APSIPA Transactions on Signal and Information Processing, XXXX, XX: 1–35

Emotion-controllable Speech Synthesis using Emotion Soft Label, Utterance-level Prosodic Factors, and Word-level Prominence

Xuan Luo[1], Shinnosuke Takamichi[1], Yuki Saito[1], Tomoki Koriyama[2] and Hiroshi Saruwatari[1]

[1] *The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan*
[2] *CyberAgent, Shibuya-ku, Tokyo, Japan*

ABSTRACT

We propose a two-stage emotion-controllable text-to-speech (TTS) model that can increase the diversity of intra-emotion variation and also preserve inter-emotion controllability in synthesized speech. Conventional emotion-controllable TTS models increase the diversity of intra-emotion variation by controlling fine-grained emotion strengths; however, such models cannot control various prosodic factors (e.g., pitch). While other methods directly condition TTS models on intuitive prosodic factors, they cannot control emotions. Our proposed two-stage emotion-controllable TTS model extends the Tacotron2 model with a speech emotion recognizer (SER) and a prosodic factor generator (PFG) to solve this problem. In the first stage, we condition our model on emotion soft labels predicted by the SER model to enable inter-emotion controllability. In the second stage, we fine-condition our model on utterance-level prosodic factors and word-level prominence generated by the PFG model from emotion soft labels, which provides intra-emotion diversity.

ISSN 2161-1823; DOI 10.1561/103.00000003

Due to this two-stage control design, we can increase intra-emotion diversity at both the utterance and word levels, and also preserve inter-emotion controllability. The experiments achieved 1) 51% emotion-distinguishable accuracy on average when conditioning on soft labels of three emotions, 2) average linear controllability scores of 0.95 when fine-conditioning on prosodic factors and prominence, respectively, and 3) comparable audio quality to conventional models.

Keywords: Emotion-controllable speech synthesis, expressive speech synthesis, controllable speech synthesis, text to speech, speech emotion recognition

1 Introduction

Text-to-speech (TTS) models synthesize human-like speech which includes linguistic and paralinguistic information. The fast development of deep learning models [12, 27, 28, 31, 37, 39] has made it possible to synthesize understandable speech from a linguistic perspective. On the other hand, synthesizing human-like speech with diverse paralinguistic information is not an easy task. The paralinguistic information of human speech, such as emotion, is expressed by various types and strengths of prosodic factors (e.g., pitch) [14, 29] or prominence (i.e., emphasis) [13, 43] at different levels. Even when the same words are spoken with the same emotion, using slightly different prosodic factors or prominence can cause listeners to have completely different perceptions of meaning and feeling. Therefore, it is important for the TTS models to control not only emotion variations but also the variations of prosodic factors or prominence on the basis of a given emotion. However, few TTS models are capable of synthesizing speech with such diverse emotion variations as that of actual human speech.

The diverse emotion variations can be broadly separated into inter- and intra-emotion variations. Inter-emotion variation primarily represents significant differences between emotions, while intra-emotion variation represents minor differences within an emotion. Conventional

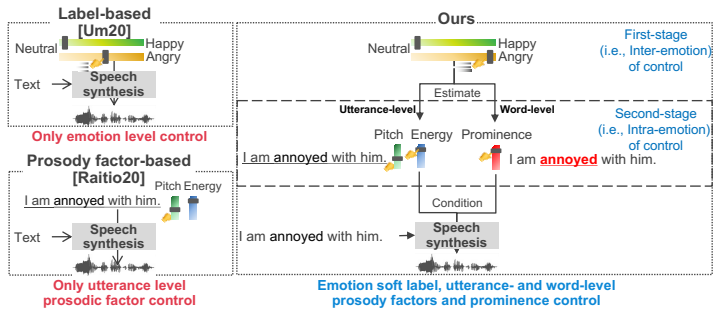

Figure 1: Overview of the proposed model

emotional TTS models generate an inter-emotion variation by integrating various inter-emotion representations, such as explicit emotion labels [18, 44] or implicit emotion embeddings [17, 21], into the TTS models. Compared to inter-emotion variation, intra-emotion variation yields richer emotion diversity. Subsequent studies generate an intra-emotion variation by controlling a finer granularity, such as emotion strength [19, 35, 44], which indicates the intensity of emotion. Such models can increase the diversity of different emotion strengths but not various prosodic factors (e.g., pitch), which are finer granularities than emotion strength. Considering emotion strength can be formed by various prosodic factors depending on different conditions (such as gender), the inability to control prosodic factors limits the generation of more diverse intra-emotion variation.

To synthesize speech with more diverse prosodic factors, studies condition TTS models on implicit prosodic factors. Subsequent studies aim to disentangle certain implicit prosodic factors in unsupervised manners by utilizing a reference encoder, a variational autoencoder (VAE), and multi-head attention-based models [32, 40, 42] from reference audio. These implicit prosodic factors are further used to condition the TTS models, which can increase the prosodic diversity in speech. However, these implicit prosodic factors are not always disentangled into desired ones, such as pitch, and energy. To solve such a problem, other studies directly condition TTS models on intuitive prosodic factors (e.g., mean of pitch) at utterance-level [26, 30] or phoneme-level [38]. These models can increase intra-emotion diversity but cannot control inter-emotions.

In this paper, we propose a two-stage emotion-controllable TTS model that can control both inter- and intra-emotion of synthesized speech. In the first stage, we control the emotion of synthesized speech by conditioning on emotion-soft labels. In the second stage, we fine-condition utterance-level prosody factors and word-level prominence estimated on the basis of emotion soft labels (i.e., emotion posterior probability). The overview of the proposed model is shown in Fig. 1. The evaluation results demonstrate that our model attains 51% emotion-distinguishable accuracy on average which is promising despite using only a narrative dataset. In addition, our model achieves average linear controllability scores of 0.95 when fine-conditioning on prosodic factors and prominence, respectively, which are comparable to the conventional method [26].

The remainder of this paper is organized as follows. In Section 2, we discuss work related to our study. In Section 3, we elaborate on the proposed model by breaking it down into the speech emotion recognizer (SER) and the prosodic factor generator (PFG), and the emotion-controllable TTS models. In Section 4, we explain the experimental setup consisting of data, preprocessing, model architecture, and training procedures. In Section 5, we first evaluate the performance of SER and PFG models. Then we evaluate the controllability of emotional soft labels and the linear controllability of prosodic factors and prominence, respectively.

2 Related Work

2.1 Emotion-controllable TTS

To control emotion, previous studies commonly conditioned their TTS models on explicit emotion labels [18, 22, 44] or implicit emotion embeddings [17, 21]. The explicit emotion labels were usually obtained from emotion-labeled speech datasets [18, 44], or emotion-predictive models [22]. The implicit emotion embeddings were generally trained by learnable models, such as a speech emotion recognition model [21] or a multi-head attention model [17]. To generate intra-emotion variation

in speech, subsequent studies conditioned on not only emotion labels or embeddings but also emotion strength, which is a finer granularity. The emotion strength was obtained by an interpolation method between emotion categories [35], a ranking function at utterance-level [44], or phoneme-level [19]. These models that were conditioned on emotion strengths can generate more diverse intra-emotion variation than the models conditioned on emotion labels. However, emotion strength is still not the smallest granularity to be controlled because it can be further represented by various prosodic factors [14, 29] or prominence [13, 43].

2.2 Prosody-controllable TTS

To control prosodic factors, previous studies conditioned their TTS models on prosodic factors using explicit prosodic factors or implicit prosodic factor tokens. The former directly conditioned TTS models on intuitive prosodic factors (e.g., pitch mean) on utterance-level [26, 30], or more fine-grained levels such as the phoneme-level [38]. On the other side, the latter aimed to learn disentangled prosodic factor tokens by training a reference encoder [32], a variational autoencoder [42], or multi-head attention-based models [40] from reference audio. The implicit methods generated more prosodic diversity than the explicit methods. However, there was no guarantee that they could disentangle these tokens into desired prosodic factors because the learned tokens contained plenty of other paralinguistic information (e.g., speaker or noise), which makes the disentanglement difficult. Meanwhile, both the explicit and implicit methods were incapable of controlling emotion in speech.

In addition to prosodic factors, Li et al. [20] condition their TTS model on word/phoneme-level prominence (i.e., emphasis) to make synthesized speech more diverse. The prominence that was used to condition the TTS model was mainly formed by specific prosodic factors, such as pitch, energy, and duration [33]. Because prominence is also related to emotion [34] and contributes to intra-emotion diversity, we also enable prominence control in this paper.

2.3 Prediction of prosodic factors and prominence from emotion

There is a strong relationship between emotions and prosodic factors [1, 14, 29] or prominence [2, 13, 43]. Akçay et al. [1] reported on the correlation between global statistics of prosodic factors including pitch, energy, and emotion states. For example, the average pitch increases in happy speech. Arias et al. [2] found that, in addition to prosodic factors, local prominence is also correlated with emotion state. For example, the intonation of happy speech usually increases at the end. Therefore, various statistics of prosodic factors [1, 14, 29] and prominence [13, 43] are utilized to predict emotions by utilizing different discriminative models.

However, few studies utilize emotion states to predict prosodic factors and prominence. Raitio et al. [26] predicted prosodic factors on which the TTS model is conditioned, from only text by utilizing a long short-term memory (LSTM) based module. Talman et al. [34] also predicted prominence from the text by utilizing BERT [9], a pre-trained language model. In controllable TTS models, predicting prosodic factors and prominence which are used for controlling without considering emotion will result in a limited variety in synthesized speech. Therefore, in our proposed emotion-controllable TTS model, we predict prosodic factors and prominence from both text and emotion, which can increase diversity in synthesized speech.

3 Proposed Method

We propose a two-stage emotion-controllable TTS model that enables conditioning on emotion soft labels in the first stage (inter-emotion) of control and fine-conditioning on the utterance-level prosodic factors (i.e., prosodic factors) and word-level prominence (i.e., prominence) in the second stage (intra-emotion) of control. To enable this two-stage control, we extend the baseline Tacotron2 model [31] with a speech emotion recognizer (SER) and a prosodic factor generator (PFG) model, as shown in Fig. 2. The SER model estimates the emotion soft labels, on which the TTS model is conditioned in the first stage of control, and the PFG model generates prosodic factors and prominence, on which is fine-conditioned in the second stage of control. We detail the SER, PFG, and the proposed emotion-controllable TTS model in the

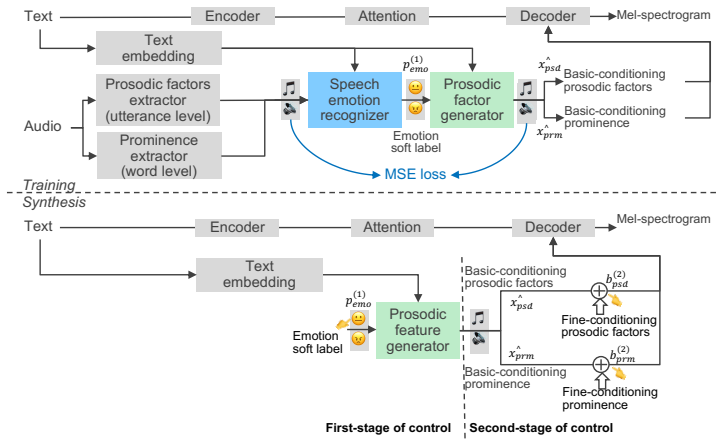

Figure 2: The architecture of the proposed emotion-controllable TTS model

following sections.

3.1 SER and PFG models

3.1.1 SER model

The SER model estimates emotion soft labels which are used for the first stage of control. The model takes multi-modal features consisting of prosodic factors, prominence, and textual features, as input. We utilize the multi-modal features as input because of their better performance on emotion estimation than unimodal features, which has been demonstrated in previous research [29]. In addition, the emotion soft labels estimated by multi-modal features in the SER model can be efficiently used to generate prosodic factors and prominence in the PFG model, which we will discuss later. The ground truth of prosodic factors, prominence, and textual features can be extracted by the following approach.

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. The pitch contour is predicted using the pYIN algorithm [23], and the energy contour is calculated by the root-mean-square value of the magnitude of each frame. These three statistics of pitch and energy,

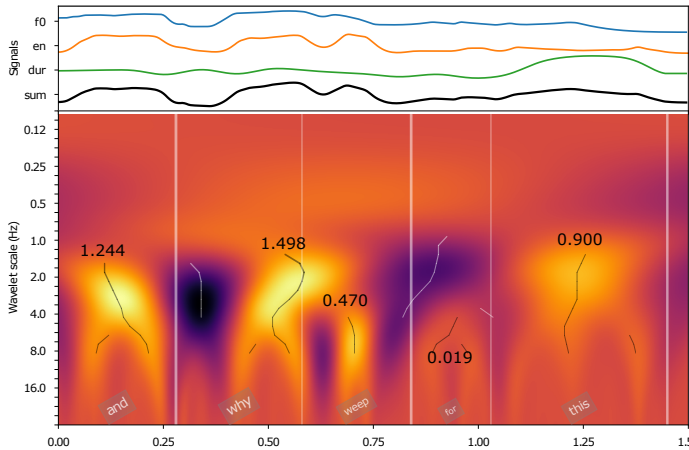

Figure 3: The word-level prominence extraction by applying the CWT-LoMA with a sum of signal contours of pitch, energy, and duration. The lines of maximum amplitude (LoMA) are shown in black, while the strength of each line indicates the word-level prominence which is shown in decimal numerical. The white lines are the minimum amplitude which indicates the boundaries of words.

a total of 6-dimensional prosodic factors, are used to condition the proposed TTS model because they are related to speech emotion [29]. Each of the extracted prosodic factors is normalized to the range from 0 to 1 by applying Min-Max normalization over the training dataset.

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 [33]. The CWT of a composition of pitch, energy, and duration contours can 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 [36] are lines that can identify and quantify word-level prominence by connecting nearby peaks in the CWT of the signal at different scales, as shown in Fig. 3. The strength of the line for each word (decimal numerical in Fig. 3) is the word-level prominence which is determined by the cumulative sum of scale values of the line

with weights, shown as follows:

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

where $\boldsymbol{x}_{\text{prm}}$ is word-level prominence, a_0 denotes the finest scale in CWT, $t_{i_j,j}$ 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 in $t_{i_j,j}$ time point at a_0a^j level scale. From this formula, we can conclude that the higher levels of the hierarchy are given more weight by the logarithmic term than the lower levels.

To extract prominence, we first align speech and its corresponding text at the word level by using the Montreal Forced Aligner (MFA) [24], a text-speech alignment tool. Second, we extract the prominence of each word to a scalar value by using a wavelet prosody toolkit which is available here [1]. The extracted word prominence indicates the degree of emphasis, which is also related to speech emotion [13]. The prominence is normalized to the range from 0 to 1 by applying the Min-Max normalization over the training dataset.

Word-level textual feature extraction We extract word-level textual features by applying the fastText [4], a word-level text embedding model, to a text embedding. The text embedding is an $L \times M$ tensor, where the L indicates the number of words in a sentence and M is the embedding dimension. Similar to prosodic factors and prominence, textual features are also related to speech emotion [29].

Multi-modal features We concatenate prosodic factors, prominence, and text embedding as multi-modal features to predict emotion soft labels. To do this, we extend prosodic factors $\boldsymbol{x}_{\text{psd}}$ from a 1×6 tensor to a word-length $L \times 6$ tensor and concatenate it to prominence $\boldsymbol{x}_{\text{prm}}$ and text embedding $\boldsymbol{x}_{\text{wrd}}$ along the L dimension. To avoid the domination of text embedding, we upsample the prosodic factors and prominence from 6 and 1 dimensions to 16 and 8 dimensions, respectively. We denote this concatenation as $\text{Concat}_{\text{wrd}}$, shown as follows:

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

The SER model is a 2-layer LSTM model followed by a softmax output layer. It estimates emotion soft labels $\boldsymbol{p}^{(1)}_{\text{emo}}$, where superscript

[1]The wavelet prosody toolkit:link

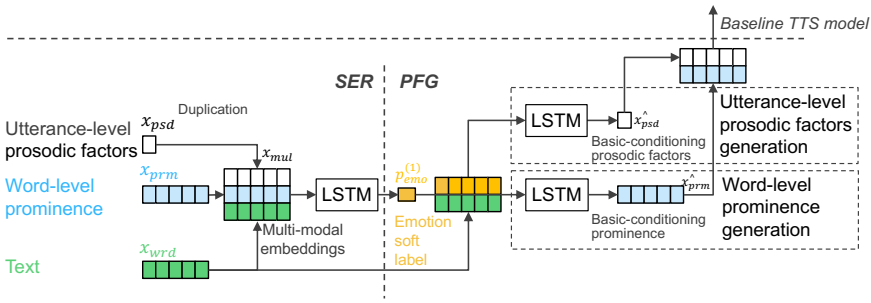

Figure 4: The SER and PFG models

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 side of Fig. 4.

Training objective The SER model 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.

3.1.2 PFG model

The PFG model generates basic-conditioning prosodic factors and prominence on which the TTS model is conditioned, to synthesize speech corresponding to a given emotion. The PFG model consists of an utterance-level prosodic factor generator and a word-level prominence generator.

Utterance-level prosodic factor generator The utterance-level prosodic factor generator $\mathrm{PFG}_{\mathrm{psd}}$ generates basic-conditioning prosodic factors $\hat{\boldsymbol{x}}_{\mathrm{psd}}$ from a concatenation of text embedding $\boldsymbol{x}_{\mathrm{wrd}}$ and emotion soft labels $\boldsymbol{p}^{(1)}_{\mathrm{emo}}$:

$$\hat{\boldsymbol{x}}_{\mathrm{psd}} = \mathrm{PFG}_{\mathrm{psd}}(\boldsymbol{x}_{\mathrm{wrd}}, \boldsymbol{p}^{(1)}_{\mathrm{emo}}), \tag{5}$$

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

The $\mathrm{PFG}_{\mathrm{psd}}$ consists of a 2-layer LSTM network followed by a fully connected (FC) layer and a sigmoid layer in sequence.

Word-level prominence generator Similarly, the word-level prominence generator $\mathrm{PFG}_{\mathrm{prm}}$ generates basic-conditioning prominence $\hat{\boldsymbol{x}}_{\mathrm{prm}}$ from a concatenation of text embedding $\boldsymbol{x}_{\mathrm{wrd}}$ and emotion soft labels $\boldsymbol{p}^{(1)}_{\mathrm{emo}}$:

$$\hat{\boldsymbol{x}}_{\mathrm{prm}} = \mathrm{PFG}_{\mathrm{prm}}(\boldsymbol{x}_{\mathrm{wrd}}, \boldsymbol{p}^{(1)}_{\mathrm{emo}}), \tag{6}$$

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

The $\mathrm{PFG}_{\mathrm{prm}}$ also consists of a 2-layer LSTM network followed by an FC layer and a sigmoid layer in sequence.

The architectures of the $\mathrm{PFG}_{\mathrm{psd}}$ and $\mathrm{PFG}_{\mathrm{prm}}$ models are shown in the upper right and lower right of Fig. 4, respectively.

Training objective The $\mathrm{PFG}_{\mathrm{psd}}$ and $\mathrm{PFG}_{\mathrm{prm}}$ models are jointly optimized by minimizing the PFG loss L_{PFG} which is calculated by the sum of L2 loss of prosodic factors L_{psd} and prominence L_{prm}. The objective function is:

$$L_{\mathrm{PFG}} = \sum_{i=1}^{N} L_{\mathrm{psd}_i} + \sum_{i=1}^{N} L_{\mathrm{prm}_i}, \tag{7}$$

where i indicates an utterance index and N is the total number of utterances.

To increase the fit of the SER and PFG models, they are first trained jointly by minimizing the sum of the SER and PFG losses on an emotion-labeled dataset. The objective function is:

$$L_{(\mathrm{SER+PFG})} = L_{\mathrm{SER}} + L_{\mathrm{PFG}}. \tag{8}$$

3.2 Emotion-controllable TTS Model

The emotion-controllable TTS model enables two-stage control by extending the baseline Tacotron2 model with the concatenated SER and PFG models, as shown in Fig. 2. The SER model takes multi-modal features as input and outputs emotion soft labels, which are fed into the PFG model along with text embedding. The PFG model outputs basic-conditioning prosodic factors and prominence, which are then fed into the TTS decoder along with phoneme embedding from the TTS encoder.

Because the basic-conditioning prosodic factors, prominence, and phoneme embedding have different shapes, to concatenate them together, we extend prosodic factors and prominence to phoneme lengths by simple duplication and alignment by the English grapheme-to-phoneme conversion algorithm [2], respectively.

The proposed TTS model emoTTS is conditioned on the concatenated embeddings $\hat{\boldsymbol{c}}_{\text{con}}$ to synthesize speech $\boldsymbol{y}_{\text{speech}}$:

$$\boldsymbol{y}_{\text{speech}} = \text{emoTTS}(\hat{\boldsymbol{c}}_{\text{con}}), \tag{9}$$

where

$$\hat{\boldsymbol{c}}_{\text{con}} = \text{Concat}_{\text{phn}}(\hat{\boldsymbol{c}}_{\text{psd}}, \hat{\boldsymbol{c}}_{\text{prm}}, \boldsymbol{x}_{\text{phn}}), \tag{10}$$

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

Specifically, $\hat{\boldsymbol{c}}_{\text{psd}}$ includes two parts: basic-conditioning prosodic factors $\hat{\boldsymbol{x}}_{\text{psd}}$ and fine-conditioning prosodic factors (i.e., prosodic factors biases or fine-conditioning biases) $\boldsymbol{b}^{(2)}_{\text{psd}}$, where the superscript 2 indicates that they are used for the second stage of control, shown in Eq. 11. We condition the TTS model on $\hat{\boldsymbol{x}}_{\text{psd}}$ to synthesize speech with a given emotion. We can achieve this because the $\hat{\boldsymbol{x}}_{\text{psd}}$ is generated from emotion soft labels by the PFG model. We fine-condition the TTS model on $\boldsymbol{b}^{(2)}_{\text{psd}}$ to enable a slight change of basic prosodic factors to provide diversity.

$$\begin{aligned} \hat{\boldsymbol{c}}_{\text{psd}} &= \text{PFG}_{\text{psd}}(\boldsymbol{x}_{\text{wrd}}, \boldsymbol{p}^{(1)}_{\text{emo}}) + \boldsymbol{b}^{(2)}_{\text{psd}} \\ &= \hat{\boldsymbol{x}}_{\text{psd}} + \boldsymbol{b}^{(2)}_{\text{psd}}, \end{aligned} \tag{11}$$

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

where the $\mathrm{PFG}_{\mathrm{psd}}$ is the utterance-level prosodic factor generator and the $\boldsymbol{x}_{\mathrm{wrd}}$ is the word-level text embedding.

Similarly, the conditioning prominence $\hat{\boldsymbol{c}}_{\mathrm{prm}}$ also comprises two parts: basic-conditioning prominence $\hat{\boldsymbol{x}}_{\mathrm{prm}}$ and fine-conditioning prominence (i.e., prominence bias or fine-conditioning bias) $\boldsymbol{b}^{(2)}_{\mathrm{prm}}$ and behaves in the same way as the $\hat{\boldsymbol{c}}_{\mathrm{psd}}$, shown in Eq. 12.

$$\begin{aligned}\hat{\boldsymbol{c}}_{\mathrm{prm}} &= \mathrm{PFG}_{\mathrm{prm}}(\boldsymbol{x}_{\mathrm{wrd}}, \boldsymbol{p}^{(1)}_{\mathrm{emo}}) + \boldsymbol{b}^{(2)}_{\mathrm{prm}} \\ &= \hat{\boldsymbol{x}}_{\mathrm{prm}} + \boldsymbol{b}^{(2)}_{\mathrm{prm}},\end{aligned} \tag{12}$$

where $\mathrm{PFG}_{\mathrm{prm}}$ is the word-level prominence generator.

According to Eq. 11 and Eq. 12, the proposed TTS model that is conditioned on $\hat{\boldsymbol{c}}_{\mathrm{psd}}$ and $\hat{\boldsymbol{c}}_{\mathrm{prm}}$ functionally enables the inter-emotion control by applying emotion soft labels $\boldsymbol{p}^{(1)}_{\mathrm{emo}}$ and the intra-emotion control by fine-conditioning prosodic factors $\boldsymbol{b}^{(2)}_{\mathrm{psd}}$ (or prominence $\boldsymbol{b}^{(2)}_{\mathrm{prm}}$), respectively. It is worth noting that although the proposed model can control both emotion and prosodic factors (or prominence), in reality, we only need to condition our model on the pure prosodic factors (or prominence). Such characteristics can efficiently avoid complicated control over correlated emotion and prosodic factors (or prominence).

Training objective The proposed TTS model is optimized by minimizing the additive loss $L_{\mathrm{emo\_TTS}}$ of $L_{\mathrm{Tacotron2}}$ and L_{PFG}:

$$L_{\mathrm{emo\_TTS}} = L_{\mathrm{Tacotron2}} + L_{\mathrm{PFG}}. \tag{13}$$

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 prosodic factors or prominence, the proposed model can synthesize specified emotional speech with slightly changed prosodic factors or prominence.

4 Experimental Setup

4.1 Data and Preprocessing

We first used the IEMOCAP dataset [5] to jointly train the SER and PFG models. Then we trained our emotion-controllable TTS model with the SER model frozen on the Blizzard Challenge 2013 (BC2013) dataset [15].

IEMOCAP is a multimedia English conversation dataset containing speech, video, etc., performed by five male and five female speakers with nine different emotions. The speech part includes 10,039 utterances (about 12 hours) recorded with a sampling rate of 16,000 Hz. Each utterance is annotated by an emotion label ranging from nine emotions, including anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral state. To focus only on the significant emotions, we utilized four emotion categories: angry (i.e., anger), sad (i.e., sadness), neutral (i.e., neutral state), and happy (either happiness or excitement). Specifically, we combined happiness and excitement into one happy emotion, following Sahu et al.'s [29] method. The speech labeled with these four emotions was then used to train the SER to preprocess the BC2013 datasets.

BC2013 is an audiobook dataset containing 340 hours of speech recorded by a professional female speaker in narrative and expressive styles. BC2013 is a high-quality dataset encoded at a sampling rate of 22,050 Hz.

BC2013 preprocessing To efficiently train the proposed emotion-controllable model, we preprocessed BC2013 to select a subset of the dataset that includes a higher percentage of expressive speech than the original one. In addition, we predicted emotion labels for each utterance. The preprocessing consists of three steps:

1. **Character utterance selection**. First, we selected all utterances spoken by characters in BC2013 because they were more likely to contain expressive speech than others. The characters' utterances were extracted by selecting the sentences enclosed in single or double quotation marks in transcripts from the BC2013 dataset. This approach is similar to that of previous study [25]. To balance character and non-character utterances, we added all utterances in Jane Eyre, Emma, A Little Princess, and Twenty Thousand Leagues Under the Seas fictions to the extracted characters' utterances.

2. **Emotion soft labels estimation**. Second, we estimated emotion soft labels for the utterances obtained in step 1 using the SER model pre-trained on the IEMOCAP dataset with the four emotion categories.

3. **Emotion category filter**. Third, we conducted a simple listening test to filter out emotion categories that are incorrectly estimated in step 2, which may result in unexpected emotion control. The details of the listening test are provided in Appendix A.

Finally, we collected a total of 18,638 utterances (about 75 hours), including 4,416 angry, 6,762 neutral, and 7,460 sad utterances. The utterances in the happy category were dropped out because they did not sound as happy in the listening test during step 3.

4.2 Model Architecture and Training

4.2.1 Model Architecture

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 SER model took 324-dimensional multi-modal features, a combination of 300-dimensional text embedding, 16-dimensional upsampled utterance-level prosodic factors from the original 6-dimensional, and 8-dimensional unsampled word-level prominence from the original 1-dimensional as input and output the 3-dimensional emotion soft labels (angry, neutral, and sad).

The PFG model consisted of an utterance-level prosodic factor generator and a word-level prominence generator. The architecture of the former consisted of a 2-layer LSTM network with 128 hidden units and 128×6 FC layers followed by a sigmoid activation function. The input was a 303-dimensional joint vector concatenated by a 300-dimensional text embedding and 3-dimensional emotion soft labels. The output was 6-dimensional predicted prosodic factors. Similarly, the latter model included a 2-layer LSTM network with 128 hidden units and a 128×1 FC layer followed by a sigmoid activation function. It also 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 prosodic representations.

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.

The decoder network included an autoregressive recurrent neural network, which consisted of a 2-layer LSTM with 1,024 hidden units, location-sensitive attention [7], which is an extension of additive attention [3], a pre-net consisted of a 2-layer FC network with 256 hidden units, and a post-net consisted of a 5-layer 1-dimensional convolutional network with 512 filters. In addition, we also introduced a psd-net which converted prosodic factors and prominence into a hidden prosodic presentation. A hidden text representation, the encoder output, was consumed by location-sensitive attention which summarized weighted hidden text representations into a 512-dimensional context vector. A previous mel-spectrogram prediction was passed through the pre-net whose output was a 256-dimensional vector, while prosodic factors and prominence were passed through the psd-net whose output was a 32-dimensional vector. We applied dropout with 0.5 dropout rate to the output of the pre-net for better audio quality in both the training and inference stages, following [31]. The context vector of attention output and the psd-net output were concatenated and passed into the autoregressive recurrent neural network which predicted the mel-spectrogram one frame at a time. The predicted mel-spectrogram was then fed into a post-net to improve the overall reconstruction.

The ground-truth mel-spectrograms were calculated by a short-time Fourier transform (STFT) on a window size of 2,048 samples and a hop length of 512 samples with a Hann window function. We then transformed the STFT magnitude to the mel scale by using an 80-channel mel filterbank spanning from 80 Hz to 7,600 Hz.

We utilized a Parallel WaveGAN model [41] as a vocoder to generate waveform samples conditioned on the predicted mel-spectrograms. The Parallel WaveGAN model was pre-trained on the LJSpeech dataset [11]

and is accessible online[3].

4.2.2 Training

We extracted prosodic factors and prominence of the IEMOCAP and BC2013 datasets using the approach described in Section 3. In the experiment, we removed 1,718 samples that could not be aligned correctly by the MFA.

The training process consisted of the SER and PFG joint training and the emotion-controllable TTS training. The former trained the SER and PFG models on the IEMOCAP datasets by optimizing the L_{SER} and L_{PFG} in a supervised manner. We used the Adam optimizer [16] with a learning rate of 0.001 and 200 epochs. The latter trained the emotion-controllable TTS model by optimizing the L_{PFG} and $L_{\text{Tacotron2}}$ with the SER frozen and the PFG fine-tuned. We also used the Adam optimizer [16], and the 0.001 learning rate started decaying exponentially to 0.00001 after 50,000 iterations.

5 Evaluation

We conducted two principal evaluations: 1) a preliminary evaluation of the SER and PFG models and 2) a controllability evaluation of the proposed emotion-controllable TTS. The preliminary evaluation of the proposed SER and PFG models was performed by comparing them with traditional SER [29] and PFG [33] models. In the latter evaluation, we first evaluated the emotion controllability of the proposed model when conditioning on emotion soft labels, and then we evaluated the linear controllability of utterance-level prosodic factors and word-level prominence when fine-conditioning them with respective biases.

5.1 SER and PFG Performance

We jointly trained the proposed SER and PFG models on the IEMOCAP dataset and further fine-tuned the trained PFG model on the BC2013 dataset while freezing the SER model. To evaluate the SER and PFG models, we randomly selected 80% of the IEMOCAP and

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

BC2013 datasets as training datasets and evaluated them on the remaining 20% of the datasets.

5.1.1 SER performance

We evaluated the SER model on the testing part of the IEMOCAP and emotion-labeled part of BC2013 datasets (described in Appendix A) on precision, recall, and F1-score. We conducted an ablation study on the effectiveness of each of the multi-modal features in predicting three emotions (angry, neutral, and sad) by training the SER models with only text, text, and prosodic factors [29], and multi-modal features of all three input. The results showed that our SER model with multi-modal features of text, prosodic factors, and prominence improved F1 scores by 7.8%, 1.3% on the IEMOCAP, and 5.7%, 2.0% on BC2013 datasets when compared with the other two benchmark models, respectively. The details of the results are shown in Table 3 in Appendix B. We argue that such results indicate the effectiveness of appending prosodic factors and prominence to text input for emotion prediction.

5.1.2 PFG performance

We also evaluated the PFG models on the testing part of the BC2013 dataset in terms of L2 loss. We utilized the conventional prosodic factor generator [26] and prominence generator [34] as our benchmark models, both of which only utilize text as input. The results indicated that our PFG models (both prosodic factor and prominence generators) outperformed the corresponding benchmark models by 0.008 (12.9%) and 0.005 (21.7%) on the absolute (relative) decrease of L2 loss, respectively, as shown in Table 4 in Appendix B. Thus, we can conclude that emotion soft labels, in addition to text, also contributed to predicting both prosodic factors and prominence.

In summary, our proposed SER model with extra prominence input and the PFG model with extra emotion soft labels input outperformed the conventional SER and PFG models in predicting emotion soft labels, and prosodic factors/prominence, respectively.

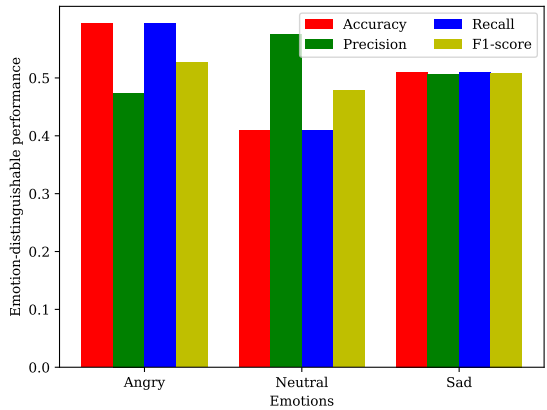

Figure 5: Emotion-distinguishable performance of synthesized speech

5.2 Emotion-controllable TTS performance

5.2.1 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 because the speech with the highest posterior probability has shown better representativeness than the others [6]. This test was conducted on the Amazon Mechanical Turk [8] 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, as shown in Fig. 5. Specifically, the accuracy of angry speech was 60% which was relatively higher than other

Table 1: Pearson correlation coefficient between fine-conditioning and measured prosody factor biases for three emotions

	Angry		Neutral		Sad	
Prosodic factors	PCC	p-value	PCC	p-value	PCC	p-value
Energy mean	0.99	2.87e-07	0.98	7.10e-07	0.99	1.69e-07
Energy range	0.98	8.23e-05	0.99	1.21e-05	0.97	2.43e-04
Energy SD	0.97	1.99e-05	0.96	3.70e-04	0.98	2.16e-05
Pitch mean	0.96	4.69e-05	0.99	6.82e-06	0.98	1.90e-05
Pitch range	0.83	1.83e-02	0.87	5.07e-03	0.91	1.74e-03
Pitch SD	0.89	7.10e-03	0.91	3.82e-03	0.97	2.73e-04

emotions. The accuracy of our model is lower than that of the conventional model (80%) [17]. We suggest the reason is two-fold. First, the conventional model was trained on an annotated private dataset recorded with good emotion distinguishability, while our model was trained on an unannotated narrative-style BC2013 dataset. Second, we made a trade-off between the accuracy of emotion distinguishability and the fine-conditioning ability of prosodic factors and prominence. Nevertheless, we still argue that our model provides good emotion controllability. Furthermore, the emotion-distinguishable accuracy of our model can be improved by training on more emotional speech, as the conventional model did.

5.2.2 Linear controllability of utterance-level prosodic factors (second stage of control)

We evaluated the linear controllability of our proposed model by fine-conditioning on prosodic factors during the second stage (intra-emotion) of control. We expected to slightly change prosodic factors of synthe-

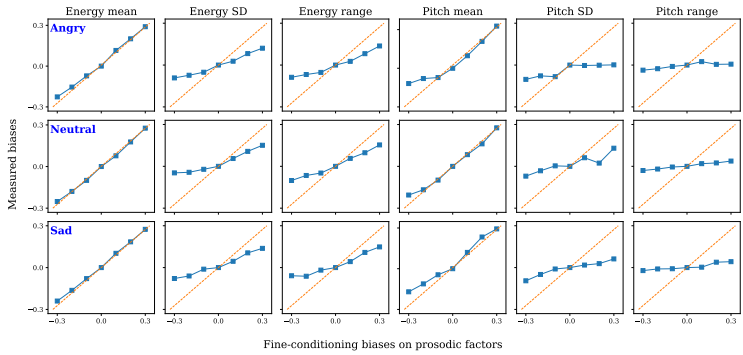

Figure 6: Correlation between fine-conditioning and observed biases when fine-conditioning on utterance-level prosodic factors for three emotions

sized speech to "biases" linearly relative to the fine-conditioning biases. To measure this linear relation, we defined a linear controllability score by utilizing the Pearson Correlation Coefficient (PCC) between the fine-conditioning biases and measured biases for each prosodic factor. This can be expressed as $\mathrm{PCC}(\boldsymbol{b}^{(2)}_{\mathrm{psd}}, \boldsymbol{b}'_{\mathrm{psd}})$, where $\boldsymbol{b}^{(2)}_{\mathrm{psd}}$ is the fine-conditioning biases and $\boldsymbol{b}'_{\mathrm{psd}}$ is the measured biases indicating the difference in prosodic factors between the speech synthesized with fine-conditioning biases and without fine-conditioning biases (fine-conditioning bias = 0), as shown in Eq. 14:

$$\begin{aligned}\boldsymbol{b}'_{\mathrm{psd}} = \mathrm{PSD}(\mathrm{emoTTS}(\hat{\boldsymbol{x}}_{\mathrm{psd}} + \boldsymbol{b}^{(2)}_{\mathrm{psd}})) \\ - \mathrm{PSD}(\mathrm{emoTTS}(\hat{\boldsymbol{x}}_{\mathrm{psd}})).\end{aligned} \tag{14}$$

PSD indicates the prosodic factor extraction, and $\hat{\boldsymbol{x}}_{\mathrm{psd}}$ denotes basic-conditioning prosodic factors, which were discussed in Section 3.1.2.

To synthesize the evaluation speech, we input 50 sentences selected from the BC2013 dataset, and for each sentence, we fine-conditioned on six prosodic factors 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 6,300 speech samples.

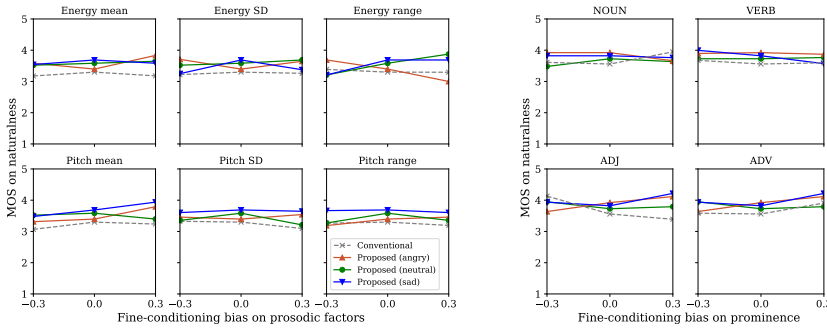

Figure 7: MOS scores of speech synthesized by the conventional model [26] and the proposed model when conditioning on angry/neutral/sad and fine-conditioning on prosodic factors (left) and prominence (right)

We calculated PCC($\boldsymbol{b}^{(2)}_{\text{psd}}, \boldsymbol{b}'_{\text{psd}}$), as shown in Table 1, and visualized the correlation between $\boldsymbol{b}^{(2)}_{\text{psd}}$ and $\boldsymbol{b}'_{\text{psd}}$ on angry, neutral, and sad evaluation speech, as shown in Fig. 6.

From the results, we can conclude that:

1. We can linearly fine-condition our model on the six prosodic factors for angry, neutral, and sad emotions, respectively. With the p-value (statistical significance) <0.05, the average PCC score for angry (0.93), neutral (0.95), sad (0.97), and overall (0.95) emotions showed strong linear controllability[4] on the proposed prosodic factors.

2. Compared with the prosodic factors of energy and pitch range/SD, the correlation lines (the blue lines) of the energy and pitch mean, as shown in Fig. 6, exhibit higher controlling slopes and are closer to the ideal controlling lines (the red dotted lines).

3. The linear controllability of our model on the prosodic factors was fairly comparable with the conventional method [26]. In addition, our model can also be conditioned on emotion, while the conventional method cannot.

[4]The definition of strong, moderate, and weak linear relationships are $1.0 \geq \text{PCC} \geq 0.6$, $0.6 > \text{PCC} \geq 0.4$, $0.4 > \text{PCC} \geq 0.0$ while p-value < 0.05 [10]

Table 2: Pearson correlation coefficient between fine-conditioning and measured prominence biases for three emotions

	Angry		Neutral		Sad	
Prominence	PCC	p-value	PCC	p-value	PCC	p-value
NOUN	0.88	8.01e-3	0.95	1.35e-4	0.98	3.44e-4
VERB	0.96	4.81e-4	0.98	4.49e-3	0.96	4.53e-4
ADJ	0.95	8.45e-4	0.95	4.39e-1	0.94	1.32e-3
ADV	0.98	6.89e-5	0.97	4.02e-2	0.96	3.95e-4

We also evaluated the quality of speech which was synthesized by fine-conditioning on prosodic factors biases for angry, neutral, and sad emotions. In detail, we synthesized speech samples for evaluation by conditioning on angry, neutral, and sad emotions and fine-conditioning on -0.3, 0, and 0.3 biases of each of six prosodic factors from 10 sentences randomly selected from the BC2013 test dataset. Finally, we collected 540 speech samples where each of 10 sentences had 54 variations (3 emotions $\times$ 6 prosodic factors $\times$ 3 biases). We conducted a mean opinion score (MOS) test on the Amazon Mechanical Turk with 50 participants, each of whom was given 54 speech sample variations of the same sentence and required to choose speech quality for each speech in five stages (1: very bad, 5: very good). The result is shown on the left side of Fig. 7. From the result, We can conclude that our model can condition on both emotion and prosodic factors without degrading audio quality (MOS = 3.5), which is comparable to the conventional method that can only condition on prosodic factors.

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

We also evaluated the linear controllability of our proposed model by fine-conditioning on prominence during the second stage (intra-emotion) of control. Similarly, we defined a linear controllability score using the PCC between the fine-conditioning biases and measured biases for prominence. This can be represented by $\mathrm{PCC}(\boldsymbol{b}_{\mathrm{prm}}, \boldsymbol{b}'_{\mathrm{prm}})$, where $\boldsymbol{b}_{\mathrm{prm}}$ is fine-conditioning biases and $\boldsymbol{b}'_{\mathrm{prm}}$ is measured biases in-

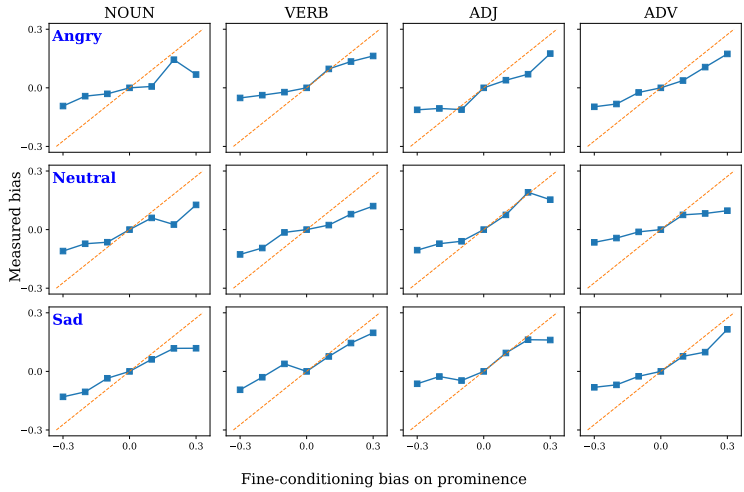

Figure 8: Correlation between fine-conditioning and observed biases when fine-conditioning on the prominence of NOUN, VERB, ADJ, and ADV words for three emotions

dicating the difference in prominence between the synthesized speech with fine-conditioning bias and without bias (fine-conditioning bias = 0), as shown in Eq. 15:

$$\begin{aligned} \boldsymbol{b}'_{\text{prm}} = &\text{PRM}(\text{emoTTS}(\hat{\boldsymbol{x}}_{\text{prm}} + \boldsymbol{b}^{(2)}_{\text{prm}})) \\ &- \text{PRM}(\text{emoTTS}(\hat{\boldsymbol{x}}_{\text{prm}})). \end{aligned} \tag{15}$$

PRM indicates the prominence measurement and $\hat{\boldsymbol{x}}_{\text{prm}}$ denotes basic-conditioning prominence, which were discussed in Section 3.1.2.

In the experiment, we found that word-level prominence was distributed differently depending on the part of speech in the training dataset, as shown in Fig. 12 in Appendix D. The prominence of NOUN, VERB, ADJ, and ADV words were distributed close to a normal distribution; however, the prominence of other parts of speech was not. Because non-normally distributed parts of speech theoretically cannot be controlled linearly, 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.

We calculated PCC($\boldsymbol{b}_{\text{prm}}, \boldsymbol{b}'_{\text{prm}}$) based on the basis of both neutral and angry emotions, as shown in Table 2, and visualized the correlation between $\boldsymbol{b}_{\text{prm}}$ and $\boldsymbol{b}'_{\text{prm}}$ on the angry, neutral, and sad emotions, as shown in Fig. 8.

From the results, we can conclude that:

1. Our model can linearly fine-condition on the prominence of certain parts of speech including NOUN, VERB, ADJ, and ADV whose prominence is distributed close to the normal distribution in the training dataset. With p-value < 0.05, the average PCC 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, and ADV words.

2. For the parts of speech aside from NOUN, VERB, ADJ, and ADV words, they may also be linearly fine-controlled if the training dataset is extended with a new one whose prominence is close to the normal distribution.

We also visualized the prominence contours of utterances, synthesized by conditioning on the angry, neutral, and sad emotions and fine-conditioning prominence on the NOUN, VERB, ADJ, and ADV words with three biases (-0.3, 0, and 0.3) from the same sentence. The sentence we chose should contain NOUN, VERB, ADJ, and ADV words at the same time. As shown in Fig. 9, the prominence of fine-conditioned words increased (or decreased) when the conditioning bias increased (or decreased). For example, by fine-conditioning on the NOUN "hotel", the prominence increased from 1.0 to 1.4 (when bias is 0.3) and decreased from 1.0 to 0.6 (when bias is -0.3) for the angry emotion. On the other side, the prominence of words that were not fine-conditioned (e.g., "They") was also slightly changed in the experiment. Such a phenomenon occurred because we enabled the dropout of pre-net even in the inference stage for better audio quality which brings a slight variation to mel-spectrogram, as described in Section 4.2. Nevertheless,

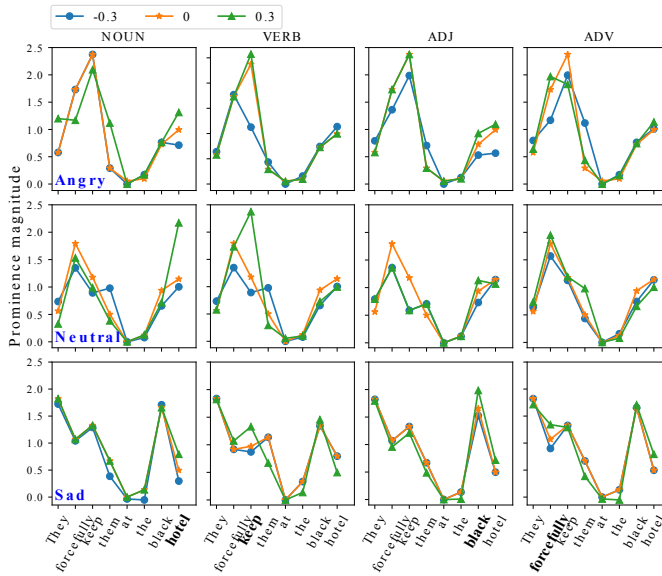

Figure 9: Prominence contours of an utterance synthesized from a given sample sentence by conditioning on the angry, neutral, and sad emotions (first stage of control) and fine-conditioning prominence on NOUN, VERB, ADJ, and ADV words with three biases (−0.3, 0, and 0.3) for each emotion (second stage of control). 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.

such slight variation did not affect the prominence controllability of our model.

To further investigate how the energy and pitch contours changed when conditioning on prominence, we also drew the energy and pitch contours when conditioning on different parts of speech for the angry emotion, as shown in Fig. 10. We can conclude that the energy and pitch of the fine-conditioned word increased (or decreased) simultaneously when the bias was increased to 0.3 (or decreased to −0.3). In particular, the pitch was more significantly affected than the energy. We interpret such phenomenon as follows: changing the pitch can achieve the desired prominence shift with minimal mel-spectrogram changes compared to changing energy or duration.

We also evaluated the quality of speech which is synthesized by fine-conditioning on the prominence of NOUN, VERB, ADJ, and ADV

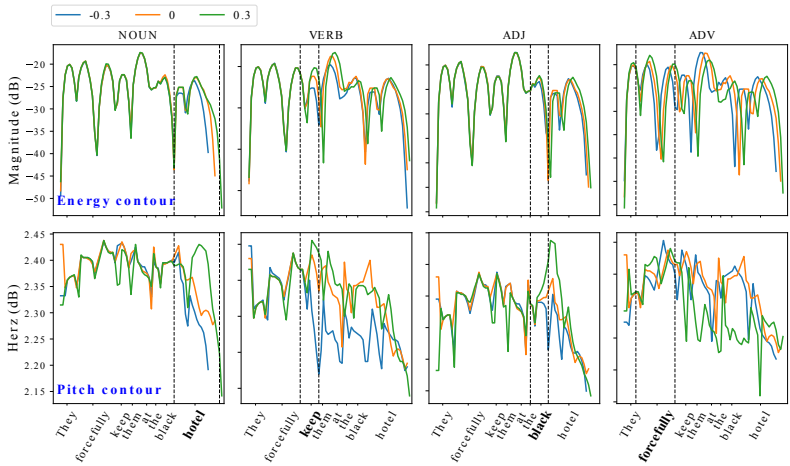

Figure 10: Energy and pitch contours of a synthesized utterance when fine-conditioning prominence on NOUN, VERB, ADJ, and ADV words with three biases (−0.3, 0, and 0.3) for angry emotion. The interval (black dashed line) of the fine-conditioned word was computed by the average intervals of corresponding words fine-conditioned with three biases.

words for angry, neutral, and sad emotions. In detail, we synthesized speech samples for evaluation by conditioning on angry, neutral, and sad emotions and fine-conditioning on −0.3, 0, and 0.3 biases of the prominence of NOUN, VERB, ADJ, and ADV words from 10 sentences randomly selected from the BC2013 test dataset. Finally, we collected 360 speech samples where each of the 10 sentences had 36 variations (3 emotions × 4 parts of speech × 3 biases). We conducted a mean opinion score (MOS) 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 is shown on the right side of Fig. 7. From this result, we can conclude that our model can fine-condition the prominence of NOUN, VERB, ADJ, and ADV words for these three emotions without degrading audio quality (MOS = 3.9), which is comparable to the method that can only condition prominence. The sample audio is accessible here[5].

[5]Sample audio:link

6 Conclusion and Discussion

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 including both the utterance-level prosodic factors (e.g., energy mean) and 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 the proposed model on the utterance-level prosodic factors for angry, neutral, and sad emotions, respectively, 3) linearly fine-condition on the prominence of NOUN, VERB, ADJ, and ADV words for the angry, neutral and sad emotions, and finally 4) synthesize speech with audio quality (MOS = 3.5) when fine-conditioning on prosodic factors and MOS = 3.9 when fine-conditioning on prominence) comparable to that of the conventional methods. Although the emotion-distinguishable score was slightly lower (= 51%) due to the usage of the narrative-style BC2013 dataset, the results are still promising. The emotion-distinguishability score may be further improved by using more emotional speech datasets.

In addition to the emotion-distinguishable score, other areas to be improved include emotion strength controllability and the linear controllability of word-level prominence, especially on the parts of speech other than NOUN, VERB, ADJ, and ADV. More generally, we can imagine that the two-stage control approach can be utilized in other domains, such as controllable image synthesis.

Acknowledgements: This work was supported by JST, Moonshot R&D Grant Number JPMJMS2011 (for experiments), and JSPS KAKENHI 21H05054, 19H01116, 21H04900 (for basic technique).

A Appendix A: Preference test for filtering out emotion categories

We conducted a simple listening preference test for selecting the speech whose annotated emotion is consistent with the estimated one. This listening preference test required three evaluators to annotate 100 ran-

Table 3: Performance of conventional and proposed SER models on the evaluation part of IEMOCAP and the preprocessed BC2013 dataset (only the annotated part) with three emotions (angry, neutral, sad). The conventional model utilized Text, Text and PSD (prosodic factors), while the proposed SER model utilized Text, PSD, and PRM (prominence) as input.

Dataset	Input	Precision	Recall	F1
IEMOCAP	Text	0.551	0.562	0.554
	Text + PSD [29]	0.621	0.618	0.619
	Text+PSD+PRM	**0.642**	**0.623**	**0.632**
BC2013	Text	0.535	0.480	0.486
	Text + PSD [29]	0.552	0.518	0.523
	Text+PSD+PRM	**0.562**	**0.536**	**0.543**

domly selected utterances from each emotion category, for a total of 400 utterances (100 utterances × four emotions). An emotion annotated more than two times is treated as the ground truth emotion of the speech. Given the ground truth and estimation, we calculated the estimation accuracy and filtered out the emotion categories of which the accuracy is lower than 60%.

B Appendix B: Performance of SER and PFG models

B.1 SER performance

We conducted an ablation study on the effectiveness of different features in predicting three emotions (angry, neutral, and sad) by training the SER models with only text, text with prosodic factors [29], and multi-modal features of text, prosodic factors, and prominence, respectively. The SER and PFG models were jointly trained on the training part (80%) of the IEMOCAP dataset. To evaluate the SER performance on both the IEMOCAP and BC2013 datasets, we evaluated it on the testing part (20%) of the IEMOCAP and the emotion-labeled part of the preprocessed BC2013 datasets (described in Appendix A) on precision, recall, and F1-score. As a result, the F1 scores of our SER model trained with multi-modal features of text, prosodic factors, and prominence input were improved by 7.8% and 1.3% on the IEMOCAP,

Table 4: Performance of conventional and proposed PFG models (prosodic factor generator and prominence generator) on the preprocessed BC2013 dataset. The conventional model utilized text, while the proposed PFG model utilized text and emotion soft labels as input. The L2 loss of utterance-level prosodic factors and word-level prominence were calculated, respectively.

Model	Input	L2 loss
Prosodic factor generator	Text [26]	0.062
	Text+EmoSoftLabel	**0.054**
Prominence generator	Text [34]	0.023
	Text+EmoSoftLabel	**0.018**

5.7% and 2.0% on BC2013 datasets when compared with the other two benchmark models [29], respectively. The results are shown in Table 3.

B.2 PFG performance

We fine-tuned the PFG model with the SER model frozen when training the proposed emotion-controllable TTS model on the training part of the BC2013 dataset and separately evaluated the prosodic factor generator and prominence generator of our PFG model on the testing part (20%) of the BC2013 dataset using the L2 loss. We compared our model with the PFG model without emotion soft label as input, followed by previous methods [26, 34]. The results are shown in Table 4.

C Appendix C: Prosodic factor distribution of angry, neutral, and sad speech

in the training part of the BC2013 dataset We visualized the distribution of six prosodic factors, including the mean, SD, and the range of energy and pitch contours, for angry, neutral, and sad speech in the training part of the BC2013 dataset. The results are shown in Fig. 11. The results demonstrated that the prominence of NOUN, VERB, ADJ, and ADV words was distributed close to normal distribution, while the prominence of ADP, AUX, PRON, PROPN, and INTJ was not.

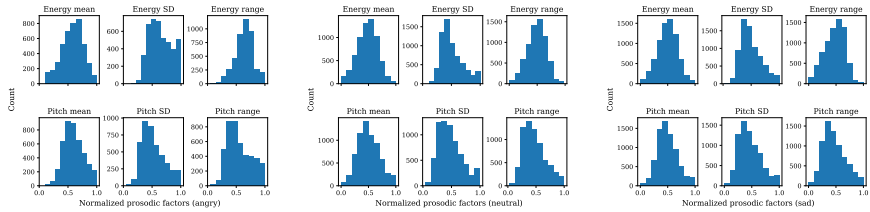

Figure 11: Prosodic factors distribution on six prosodic factors for three emotions (left: angry, center: neutral, right: sad)

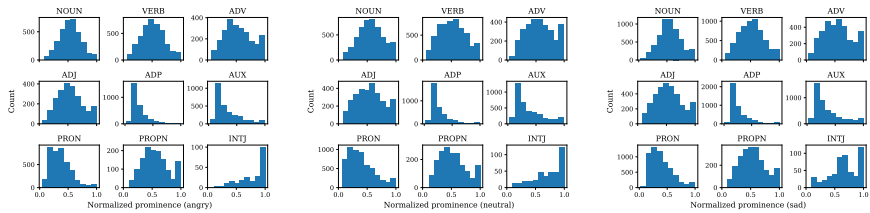

Figure 12: Prominence distribution on different parts of speech for three emotions (left: angry, center: neutral, right: sad)

D Appendix D: Prominence distribution of angry, neutral, and sad speech in the training part of the BC2013 dataset

We visualized the distribution of word-level prominence on NOUN, VERB, ADJ, and ADV for angry, neutral, and sad speech in the training part of the BC2013 dataset. The results are shown in Fig. 12.

References

[1] M. B. Akçay and K. Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers". *Speech Communication*. 116: 56–76.

[2] J. P. Arias, C. Busso, and N. B. Yoma. "Shape-based modeling of the fundamental frequency contour for emotion detection in speech". *Computer Speech & Language*. 28(1): 278–294.

[3] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate". *arXiv preprint arXiv:1409.0473*.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching word vectors with subword information". *Transactions of the association for computational linguistics*. 5: 135–146.

[5] 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*. 42(4): 335–359.

[6] 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*.

[7] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. "Attention-based models for speech recognition". *Advances in neural information processing systems*. 28.

[8] 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. 210–221.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*.

[10] J. D. Evans. *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

[11] K. Ito. "The lj speech dataset". URL: https://keithito.com/LJ-Speech-Dataset/.

[12] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim. "Diff-tts: A denoising diffusion model for text-to-speech". *arXiv preprint arXiv:2104.01409*.

[13] S. Jing, X. Mao, and L. Chen. "Prominence features: Effective emotional features for speech emotion recognition". *Digital Signal Processing*. 72: 216–231.

[14] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. "Speech emotion recognition using deep learning techniques: A review". *IEEE Access*. 7: 117327–117345.

[15] S. King and V. Karaiskos. "The Blizzard Challenge 2013". In: *Blizzard challenge workshop*.

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

[17] 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*. 26(9): 1383–1387.

[18] 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.

[19] 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. 423–430.

[20] H. Li, Y. Kang, and Z. Wang. "EMPHASIS: An emotional phoneme-based acoustic model for speech synthesis system". *arXiv preprint arXiv:1806.09276*.

[21] T. Li, S. Yang, L. Xue, and L. Xie. "Controllable emotion transfer for end-to-end speech synthesis". In: *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE. 1–5.

[22] X. Luo, S. Takamichi, T. Koriyama, Y. Saito, and H. Saruwatari. "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. 794–799.

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

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. "Montreal forced aligner: Trainable text-speech alignment using kaldi." In: *Interspeech*. Vol. 2017. 498–502.

[25] W. Nakata, T. Koriyama, S. Takamichi, Y. Saito, Y. Ijima, R. Masumura, and H. Saruwatari. "Predicting VQVAE-based Character Acting Style from Quotation-Annotated Text for Audiobook Speech Synthesis". In: *Proc. Interspeech*. 4551–4555.

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

[27] 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*.

[28] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. "Fastspeech: Fast, robust and controllable text to speech". *arXiv preprint arXiv:1905.09263*.

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

[30] S. Shechtman and A. Sorin. "Sequence to sequence neural speech synthesis with prosody modification capabilities". *arXiv preprint arXiv:1909.10302*.

[31] 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. 4779–4783.

[32] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous. "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron". In: *international conference on machine learning*. PMLR. 4693–4702.

[33] A. Suni, J. Šimko, D. Aalto, and M. Vainio. "Hierarchical representation and estimation of prosody using continuous wavelet transform". *Computer Speech & Language*. 45: 123–136.

[34] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio. "Predicting prosodic prominence from text with pre-trained contextualized word representations". *arXiv preprint arXiv:1908.02262*.

[35] 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. 7254–7258.

[36] M. Vainio, A. Suni, and D. Aalto. "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. 173–188.

[37] R. Valle, K. Shih, R. Prenger, and B. Catanzaro. "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis". *arXiv preprint arXiv:2005.05957*.

[38] A. Vioni, M. Christidou, N. Ellinas, G. Vamvoukakis, P. Kakoulidis, T. Kim, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis. "Prosodic clustering for phoneme-level prosody control in end-to-end speech synthesis". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 5719–5723.

[39] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.* "Tacotron: Towards end-to-end speech synthesis". *arXiv preprint arXiv:1703.10135*.

[40] 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. 5180–5189.

[41] 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. 6199–6203.

[42] 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. 6945–6949.

[43] J. Zhao, X. Mao, and L. Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks". *Biomedical Signal Processing and Control*. 47: 312–323.

[44] 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. 192–199.