EMOTION-CONTROLLABLE SPEECH SYNTHESIS USING EMOTION SOFT LABELS AND FINE-GRAINED PROSODY FACTORS

Xuan Luo, Shinnosuke Takamichi, Tomoki Koriyama, Yuki Saito, Hiroshi Saruwatari
Graduate School of Information Science and Technology, The University of Tokyo, Japan.
luo-xuan@g.ecc.u-tokyo.ac.jp, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

***Abstract*—We propose an emotion-controllable text-to-speech (TTS) model that allows both emotion-level (i.e., coarse-grained) and prosody-factor-level (i.e., fine-grained) control of speech using both emotion soft labels and prosody factors. Conventional methods control speech only by using emotion labels, emotion strength, or prosody factors (e.g., mean and standard deviation of pitch), which cannot express diverse emotions. Our model is based on a speech emotion recognizer (SER) and a prosody factor generator (PFG) model that encodes utterance-level prosody factors into emotion soft labels and decodes encoded emotion soft labels back into utterance-level prosody factors. Our model enables emotion labels and prosody factors to control synthetic speech emotion. Experiment results show that the emotion-perceptual accuracy of synthetic speech reached 66 %, and the mean opinion score for the naturalness of emotionally controlled synthetic speech was 3.9, which is comparable to a conventional method that only uses prosody factors.**

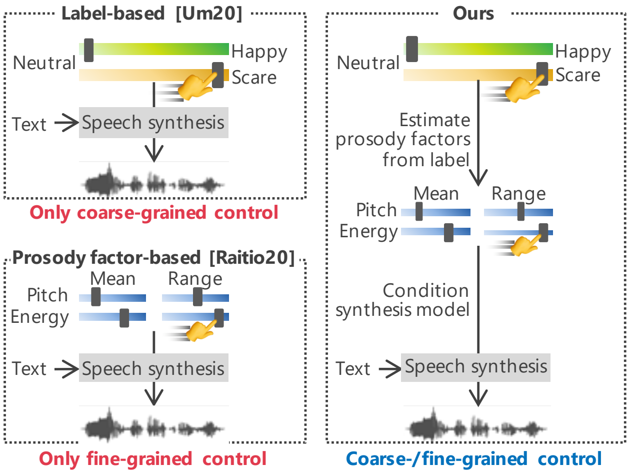
Label-based [Um20]
Neutral
Happy
Scare
Text
Speech synthesis
Only coarse-grained control
Prosody factor-based [Raitio20]
Mean
Range
Pitch
Energy
Text
Speech synthesis
Only fine-grained control
Ours
Neutral
Happy
Scare
Estimate prosody factors from label
Mean
Range
Pitch
Energy
Condition synthesis model
Text
Speech synthesis
Coarse-/fine-grained control

Fig. 1. Proposed emotion-controllable TTS that enables both coarse-grained and fine-grained emotion control.

I. Introduction

Text-to-speech (TTS) technology aims to generate human-like speech that includes both linguistic and para-linguistic information. The fast development of deep learning models has already made synthetic speech very understandable from a linguistic perspective [1]. The next challenge for TTS models is reproducing and controlling a diverse variety of para-linguistic information (e.g., emotions) in natural speech. Therefore, we aim to develop emotion-controllable TTS that can express diverse emotions.

Diverse speech emotions are mainly produced in two different variations: inter-category and intra-category. The inter-category variation intuitively produces diverse speech emotions because different speech emotions are expressed in completely different ways. Corresponding typical approaches, called coarse-grained emotion control approaches, reproduces the inter-category variation in a speech by conditioning TTS models with different emotion labels [2]–[4]. These emotion labels can be assigned in a supervised [2] or an unsupervised [4], [5] manner. Meanwhile, the intra-category variation also produces diverse speech emotions considerably. Corresponding typical approaches, called ”fine-grained” emotion control approaches, condition TTS models using emotion strength or weight [6]–[8]. These TTS models can furthermore control emotion intensity in a speech by conditioning using emotion strength under given emotion labels. However, these models cannot control how emotion intensity is represented in speech. Prosody factors (e.g., energy or pitch) that represented emotion intensity cannot be controlled by these fine-grained emotion control models. The prosody factors vary from utterance to utterance even when the emotion strength is the same. For example, anger strength 0.90 can be expressed in different prosody factors, such as a -22.0 dB energy mean or a 10 dB-Hz pitch range. That is, current fine-grained emotion controlling models that reproduce the intra-category variation by emotion strength are not ”fine” enough. Another method of representing an intra-category variation in speech directly conditions a TTS model using prosody factors [9], but it cannot control emotion.

To summarize, the conventional approaches cannot express both inter-category and intra-category variations of speech emotion at a coarse-grained (i.e., emotion-level) and a fine-grained level (i.e., prosody factors) at the same time.

We propose an emotion-controllable TTS model that en-

978-988-14768-9-0/21/$31.00 ©2021 APSIPA

ables both coarse-grained and fine-grained emotion control, as shown in Fig. 1. To achieve coarse-grained emotion control, we introduce a speech emotion recognizer (SER) that estimates the speech emotion soft labels used for coarse-grained emotion control from the utterance-level prosody factors. To achieve fine-grained emotion control, we also introduce a prosody factor generator (PFG) that estimates the utterance-level prosody factors used for fine-grained emotion control from the estimated speech emotion soft labels. The proposed TTS model based on Tacotron2 [1] is conditioned by the emotion soft labels and the estimated prosody factors. This design enables our method to provide coarse-grained control by using emotion soft labels and fine-grained control by using both emotion soft labels and prosody factors. The experiment results show that our proposed method can achieve coarse-grained control of speech emotion with 66 % accuracy and linearly fine-grained control by proposed prosody factors without deterioration of audio quality compared with the conventional method.

In short, this paper's contributions are summarized as follows:

1) We proposed a paired SER–PFG model that estimates emotion soft labels from utterance-level prosody factors and the utterance-level prosody factors from the emotion soft labels, respectively.
2) We proposed an emotional TTS model based on the paired SER–PFG model that can produce coarse-grained and fine-grained control of speech emotion.

II. Related Work

The emotion information in the emotional TTS model can be represented in various ways, such as emotion labels, prosody factors, and implicitly hidden states. The approaches to representing emotion as hidden states include a multi-head attention model [10], an unconditioned model [11], and a variational autoencoder model [12] that embed emotion in a hidden state vector trained in an unsupervised way. However, such an unsupervised learned hidden state does not correspond to human perceptual emotions, which leads to an inability to control the synthetic speech using a specific emotion. The approaches to representing emotion as emotion labels include assigning a one-hot emotion label (i.e., using a labeled dataset) [2], elaborately selecting the centroid weight of style tokens trained by a labeled emotional speech dataset as hard emotion label [3], and using emotion soft labels obtained by an emotion interpolation approach [6]. However, none of them can control speech emotion at a fine-grained level under given emotion labels.

Research such as [9] could properly control speech at a fine-grained level by adjusting five prosody factors obtained from the Long short-term memory (LSTM) [13] based prosody encoder, and FastSpeech2 [14] also controls speech by predicted pitch and energy. However, neither can control the coarse-grained emotion of synthetic speech.

To control speech emotion at both coarse-grained and fine-grained levels, research [7] conditioned speech using an emotion label and a continuous emotion-strength scalar value. Another research [8] conditioned speech using phoneme-level emotion strength instead of sentence-level emotion strength for better controlling ability. However, fine-grained emotion control is defined quite differently in those papers from our definition. To be clear, the fine-grained emotion control mentioned in those papers denotes to control speech by sentence-level or phoneme-level emotion strength, however, it means to control speech by prosody factors in our paper. Considering emotion strength is formed by prosody factors, we believe our ”fine-grained” controls can control speech more finely.

We propose a model that enables us to control the speech emotion at the coarse-grained level using emotion soft labels and at the fine-grained level using prosody factors. For coarse-grained control, we utilize an SER model to predict emotion soft labels using utterance-level prosody factors and textual features. Unlike the latest SER models fed with phoneme-level prosody factors [15], we use utterance-level prosody factors because they can predict utterance-specific emotion soft labels more easily. For fine-grained control, we use a similar (deep neural network) DNN architecture to estimate utterance-level prosody factors from emotion soft labels.

III. Proposed Method

We propose an emotion-controllable TTS model based on SER and PFG models.

A. SER and PFG Models

The SER and PFG models are used to achieve coarse-grained and fine-grained emotion speech control, respectively. In training, the SER model estimates emotion soft labels for coarse-grained controlling from textual and prosody factors of input text and speech, and the PFG model estimates the prosody factors for fine-grained controlling from the emotion soft labels.

To extract prosody factors, we use the means and standard deviations of pitch, energy, and harmonics of speech as the prosody factors, which are originally proposed in [16] and believed to be strongly related to speech emotion. In addition to these six features, we also introduce ranges of energy and pitch for better prosody controlling. In summary, we extracted eight prosody factors. To achieve better performance of the SER model, we use term frequency-inverse document frequency (TF-IDF) [17] as the textual features because it improves the performance of multi-modal speech emotion classification [16].

1) Speech Emotion Recognizer (SER): We construct a DNN-based SER model that predicts an emotion posterior probability (i.e., emotion soft labels) from given emotional features that include textual and prosodic information. The predicted emotion soft labels can be used as features for achieving emotion-level control of synthetic speech in the emotional TTS model.

2) prosody factor Generator (PFG): We also construct a DNN-based PFG model that estimates prosody factors from the soft labels, obtained from the SER model. Compared with the conventional approaches [9] that generate prosody factors

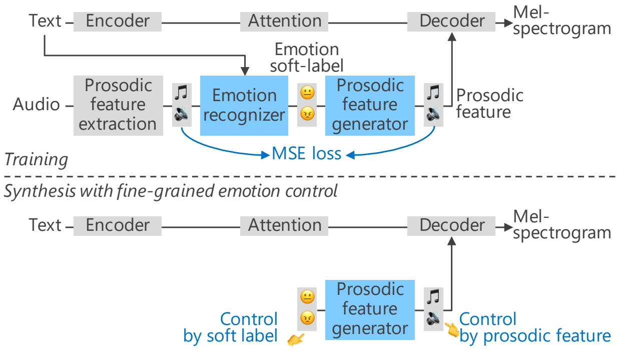

Fig. 2. The overall architecture of the proposed emotional TTS based on SER and PFG.

from the text input, our PFG model generates utterance-level prosody factors from emotion soft labels, instead of commonly used hard labels.

First, we argue that the utterance-level prosody factors in emotional TTS are mainly influenced by emotion rather than text. The stronger the emotion is, the more prosody factors can be influenced by emotion. Therefore, we generate utterance-level prosody factors from emotion.

Second, we assume that the soft labels representation of emotion can introduce more emotion diversity into synthetic speech compared with the hard-label representation.

3) Objective Function of Joint Pretraining: The SER and PFG models are jointly pre-trained using a corpus consisting of text, emotional speech, and corresponding emotion labels. The objective function to be minimized in training is

$$L_{\mathrm{EMO}} = L_{\mathrm{SER}} + L_{\mathrm{PFG}}, \tag{1}$$

where the first term L_{SER} is a cross-entropy loss between the estimated and reference emotion labels for training the SER model. The second term L_{PFG} is the mean squared error (MSE) between the estimated and reference prosody factors for training the PFG model.

B. Emotion-controllable TTS Model

Our emotion-controllable TTS model embeds the SER and PFG models into a Tacotron2 network as prosody factor controllers, as shown in Fig. 2.

1) Model Structure: The backbone TTS model is inspired by Tacotron2 [1], which consists of the encoder, a decoder, a prenet, and a postnet model. The proposed SER and PFG models are embedded as decoder input in the Tacotron2 model, shown in Fig. 2. In detail, the prosody factors, obtained from the PFG model, are concatenated with the output of the Tacotron2 attention and then fed to the Tacotron2 decoder.

2) Training: The emotion-controllable TTS model was trained using an emotional speech corpus without emotion labels. The top of Fig. 2 shows this training. Because typical corpora do not have an emotion label, we follow a previous work [5] and introduce an unsupervised way using the pre-trained SER model. We fed the textual features and prosody

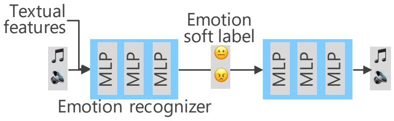

Fig. 3. The architecture of SER and PFG. “MLP” in this figure denotes multi-layer perceptron.

factors of the training data into the pre-trained SER model and obtain emotion soft labels. The prosody factors are estimated from the emotion soft label and used to condition the TTS model. The objective function L_{TTS} to be minimized for training is

$$L_{\mathrm{TTS}} = L_{\mathrm{Tacotron2}} + L_{\mathrm{PFG}}, \tag{2}$$

where $L_{\mathrm{Tacotron2}}$ is the objective function described in the Tacotron2's paper [1]. The SER model was frozen during the TTS model training because we used an unlabeled emotional speech dataset,

3) Inference: In the synthesis process, we had two options for controlling the emotion of synthetic speech: controlling by the emotion soft labels or prosody factors. The bottom of Fig. 2 shows this inference. For the former option, the emotion soft labels are manually assigned, and then prosody factors are predicted by the PFG model and fed into the TTS model. For the latter option, the estimated prosody factors are fine-adjusted manually by assigned prosody factor biases. The fine-adjusted features are then fed into the Tacotron2 decoder to estimate mel-spectrogram, which are used to generate waveforms by applying the Parallel WaveGAN model [18].

IV. Experiment Evaluation

We conducted two experiments: 1) coarse-grained emotion control by emotion labels, and 2) fine-grained emotion control by prosody factors under given emotion labels.

A. Experimental Setup

1) Data: We used the IEMOCAP corpus [19] to pre-train the SER and PFG models and the Blizzard2013 corpus [20] to train the TTS model. The IEMOCAP corpus has 12 hours of transcripts and speech recorded from emotional dialogues acted or improvised by five men and five women. We randomly split the corpus into 80 % for training and 20 % for testing the SER and PFG models. The Blizzard2013 corpus contains unlabelled emotional speech uttered by a single English speaker. Because most of the speech in the Blizzard2013 corpus has a narrative (e.g., close to neutral) style, we only filtered out only emotional speech part for training and testing using the following approach. First, we selected character-speaking sentences surrounded by single or double quotation marks. Next, we filtered out weak-emotional speeches with a score above 0.8 estimated by the SER model in each category. Finally, we had three human annotators listen to 100 randomized speeches in each emotional category of filtered data and removed perceptually non-emotional categories. As a result, we obtained 28 hours of neutral and angry speech and split it into 80 % for training and 20 % for testing the TTS model.

2) Model Parameter and Features: The SER model consisted of 3×512 multi-layer perceptrons and the PFG model consisted of 3×512 multi-layer perceptrons, as shown in Fig. 3. The TTS model was based on Tacotron2 [1], which consists of the encoder networks that include 3-layer 1-dimensional convolutions with 512 filters and a 5×1 window size, decoder networks that include a 2-layer LSTM with 1,024 hidden states, a prenet that includes 2-layer fully connected networks with 256 hidden units, and a postnet that includes 5-layer 1-dimensional convolutional networks with 512 filters. The SER model predicted 2-dimensional emotion posterior probability from a joint vector of 2030-dimensional textual features and 8-dimensional prosody factors.

As in the TTS model, Mel-spectrograms were computed through a short-time Fourier transform (STFT) using a 46 ms frame size, an 11.5 ms frame hop, and a Hann window function. And the mel scale was transformed using an 80-channel mel filterbank spanning from 80 Hz to 7,600 Hz.

The Parallel WaveGAN model is pre-trained on the LJSpeech dataset [21] and we utilized it as a neural vocoder. The pre-trained model is accessible online[1].

Textual features were extracted by the TF-IDF [17] approach for each word in the Blizzard2013 corpus, and prosody factors were extracted as the mean and standard deviation (std.) of pitch, energy, and harmonics [16] and also the range of energy and pitch based on the utterance-level. The prosody factors were normalized to $[0, 1]$, and Table I lists the range of prosody factors, and their min and max values corresponding to 0 and 1 of normalized features, respectively. The PFG model predicted 8-dimensional prosody factors from 2-dimensional emotion posterior probability.

TABLE I
RANGE AND UNIT OF PROSODIC FEATURES IN FILTERED BLIZZARD2013 DATASET

	Unit	Min	Max
Energy mean	dB	-28.5	-22.0
Energy std.	dB	-6.9	13.1
Energy range	dB	37.6	54.0
Harmonic mean	dB	-0.01	0
Harmonic std.	dB	52.1	95.3
Pitch mean	dB-Hz	44.4	47.0
Pitch std.	dB-Hz	1.0	2.2
Pitch range	dB-Hz	2.4	10.1

The SER and PFG models were firstly pre-trained with the IEMOCAP data before being used as the initial parameters in the following TTS model trained on the Blizzard2013 corpus. The parameters of the SER model and the PFG model were frozen and fine-tuned, respectively during the training. We used the Adam optimizer [22] and the 0.001 learning rate started decaying exponentially to 0.00001 after 50,000 iterations.

B. Performance of SER and PFG Models

We evaluated the performance of the SER and PFG models pre-trained with the IEMOCAP dataset before evaluating our TTS model. We evaluated the accuracy of the SER model using both IEMOCAP and Blizzard2013 (Angry/Neutral) test data, as shown in Table II. The results indicate that the pre-trained SER model showed a fair performance using the Blizzard2013 dataset(accuracy $= 0.71$), although there was inevitably a problem with domain adaption.

TABLE II
SER MODEL PERFORMANCE

Evaluation data	Accuracy
IEMOCAP test data (2 emos)	0.90
Blizzard2013 test data (2 emos)	0.71

We evaluated the PFG model using Blizzard2013 (Angry/Neutral) test data based on the MSE between origin and estimated prosody factors from the pre-trained PFG model. The results in Table III indicate that the PFG model can accurately estimate prosody factors with a low MSE.

TABLE III
PSD MODEL PERFORMANCE BY MSE (MEAN SQUARED ERROR)

Prosody factor	MSE
Energy mean	0.03
Energy std.	0.008
Energy range	0.02
harmonic mean	0.03
harmonic std.	0.006
Pitch mean	0.01
Pitch std.	0.02
Pitch range	0.01

C. Control by Emotion soft labels

We conducted a listening test for emotion distinctness to evaluate coarse-grained emotion control using emotion labels. We generated 130 angry and 130 neutral synthetic speech from randomly selected 130 test sentences from the Blizzard2013 corpus. Each of the 50 listeners evaluated 20 angry–neutral paired synthetic speech and selected an angry speech for each. The test was done in our evaluation system on the Amazon Mechanical Turk [23].

The results show that the accuracy of perceptual emotion reached 66 %. Although our results are inferior to the more than 80 % accuracy of the conventional method [3], the synthetic speech emotion achieved using our method is still distinguishable using only weak-emotional and unlabeled speech data.

D. Fine-grained Control by Emotion Label and prosody factors

We conducted objective and subjective evaluations to evaluate fine-grained emotion control by prosody factors. The objective evaluation calculated the correlation coefficients between the controlled and observed prosody factors. The subjective evaluation investigated whether the fine-grained control degrades the speech naturalness or not.

[1] The pre-trained Parallel WaveGAN model:model link

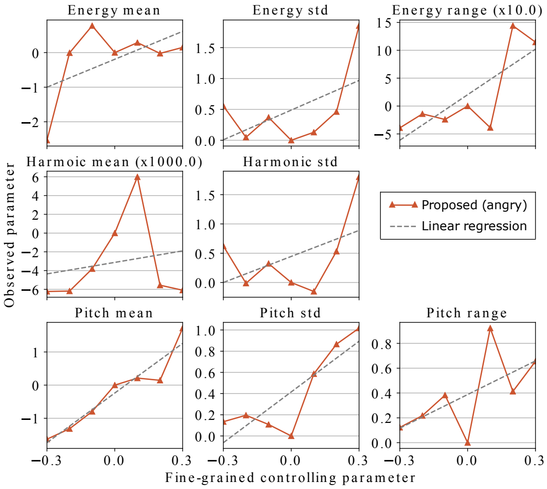

Fig. 4. Relation between controlling and observed prosody features (angry). The x-axis is prosody factor bias ranged $[-0.3, 0.3]$ by 0.1 step size and the y-axis is the observed value of each of the eight prosody factors. The dashed lines are linear approximations of the observed prosody features.

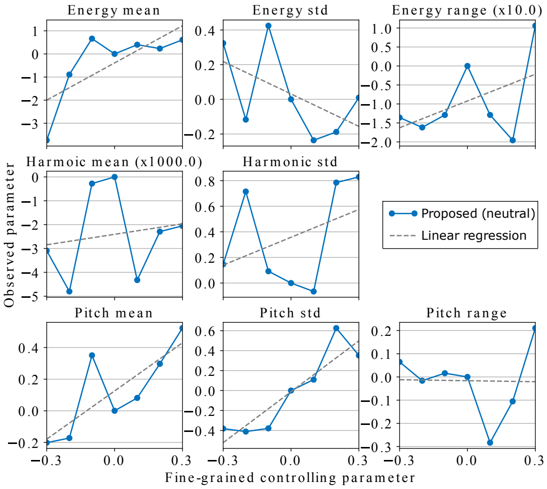

Fig. 5. Relation between controlling and measured prosody features (neutral). The x-axis is prosody factor bias ranged $[-0.3, 0.3]$ by 0.1 step size and the y-axis is the observed value of each of the eight prosody factors. The dashed lines are linear approximations of the observed prosody features.

1) Objective Evaluation: In the objective evaluation, we generated 1,120 audio files multiplied by ten randomly selected test sentences, two emotions, the eight prosody factors mentioned in Section 3.1, and seven prosody factor biases ranged $[-0.3, 0.3]$ by 0.1 step size. Under the given neutral or angry labels, we biased 8-dimensional prosody factors that were estimated from the emotion label for fine-grained controlling. We extracted the prosody factors of synthetic speech and evaluated the relation between the controlling bias and observed prosody factors.

TABLE IV
PEARSON CORRELATION COEFFICIENT FOR CONTOLLED AND OBSERVED PROSODY FACTORS. UNDERLINED PROSODY FACTORS SHOW MEDIUM OR STRONG CORRELATION (PEARSON CORRELATION COEFFICIENT > 0.3 AND p-VALUE < 0.05)

Prosody	**Pearson**	**p-value**
Energy mean	0.56	3.9e-07
Energy std.	0.27	0.56
Energy range	0.29	0.01
Harmonic mean	-0.02	0.81
Harmonic std.	0.35	0.002
Pitch mean	0.58	9.2e-08
Pitch std.	0.41	0.007
Pitch range	0.27	0.5

Figures 4 and 5show the relation, where the labels 0.0 and ±0.3 of the horizontal axis indicate no modifications or max biases on the negative and positive sides, respectively. Table IV lists their Pearson correlation coefficient (PCC) for quantitative evaluation. the results show that 1) pitch-related features (i.e., pitch mean, std., and range), energy mean, and harmonics std. show a medium or strong linear relationship between the controlled bias and the observed value (PCC > 0.3 and p-value $<$ 0.05), and 2) energy std. (0.27), energy range (0.29), and pitch range (0.27) show a linear relationship at a nearly medium level between the controlled bias and the observed value, and 3) the harmonics std. is not controlled at all. Therefore, we can say that our system can accurately control most of the proposed prosody factors in the synthesized emotional speech. Compared with the conventional research [9] that shows a better linear relation with only control prosody factors, we argue that our model only partly achieved linear controllability for gaining emotion controlling ability. We will furthermore investigate how to improve the performance in future work[2]

2) Subjective Evaluation: To evaluate the quality of synthetic speech by comparing our method with a conventional method, we synthesized 51 types of speech audio using our method with two emotion labels (neutral and angry) and the conventional prosody factor controlling method [9] for each of 200 listeners from eight prosody factors, three biases (-0.3, 0, 0.3), and ten randomly selected test sentences from the blizzard2013 test dataset. We carried out mean opinion score (MOS) tests on the naturalness of emotion-controlled synthetic speech. Figure 6 shows the encouraging result that despite the controllability on both the emotional-level and the prosodic-feature level, our model shows equal performance (MOS $= 3.9$) synthetic speech quality equal with the conventional method (which can only control speech in prosody factors).

V. Conclusion

We proposed a method of achieving coarse-grained and fine-grained control of an emotional text-to-speech (TTS) model by using emotion soft labels and prosody factors. Our

[2] sample audio:sample audio link

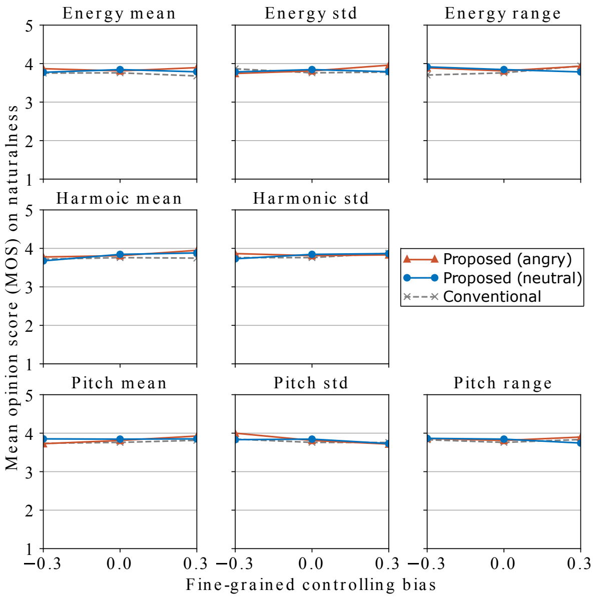

Fig. 6. MOS measured over eight controlling prosody factors.

method is based on an SER, which estimates the emotion posterior possibility from emotional features, and a PFG, which estimates prosody factors from the emotion posterior possibility. The estimated prosody factors can be used for the fine-grained control by assigned biases. Our experiment showed that our model performed as well as a conventional approach that cannot control the emotion of the synthetic speech in speech quality tests. However, our method was slightly inferior regarding emotion-perceptual accuracy and the objective measures of the controlled and observed prosody factors. The result is still encouraging considering we only used a weak emotional training dataset.

Acknowledgement: This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] Y. Lee, S.-Y. Lee, and A. Rabiee, "Emotional end-to-end neural speech synthesizer," in *Neural Information Processing Systems(NIPS) 2017*. Neural Information Processing Systems Foundation, 2017.

[3] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.

[4] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018.

[5] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *arXiv preprint arXiv:2010.13350*, 2020.

[6] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[7] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 192–199.

[8] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 423–430.

[9] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[10] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[11] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[12] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[15] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[16] G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," *arXiv preprint arXiv:1904.06022*, 2019.

[17] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Blizzard challenge workshop*, 2014.

[21] K. Ito., "The lj speech dataset." [Online]. Available: https://keithito.com/LJ-Speech-Dataset/,2017.

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

[23] K. Crowston, "Amazon Mechanical Turk: A research tool for organizations and information systems scholars," in *Shaping the future of ict research. methods and approaches*. Springer, 2012, pp. 210–221.