KaraSinger: Score-Free Singing Voice Synthesis with VQ-VAE using Mel-spectrograms

Chien-Feng Liao[*], *Jen-Yu Liu*[*], *Yi-Hsuan Yang*

Taiwan Ai Labs, Taiwan

AI Labs.tw 台灣人工智慧實驗室

Abstract

- **KaraSinger** is proposed for a task named ***score-free*** **singing voice synthesis** (SVS), in which the prosody and melody are spontaneously decided by machine.
- **Why?**
  1. SVS is the task of computationally generating singing voices from **music scores** and **lyrics**. For ordinary users, composing a reasonable melody is much harder than writing lyrics.
  2. Here, we explore the realm of **score-free SVS** (or, **text-to-music**), where a model learns the prosody and melody of music implicitly from data. At inference time, **the model sings without the guidance of human input other than lyrics**.
- A **VQ-VAE** is first trained to compress singing voices into a discrete space, which is then modeled by an autoregressive Transformer conditioned on lyrics.

Method

- **Hierarchical VQ-VAE over Mel-spectrograms**
  1. Work on Mel-spectrograms to reduce computational time.
  2. The objective function is a combination of **reconstruction loss**, **commitment loss** and an additional **CTC loss**, which is crucial for the system to work.

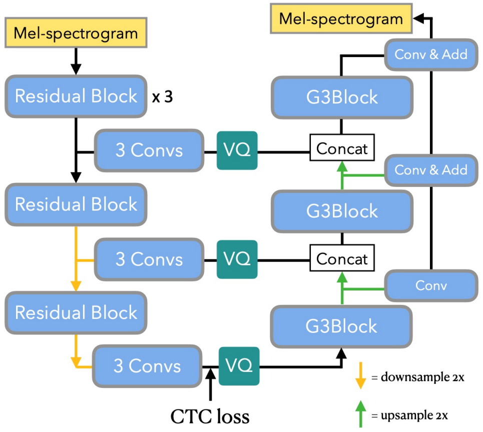

Figure 1. *Schematic diagram of the VQ-VAE part of our model. The 'G3 Block' comprises 1-D convolutions, group-normalization, and bi-directional GRUs.*

- **Language Model: GRU & Transformer**
  1. Modeling the compressed discrete latent codes via multi-level autoregressive Transformer.
  2. **For the top-level codes, we employ a seq2seq model to account for the input lyrics.** As shown in the left and middle parts of Fig. 2, the encoder-decoder is trained to map a sequence of phonemes to a sequence of top-level codes. The **location-sensitive attention** is used to encourage a monotonic alignment between the two sequences.
  3. **For the middle- and bottom-level codes, we combine them into a single sequence by "interleaving" and use a single Transformer to jointly model them**, as shown in the rightmost part of Fig. 2. We propose this setting because the output Mel-spectrogram is connected to each level of the VQ decoder via residual connections, and accordingly codes from both levels affect each other bilaterally. In this way, the generation of the middle-level codes is conditioned on the whole sequence of the top-level codes.
  4. **Linear Transformer is adopted to model the mixed sequence**, for its effectiveness against autoregressive prediction on long sequences.
  5. **The tick embedding**, tick ∈ {t1, t2, ..., t6}, where t is a learnable vector, is repeated and added to the mixed sequence by the order of (t1, t2, ..., t6, t1, ...), to signal the current position in the order of (M, M, B, B, B, B).

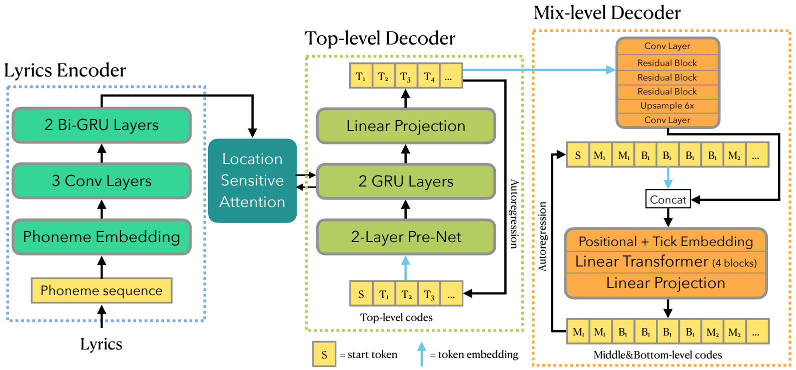

Figure 2. *The LM part of our model that models the three layers of VQ codes (T: top, M: middle, B: bottom). The combination of the lyrics encoder and the top-level decoder follows the same design as the acoustic model in Tacotron2, except that the decoder predicts discrete tokens instead of Mel-spectrograms. The rightmost decoder consists of an upsampler and a linear Transformer. Finally, the decoded tokens are rearranged and combined with the top-level tokens to be taken as input for the VQ-VAE decoder shown in Fig. 1.*

Experiments and results

- **Dataset:** We purchased 550 songs from an English karaoke website to build a **multi-singer, pop genre dataset**. We run NUS AutoLyricsAlign on the vocals to obtain **sentence-level alignments of the lyrics**. Songs are split into segments between 5 to 15 seconds. The final dataset consists of 10,589 short segments, amounting to roughly **20 hours worth of data, sampled at 44.1 kHz**. MelGAN is used for vocoder.
- **Experiments:**
  1. **noCTC** is the case without the CTC loss during the VQ-VAE training. **3-level** is the case where the CTC loss is used, but the middle- and bottom-level codes are separately predicted, like in Jukebox.
  2. **Subjective listening test:** participants are asked to rate the samples considering the following three aspects, **intelligibility**, **musicality**, and **overall quality**, all on a 5-point Likert scale.

Model	**Intelligibility**	**Musicality**	**Overall**
noCTC	1.30 ± 0.14	2.00 ± 0.16	1.64 ± 0.14
3-level	3.30 ± 0.18	3.43 ± 0.16	3.19 ± 0.17
Proposed	**4.23 ± 0.14**	**3.85 ± 0.14**	**3.67 ± 0.15**

Table 1. Mean scores (plus the 95% confidence interval) from the participants of the subjective evaluation in three different aspects.

Conclusion

- We have introduced in this paper **KaraSinger**, an SVS model based on a hierarchical VQ-VAE over Mel-spectrograms and a lyrics-conditioned LM to achieve **score-free SVS, which can generate novel singing audio given lyrics but no score input**.
- This opens up a new direction for SVS, where music creators can get inspirations from the synthesized singing, and common users can experience the creativity of AI.
- Examples of audio samples can be found on at the following demo page: https://jerrygood0703.github.io/KaraSinger