INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Singing Voice Synthesis Using Deep Autoregressive Neural Networks for Acoustic Modeling

Yuan-Hao Yi, Yang Ai, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R. China
{yiyh,ay8067}@mail.ustc.edu.cn, {zhling,lrdai}@ustc.edu.cn

Abstract

This paper presents a method of using autoregressive neural networks for the acoustic modeling of singing voice synthesis (SVS). Singing voice differs from speech and it contains more local dynamic movements of acoustic features, e.g., vibratos. Therefore, our method adopts deep autoregressive (DAR) models to predict the F0 and spectral features of singing voice in order to better describe the dependencies among the acoustic features of consecutive frames. For F0 modeling, discretized F0 values are used and the influences of the history length in DAR are analyzed by experiments. An F0 post-processing strategy is also designed to alleviate the inconsistency between the predicted F0 contours and the F0 values determined by music notes. Furthermore, we extend the DAR model to deal with continuous spectral features, and a prenet module with self-attention layers is introduced to process historical frames. Experiments on a Chinese singing voice corpus demonstrate that our method using DARs can produce F0 contours with vibratos effectively, and can achieve better objective and subjective performance than the conventional method using recurrent neural networks (RNNs).

Index Terms: singing voice synthesis, deep autoregressive model, self-attention, recurrent neural network

1. Introduction

Singing voice synthesis (SVS) converts lyrics and musical score information (e.g., tempo, pitch, etc.) into songs, which differs from traditional text-to-speech (TTS) synthesis. Some song synthesizers have been developed based on the unit selection speech synthesis approach [1, 2]. Although this approach can achieve high sound quality, it relies on large corpora and its flexibility is limited. On the other hand, a statistical parametric approach to SVS based on hidden Markov models (HMMs) [3] has also been studied. However, this method can't generate singing voice with high naturalness because of the over-smoothing issue of HMM modeling.

Recently, various neural networks, such as deep neural network (DNN) [4] and recurrent neural network with long-short term memory (LSTM-RNN) [5], have been applied to speech synthesis and demonstrated their superiority over traditional HMM-based ones. Some new neural models, such as Tacotron [6], WaveNet [7] and WaveRNN [8], have also been proposed to improve the acoustic modeling and waveform generation of statistical parametric speech synthesis. For SVS, DNN and LSTM-RNN have been adopted for acoustic modeling [9, 10]. A neural network model similar to WaveNet has also been proposed for modeling spectral features in SVS [11].

The differences between singing voice and common speech should be considered when designing SVS methods. First, there are plenty of linguistic-independent dynamic movements of acoustic features in singing voice. For example, the F0 contours of singing voice contain a lot of *singingness*-related dynamic F0 patterns, such as vibrato, overshoot, preparation, and fine-fluctuation [12].

The spectral features of singing voice are also affected by these kinds of F0 movements. However, it is difficult to model these local dynamic characteristics of acoustic features using conventional DNNs or LSTM-RNNs directly. Second, the predicted F0 contours should be consistent with the input music notes, which can not be guaranteed by state-of-the-art acoustic models for SVS. The synthetic voice may be perceived as out of tune if the predicted F0 contours deviate too much from the pitch determined by music notes.

Therefore, this paper proposes to adopt deep autoregressive (DAR) [13] models for predicting the F0 and spectral features of singing voice in order to better describe the dependencies among the acoustic features of consecutive frames. For F0 modeling, discretized F0 values are used and the influences of the history length in DAR are analyzed by experiments. An F0 post-processing strategy is also designed to alleviate the inconsistency between the predicted F0 contours and the stair-like F0 contours determined by music notes. For spectral modeling, the original DAR model is extended to deal with continuous spectral features, and a prenet module with self-attention [14] layers is introduced to process historical frames. Finally, a WaveRNN vocoder [8] is built to synthesize the waveforms of singing voice from the predicted F0 and spectral features. Experiments on a Chinese singing voice corpus show that our method using DARs can produce F0 contours with vibratos effectively, and can achieves better objective and subjective performance than RNN-based acoustic modeling.

This paper is organized as follows. In Section 2, we briefly review the basic DAR model and describe the details of our proposed method. Section 3 reports our experimental results. Conclusions are given in Section 4.

2. DAR-based Singing Voice Synthesis

2.1. Basic DAR models

The autoregressive (AR) dependency has been widely studied for many signal modeling and generation tasks. Deep autoregressive (DAR) models [13] follow the idea of feeding the target data of previous frames as additional input to a uni-directional recurrent layer [15]. Assume that $\boldsymbol{o}_t$ stands for the model output at the t-th frame. At the t-th time step of a DAR model,

http://dx.doi.org/10.21437/Interspeech.2019-1563

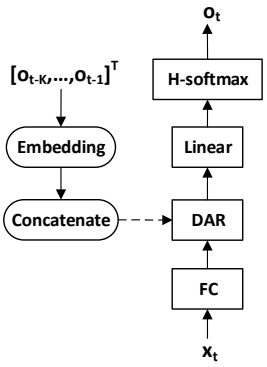

Figure 1: *The structure of our DAR-based F0 model for SVS.*

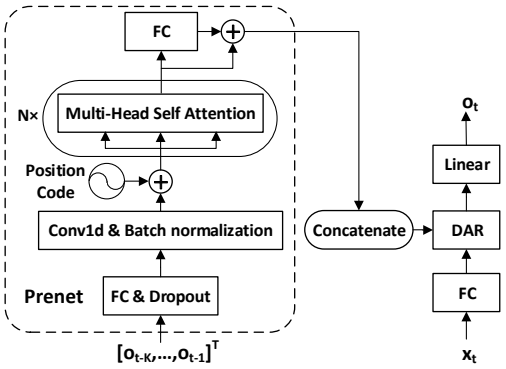

Figure 2: *The structure of our DAR-based spectral model for SVS.*

the sequence of K history outputs, i.e., $[\boldsymbol{o}_{t-k}, ..., \boldsymbol{o}_{t-1}]$, are concatenated, and are sent into the recurrent unit together with the hidden representations of the $(t-1)$-th time step. The feedback path of history outputs is referred to as feedback link in this paper.

The DAR model for speech synthesis was initially proposed for modeling the sequences of quantized F0 values [13]. The F0 quantization is achieved by first mapping each original F0 into Mel scale and then quantizing it into N levels. Thus, the F0 output at each frame can be encoded into a one-hot vector $\boldsymbol{o_t} = [o_{t,0}, o_{t,1}, \ldots, o_{t,N}]^\top$, where $o_{t,i} \in \{0, 1\}$. For unvoiced frames, we have $o_{t,0} = 1$. At the training stage, H-softmax (hierarchical softmax) [16] is adopted as the final output layer to handle the unbalanced data distribution caused by the large amount of unvoiced frames. A data dropout strategy is designed to alleviate the issue that the DAR may only copy the feedback link while ignore the input features at current frame. At the generation stage, the t-th frame is classified as an unvoiced one if $o_{t,0} \geq 0.5$. Otherwise, this frame is determined as a voiced one and its F0 quantization level is predicted.

2.2. DAR-based F0 model for SVS

The structure of our DAR-based F0 model for SVS is shown in Figure 1. It is almost the same as the original DAR for F0 modeling [13] and one difference is that GRU instead of LSTM is adopted at recurrent layers for simplification. The frame-level context features $\boldsymbol{x_t}$, which contain both linguistic and music score information, is first passed through fully connected (FC) layers. Then, the DAR module consists of a bidirectional GRU layer followed by a unidirectional GRU layer with data dropout strategy. The K history outputs of quantized F0 values are first passed through an embedding layer. The embedding vectors are then concatenated and fed into the unidirectional GRU layer as additional inputs. At the generation stage, the F0 values of voiced frames are predicted by mean-based generation [13] in our method.

In order to reduce the deviation between the predicted F0 contours and the pitch determined by music notes, and to alleviate the out-of-tune issue in synthetic voice, an F0 post-processing strategy is proposed in this paper. This strategy is achieved by performing a moving average on the F0 contours predicted by models and replacing the slow-change components with melody components [17], i.e., the stair-like F0 contours determined by music notes. Mathematically, let f_t denote the predicted F0 value at the t-th frame after dequantizaiton. The post-processed F0 value $\widehat{f}_t$ can be calculated as

$$\widetilde{f}_t = \frac{1}{2w+1} \sum_{i=t-w}^{t+w} f_i, \tag{1}$$

$$\widehat{f}_t = f_t - \widetilde{f}_t + f_t^{(n)}, \tag{2}$$

where $2w+1$ represents the window size for moving average, $\widetilde{f}_t$ denotes the output of moving average and and $f_t^{(n)}$ represents the t-th frame of the stair-like F0 contours determined by music notes. The above operations are performed independently for each voiced segment. The first and last frames in each voiced segment are copied w times to provide data for calculating moving average and the beginning and the end of the segment.

2.3. DAR-based spectral model for SVS

In contrast to our DAR-based F0 model, continuous spectral features, i.e., mel-spectral coefficients (MCCs) and energy, are modelled directly in our DAR-based spectral model. As depicted in Figure 2, the most significant difference to Figure 1 is that a prenet module is designed to process the history outputs of continuous spectral features. In our preliminary experiments, we found that directly feeding the history spectral features into the DAR module didn't work well. One possible reason is the significant difference between the distribution spaces of spectral features and the input context features processed by FC layers. In the F0 model shown in Figure 1, the embedding layer contributes to unifying these two spaces while it can't be employed directly for continuous features. Therefore, a prenet module with self-attention [14] layers is introduced to process historical frames of spectral features and to extract high-level representations as our feedback link for spectral modeling.

In the prenet module, the sequence of K history frames $[\boldsymbol{o}_{t-k}, \cdots, \boldsymbol{o}_t]$ first pass through FC layers with dropout [18] and convolution layers with batch normalization [19]. Then, a d-dimensional position code [14] $\boldsymbol{p}_t = [p_t(0), \ldots, p_t(d-1)]^\top$ is added to the output of batch normalization to provide explicit position information for each frame in the history. The elements in $\boldsymbol{p}_t$ are calculated as

$$p_t(2i) = sin(n/10000^{2i/d}), \tag{3}$$

$$p_t(2i+1) = cos(n/10000^{2i/d}), \tag{4}$$

where $i \in [0, \ldots, d/2-1]$ is the dimension index. Then, N multi-head self attention layers are stacked. Each layer has h heads with scaled dot-product attention and adopts masks in

Table 1: *The performance of F0 prediction on the validation set with different history length K.*

	K=1	K=2	K=3	K=4
F0 RMSE (Hz)				
-Natural	21.51	**20.78**	21.18	22.90
-Music Note	19.21	**19.11**	19.14	21.62
CORR				
-Natural	0.95	0.96	0.96	0.95
-Music Note	0.96	**0.97**	0.96	0.96
V/UV ERROR (%)	2.37	**2.35**	2.36	2.38

the form of upper triangular matrix to ensure the autoregressive causality. Finally, a FC layer with residual connections is utilized to produce the outputs of the prenet.

3. Experiments

3.1. Experimental conditions

A Chinese singing voice corpus was adopted in our experiments. This corpus contained 3290 utterances (100 songs about 220 minutes) without background music from a male singer. The recordings were sampled at 16kHz with 16-bit quantization. This dataset was separated into a training set with 2976 utterances (91 songs), a validation set with 82 utterances (2 songs), and a test set with 232 utterances (7 songs). The 43-dimensional acoustic features at each frame, including 40 MCCs, 1 energy, 1 F0, and 1 voiced/unvoicde (V/UV) flag, were extracted by STRAIGHT [20] with 40ms window size and 5ms frame shift.

3.2. System construction

Two acoustic models, one RNN-based baseline model and one proposed DAR model, were built for comparison. In these models, the input context features at each frame were 1969-dimensional, including 1959 binary answers to context-related questions, 9 numerical values describing the position of current frame, and 1 numerical value describing the music note that current frame belonged to. The phone and state boundaries were obtained by HMM-based force alignment [21]. This paper doesn't investigate the duration modeling for SVS. Thus, the segmentation results of natural recordings were used at synthesis time for both acoustic models. A WaveRNN-based vocoder [8, 22] was built to reconstruct 16-bit waveforms given the predicted frame-level acoustic features.

3.2.1. Baseline model

The baseline model had 3 bidirectional GRU layers with 1024 units per layer and 1 fully connected output layer. This structure was determined after tuning on the validation set. The output acoustic features were composed of the static, delta and delta-delta components of MCCs, energies and F0s, together with a V/UV flag. An Adam optimizer [23] with a learning rate of 1e-3 was used to update the parameters to minimize the mean square error (MSE) of model prediction on the training set. The final acoustic features were generated from the model outputs by maximum likelihood parameter generation (MLPG) algorithm.

3.2.2. DAR model

For building the DAR-based F0 model, the F0 values in Hz were first transformed to Mel scale using $mel = 1127log(1 + F0/700)$, and were then quantized into 255 levels between 106 and 831 on the Mel scale according to the data distribution of training set. The dynamic F0 components were not used here. The model structure shown in Figure 1 contained 2 FC layers with 512 tanh units per layer. The bidirectional GRU layer in the DAR module contained 256 units and the unidirectional GRU layer contained 128 units. The following linear layer had 256 units. The dropout rate of the feedback link was set as 0.75. An Adam optimizer [23] with exponential decay of learning rate was used to update the parameters by minimizing the cross-entropy of F0 prediction on training set. The initial learning rate was 0.01 and the decay rate was 0.9886 per 5000 learning steps. The history length K was tuned on the validation set and the results are shown in Table 1. The F0 post-processing strategy introduced in Section 2.2 was not applied here. In Table 1, **F0 RMSE** and **CORR** mean the root mean square error (RMSE) and the Pearson correlation coefficient between the predicted and the reference F0 contours. **V/UV error** denotes the percentage of frames with incorrect V/UV flag prediction. *Natural* and *Music Notes* stands for using the F0 contours extracted from natural recordings and the F0 contours determined by music notes as references respectively. From this table, we can see that using more history frames may not always improve the accuracy of F0 prediction. $K = 2$ achieved the best performance on all metrics. For the post-processing, the window size for moving average was also tuned on the validation set and $w = 15$ achieved the best performance.

Table 2: *The MCD (dB) of spectral prediction on the validation set with different history length K, head number h and self-attention layer number N.*

	N=1	N=2	N=3	N=4
K=1, h=1	4.42	4.02	4.03	4.23
K=1, h=2	4.82	4.72	4.00	4.63
K=1, h=4	4.96	4.83	4.67	4.74
K=1, h=8	5.00	4.89	4.72	4.98
K=2, h=1	3.96	3.78	3.67	3.70
K=2, h=2	3.82	3.72	**3.52**	3.61
K=2, h=4	3.73	3.80	3.71	3.76
K=2, h=8	3.74	3.87	3.89	3.79
K=3, h=1	5.47	5.33	5.12	5.46
K=3, h=2	5.39	5.29	5.05	5.44
K=3, h=4	5.73	5.24	5.67	5.78
K=3, h=8	5.93	5.86	5.42	5.79

For building the DAR-based spectral model, its continuous output at each frame included 40 MCCs and 1 energy. The DAR module was the same as one for F0 modeling. The only difference was that a linear output layer was adopted. In the prenet module, the first two FC layers had 64 units with ReLU activation per layer, and were followed by 2 dropout layers with 0.1 dropout rate. Then, the following convolution layer used kernel size of 2 and 64 output channels, and were followed by a batch normalization layer with ReLU activation. For multi-head self-attention layers, the outputs of linear projection were 64-dimensional. The final FC layer had 64 units and residual connections. An Adam optimizer [23] with exponential decay of learning rate was used to update the parameters by minimize the MSE of spectral prediction on training set. The initial learning rate was 0.001 and the decay rate was 0.9886 per 250 learning steps. In Table 2, we compared the performance of spectral prediction on the validation set with different history length K, head number h and self-attention layer number N.

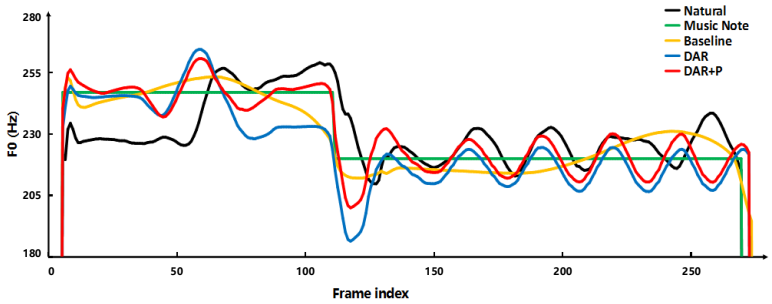

Figure 3: *F0 contours generated by different models for a voiced segment in our test set, where "Natural" and "Music Note" are two references.*

Table 3: *Accuracies of acoustic feature prediction using the baseline model (Baseline), the DAR model without F0 post-processing (DAR), and the DAR model with F0 post-processing (DAR+P) on test set.*

	Baseline	**DAR**	**DAR+P**
F0 RMSE (Hz)			
-Natural	35.09	20.71	**20.36**
-Music Note	34.42	19.14	**8.45**
CORR			
-Natural	0.89	0.96	0.96
-Music Note	0.89	0.97	**0.99**
V/UV ERROR (%)	2.57	**2.35**	**2.35**
MCD (dB)	4.16	**3.51**	**3.51**

Similar to F0 modeling, the optimal history length was $K = 2$ and the optimal mel-cepstral distortion (MCD) was achieved by $h = 2$ and $N = 3$.

3.3. Objective evaluation

We compared the accuracies of acoustic feature prediction using the baseline model (**Baseline**), the DAR model without F0 post-processing (**DAR**), and the DAR model with F0 post-processing (**DAR+P**) on test set. The results are shown in Table 3. From this table, we can see that the DAR-based F0 model achieved lower F0 RMSE and higher correlation coefficient than the baseline model when either natural F0 contours or the F0 contours determined by music notes were used as references. The DAR-based models also achieved lower V/UV error and MCD than the baseline model. After applying the F0 post-processing strategy introduced in Section 2.2, the F0 prediction accuracy got further improved, especially for the metrics using the F0 contours determined by music notes as references.

Figure 3 shows the F0 contours generated by different models for a voiced segment in our test set. We can observe that the F0 contour determined by music notes was stair-like and there were plenty of local dynamic movements, e.g., vibratos, in the F0 contour extracted from natural speech. The F0 contour predicted by the baseline model was over-smoothed and failed to reproduce vibratos. In contrast, vibratos can be generated effectively by the DAR-based F0 model. The proposed F0 post-processing strategy further alleviated the inconsistency between the overall shape of the predicted F0 contour and the F0 values determined by music notes.

3.4. Subjective evaluation

Subjective listening tests were carried out to evaluate the preference scores between the songs synthesized by different methods. 5 songs and 6 utterances in each song were randomly selected from our test set and were synthesized by the three methods listed in Table 3. Two preference tests were conducted to compare **Baseline** with **DAR**, and **DAR** with **DAR+P** respectively. 10 Chinese native listeners participated in each test using headphones. Each pair of synthetic utterances were presented to a listener by random order, and the listener was asked to make a choice among, 1) the former was better, 2) the latter was better, and 3) there was no preference. The average preference scores are shown in Figure 4. It shows that our proposed method using DARs for acoustic modeling was significantly preferred than the RNN-based baseline method ($p < 0.001$). This can be attributed to the advantages of DARs at modeling the temporal dependency of acoustic features across frames. The preference scores between **DAR** and **DAR+P** show that the proposed F0 post-processing strategy further improved the performance of DAR-based F0 prediction ($p < 0.01$). These results are consistent with the objective ones shown in Table 3.

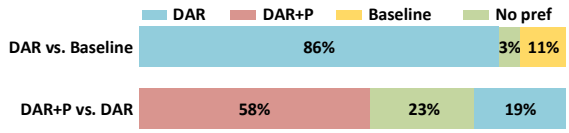

Figure 4: *The subjective preference scores among Baseline, DAR and DAR+P.*

4. Conclusions

In this paper, we have presented a method of using deep autoregressive (DAR) neural networks to model F0s and spectral features in singing voice synthesis (SVS). For F0 modeling, discretized F0 values are used and a moving average-based F0 post-processing strategy is designed to alleviate the inconsistency between the predicted F0 contours and the F0 values determined by music notes. Furthermore, a DAR-based spectral model is proposed by designing a prenet module with self-attention layers. Objective and subjective experimental results have demonstrated the effectiveness of our proposed method. To investigate the neural network-based methods of duration modeling for SVS will be our work in the future.

5. Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2017YFB1002202) and the National Nature Science Foundation of China (Grant No. 61871358 and U1613211).

Some samples of generated speech can be found at `http://home.ustc.edu.cn/~yiyh/interspeech2019`.

6. References

[1] H.-Y. Gu and J.-K. He, "Singing-voice synthesis using demi-syllable unit selection," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2. IEEE, 2016, pp. 654–659.

[2] J. Bonada, M. Umbert, and M. Blaauw, "Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016." in *INTERSPEECH*, 2016, pp. 1230–1234.

[3] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Ninth International Conference on Spoken Language Processing*, 2006.

[4] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 3844–3848.

[5] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[7] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[8] Y. Ai, J.-X. Zhang, L. Chen, and Z.-H. Ling, "DNN-based spectral enhancement for neural waveform generators with low-bit quantization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[9] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on deep neural networks." in *Interspeech*, 2016, pp. 2478–2482.

[10] J. Kim, H. Choi, J. Park, S. Kim, J. Kim, and M. Hahn, "Korean singing voice synthesis system based on an LSTM recurrent neural network," in *INTERSPEECH 2018*. International Speech Communication Association, 2018.

[11] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," *arXiv preprint arXiv:1704.03809*, 2017.

[12] T. Saitou, M. Unoki, and M. Akagi, "Extraction of F0 dynamic characteristics and development of F0 control model in singing voice." Georgia Institute of Technology, 2002.

[13] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive neural F0 model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1406–1419, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] M. Schuster, "Better generative models for sequential data problems: Bidirectional recurrent mixture density networks," in *Advances in Neural Information Processing Systems*, 2000, pp. 589–595.

[16] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model." in *Aistats*, vol. 5. Citeseer, 2005, pp. 246–252.

[17] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech communication*, vol. 46, no. 3-4, pp. 405–417, 2005.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[21] N. Ueda and R. Nakano, "Deterministic annealing em algorithm," *Neural networks*, vol. 11, no. 2, pp. 271–282, 1998.

[22] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.