Speech-like Emotional Sound Generator by WaveNet

Kento Matsumoto*, Sunao Hara† and Masanobu Abe‡
* Okayama University, Japan
E-mail: k_matsu@a.cs.okayama-u.ac.jp
† Okayama University, Japan
E-mail: hara@okayama-u.ac.jp
‡ Okayama University, Japan
E-mail: abe@cs.okayama-u.ac.jp

***Abstract*—In this paper, we propose a new algorithm to generate Speech-like Emotional Sound (SES). Emotional information plays an important role in human communication, and speech is one of the most useful media to express emotions. Although, in general, speech conveys emotional information as well as linguistic information, we have undertaken the challenge to generate sounds that convey emotional information without linguistic information, which results in making conversations in human-machine interactions more natural in some situations by providing non-verbal emotional vocalizations. We call the generated sounds "speech-like", because the sounds do not contain any linguistic information. For the purpose, we propose to employ WaveNet as a sound generator conditioned by only emotional IDs. The idea is quite different from WaveNet Vocoder that synthesizes speech using spectrum information as auxiliary features. The biggest advantage of the idea is to reduce the amount of emotional speech data for the training. The proposed algorithm consists of two steps. In the first step, WaveNet is trained to obtain phonetic features using a large speech database, and in the second step, WaveNet is re-trained using a small amount of emotional speech. Subjective listening evaluations showed that the SES could convey emotional information and was judged to sound like a human voice.**

I. Introduction

Recently, the intelligibility and naturalness of synthetic speech have been greatly improved, and synthetic speech will likely be widely used in commercial products in the coming years. Such products include smart speakers (e.g., Amazon Echo and Google Home) and voice assistant applications (e.g., Apple Siri). Although synthetic speech plays an important role in these products, most users, unfortunately, are not satisfied with the quality of the synthetic speech. One of the reasons is the lack of emotional expressions; in the situation of a human-machine interaction (HCI), emotional expressions could be the most important factor in generating natural responses. Even though intensive studies have been done in the last two decades, synthetic speech has failed in expressing a range of emotions[1] [2].

To develop emotional speech, one of the most difficult problems is speech data collection. Although a large amount of data is necessary to generate high-quality speech, it is difficult to collect emotional speech data as needed because people other than professional narrators or radio performers cannot utter speech while keeping with a particular emotion for a long time. Moreover, because emotional expressions depend heavily on situations, contexts, and phrases, it is also difficult to design appropriate text materials or setups for recordings. For these reasons, waveform-based speech synthesis[3][4] is not appropriate for emotional speech synthesis and has failed to synthesize expressive speech [5]. On the other hand, HMM-based speech synthesis [6] is promising as it can provide better flexibility in parameter control using interpolation or adaptation [7] [8] [9]. However, it is reported that emotions could not be reproduced well using this approach either [10].

Speech conveys emotional information through two channels: (1) a linguistic channel of words and phrases, and (2) a non-linguistic channel of acoustic features in voice quality and intonation. In some cases, emotional information can be conveyed only through the non-linguistic channel. For example, we can sometimes feel emotions by hearing speech spoken in a language that we do not know at all. It is also reported that non-verbal emotional vocalizations are important when expressing emotions, especially in conversations, and an approach to synthesize emotional speech by Text-To-Speech (TTS) is not applicable because TTS always requires linguistic information [11] [12]. This implies that, to synthesize emotional speech, it is necessary to develop a new algorithm that can deal independently with the linguistic and non-linguistic channels.

In this paper, we propose a new algorithm to generate Speech-like Emotional Sound (SES) that conveys emotional information from the non-linguistic channel, which results in making it possible to deal with emotional information from the non-linguistic channel independently of the linguistic channel. SES might be useful for dialog systems to express simple reactions with emotions, because designing appropriate phrases for simple reactions is sometimes difficult. We call the generated sounds "speech-like" because they include no linguistic information, but they sound as if they were uttered by human beings. For the purpose, we propose to employ WaveNet as a sound generator conditioned by only emotional IDs. The idea is quite different from WaveNet Vocoder that synthesizes speech using spectrum information as auxiliary features. The biggest advantage of the idea is to reduce the amount of emotional speech data for the training.

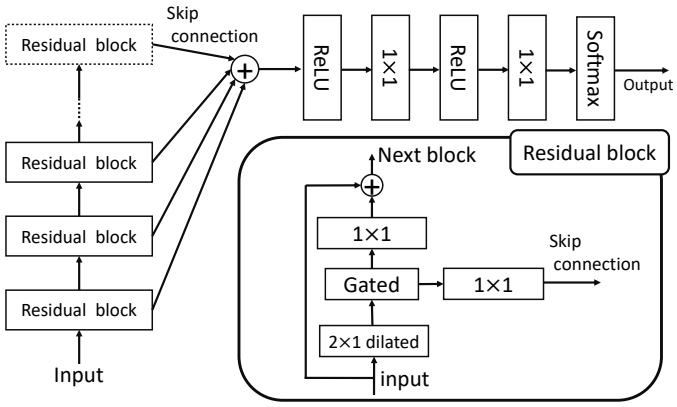

Fig. 1. WaveNet

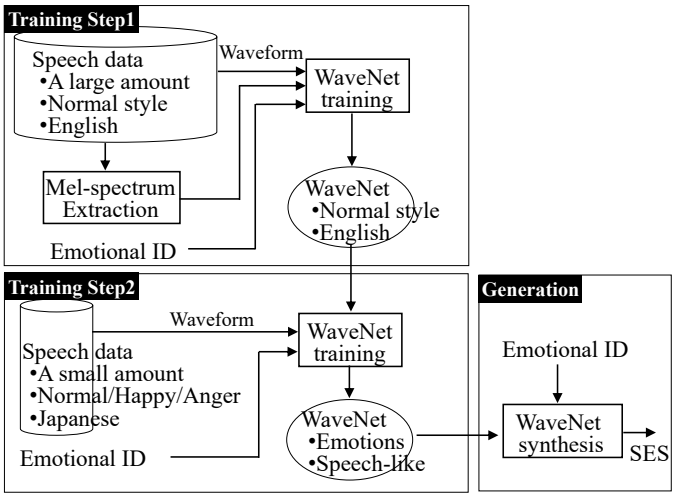

Fig. 2. Outline of the proposed algorithm

The rest of the paper is organized as follows. In Section 2, we describe the basics of WaveNet. In Section 3, we explain the proposed algorithm. In Section 4, we show our evaluation results and provide a discussion. Finally, in Section 5, we present our conclusions and suggest avenues for future work.

II. Basics of WaveNet

A. WaveNet

WaveNet [13] is a Convolutional Neural Network that directly predicts the next sample point using preceding R sample points. Joint probability $p(\mathbf{x})$ of a waveform $\mathbf{x} = \{x_1\ x_2\ \cdots\ x_T\}$ is expressed by the following equation:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}) \quad (1)$$

Eq. (1) shows that each sample point x_t is conditioned on preceding all sample points. Because a large receptive field is needed to predict the waveform, WaveNet uses a dilated causal convolution, which is a convolution where the filter is applied over an area larger than its length by skipping input values with a certain step.

Figure 1 shows a layout of WaveNet where several residual blocks are stacked in the network. Each residual block has a dilated convolution layer. In Fig. 1, "1×1" represents a 1×1 convolution calculation, and "Gated" represents the gated activation function that is defined as follows:

$$z = \tanh(W_{f,k} * \mathbf{x}) \odot \sigma(W_{g,k} * \mathbf{x}) \quad (2)$$

The symbol $*$ denotes a convolution operator, $\odot$ denotes an element-wise product operator, and $\sigma(\cdot)$ denotes a sigmoid function. W_f and W_g refer to the weight of convolution, k is the layer index, f and g denote the filter and gate, respectively. In the output layer, WaveNet predicts a sample point which is quantized by the 8-bit μ-law algorithm as a classification of $2^8 = 256$ classes.

B. Conditional WaveNet

By adding auxiliary features $\mathbf{h}$, we can control WaveNet's outputs. Then, Eq. (2) is modified as follows:

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}) \quad (3)$$

In this case, $\mathbf{y}$ denotes a feature vector which is transformed by $\mathbf{y} = f(\mathbf{h})$ so that the length of the auxiliary features $\mathbf{h}$ is matched to that of the input audio signal $\mathbf{x}$. Now, $V * \mathbf{y}$ is a 1×1 convolution.

III. SES generation algorithm using WaveNet

The proposed algorithm employs the conditional WaveNet and has two steps (Step 1 and Step 2) for model training. In both steps, the WaveNet structure is the same, but the auxiliary inputs and training data are different. An outline of the proposed algorithm is shown in Fig. 2.

A. Training Step 1

Step 1 is a key procedure to generate "speech-like" sounds. WaveNet is trained using speech data uttered with the normal speaking style of a language and the size of the speech data is relatively large. As auxiliary inputs, mel-spectrum parameters and Emotion ID (EID) are used. EID is represented by one-of-k vector, and in Step 1 the normal emotion is always set to 1. In this case, Eq. (3) is rewritten as follows:

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}_{\mathbf{EID}} + U_{f,k} * \mathbf{y}_{\mathbf{mel-spectrum}}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}_{\mathbf{EID}} + U_{f,k} * \mathbf{y}_{\mathbf{mel-spectrum}}) \quad (4)$$

In this equation, $\mathbf{y}_{\mathbf{EID}}$ denotes a feature matrix, which is copied so that the length of the EID vector is matched to that of the input audio signal $\mathbf{x}$, and $\mathbf{y}_{\mathbf{mel-spectrum}}$ denotes a feature matrix which is transformed by transposed convolution so that the length of $\mathbf{y}_{\mathbf{mel-spectrum}}$ is matched to that of the input audio signal $\mathbf{x}$. Also, we assumed that the emotion is kept during an utterance. Now, $U * \mathbf{y}$ is a 1×1 convolution.

According to our preliminary experiments, the size of the speech data should equate to more than 20 hours. When the size is small, WaveNet cyclically generates similar sounds resulting in sounds that are far from "speech-like." The large amount of training data makes it possible for WaveNet to generate sounds with various spectrum patterns. In other

words, we cannot train WaveNet using only emotional speech because it is difficult to collect a large amount of emotional speech data, as mentioned in the introduction.

B. Training Step 2

Step 2 is a key procedure used to generate emotional sounds and to discard linguistic information. WaveNet is re-trained after Step 1 using speech data uttered in several emotional styles and using correspondent EIDs as auxiliary inputs. However, mel-spectrum parameters are not used for the training because we use only EIDs for sound generation. Furthermore, the emotional speech is uttered in a different language from the language used in Step 1. In this case, the gated activation function, shown in the following equation, has removed the term of mel-spectrum from Eq. (4).

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y_{EID}}) \\ \odot\sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y_{EID}}) \quad (5)$$

C. Sound generation

As auxiliary input, the EIDs are set and the first data point is fed to WaveNet. With the data, WaveNet successively generates signals.

IV. Evaluation experiments

To evaluate the performances for generating SES, subjective evaluations were performed from the view points of a speech-like aspect and emotional expression.

A. Model construction

WaveNet was constructed using the conditions shown in Table I. In Step 1, we used the LJ Speech Dataset[14], which contains about 24 hours of speech data uttered by an English female professional narrator. She read English texts aloud using a normal speaking style. In Step 2, the Voice-Actress Corpus [15] was used. A Japanese actress read Japanese texts aloud using three emotions: normal, angry, and happy. Duration of each emotional speech recording is about 17 minutes, and 51 minutes in total. As auxiliary inputs, mel-spectrum parameters were calculated using short-time Fourier transform.

B. Stimuli used in subjective evaluations

For listening tests, we prepared emotional speech uttered in Italian and German. For Italian, we used speech labeled as "normal," "happy," and "angry" from the EMOVO Corpus [17] uttered by a female speaker, "f2." For German, we used speech labeled as "normal," "happy," and "angry" from Berlin Emotional Speech (EMO-DB) [18] uttered by a female speaker, "08." From the two databases, we selected five utterances for each emotion, 30 utterances in total: 2 languages $\times 5$ utterances $\times 3$ emotions. For the 30 utterances, synthesis-by-analysis speech were generated by the WORLD vocoder [19] and MLSA vocoder [20] (hereinafter, they will be referred to as WORLD and MLSA, respectively). In the MLSA vocoder, spectral envelopes extracted by WORLD with a 20 msec window were parameterized to the 0-39th mel-cepstral coefficients and speech was synthesized by a Mel log spectrum approximation (MLSA) filter. Stimuli used in the following evaluations were original speech (hereinafter, ORIGINAL), WORLD, MLSA and speech generated by the proposed algorithm (hereinafter, WAVENET). In terms of WAVENET, the three emotions ("normal," "happy," and "angry") were generated using EID. For each emotion, ten utterances were generated with an average duration of ORIGINAL. In the end, there were 120 stimuli in total. Also generated speeches can be found on our web page[1].

TABLE I
Experimental conditions

Training data	
Corpus	Step 1: The LJ Speech Dataset (24 hours) Step 2: Voice-Actress Corpus (137 minutes)
Sampling freq.	16 kHz
Training data	Step 1: 13,100 utterances (24 hours) Step 2: 285 utterances (51 minutes)

Speech analysis	
Window length	64 msec
Frame shift	16 msec

WaveNet configuration	
Iterations	Step 1: 770,000 iterations Step 2: 40,000 iterations
Mini batch size	4
Optimization	Adam[16]
Residual blocks	30 blocks
Dilations	$[2^0, 2^1, 2^2, ..., 2^9]$ was repeated three times
Input(Step 1)	Waveform: 256 classes $\times$ 7680 samples Mel-spectrum: 80 band $\times$ 30 frames EID: 3 types $\times$ 1 samples
Input(Step 2)	Waveform: 256 classes $\times$ 7680 samples EID: 3 types $\times$ 1 samples
Output	256 classes $\times$ 1 samples

C. Evaluation for emotional expression

1) Experimental procedures: To evaluate emotional expression of the generated speech, emotional identification tests were carried out. The stimuli used in the experiments are the 120 utterances described in Section IV-B and they were presented to participants in a random order. Each utterance was used twice, so the total number of stimuli was 240. To discard influence of the linguistic channel and to focus on emotional information of non-linguistic channel, we selected participants who did not know both Italian and German. The total number of participants was 11, and they were asked to select an emotion representing each stimulus from "normal," "happy," and "angry."

2) Experimental results: Tables II shows the confusion matrixes of ORIGINAL, MLSA, WORLD, and WAVENET, respectively. In ORIGINAL, MLSA, and WORLD, "angry" was correctly identified more than 75%. However, "happy" was around 40%. Judging from the results of ORIGINAL, "happy" is considered to be similar to "normal." This might be a central reason why the discrimination rate was so low for "happy". On the other hand, WAVENET showed much better and slightly worse discrimination rates regarding "happy" and "angry," respectively. To verify the reason, the

[1] https://ktmatu.github.io/SES-sample/

TABLE II
CONFUSION MATRIXES OF EXPERIMENTAL RESULTS

Confusion matrix of ORIGINAL			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.900**	0.073	0.027
Angry	0.232	**0.750**	0.027
Happy	0.523	0.023	**0.455**

Confusion matrix of MLSA			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.909**	0.082	0.009
Angry	0.195	**0.782**	0.023
Happy	0.582	0.045	**0.368**

Confusion matrix of WORLD			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.900**	0.082	0.018
Angry	0.150	**0.827**	0.023
Happy	0.564	0.023	**0.414**

Confusion matrix of WAVENET			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.927**	0.050	0.023
Angry	0.300	**0.600**	0.100
Happy	0.200	0.077	**0.723**

TABLE III
CONFUSION MATRIX OF TRAINING DATA

	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**1.000**	0.000	0.000
Angry	0.000	**0.994**	0.006
Happy	0.000	0.011	**0.989**

same emotional identification tests were carried out using the training data of WAVENET. The results are shown in Table III. Judging from the results, WAVENET showed better performance for "happy" because of the training data. This fact indicates that WAVENET effectively obtained features of emotional expression in Step 2, and training data sets should have discrimination rates of emotional expressions as high as possible.

Figure 3 shows examples of the generated sounds by the proposed algorithm. In general, the spectrograms seem like those of speech signals, and appropriate characteristicss are observed for each emotion. In terms of the spectrogram, the "angry" version has relatively more power in the high frequency band and a weaker formant structure in the low frequency band than the "normal" version. The "happy" version has almost no formant structure in the low frequency band. In terms of F_0 contours, the "happy" version has relatively high values and shows large changes with the passage of time. As shown in these figures, we can say that the WaveNet simultaneously obtained both spectrum and source features of emotional speech. This might be one of the biggest advantages for WaveNet to generate high quality speech. In terms of the reason why WAVENET showed slightly worse discrimination rates regarding "angry," F_0 contours of "angry" are similar to those of "normal" as shown in Fig. 3

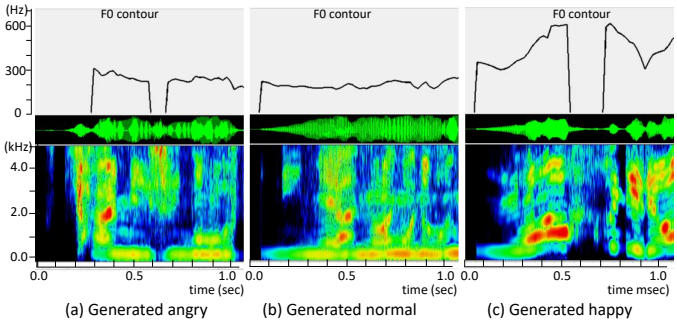

Fig. 3. Examples of the generated sound by the proposed algorithm

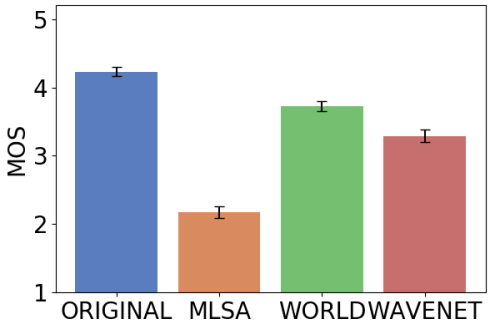

Fig. 4. Average MOS score for all emotions

D. Evaluation for speech-like aspects

1) Experimental procedures: To evaluate speech-like aspects of the generated speech, Mean Opinion Score (MOS) tests were carried out. The stimuli used in the experiments were the 120 utterances described in Section IV-B. The same participants described in Section IV-C1 were asked to judge the stimuli using a 5-point scale. (5: The speech is definitely uttered by a human, 3: The speech has an even chance of being uttered by a human or synthesized by a computer, 1: The speech is definitely synthesized by a computer). The judgments were carried out emotion-by-emotion; i.e. from speech uttered with an emotion, ORIGINAL, WORLD, MLSA, and WAVENET were randomly selected and presented to participants for judgments, and then other judgments were performed for speech uttered with other emotions.

2) Experiment results: Figure 4 shows the experiment results where MOS values are averaged for all emotions. The error bar represents a 95% confidence interval. Interestingly, ORIGINAL does not always get five points. Based on this fact, we can say that participants judged that WAVENET sounds as if it is uttered by a human. WORLD must have the best match between the spectrum envelope parameter and source parameter because WORLD is synthesis-by-analysis speech. On the other hand, WAVENET is generated using only EIDs. However, there is little difference in the MOS score between WORLD and WAVNET. This indicates that the proposed algorithm successfully trained for emotional speech by taking interactions of a spectrum envelope parameter and source parameter into account.

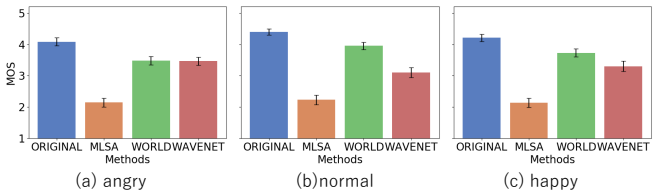

Fig. 5. MOS score for each emotion

Figure 5 shows the experimental results for each emotion. In terms of the MOS score, there is little difference between WORLD and WAVNET regarding "angry" and "happy," but relatively large differences regarding "normal." This is mainly because there are abrupt changes in both the spectrum and source regarding "angry" and "happy." WORLD could extract good parameters for "normal."

V. Conclusions

In this paper, we proposed a new algorithm to generate Speech-like Emotional Sound (SES) using WaveNet. The final goal is to generate non-verbal emotional vocalizations for human-machine interaction (HCI) by developing an algorithm to deal with emotional information of non-linguistic channels independently of linguistic channels. The algorithm consists of two steps. The first step involves learnings "speech-like" sounds using a large amount of normal speech data, and the second step in learnings emotional expressions using a relatively small amount of emotional speech data. Judging from the experiment results, we can say that the proposed algorithm makes it possible to generate sounds with emotional information without linguistic information. Moreover, the quality of the sounds is as high as if they were uttered by human beings.

As part of our future work, we have plans to apply the proposed algorithm to various type of emotional speech and reveal the relationship between the quality and size of the training data. Moreover, we would like to use the generated sounds for HCI.

References

[1] M. Schröder, "Emotional speech synthesis: A review," in *Proc. of EUROSPEECH*, 2001, pp. 561–564.
[2] M. Schröder, "Expressive speech synthesis: past, present, and possible futures," in *Affective Information Processing*, 2009, pp. 111–126.
[3] A. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proc. of EUROSPEECH*. International Speech Communication Association, 1995, pp. 581–584.
[4] H. Mizuno, H. Asano, M. Isogai, M. Hasebe, and M. Abe, "Text-to-speech synthesis technology using corpus-based approach," *NTT Technical Review*, pp. 70–75, 2004.
[5] A. Iida and N. Campbell, "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 379–392, 2003.
[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[7] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 2461–2464.
[8] T. Masuko, T. Kobayashi, and K. Miyanaga, "A style control technique for HMM-based speech synthesis," in *Proceedings of the 8th International Conference of Spoken Language Processing*, 2004.
[9] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP*, 2007, pp. 1233–1236.
[10] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech communication*, vol. 52, no. 5, pp. 394–404, 2010.
[11] J. Trouvain and M. Schröder, "How (not) to Add Laughter to Synthetic Speech," *Proc. Workshop on Affective Dialogue Systems*, pp. 229–232, 2004.
[12] M. Schröder, D. K. Heylen, and I. Poggi, "Perception of non-verbal emotional listener feedback," *Proc. of Speech Prosody*, pp. 43–46, 2006.
[13] A. van den 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," the Computing Research Repository(CoRR) abs/1609.03499, 2016.
[14] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017, accessed Nov. 2018.
[15] y_benjo and MagnesiumRibbon, "Voice-Actress Corpus," http://voice-statistics.github.io/, accessed Nov. 2018.
[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[17] G. Costantini, I. Iadarola, A. Paoloni, and M. Todisco, "EMOVO Corpus: an Italian Emotional Speech Database," https://core.ac.uk/download/pdf/53857389.pdf, accessed March. 2019.
[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *INTERSPEECH*, 2005, pp. 1517–1520.
[19] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.
[20] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.