Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/68/20241430

Animal speech and singing synthesis model based on So-VITS-SVC

Yutang Gong

London College of Communication, University of the Arts London, London, the United Kingdom

The corresponding author's e-mail address: y.gong0620231@arts.ac.uk

Abstract. Currently, when researchers in deep learning and neural network technology have made significant progress, the author makes a new bold attempt to apply the technical principles of speech and singing synthesis with artificial intelligence to the field of animal speech and singing synthesis, using So-VITS-SVC4.0 framework, which was originally designed for human voice synthesis. Taking dogs as an example of a species and putting datasets of their sounds to use, the author is committed to capturing its sound characteristics and vocalization through model training and generating synthetic sounds with a high degree of similarity. This research may not only contribute to a deeper understanding of how animals communicate, but also open up new possibilities for animal sound art and music creation. With the continuous progress and improvement of technology, synthetic animal speech and singing by artificial intelligence may play an increasingly important role in zoological research and entertainment, bringing new perspectives and possibilities for communication between humans and animals.

Keywords: Animal, sound synthesis, So-VITS-SVC

1. Introduction

Humans and other animals produce complex acoustic signals for various purposes. Speech, song and music serve a variety of functions including communication and entertainment [1]. Long and varied vocalizations in certain whale and bird species are used for courtship, territorial establishment and social affiliation [2]. The history of this research could go back to ancient times, but only in recent times, it starts to be systematically explored scientifically. In the past, people relied mainly on observing animal behavior and sounds to infer their intentions and feelings, and the limitations of this approach are obvious. Various groups of investigators that believe in the existence of shared characteristics between human language and non-human animal communication have been doing extensive research over the past 25 years, on humans, apes, rodents, and birds, among others [3]. With the development in technology, especially computer and neuroscience, researchers have begun to experiment with more advanced tools and methods to decode how animals communicate. Modern research methods including advanced recording equipment, neuroimaging techniques, and machine learning algorithms, show excellent results in analyzing and interpreting animal language and behavior. For example, by recording and analyzing the voices of large marine mammals such as whales and dolphins, scientists are attempting to decode their possible language systems and modes of communication. Their vocals may have a certain complexity and diversity. To some extent, a lot of animals make singing-like sounds, which may exhibit

DOI: 10.54254/2755-2721/68/20241430

certain melodies and rhythms. Birds are one of the most famous singing animals: Males usually sing to attract females, to claim territory, or to communicate with competitors. The songs of many birds have complex structures and varied pitches, such as the robin and nightingale. Their songs are often celebrated as music in nature.

With the rapid advances in artificial intelligence and machine learning, some researchers are beginning to use these techniques to build animal language translation systems. These systems analyze animal sounds and other forms of their communication in order to translate them into human-understandable language or symbols. Although these systems are still in their early stages, they offer new ways of understanding animal thinking and emotions. Animal language translation research breaks down communication barriers between humans and other species, helping us to better understand the animal world and promote a more harmonious relationship between humans and animals. While there are still many challenges in this field, the ability to interpret animal language will continue to improve as technology and theory continues to advance.

Initial works of singing voice synthesis generate the sounds using concatenated methods, which are kind of cumbersome and lack flexibility and harmony. Thanks to the rapid evolution of deep learning, several SVS systems based on deep neural networks have been proposed in the past few years [4]. AI synthesized speech and singing, as an important application of AI technology in the field of audio processing, which successfully utilizes algorithms and machine-learning models to generate synthesized singing based on human voices, has made great strides and is able to achieve natural and smooth musical performance to a large extent. The techniques included are, but not limited to: deep learning and neural networks, generalized speech modelling, accumulation and opening of datasets, and advances in sound synthesis techniques. As this type of acoustic model can well capture the common features and changing patterns of the sounds provided by the datasets, the author holds a great interest in combining the two different fields of animal language translation science and algorithmic speech synthesis, hoping that the study of animal language translation will not only stay at the literary level, but will directly through the animal's own timbre and vocalizations to the part that fits into the human linguistic system in the voice or song and interact with human beings. Therefore, the author tried to collect the voice samples of the dog, which is the most closely related species to human beings, and produced a datasets to be put into the existing pre-training model of speech and singing synthesis for training, then eventually gained successful training results. The datasets production, model training and evaluation sessions will also be described in detail later.

2. Data and method

2.1. Data

Neural networks are able to generate sound based on the statistical distribution of the training data. The more uniform the input data to the network is, the more natural outcome can be achieved [5]. The datasets used for training in this study was derived entirely from the author's in-person sound collection of small, medium, and large sized dogs near her residence in the community via recording devices, with sound elements including barking, panting, howling, whining, sniffing, breathing, and grunting. Choosing the species of dog as an example is due to they can almost be considered as the most influential species in human life and social activities (guide dogs, comfort dogs, police dogs, etc.), and also they have a good level of intelligence. After the sound collection was complete, the author performed some screening and noise reduction on the audio, and then used Audio Slicer to cut them into shorter duration slices suitable for training and packaged for integration.

2.2. Model

The model trained by the author is based the open-source framework So-VITS-SVC 4.0 from GitHub. The author has already used it to train the human voice several times before conducting experiments on animals, and found that it has excellent learning ability for the timbre, vocal style, and habits of the datasets, which is very much in line with the author's expectation of attempting to capture the

DOI: 10.54254/2755-2721/68/20241430

characteristics of the animal's voice information from human language text non-contained voices. Singing voice conversion (SVC) is considered more challenging because: (1) compared to normal speech, it involves a wider range of varieties in pitch, energies, expressions, and singing style, (2) from the pitch information perspective, while the generated singing voice needs to follow the notes of the song, the singing style can vary from singer to singer, thus the level of disentanglement needs to be properly modeled [6]. These may be the reason of why So-VITS-SVC can better capture the details of animal vocalizations not only on singing synthesis but also speech synthesis than normal speech conversion model.

3. Results and discussion

3.1. Model construction

So-VITS-SVC4.0 serves as a framework only and does not possess speech synthesis functionality by itself. All functionalities require users to train the models independently. So the author first resample the audio in datasets to 44.1k Hz, then partition the training set, validation set, test set, and config.json files, including all the information during training. After that, it is essential to load pre-trained models. Finally, the training can be started, and training parameters such as loss values are observed at this time. The author checked data trends repeatedly to ensure final training results. The full details of the training process is written in the training log. The training stopped at 163600 steps. Then a cluster model started to be trained, which could reduce tone leakage, making the model trained to sound more like the target tone. During the training process, the author used the following parameters:

- log_interval: 200
- eval_interval: 800
- epochs: 10000
- learning_rate: 0.0002
- eps: 1e-09
- batch_size: 12
- lr_decay: 0.999875
- segment_size: 10240
- init_lr_ratio: 1
- c_mel: 45
- c_kl: 1.0
- max_speclen: 512
- keep_ckpts: 20

3.2. Evaluation results

The loss total results are shown in Fig. 1. As one of the important research topics in machine learning, loss function plays an important role in the construction of machine learning algorithms and the improvement of their performance, which has been concerned and explored by many researchers [7]. The Loss Function is a function that measures the difference between the predicted and true values of a model [8]. Typically, one calculates the loss values for each sample and add them together to get the overall loss. It is expected to minimize the overall loss, which means that the model fits the training data better, thus improving its generalization to new data.

DOI: 10.54254/2755-2721/68/20241430

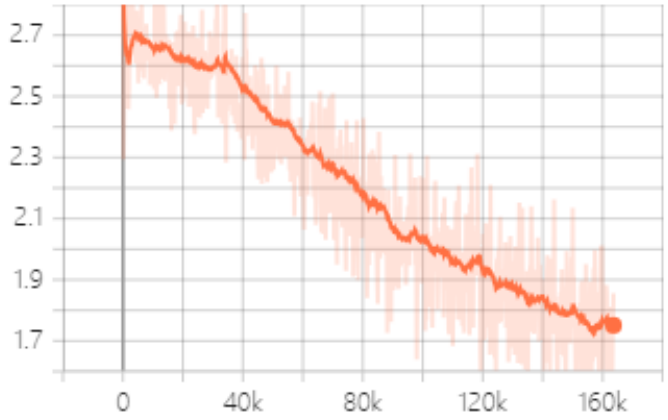

Figure 1. Loss total (Photo/Picture credit: Original).

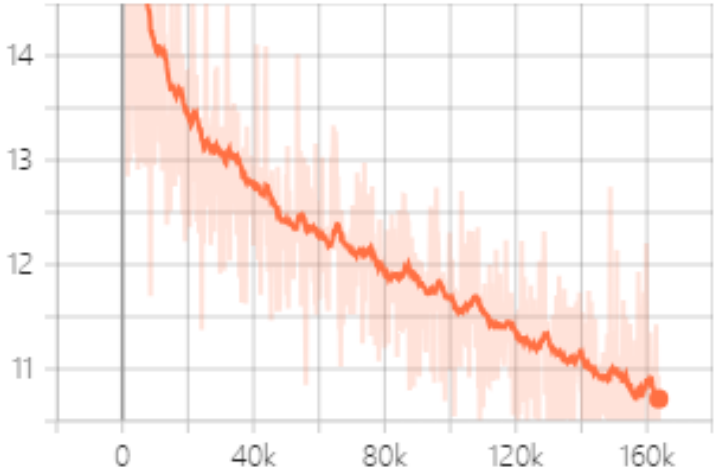

Figure 2. Loss_mel (Photo/Picture credit: Original).

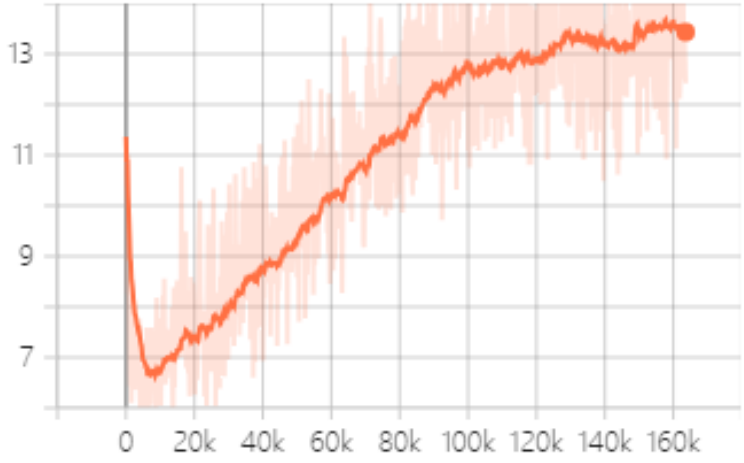

Figure 3. Loss_fm (Photo/Picture credit: Original).

In addition, loss_mel is presented in Fig. 2, which is reconstruction loss, derived from L1 loss. Roughly describing the model pronunciation accuracy. For loss_fm (presented in Fig. 3), the loss of distance between the generated waveforms and the real waveforms in the discriminator intermediate features in adversarial training. Besides the loss value, other metrics also need to be considered to assess how good the model is, such as accuracy, precision, recall, F1 score.

DOI: 10.54254/2755-2721/68/20241430

3.3. Limitations and prospects

Although this model has been partially successful in using animal voices for speech and singing synthesis, it needs to be combined with a complete animal language translation system and speech recognition technology, such as Neural networks, which are capable of translating between languages. i.e., in some cases even between two languages where there is little or no access to parallel translations, in what is known as Unsupervised Machine Translation (UMT), if it is going to be truly effective for real-life applications to facilitate human-animal communication and interaction [9]. This model can only enable humans to understand animals more intuitively, but how to enable animals to understand humans well, such as using human language to correspond to animal semantics for animal sound synthesis, including some ultrasonic or infra-sound, etc., will be a huge and complex subject that still needs to be explored.

As technology has become more ubiquitous, spreading from workplaces to everyday life, the field of HCI has adopted new ways to investigate how one interacts with such pervasive technology [10]. When some animals are raised as pets and their keepers have a special emotional relationship to their single individuals, people may prefer to hear their pet's distinctive timbre rather than the average timbre of the species as a whole. In this case, there should also be a more refined approach in the future on how to do more fine-grained tonal mimicry of individual animals. Firstly, a deeper understanding of the acoustic features that contribute to the uniqueness of each animal's voice is required. This may involve studying factors such as pitch, rhythm and timbre, as well as the effects of individual anatomical and physiological differences. By training models on larger datasets of individual animal sounds, researchers can develop algorithms that capture the subtle details and individualized characteristics of each animal's voice.There are also ethical issues that need to be considered when developing finer-grained individual animal tone imitation techniques. It is important to ensure the responsible and ethical use of such technologies. However, one can really expect to achieve deeper and more comprehensive research on animal language and expression with the continuous advancement of technology and the expansion of its application scope.

Animal speech and singing synthesis may also play a significant role in entertainment projects such as movies, television, and games. Perhaps one day, human beings won't need voice actors to mimic animals word for dubbing. To give those characters based on animals more vivid expression, one could directly use the species' vocalizations for dialogue, or even use the singing synthesis for creating characters' songs. Collaborating with machines to create art will be a fascinating attempt for all the artist and audience. By synthesizing realistic animal sounds, it can enhance the authenticity of the works and expands creative possibilities. It adds depth to interactive experiences, allowing players to interact with virtual animals. In the educational field, animal voice synthesis may also be a useful tool to help children to learn about different species of animals. Animal speech and singing synthesis may offer diverse possibilities for entertainment projects, bringing new opportunities for narrative enhancement, creativity, and audience engagement.

4. Conclusion

Overall, the mimicry of animal voices using existing speech and singing synthesis techniques may be an unexpected new area of convergence that appears to be feasible and evolving. From the collection and editing of the datasets in the early stage, to the model training and parameter tuning attempts in the middle stage, as well as the inference and evaluation at the end, the process has encountered difficulties and doubts, but they were all solved in the end. The attempt and success of this model training may open up a more intuitive way of thinking for animal language translation, and will also enable more people to really hear animals voices and emotions, promote the harmonious relationship between animals and humans, and raise the awareness of animal protection in a subtle way.

DOI: 10.54254/2755-2721/68/20241430

References

[1] Pepperberg I M. 2017 Animal language studies: What happened?. Psychonomic Bulletin & Review vol 24(1) pp 181-185.

[2] Christopher K, Simone B, Butovens M and Ramesh B. 2017 Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. Journal of The Royal Society Interface vol 14 p 135.

[3] Barón B, Leonardo. 2016 Animal Communication and Human Language: An overview. International Journal of Comparative Psychology vol 29 p 1.

[4] Liu J, Li C, Ren Y, Chen F and Zhao Z. 2022 DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. Proceedings of the AAAI conference on artificial intelligence 36(10) pp 11020-11028.

[5] Natsiou A and O'Leary S. 2021 Audio representation for deep learning in sound synthesis: A review. IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA) pp 1-8.

[6] Huang W C, Violeta L P, Liu S, Shi J and Toda T. 2023 The Singing Voice Conversion Challenge 2023. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) pp 1-8.

[7] Wang Q, Ma Y, Zhao K and Tian Y. 2020 A Comprehensive Survey of Loss Functions in Machine Learning. Annals of Data Science pp 1-26.

[8] Niu M, He L, Fang Z, Zhao B and Wang K. 2022 Pseudo-Phoneme Label Loss for Text-Independent Speaker Verification. Applied Sciences vol 12(15) p 7463.

[9] Goldwasser S, Gruber D, Kalai A T and Paradise O. 2024 A Theory of Unsupervised Translation Motivated by Understanding Animal Communication. Advances in Neural Information Processing Systems vol 36.

[10] Caramiaux B and Donnarumma M. 2021 Artificial Intelligence in Music and Performance: A Subjective Art-Research Inquiry. Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity pp 75-95.