INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Controlling the Strength of Emotions in Speech-like Emotional Sound Generated by WaveNet

Kento Matsumoto[1], *Sunao Hara*[1], *Masanobu Abe*[1]

[1]Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Japan

k_matsu@s.okayama-u.ac.jp, hara@okayama-u.ac.jp, abe-m@okayama-u.ac.jp

Abstract

This paper proposes a method to enhance the controllability of a Speech-like Emotional Sound (SES). In our previous study, we proposed an algorithm to generate SES by employing WaveNet as a sound generator and confirmed that SES can successfully convey emotional information. The proposed algorithm generates SES using only emotional IDs, which results in having no linguistic information. We call the generated sounds "speech-like" because they sound as if they are uttered by human beings although they contain no linguistic information. We could synthesize natural sounding acoustic signals that are fairly different from vocoder sounds to make the best use of WaveNet. To flexibly control the strength of emotions, this paper proposes to use a state of voiced, unvoiced, and silence (VUS) as auxiliary features. Three types of emotional speech, namely, neutral, angry, and happy, were generated and subjectively evaluated. Experimental results reveal the following: (1) VUS can control the strength of SES by changing the durations of VUS states, (2) VUS with narrow F0 distribution can express stronger emotions than that with wide F0 distribution, and (3) the smaller the unvoiced percentage is, the stronger the emotional impression is.

Index Terms: speech synthesis, emotional speech, WaveNet

1. 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 its quality. One of the reasons is the lack of emotional expressions. In a human-computer interaction (HCI), emotional expressions could be the most important factor in generating natural responses. Even though intensive studies have been conducted in the last two decades, synthetic speech has failed in expressing a range of emotions [1] [2]. For instance, waveform-based speech synthesis needs a large amount of emotional speech data, but collecting them is difficult [3] [4] [5]. HMM-based speech synthesis can generate emotional speech, but their quality is not good enough [6] [7] [8] [9] [10].

We proposed to generate Speech-like Emotional Sound (SES), as a new approach to synthesize emotional speech, using WaveNet [11] and confirmed that SES can successfully convey emotional information. We call the generated sounds "speech-like" because they include no linguistic information, but they sound as if they are uttered by human beings. An idea is to focus on synthesizing a non-linguistic emotional aspect that is represented by acoustic features in voice quality and intonation. This could be possible because we can sometimes feel emotions by hearing speech spoken in a language that we do not know at all. The advantages of the proposed approach are the following: (1) by independently dealing with the non-linguistic and linguistic aspects of emotional speech, we can reduce the amount of emotional speech data for the training, and (2) by employing WaveNet as a sound generator, we can synthesize natural sounding acoustic signals that are fairly distinct from vocoder sounds. SES might be useful for dialog systems to express simple reactions with emotions because it is reported that non-verbal emotional vocalizations are important when expressing emotions, especially in conversations. Moreover, an approach to synthesize emotional speech by Text-To-Speech (TTS) is inappropriate because TTS always requires linguistic information [12] [13].

This paper proposes to increase the controllability of SES. Considering that controlling the degree of emotional expressions for HCI is necessary, we propose to control it by auxiliary features. The proposed SES generation algorithm comprises two steps. In the first step, WaveNet is trained to generate "speech-like" sounds using a large amount of neutral speech, and in the second step, the trained model is retrained to express emotions using a small amount of emotional speech. In our previous study, in the first step, mel-spectrum parameters were used as auxiliary features to train spectrum variations as much as possible. This is mainly because we observed that similar sound was repeatedly generated without auxiliary features. It succeeded to generate SES, but it had no controllability because the proposed algorithm needed only emotional ID (EID). Other auxiliary features are crucial to increase the controllability. As a possible candidate, we propose to use a state of voiced, unvoiced, and silence (VUS).

The rest of the paper is organized as follows: Section 2 describes the basics of WaveNet, Section 3 explains the proposed method, Section 4 shows our evaluation results and provides a discussion, and Section 5 presents our conclusions and suggests avenues for future works.

2. Basics of WaveNet

2.1. WaveNet

WaveNet [14] 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:

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

Equation (1) shows that each sample point x_t is conditioned on preceding all sample points. Given that a large receptive field is

http://dx.doi.org/10.21437/Interspeech.2020-2064

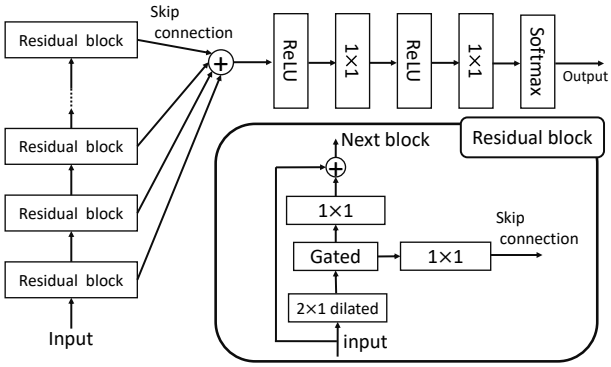

Figure 1: *WaveNet*

essential 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 Figure 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. $W_{f,k}$ and $W_{g,k}$ refer to the weight of convolution, k is the layer index, and f and g denote the filter and gate, respectively. In the output layer, WaveNet predicts a sample point that is quantized by the 8-bit μ-law algorithm as a classification of $2^8 = 256$ classes.

2.2. Conditional WaveNet

By adding auxiliary features $\mathbf{h}$, we can control the outputs of WaveNet. Equation (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 that 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 refers to the weight of convolution, and $V * \mathbf{y}$ is a 1×1 convolution.

3. SES generation algorithm using WaveNet

The proposed algorithm employs the conditional WaveNet and has two steps (Steps 1 and 2) for model training. In both steps, the WaveNet structure is the same as in our previous study [11] except the following. The conditional feature module proposed in [15] was used as an upsampling method $f(\mathbf{h})$. Figure 2 illustrates the outline of the proposed algorithm.

3.1. Training Step 1

Step 1 is a key procedure to generate "speech-like" sounds. WaveNet is trained using speech data uttered with the neutral speaking style, and the size of the speech data is relatively large. Figure 2 also depicts that Step 1 has the necessary auxiliary feature of Emotion ID (EID) and two optional auxiliary features. The one is mel-spectrum parameters that were proposed in the previous study, and the other is a state of voiced, unvoiced, and silence that are represented by one-hot vector $\mathbf{c}_n \in \mathbb{R}^3$, where n denotes the n-th frame of a waveform. They are denoted by

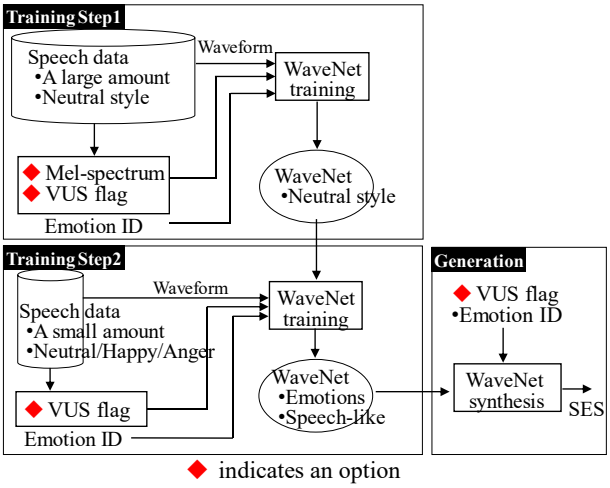

Figure 2: *Outline of the proposed method*

a VUS flag in Figure 2. EID is represented by one-hot vector, and the neutral emotion is always set to 1 in Step 1.

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

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 $\mathbf{x}$. The $\mathbf{y}_{\mathbf{condition}}$ denotes a feature matrix that is transformed by the conditional feature module so that the length of the $\mathbf{y}_{\mathbf{condition}}$ is matched to that of the input audio signal $\mathbf{x}$. The feature matrix was determined by the selection of optional auxiliary features, i.e., both mel-spectrum and VUS flags or only VUS flags. Furthermore, we assumed that the emotion is kept during an utterance. Now, U refers to the weight of convolution, and $U * \mathbf{y}$ is a 1×1 convolution.

3.2. Training Step 2

Step 2 is a key procedure to generate emotional sounds and to discard linguistic information. WaveNet is retrained after Step 1 using the speech data uttered in several emotional styles and using correspondent EIDs and VUS flags if necessary. However, mel-spectrum parameters were not used for the training to discard linguistic information.

3.3. Sound generation

WaveNet generates SES with EIDs and the first data point. The VUS flags were also used as auxiliary features in the systems that exploit them. Furthermore, WaveNet predicts the next sample point using random sampling from softmax distribution that is the output of WaveNet. Therefore, generated speeches are different each time.

4. Evaluation experiments

The subjective evaluations were performed from the viewpoints of emotional identification and overall performance to evaluate the performances for generating SES.

4.1. Model construction

Three types of WaveNets models were trained. Table 1 shows the model names and combinations of the selected auxiliary fea-

Table 1: *Conditions about auxiliary features*

	Step 1		Step 2
	Mel-spectrum	VUS flags	VUS flags
Spe-Vus	○	○	○
Vus	-	○	○
No	-	-	-

tures. Mel-spectrum parameters were used to train spectrum variations in the previous study [11]. We prepared Spe-Vus and Vus to compare the differences in performance with and without the spectrum. In Step 1, we used the JSUT corpus [16], which contains 7,696 utterances (10 hours) of speech data uttered by a native Japanese female speaker in a neutral speaking style. In Step 2, we used the Voice-Actress Corpus [17], where another Japanese actress read Japanese texts aloud using three emotions: neutral, angry, and happy. It contains 100 utterances (17 minutes) of each emotion, 300 utterances (51 minutes) in total. For training, we used 7,646 utterances and 285 utterances from the JSUT corpus and the Voice-Actress Corpus, respectively. All utterances were downsampled to 16 kHz, and the acoustic features were extracted with a window length of 64 msec and a frame shift of 5 msec. As auxiliary features, mel-spectrum parameters were calculated using short-time Fourier transform, and VUS flags were determined by the power of waveform and fundamental frequency (F0) extracted using WORLD [18]. In addition, we used the Adam optimizer [19], and the WaveNets models were trained for 800 K iterations in Step 1. In Step 2, the WaveNets models were trained for 200 K iterations, 120 K iterations, and 40 K iterations that correspond to Spe-Vus, Vus, and No, respectively. The batch size is 4, and the batch length is 16,000 samples. WaveNets consists of 30 residual blocks. The dilations were set to $[2^0, 2^1, 2^2, ..., 2^9]$, and this was repeated three times.

4.2. Evaluation by emotional identification

4.2.1. Experimental procedures

To evaluate the generated models, the emotional identification tests were carried out. As explained in Section 3.3, VUS flags are crucial when they are used as auxiliary features. Hence, in the experiments, VUS flags are generated by two methods, i.e., (Method 1) to extract from natural speech and (Method 2) to extract from SES. The stimuli were then generated using six types of models: Spe-Vus-1, Spe-Vus-2, Vus-1, Vus-2, No, and Conv. Here, Conv means the model proposed in the previous study [11], and the numbers 1 and 2 correspond to the method of VUS flag generation. In the Method 2, Conv is used as SES. We prepared 3 emotions $\times$ 5 utterances $=$ 15 utterances for each method. In the end, there are 15 utterances $\times$6 methods $=$ 90 stimuli in total, and they were presented to participants in a random order. Each utterance was used twice; thus, the total number of stimuli is 180. Twelve Japanese native speakers participated, and they were asked to select an emotion representing each stimulus from "neutral," "happy," and "angry." Generated speeches can be found on our web page [1].

4.2.2. Experimental results

Table 2 shows the confusion matrices of Spe-Vus-1, Spe-Vus-2, Vus-1, Vus-2, No, and Conv. As a total performance, Conv is the best, and all the three emotions are correctly judged at over 93%. The results indicate that Conv can generate the strongest emotional expressions, but there is no way to control the degree of emotions because SES is generated using only EIDs. Figure 3 shows the distribution of F0 and differential F0 extracted from Vus-1, Vus-2, and Conv that were used in the subjective tests. Evidently, the distribution from Conv is narrower than those from the others, and there is a little overlap between the distributions of three emotions in Conv, which might be a reason why the Conv showed the best performance in emotional identification. In addition, Figure 3 shows that the distributions of F0 are similar in "Angry" and "Neutral," which is a reason why all models, except for Conv, have low identification rates in "Angry" and "Angry" is misjudged as "Neutral." Another finding is that identification performance is related to the distribution, i.e., Vus-2 has narrower F0 distribution than Vus-1, and the identification performance rates increase by 10% and 4.2% in "Angry" and "Happy," respectively. Another reason for the identification performance differences in Vus-2 and Vus-1 might come from the state durations of voiced and unvoiced as follows. Figure 4 shows the percentage of voiced and unvoiced in the total duration. By comparing the identification performance of Vus-2 with that of Vus-1, for "Angry" and "Happy," the smaller the unvoiced percentage is, the stronger the emotional impression is. The results validate that the states of VUS can control the degree of emotional expressions.

4.3. Evaluation for overall performance

4.3.1. Experimental procedures

To evaluate the overall performance for naturalness and emotional expressions, Mean Opinion Score (MOS) tests were carried out. The stimuli used in the experiments are the 60 utterances of Spe-Vus-1, Vus-1, No, and Conv, and they were presented to the participants in a random order. Each utterance was used three times; hence, the total number of stimuli is 60 utterances $\times$3 emotions $=$ 180. The same participants described in Section 4.2 were asked to judge the stimuli using a 5-point scale (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad). The MOS test was separately performed for each emotion. Note that the naturalness in this experiment means that how close generated speeches are to the speeches uttered by a human, and this experiment didn't evaluate linguistic naturalness.

4.3.2. Experiment results

Figure 5 shows the experimental results for each emotion. The error bar represents a 95% confidence interval. In terms of Conv, "Neutral" does have worse MOS score than "Angry" and "Happy." Moreover, for "Neutral", Conv and No have almost the same scores and are the worst in all the stimuli. The reason can be easily observed in spectrograms shown in Figure 6 (a) that is obtained from Conv for "Neutral." As shown here, similar spectrogram patterns are periodically repeated between 0.5 seconds and 2.5 seconds. This sounds fairly unnatural because we have little chance to hear this type of sound in neutral speaking style. However, the sounds might be acceptable as strong emotional expressions, which can be another reason why Conv obtains a high score in Table 2. Furthermore, Fig. 6 (b) shows the spectrums obtained from Vus-1. Given that Vus-1 is synthesized using the VUS flags extracted from natural speech, the spectrogram change is not so fast and seems to be moderate. As shown here, VUS flags can control the duration of voiced and unvoiced. This implies that VUS flags can change the degree of emotional expressions to some extent. In terms of the

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

Table 2: *Confusion matrices of experimental results*

(a) Spe-Vus-1

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.967**	0.025	0.008
Angry	0.383	**0.600**	0.017
Happy	0.058	0.083	**0.858**

(b) Spe-Vus-2

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.950**	0.008	0.042
Angry	0.383	**0.617**	0.000
Happy	0.092	0.042	**0.867**

(c) Vus-1

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.992**	0.000	0.008
Angry	0.317	**0.683**	0.000
Happy	0.108	0.183	**0.708**

(d) Vus-2

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.992**	0.008	0.000
Angry	0.208	**0.783**	0.008
Happy	0.192	0.058	**0.750**

(e) No

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.925**	0.075	0.000
Angry	0.312	**0.688**	0.000
Happy	0.142	0.108	**0.750**

(f) Conv

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.942**	0.058	0.000
Angry	0.050	**0.933**	0.017
Happy	0.017	0.042	**0.942**

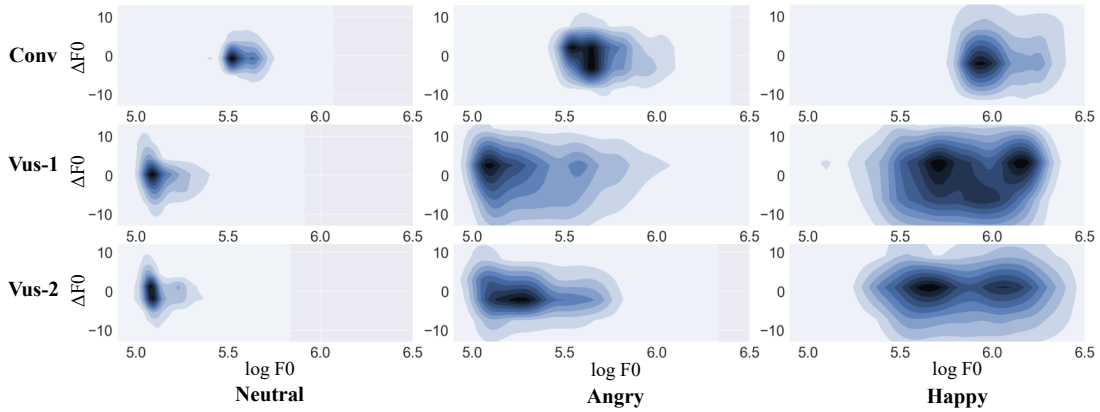

Figure 3: *The distribution of F0 and differential F0*

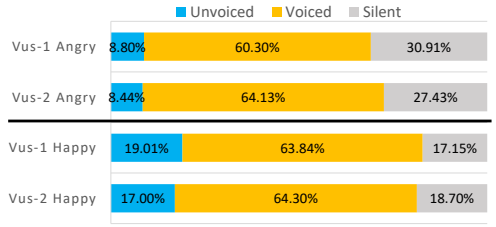

Figure 4: *The percentage of voiced and unvoiced*

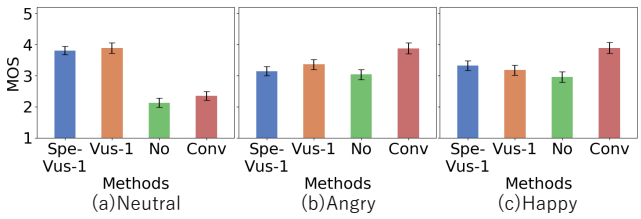

Figure 5: *MOS score for each emotion*

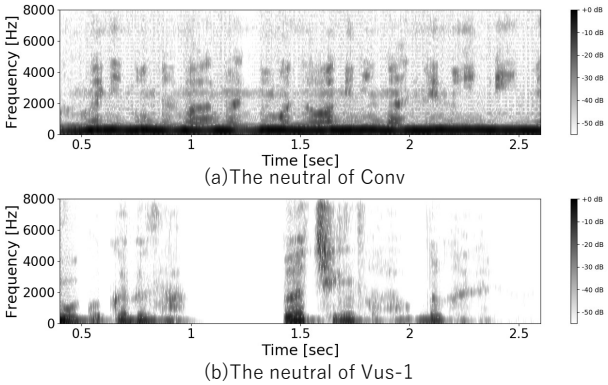

Figure 6: *Examples of Conv and Vus-1*

effects of VUS flags, for "Neutral," Spe-Vus-1 and Vus-1 have greatly higher MOS scores than Conv. This indicates another effect of VUS flags that improve the naturalness of "Neutral" by reducing the chance to generate periodical patterns.

5. Conclusions

This paper proposed to use a state of voiced, unvoiced, and silence (VUS) as auxiliary features to flexibly control the strength of emotions. Experimental results showed that: (1) VUS can control the strength of SES by changing the durations of VUS states, (2) VUS with narrow F0 distribution can express stronger emotions than that with wide F0 distribution, and (3) the smaller the unvoiced percentage is, the stronger the emotional impression is. In addition, we found VUS reduces periodic similar sounds, and improves a naturalness for neutral of SES. However, it makes an emotional expression SES restrained.

As part of our future work, we have plans to examine VUS flags that can strongly express emotions and improve the emotional expressions of SES. Moreover, we would like to use the generated sounds for HCI; e.g. conversational agents that respond using emotional expressions without linguistic information.

6. 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] K. Matsumoto, S. Hara, and M. Abe, "Speech-like Emotional Sound Generator by WaveNet," in *APSIPA Annual Summit and Conference 2019*, 2019, pp. 143–147.

[12] J. Trouvain and M. Schröder, "How (not) to Add Laughter to Synthetic Speech," *Proc. Workshop on Affective Dialogue Systems*, pp. 229–232, 2004.

[13] M. Schröder, D. K. Heylen, and I. Poggi, "Perception of non-verbal emotional listener feedback," *Proc. of Speech Prosody*, pp. 43–46, 2006.

[14] 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.

[15] X. Wang, S. Takaki, and J. Yamagishi, "Investigation of WaveNet for Text-to-Speech Synthesis," *IPSJ SIG Technical Report*, vol. 2018-SLP-120, no. 6, pp. 1–6, Feb. 2018.

[16] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[17] y_benjo and MagnesiumRibbon, "Voice-Actress Corpus," http://voice-statistics.github.io/, accessed Nov. 2018.

[18] 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.

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