Deep Singing Voice Conversion

MDS6002 Default Project

Jiahao Chen (222041038)
School of Data Science
Chinese University of Hong Kong, Shenzhen
222041038@link.cuhk.edu.cn

1 Key Information to include

- Mentor and external collaborator (if any): Zhizheng Wu, Xueyao Zhang...
- Complete the project as an individual.

2 Background and Introduction

Singing Voice Conversion (SVC) aims at converting a piece of singing voice produced by an individual to another one as if it's produced by another individual. The motivation for me to choose this topic to dive into is out of both utility and interest. It's useful in that it can be used to polish the pieces produced by immature singers to a somewhat perfect state under the criterion of sing voice. To achieve that, the only thing we need to do is to feed the model with the singing voice sample of a renowned singer for it to learn. In the meanwhile, it's a fascinating idea that as long as you have the samples of a singer, you can produce his/her singing voice with any textual content as long as another voice with designated textual content is provided, regardless of it's mastery or skills of singing.

The components of a piece of singing voice are divided into textual content and acoustic properties. In terms of human's perception, the acoustic properties are comprised of volume, pitch, and timbre. What determines the identity of the source of a piece of singing voice most is its timbre, which is actually defined as the remaining voice of a piece of human voice except volume and pitch. All those concepts mentioned above are from the perspective of subjective perceptual. There are also objective counter-part in the domain of acoustics, e.g., spectral envelope for timbre. Therefore, by assigning each singer a characteristic vector further extracted from these acoustic properties (identity vector), we can represent the timbre of a singer.

3 Approach Proposal

Methodology Currently, my personal conception of the process of SVC is that, firstly, we extract both the textual content and the acoustic properties of a piece of singing voice. Further feeding these acoustic properties to some higher-level feature extractor (the Identity Extractor as is mentioned in the following paragraph), we obtain unique identity vector for each singer. Then, by replacing the source identity vector to the target identity vector (i.e., the identity vector of the singer we are targeting at), we have essentially finished the conversion part of SVC. Finally, we concatenate the original vectors (or matrices/tensors) representing the textual content, original vector representing other acoustic properties and the substituted identity vector into a vocoder (which synthesizes those materials above into a piece of voice), we obtained the desired singing voice.

Shallow Features Extractor In this project, I would like to define those low-level textual and acoustic properties as the shallow features, which are supposed to be extracted from the raw singing

voice at the beginning stage of our model pipeline. Given that the raw input is in the form of time series, the most popular choice of shallow features extractor is Transformer. There is also upgraded version of Transformer, for example, Conformer. Meanwhile, we can directly utilize the pre-trained extractor as well, e.g., Whisper from OpenAI, Wav2Vec2 from Facebook and HuBERT etc.

Identity Extractor After the acoustic properties representing a singer's timbre are extracted, an unsupervised **Generative Model** may be used to project these properties as higher-level feature vector. I will call this part of the model the Identity Extractor. Classic deep generative models include **Variational AutoEncoder (VAE), Conditional VAE (CVAE) or Adversarial Generative Neural Network (GAN)**. For VAE, we input the timbre feature vector, and the bottleneck will exactly be the identity vector we want. And for CVAE, maybe we can use shallow features as input, and condition on each singer's identity.

Vocoder I have not investigated into this part yet. A baseline can be the WORLD model as is mentioned in the project specification.

Possible Directions of Research Exploration of the choices of each part of the model pipeline, i.e., shallow features extractor, identity extractor and vocoder as I mention above, will be good research directions. More cutting-edge model or architecture may be utilized. Apart from that, the choice of shallow features is worth considering as well. This requires further study into the acoustics behind singing voice.

4 Related Work

Model Framework Yin-Jyun Luo et al. use VAE as the identity extractor as is mentioned in the part of identity extractor. [8] Eliya Nachma et al. put forwards a model pipeline which adopts the idea resembling CVAE. [9] Berrak Sisman et al. adopt GAN as the identity extractor. [11] Xin Chen et al. breakthrough with the non-parallel data, using Bi-directional LSTM as the identity extractor. [1] There are also other works which explore frontier Deep Learning techniques and exploit them to perform SVC. [2][10][12][5][4][3][6][7]

Feature Engineering An important part of SVC is the choice of shallow acoustic features. Hence, potentially, some articles analyzing the acoustics behind singing voice in depth are worth reading. Some more old-aged yet classical works may be investigated into as well, due to their work of feature engineering in the absence of deep learning.

References

[1] Xin Chen et al. "Singing voice conversion with non-parallel data". In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2019, pp. 292–296.

[2] Chengqi Deng et al. "Pitchnet: Unsupervised singing voice conversion with pitch adversarial network". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7749–7753.

[3] Haohan Guo et al. "Improving adversarial waveform generation based singing voice conversion with harmonic signals". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6657–6661.

[4] Zhonghao Li et al. "Ppg-based singing voice conversion with adversarial representation learning". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7073–7077.

[5] Songxiang Liu et al. "Diffsvc: A diffusion probabilistic model for singing voice conversion". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2021, pp. 741–748.

[6] Songxiang Liu et al. “Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation”. In: *2021 ieee international conference on multimedia and expo (icme)*. IEEE. 2021, pp. 1–6.

[7] Junchen Lu et al. “Vaw-gan for singing voice conversion with non-parallel training data”. In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 514–519.

[8] Yin-Jyun Luo et al. “Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3277–3281.

[9] Eliya Nachmani and Lior Wolf. “Unsupervised singing voice conversion”. In: *arXiv preprint arXiv:1904.06590* (2019).

[10] Adam Polyak et al. “Unsupervised cross-domain singing voice conversion”. In: *arXiv preprint arXiv:2008.02830* (2020).

[11] Berrak Sisman et al. “SINGAN: Singing voice conversion with generative adversarial networks”. In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2019, pp. 112–118.

[12] Liqiang Zhang et al. “Durian-sc: Duration informed attention network based singing voice conversion system”. In: *arXiv preprint arXiv:2008.03009* (2020).