Vol.2018-MUS-118 No.4
Vol.2018-SLP-120 No.4
2018/2/20

Recent Advances in our Neural Parametric Singing Synthesizer

Jordi Bonada[1,a)] Merlijn Blaauw[1,b)]

Abstract: We recently presented a new model for singing synthesis based on a modified version of the WaveNet architecture. Instead of modeling raw waveform, we model features produced by a parametric vocoder that separates the influence of pitch and timbre. This allows conveniently modifying pitch to match any target melody, facilitates training on more modest dataset sizes, and significantly reduces training and generation times. Nonetheless, compared to modeling waveform directly, ways of effectively handling higher-dimensional outputs, multiple feature streams and regularization become more important with our approach. We include additional components for predicting F0 and phonetic timings from a musical score with lyrics. These expression-related features are learned together with timbrical features from a single set of natural songs. Here we describe our recent advances on multisinger and multiple voice quality models.

Keywords: Singing synthesis, Deep learning, Conditional generative models, Autoregressive models

1. Introduction

Singing voice is one of the most difficult instruments to model and synthesize. It features a rich and varied palette of timbres, and its expressivity depends on a complex mixture of musical aspects, emotional state and lyrics semantics. Most of the current successful singing synthesizers are based on concatenative approaches [1], [2]. In these approaches, short audio units are selected from a collection of singer recordings, transformed to match a musical target, and concatenated to generate continuous singing. In general, these systems offer good sound quality and naturalness, although they tend to be limited in terms of flexibility. A significant limitation is that usually timbre and expression must be modeled separately from different and specialized corpora.

By contrast, machine learning approaches offer an increased flexibility, and are able to jointly model timbre and expression from a single corpus of natural songs (e.g. statistical parametric methods [3], [4]). However, until recently, these approaches have not been able to match the sound quality of concatenative methods, mostly due to suffering from oversmoothing in time and frequency.

Recent advances in generative models for Text-to-Speech Synthesis (TTS) using Deep Neural Networks (DNNs) have recently proved that model-based approaches can match or even go beyond the sound quality achieved by concatenative methods. In particular, the Wavenet model [5] is able to accurately generate sample-by-sample raw speech waveform without suffering from oversmoothing.

[1] Universitat Pompeu Fabra, Barcelona, Spain
[a)] jordi.bonada@upf.edu
[b)] merlijn.blaauw@upf.edu

2. A Neural Parametric Synthesizer

We recently presented a new model for singing synthesis based on the Wavenet architecture, reported in detail in [6] and [7]. Our goal is to mimic the task of a singer interpreting a musical score with lyrics. The entire system is depicted in **Fig. 1**. It consists of three components for predicting timing, F0 and timbre.

The core architecture of our system is based on the Wavenet model, a probabilistic and autoregressive model that uses dilated gated convolutional units with residual and skip connections. That means that predicted outputs at each timestep depend on past predictions, and that the system predicts a probabilistic distribution instead of output values. Additionally, in order to control the synthesis the system is conditioned on a sequence of phonetic and musical control inputs besides past predictions. An important difference with the original Wavenet model is that in our approach we model vocoder features instead of raw waveform. One reason for that is that a vocoder decomposes the voice signal into phonetic and pitch components, and this allows us to train the models with less data to sufficiently cover the pitch-timbre space, while still being able to synthesize any melody and lyrics. Another reason is that we consider that the degradation introduced by the model itself is the dominant factor in the sound quality. Therefore, we expect our system could ideally achieve a sound quality close to the upper bound the vocoder can provide, i.e. a resynthesis without modification.

Since our model predicts an entire frame at once, we opted to use a mixture density output (similar to [8]) instead of categorical distributions (using a softmax output) that would largely increase the parameter count. Additionally, for minimizing the exposure bias of our model we found necessary to regularize it, in our case by simply adding gaussian noise to the input.

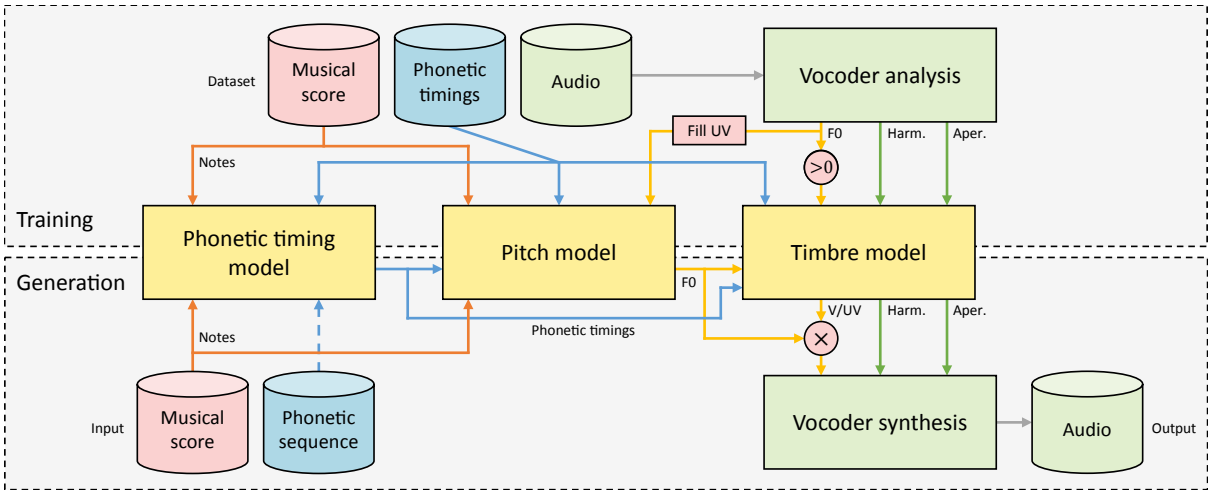

Fig. 1 Diagram depicting an overview of the system with its different components. Here, V/UV is the predicted voice/unvoiced decision, and the Fill UV block fills unvoiced segments by interpolation.

As detailed in [6] and [7], we found our approach to produce more natural variation of predicted parameters over time than statistical parametric systems, and to a be more robust to misalignments between phonetic and acoustic features in the training data than concatenative systems. Additionally, listening tests showed our system to reduce the gap between the reference and the second best system by about a third.

3. Recent advances

Recent ongoing research shows promising results in the area of multisinger models [9]. Concretely, we trained a single model with recordings of 12 choir singers (3 per section: bass, tenor, alto, soprano). For both training and inference, a singer identity (one-hot encoded) is appended to the conditioning vector. For achieving more realistic timing and F0 models in the scope of choir singing, we are currently preparing a recording of a full choir where one directional microphone will be placed in front of each singer. We expect this approach will allow us to simultaneously capture the utterances of each individual in a realistic choir singing scenario, and to gather sufficient data for modeling interdependencies between singers.

Other ongoing research on multiple voice quality shows also promising results. Concretely, we did experiments with recordings containing a limited coverage of soft, modal and powerful singing. Preliminary results show that the model can successfully generalize enough to synthesize voice qualities in unseen phonetic and musical contexts.

We hope in the near future to explore the flexibility offered by this neural approach to model more challenging natural and expressive singing including non-modal voice qualities (such as rough voices) and expressive resources (such as growls).

References

[1] Bonada, J., Umbert, M. and Blaauw, M.: Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016, *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*, San Francisco, CA, USA, pp. 1230–1234 (September 8–12, 2016).

[2] Bonada, J. and Serra, X.: Synthesis of the Singing Voice by Performance Sampling and Spectral Models, *IEEE Signal Processing Magazine*, Vol. 24, No. 2, pp. 67–79 (2007).

[3] Saino, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K.: An HMM-based singing voice synthesis system, *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, Pittsburgh, PA, USA, pp. 2274–2277 (September 17–21, 2006).

[4] Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y. and Tokuda, K.: Recent development of the HMM-based singing voice synthesis system - Sinsy, *Proceeedings of the 7th ISCA Workshop on Speech Synthesis (SSW7)*, Kyoto, Japan, pp. 211–216 (September 22–24, 2010).

[5] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *CoRR*, Vol. abs/1609.03499 (2016).

[6] Blaauw, M. and Bonada, J.: A Neural Parametric Singing Synthesizer, *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*, Stockholm, Sweden, pp. 1230–1234 (August 20–24, 2017).

[7] Blaauw, M. and Bonada, J.: A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs, *Applied Sciences 2017, 7, 1313.*

[8] Salimans, T., Karpathy, A., Chen, X. and Kingma, D. P.: PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications, *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France (April 24–26, 2017).

[9] CASAS project (Community-Assisted Singing Analysis and Synthesis) Available online: www.upf.edu/web/mtg/casas (accessed January 22, 2018).

Jordi Bonada received the Ph.D. degree in Computer Science and Digital Communications from the Universitat Pompeu Fabra (UPF). Since 1996 he has been a researcher at the Music Technology Group of the UPF while leading several projects funded by public and private institutions. Dr. Bonada has a long research experience supported by more than 80 scientific publications and over 50 patents. Some of the algorithms he has proposed have been integrated into successful commercial products such as Vocaloid.

Merlijn Blaauw is currently pursuing his Ph.D. degree in Information and Communication Technology at the Universitat Pompeu Fabra (Barcelona, Spain). He has been working as a researcher and developer at the Music Technology Group (MTG) of same institute after completing is Master degree in Music Technology from the HKU (Utrecht, The Netherlands) in 2004. His research focuses on synthesis and modeling of singing voice using machine learning. Notable industrial research projects on which he has collaborated include VOCALOID with Yamaha (Japan).