Controllable Text-to-speech Synthesis Using Prosodic Features and Emotion Soft-label *

☆ Xuan Luo, Shinnosuke Takamichi, Tomoki Koriyama,
Yuki Saito, Hiroshi Saruwatari (The University of Tokyo)

1 Introduction

Text-to-speech (TTS) technology aims to generate human-like speech including both linguistic and para-linguistic information. Due to the fast development of deep learning models, synthetic speech is already well understandable from a linguistic perspective [1]. The next challenge is how to reproduce and control a diverse variety of para-linguistic information, such as emotions, in natural speech. In this paper, we aim to develop emotion-controllable TTS that can express diverse emotions. The conventional approaches to controlling emotion in synthetic speech can be separated into 2 categories: coarse-grained and Fine-grained emotion control. The "coarse-grained" emotion-control models condition TTS models with emotion labels [2, 3, 4]. And the "fine-grained" emotion-control models condition TTS models with emotion strength or weight [5, 6, 7]. However, these "fine-grained" emotion control models cannot control how emotion intensity is represented in speech. The prosodic features, such as energy or pitch, that represented emotion intensity cannot be controlled by these "fine-grained" emotion control models. The prosodic features vary from utterance to utterance even when the emotion strength is the same. In another word, current "fine-grained" emotion controlling models are not "fine" enough. Another method conditions a TTS model on prosodic features [8], but it cannot control emotion. To summarize, the conventional approaches can not express both inter-category and intro-category variation of speech emotion with at coarse-grained (i.e., emotion-level) and fine-grained level (i.e., prosodic features) at the same time.

In this paper, we propose an emotion-controllable TTS model that enables both coarse-grained and fine-grained emotion control. To achieve coarse-grained emotion control, we introduce a speech emotion recognizer (SER) that estimates the speech emotion soft-label, which used for coarse-grained emotion control in inference stage, from the utterance-level prosodic features. To achieve fine-grained emotion control, we also introduce a prosodic feature generator (PFG) that estimates the utterance-level prosodic features, which used for fine-grained emotion control in inference stage, from the estimated speech emotion soft-label.

2 Related Work

The emotion information in the emotional TTS model can be represented in various ways, such as emotion labels, prosodic features and implicitly hidden states. The approaches to representing emotion as hidden states include variational autoencoder, etc [9, 10, 11]. Either of them embedded emotion into a hidden state vector trained in an unsupervised way. However, such models are incapability to controlling the synthetic speech by a specific emotion.

The approaches to representing emotion as emotion labels include assigning a one-hot emotion label manually to each speech (i.e., using labeled dataset) [2], elaborately selecting centroid weight of style tokens trained by labeled emotional speech dataset as hard emotion label [3], and using emotion soft-label obtained by an emotion interpolation approach [5]. These researches can control synthetic speech at a coarse-grained level; however, neither can control speech emotion at a fine-grained level under given emotion labels.

The research such as [8] can properly control speech in fine-grained level by adjusting 5 prosodic features obtained from recurrent neural network (RNN) based prosody encoder [12], and FastSpeech2 [13] 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 [6, 7] conditions speech on discrete emotion category and continuous emotion-strength scalar value or phoneme-level emotion strength. However, the definitions of "fine-grained" emotion control in these papers are sentence-level or phoneme-level emotion strength which is not "fine" enough compared with prosodic features controlling.

3 Proposed Method

3.1 SER and PFG Models

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

To extract prosodic features, we use the means

---

*韻律特徴と感情ソフトラベルを利用した制御可能な音声合成モデル, 羅 旋, 高道 慎之介, 郡山 知樹, 齋藤 佑樹, 猿渡 洋 (東大院・情報理工)

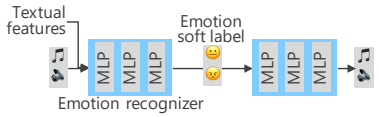

Fig. 1 Architecture of SER and PFG. "MLP" in this figure denotes multi-layer perceptron.

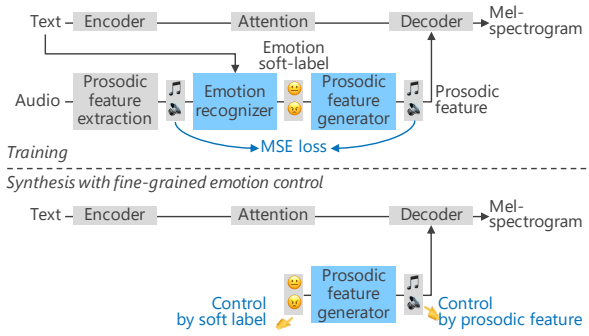

Fig. 2 Overall architecture of proposed TTS.

and standard deviations of pitch, energy, and harmonics of speech as the prosodic features, which are originally proposed in [14] and believed strongly related to speech emotion. In addition to these 6 features, we also introduce 2 new features, the ranges of energy and pitch for better prosody controlling. In summary, 8 prosodic features are extracted. To achieve better performance of the SER model, we use Term Frequency-Inverse Document Frequency (TF-IDF) [15] as the textual features since it improves the performance of multi-modal speech emotion classification [14].

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

Prosodic Feature Generator (PFG) We construct a DNN-based PFG model that estimates prosodic features from the soft-label, obtained from the SER model. Compared with the conventional approaches [8] that generate prosodic features from the text input, our PFG model is different, since 1) it generates utterance-level prosodic features from emotion and 2) the emotion here is represented by a soft-label, instead of a commonly used hard-label.

The SER and PFG models, as shown in Fig. 1 are jointly pre-trained using a corpus consisting of text, emotional speech, and corresponding emotion labels.

3.2 Emotion-controllable TTS Model

3.2.1 Model Structure

The backbone TTS model is inspired by Tacotron2 [1] consisting of the encoder, decoder, attention, prenet, and postnet. The proposed SER and PFG models are embedded as input of decoder in the Tacotron2 model, shown in Fig. 2. In detail, the prosodic features, obtained from the PFG model, are concatenated with the output of the Tacotron2 attention and then fed to the Tacotron2 decoder.

3.2.2 Training

The emotion-controllable TTS model is trained on emotional speech corpus without emotion labels. The top of Fig. 2 shows this training. Since typical corpora do not have an emotion label, we follow the previous work [16] and introduce an unsupervised way using the pre-trained SER model. We feed the textual features and prosodic features of the training data into the pre-trained SER model and obtain an emotion soft-label. From the emotion soft-label, the prosodic features are back-estimated and used for conditioning the TTS model. The objective function L_{TTS} to be minimized for training is $L_{\mathrm{TTS}} = L_{\mathrm{Tacotron2}} + L_{\mathrm{PFG}}$, where $L_{\mathrm{Tacotron2}}$ is the objective function described in the Tacotron2's paper [1]. Since we use unlabeled emotional speech dataset, the SER model is frozen during the TTS model training.

3.2.3 Inference

In the synthesis process, we have two options for controlling the emotion of synthetic speech: the emotion soft-label and prosodic features. The bottom of Fig. 2 shows this inference. For the former option, the emotion soft-label is manually assigned and prosodic features are then predicted by the PFG model and fed into the TTS model. For the latter option, the back-estimated prosodic features can be fine-adjusted manually by assigned prosodic feature biases. The fine-adjusted features are then fed into the Tacotron2 decoder to estimate mel spectrograms, which are used for generating waveform by applying the Parallel WaveGAN model [17].

4 Experiment Evaluation

4.1 Experimental Setup

4.1.1 Data

We used the IEMOCAP corpus [18] for pre-training SER and PFG models and the Blizzard2013 corpus [19] for training TTS model. The IEMOCAP corpus has 12 hours of transcript and speech, recording from emotional dialogues of five males and five females in both acting and improvising way. We randomly split it into 80 % and 20 % for training and testing the SER and PFG models. The Blizzard2013 corpus contains emotion-unlabelled emotional speech uttered by a single English speaker. We filtered out only emotional speech part for training and testing by the following approach. First, we selected character-speaking sentences surrounded by a single or double quotation mark. Then, we filtered out weak-emotional speech which is estimated with more than 0.8 score by the SER model in each category. Finally, we required 3 human annotators to randomly listen to 100 speech 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 followed by splitting into 80 % and 20 % for training and testing the TTS model.

4.1.2 Model Parameter and Features

As in the TTS model, mel-spectrograms were computed through a short-time Fourier transform (STFT) using a 46 ms frame size, 11.5 ms frame hop, and a Hann window function. The mel scale was transformed using an 80 channel mel-filterbank spanning from 80 Hz to 7600 Hz. As in neural vocoder, we applied the Parallel WaveGAN model which is pre-trained on LJSpeech dataset [20].

Textual features were extracted by the TF-IDF [15] approach for each word in the Blizzard2013 corpus, and prosodic features were extracted as the mean and standard deviation (std.) of pitch, energy, and harmonics [14] and also the range of energy and pitch in utterance-level. The minimum and maximum values of the prosodic were normalized to $[0, 1]$, respectively. The PFG model predicted 8-dimensional prosodic features from 2-dimensional emotion posterior probability.

The SER and PFG models were firstly pre-trained on the IEMOCAP data and then were used as initial parameters in the following TTS model training on the Blizzard2013 corpus. The parameters of the SER model and PFG model were frozen and fine-tuned, respectively during the training.

4.2 Control by Emotion Soft-label

To evaluate coarse-grained emotion control by emotion labels, we conducted a listening test for emotion distinctness. We generated 130 angry and 130 neutral speech from randomly selected 130 test sentences. Each of 50 listeners evaluated 20 angry–neutral paired speech and select an angry one for each. The test was done in our evaluation system on the Amazon Mechanical Turk [21].

The result shows that the accuracy of perceptual emotion reached 66 %. Although our result is inferior compared with the conventional research [3] showing higher than 80 % accuracy, the emotion of synthetic speech by our method is still distinguishable by only using weak-emotional and unlabeled speech data.

4.3 Objective Evaluation

In the objective evaluation, we generated 1120 audio totally multiplied by 10 randomly selected test sentences, 2 emotions, 8 prosodic features mentioned in Section 3.1, and 7 prosodic feature biases ranged $[-0.3, 0.3]$ by 0.1 step size. Under the given neutral or angry label, we biased 8-dimensional prosodic features which were back-estimated from the emotion label for fine-grained controlling. We extracted the prosodic features of synthetic speech and evaluated the relation between the controlling bias and observed prosodic features.

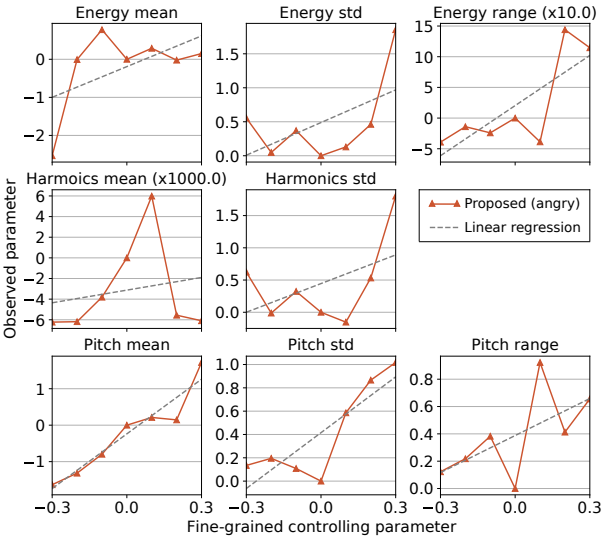

Fig. 3 Controlling vs. observed prosodic features (angry). The dashed lines are linear approximation.

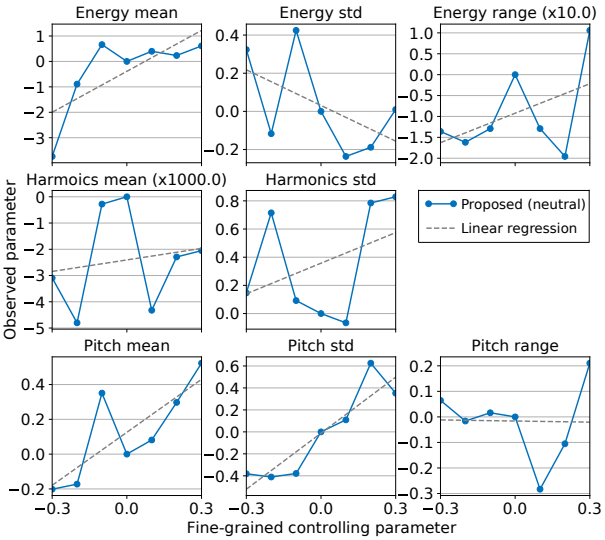

Fig. 4 Controlling vs. observed prosodic features (neutral). The dashed lines are linear approximation.

Figures 3 and 4 show the relation, where the labels 0.0 and ± 0.3 of the horizontal axis indicate no modification and max biases in negative and positive sides, respectively. Table 1 lists their Pearson correlation coefficient (PCC) for quantitative evaluation. From the results, it is observed that 1) pitch-related features (i.e., pitch mean, std., and range), energy mean and harmonics std. show medium or strong relation between controlling bias and observed value (PCC > 0.3 and p-value < 0.05), 2) according to the PCC values of energy std. (0.27), Energy range (0.29) and pitch range (0.27), these prosodic features are nearly medium relation, and 3) harmonics std. is not controlled at all. Therefore, we can say that our system can accurately control most of proposed prosodic features in the synthesized emotional speech. Compared with the conventional research [8] which shows better linear relation by only control prosodic features, we argue that our model satisfied part of linear controllability in exchange

Table 1 PCC of controlling and observed prosodic features. Underlined prosodic features show medium or strong correlation (PCC $>$ 0.3 and p-value $<$ 0.05)

Prosody	PCC	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

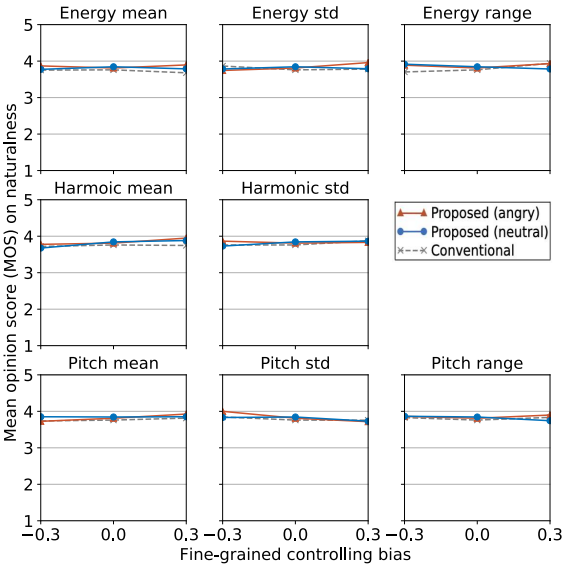

Fig. 5 MOS measured over 8 controlling features.

for emotion controlling ability. We will furthermore investigate on the approach to improve the performance in future work[1]

4.4 Subjective Evaluation

To evaluate the quality of synthetic speech by comparing with conventional approach, we synthesized 51 types of speech audio by our method with two emotion labels (neutral and angry) and the conventional prosodic feature controlling approach [8] for each of 200 listeners from 8 prosodic features, 3 biases (−0.3, 0, 0.3) and randomly selected 10 test sentences. We carried out mean opinion score (MOS) tests on naturalness of emotion-controlled synthetic speech. Figure 5 shows the encouraging result that despite the controllability in both emotional-level and prosodic-feature level, our model shows equal performance (MOS = 3.9) synthetic speech quality with the conventional research that can only control speech in prosodic-feature.

5 Conclusion

In this paper, we proposed a method that can achieve coarse-grained and fine-grained control of emotional text-to-speech (TTS) model by using emotion soft-label and prosodic features. Our model shows a slight inferior when compared with conventional approach, considering weak-emotional training dataset and doubled controlling ability, the result is still encouraging.

[1]sample audio:sample audio link

References

[1] J. Shen 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 et al., "Emotional end-to-end neural speech synthesizer," in *Neural Information Processing Systems(NIPS) 2017*. Neural Information Processing Systems Foundation, 2017.

[3] O. Kwon et al., "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 et al., "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018.

[5] S.-Y. Um et al., "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.

[6] X. Zhu et al., "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.

[7] Y. Lei et al., "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.

[8] T. Raitio et al., "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[9] Y. Wang et al., "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

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

[13] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

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

[15] 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.

[16] X. Cai et al., "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *arXiv preprint arXiv:2010.13350*, 2020.

[17] R. Yamamoto et al., "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.

[18] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

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

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

[21] 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.