Emotion-controllable Speech Synthesis using Emotion Soft Label and Word-level Prominence*

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

1 Introduction

Text-to-speech (TTS) models aim to synthesize human-like speech including linguistic and paralinguistic information. Current TTS models [1, 2] can synthesize understandable speech from a linguistic perspective. On the other hand, synthesizing human-like speech with diverse paralinguistic information, such as emotion and prominence, is still not an easy task. Conventional emotion-controllable TTS studies enable emotion controllability by conditioning on explicit emotion labels [3, 4] or implicit emotion embeddings [5, 6]. To enable more diverse controllability, subsequent studies conditioned on emotion strength which is usually obtained by an interpolation method [7] or a ranking function at utterance- or phoneme-level [3, 8].

However, emotion strength is a limited concept to only emotional speech (not for normal speech), and it refers to the intensity of only emotion (not for others, like intention), which limits its broader use. In addition, predicting emotion strength of speech requires specific emotional datasets [3], which also limits its application. On the other side, word-level prominence is a more general concept for emotional and normal speech. It refers to the perceptual quantity of standing out from other words [9], which can be calculated by speech processing techniques [10] without training on specific emotional datasets. Previous controllable TTS model [11] conditions their TTS model on word/phoneme-level prominence (i.e., emphasis) to enable diverse synthesized speech. However, such models cannot control emotion.

In this paper, we propose a two-stage emotion-controllable TTS model that we can condition on emotion soft labels and fine-condition on word-level prominence, which conventional models cannot. Our proposed model extends the Tacotron2 model with a speech emotion recognizer (SER) and a prominence predictor (PP) to enable this dual controllability. In the first stage, we condition on emotion soft labels predicted by the SER. In the second stage, we fine-condition on the prominence predicted by the PP model. The experiments achieved 1) 51% emotion-distinguishable accuracy, and 2) 0.95 linear controllability on prominence.

2 Proposed method

2.1 SER model

The SER model estimates emotion soft labels which are used for the first stage of control. It takes multi-modal features as input because of better performance than single features [12]. The multi-modal features consist of prosodic factors, prominence, and textual features.

Utterance-level prosodic factors extraction We extract pitch and energy contours of speech at the frame level and calculate their means, standard deviations (SD), and range as utterance-level prosodic factors (6-dimension) because they are expected to relate to speech emotion [12]. The pitch contour is predicted using the pYIN algorithm [13], and the energy contour is calculated by the root-mean-square value of the magnitude of each frame.

Word-level prominence extraction We extract word-level prominence by using the lines of maximum amplitude (LoMA) in the continuous wavelet transform (CWT) of a sum of signal contours of pitch, energy, and duration with weights [10]. The CWT is expected to approximate human processing of a complex signal relevant to prominence by resembling the perceptual hierarchical structures (phoneme, syllable, word) related to prosody. This ability is more difficult to achieve with traditional spectrograms. The LoMA [14] are lines that can identify and quantify word-level prominence by connecting nearby peaks in the CWT of the signal at different scales. The strength of the line for each word is the word-level prominence which is determined by the cumulative sum of scale values of the line with weights, shown as follows:

$$\begin{aligned}\boldsymbol{x}_{\mathrm{prm}} = W_s(a_0, t_{i_0,0}) + \ldots + \\ \log(j+1)a^{-j/2}W_s(a_0a^j, t_{i_j,j}),\end{aligned} \tag{1}$$

where $\boldsymbol{x}_{\mathrm{prm}}$ is word-level prominence, a_0 denotes the finest scale in CWT, a defines the spacing between chosen scales, j denotes sale, $t_{i_j,j}$ is a time point where the local maxima occurred in the a_0a^j scale. $W_s(a_0a^j, t_{i_j,j})$ denotes the CWT amplitude in $t_{i_j,j}$ time point at a_0a^j level scale.

Word-level textual feature extraction We extract word-level textual features by applying the fastText [15], a word-level text embedding model, to a text embedding.

We then concatenate prosodic factors $\boldsymbol{x}_{\mathrm{psd}}$, prominence $\boldsymbol{x}_{\mathrm{prm}}$, and text embedding $\boldsymbol{x}_{\mathrm{wrd}}$ to multi-modal features $\boldsymbol{x}_{\mathrm{mul}}$ as the SER input, shown as

* 単語レベル強調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, 齋藤 佑樹, 猿渡 洋 (東大院・情報理工)

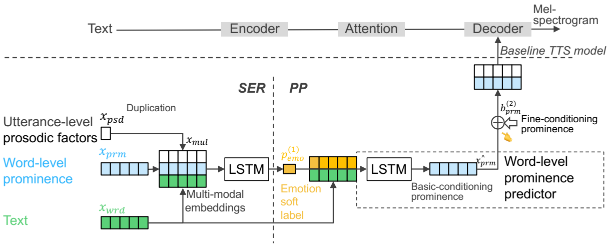

Fig. 1 The proposed emotion-controllable model

follows:

$$\boldsymbol{x}_{\text{mul}} = \text{Concat}_{\text{wrd}}(\boldsymbol{x}_{\text{psd}}, \boldsymbol{x}_{\text{prm}}, \boldsymbol{x}_{\text{wrd}}). \tag{2}$$

where $\text{Concat}_{\text{wrd}}$ is concatenation at word-level.

The SER model architecture The SER model is a 2-layer LSTM model followed by a softmax output layer. It estimates emotion soft labels $\boldsymbol{p}_{\text{emo}}^{(1)}$, where superscript 1 indicates that it is used for the first stage of control. The emotion soft labels are the posterior probabilities for predicting the emotion labels $\boldsymbol{y}_{\text{emo}}$, conditional on multi-modal features $\boldsymbol{x}_{\text{mul}}$:

$$\boldsymbol{p}_{\text{emo}}^{(1)} = \text{SER}(\boldsymbol{x}_{\text{mul}}) = P(\boldsymbol{y}_{\text{emo}}|\boldsymbol{x}_{\text{mul}}). \tag{3}$$

The SER architecture is shown on the left below side of Fig. 1. It is trained by minimizing the cross-entropy loss L_{SER} between the groud-truth emotion labels and estimated emotion soft labels:

$$L_{\text{SER}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{\text{emo}_{i,c}}), \tag{4}$$

where $y_{i,c}$ is an emotion label indicator, assigned 0 or 1, indicating whether the i-th utterance belongs to the c-th emotion (1) or not (0). N and C are the total numbers of utterances and emotion categories, respectively. $p_{\text{emo}_{i,c}}$ is the estimated emotion soft label of the i-th utterance for the c-th emotion.

2.2 PP model

The word-level prominence predictor PP_{prm} predicts basic-conditioning prominence $\hat{\boldsymbol{x}}_{\text{prm}}$ from a concatenation of text embedding $\boldsymbol{x}_{\text{wrd}}$ and emotion soft labels $\boldsymbol{p}_{\text{emo}}^{(1)}$:

$$\hat{\boldsymbol{x}}_{\text{prm}} = \text{PP}_{\text{prm}}(\boldsymbol{x}_{\text{wrd}}, \boldsymbol{p}_{\text{emo}}^{(1)}), \tag{5}$$

where $\boldsymbol{p}_{\text{emo}}^{(1)}$ is also the SER output in training and manually assigned in inference.

The PP_{prm} also consists of a 2-layer LSTM network followed by an FC layer and a sigmoid layer in sequence, as shown in Fig. 1.

The SER and PP models are jointly trained by minimizing the sum of the SER and PP losses $L_{(\text{SER+PP})}$ on an emotion-labeled dataset. The objective function is:

$$L_{(\text{SER+PP})} = L_{\text{SER}} + L_{\text{PP}}. \tag{6}$$

where L_{PP} indicates the L2 loss of and predicted and ground truth prominence.

2.3 Emotion-controllable TTS Model

The emotion-controllable TTS model enables two-stage control by extending the baseline Tacotron2 model with the concatenated SER and PP models, as shown in Fig. 1. The SER model takes multi-modal features as input and outputs emotion soft labels, which are fed into the PP model along with text embedding. The PP model outputs basic-conditioning prosodic factors and prominence, which are then fed into the TTS decoder along with phoneme embedding from the TTS encoder.

We concatenate basic-conditioning prominence and phoneme embedding by alignment with the English grapheme-to-phoneme conversion algorithm [1]. The proposed TTS model emoTTS is conditioned on the concatenated embeddings $\hat{\boldsymbol{c}}_{\text{con}}$ to synthesize speech $\boldsymbol{y}_{\text{speech}}$:

$$\boldsymbol{y}_{\text{speech}} = \text{emoTTS}(\hat{\boldsymbol{c}}_{\text{con}}), \tag{7}$$

where

$$\hat{\boldsymbol{c}}_{\text{con}} = \text{Concat}_{\text{phn}}(\hat{\boldsymbol{c}}_{\text{prm}}, \boldsymbol{x}_{\text{phn}}), \tag{8}$$

where $\text{Concat}_{\text{phn}}$ is a concatenation of prominence $\hat{\boldsymbol{c}}_{\text{prm}}$, and phoneme embedding $\boldsymbol{x}_{\text{phn}}$.

The conditioning prominence $\hat{\boldsymbol{c}}_{\text{prm}}$ comprises two parts: basic-conditioning prominence $\hat{\boldsymbol{x}}_{\text{prm}}$ and fine-conditioning prominence (i.e., prominence bias or fine-conditioning bias) $\boldsymbol{b}_{\text{prm}}^{(2)}$, shown in Eq. (9).

$$\begin{aligned}\hat{\boldsymbol{c}}_{\text{prm}} &= \text{PP}_{\text{prm}}(\boldsymbol{x}_{\text{wrd}}, \boldsymbol{p}_{\text{emo}}^{(1)}) + \boldsymbol{b}_{\text{prm}}^{(2)} \\ &= \hat{\boldsymbol{x}}_{\text{prm}} + \boldsymbol{b}_{\text{prm}}^{(2)},\end{aligned} \tag{9}$$

where PP_{prm} is the word-level prominence predictor.

According to Eq. (7) and Eq. (9), the proposed TTS model enables the inter-emotion and intra-emotion control by converting emotion soft labels $\boldsymbol{p}_{\text{emo}}^{(1)}$ and fine-conditioning prominence $\boldsymbol{b}_{\text{prm}}^{(2)}$ to $\hat{\boldsymbol{c}}_{\text{prm}}$.

The proposed TTS model is optimized by minimizing the additive loss $L_{\text{emo\_TTS}}$ of $L_{\text{Tacotron2}}$ and L_{PP}:

$$L_{\text{emo\_TTS}} = L_{\text{Tacotron2}} + L_{\text{PP}}. \tag{10}$$

In inference, the proposed two-stage control TTS model can synthesize speech in the following ways:

1. Enabling only the first stage of control. Given emotion soft labels, the proposed model can synthesize speech with a specified emotion.

2. Enabling both the first and second stages of controls. Given emotion soft labels and fine-conditioning prominence, the proposed model can synthesize specified emotional speech with slightly changed prominence.

3 Experimental setup

3.1 Data

We used the IEMOCAP corpus [16] for pre-training SER and PP models and the Blizzard2013 corpus [17] for training the proposed 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 %

[1] The English grapheme-to-phoneme conversion package:link

and 20 % for training and testing the SER and PP 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.

3.2 Model parameter and features

The SER model consisted of a 3-layer LSTM network with 128 hidden units and a 128 × 3 fully connected (FC) layer, followed by a softmax activation. The PP model included a 2-layer LSTM network with 128 hidden units and a 128 × 1 FC layer followed by a sigmoid activation function. It took a 303-dimensional joint vector as input and output a 1-dimensional prominence.

The backbone Tacotron2 consisted of an encoder network that converted phoneme embedding into a hidden text representation and a decoder network that predicted mel-spectrograms from hidden text and prominence with attention. Specifically, the encoder network consisted of 3-layer 1-dimensional convolutions with 512 filters and a 5 × 1 window size. A phoneme embedding represented by a 512-dimensional vector was passed through the encoder network whose output was a hidden text representation.

4 Evaluation

4.1 Controllability of emotion soft labels (first stage of control)

We first evaluated the emotion controllability of our proposed model when conditioning on emotion soft labels during the first stage (inter-emotion) of control. To obtain the perceived emotion of synthesized speech, we conducted a preference test in which each participant was required to choose angry, neutral, and sad speech, respectively, from a set of three synthesized speech with angry, neutral, and sad emotions. We synthesized 10 utterances for each emotion (angry, neutral, and sad) as test speech from randomly selected sentences in the BC2013 dataset by conditioning corresponding emotion soft labels to 1.0. We applied the emotion soft label to 1.0 for better representativeness [18]. This test was conducted on the Amazon Mechanical Turk with 50 participants and 10 sets of speech for each participant. The performance of emotion controllability was evaluated by the accuracy, precision, recall, and F1-score which indicates the distinguishability for each emotion category. The results demonstrated that the accuracy, precision, recall, and F1-score were 51%, 52%, 50%, and 51% on average of three emotions. Specifically, the accuracy of angry speech was 60% which was relatively higher than other emotions.

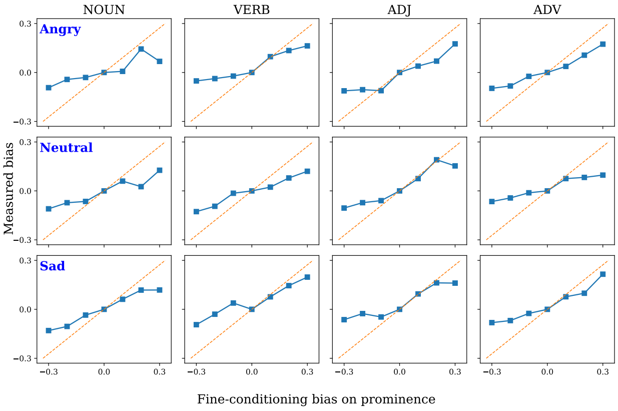

Fig. 2 Correlation between fine-conditioning and observed biases when fine-conditioning on the prominence of NOUN, VERB, ADJ, and ADV words for three emotions

4.2 Linear controllability of word-level prominence (second stage of control)

We evaluated the linear controllability of our proposed model by fine-conditioning on prominence during the second stage (intra-emotion) of control. We defined a linear controllability score using the Pearson Correlation Coefficient (PCC) between the fine-conditioning biases and measured biases for prominence. This can be represented by PCC($\boldsymbol{b}_{\mathrm{prm}}, \boldsymbol{b}'_{\mathrm{prm}}$), where $\boldsymbol{b}_{\mathrm{prm}}$ is fine-conditioning biases and $\boldsymbol{b}'_{\mathrm{prm}}$ is measured biases indicating the difference in prominence between the synthesized speech with fine-conditioning bias and without bias (fine-conditioning bias $= 0$).

In the experiment, we found that word-level prominence was distributed differently depending on the part of speech in the training dataset where the prominence of NOUN, VERB, ADJ, and ADV words was distributed close to a normal distribution. Therefore, we only experiment on the NOUN, VERB, ADJ, and ADV words. To synthesize the evaluation speech, we also input 50 sentences selected from the BC2013 dataset, and for each sentence, we fine-conditioned on the prominence of NOUN, VERB, ADJ, and ADV words, respectively, with seven biases for each, ranging from -0.3 to 0.3 with a 0.1 step, for angry (angry $= 1.0$), neutral (neutral $= 1.0$), and sad emotion (sad $= 1.0$). In total, we synthesized $2,100$ speech samples.

As the result, the average PCC($\boldsymbol{b}_{\mathrm{prm}}, \boldsymbol{b}'_{\mathrm{prm}}$) score of angry (0.93), neutral (0.97), sad (0.96), and overall (0.95) emotions showed strong linear controllability on the prominence of the NOUN, VERB, ADJ,

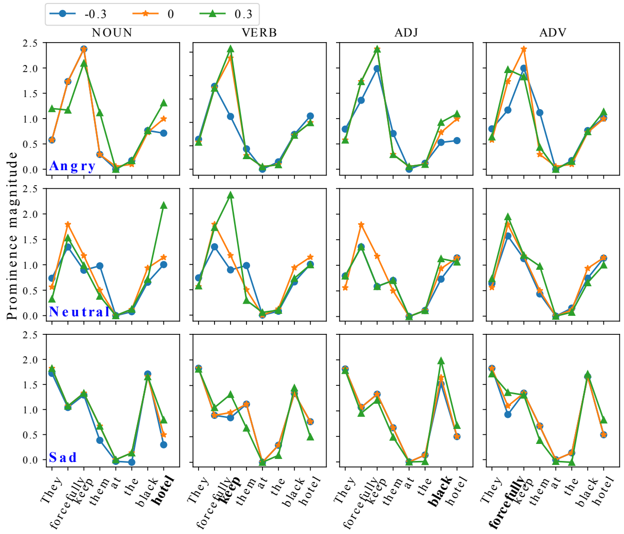

Fig. 3 Prominence contours of an utterance synthesized by conditioning on the angry, neutral, and sad emotions and fine-conditioning prominence on NOUN, VERB, ADJ, and ADV words with three biases (-0.3, 0, and 0.3) for each emotion. The sample sentence is "They forcefully keep them at a black hotel". The NOUN, VERB, ADJ, and ADV words correspond to "hotel", "keep", "black", and "forcefully", respectively.

and ADV words with p-value < 0.05. We also visualized the correlation between $\boldsymbol{b}_{\mathrm{prm}}$ and $\boldsymbol{b}'_{\mathrm{prm}}$ on the three emotions, as shown in Fig. 2.

We also visualized the prominence contours of utterances, synthesized by conditioning on the different emotions and prominence, as shown in Fig. 3. From the result, the prominence of fine-conditioned words increased (or decreased) when the conditioning bias increased (or decreased).

4.3 Subjective evaluation

We evaluated the quality of speech by a mean opinion score (MOS) test on 360 synthesized speech where each of the 10 sentences had 36 variations (3 emotions × 4 parts of speech × 3 biases). We conducted the test on the Amazon Mechanical Turk with 50 participants, each of whom was given 36 speech samples and required to choose speech quality for each speech in five stages (1: very bad, 5: very good). The result shows the MOS = 3.9, which is comparable to the method that can only condition prominence.

5 Conclusion

We proposed a two-stage emotion-controllable text-to-speech (TTS) model that can condition on inter-emotion (e.g., angry) in the first stage and fine-condition on intra-emotion with word-level prominence in the second stage of control. Due to the two-stage design, our model enables inter-emotion controllability and increases intra-emotion diversity. The results show that we can 1) condition the proposed model on emotion and synthesize adequately emotion-distinguishable speech (emotion-distinguishable score = 51%), 2) linearly fine-condition on the prominence of NOUN, VERB, ADJ, and ADV words for the angry, neutral and sad emotions, and finally 3) synthesize speech with comparable audio quality (MOS = 3.9) to that of the conventional methods. In future, we plan to enable more emotion controllability and their better combination with prominence.

Acknowledgement: This work is supported by Moonshot R&D Grant Number JPMJPS2011.

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

[3] 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.

[4] X. Luo et al., "Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 794–799.

[5] T. Li et al., "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[6] 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.

[7] 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.

[8] 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.

[9] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[10] A. Suni et al., "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[11] H. Li et al., "Emphasis: An emotional phoneme-based acoustic model for speech synthesis system," *arXiv preprint arXiv:1806.09276*, 2018.

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

[13] M. Mauch, S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[14] M. Vainio et al., "Emphasis, word prominence, and continuous wavelet transform in the control of hmm-based synthesis," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015, pp. 173–188.

[15] P. Bojanowski et al., "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

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

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

[18] 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.