Speech Synthesis With Mixed Emotions

Kun Zhou, *Student Member, IEEE*, Berrak Sisman, *Member, IEEE*, Rajib Rana, *Member, IEEE*, Björn W. Schuller, *Fellow, IEEE*, and Haizhou Li, *Fellow, IEEE*

Abstract—Emotional speech synthesis aims to synthesize human voices with various emotional effects. The current studies are mostly focused on imitating an averaged style belonging to a specific emotion type. In this paper, we seek to generate speech with a mixture of emotions at run-time. We propose a novel formulation that measures the relative difference between the speech samples of different emotions. We then incorporate our formulation into a sequence-to-sequence emotional text-to-speech framework. During the training, the framework does not only explicitly characterize emotion styles but also explores the ordinal nature of emotions by quantifying the differences with other emotions. At run-time, we control the model to produce the desired emotion mixture by manually defining an emotion attribute vector. The objective and subjective evaluations have validated the effectiveness of the proposed framework. To our best knowledge, this research is the first study on modelling, synthesizing, and evaluating mixed emotions in speech.

Index Terms—Emotional speech synthesis, mixed emotions, sequence-to-sequence, the ordinal nature of emotions, relative difference, emotion attribute vector

1 Introduction

Humans can feel multiple emotional states at the same time [1]. Consider some bittersweet moments such as remembering a lost love with warmth or the first time leaving home for college, it is possible to experience the co-occurrence of different types of emotions - even two oppositely valenced emotions (e. g., happy and sad) [2], [3]. Emotional speech synthesis aims to add emotional effects to a synthesized voice [4]. Synthesizing mixed emotions will mark a milestone for achieving human-like emotions in speech synthesis, thus enabling a higher level of emotional intelligence in human-computer interaction [5], [6], [7].

Speech synthesis aims to generate human-like voices from input text [8], [9], [10]. With the advent of deep learning, the state-of-the-art speech synthesis systems [11], [12], [13] are able to produce speech of high naturalness and intelligibility. However, most of them do not convey the omnipresent emotional contexts in human-human interaction [14], [15], [16]. The lack of expressiveness limits the emotional intelligence of current speech synthesis systems [17]. Emotional speech synthesis aims to fill this gap [18], [19], [20].

Synthesizing a mixed emotional effect is a challenging task. One of the reasons is the subtle nature of human emotions [21]. Therefore, it is not straightforward to precisely characterize speech emotion. Besides, speech emotion is inherently suprasegmental, complex with multiple acoustic cues such as timbre, pitch and rhythm [22], [23]. Both spectral and prosodic variants need to be studied when modelling speech emotion. The early studies on emotional speech synthesis rely on statistical modelling of different speech parameters with hidden Markov models (HMM) [24], [25] and Gaussian mixture model (GMM) [26], [27]. Deep neural networks (DNN) [28], [29] and deep bi-directional long-short-term memory network (DBLSTM) [30], [31] represent the recent advances. The end-to-end neural architecture [32], [33] becomes popular because of its superior performance. We note that there are generally two types of methods in the literature to learn emotion information: one uses auxiliary emotion labels as the condition of the framework [34], [35], and the other imitates the emotion style of the reference speech [36], [37]. However, these methods learn the global temporal structure of speech emotion, resulting in a monotonous expressiveness in synthesized speech. In this way, these frameworks can only synthesize

- *Kun Zhou is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077. E-mail: zhoukun@u.nus.edu.*
- *Berrak Sisman is with the Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX 75080 USA. E-mail: berraksisman@u.nus.edu.*
- *Rajib Rana is with the University of Southern Queensland, Toowoomba, QLD 4350, Australia. E-mail: Rajib.Rana@usq.edu.au.*
- *Björn W. Schuller is with the GLAM – the Group on Language, Audio, & Music, Imperial College London, SW7 2BX London, U.K., and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany. E-mail: bjoern.schuller@imperial.ac.uk.*
- *Haizhou Li is with the Shenzhen Research Institute of Big Data, and the School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong Province 518172, China, also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, and also with the University of Bremen, 28359 Bremen, Germany. E-mail: haizhouli@cuhk.edu.cn.*

Manuscript received 11 August 2022; revised 14 December 2022; accepted 20 December 2022. Date of publication 29 December 2022; date of current version 29 November 2023.
*The work of Kun Zhou was supported in part by the Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 under Grant 192 25 00054); in part by Human-Robot Collaborative AI under its AME Programmatic Funding Scheme under Grant A18A2b0046; and in part by the National Research Foundation Singapore under its AI Singapore Programme under Grant AISG-100E-2018-006. The work of Haizhou Li was supported in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant B10120210117-KP02, in part by the Research Foundation of Guangdong Province under Grant 2019A050505001 and in part by the National Natural Science Foundation of China under Grant 62271432.*
(Corresponding author: Kun Zhou.)
Recommended for acceptance by F. Metze.
Digital Object Identifier no. 10.1109/TAFFC.2022.3233324

several emotion types exhibited in the database. These disadvantages limit the flexibility and controllability of the above frameworks. For example, it is hard to synthesize mixed emotional effects with existing emotional speech synthesis frameworks.

For the first time, we study the modelling of mixed emotions in speech synthesis. In psychology, there have been studies [38], [39] to understand the paradigms and measures of mixed emotions. However, the study of mixed emotions in speech synthesis is not given attention yet, where there exist two main research problems: (1) how to characterize and quantify the mixture of speech emotions, and (2) how to evaluate the synthesized speech. In this article, we will address these two challenges.

The main contributions of this article are listed as follows:

- For the first time, we study the modelling of mixed emotions for speech synthesis, which brings us a step closer to achieving emotional intelligence;
- We introduce a novel scheme to measure the relative difference between emotion categories, with which the emotional text-to-speech framework learns to quantify the differences between the emotion styles of speech samples during the training. At run-time, we control the model to produce the desired emotion mixture by manually defining an emotion attribute vector;
- We carefully devise objective and subjective evaluations to confirm the effectiveness of the proposed framework and the emotional expressiveness of the speech.

This paper is organized as follows: In Section 2, we motivate our study by introducing the background and related work. In Section 3, we present the details of our proposed framework, and we introduce our experiments in Section 4. We provide further investigations in Section 5. The study is concluded in Section 6.

2 Background and Related Work

This work is built on several previous studies on the characterization of emotions, sequence-to-sequence emotion modelling for speech synthesis and controllable emotional speech synthesis. We briefly introduce the related studies to set the stage for our research and rationalize the novelty of our contributions.

2.1 Characterization of Emotions

Understanding human emotions (e. g., their nature and functions) has been gaining lots of attention in psychology [40], [41], [42]. This study is inspired by several previous research, including the theory of the emotion wheel and the ordinal nature of emotions.

2.1.1 Theory of the Emotion Wheel

Humans can experience around 34,000 different emotions [43]. While it is hard to understand all these distinct emotions, Plutchik proposed 8 primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust and joy, and arranged them in an emotion wheel [44] as shown in Fig. 1. All other emotions can be regarded as mixed or derivative states of these primary emotions [44]. According to the theory of the emotion wheel, the changes in intensity could produce the diverse amount of emotions we can feel. Besides, the adding up of primary emotions could produce new emotion types. For example, delight can be produced by combining joy and surprise [45].

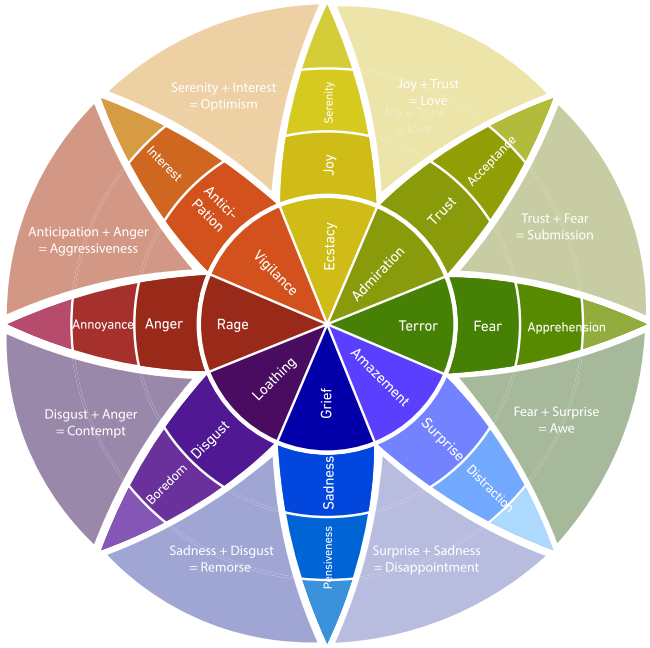


Fig. 1. An illustration of the theory of the emotion wheel [44], where all emotions occur as the mixed or derivative states of eight primary emotions.

Despite these efforts in psychology, there is almost no attempt to model the mixed emotions in the literature of speech synthesis. Inspired by the theory of the emotion wheel, we believe it is possible to combine different primary emotions and synthesize mixed emotions in speech. This technique will also allow us to create new emotion types that are hard to collect in real life, which could help us better mimic human emotions and further enhance the engagement in human-robot interaction.

2.1.2 The Ordinal Nature of Emotions

Emotions are intrinsically relative, and their annotations and analysis should follow the ordinal path [46], [47]. Instead of assigning an absolute score or an emotion category, ordinal methods characterize emotions through comparative assessments (e. g., is sentence one *happier* than sentence two?). Ordinal methods have shown remarkable performance, especially in speech emotion recognition [48], [49], [50].

The key idea of ordinal methods is to learn a ranking according to the given criterion. An example is preference learning [51], where the task is to establish preferences between samples. Once the preferences are established, ranking samples [52], [53], [54] is straightforward. Other rank-based methods [55], [56], [57] also show the effectiveness of modelling the affect for speech emotion recognition. As for emotional speech synthesis, researchers also explore the ordinal nature of emotions to model the emotion intensity [58], [59], [60], [61], where the intensity of an emotion is treated as the relative difference between neutral and

emotional samples. Inspired by the previous studies, we aim to study rank-based methods to quantify the relative differences between the speech samples from different emotion categories, which we discuss later.

2.2 Sequence-to-Sequence Emotion Modelling for Speech Synthesis

The sequence-to-sequence model with attention mechanism was first studied in machine translation [62] and later on found effective in speech synthesis [12], [63]. We consider that sequence-to-sequence models are suitable for modelling speech emotion. Sequence-to-sequence models are more effective in modelling the long-term dependencies at different temporal levels such as word, phrase and utterance [64]. By learning attention alignment, sequence-to-sequence models can capture the dynamic prosodic variants within an utterance [65]. They also allow for the prediction of the speech duration at run-time, which is a critical prosodic factor of the speech emotion [66].

There are generally two types of methods in the literature to model speech emotions: 1) explicit label-based and 2) reference-based approaches. Next, we will briefly introduce these two approaches in sequence-to-sequence modelling.

2.2.1 Learn to Associate With Explicit Labels

It is the most straightforward to characterize emotion by using explicit emotion labels [34], [35], where the model learns to associate labels with emotion styles. In [34], an emotion label vector is taken by the attention-based decoder to produce the desired emotion. In [35], a low-resourced emotional text-to-speech is built using model adaptation with a few emotion labels. In addition to the explicit labels of discrete emotion categories, there are attempts to condition the decoder with continuous variables [67].

2.2.2 Learn to Imitate a Reference

Another approach is to use a style encoder to imitate and transplant the reference style [32]. Global style token (GST) [36] is an example to learn style embeddings from the reference audio in an unsupervised manner. Some studies incorporate additional emotion recognition loss [33], [68], perceptual loss [60], [69] or adversarial training [70] to help with the emotion rendering. Other studies [71], [72], [73], [74] replace the global style embedding with phoneme or segmental level prosody embedding to capture multi-scale emotion variants. Similar approaches have also been applied to emotional voice conversion research. In [75], the style encoder further acts as the emotion encoder to learn actual emotion information through a two-stage training. In [76], a speaker encoder is further introduced to preserve the speaker information.

These successful attempts motivate us to leverage the sequence-to-sequence mechanism to enable emotion modelling for speech synthesis.

2.3 Controllable Emotional Speech Synthesis

Speech emotion is often manifested in various prosody aspects [77]. Emotion rendering can be controlled by modifying different prosodic cues. Current studies [78], [79] mainly focus on designing the prosody embedding as a control vector that is derived from a representation learning framework. For example, style tokens [36] are designed to represent high-level styles such as speaker style, pitch range and speaking rate. Emotion rendering can be controlled by choosing specific tokens. Recent attempts [80], [81] study a way to include a hierarchical, fine-grained prosody representation into the style token-based diagram [36]. Some other studies also use variational autoencoders (VAE) [82] to control the speech style by learning, scaling or combining disentangled representations [83], [84].

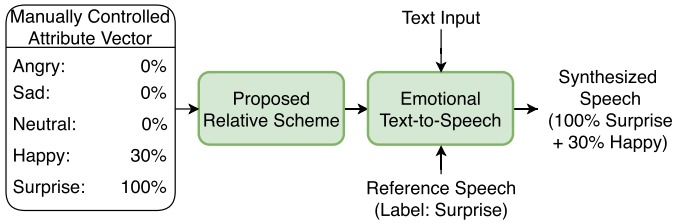


Fig. 2. Block diagram of our proposed relative scheme applied to emotional text-to-speech at run-time.

Recently, emotion intensity control has attracted much attention in emotional speech synthesis. Emotion intensity is considered to be correlated with all the acoustic cues that contribute to speech emotion [85], which makes itself even more subjective and challenging to model. Some studies use auxiliary features such as a state of voiced, unvoiced and silence (VUS) [86], attention weights or a saliency map [87] to control the emotion intensity. Other studies manipulate the internal emotion representations through interpolation [88], scaling [76] or distance-based quantization [89]. In [58], [59], [60], [61], relative attributes are introduced to learn a more interpretable representation of emotion intensity. However, none of these frameworks studied the correlation and interplay between different emotions. This contribution aims to fill this research gap.

2.4 Summary of Research Gap

We briefly summarize the gaps in the current literature on speech synthesis that we aim to address in this study:

- The synthesis of mixed emotions has not been studied in speech synthesis, which limits the capability of current systems to imitate human emotions;
- Despite much progress in psychology, it is still challenging to characterize and quantify the mixture of emotions in speech;
- Current evaluation methods are inadequate to assess mixed emotional effects. The rethinking of the current evaluation for mixed emotions is needed.

This study is a departure from the current studies on emotional speech synthesis. We seek to display the possibilities to synthesize mixed emotions that are subtle but do exist in our real life.

3 Mixed Emotion Modelling and Synthesis

We propose a novel relative scheme that allows for manually manipulating the synthesized emotion, i.e. mixing multiple different emotion styles. As shown in Fig. 2, the proposed scheme allows for flexible control of the extent of each contributing emotion in the speech. At run-time, the framework transfers

the reference emotion into a new utterance with the text input, also known as emotional text-to-speech.

We first describe our method of characterizing mixed emotions in speech and highlight our contributions to designing a novel relative scheme. Then, we present the details of the sequence-to-sequence emotion training with the proposed relative scheme. Lastly, we show the flexible control of the proposed framework for synthesizing mixed emotions.

3.1 Characterization of Mixed Emotions in Speech

Emotion can be characterized with either categorical [90], [91] or dimensional representations [92], [93]. With designated emotion labels, the emotion category approach is the most straightforward way to represent emotions. However, such representation ignores the subtle variations of emotions. Another approach seeks to model the physical properties of speech emotion with dimensional representations. An example is Russell's circumplex model [92], where emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions.

One of the most straightforward ways to characterize mixed emotions is to inject different emotion styles into a continuous space. Mixed emotions could be synthesized by adjusting each dimension carefully. However, only a few emotional speech databases [94], [95] provide such annotations. These dimensional annotations are subjective and expensive to collect. Therefore, we only utilize discrete emotion labels available in most databases. We first make an assumption based on the theory of the emotion wheel [44]: Mixed emotions are characterized by combinations, mixtures, or compounds of primary emotions. While it is not straightforward to add up emotions, we explore the ordinal nature of emotions instead.

We propose a rank-based relative scheme to quantify the relative difference between speech recordings with different emotion types. Mixed emotions can be characterized by adjusting the relative difference with other emotion types. The relative difference value can also quantify the level of engagement of each emotion. We introduce our design of a novel relative scheme next.

3.2 Design of a Novel Relative Scheme

One of the challenges of synthesizing mixed emotions is quantifying the association or the interplay between different emotions. Inspired by the ordinal nature of emotions, we propose a novel relative scheme to address this challenge. We first make two assumptions according to the theory of the emotion wheel: (1) all emotions are related to some extent; (2) each emotion has stereotypical styles. In our proposal, we not only characterize the identifiable styles of each emotion but also seek to quantify the similarity between different emotion styles.

We study a rank-based method to measure the relative difference between emotion categories, which can offer more informative descriptions and thus be closer to human supervision [96]. In computer vision, the relative attribute [96] represents an effective way to model the relative difference between two categories of data. Inspired by the success in various computer vision tasks [97], [98], [99], we believe relative attributes bridge between the low-level features and high-level semantic meanings, which allows us to model the relative difference between emotions only with discrete emotion labels. In this way, we regard the identifiable emotion style as an attribute of speech data, which can be represented with a rich set of emotion-related acoustic features. The relative difference of the emotion styles can be modelled as a relative attribute, which is called "*emotion attribute*" in this article. The emotion attribute can be learned through a max-margin optimization problem as explained below:

Given a training set $T = \{\mathbf{x}_n\}$, where $\mathbf{x}_n$ is the acoustic features of the n^{th} training sample, and $T = A \cup B$, where A and B are two different emotion sets, we aim to learn a ranking function given as below:

$$f(x_n) = \mathbf{W}\mathbf{x}_n, \tag{1}$$

where $\mathbf{W}$ is a weighting matrix indicating the difference in emotion styles. According to hypotheses (1) and (2), we propose the following constraints:

$$\forall \mathbf{x}_i \in A, \forall \mathbf{x}_j \in B : \mathbf{W}\mathbf{x}_i > \mathbf{W}\mathbf{x}_j \tag{2}$$

$$\forall (\mathbf{x}_i, \mathbf{x}_j) \in A, \forall (\mathbf{x}_i, \mathbf{x}_j) \in B : \mathbf{W}\mathbf{x}_i = \mathbf{W}\mathbf{x}_j, \tag{3}$$

The weighting matrix $\mathbf{W}$ is estimated by solving the following problem similar to that of a support vector machine [100]:

$$\min_{\mathbf{W}} \left(\frac{1}{2} \parallel \mathbf{W} \parallel_2^2 + C \left(\sum \xi_{i,j}^2 + \sum \gamma_{i,j}^2 \right) \right) \tag{4}$$

$$\text{s.t. } \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{i,j}; \forall \mathbf{x}_i \in A, \forall \mathbf{x}_j \in B \tag{5}$$

$$|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)| \leq \gamma_{i,j}; \forall (\mathbf{x}_i, \mathbf{x}_j) \in A, \forall (\mathbf{x}_i, \mathbf{x}_j) \in B \tag{6}$$

$$\xi_{i,j} \geq 0; \gamma_{i,j} \geq 0, \tag{7}$$

where C is the trade-off between the margin and the size of slack variables $\xi_{i,j}$ and $\gamma_{i,j}$.

Through Eq. (4)–(7), we learn a wide-margin ranking function that enforces the ordering on each training point. As shown in Fig. 3a, we train a relative ranking function $f(x)$ between each emotion pair. At the inference phase, the trained function can estimate an emotion attribute of unseen data, as shown in Fig. 3b. In practice, each emotion attribute value is normalized to [0,1], where a smaller value indicates a similar emotional style. All the normalized emotion attributes form an emotion attribute vector. The emotion attribute vector bridges the discrete primary emotion labels and is further incorporated in sequence-to-sequence emotion training.

3.3 Training Strategy

We adopt an emotional text-to-speech framework with the joint training of voice conversion as in [75]. As both text-to-speech and voice conversion share a common goal of generating realistic speech from the internal representations, the joint training was shown effective [101], [102], [103], [104]. The text-to-speech task could benefit from the phone-embedding vectors [105], [106], or the prosody style introduced by a reference encoder [32]. A shared decoder between text-to-speech and voice conversion contributes to a robust decoding process [107], [108], [109].

The overall emotional text-to-speech framework is an encoder-decoder model that is trained as a sequence-to-sequence system, as shown in Fig. 4, where the text encoder

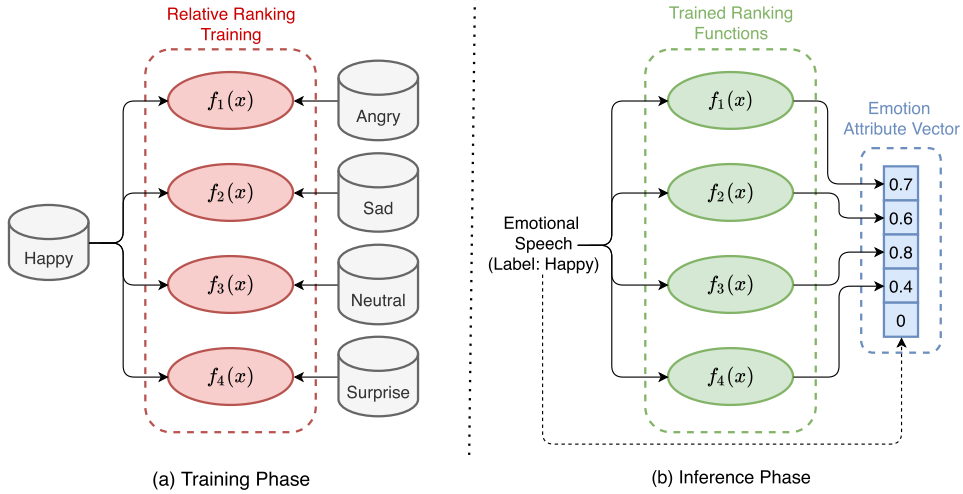


Fig. 3. The illustration of the proposed relative scheme at (a) training and (b) run-time phase. A relative ranking function is trained between each emotion pair and automatically predicts an emotion attribute at run-time. A smaller emotion attribute value represents a similar emotional style between the pairs. All the emotion attributes form an emotion attribute vector.

and linguistic encoder generate an embedding sequence for the input, while the emotion encoder generates one embedding that encapsulates the whole reference speech sample.

Given the text or speech as input, the text and the linguistic encoder learn to predict the linguistic embedding from the text or speech, respectively. The decoder takes the linguistic embedding from the text or speech in an alternative manner, depending on whether the epoch number is odd or even. Similar to [102], a contrastive loss is used to ensure the similarity between these two types of linguistic embeddings. The adversarial training strategy with an emotion classifier is employed on the acoustic linguistic embedding to eliminate the residual emotion information.

An emotion encoder is used to extract an emotion embedding vector from the input speech under the supervision of an emotion label. Meanwhile, an emotion attribute vector is generated by the pre-trained relative scheme described in Section 3.2, and then produced by a fully connected (FC) layer, resulting in a relative embedding. The emotion embedding describes the emotion styles of the input speech, while the emotion attribute vector indicates the difference between the input emotion style and other emotion styles. Finally, the decoder learns to reconstruct the input emotion style from a combination of emotion and relative embeddings.

The whole training procedure can be viewed as a recognition-synthesis process at the sequence level. Our proposed framework does not only learn the abundant emotion variance that is exhibited in a database but also the correlation or association across different emotion categories. It allows us to explicitly adjust the difference level at run-time and further enables mixed emotion synthesis and the flexible control of emotion rendering at the same time, which will be discussed next.

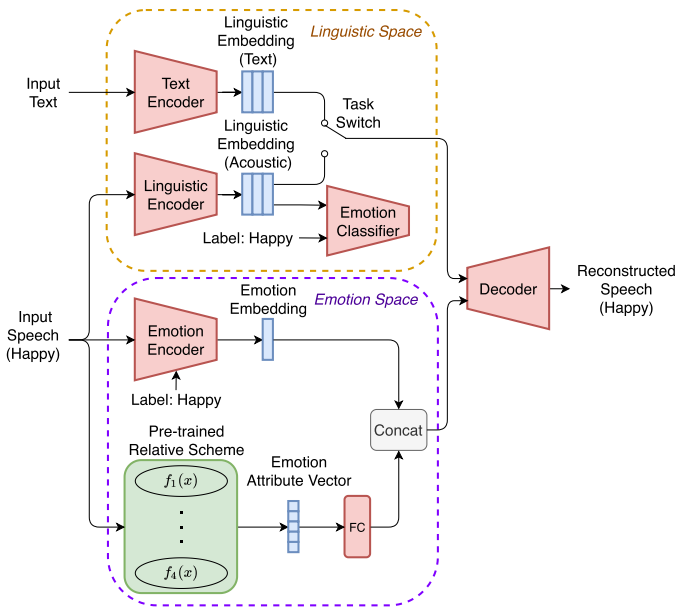


Fig. 4. The training diagram of the proposed framework. The pre-trained relative scheme learns to generate an emotion attribute vector that measures the relative difference between the input emotion style ('Happy') and other primary emotion styles ('Angry', 'Sad', 'Surprise' and 'Neutral').

3.4 Control of Emotion Rendering

We illustrate our proposed emotional text-to-speech framework in Fig. 5, which renders controllable emotional speech at run-time. The framework consists of three main modules, the content encoder, the emotion controller, and the decoder.

The text encoder projects the linguistic information from the input text into an internal representation. The emotion encoder captures the emotion style in an embedding from the reference speech, while the relative scheme further introduces the characteristics of other emotion types with a manually assigned attribute vector. By varying the percentage for each primary emotion in the attribute vector, we can easily synthesize the desired emotional effects and control the emotion rendering in synthesized speech.

4 Experiments and Evaluations

In this section, we report our experimental settings and results. As shown in Table 1, for all the experiments, we synthesize mixed emotional effects by mixing a primary emotion (*Surprise*) with three reference emotions (*Happy*, *Angry* and *Sad*) respectively. We expect to synthesize mixed emotional effects similar with the secondary emotions such

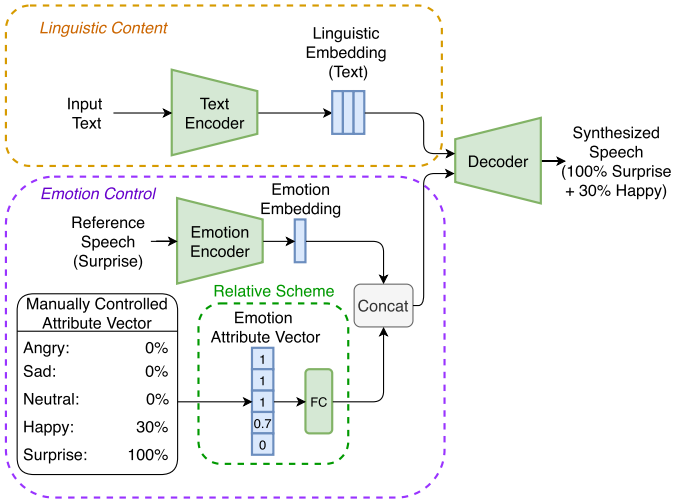


Fig. 5. The run-time diagram of the proposed emotional text-to-speech framework. The emotion rendering can be manually controlled via the relative scheme. By assigning the appropriate percentage to the attribute vector, we produce a target emotion mixture.

as *Delight*, *Outrage* and *Disappointment*, respectively. We choose these three combinations because they are thought to be easier to perceive for the listeners and have been studied in psychology [1], [44].

Since this contribution serves as a pioneer in related fields, there is no literature or reference method before this study, to our best knowledge. Therefore, we could not include any baselines in our experiments. Instead, we adopt objective and subjective metrics widely used in previous literature and carefully design evaluation methods to show the effectiveness of our proposal. We have made the source codes and speech demos available to the public[1]. We encourage readers to listen to the speech samples on our demo website to best understand this work.

4.1 Experimental Setup

We use acoustic features and phoneme sequences as inputs to the proposed framework during the training. The acoustic features are 80-dimensional logarithm Mel-spectrograms extracted every 12.5 ms with a frame size of 50 ms for short-time Fourier transform (STFT). We convert text to phoneme with the Festival [110] G2P tool to serve as the input to the text encoder. At run-time, we synthesize emotional speech from the text input.

4.1.1 Network Configuration

Our proposed framework can be regarded as a sequence-level recognition-synthesis structure similar to that of [102], [111]. Both the linguistic encoder and the decoder have a sequence-to-sequence encoder-decoder structure. The linguistic encoder consists of an encoder, a 2-layer 256-cell BLSTM and a decoder, a 1-layer 512-cell BLSTM with an attention layer followed by a full-connected (FC) layer with an output channel of 512. The decoder has the same model architecture as that of Tacotron [12].

The text encoder is a 3-layer 1D CNN with a kernel size of 5 and a channel number of 512. The text encoder is followed by a 1-layer of 256-cell BLSTM and an FC layer with an output channel number of 512. The style encoder is a 2-layer 128-cell BLSTM followed by an FC layer with an output channel number of 64. The classifier is a 4-layer FC with channel numbers of {512, 512, 512, 5}.

TABLE 1
Our Experimental Settings of One Primary Emotion (A), Three Reference Emotions (B) and the Expected Mixed Emotional Effects (A+B)

Primary Emotion (A)	Reference Emotion (B)	Mixed Effects (A+B)
Surprise	Happy	Delight
Surprise	Angry	Outrage
Surprise	Sad	Disappointment

4.1.2 Training Pipeline

We first pre-train a relative ranking function between each emotion pair using an emotional speech dataset. We implement the relative ranking function following an open-source repository[2]. We use a standardized set of 384 acoustic features extracted with openSMILE [112] as the input features. These features include zero-crossing rate, frame energy, pitch frequency, and Mel-frequency cepstral coefficient (MFCC) used in the Interspeech Emotion Challenge [113]. The trained ranking functions reported a classification accuracy of 97% on the test set.

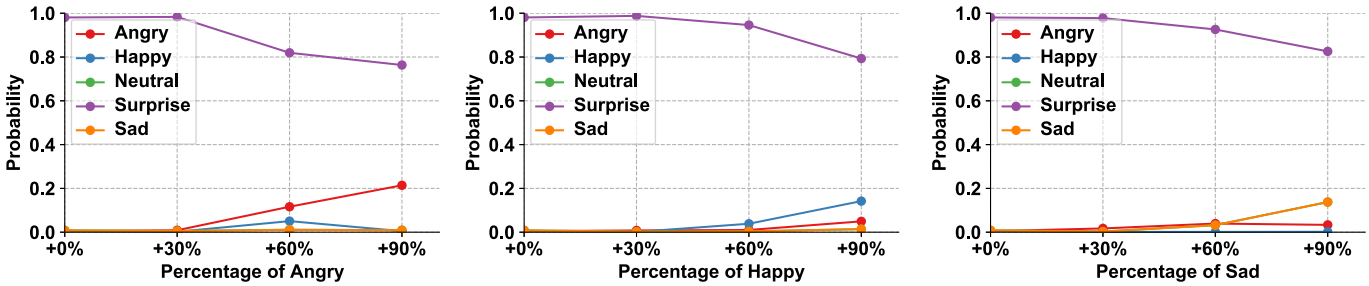
We then conduct a two-stage training strategy to train our text-to-speech framework, which consists of (1) Multi-speaker text-to-speech training with the VCTK Corpus [114] and (2) Emotion Adaptation for text-to-speech with a single speaker from the ESD dataset [115], [116]. The proposed text-to-speech framework learns abundant speaker styles with a multi-speaker corpus and then learns the actual emotion information with a small amount of emotional speech data. The training strategy we used is similar to that of [75]. During the training, we use the Adam optimizer [117] and set the batch size to 64 and 4 for multi-speaker text-to-speech training and emotion adaptation, respectively. We set the learning rate to 0.001 and the weight decay to 0.0001 for multi-speaker text-to-speech training. We halve the learning rate every seven epochs during the emotion adaptation.

4.1.3 Data Preparation

We select the VCTK Corpus [114] to perform multi-speaker text-to-speech training, where we use 99 speakers and the total duration of training speech data is about 30 hours. We select the ESD dataset [115], [116] to perform emotion adaptation and relative ranking training. We choose one English male ('0013') and one English female ('0019') speaker from the ESD. We consider five emotions: *Neutral*, *Angry*, *Happy*, *Sad* and *Surprise*, and for each emotion, we follow the data partition given in the ESD. For each speaker and each emotion, we use 300, 30 and 20 utterances for training, testing,

1. **Codes & Speech Demos**: https://kunzhou9646.github.io/Mixed_Emotions_Demo/

2. https://github.com/chaitanya100100/Relative-Attributes-Zero-Shot-Learning



(a) Mixing Surprise (100%) with different percentages of Angry (0%, 30%, 60%, 90%)

(b) Mixing Surprise (100%) with different percentages of Happy (0%, 30%, 60%, 90%)

(c) Mixing Surprise (100%) with different percentages of Sad (0%, 30%, 60%, 90%)

Fig. 6. Classification probabilities derived from the pre-trained SER model for a male speaker ('0013') from the ESD dataset. Each point represents an averaged probability value of 20 utterances with mixed emotions.

and evaluation, respectively. The total duration of emotional speech training data is around 50 minutes.

4.2 Objective Evaluation

We first perform objective evaluations to validate the proposed mixed emotion synthesis. We demonstrate the effectiveness of our proposals and provide analysis with a pre-trained speech emotion recognition (SER) model. We calculate Mel-cepstral distortion (MCD) and Pearson correlation coefficient (PCC) as objective evaluation metrics.

4.2.1 Analysis With Speech Emotion Recognition

We train a speech emotion recognition model on the ESD dataset [115] with the same data partition described in Section 4.1.3. To improve the robustness of SER, data augmentation is performed by adding white Gaussian noise during the SER training [118], [119], [120], [121].

The SER architecture is the same as that in [122], which includes: 1) a three-dimensional (3-D) CNN layer; 2) a BLSTM; 3) an attention layer; and 4) a fully connected (FC) layer. We evaluate our synthesized mixed emotions with the pre-trained SER. We use the classification probabilities derived from the softmax layer of the SER to analyze the effects of mixed emotions. As a high-level feature, the classification probabilities summarize the useful emotion information from the previous layers for final decision-making. The classification probabilities offer us an effective tool to justify how well each emotional component can be perceptually recognized by the SER from the emotion mixture.

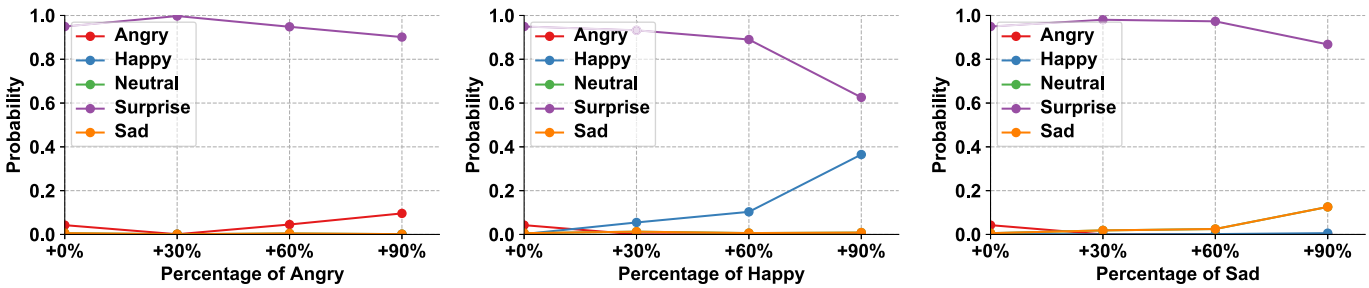
We first report the classification probabilities for a male speaker ('0013') in Fig. 6. We evaluate four different combinations where we gradually increase the percentage (0%, 30%, 60%, 90%) of *Angry*, *Happy* or *Sad* while keeping that of *Surprise* always being 100%. As shown in Fig. 6a, we observe that the probability of *Angry* increases while we increase the percentage of *Angry* from 0% to 90%. In the meanwhile, the probability of *Surprise* decreases but still remains to be higher than for others. The probability of *Angry* achieves 0.25 when the percentage of *Angry* reaches 90%. We also note similar observations for *Happy* and *Sad* as shown in Figs. 6b and 6c.

We then report the classification probabilities for a female speaker ('0019') in Fig. 7. Similar to that of the male speaker, we report four different percentages (0%, 30%, 60%, 90%) of *Angry*, *Happy* or *Sad* while keeping that of *Surprise* being 100%. For *Happy*, we observe the probability of *Happy* considerably increases while we increase the percentage of *Happy* in mixed emotions as shown in Fig. 7b. For *Angry* and *Sad*, we find similar observations as in Figs. 7a and 7c. These observations indicate that the mixed emotions can be perceptually recognized by a pre-trained SER.

4.2.2 Mel-Cepstral Distorion

Spectral features, based on the short-term power spectrum of sound, such as Mel-cepstral coefficients (MCEPs), contain rich information about expressivity and emotion [123]. Mel-cepstral Distortion (MCD) [124] is a widely adopted metric to measure the spectrum similarity, which is calculated between the synthesized ($\hat{\mathbf{y}} = \{\hat{\mathbf{y}}_m\}$) and the target MCEPs ($\mathbf{y} = \{\mathbf{y}_m\}$):

$$\text{MCD [dB]} = \frac{10\sqrt{2}}{\ln 10}\frac{1}{M}\sqrt{\sum_{m=1}^{M}(\mathbf{y_m} - \hat{\mathbf{y}}_m)^2}, \tag{8}$$

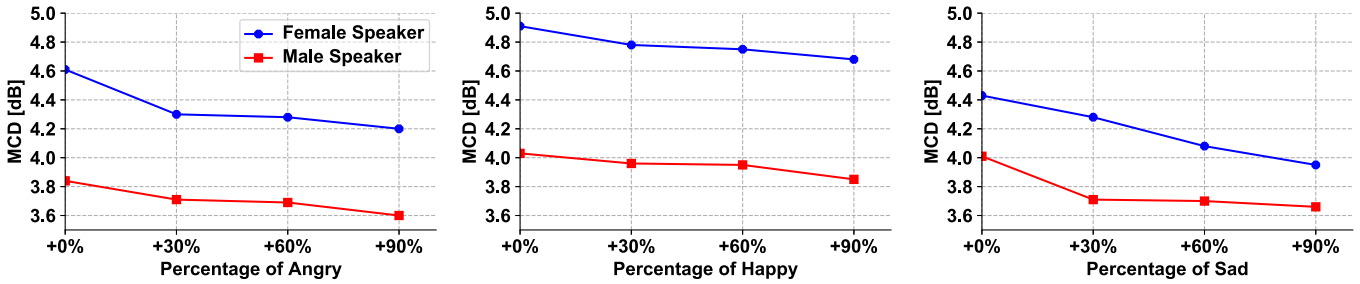


(a) Mixing Surprise (100%) with different percentages of Angry (0%, 30%, 60%, 90%)

(b) Mixing Surprise (100%) with different percentages of Happy (0%, 30%, 60%, 90%)

(c) Mixing Surprise (100%) with different percentages of Sad (0%, 30%, 60%, 90%)

Fig. 7. Classification probabilities derived from the pre-trained SER model for a female speaker ('0019') from the ESD dataset. Each point represents an averaged probability value of 20 utterances with mixed emotions.



(a) Mixing Surprise (100%) with different percentages of Angry (0%, 30%, 60%, 90%)

(b) Mixing Surprise (100%) with different percentages of Happy (0%, 30%, 60%, 90%)

(c) Mixing Surprise (100%) with different percentages of Sad (0%, 30%, 60%, 90%)

Fig. 8. Mel-cepstral distortion (MCD) [dB] calculated between the Mel-cepstral coefficients (MCEPs) of mixed emotions and the reference emotions (Angry, Happy and Sad). Each point represents an averaged MCD value of 20 utterances with mixed emotions.

where M represents the dimension of the MCEPs. A lower value of MCD indicates a higher degree of in the spectrum.

4.2.3 Pearson Correlation Coefficient

Pitch is considered a major prosodic factor contributing to speech emotion, closely correlated to the activity level [125], [126]. In practice, the pitch is often represented by the fundamental frequency (F0), which can be estimated with the harvest algorithm [127]. We calculate the Pearson Correlation Coefficient (PCC) of F0 to measure the linear dependency between two F0 sequences, which has been used in previous studies [128], [129], [130]. The PCC between two F0 sequences is given as:

$$\rho(F_0^s, F_0^t) = \frac{cov(F_0^s, F_0^t)}{\sigma_{F_0^s}\sigma_{F_0^t}}, \tag{9}$$

where $cov(\cdot)$ represents the covariance function, $\sigma_{F_0^s}$ and $\sigma_{F_0^t}$ are the standard deviations of the synthesized sequences (F_0^s) and the target F0 sequences (F_0^t), respectively. A higher PCC value represents a higher degree of similarity in prosody.

4.2.4 Discussion of the MCD and PCC Results

To show the effectiveness of synthesizing mixed emotions, we calculate MCD and PCC between the synthesized results and the reference emotions (*Angry*, *Happy* and *Sad*). We choose one male ('0013') and one female speaker ('0019') from the ESD dataset [115]. For each speaker, we use 20 utterances for evaluation. We report four different percentages of *Angry*, *Happy* and *Sad* that are: 0%, 30%, 60% and 90%. Again, we keep *Surprise* as the primary emotion that has a percentage of *Surprise* is always 100%.

We first compare spectrum similarity as shown in Fig. 8. For all three different combinations, we observe that the MCD values decrease as the percentage of reference emotions (*Angry*, *Happy* and *Sad*) increases as shown in Fig. 8a, 8b and 8c. These results show that the synthesized emotion becomes more similar to the reference emotions in the spectrum as we increase the percentage of the reference emotions.

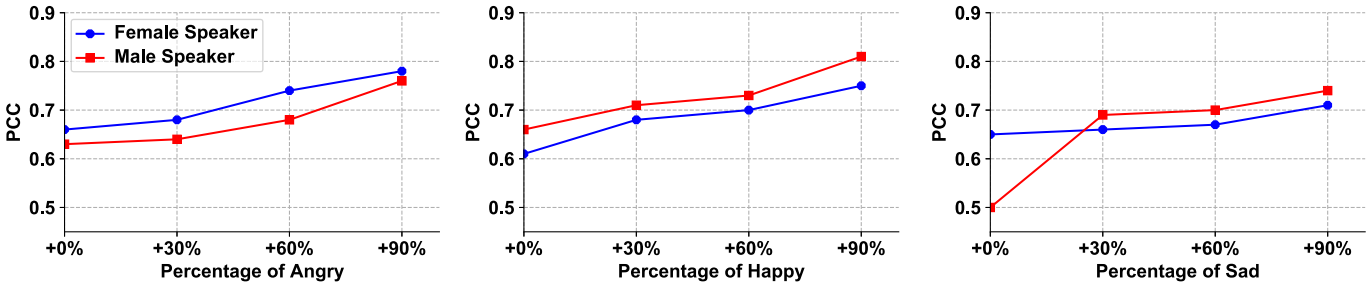
We have similar observations for prosody similarity as shown in Fig. 9. As the percentage of reference emotions (*Angry*, *Happy* and *Sad*) increases, we observe that the PCC value consistently increases. It indicates that the synthesized mixed emotions have a stronger correlation with the reference emotions (*Angry*, *Happy* and *Sad*) in terms of the prosody variance. These results show that we can effectively synthesize and further control the rendering of mixed emotions in terms of the spectrum and prosody.

4.3 Subjective Evaluation

We conduct subjective evaluations with human listeners, whom we ask to focus on two aspects: (1) Speech Quality and (2) Emotion Perception.

4.3.1 Speech Quality

We first conduct the Mean Opinion Score (MOS) test to evaluate speech quality, covering the speech's naturalness, intelligibility and listening efforts. All participants



(a) Mixing Surprise (100%) with different percentages of Angry (0%, 30%, 60%, 90%)

(b) Mixing Surprise (100%) with different percentages of Happy (0%, 30%, 60%, 90%)

(c) Mixing Surprise (100%) with different percentages of Sad (0%, 30%, 60%, 90%)

Fig. 9. Pearson Correlation Coefficient (PCC) calculated between the fundamental frequency (F0) of mixed emotions and the reference emotions (Angry, Happy and Sad). Each point represents an averaged PCC value of 20 utterances with mixed emotions.

TABLE 2
Mean Opinion Score (MOS) With 95% Confidence Interval to Evaluate the Speech Quality of Synthesized Mixed Emotions

Configuration		MOS
Ground truth (Surprise)		4.83 ± 0.16
Mixing Surprise (100%) with Angry	Ground truth (Angry)	4.81 ± 0.19
	+ 0% Angry	3.51 ± 0.36
	+ 30% Angry	3.79 ± 0.37
	+ 60% Angry	3.81 ± 0.35
	+ 90% Angry	3.76 ± 0.35
Mixing Surprise (100%) with Happy	Ground truth (Happy)	4.93 ± 0.05
	+ 0% Happy	3.21 ± 0.41
	+ 30% Happy	3.36 ± 0.36
	+ 60% Happy	3.39 ± 0.39
	+ 90% Happy	3.52 ± 0.42
Mixing Surprise (100%) with Sad	Ground truth (Sad)	4.84 ± 0.15
	+ 0% Sad	3.64 ± 0.35
	+ 30% Sad	3.73 ± 0.32
	+ 60% Sad	3.74 ± 0.31
	+ 90% Sad	3.60 ± 0.38

TABLE 3
Best-Worst Scaling (BWS) Test Results to Evaluate the Perception of the Reference Emotions (*Angry*, *Happy*, and *Sad*) in Synthesized Mixed Emotions

(a) Perception of *Angry*

Configuration		Best (%)	Worst (%)
Mixing Surprise (100%) with Angry	+ 0% Angry	8.3	61.7
	+ 30% Angry	6.0	19.5
	+ 60% Angry	24.8	11.3
	+ 90% Angry	**60.9**	**7.5**

(b) Perception of *Happy*

Configuration		**Best (%)**	**Worst (%)**
Mixing Surprise (100%) with Happy	+ 0% Happy	8.3	44.4
	+ 30% Happy	24.0	25.6
	+ 60% Happy	27.1	**11.2**
	+ 90% Happy	**40.6**	18.8

(c) Perception of *Sad*

Configuration		**Best (%)**	**Worst (%)**
Mixing Surprise (100%) with Sad	+ 0% Sad	9.0	57.1
	+ 30% Sad	9.0	29.3
	+ 60% Sad	20.3	**3.8**
	+ 90% Sad	**61.7**	9.8

are asked to listen to the reference speech ("Ground truth") and the synthesized speech with mixed emotions and score the "quality" of each speech sample on a 5-point scale ('5' for excellent, '4' for good, '3' for fair, '2' for poor, and '1' for bad). 20 subjects listened to 80 speech samples in total (80 = 5 x 4 (# of percentages) x 3 (*Angry*, *Happy* and *Sad*) + 20 (# of Ground truth)). The actual speech samples can be found in our demo website. We report the MOS results in Table 2, which show that our synthesized mixed emotions retain the speech quality between fair and good.

4.3.2 Emotion Perception

We then conduct the best-worst scaling (BWS) test to evaluate the emotion perception of synthesized mixed emotions. All participants are asked to listen to the speech samples and choose the best and the worst one according to their perception of a specific emotion type. 20 subjects listened to 168 speech samples in total (168 = 7 x 4 (# of percentages) x 6 (*Angry*, *Happy*, *Sad*, *Outrage*, *Delight* and *Disappointment*)). The actual speech samples can be found on our demo website.

We first evaluate the perception of the reference emotions (*Angry*, *Happy* and *Sad*) that are mixed with *Surprise*. As shown in Tables 3 a, 3 b and 3 c, the mixed emotion with 90% of the reference emotions consistently achieves the highest percentage of the "Best" score; also, the "Best" score increases as the percentage of reference emotion increases. Similarly, the highest "Worst" score is observed when the reference emotion is added at the lowest percentage (0%). These results confirm the effectiveness of controlling the rendering of mixed emotions. We also observe a slight rise of the worst rating when the percentage of *Happy* and *Sad* exceeds 60% in Tables 3 b, and 3 c. This observation we attribute to the unnatural emotional expressions that may be created to influence listeners' preferences.

We then take one step further to evaluate the perception of *Outrage*, *Delight* and *Disappointment* in synthesized speech. In psychology, there is evidence that those feelings could be produced by combining several emotions. We observe that participants can perceive such feelings, and most of them choose those with 90% of reference emotions as the "Best", as shown in Tables 4 a, 4 b and 4 c. As for the rating of "Worst", we also have similar observations to those in Table 3. These results show that we can synthesize new emotion types that are subtle and hard to collect in real life, which will significantly benefit the research community.

4.4 Ablation Study

We further conduct ablations studies to validate the contributions of the proposed relative scheme on emotional expression. We compare the proposed framework with or without the relative scheme through several XAB preference tests, where the participants are asked to listen to the reference emotional speech first, then choose the one closer to the reference in terms of emotional expression. 20 subjects listened to 60 speech samples in total (60 = 5 x 2 (# of frameworks) x 4 (# of emotions) + 20 (# of ground truth)).

We report the XAB results in Fig. 10 where we observe that "Proposed w/ Relative Scheme" consistently and considerably outperforms "Proposed w/o Relative Scheme" for all emotions (*Angry*, *Happy*, *Sad* and *Surprise*). Besides, the p values calculated between those two pairs ("Proposed w/ Relative Scheme" and "Proposed w/o Relative Scheme") are always lower than 0.05, indicating that the out-performance did not occur by chance. These results demonstrate that our relative scheme can improve emotional intelligibility in synthesized emotional speech.

5 Further Investigations and Discussion

In this section, we expand our experiments and show the ability of our proposed methods on other interesting topics. We

TABLE 4
Best-Worst Scaling (BWS) Test Results to Evaluate the Perception of Mixed Emotional Effects (*Outrage*, *Delight*, and *Disappointment*) in Synthesized Mixed Emotions

Configuration		Best (%)	Worst (%)
(a) Perception of *Outrage*			
Mixing Surprise (100%) with Angry	+ 0% Angry	6.8	61.7
	+ 30% Angry	4.5	23.3
	+ 60% Angry	15.8	9.8
	+ 90% Angry	**72.9**	**5.2**
(b) Perception of *Delight*			
Configuration		**Best (%)**	**Worst (%)**
Mixing Surprise (100%) with Happy	+ 0% Happy	5.3	66.2
	+ 30% Happy	10.5	21.8
	+ 60% Happy	30.1	**3.0**
	+ 90% Happy	**54.1**	9.0
(c) Perception of *Disappointment*			
Configuration		**Best (%)**	**Worst (%)**
Mixing Surprise (100%) with Sad	+ 0% Sad	11.3	54.9
	+ 30% Sad	12.0	26.4
	+ 60% Sad	14.3	**9.0**
	+ 90% Sad	**62.4**	9.7

first investigate the mixed emotional effects of *Happy* and *Sad*, which are two oppositely valenced emotions. We then build an emotion transition system with our proposed method. We do not seek to conduct comprehensive evaluations but to provide some interesting insights into mixed emotion synthesis and its applications. All the speech samples are provided on the demo page.

5.1 Oppositely Valenced Emotions: Happy and Sad

In our experiments, we mostly focus on mixing *Surprise* with other emotions (*Angry*, *Happy* and *Sad*), which is thought to be easier to perceive for human listeners. Here, we move one step further to study a more challenging task, which is to synthesize mixed effects of *Happy* and *Sad*. In Russell's valence-arousal model [92], *Happy* and *Sad* are two conflicting emotions with opposite valance (*Pleasant* and *Unpleasant*). There are some debates that agree with the co-existence of conflicting emotions [131], [132]. In real life, there are also some terms to describe such feelings in different cultures, for example, "*Bittersweet*" in English. Professional actors are thought to be able to deliver such feelings to the audience through both actions and speech. With our proposed methods, we are able to synthesize such mixed feelings of the oppositely valenced emotions such as *Happy* and *Sad*. Readers are suggested to refer to the demo page.

5.2 An Emotion Transition System

One potential application of mixed emotion synthesis is building an emotion transition system [133]. Emotion transition aims to gradually transition the emotion state from one to another. One similar study is emotional voice conversion [116], which aims to convert the emotional state. Compared with emotional voice conversion, the key challenge of emotion transition is to synthesize internal states between different emotion types. With our proposed methods, we are able to model these internal states by mixing them with different emotions. To achieve this, the sum up of the percentages of each emotion needs to be 100% (e. g., 80% *Surprise* with 20% *Angry*; 40% *Happy* with 60% *Sad*). Then, we can synthesize various internal emotion states by adjusting the percentages. Compared with traditional methods such as interpolation, our proposed system is data-driven, and the synthesized emotions are more natural.

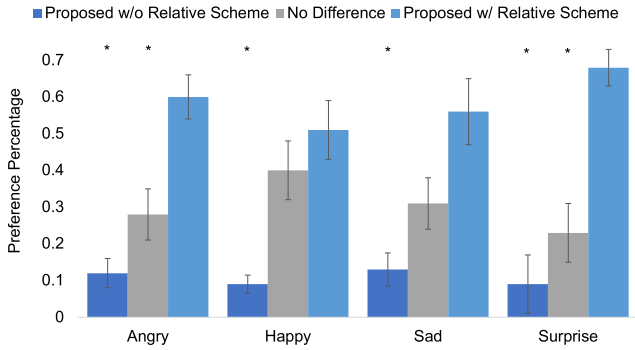


Fig. 10. XAB preference test results with 95% confidence interval to evaluate the emotion similarity with the ground truth emotions. The marker * indicates $p < 0.05$ for paired t-test scores (pairs between "Proposed w/ Relative Scheme" and the others).

5.3 Discussion

This study serves as the first attempt to model and synthesize mixed emotions for speech synthesis. Although we have shown the effectiveness of our methods, the related problems have not been completely solved. We provide a discussion to address the concerns, show our findings, and inspire future studies.

5.3.1 Category Versus Dimensional Emotion Models

Our assumptions, formulation, and evaluation of mixed emotions are all based on categorical emotion studies. We note that mixed emotions can also be modelled with dimensional representations such as arousal, valence, and dominance. A dimensional model can capture a wide range of emotional concepts, which offers a means of measuring the similarity of different emotional states [134]. However, several problems need to be adequately dealt with when modelling mixed emotions with a dimensional model. As mentioned in Section 3.1, the significant challenge for using dimensional representations comes from the lack of labels. Besides, humans are more efficient at discriminating among options than giving an absolute score [135], which adds challenges to the evaluation process. Furthermore, dimensional models are restricted to modelling the co-occurrence of like-valenced discrete emotions [136]. For these reasons, we refrain from applying dimensional emotions to the current framework.

5.3.2 Remaining Challenges

There are a few remaining challenges that need attention from the community. As mentioned in Section 4.3.2, increasing the percentage of adding emotions may result in unnatural emotional expressions. If the synthesized

emotion sounds unnatural or is difficult to understand, it may not be effective in achieving the desired outcome. Additionally, the human voice is a complex and highly variable instrument, and different people can produce the same emotional state in very different ways. This can make it difficult to accurately capture and reproduce a desired mix of emotions. At last, human raters are asked to evaluate the mixed emotions totally based on their personal experiences because of the lack of "ground truth" emotions. People from different cultures may have different experiences and backgrounds that can influence their emotional responses, and having a diverse group of evaluators can provide a more well-rounded perspective on the synthesized emotions.

5.3.3 Potential Improvements

We discuss several potential improvements to inspire future studies on mixed emotion synthesis: 1) Selection of ranking functions: adopt deep learning-based ranking methods [137] to improve the performance of ranking; 2) Multi-speaker studies: add training data from multiple speakers; 3) Non-autoregressive backbone frameworks: use non-autoregressive TTS framework as the backbone to avoid the misalignment of attention and improve the naturalness of synthesized speech.

6 Conclusion

This contribution fills the gap on mixed emotion synthesis in the literature on speech synthesis. We proposed an emotional speech synthesis framework that is based on a sequence-to-sequence model. For the first time, with the proposed framework, we are able to synthesize mixed emotions and further control the rendering of mixed emotions at run-time. The key highlights are as follows:

1) We proposed a novel relative scheme to measure the difference between each emotion pair. We demonstrate that our proposed relative scheme enables the effective synthesis and control of the rendering of mixed emotions. Through ablation studies, we also show that the proposed relative scheme improves emotional intelligibility in synthesized speech;
2) We presented a comprehensive study to evaluate mixed emotions for the first time. Through both objective and subjective evaluations, we validated our idea and showed the effectiveness of our proposed framework in terms of synthesizing mixed emotions;
3) We present further investigations on synthesising a bittersweet feeling and an emotion triangle. The investigation study serves as an additional contribution to the article, which could broaden the scope of the study.

In this article, we only focused on studying mixed emotions for emotional text-to-speech. We believe that our proposed relative scheme could enable mixed emotion synthesis in most existing emotional speech synthesis frameworks, including but not limited to emotional text-to-speech. We will expand our experiments to include emotional voice conversion in our future studies.

The future work includes: 1) a comparison with other ranking methods such as metric learning [138] and Siamese neural networks [137]; 2) conducting experiments for more emotion combinations, speakers, and other languages. Our future directions also include the study of cross-lingual emotion style modeling and transfer. Besides, a closer look at linguistic prosody for emotional speech synthesis is foreseen; for example, different semantic meanings can affect the way of expressing an emotion.

References

[1] A. Braniecka, E. Trzebińska, A. Dowgiert, and A. Wytykowska, "Mixed emotions and coping: The benefits of secondary emotions," *PLoS One*, vol. 9, no. 8, 2014, Art. no. e103940.
[2] J. T. Larsen and A. P. McGraw, "The case for mixed emotions," *Social Pers. Psychol. Compass*, vol. 8, no. 6, pp. 263–274, 2014.
[3] J. T. Larsen and A. P. McGraw, "Further evidence for mixed emotions," *J. Pers. Social Psychol.*, vol. 100, no. 6, 2011, Art. no. 1095.
[4] M. Schröder, "Emotional speech synthesis: A review," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 561–564.
[5] J. Pittermann, A. Pittermann, and W. Minker, *Handling Emotions in Human-Computer Dialogues*. Berlin, Germany: Springer, 2010.
[6] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *Int. J. Social Robot.*, vol. 8, no. 2, pp. 271–285, 2016.
[7] A. Rosenberg and J. Hirschberg, "Prosodic aspects of the attractive voice," in *Voice Attractiveness*. Berlin, Germany: Springer, 2021, pp. 17–40.
[8] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021, *arXiv:2106.15561*.
[9] J. P. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, *Progress in Speech Synthesis*. Berlin, Germany: Springer, 2013.
[10] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 132–157, 2021.
[11] J. Sotelo et al., "Char2wav: End-to-end speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2017.
[12] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.
[13] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.
[14] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7962–7966.
[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.
[16] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: Where do the improvements come from?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5505–5509.
[17] D. Schuller and B. W. Schuller, "The age of artificial emotional intelligence," *Computer*, vol. 51, no. 9, pp. 38–46, 2018.
[18] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to english," in *Proc. IEEE Speech Synth. Workshop*, 2002, pp. 227–230.
[19] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 274–278.
[20] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in DNN-based text-to-speech," *Speech Commun.*, vol. 126, pp. 35–43, 2021.
[21] R. Plutchik, *The Emotions*. Lanham, MD, USA: Univ. Press America, 1991.
[22] Y. Xu, "Speech prosody: A methodological review," *J. Speech Sci.*, vol. 1, no. 1, pp. 85–115, 2011.
[23] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 2274–2277.

[24] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2461–2464.
[25] F. Eyben et al., "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4009–4012.
[26] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *Amer. J. Signal Process.*, vol. 2, pp. 134–138, 2012.
[27] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. IEEE Conf. Trans. Speech Audio*, 2003, pp. 2401–2404.
[28] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, 2018.
[29] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," in *Proc. INTERSPEECH Conf.*, 2017, pp. 3399–3403.
[30] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. Interspeech Conf.*, 2016, pp. 2453–2457.
[31] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 1613–1616.
[32] R. Skerry-Ryan et al., "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.
[33] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 623–627.
[34] Y. Lee, S.-Y. Lee, and A. Rabiee, "Emotional end-to-end neural speech synthesizer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017.
[35] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional TTS," in *Proc. SAI Intell. Syst. Conf.*, 2019, pp. 52–60.
[36] Y. Wang et al., "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.
[37] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 595–602.
[38] S. D. Kreibig and J. J. Gross, "Understanding mixed emotions: Paradigms and measures," *Curr. Opin. Behav. Sci.*, vol. 15, pp. 62–71, 2017.
[39] R. Berrios, P. Totterdell, and S. Kellett, "Eliciting mixed emotions: A meta-analysis comparing models, types, and measures," *Front. Psychol.*, vol. 6, 2015, Art. no. 428.
[40] P. M. Niedenthal and M. Brauer, "Social functionality of human emotion," *Annu. Rev. Psychol.*, vol. 63, pp. 259–285, 2012.
[41] P. C. Hogan, *The Mind and its Stories: Narrative Universals and Human Emotion*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[42] P. E. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. London, U.K.: Oxford Univ. Press, 1994.
[43] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Amer. Scientist*, vol. 89, no. 4, pp. 344–350, 2001.
[44] R. Plutchik and H. Kellerman, *Theories of Emotion*, vol. 1. New York, NY, USA: Academic Press, 2013.
[45] M. Cross and C. Hanrahan, *Changing Minds: The Go-to Guide to Mental Health for Family and Friends*. New York, NY, USA: HarperCollins, 2016.
[46] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proc. IEEE 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 248–255.
[47] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 16–35, First Quarter 2021.
[48] J. B. Harvill, S.-G. Leem, M. Abdelwahab, R. Lotfian, and C. Busso, "Quantifying emotional similarity in speech," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3127390.
[49] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, 2015.
[50] J. Harvill, M. A. Wahab, R. Lotfian, and C. Busso, "Retrieving speech samples with similar emotional content using a triplet loss function," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7400–7404.
[51] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *Proc. Eur. Conf. Mach. Learn.*, 2003, pp. 145–156.
[52] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech Conf.*, 2016, pp. 490–494.
[53] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2108–2121, Nov. 2016.
[54] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 314–326, Third Quarter 2014.
[55] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 574–580.
[56] J. Huang et al., "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proc. Interspeech Conf.*, 2018, pp. 3673–3677.
[57] K. Feng and T. Chaspari, "A siamese neural network with modified distance loss for transfer learning in speech emotion recognition," 2020, *arXiv:2006.03001*.
[58] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 192–199.
[59] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 423–430.
[60] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2022.3175578.
[61] Y. Lei, S. Yang, X. Wang, and L. Xie, "MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 853–864, 2022.
[62] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[63] K. K. J. F. S. Kyle, K. A. C. Y. B. Jose, and S. M. Sotelo, "Char2wav: End-to-end speech synthesis," in *Proc. Int. Conf. Learn. Representations, Workshop*, 2017.
[64] D. M. Schuller and B. W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Rev.*, vol. 13, pp. 44–50, 2020.
[65] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," 2021, *arXiv:2103.16809*.
[66] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 785–788.
[67] A. Rabiee, T.-H. Kim, and S.-Y. Lee, "Adjusting pleasure-arousal-dominance for continuous emotional text-to-speech synthesizer," in *Proc. INTERSPEECH Conf.*, 2019, pp. 3693–3694.

[68] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5734–5738.

[69] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[70] S. Ma, D. Mcduff, and Y. Song, "Neural TTS stylization with adversarial and collaborative games," in *Proc. Int. Conf. Learn. Representations*, 2018.

[71] T. Cornille, F. Wang, and J. Bekker, "Interactive multi-level prosody control for expressive speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8312–8316.

[72] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *Proc. Interspeech Conf.*, 2019, pp. 4440–4444.

[73] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," in *Proc. Interspeech Conf.*, 2021, pp. 4673–4677.

[74] G. Zhang, Y. Qin, and T. Lee, "Learning syllable-level discrete prosodic representation for expressive speech generation," in *Proc. INTERSPEECH Conf.*, 2020, pp. 3426–3430.

[75] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," in *Proc. Interspeech Conf.*, 2021, pp. 811–815.

[76] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.

[77] S. Mozziconacci, "Prosody and emotions," in *Proc. Int. Conf. Speech Prosody*, 2002.

[78] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5911–5915.

[79] D. Tan and T. Lee, "Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement," in *Proc. Interspeech Conf.*, 2021, pp. 4683–4687.

[80] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6264–6268.

[81] G. Sun et al., "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6699–6703.

[82] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, *arXiv:1312.6114*.

[83] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6945–6949.

[84] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3331–3340.

[85] N. H. Frijda, A. Ortony, J. Sonnemans, and G. L. Clore, "The complexity of intensity: Issues concerning the structure of emotion intensity," 1992.

[86] K. Matsumoto, S. Hara, and M. Abe, "Controlling the strength of emotions in speech-like emotional sound generated by WaveNet," in *Proc. Interspeech Conf.*, 2020, pp. 3421–3425.

[87] B. Schnell and P. N. Garner, "Improving emotional TTS with an emotion intensity input from unsupervised extraction," in *Proc. 11th ISCA Speech Synth. Workshop*, 2021, pp. 60–65.

[88] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7254–7258.

[89] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "EMOQ-TTS: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6317–6321.

[90] C. M. Whissell, "The dictionary of affect in language," in *The Measurement of Emotions*. Amsterdam, The Netherlands: Elsevier, 1989, pp. 113–131.

[91] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, pp. 169–200, 1992.

[92] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.

[93] M. Schroder, "Expressing degree of activation in synthetic speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1128–1136, Jul. 2006.

[94] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.

[95] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 67–80, First Quarter 2017.

[96] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 503–510.

[97] A. Kovashka, D. Parikh, and K. Grauman, "WhittleSearch: Image search with relative attribute feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2973–2980.

[98] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Anal. Appl.*, vol. 18, no. 1, pp. 157–171, 2015.

[99] Q. Fan, P. Gabbur, and S. Pankanti, "Relative attributes for large-scale abandoned object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2736–2743.

[100] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.

[101] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," in *Proc. Interspeech Conf.*, 2019, pp. 1298–1302.

[102] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 540–552, 2019.

[103] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer learning from speech synthesis to voice conversion with non-parallel training data," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1290–1302, 2021.

[104] A. Polyak and L. Wolf, "Attention-based wavenet autoencoder for universal voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6800–6804.

[105] W. Ping et al., "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. Int. Conf. Learn. Representations*, 2018.

[106] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," 2018, *arXiv:1809.08895*.

[107] H.-T. Luong and J. Yamagishi, "Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 200–207.

[108] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech Conf.*, 2020, pp. 4676–4680.

[109] H.-T. Luong and J. Yamagishi, "NAUTILUS: A versatile voice cloning system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2967–2981, 2020.

[110] A. Black et al., "The festival speech synthesis system, version 1.4.2," Unpublished document available via, 2001. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival.html

[111] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Recognition-synthesis based non-parallel voice conversion with adversarial learning," in *Proc. Interspeech Conf.*, 2020, pp. 771–775.

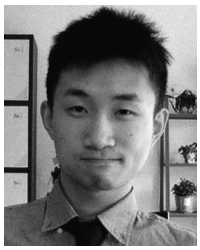
[112] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[113] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

[114] C. Veaux et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[115] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 920–924.

[116] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Commun.*, vol. 137, pp. 1–18, 2022.
[117] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization,".
[118] P. Heracleous, K. Yasuda, F. Sugaya, A. Yoneyama, and M. Hashimoto, "Speech emotion recognition in noisy and reverberant environments," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 262–266.
[119] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7194–7198.
[120] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, 2021, Art. no. 1249.
[121] H. Muthusamy, K. Polat, and S. Yaacob, "Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals," *Math. Problems Eng.*, vol. 2015, 2015, Art. no. 394083.
[122] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[123] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, no. 7/8, pp. 613–625, 2010.
[124] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128.
[125] W. F. Johnson, R. N. Emde, K. R. Scherer, and M. D. Klinnert, "Recognition of emotion from vocal cues," *Arch. Gen. Psychiatry*, vol. 43, no. 3, pp. 280–283, 1986.
[126] M. J. Owren and J.-A. Bachorowski, "Measuring emotion-related vocal acoustics," *Handbook of Emotion Elicitation and Assessment*. London, U.K.: Oxford Univ. Press, 2007, pp. 239–266.
[127] M. Morise et al., "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH Conf.*, 2017, pp. 2321–2325.
[128] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," in *Proc. Interspeech Conf.*, 2020, pp. 3416–3420.
[129] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Speaker Lang. Recognit. Workshop*, 2020, pp. 230–237.
[130] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 415–422.
[131] P. Williams and J. L. Aaker, "Can mixed emotions peacefully coexist?," *J. Consum. Res.*, vol. 28, no. 4, pp. 636–649, 2002.
[132] Y. Miyamoto, Y. Uchida, and P. C. Ellsworth, "Culture and mixed emotions: Co-occurrence of positive and negative emotions in japan and the united states," *Emotion*, vol. 10, no. 3, 2010, Art. no. 404.
[133] S. Hareli, S. David, and U. Hess, "The role of emotion transition for the perception of social dominance and affiliation," *Cogn. Emotion*, vol. 30, no. 7, pp. 1260–1270, 2016.
[134] S. PS and G. Mahalakshmi, "Emotion models: A review," *Int. J. Control Theory Appl.*, vol. 10, no. 8, pp. 651–657, 2017.
[135] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, 1956, Art. no. 81.
[136] L. F. Barrett, "Discrete emotions or dimensions? The role of valence focus and arousal focus," *Cogn. Emotion*, vol. 12, no. 4, pp. 579–599, 1998.
[137] G. Koch et al., "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Deep Learn. Workshop*, 2015.
[138] B. McFee and G. R. Lanckriet, "Metric learning to rank," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 775–782.



Kun Zhou (Student Member, IEEE) received the BEng degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018, and the MSc degree from the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, in 2019. He is currently working toward the PhD degree with the National University of Singapore. He was a visiting PhD student with The Center for Robust Speech Systems (CRSS), the University of Texas at Dallas, United States (2022). His research interests mainly focus on emotion analysis and synthesis in speech, including emotional voice conversion and emotional text-to-speech. He is the recipient of the PREMIA best student paper award 2022. He served on the organizing committee for IEEE ASRU 2019, SIGDIAL 2021, IWSDS 2021, O-COCOSDA 2021 and ICASSP 2022. He is a reviewer for multiple leading conferences and journals including INTERSPEECH, ICASSP, IEEE SLT, *IEEE Signal Processing Letters*, *Speech Communication* and *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

Berrak Sisman (Member, IEEE) received the PhD degree in electrical and computer engineering from the National University of Singapore in 2020, fully funded by A*STAR Graduate Academy under Singapore International Graduate Award (SINGA). She is currently working as a tenure-track assistant professor with the Erik Jonsson School Department of Electrical and Computer Engineering, the University of Texas at Dallas, United States. Prior to joining UT Dallas, she was a tenure-track faculty with the Singapore University of Technology and Design (2020-2022). She was a postdoctoral research fellow with the National University of Singapore (2019-2020), and a visiting researcher with Columbia University, New York, United States (2020). She was an exchange PhD student with the University of Edinburgh and a visiting scholar with The Centre for Speech Technology Research (CSTR), University of Edinburgh (2019). She was attached to RIKEN Advanced Intelligence Project, Japan (2018). Her research is focused on machine learning, signal processing, emotion, speech synthesis and voice conversion. She plays leadership roles in conference organizations and is also active in technical committees. She has served as the general coordinator of the Student Advisory Committee (SAC) of the International Speech Communication Association (ISCA). She has served as the area chair for INTERSPEECH 2021, INTERSPEECH 2022, IEEE SLT 2022 and as the publication chair for ICASSP 2022. She has been elected as a member of the IEEE Speech and Language Processing Technical Committee (SLTC) in the area of Speech Synthesis for the term from Jan. 2022 to Dec. 2024.

Rajib Rana (Member, IEEE) received the BSc degree in computer science and engineering from Khulna University, with the Prime Minister and President's Gold Medal for outstanding achievements, and the PhD degree in computer science and engineering from the University of New South Wales, Sydney, Australia, in 2011. He received his postdoctoral Training with the Autonomous System Laboratory, CSIRO, before joining the University of Southern Queensland, as a faculty member, in 2015. He is currently a senior advance queensland research fellow and an associate professor with the University of Southern Queensland. He is also the director of the IoT Health Research Program with the University of Southern Queensland, which capitalizes on advancements in technology and sophisticated information and data processing to understand disease progression in chronic health conditions better and develop predictive algorithms for chronic diseases, such as mental illness and cancer. His current research interests include unsupervised representation learning, Adversarial Machine Learning, Re-enforcement Learning, Federated Learning, Emotional Speech Generation, and Domain Adaptation.

Björn W. Schuller (Fellow, IEEE) received the diploma degree, the doctoral degree in automatic speech and emotion recognition, and the habilitation and adjunct teaching professor in signal processing and machine intelligence from Technische Universität München (TUM), Munich, Germany, in 1999, 2006, and 2012, respectively, all in electrical engineering and information technology. He is currently a professor of Artificial Intelligence with the Department of Computing, Imperial College London, U.K., where he heads the Group on Language, Audio, & Music (GLAM), a full professor and the head of the chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg, Germany, and the founding CEO/CSO of audEERING. He was previously a full professor and the head of the chair of Complex and Intelligent Systems at the University of Passau, Germany. He has (co)authored five books and more than 1 000 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 40,000 citations (H-index=97). He was an Elected member of the IEEE Speech and Language Processing Technical Committee. He is a Golden Core member of the IEEE Computer Society, a fellow of AAAC, BCS, and ISCA, as well as a senior member of the ACM, and the president-emeritus of the Association of the Advancement of Affective Computing (AAAC). He was the general chair of ACII 2019, a co-program chair of Interspeech, in 2019, and ICMI, in 2019, a repeated area chair of ICASSP, next to a multitude of further associate and guest editor roles and functions in Technical and Organisational Committees. He is the field chief editor of the Frontiers in Digital Health and a former editor-in-chief of the *IEEE Transactions on Affective Computing*.

Haizhou Li Fellow, IEEE) received the BSc, MSc, and PhD degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987 and 1990 respectively. He is currently a presidential chair professor and the executive dean with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also with the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition and natural language processing. Prior to joining NUS, he taught with the University of Hong Kong (1988-1990) and the South China University of Technology (1990-1994). He was a visiting professor with CRIN in France (1994-1995), research manager with the Apple-ISS Research Centre (1996-1998), research director with Lernout & Hauspie Asia Pacific (1999-2001), vice president with InfoTalk Corp. Ltd. (2001-2003) and the principal scientist and Department head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003-2016). He served as the editor-in-chief of *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2015-2018) and as a member of the Editorial Board of Computer Speech and Language (2012-2018). He was an elected member of the IEEE Speech and Language Processing Technical Committee (2013-2015), the president of the International Speech Communication Association (2015-2017), the president of the Asia Pacific Signal and Information Processing Association (2015-2016) and the president of the Asian Federation of Natural Language Processing (2017-2018). He was the general chair of ACL 2012, INTERSPEECH 2014, ASRU 2019 and ICASSP 2022. He is a fellow of the ISCA and the Academy of Engineering Singapore. He was a recipient of the National Infocomm Award in 2002 and the President's Technology Award in 2013 in Singapore. He was named one of the two Nokia visiting professors in 2009 by the Nokia Foundation, and Bremen Excellence chair professor in 2019.