Basic Information

**Title**: SLMGAN: Exploiting Speech Language Model Representations for Unsupervised Zero-Shot Voice Conversion in GANs

**Author**: Yinghao Aaron Li, Cong Han, Nima Mesgarani

**ID**: arXiv:2307.09435

Introduction:

**\* (1) What is the research background of this article?:** The research background of this article is in the field of voice conversion, specifically the development of a model that can perform voice conversion without requiring text labels.

**\* (2) What are the pastmethod? What are the problems with them? Is the approach well motivated?:** The past methods for voice conversion rely on text labels and can be time-consuming and require significant computational resources. The proposed approach is well-motivated as it eliminates the need for text labels and has the potential for real-time applications.

**\* (3) What is the research methodology proposed in this paper?:** The research methodology proposed in this paper is a self-supervised learning approach using a large speaker language model (SLM) to learn speaker-related representations, which is then used to perform voice conversion.

**\* (4) On what task and what performance is achieved by the methods in this paper? Can the performance support their goals?:** The task addressed in this paper is voice conversion, and the performance achieved by the proposed method is better than two baseline models in terms of naturalness and similarity, and it also outperforms two of the best-performing publicly available models for zero-shot voice conversion that rely on text labels. The achieved performance supports the goal of developing a voice conversion model that eliminates the need for text labels and has the potential for real-time applications.

Method:

\* (1): The paper proposes a speaker verification system called SLMGAN, which aims to learn a mapping that converts a speech sample from the source speaker to a speech sample in the target speaker without parallel data and text labels.

\* (2): The system uses a mel-discriminator to determine whether the converted speech is real or fake and a mel-based source classifier to identify the source speaker of the converted speech.

\* (3): The system is modified from StarGANv2-VC by adding a SLM-based discriminator in addition to the mel-based discriminator, substituting the mel-based source classifier with the SLM-based one, removing the mapping network and domain-specific projections, and replacing the ASR loss with the SLM loss.

\* (4): The framework consists of four modules: the generator, the F0 network, the style encoder, and the discriminators.

\* (5): The generator transforms an input mel-spectrogram into a feature vector that reflects the speaker identity in the style vector encoded by the style encoder and the pitch represented in the convolution layers in the F0 extraction network.

\* (6): The F0 network is a pre-trained JDC network that extracts the fundamental frequency from an input mel-spectrogram.

\* (7): The style encoder is given a reference mel-spectrogram and extracts the style vector.

\* (8): The discriminators include the mel-based discriminator and the SLM-based discriminator.

\* (9): The SLM-based discriminator is introduced on top of the mel-based domain-specific discriminator and replaced the source classifier in StarGANv2-VC with a SLM-based classifier.

\* (10): The model is trained with the following objectives: adversarial loss, adversarial source classifier loss, style reconstruction loss, F0 loss, norm consistency loss, style consistency loss, and cycle consistency loss.

Conclusion:

**\* (1) What is the significance of this piece of work?:** The significance of this piece of work is that it proposes a novel zero-shot voice conversion (VC) model called SLMGAN, which eliminates the need for text labels and achieves state-of-the-art performance in naturalness and similarity.

**\* (2) Summarize the strengths and weaknesses of this article in three dimensions: innovation point, performance, and workload:** Innovation point: The SLM-based discriminators and speech consistency loss are novel contributions that improve the performance of the VC model. Performance: The proposed model outperforms two baseline models and two publicly available models for zero-shot VC that rely on text labels. Workload: The model is trained with multiple objectives, which may increase the training workload.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Author | Year | Method | Main Result | Advantages and Disadvantages | Innovation |
| column\_1 | authors | published | review\_table\_information\_from\_method | review\_table\_information\_from\_conclusion | column\_2 | column\_3 |
| column\_12 | authors2 | published2 | review\_table\_information\_from\_method2 | review\_table\_information\_from\_conclusion2 | column\_22 | column\_32 |

Tablecaption