Artificial Intelligence 329 (2024) 104076

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Emotion selectable end-to-end text-based speech editing

Tao Wang [a,b], Jiangyan Yi [a,*], Ruibo Fu [a,*], Jianhua Tao [c,*], Zhengqi Wen [b], Chu Yuan Zhang [a,b]

[a] Institute of Automatic Chinese Academy of Sciences, Beijing, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[c] Department of Automation, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Emotion selectable
Text-based speech editing
Emotion decoupling
Mask prediction
Few-shot learning
Text-to-speech

ABSTRACT

Text-based speech editing is a convenient way for users to edit speech by intuitively cutting, copying, and pasting text. Previous work introduced CampNet, a context-aware mask prediction network that significantly improved the quality of edited speech. However, this paper proposes a new task: adding emotional effects to the edited speech during text-based speech editing to enhance the expressiveness and controllability of the edited speech. To achieve this, we introduce Emo-CampNet, which allows users to select emotional attributes for the generated speech and has the ability to edit the speech of unseen speakers. Firstly, the proposed end-to-end model controls the generated speech's emotion by introducing additional emotion attributes based on the context-aware mask prediction network. Secondly, to prevent emotional interference from the original speech, a neutral content generator is proposed to remove the emotional components, which is optimized using the generative adversarial framework. Thirdly, two data augmentation methods are proposed to enrich the emotional and pronunciation information in the training set. Experimental results[1] show that Emo-CampNet effectively controls the generated speech's emotion and can edit the speech of unseen speakers. Ablation experiments further validate the effectiveness of emotional selectivity and data augmentation methods.

1. Introduction

With the rapid development of the internet, various media platforms have emerged, enabling us to learn, entertain, and communicate. Speech plays a critical role in many of these media formats. Text-based speech editing, which involves modifying speech by directly editing the transcript, has the potential to greatly streamline the audio generation process [11,26,35,42,2]. This technique allows content creators to quickly edit transcripts using familiar word processing operations such as cut, copy, and paste, and automatically propagate changes to the corresponding audio recording without manually editing the original waveform.

Several studies have focused on improving the naturalness of edited speech in text-based speech editing. One such approach is the VoCo pipeline system [11], which leverages a speech synthesis system and a voice conversion system to produce edited speech that sounds more natural. Although the full name of VoCo was not explicitly mentioned in the original paper, it represents an innovative approach to text-based manipulation of speech content. Another method, called context-aware prosody correction [26], modifies the

* Corresponding authors.
E-mail address: jhtao@tsinghua.edu.cn (J. Tao).

[1] Examples of generated speech can be found at https://hairuo55.github.io/Emo-CampNet.

https://doi.org/10.1016/j.artint.2024.104076
Received 10 March 2023; Received in revised form 2 December 2023; Accepted 5 January 2024

Available online 23 January 2024

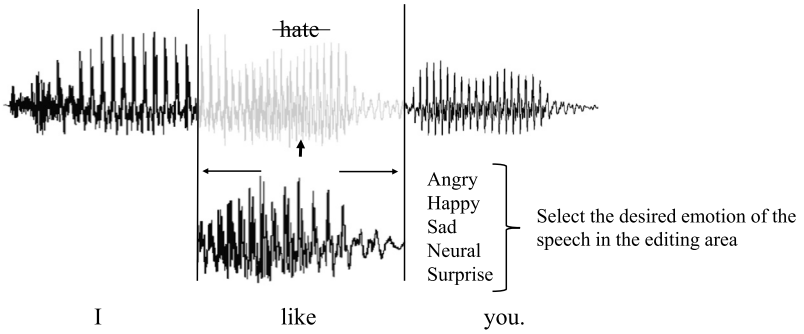


Fig. 1. The replacement operation of emotion selectable text-based speech editing. It involves replacing words in the text and selecting the corresponding emotional attributes to achieve the desired emotional effect. The model can then automatically calculate the speech required for the corresponding editing area.

prosodic information of the target segment to further enhance the overall prosody of the edited speech. To address the complexities and potential error accumulation associated with pipeline systems, EditSpeech [35,36] utilizes partial inference and bidirectional fusion mechanisms. Another framework, Alignment-Aware Acoustic-Text Pretraining [2], reconstructs masked acoustic signals with text input and acoustic-text alignment during training. Similar work includes Speechpainter [3], RetrieverTTS [47]. In our previous work, we proposed a context-aware mask and prediction network (CampNet) [42] that can simulate the text-based speech editing process and be trained end-to-end without relying on duration information.

The existing research on text-based speech editing has made significant strides in improving the naturalness of edited speech. However, there are still limitations, such as monotonous rhythms and a lack of emotional expressiveness. To address this gap, adding emotional effects to the generated speech has become increasingly important [29,7]. As the quality of generated speech continues to improve, there is a growing demand for speech styles that convey emotion [22,20,44,19,38,51].

In light of these developments, we propose a new task for text-based speech editing, which we refer to as "Emotional Selectable Text-based Speech Editing", as shown in Fig. 1. This approach aims to enable users to select the desired emotional style for edited speech, thereby enhancing the expressiveness and naturalness of the generated audio. Firstly, we emphasize the need for enhanced expressiveness and controllability in text-based speech editing. While existing methods, such as the VOCO pipeline, have shown effectiveness in transforming speech based on textual inputs, they often lack the ability to incorporate emotional effects into the edited speech. This limitation motivates the development of our proposed innovation module that specifically focuses on adding emotional effects during text-based speech editing.

Furthermore, our innovation model offers distinct advantages in terms of decoupling and controlling emotions independently from the textual content. By decoupling emotions from the text, users have the flexibility to modify emotional characteristics, such as happiness, sadness, excitement, and more. This decoupling not only enhances the expressiveness of the generated audio but also provides greater granularity in tailoring the emotional aspects of the speech to specific contexts, target audiences, or desired outcomes. By enabling users to precisely manipulate emotional effects, our module empowers them to create more expressive and emotionally engaging speech content. One significant challenge lies in ensuring the effective decoupling of emotional information from speech.

To address this new task "Emotional Selectable Text-based Speech Editing", we build upon the context-aware mask and prediction network proposed in [42] and make several enhancements. Firstly, we propose an end-to-end emotion selectable text-based speech editing network that enables users to specify the desired emotional attribute for the generated speech. Then, to address the challenge of preserving the desired emotion while also removing the original emotional content of the edited speech, we introduce a training framework based on generative adversarial methods that extracts emotion-independent content information from the original speech. Additionally, we propose two data augmentation methods to overcome the limitations of small-scale emotional speech datasets and the poor performance of few-shot learning. Together, these innovations enable us to achieve more expressive, controllable and natural-sounding edited speech with emotional attributes.

Overall, the main contributions of this paper are:

- This paper proposes an emotion-selectable text-based speech editing task and designs an end-to-end model for it, which is called Emo-CampNet. The model utilizes decoupling and reconstruction methods of emotions and can effectively control the emotional attributes of the edited speech region.
- To ensure that the generated speech's emotion is controlled solely by the input emotion attributes and not affected by the emotional components in the original speech, we propose a neutral content generator that is optimized using the generative adversarial network. Experimental results demonstrate that this method effectively removes emotional components from the original speech.
- We propose two data augmentation methods to enrich the emotional information and pronunciation information of the training data. Experimental results show these methods can effectively improve the model's performance and enable it to edit unseen speakers' speech.

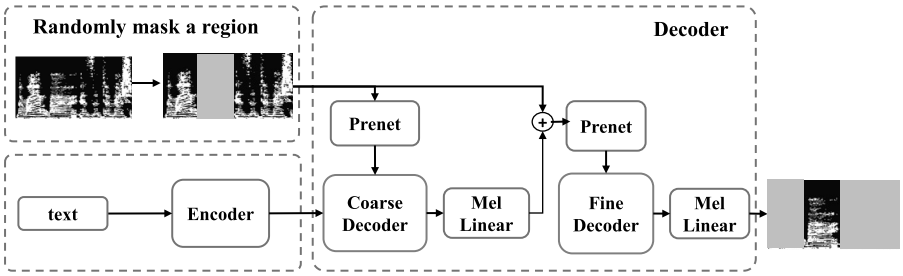


Fig. 2. The structure of context-aware mask prediction network (CampNet). To simulate text-based speech editing, the key idea of the model is randomly masking a region of speech and then predicting the masked speech according to the text and speech context.

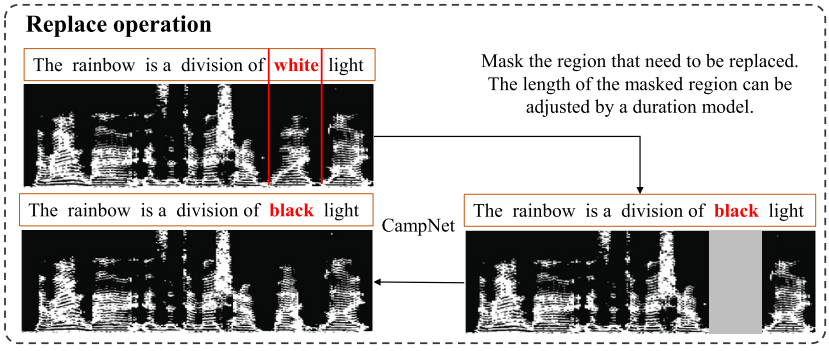


Fig. 3. Replacement operation of text-based speech editing based on CampNet. The operation can be divided into two steps: first, masking the region to be edited and then predicting new speech according to the modified text and speech context. Deletion and insertion operations can also be divided into similar masking and prediction processes.

This paper is structured as follows. Section 2 introduces the related work. The proposed Emo-CampNet, its operations for emotion-selectable text-based speech editing, emotion decoupling with generating adversarial networks, and the data augmentation methods are described in Section 3. After explaining the experiments and results in Section 4 and Section 5, we draw a conclusion in Section 6.

2. Relate work

This paper is based on the framework of CampNet [42] and adds the emotion selectable function for the generated speech. Therefore, we will briefly introduce the CampNet, then present the work related to emotion speech generation.

2.1. Context-aware mask and prediction network (CampNet)

In general, there are three modes of operation for text-based speech editing, which include deletion, replacement, and insertion. While there are various modes of operation, they can be simplified into two primary steps, which have been thoroughly analyzed in [42]. The first step involves masking the region of the original speech that requires editing, and the second step involves combining the masked speech and the edited text to predict the masked area. Consequently, the CampNet framework is developed based on the context-aware mask and predicting concept, as shown in Fig. 2. It consists of two processing stages: an encoder and a decoder. The encoder module initially processes the input sentence and converts it into a hidden representation, which then guides the decoder in predicting the acoustic features of the edited speech. At the training stage, a random acoustic feature region is masked as the ground truth to condition the decoder. The decoder is divided into two steps. The first step involves learning the alignment between the masked ground truth and the text representation using the multi-head attention mechanism [40] to predict coarse acoustic features. In the second step, the decoder predicts finer acoustic features based on the coarse acoustic features and the original speech context. This further fuses the contextual information of speech to produce more natural-sounding predicted speech.

In the test stage, we can use CampNet to complete various operations, such as deletion, operation, and replacement. Here we take the replacement operation as an example, as shown in Fig. 3. The first step is to define the word boundary to be replaced, mask it according to the word boundary and then modify the text. The second step is to input the masked speech and the modified text into CampNet. The model will predict the replaced speech according to the modified text. If there is a big difference between the length of the replaced speech and the original speech, such as adding or deleting some words, a pre-trained duration model [45] can be used to predict the length of the replaced region.

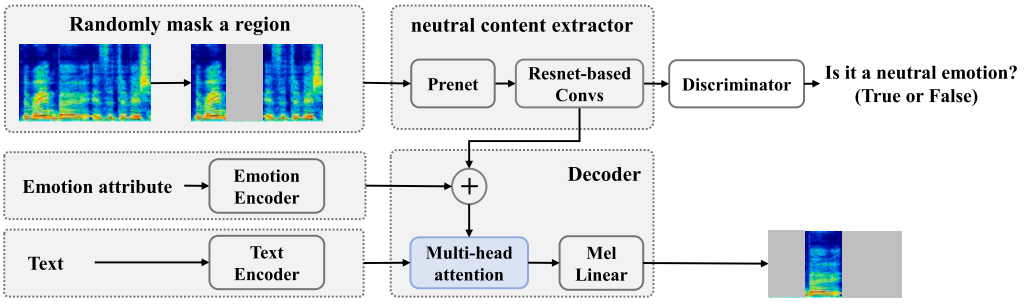


Fig. 4. Structure of emotion selectable context aware mask prediction network (Emo-CampNet). Based on the framework of context-aware masking and prediction, emotion attributes are used as input to control the emotion of generated speech. In addition, generative adversarial network is introduced to remove the emotional components from the input speech to prevent the interference to the generated speech.

2.2. Emotion speech generation

Attempts to add emotional effects to generated speech have existed for more than a decade. Two tasks have been widely studied, emotional text-to-speech [19,4] and emotional voice conversion [51]. For the former task, researchers focused on obtaining expressive emotion representations to guide the system to synthesize expressive speech. Global style tokens (GST) [44] have been proposed to learn the style information in speech, including emotional attributes automatically. In addition, to more finely control the emotion representation of each phoneme, emotional text-to-speech methods based on the phoneme level are also proposed [21,10,5]. Some other studies based on variational autoencoders (VAE) [17] show the effectiveness of controlling the speech style by learning, scaling, or combining disentangled representations [48,14]. For emotional voice conversion, the study focuses on preserving the speech content and converting emotional information. The early studies on emotion voice conversion include GMM and HMM [1,13,51]. The recent studies on deep learning have seen a remarkable performance, such as DNN, sequence-to-sequence models [22–24]. With the development of generative adversarial networks (GAN), CycleGAN [46,49,34] and StarGAN [33] have been proposed to disentangle the emotional elements from speech.

In addition to the two aforementioned tasks, this paper proposes a novel method to incorporate emotional effects into synthesized speech through emotion-selectable text-based speech editing, which is discussed in the following section.

3. Emotion selectable context-aware mask prediction network

The key distinction between emotion selectable text-based speech editing and the original text-based speech editing is that the former incorporates a selectable emotion into the synthesized speech based on the latter. One significant challenge lies in ensuring the effective decoupling of emotional information from speech. The effectiveness of this part is a bottleneck of the model. If the model can't sufficiently remove emotional information, the model's ability to control emotions may significantly degrade. To achieve this functionality, we must first introduce additional emotional attribute information to govern the emotional aspects of the generated speech. Next, we need to disentangle the emotional component from the edited speech to enable better control of the generated speech by the input emotional attributes. Moreover, to enhance the model's ability to edit any speaker's speech in a single-shot manner, we propose two data augmentation techniques to augment the emotional and pronunciation information of the dataset.

In this section, we first introduce the framework of Emo-CampNet. Secondly, we introduce the generative adversarial training method for emotion decomposition and control. Thirdly, we introduce the two proposed data argumentation methods.

3.1. Emo-CampNet and its operations

To make the emotion of the synthesized speech controlled only by the selected emotion attribute and not disturbed by the emotional components in the original speech, we first need to remove the emotion components in the original speech. To remove emotional components from speech, we assume that all emotional speech can be converted to and constructed from neutral emotional speech. Based on the assumption, the emotion selectable speech editing system includes four modules: text encoder, emotion encoder, neutral content generator (NCG), and decoder, as shown in Fig. 4.

Firstly, to realize the basic function of text-based speech editing, we follow the idea of CampNet, i.e., mask part of the speech and then predict the masked area based on the input text and the remaining portions of the speech. However, in contrast to directly feeding the masked speech into the decoder, a neutral content generator is used to extract the emotion-independent content information to prevent the interference of the emotional information in the original speech to the decoder. Furthermore, to ensure that the extracted content information is truly independent of emotional information, we use the generative adversarial network to supervise the content information, which will be described in detail in the next section. In the decoding stage, the decoder module predicts the masked area according to the masked neutral content information, text, and emotion attribute. We will describe this process in detail below.

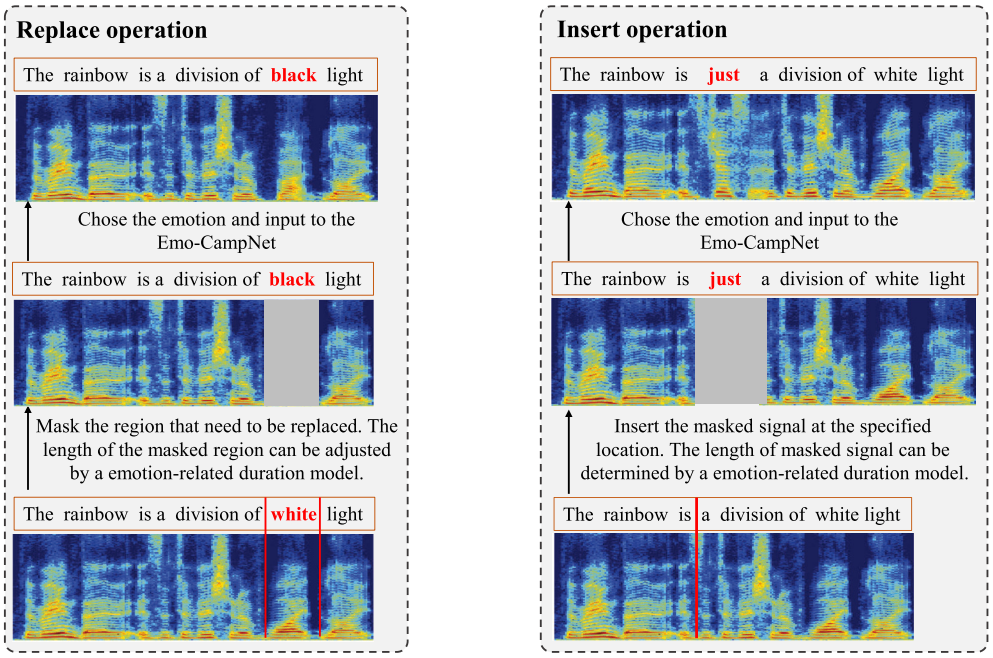


Fig. 5. At the inference stage, the replace and insert operations of emotion selectable text-based speech editing based on Emo-CampNet.

First, given a speech y, its text information $x = (x_1, \ldots, x_2, \ldots, x_M)$ and emotional attribute *emo* — a one-hot vector corresponding to different emotions (neutral, happy, sad, angry, surprise), M represents the length of the text sequence. A text encoder processes the input sentence x and converts it into a hidden representation h_x in the following way:

$$h_x = \left(h_{x_1}, h_{x_2}, \ldots, h_{x_M}\right) = \text{encoder}_{\theta_x}(x) \tag{1}$$

where θ_x denotes the parameters of text encoder network.

At the same time, the emotional attribute *emo* is mapped to a learnable embedding feature, so as to transform the emotional information into the same dimension as the hidden representation of the text h_x, this learned embedding can be expressed as:

$$h_{emo} = \text{encoder}_{\theta_{emo}}(emo) \tag{2}$$

To simulate the text-based speech editing process, we randomly mask a portion of the speech and then use the network to predict the masked region. Furthermore, to make the generated speech controlled only by the input emotional attributes *emo*, we need to remove the emotional information from the masked speech. Suppose that the speech after randomly masking a region from y is denoted as y_{mask}, the NCG module is used to extract the emotion-independent content information, which can be expressed as:

$$h_c = \text{NCG}_{\theta_c}(y_{mask}) \tag{3}$$

where θ_c denotes the parameters of NCG. The NCG module has two functions. One is to project y_{mask} into the same dimension as the hidden representation of the text h_x and emotion information h_{emo}. Second, to remove the emotion components in y_{mask}.

Finally, combined with the hidden features of text h_x, emotion h_{emo} and the masked neutral content information h_c, the masked area of speech is predicted by the decoder, which can be expressed as:

$$y_{pre} = \text{decoder}_{\theta_d}(h_x, h_{emo}, h_c) \tag{4}$$

where θ_d denotes the parameters of decoder network.

With a pre-trained Emo-CampNet model, some operations of speech editing, such as deletion, insertion, and replacement, can be carried out. These operations are similar to that described in CampNet. Since the deletion operation cannot add any emotional effect, the process is the same as CampNet. Here, we will briefly describe the replacement and insertion operations, as shown in Fig. 5. These operations differ from their counterparts in CampNet is that when we input masked speech and text information into Emo-CampNet, we can additionally select the desired emotional attribute *emo* of the generated speech, which includes neutral, happy, sad, angry, and surprise. When the Emo-CampNet predicts the speech in the mask area, the speech in this area has the emotional attribute *emo*.

In the decoding stage, ensuring that the extracted content information is independent of emotion is the key to Emo-CampNet. We propose a generative adversarial training method, which will be introduced in the following section.

It should be noted that the emotion modification is focused on the targeted segments and does not impact the emotion in the non-edited portions of the speech. This approach allows for more precise emotion control in specific parts of the speech while

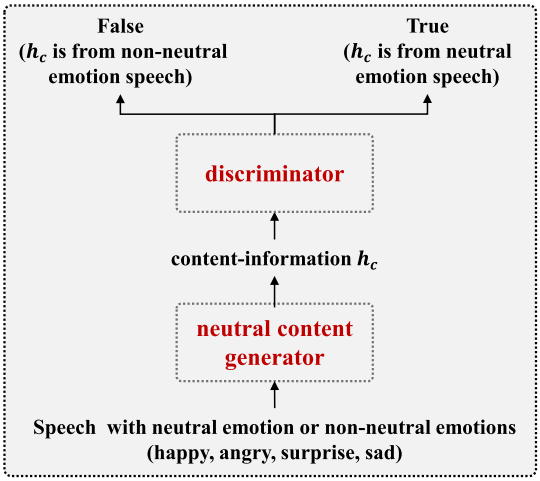


Fig. 6. The generative adversarial network is proposed to remove the emotional components in the input speech, which can prevent interference to the emotional information of the generated speech. All the speeches are divided into two categories: neutral emotion speech and non-neutral emotion speech. The discriminator is used to judge whether the content information extracted by the generator comes from neutral speech or non-neutral speech. The generator's task is to confuse the discriminator's judgment.

maintaining the original emotion in the rest of the utterance. While this targeted emotion editing may introduce some challenges in terms of similarity and continuity, it offers flexibility in various applications. For example, if we have a neutral audio segment and want to emphasize certain words with different emotions, such as transforming them into sad or happy emotions, this method can be employed. Similarly, if a speech segment has predominantly happy emotion but contains some relatively neutral parts, this approach can be used to modify those specific areas.

3.2. Generative adversarial network for emotion decomposition and control

To extract emotion-independent content information, we hypothesize that emotional speech can be converted to and constructed from neutral emotional speech. Based on this assumption, we propose a generative adversarial training framework [8] to remove the emotional components in speech, as illustrated in Fig. 6. Firstly, we utilize a neutral content generator (NCG) to extract the content information h_c from the masked speech y_{mask}. To ensure that the content information does not contain emotional components, we introduce a discriminator D. D is responsible for distinguishing whether h_c represents the content information extracted from neutral emotional speech or non-neutral emotional speech. Specifically, we mark the discriminator target of h_c extracted from neutral emotional speech as $True$, and from non-neutral emotional speech (such as happy, angry, surprise, and sad) as $False$.

Our overall objective contains two components: adversarial losses for transforming emotional speech into neutral content information; and reconstruction loss enables the model to predict the information of the masked area.

3.2.1. Adversarial loss

We apply adversarial losses to remove the emotional component and retain the content information in speech. For the mapping function $NCG : y \rightarrow h_c$ and its discriminator D, we express the objective as:

$$\begin{aligned}\mathcal{L}_{\text{adv}}(NCG, D, Y) = &\mathbb{E}_{y\sim p_{\text{neutral}}(y)}\left[\log D(NCG(y))\right] \\ &+ \mathbb{E}_{y\sim p_{\text{non-neutral}}(y)}\left[\log(1 - D(NCG(y))\right]\end{aligned} \tag{5}$$

where $y \sim p_{\text{neutral}}(y)$ denotes y is neutral emotion speech and $y \sim p_{\text{non-neutral}}(y)$ denotes y is non-neutral emotion speech. NCG tries to generate h_c that is independent of emotion information, while D aims to discern whether h_c contains non-neutral emotional components. When h_c comes from non-neutral emotion speech, the target value of the discriminator is True (denoted with 1). When h_c comes from neutral emotion speech, the target value of the discriminator is False (denoted with 0). NCG aims to trying to minimize this objective, while the adversarial discriminator D tries to maximize it, i.e., $\min_{NCG}\max_{D}\mathcal{L}_{\text{adv}}(NCG, D, Y)$.

3.2.2. Reconstruction loss

By minimizing the adversarial loss, the NCG can generate neutral content information. However, the final goal of the Emo-CampNet is to predict the masked region of acoustic features. Therefore, a reconstruction loss is used in the prediction of the masked region, which is defined as

$$\mathcal{L}_{\text{rec}}\left(y_{pre}, y - y_{mask}\right) = MSE(y_{pre}, y - y_{mask}) \tag{6}$$

Where $y - y_{mask}$ represents the ground truth of speech's masked region, and y_{pre} is the speech predicted by the decoder in Eq. (4). MSE stands for mean square error.

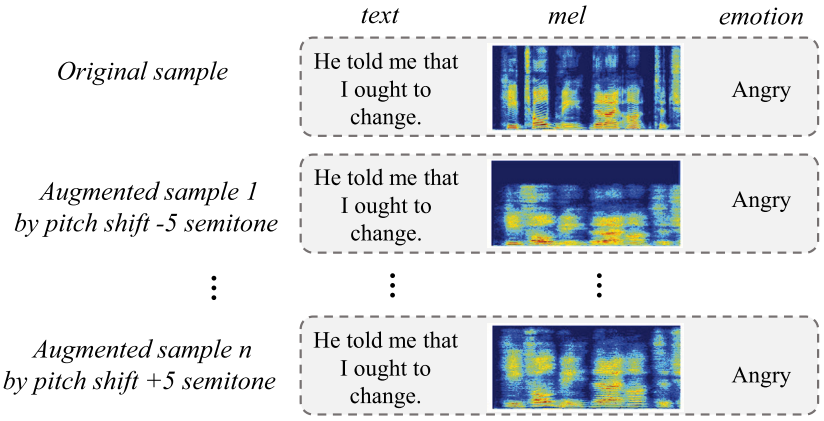


Fig. 7. An example of emotion data augmentation. By shifting the F0 of speech, more speech with the same emotion as the original speech can be obtained.

3.2.3. Full objective

Finally, the objective functions to optimize emo-campnet and discriminator are written, respectively, as

$$\begin{aligned} \mathcal{L}_{discriminator} &= -\mathcal{L}_{adv} \\ \mathcal{L}_{emo-campnet} &= \lambda_{adv}\mathcal{L}_{adv} + \mathcal{L}_{rec} \end{aligned} \tag{7}$$

Where λ_{adv} is a hyperparameter that controls the relative importance of the adversarial loss. In our experiments, we carefully examined hyperparameter λ_{adv}. We acknowledge that this parameter plays a crucial role in achieving a balance between preserving emotional attributes and maintaining overall speech quality. Higher values of λ_{adv} tend to prioritize the preservation of emotional attributes, potentially resulting in overemphasis and distortion of the desired emotion. Conversely, lower values of λ_{adv} may lead to insufficient emphasis on emotional content, resulting in edited speech that lacks expressiveness. Through comparisons, we found that setting $\lambda_{adv} = 0.5$ consistently yielded the best results in terms of both emotional expressiveness and speech fidelity. This value strikes a favorable trade-off, allowing the model to effectively capture and reproduce emotional attributes while ensuring the overall naturalness and quality of the generated speech.

3.3. Data augmentation for one shot learning

One of the critical challenges that impede the progress of emotional speech generation research is the scarcity of emotional speech data. A limited amount of emotional speech data can restrict the emotional, speaker, and pronunciation information available for model training, leading to overfitting [31] and reduced expressiveness when editing speech from new speakers. Therefore, to enhance the Emo-CampNet model's ability to edit speech from unseen speakers, it is crucial to augment the training data with additional emotional, speaker, and pronunciation information. This section presents two data augmentation methods that can effectively prevent overfitting and enhance the model's expressiveness.

3.3.1. Emotional data augmentation

We employ the fundamental frequency (F0) perturbation method to augment emotional speech data. Relative changes in pitch significantly impact emotional information. Through our experiments, we observed that synchronously shifting all pitch values in a speech segment maintains emotional attributes. The works that adopted similar approaches include references [37,28,12]. This involves perturbing the F0 of emotional speech to alter its pitch while retaining its emotional attributes. Let's consider the following example using a simplified mathematical representation. Suppose we have an input speech signal x with a corresponding F0 contour denoted as $F0\_original(t)$. We aim to apply pitch perturbation by shifting the F0 values by a certain number of steps, denoted as n_steps. The modified F0 contour, denoted as $F0\_perturbed(t)$, can be calculated as follows:

$$F0\_perturbed(t) = F0\_original(t) * 2^{(n\_steps/12)} \tag{8}$$

In this equation, the term $2^{(n\_steps/12)}$ represents the frequency ratio corresponding to the number of semitones, where a semitone corresponds to a frequency ratio of $2^{(1/12)}$. By multiplying the original F0 values by this ratio, we effectively shift the pitch of the speech signal. For instance, if n_steps is set to -4 (representing a downward shift of four semitones), the F0 values in the modified contour will be reduced by a factor of $2^{(-4/12)}$, resulting in a lower pitch for the perturbed speech signal. Similarly, positive values of n_steps will lead to an upward pitch shift. In the specific implementation, to generalize to different pitch variations while still ensuring that the perturbations remain within the bounds of natural human speech, we performed data augmentation on a given audio segment using the following range of values (n_steps): $[-5, -4, -3, -2, -1, 1, 2, 3, 4, 5]$. Each value represents the number of semitones by which the F0 of the speech signal was shifted. We utilized the $librosa$ library, which provides a function called $pitch\_shift()$ to perform pitch perturbation. Fig. 7 displays acoustic features (mel spectrogram) excerpts resulting from various amounts of F0 shifting. This

method offers the advantage of enriching F0 data in different frequency bands without altering emotional attributes, which helps expand emotional and speaker information in the training data.

It should be noted that, when we apply F0 perturbation by raising or lowering the entire F0 contour, we are introducing a global shift to the pitch of the speech segment. This global shift affects the overall pitch level but preserves the relative pitch variations that carry emotional information. In other words, the emotional nuances conveyed through the relative pitch differences remain intact despite the change in the overall pitch. Pitch-shift augmentation has been previously used in other tasks such as emotional voice conversion [37] and singing voice synthesis [25], it is important to note that in our study, employing this method ensure the model can be trained to generate semantically meaningful speech. We would encounter a limitation in the availability of open-source emotional datasets, which led to insufficient speaker information, emotional information, and content information to support effective model training. To address this challenge and enable successful model training, we explored two data augmentation methods to expand our dataset, which could provide valuable insights for researchers, enabling them to address the challenges posed by limited emotional datasets and effectively expand their data to achieve successful training.

3.3.2. Neutral data enrichment

Although emotional data can be enriched by F0 disturbance, this method only changes the voice timbre and does not expand speech content information. To make the model learn more pronunciation information, we use the dataset designed for text-to-speech to expand the training data of Emo-CampNet. This paper uses the open-source dataset VCTK as an example. Since the speech in VCTK is a recording style, we regard its emotional attribute as neutral emotion. In this way, we can introduce many neutral emotional speeches and enrich the speaker and pronunciation information.

This method has two advantages: (1) it can introduce rich pronunciation and speaker information into the training dataset, and (2) it can alleviate the imbalance between the number of neutral emotion speech and non-neutral emotion speech during discriminator training. Since the data of non-neutral emotion speech are much more abundant than that of neutral emotion speech in the emotion dataset, introducing additional neutral emotion data can better train the discriminator.

4. Experimental procedures

4.1. Dataset and task

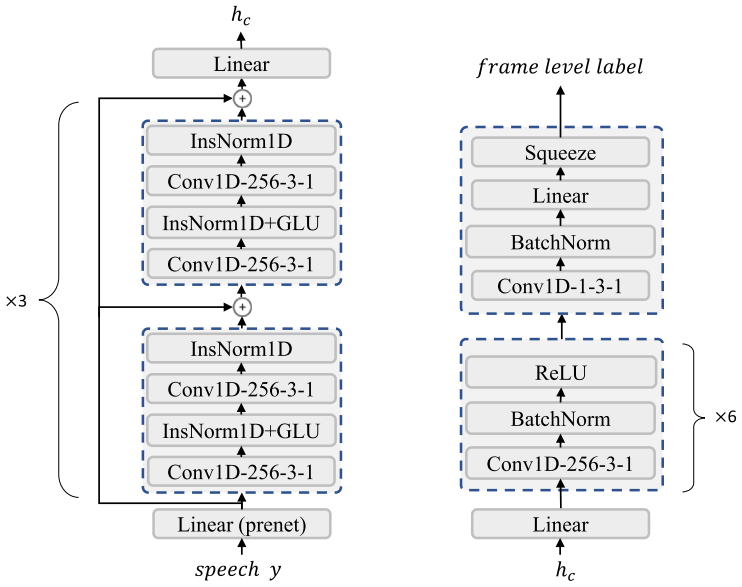
In this section, we conduct experiments on the VCTK [41] and ESD [51] corpora to evaluate our proposed method.[2] The ESD database consists of 350 parallel utterances spoken by 20 speakers and covers 5 emotion categories (neutral, happy, angry, sad, and surprise). The VCTK corpus includes speech data uttered by 110 English speakers with different accents. Each speaker reads out about 400 sentences. Specifically, we select four speakers from the VCTK dataset as the test set, and the remaining utterances are divided into 90% training set and 10% validation set. We expand the ESD dataset using the emotion data augmentation method introduced in Section 3.3. We take all the training data in VCTK as neutral emotion speech to expand the neutral speech. Finally, we take the training sets of ESD and VCTK as the total training set to train the Emo-CampNet model and take the test set in ESD data and the test set in VCTK as the total test set. All wav files are sampled at 16 KHz. It is worth noting that while introducing neutral emotional speech, we can also use large-scale neutral emotional data to pre-train the model, and then use emotional speech to fine-tune the model, which can also improve the model's performance. This method can serve as a trick for improving model performance in the future, and this paper will focus on comparing the implementation of directly trained models.

4.2. Model details

LPCNet [39] is utilized to extract 32-dimensional acoustic features, including 30-dimensional BFCCs (Bark-Frequency Cepstrum Coefficients) [9], 1-dimensional pitch, and 1-dimensional pitch correlation parameter. To calculate the BFCCs, spectrum analysis was performed with a window length of 20 ms and a frame shift of 10 ms, and the Bark-scale filter bank was applied. Pitch calculation was based on an open-loop cross-correlation search. Since this is a new task and there has been no relevant work before, we have constructed different baseline systems according to the relevant innovations of this paper. Based on these systems, we can compare the impact of various modules and settings on the performance of the model. There are 5 systems for comparison:

- **Emo-CampNet** First, we train the proposed model Emo-CampNet according to the framework of Fig. 4 with the generating adversarial training method and the data augmentation methods. The structure of the neutral content extractor is shown on in Fig. 8(a). The structure of the discriminator is shown in Fig. 8(b). It should be noted that the output of the discriminator is a frame-level label, which is helpful for the convergence of the model. The decoder and text encoder structures are based on the Transformer structure and are similar to the structure of CampNet. The emotion encoder maps emotion attributes into a learnable embedding. The phoneme sequence is input to a 3-layer CNN to learn the context information of the text. Each phoneme has a trainable embedding of 256 dims. The text encoder and the decoder contain 3 and 6 transformer blocks, respectively. The Prenet in our model is designed to transform the dimensionality of the acoustic features to match the dimensionality of the hidden layer features. It consists of a single fully connected layer in our implementation, which is responsible for reducing the dimensionality

[2] Examples of generated speech can be found at https://hairuo55.github.io/Emo-CampNet.



(a) The structure of neutral content generator (b) The structure of discriminator

Fig. 8. Structures of neutral content generator and discriminator in Emo-CampNet.

of the input. Resnet-based Convs are a stack of 5 1D convolutional layers in the network architecture that follow the principles of residual connections from the ResNet framework [1]. Each convolutional layer uses 256 filters of size 3. The hidden dimension of the transformer is 256. Conv1D-256-3-1 represents a 1D convolutional layer with 256 filters, a kernel size of 3, and a stride of 1. We set the masked region to be 12% of the total speech length at the training stage, which is the same configuration as the CampNet. Adam [16] is used as the optimizer, with an initial learning rate of $1e-3$. The model is then trained for 2 million steps with a batch size of 16.

- **p-w/o-dis** The key of Emo-CampNet is removing emotional information from the original masked speech by using the generative adversarial training method. In this way, the generated speech's emotion is controlled only by the input emotion attribute. Therefore, to verify the effect of the generative adversarial method, we remove the discriminator and retrain the model, denoted as p-w/o-dis. The model's loss function is only the reconstruction loss, as shown in Eq. (6). Other configurations are the same as the Emo-CampNet model.
- **p-w/o-ncg** To verify whether the neutral content generator structure we designed is useful for the model, we replace the neutral content generator module with the linear module to extract h_c information. We keep other experimental configurations unchanged to retrain the model, which is denoted as p-w/o-ncg.
- **p-w/o-eda** To verify the impact of the emotional data augmentation method proposed in Sec. 3.3, we remove the additional data from the emotional data augmentation method in the training set, keep other experimental settings unchanged, and retrain Emo-CampNet, which is denoted as p-w/o-eda.
- **p-w/o-nda** To verify the impact of the neutral data enrichment method proposed in Sec. 3.3, we remove the additional data from the neutral data enrichment method in the training set, keep other experimental settings unchanged, and retrain Emo-CampNet, which is denoted as p-w/o-nda.

5. Results

In this section, we will begin by comparing the performance of the proposed Emo-CampNet to other systems, including both objective and subjective metrics. As the primary purpose of Emo-CampNet is the synthesis of emotional speech, we will also pay close attention to its emotional expressiveness, particularly with regards to fundamental frequency and the results of emotion classification.

5.1. Objective evaluation for the quality of speech

First of all, we will use the insertion operation as an example. We will insert a piece of text at a specific position in a neutral emotional speech and use the proposed Emo-CampNet to synthesize speech with different emotions as input. The resulting spectrums of the generated speech with varying emotions are presented in Fig. 9. Notably, despite inserting the same text, the spectrums corresponding to different emotions exhibit significant differences, and the generated spectrums demonstrate a natural prosodic connection. This observation serves as evidence of the effectiveness of the emotion selection function.

Secondly, to objectively compare the performance of different models, we conduct an objective evaluation based on different systems. We take the replacement operation as an example because it is convenient for us to obtain the ground truth of generated

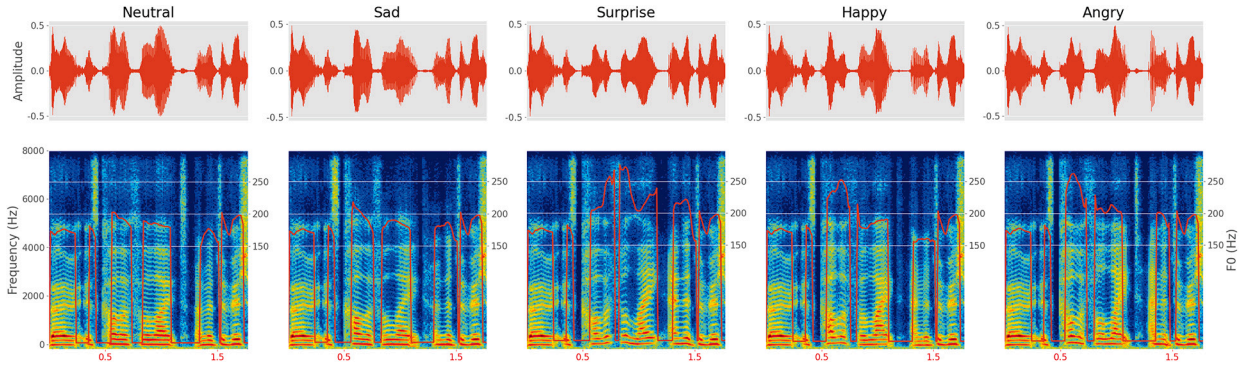


Fig. 9. The waveforms and spectrograms of insertion operation with the same text and different emotion attributes based on the proposed Emo-CampNet. The region marked with time (0.5 s ∼ 1.5 s) is the inserted region. The text of the masked region is 'colorful light'.

Table 1
Objective evaluation (MCD value) results of different systems on the test set.

Systems	Neutral	Surprise	Angry	Happy	Sad
p-w/o-ncg (fail synthesis)	-	-	-	-	-
p-w/o-eda	3.117	3.556	3.572	3.460	3.646
p-w/o-nda	3.183	3.806	3.679	3.787	3.528
p-w/o-dis	3.149	3.617	3.724	3.452	3.481
Emo-CampNet	**3.078**	**3.495**	**3.528**	**3.425**	**3.332**

speech to calculate the objective metrics. To prevent the interference of the emotional information in the original speech to the generated speech, we select 20 neutral emotional speeches from the test set in the emotional dataset ESD. Then we randomly select 20 words that span 3 to 10 phonemes from these 20 sentences. For each sentence, we remove the region of the corresponding words in the speech. Then we use different systems to predict the removed region with the five emotions (neutral, sad, angry, surprise, and happy). Therefore, since each test sentence can be used to synthesize five sentences of the same text but with different emotions, we edited 100 sentences as the test sample to calculate the objective metrics. Because the ESD dataset contains parallel corpora with different emotions, we can obtain the ground truth of different emotional speech in the removed region.

Following papers [51,50], we use the Mel-Cepstral Distortion (MCD) to evaluate emotional speech. The calculation method is as follows:

Given two mel-cepstra $\hat{\mathbf{x}} = \left[\hat{x}_1, \ldots, \hat{x}_M\right]^\top$ and $\mathbf{x} = \left[x_1, \ldots, x_M\right]^\top$, we use the mel-cepstral distortion (MCD):

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{M}\left(\hat{x}_i - x_i\right)^2} \tag{9}$$

to measure their difference. Where M is the order of mel-cepstrum and equals 28 in our implementation. Here, we used the average of the MCDs [18] taken along the DTW [27] path between the edited and reference feature sequences as the objective performance measure for each test utterance.

The objective results of the test set are listed in Table 1. We elaborate on our conclusions from the following three aspects.

5.1.1. *Comparison of metrics for different emotions*

Firstly, the results for the Emo-CampNet system reveal that the MCD value of neutral emotional speech is the lowest, which is consistent with the findings of other systems (p-w/o-dis, p-w/o-eda, p-w/o-nda). These results suggest that the acoustic feature of neutral emotional speech is the easiest to learn, possibly due to the smaller emotional change in neutral emotional speech and the more regular spectrum. Additionally, the MCD of sad emotional speech is relatively small, likely because of the smaller prosodic changes in sad emotional speech, which make it easier for the model to predict. Conversely, for emotions with significant prosodic changes, such as anger and surprise, their MCD values are higher than those of other emotions.

5.1.2. *Comparison with generative adversarial network*

When compared to the p-w/o-dis and p-w/o-ncg model, it becomes apparent that the two modules in Emo-CampNet, namely the neutral content generator and discriminator, are essential for its performance. Specifically, when the NCG module is removed from Emo-CampNet, the model cannot generate natural speech. We observed that the model's speech generation output was not natural in terms of its quality and intelligibility. The generated speech may have exhibited artifacts, distortions, or unnatural prosody that affected its overall naturalness. This can be explained by the fact that without the NCG module, the model cannot extract speech content information that is separate from emotion. However, the discriminator expects the model to generate emotion-independent content information, leading to the model's collapse.

Table 2
Objective evaluation (MCD value) results of different systems (different amounts of additional neutral dataset) on the test set.

Systems	Neutral	Surprise	Angry	Happy	Sad
p-w/o-nda	3.183	3.806	3.679	3.787	3.528
p-w-1/5nda	3.152	3.733	3.683	3.633	3.581
p-w-2/5nda	3.123	3.697	3.674	3.534	3.517
p-w-3/5nda	3.113	3.744	3.608	3.518	3.475
p-w-4/5nda	3.093	3.557	3.569	3.409	3.407
p-w-nda	3.078	3.495	3.528	3.425	3.332

When the discriminator is not used in Emo-CampNet, the model can synthesize neutral emotional speech with decent fidelity, but its objective measurement results are poor when generating speech with other emotions. This is due to the lack of a discriminator, which prevents the removal of emotional information from the speech. As a result, the generated speech's emotion is influenced not only by the input emotion attributes but also by the emotional information of the original speech. This leads to difficulty in controlling the emotion of the generated speech. Without the discriminator, the synthesized speech in the masked region would be neutral because the model can only synthesize speech based on the emotional information present in the unmasked regions of the speech. Consequently, the performance in synthesizing different emotions is compromised when the discriminator is removed, leading to inferior metrics compared to the model with the discriminator. Moreover, even if the test set contains different emotional categories, the model without a discriminator struggles to synthesize speech with the desired emotions, leading to inferior performance compared to the proposed method.

5.1.3. Comparison of data augmentation method

Comparing the Emo-CampNet model with the p-w/o-eda and p-w/o-nda models allows us to examine the impact of data augmentation methods on objective metrics.

When emotional data augmented by F0 disturbance is removed from the training data, we observe a significant reduction in objective metrics. Specifically, the objective metric of non-neutral emotional speech decreases more significantly than that of neutral emotional speech. This is because the F0 perturbation can alter the speaker's information while maintaining the emotional attributes, which facilitates the expansion of emotional data and enhances the model's generalization.

Similarly, when the neutral emotional data from the VCTK dataset is removed from the training set, the objective metric of the p-w/o-nda model significantly decreases. This is because the neutral emotional dataset from VCTK can help the model learn richer pronunciation and speaker information.

To further investigate the impact of different amounts of additional neutral dataset on performance, we have design an addition experiment as follows:

- Start by training the Emo-CampNet model using only the base data (without any additional neutral dataset).
- Record the performance metrics achieved by the model on the test set.
- Repeat the training process using the base data combined with different amounts of additional neutral dataset.
- For each iteration, record the model's performance on the test set.

We divided the additional neutral sentiment dataset into 5 parts and added one part in sequence, which are denoted as p-w-1/5nda, p-w-2/5nda, p-w-3/5nda, p-w-4/5nda, and p-w-nda. Overall, a total of six models were trained (including one model without any additional neutral emotions, denoted as p-w/o-nda).

We assess the models' performance on the test set using MCD metrics. The result it shown in Table 2.

Based on the results, in summary, based on the provided data, the inclusion of a certain proportion of NDA data appears to have a positive effect on the metrics. Besides, the inclusion of additional neutral speech in the training dataset can introduce a wealth of phonetic information. The original emotional speech dataset, such as the Emotional Speech Database (ESD), is parallel corpus data, which may not provide comprehensive coverage of phonetic variations. However, neutral speech data can be easily collected and can encompass a wide range of phonemes and pronunciation variations. This inclusion of neutral speech data can significantly enhance the model's robustness to different phonetic contexts and pronunciation patterns. By incorporating neutral speech data, the model can learn to generalize better and adapt to various speech patterns, thereby improving its overall performance. The availability of diverse phonetic information helps the model capture a broader range of acoustic and linguistic features, making it more adept at handling different speech styles and accents.

Besides, during the experiment, we found that as the number of speakers increases in the neutral emotion dataset, the model becomes more proficient at preserving the unique characteristics of the original speaker's voice. This suggests that incorporating a diverse range of speaker voices can improve the model's ability to generate speech that closely matches the characteristics of the desired speaker.

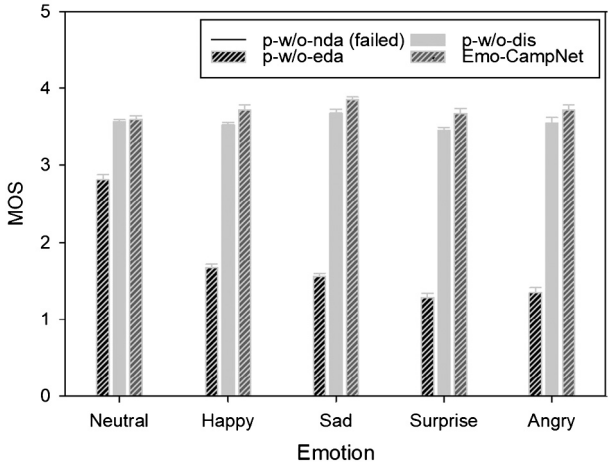


Fig. 10. The MOS score with 95% confidence intervals of the different systems.

Table 3
Average preference score of similarity between generated speech emotion and target emotion attribute of different methods (%), where n/p stands for "no Preferences", and p represents the p value of the t test.

	p-w/o-eda	p-w/o-dis	Emo-CampNet	N/P	p
p-w/o-eda vs Emo-CampNet	9.00	–	**73.00**	18.00	<0.01
p-w/o-dis vs Emo-CampNet	–	6.50	**79.00**	14.50	<0.01

5.2. Subjective evaluation for the speech

In this section, subjective evaluations are conducted to compare the performance of Emo-CampNet and other systems in terms of speech quality and emotional similarity of edited speech. We take the insertion operation as an example. The replacement operation is similar to the insertion operation, except that there are additional steps to mask part of the speech. To prevent the interference of the emotional information of the original speech to the generated speech, we select 40 neutral emotional speeches from the VCTK test set. For each utterance, we insert some words that span 8 to 20 phonemes into these sentences. Then, we use different systems to predict the inserted region with the five emotions (neutral, sad, angry, surprise, and happy). Therefore, each test sentence can be used to synthesize five sentences of the same text but with different emotions. We randomly select 20 parallel sentences for each emotion and collect the mean opinion score (MOS) separately. Ten listeners are asked to listen and rate the quality and the speaker identity of the edited sentence on a Likert scale: 1 = bad (very bad), 2 = poor (annoying), 3 = fair, 4 = good, and 5 = excellent (imperceptible, almost real). The result is shown in Fig. 10.

First, it can be found that p-w/o-nda fails to synthesize speech. This is because much pronunciation information is missing when there is no VCTK dataset in the training set, which combined with the arbitrary text insertion in the test stage, leads to poor performance of the model in the text outside the training set and failure of speech synthesis. In the previous objective evaluation, we have used text data that appears in the training set to obtain the ground truth speech for comparison. The results show that model p-w/o-nda can synthesize speech normally in the text within the training set, but it fails in the text outside the training set.

Secondly, it can be found that model p-w/o-eda performs well in neutral emotion, but the speech synthesized with other emotions is very poor. The MOS of the Emo-CampNet is slightly higher than that of p-w/o-dis. Since the MOS score mainly evaluates the speech quality and speaker similarity of synthesized speech, we will then evaluate the similarity between the emotion of generated speech and the given emotion attribute.

To this end, to assess the emotional similarity of the generated speech, we conducted an ABX test. For each subjective test, twenty sentences were randomly selected, and ten listeners evaluated each pair of generated speech. Prior to the evaluation, the listeners were informed of the target word and emotion attribute and asked to select the generated sample in each pair that was closer to the target emotion, or if they had no preference. The results are presented in Table 3. Overall, we found that Emo-CampNet outperformed the baseline system in all evaluations. Specifically, the model labeled as p-w/o-dis received the lowest score. This indicates that when the discriminator is removed, the model is unable to accurately predict the speech with a specific emotion based on the input emotion attributes.

In addition, it is worth mentioning that readers may find some differences in the speech quality between synthesized audio and real audio. We analyze the reasons for this could be the separate training of the vocoder and acoustic model, which may lead to cumulative errors during synthesis. Similar challenges are encountered in the field of speech synthesis, where directly inputting acoustic features predicted by models like Tacotron [43] or FastSpeech2 [32] into the vocoder can result in lower audio quality. To address this, various approaches have been proposed to improve the speech quality. For example, joint training of the acoustic model

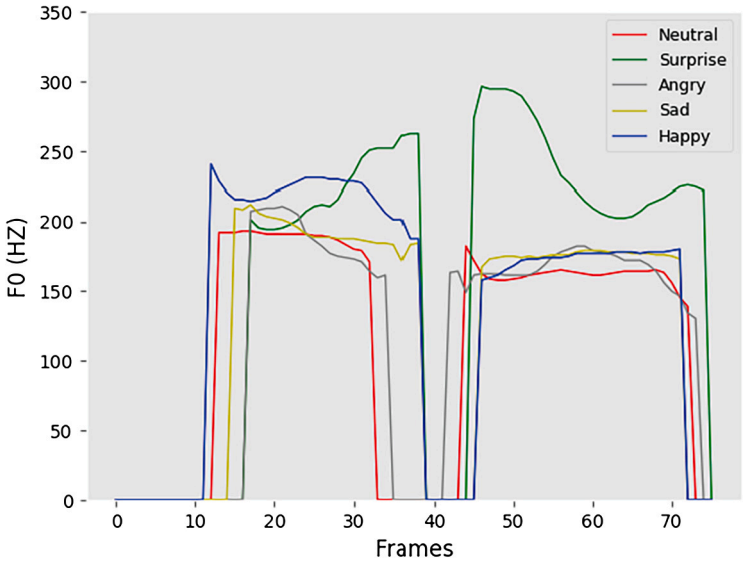


Fig. 11. An example of F0 curve of generated speech with the same text but with different emotion attributes. It can be found that the F0 curve of speech with different emotions has obvious distinction. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 4
A summary of mean and standard variance (std) of F0 of generated speech with different emotion.

Gender	Metrics	Neutral	Surprise	Angry	Happy	Sad
Female	F0 mean (Hz)	197.34	242.27	211.42	219.46	200.51
	F0 std (Hz)	44.50	62.42	48.13	59.19	40.37
Male	F0 mean (Hz)	130.92	176.96	155.67	169.49	136.24
	F0 std (Hz)	23.06	49.14	23.43	39.62	20.73

and vocoder (similar to the VITS [15] framework in speech synthesis), or fine-tuning the vocoder with acoustic features predicted by the acoustic model, or incorporating diffusion modules in the acoustic model's decoder (similar to Grad-tts [30] in speech synthesis) have shown promising results. The above methods have been proven to have significant effects in the field of speech synthesis. Therefore, if readers want to further improve the speech quality of speech, they can refer to the above methods. This paper will continue to validate the similarity and emotional controllability of speech.

5.3. Expression of fundamental frequency

The fundamental frequency (F0) is a crucial prosodic feature of speech. In order to examine whether the F0 of speech varies according to different emotions, we plotted the F0 curves of generated speech by inserting the same text and selecting different emotions. The F0 value was extracted using the Yin algorithm implemented in the Librosa tool [6], and the results are presented in Fig. 11. The analysis revealed that the F0 of each emotional speech differs. The F0 of neutral and sad emotional speech is relatively stable, while the F0 of speech with surprise, happiness, and anger undergoes significant changes.

Second, we computed the mean and standard deviation of F0 in speech editing regions in the test set by emotion and gender. The results are presented in Table 4. It can be observed that significant differences in the mean and variance of F0 among synthesized speech with different emotions. Specifically, the neutral emotional speech had the lowest mean F0, while surprise emotional speech had the highest mean F0. Moreover, the variance of F0 in sad emotion speech was the smallest, indicating that changes in F0 in sad emotion speech were minimal. On the other hand, the variance of F0 in angry emotion speech was the largest, suggesting that F0 in angry emotion speech underwent significant changes.

5.4. Emotion classification

To assess the emotional expression quality of the generated speech, we followed the approach in [51] and developed a speaker-independent speech emotion recognition (SER) model to evaluate the emotional attributes of the synthesized speech. We used the data from the Emotional Speech Dataset (ESD) as the training set, and applied the emotional data augmentation method discussed in Section 3.3 to augment the training set and improve the performance of the SER. We extracted the mel spectrum of speech as the acoustic feature, with a window size of 25 ms and a hop size of 10 ms.

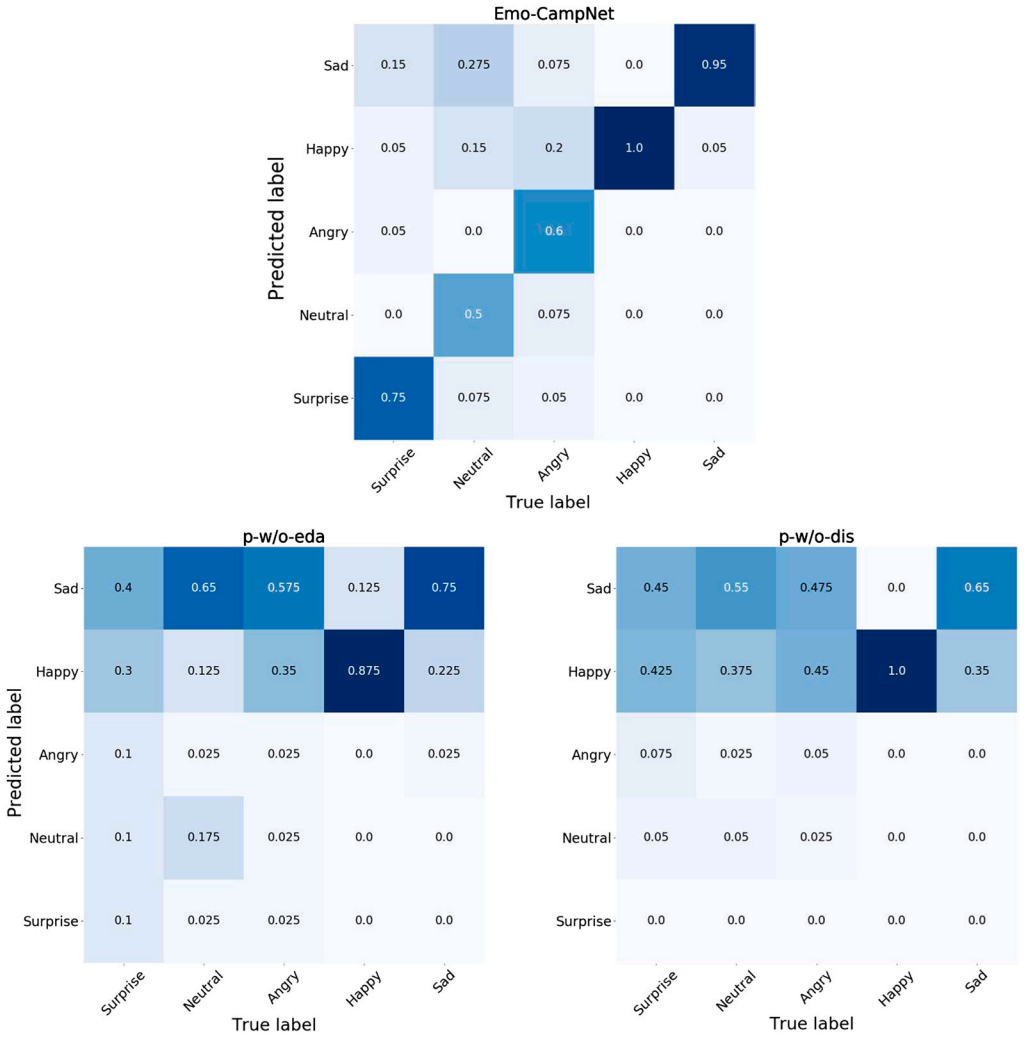


Fig. 12. Confusion matrix of different systems in a SER experiment on the test dataset. The diagonal entries represent the recall rates of each emotion.

The SER model consists of an LSTM layer, followed by a RELU activated fully connected layer with 256 nodes. Dropout is applied on the LSTM layer with a keep probability of 0.5. Finally, the resulting 256 feature vector is fed to a softmax classifier, an FC layer with 5 nodes. We randomly select 40 neutral emotional speeches from the test set. Then we insert some words into these 40 sentences and synthesize them with five emotions. Since the generated speech with long text is more helpful in reflecting the emotional information, the number of words to be inserted is more than 5, and the corresponding speech length is more than 2 seconds. Finally, we fed the synthesized speech in the editing area into the pre-trained SER model to determine whether the predicted emotion is consistent with the input emotion attribute.

The confusion matrices of SER results for different systems are presented in Fig. 12. The values on the diagonal represent the proportion of input emotional attributes consistent with the predicted emotional attributes. It can be observed that the Emo-CampNet model has the highest accuracy in each emotion. Moreover, it was observed that in Emo-CampNet, 27.5% of the synthesized speech controlled by the neutral emotion attribute was recognized as sad emotion. This finding suggests that the two emotions are easily confused during the synthesis process, perhaps due to small prosodic changes between them resulting in little difference.

At the same time, it can be observed that when the model loses the discriminator, most of the synthesized speech is recognized as happy or sad emotions. This is due to the emotional components in the input speech during the training stage, which disturbs the model, making it difficult for the input emotional attributes to effectively control the model. Moreover, removing the two data augmentation methods from the model results in a significant decrease in performance, further proving the effectiveness of our proposed method.

Finally, we compare the accuracy of all emotion recognition of different systems, as shown in Table 5. Similar to the results in Fig. 12, it can be observed that Emo-CampNet achieved the highest accuracy (76%), indicating that the proposed training framework of generative adversarial networks and data augmentation methods effectively improved the model's performance.

Table 5
Accuracy of different systems in all emotions in the SER experiment.

Systems	Accuracy
p-w/o-ncg	failed synthesis
p-w/o-nda	failed synthesis
p-w/o-eda	0.385
p-w/o-dis	0.35
Emo-CampNet	**0.76**

6. Conclusion

This paper introduces an end-to-end emotion-selectable text-based speech editing network, which enables the control of emotional attributes in synthesized speech during speech editing. This is a novel task, and we propose three innovations to achieve this goal. Firstly, we introduce emotion attribute information into the model using a context-aware mask and prediction framework, guiding the decoder to predict speech with a specific emotion. Secondly, we use a generative adversarial network to remove emotional components from the original speech to prevent interference with the decoder. Finally, we propose two data augmentation methods tailored to this task, which effectively improve the robustness and one-shot ability of the model. Experimental results show that our proposed method outperforms the baseline system in subjective and objective evaluations, as well as emotional expressiveness for emotion-selectable text-based speech editing tasks. Ablation experiments further confirm the effectiveness of our proposed method. Future work will focus on enhancing speech quality and expressiveness.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jianhua Tao, Tao Wang, Jiangyan Yi, Ruibo Fu has patent Method and apparatus for editing audio, electronic device and storage medium issued to US11462207B1.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by the Scientific and Technological Innovation Important Plan of China (No. 2021ZD0201502), the National Natural Science Foundation of China (NSFC) (No. 62322120, No. 62306316, No. 61831022, No. U21B2010, No. 62101553, No. 61971419, No. 62006223, No. 62206278).

References

[1] R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki, Gmm-based emotional voice conversion using spectrum and prosody features, Am. J. Signal Process. 2 (2012) 134–138.
[2] H. Bai, R. Zheng, J. Chen, M. Ma, X. Li, L. Huang, A3t: alignment-aware acoustic and text pretraining for speech synthesis and editing, in: International Conference on Machine Learning, PMLR, 2022, pp. 1399–1411.
[3] Z. Borsos, M. Sharifi, M. Tagliasacchi, Speechpainter: text-conditioned speech inpainting, arXiv preprint, arXiv:2202.07273, 2022.
[4] H. Choi, S. Park, J. Park, M. Hahn, Multi-speaker emotional acoustic modeling for cnn-based speech synthesis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6950–6954.
[5] C. Cui, Y. Ren, J. Liu, F. Chen, R. Huang, M. Lei, Z. Zhao, Emovie: a Mandarin emotion speech dataset with a simple emotional text-to-speech model, arXiv preprint, arXiv:2106.09317, 2021.
[6] A. De Cheveigné, H. Kawahara, Yin, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111 (2002) 1917–1930.
[7] D. Erickson, Expressive speech: production, perception and application to speech synthesis, Acoust. Sci. Technol. 26 (2005) 317–325.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[9] T. Gulzar, A. Singh, S. Sharma, Comparative analysis of lpcc, mfcc and bfcc for the recognition of Hindi words using artificial neural networks, Int. J. Comput. Appl. 101 (2014) 22–27.
[10] C.B. Im, S.H. Lee, S.B. Kim, S.W. Lee, Emoq-tts: emotion intensity quantization for fine-grained controllable emotional text-to-speech, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6317–6321.
[11] Z. Jin, et al., Speech Synthesis for Text-Based Editing of Audio Narration, 2018.
[12] P.R. Kammili, B. Ramakrishnam Raju, A.S. Krishna, Handling emotional speech: a prosody based data augmentation technique for improving neutral speech trained asr systems, Int. J. Speech Technol. 25 (2022) 197–204.
[13] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, K. Shikano, Gmm-Based Voice Conversion Applied to Emotional Speech Synthesis, 2003.
[14] T. Kenter, V. Wan, C.A. Chan, R. Clark, J. Vit, Chive: varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network, in: International Conference on Machine Learning, PMLR, 2019, pp. 3331–3340.
[15] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 5530–5540.

[16] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.
[17] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint, arXiv:1312.6114, 2013.
[18] R. Kubichek, Mel-cepstral distance measure for objective speech quality assessment, in: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, IEEE, 1993, pp. 125–128.
[19] O. Kwon, I. Jang, C. Ahn, H.G. Kang, An effective style token weight control technique for end-to-end emotional speech synthesis, IEEE Signal Process. Lett. 26 (2019) 1383–1387.
[20] Y. Lee, A. Rabiee, S.Y. Lee, Emotional end-to-end neural speech synthesizer, arXiv preprint, arXiv:1711.05447, 2017.
[21] Y. Lei, S. Yang, L. Xie, Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 423–430.
[22] J. Lorenzo-Trueba, G.E. Henter, S. Takaki, J. Yamagishi, Y. Morino, Y. Ochiai, Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis, Speech Commun. 99 (2018) 135–143.
[23] Z. Luo, J. Chen, T. Takiguchi, Y. Ariki, Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (2019) 1535–1548.
[24] Z. Luo, T. Takiguchi, Y. Ariki, Emotional voice conversion using deep neural networks with mcc and f0 features, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE, 2016, pp. 1–5.
[25] A. Mase, K. Oura, Y. Nankaku, K. Tokuda, Hmm-based singing voice synthesis system using pitch-shifted pseudo training data, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.
[26] M. Morrison, L. Rencker, Z. Jin, N.J. Bryan, J.P. Caceres, B. Pardo, Context-aware prosody correction for text-based speech editing, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 7038–7042.
[27] M. Müller, Dynamic time warping, in: Information Retrieval for Music and Motion, 2007, pp. 69–84.
[28] N.T. Pham, D.N.M. Dang, N.D. Nguyen, T.T. Nguyen, H. Nguyen, B. Manavalan, C.P. Lim, S.D. Nguyen, Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition, Expert Syst. Appl. (2023) 120608.
[29] O. Pierre-Yves, The production and recognition of emotions in speech: features and algorithms, Int. J. Hum.-Comput. Stud. 59 (2003) 157–183.
[30] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-tts: a diffusion probabilistic model for text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 8599–8608.
[31] G.J. Qi, J. Luo, Small data challenges in big data era: a survey of recent progress on unsupervised and semi-supervised methods, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 2168–2187, https://doi.org/10.1109/TPAMI.2020.3031898.
[32] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.Y. Liu, Fastspeech 2: fast and high-quality end-to-end text to speech, arXiv preprint, arXiv:2006.04558, 2020.
[33] G. Rizos, A. Baird, M. Elliott, B. Schuller, Stargan for emotional speech conversion: validated by data augmentation of end-to-end emotion recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3502–3506.
[34] R. Shankar, J. Sager, A. Venkataraman, Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator, arXiv preprint, arXiv:2007.12932, 2020.
[35] D. Tan, L. Deng, Y.T. Yeung, X. Jiang, X. Chen, T. Lee, Editspeech: a text based speech editing system using partial inference and bidirectional fusion, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 626–633.
[36] D. Tan, L. Deng, N. Zheng, Y.T. Yeung, X. Jiang, X. Chen, T. Lee, Correctspeech: a fully automated system for speech correction and accent reduction, in: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2022, pp. 81–85.
[37] R. Terashima, R. Yamamoto, E. Song, Y. Shirahata, H.W. Yoon, J.M. Kim, K. Tachibana, Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation, in: Proc. Interspeech 2022, 2022, pp. 3018–3022.
[38] S.Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, H.G. Kang, Emotional speech synthesis with rich and granularized control, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7254–7258.
[39] J.M. Valin, J. Skoglund, Lpcnet: improving neural speech synthesis through linear prediction, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5891–5895.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[41] C. Veaux, J. Yamagishi, K. MacDonald, et al., Cstr Vctk Corpus: English Multi-Speaker Corpus for Cstr Voice Cloning Toolkit, University of Edinburgh. the Centre for Speech Technology Research (CSTR), 2017.
[42] T. Wang, J. Yi, R. Fu, J. Tao, Z. Wen, Campnet: context-aware mask prediction for end-to-end text-based speech editing, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 2241–2254.
[43] 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, 2017.
[44] Y. Wang, D. Stanton, Y. Zhang, R.S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, R.A. Saurous, Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis, in: International Conference on Machine Learning, PMLR, 2018, pp. 5180–5189.
[45] Z. Wu, O. Watts, S. King, Merlin: an open source neural network speech synthesis system, in: SSW, 2016, pp. 202–207.
[46] W. Xia, Y. Zhang, Y. Yang, J.H. Xue, B. Zhou, M.H. Yang, Gan inversion: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 1–17 (2022), https://doi.org/10.1109/TPAMI.2022.3181070.
[47] D. Yin, C. Tang, Y. Liu, X. Wang, Z. Zhao, Y. Zhao, Z. Xiong, S. Zhao, C. Luo, Retrievertts: modeling decomposed factors for text-based speech insertion, arXiv preprint, arXiv:2206.13865, 2022.
[48] Y.J. Zhang, S. Pan, L. He, Z.H. Ling, Learning latent representations for style control and transfer in end-to-end speech synthesis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6945–6949.
[49] K. Zhou, B. Sisman, H. Li, Transforming spectrum and prosody for emotional voice conversion with non-parallel training data, arXiv preprint, arXiv:2002.00198, 2020.
[50] K. Zhou, B. Sisman, H. Li, Limited data emotional voice conversion leveraging text-to-speech: two-stage sequence-to-sequence training, in: Proc. Interspeech 2021, 2021, pp. 811–815.
[51] K. Zhou, B. Sisman, R. Rana, B.W. Schuller, H. Li, Emotion intensity and its control for emotional voice conversion, IEEE Trans. Affect. Comput. 14 (2022) 31–48.