the original and self-voice-converted audio. In contrast, SelfVC does not introduce any additional loss terms to our feature extractor or synthesis model. Instead, we perturb the inputs to our synthesizer while keeping the reconstruction objective exactly the same.
2. Instead of enforcing perfect feature disentanglement, we make the model robust to imperfectly disentangled representations. That is, even if the speaker information leaks into the content representation, our synthesis model learns to ignore it and capture voice characteristics only from the speaker embedding, using the proposed information perturbation technique.

SSL vs PPG As opposed to PPG (obtained from an ASR model trained on speech and text pairs), SSL representations do not require text transcripts at any stage of the training pipeline. Moreover, SSL representations are richer than PPGs since they capture aspects such as speaker, emotions and style besides the phonetic content in the audio, as shown by their effectiveness for downstream tasks besides ASR (Hussain et al ICASSP 2022, Huang et al ICASSP 2022). We compare against two prior VC models that rely on PPG, namely ACE-VC and YourTTS. As demonstrated by the results in Table 2 and Table 3, SelfVC outperforms these prior approaches, and the improvement in intelligibility is even more significant on the cross-lingual voice-conversion task (Table 3)

Other Questions



1. SSL durations vs phonetic durations: Since our model is textless, durations are extracted in an unsupervised manner using the procedure described in Appendix A. Prior work ACE-VC extracts phonetic durations as opposed to our technique using a similar model architecture and dataset as ours. As shown SelfVC outperforms ACE-VC on all benchmarks.
2. Unvoiced Speech modelling: The unnormalized F0 contour from Pyin in the unvoiced regions of the waveform is all zeros. The synthesizer implicitly learns to model voiced/unvoiced regions of the source utterance from this pattern. We encourage the readers to try the VC demo linked in the paper to judge the prosodic similarity/modifications





Replying to Rebuttal by Authors


Rebuttal by Authors

Rebuttal

 Authors ( Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

 29 Mar 2024, 00:54 (modified: 29 Mar 2024, 05:40)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=uBGhQKoAFX)


Rebuttal:

Additional Prior Work Comparison

Thanks for pointing out additional VC models (Diff-HierVC and NaturalSpeech 2). From the mentioned papers, Diff-HierVC is officially open-sourced and we were able to evaluate the model using the released checkpoints. Additionally, we evaluate DDDM-VC mentioned by Reviewer e8vQ on the same objective benchmark as Table 2 of our paper. As reported in the results below, SelfVC outperforms Diff-HierVC and DDDM-VC on objective speaker similarity and intelligibility metrics.

Technique	SV-EER ↓	SV-SIM ↑	CER ↓
AdaIN-VC	28.7%	0.36	15.5%
MediumVC	27.4%	0.40	29.1%
FragmentVC	23.3%	0.39	31.1%
S3PRL-VC	20.5%	0.38	9.6%
YourTTS	6.6%	0.54	4.9%
ACE-VC	6.6%	0.49	3.8%
Diff-HierVC	10.9%	0.48	2.7%
DDDM-VC	13.7%	0.45	2.6%
SelfVC	3.4%	0.58	1.6%



 *Replying to Rebuttal by Authors*

Official Comment by Reviewer WvW2

Official Comment  Reviewer WvW2  03 Apr 2024, 08:55

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I do not think the proposed method is sufficiently different than Wang et al.'s proposed method. AIC loss is used in conjunction with the speaker-based transformations that you propose. Admittedly, Wang et al. could ablate this better. As well, you mention jointly performing disentanglement as a limitation of Wang et al. relative to your system. How is that a limitation? Isn't that adding strictly more capability--including state-of-the-art pitch-shifting accuracy (still the case in 2024) and with a smaller (faster) model (HiFi-GAN vs FastPitch)? You mention "even if speaker content leaks...", but have you actually verified speaker content leaks in their model such that this concern is warranted? Wang et al. was also not cited in the original submission; I have a hard time believing these "differences" are deliberate decisions made relative to Wang et al.

The availability of thousands to tens-of-thousands of hours of paired text-speech data as well as high-quality ASR makes a lack of text transcript a solvable problem for PPGs. SSL and PPG models are both improving: comparing an older PPG to a newer SSL is as fair as comparing a newer PPG to an older SSL model. As well, your primary

contribution is the self-transformation, which you only apply to SSL and not PPG models. The authors' dismissal of PPGs seems to indicate that they believe PPG models cannot be further improved.

Regarding phoneme alignment, my comment was asking for a performance comparison. Comparing SelfVC vs ACE-VC is not sufficient as an isolated ablation of this design decision.



SSL and PPG representations

Official Comment

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

03 Apr 2024, 21:06 (modified: 03 Apr 2024, 21:23)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=3ZHV2cgIIu)

Comment:

We would first like to clarify that the main focus of our paper is not to argue the superiority of SSL based voice conversion over PPG based voice conversion. Rather, our work focuses on improving SSL based voice conversion, which is a widely adopted technique for performing speech synthesis, since it is not dependent on text for performing voice conversion (does not require separately trained ASR models) and thereby allows for effective voice conversion for low resource languages. Additionally, we respectfully draw attention to prior studies that have highlighted the advantages of SSL-based speech synthesis over PPG-based alternatives:

1. "Being able to achieve 'textless NLP' would be beneficial for the majority of the world's languages which do not have large textual resources or even a widely used standardized orthography (Swiss German, dialectal Arabic, Igbo, etc.), and which, despite being used by millions of users, have little chance of being served by current text-based technology. It would also be useful for 'high-resource' languages, where the oral and written forms often mismatch in terms of lexicon and syntax, and where some linguistically relevant signals carried by prosody and intonation are basically absent from text. While text is still the dominant form of language on the web, a growing amount of audio resources like podcasts, local radios, social audio apps, on-line video games provide the necessary input data to push NLP to an audio-based future and thereby expand the inclusiveness and expressivity of AI systems", On Generative Spoken Language Modeling from Raw Audio, Transactions of the Association for Computational Linguistics, 2021
2. "Additionally, the language dependency of the ASR network limits the model's capability to be extended to multilingual settings or languages with low-resources. To address these concerns, efforts have been made to divert from using the text information", Neural Analysis and Synthesis (NANSY): Reconstructing Speech from Self-Supervised Representations, Advances in Neural Information Processing Systems, 2021 (and follow up work : NANSY++: Unified voice synthesis with neural analysis and synthesis, ICLR, 2023).



We kindly clarify that self-transformation based information perturbation techniques can be generally applied to either PPGs or SSL. However, the inherent text-dependency of PPG based setups hinders cross-lingual voice conversion on low resource languages. As clarified earlier, ACE-VC (Hussain et al, 2023) is the closest text-dependent (PPG based) voice conversion model (that uses same network architectures as SelfVC for feature extractor and synthesizer); and SelfVC outperforms ACE-VC on both zero-shot English voice conversion and zero-shot cross-lingual voice conversion.



Replying to SSL and PPG representations

Comparison to new SSL based VC systems.

Official Comment

 Authors ( Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

 03 Apr 2024, 21:30  Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

As we stated before, we have compared our work to several newly proposed SSL based voice conversion systems. At the time of our submission, we have compared our work to state-of-the-art SSL based speech synthesis papers such as S3PRL-VC(Huang et al, 2021,2022), ACE-VC (Hussain et al, 2023) and reimplemented version of NANSY (Choi et al, 2021,2023).

In our rebuttal we included comparison of our work to two additional recently proposed VC frameworks (suggested by reviewers) with open source implementations: Diff-HierVC (Choi et al, Interspeech 2023) and DDDM-VC (Choi et al, AAAI 2024) and demonstrate the superiority of our proposed SelfVC framework to these models.



Official Comment by Reviewer WvW2

Official Comment  Reviewer WvW2  04 Apr 2024, 09:53

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WvW2

Comment:

The point about PPGs vs SSL is minor relative to the similarity of the primary proposed contribution with Wang et al. I acknowledge that the results are impressive, and the additional ablations with recent VC methods make that even clearer. However, I consider the primary proposed contribution to be already published, which means that your good results at least need to be reframed.

Regarding SSL vs PPGs (which, again, is minor), the language dependency you mention from NANSY can be solved with multi-lingual ASR representations and the text dependency you mention from textless NLP only applies during training, not inference. But, you are right that it's easier to scale SSL training using untranscribed speech than finding transcribed speech or doing multi-lingual ASR to transcribe.

I thank the authors for their work. I maintain my current review score.



Official Review of Submission2905 by Reviewer rNGf

Official Review  Reviewer rNGf  14 Mar 2024, 01:37 (modified: 21 Mar 2024, 05:17)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer rNGf

 Revisions (/revisions?id=Uhl2CCxkVI)

Summary:

The authors propose a model for zero-shot voice conversion based on disentangled content, prosody, and speaker representations. To improve speaker disentanglement, the model is trained first using heuristically perturbed audio to obtain content features, and then on its own outputs with randomly assigned speaker identities. The authors show that this iterative improvement method yields superior voice-conversion performance when compared to training with only heuristically-perturbed audio.

Strengths And Weaknesses:

Strengths:

- The proposed iterative refinement method is clever and intuitive, and seems applicable to a wide range of voice conversion architectures

- The authors show via ablations that training their model with iterative refinement outperforms training with only heuristic perturbations
- Duration extraction using cosine similarity on content features is interesting
- The authors' model obtains state-of-the-art performance in a zero-shot voice conversion task as measured by automated intelligibility and speaker similarity metrics and a human listener study

Weaknesses:

- The effect of self-transformations versus heuristic perturbations seems to be consistent but small, as illustrated in tables 2-3. Given that the overall architecture and approach beyond the core self-transformation idea are fairly standard, it would be nice to see more evidence of the impact of this core idea -- for instance, does adding self-transformations to the training of other analogous voice-conversion models yield similar improvements? What about training on transformations from a (separate, pretrained, potentially inferior) voice conversion model? How does performance vary when including both self- and heuristic transformations versus self-transformations only? As things stand, I think the paper could make a much stronger case for self-transformation training as a general technique in voice conversion.

Questions:

- Is there a reason the authors use PYin for pitch estimation? As far as I'm aware, PYin performs significantly worse on pitch accuracy compared to recent neural pitch estimators
- How is the distinction between voiced/unvoiced frames modeled? Does PYin output a zero pitch in unvoiced frames?
- Is there a significant difference in training wall-time when using self-transformations versus heuristic transformations (due to additional forward passes)? If self-transformation is significantly slower, does training on heuristic transformations up to an equivalent wall-time close the gap in voice-conversion performance at all?
- Are self-transformations computed using the "current" version of the model for each batch, or a cached previous version of the model that is updated periodically?

Limitations:

Yes.

Ethics Flag: No

Soundness: 3: good

Presentation: 3: good

Contribution: 3: good

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

28 Mar 2024, 18:33 (modified: 29 Mar 2024, 05:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=iHjfH3ZR5X)

Rebuttal:

Self Transformation Effectiveness: We initially train the model with heuristic transformations for a reasonable initialization of the VC model that can be used for self-transformation. It is indeed possible to train the model without any heuristic transformations, however, in this case, the initial VC model used for self-transformation will

be inferior (approaching the performance of Baseline-No Transform experiment in the paper), and the training time will be longer. However, we concur that this is a valuable experiment to see how far we can get without using any heuristic transformation and is worth including in the final version of the paper.

Heuristic + self-transformation vs only Self-transformations: In the experiments reported in the paper, the heuristic transformation is used for the first 100k mini-batch interactions, thereafter we only use self-transformation. We did try a preliminary experiment in which we sample randomly between heuristic and self-transformation after 100k training steps but did not observe any significant difference with a model trained only with self-transformations after 100k steps. Therefore, for simplicity, we only use self-transformation after the reasonable initialization of the VC model with heuristic transforms.

Latency and Longer Only-heuristic model training: Since the conformer and synthesis model are both transformer-based architectures (non-autoregressive), the forward-pass through for self-transformation does not introduce a significant latency (since it can be accelerated on GPU). Without using any asynchronous processing for self-transformations, we observe a roughly 25% increase in training time per iteration. For heuristic transformations, we use asynchronous processing on CPUs which does not introduce additional latency. While the SelfVC model improves until around 500 epochs, the baseline-heuristic model performance saturates at roughly 300 epochs and we do not observe improvement until 500 epochs. As reported in Appendix B, we use 500 epochs for each model for a fair comparison in our paper.

Additional Clarifications

1. We use PYin as a simple F0 estimation technique with demonstrated success in TTS models such as FastPitch. We agree that more accurate neural pitch estimators can further enhance prosodic modelling.
2. The unnormalized F0 contour from Pyin outputs zero for unvoiced frames.
3. Self-transformations are computed using the current version of the model for each batch.



Official Review of Submission2905 by Reviewer e8vQ

Official Review Reviewer e8vQ 03 Mar 2024, 00:27 (modified: 21 Mar 2024, 05:17)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer e8vQ

Revisions (/revisions?id=KxMv45218n)

Summary:

This paper introduces a zero-shot voice conversion system, called SelfVC. The proposed model achieves zero-shot VC by integrating pitch contour and several entangled speech representations from self-supervised learning (SSL) and speaker verification models. To address the entangled problems, not only speech perturbation but also a proposed self-transformation training strategy are employed before SSL extraction. To evaluate SelfVC, the authors employ multiple metrics, including subjective and objective measures, providing comprehensive comparisons with several previous VC models.

Strengths And Weaknesses:

Strengths

- This paper presents their method very clearly and easy to follow. One main contribution is a novel training strategy that utilizes synthesized examples for riching speech perturbation and iterative improvement of the VC model. Benefiting the previous and proposed heuristic transformations, the proposed method can achieve good zero-shot performance.
- The paper conducts comprehensive experiments from different aspects, demonstrating the superiority and efficacy of the proposed method.

Weaknesses

- The main framework used in this paper is like the combination of ACE-VC and NANSY. And the proposed Self Transformations share the similar idea to the previous cycle-consistency series methods, like cyclegan-VC, and stargan-VC (<https://arxiv.org/pdf/1806.02169.pdf> (<https://arxiv.org/pdf/1806.02169.pdf>)). Also some recent studies like DDPM-

VC(<https://arxiv.org/pdf/2305.15816.pdf> (<https://arxiv.org/pdf/2305.15816.pdf>)) and VoiceMixer (<https://proceedings.neurips.cc/paper/2021/file/0266e33d3f546cb5436a10798e657d97-Paper.pdf>) (<https://proceedings.neurips.cc/paper/2021/file/0266e33d3f546cb5436a10798e657d97-Paper.pdf>) also present this similar training process.

Questions:

Q1 In the second paragraph of Section 1 (line 61-63), it would be beneficial for the authors to provide appropriate citations or explanations about how to get inspiration about integrating the neural network-generated augmentations for better augmentation.

Q2 In the third paragraph of Section 1 (line 67), how to define the prosody since only duration is mentioned?

Q3 Figure 1 and 2 are a bit small. And the font of the full paper seems different from that of other papers in ICML.

Limitations:

N/A

Ethics Flag: No

Soundness: 3: good

Presentation: 3: good

Contribution: 2: fair

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Boris Ginsburg (/profile?id=~Boris_Ginsburg1), Rishabh Ranjan (/profile?id=~Rishabh_Ranjan5), Shehzeen Samarah Hussain (/profile?id=~Shehzeen_Samarah_Hussain1), +4 more (/group/info?id=ICML.cc/2024/Conference/Submission2905/Authors))

29 Mar 2024, 00:40 (modified: 29 Mar 2024, 05:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=xxBvixT0qj)

Rebuttal:

Thank you for your detailed assessment and feedback. We address the comments below:

Difference with cycle-consistency methods like StarGAN-VC, CycleGAN-VC and Voice-Mixer: In contrast to cycle-consistency methods, which approach voice conversion using a conditional GAN model, we propose to utilize the synthesis model as a self-improving data-augmentation method to create increasingly diverse VC training examples. Our mel-spectrogram synthesizer does not require any adversarial training objective and is trained using only reconstruction loss (L1/L2 loss on mel-spectrogram). In our framework, there is no discriminator network or domain classification network that distinguishes real/synthetic audio and identifies speakers. Instead, we only use a reconstruction objective to encourage accurate reconstruction of the original signal from self-perturbed input representations.

Moreover, StarGAN-VC and CycleGAN-VC can only perform conversion among a predefined set of speakers and cannot perform any-to-any/zero-shot voice conversion. This limitation is discussed in detail in prior work FragmentVC (Lin et al, ICASSP, 2021).

Additionally, we compare our model against DDDM-VC from the official checkpoints and inference code and find that SelfVC significantly outperforms DDDM-VC and Diff-HierVC on speaker similarity metrics. We will update Table 2 of our paper to include these results.

Technique	SV-EER ↓	SV-SIM ↑	CER ↓
Diff-HierVC	10.9%	0.48	2.7%
DDDM-VC	13.7%	0.45	2.6%
SelfVC	3.4%	0.58	1.6%

Improvements over ACE-VC and NANSY As you noted, ACE-VC and NANSY, do not employ any self-transformation techniques for performing voice conversion. We describe the differences in techniques in our paper's Related Work Section 2 - Voice Conversion. For effective comparison of results, we also develop a NANSY-like framework that utilizes heuristic transformations from NANSY and our model architectures. We demonstrate the superiority of our approach to such a setup (Baseline-Heuristic) indicating the effectiveness of self-transforms and textless content representations.

Additional Clarifications Thanks for pointing out the formatting issues. We will double-check our LaTeX packages and fix any discrepancies. We will add further clarifications regarding prosody in the introduction which refers to phonetic durations and pitch modulation of the speaker (duration and F0 contour).

[About OpenReview \(/about\)](#)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](#)

[All Venues \(/venues\)](#)

[Sponsors \(/sponsors\)](#)

[Frequently Asked Questions](#)

(<https://docs.openreview.net/getting-started/frequently-asked-questions>)

[Contact \(/contact\)](#)

[Feedback](#)

[Terms of Use \(/legal/terms\)](#)

[Privacy Policy \(/legal/privacy\)](#)

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2024 OpenReview