Emotion intensity and its control for emotional voice conversion

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, Haizhou Li

Angaben zur Veröffentlichung / Publication details:

Zhou, Kun, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023. "Emotion intensity and its control for emotional voice conversion." *IEEE Transactions on Affective Computing* 14 (1): 31–48. https://doi.org/10.1109/taffc.2022.3175578.

Emotion Intensity and its Control for Emotional Voice Conversion

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 voice conversion (EVC) seeks to convert the emotional state of an utterance while preserving the linguistic content and speaker identity. In EVC, emotions are usually treated as discrete categories overlooking the fact that speech also conveys emotions with various intensity levels that the listener can perceive. In this paper, we aim to explicitly characterize and control the intensity of emotion. We propose to disentangle the speaker style from linguistic content and encode the speaker style into a style embedding in a continuous space that forms the prototype of emotion embedding. We further learn the actual emotion encoder from an emotion-labelled database and study the use of relative attributes to represent fine-grained emotion intensity. To ensure emotional intelligibility, we incorporate *emotion classification loss* and *emotion embedding similarity loss* into the training of the EVC network. As desired, the proposed network controls the fine-grained emotion intensity in the output speech. Through both objective and subjective evaluations, we validate the effectiveness of the proposed network for emotional expressiveness and emotion intensity control.

Index Terms—Emotional voice conversion, emotion intensity, sequence-to-sequence, perceptual loss, limited data, relative attribute

1 Introduction

Emotional Voice Conversion (EVC) is a technique that seeks to manipulate the emotional state of an utterance while keeping other vocal states unchanged [1]. It allows for the projection of the desired emotion into the synthesized voice. Emotional voice conversion poses a tremendous potential for human-computer interaction, such as enabling emotional intelligence into a dialogue system [2], [3], [4].

Voice conversion aims to convert the speaker-dependent vocal attributes such as the speaker identity while preserving the linguistic information [5]. Since the speaker information is characterized by the physical structure of the vocal tract and manifested in the spectrum [6], spectral mapping has been the main focus of voice conversion [7]. However, speech also conveys emotions with various intensity levels that can be perceived by the listener [8]. For example, happy can be perceived as happy or elation [9], while angry can be divided into a 'mild' angry and the 'full-blown' angry [10]. In particular, intensity of emotion is described as the magnitude of factor to attain the goal of the emotion [11]. Therefore, emotion intensity is not just the loudness of a voice, but correlates to all the acoustic cues that contribute to achieving an emotion [12]. Moreover, speech emotion is hierarchical and supra-segmental in nature, varying from syllables to utterances [13], [14], [15], [16]. Thus, it is insufficient to only focus on frame-wise spectral mapping for emotional voice conversion. Both intensity variations and prosodic dynamics need to be considered for speech emotion modelling.

Synthesizing various intensities of an emotion is a challenging task for emotional voice conversion studies. One of the reasons is the lack of explicit intensity labels in most emotional speech datasets. Besides, emotion intensity is even more subjective and complex than just considering discrete emotion categories, which makes it challenging to model [12]. There are generally two types of methods in the literature for emotion intensity control. One uses auxiliary features such as a state of voiced, unvoiced, and silence (VUS) [17], attention weights or a saliency map [18].

- *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 Singapore University of Technology and Design, Singapore 487372. 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 GLAM – The Group on Language, Audio, and 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 Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, and with the University of Bremen, 28359 Bremen, Germany, and also with the Chinese University of HongKong (Shenzhen), Shenzhen 518172, China. E-mail: haizhou.li@nus.edu.sg.*

Manuscript received 22 Jan. 2022; revised 16 Apr. 2022; accepted 13 May 2022. Date of publication 19 May 2022; date of current version 28 Feb. 2023.
*The work of Kun Zhou and Haizhou Li 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 through AME Programmatic Funding Scheme under Grant A18A2b0046, in part by National Research Foundation Singapore through AI Singapore Programme under Grant AISG-100E-2018-006, and in part by A*STAR through RIE2020 Advanced Manufacturing and Engineering Domain (AME) Programmatic Grant under Grant A1687b0033, Project Title: Spiking Neural Networks. The work of Berrak Sisman was supported in part by the Ministry of Education, Singapore, through MOE Tier 2 Funding Programme under Grant MOE-T2EP50220-0021, in part by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion under Grant SRG ISTD 2020 158, and in part by SUTD AI Grant – Thrust 2 Discovery by AI under Grant SGPAIRS1821.*
(Corresponding author: Kun Zhou.)
Recommended for acceptance by N. Asghar.
Digital Object Identifier no. 10.1109/TAFFC.2022.3175578

Another manipulates the internal emotion representations through interpolation [19] or scaling [20]. Despite these methods, emotion intensity control is still an under-explored topic in emotional voice conversion.

Previous emotional voice conversion studies mainly focus on learning a feature mapping between different emotion types. Most of them, model the mappings of spectral and prosody parameters with a Gaussian mixture model (GMM) [21], [22], sparse representation [23], or hidden Markov model (HMM) [24]. Recent deep learning methods such as deep neural networks (DNN) [25], [26] and deep bi-directional long-short-term memory network (DBLSTM) [27] have advanced the state-of-the-art. New techniques using generative adversarial network (GAN)-based [28], [29], [30] or auto-encoder-based models [31], [32], [33] make it possible for non-parallel training. We note that these frameworks convert the emotion on a frame basis, so speech duration cannot be modified. Moreover, since the spectrum and prosody are not independent of each other, a separate study of them may cause a mismatch during the conversion [34], [35]. It would be advantageous to have a model to transfer the correlated vocal factors end-to-end, producing more realistic emotions in synthetic speech.

Recently, sequence-to-sequence (Seq2Seq) models have attracted much interest in speech synthesis [36], [37] and voice conversion [38], [39], [40], [41]. With the attention mechanism, Seq2Seq frameworks jointly learn the feature mapping and alignment and automatically predict the speech duration at run-time. Inspired by these successful attempts, researchers introduce Seq2Seq modelling into emotional voice conversion. For example, a Seq2Seq model to jointly model pitch and duration is proposed in [42]. In [43], a multi-task learning for both emotional voice conversion and emotional text-to-speech is studied. We note two limitations of these studies: First, they learn an averaged emotional pattern during the training, while emotional expressive speech presents abundant variations of emotion intensity in real life. Second, these frameworks require enormous emotional speech data to train. *But in practice, such a large emotional speech database is not widely available, which limits the scope of applications.*

In this article, we aim to address the above challenges. The main contributions of this paper are listed as follows.

- We introduce *Emovox*, a Seq2Seq emotional voice conversion framework, which jointly transfers the spectrum and duration in an end-to-end way for emotional voice conversion.
- *Emovox* automatically learns the abundant variations of intensity that are exhibited in an emotional speech dataset, without the need for any explicit intensity labels and enables effective control of the emotion intensity in the converted emotional speech at the run-time;
- *Emovox* eliminates the need for a large amount of emotional speech data for Seq2Seq EVC training and still achieves remarkable performance under limited data conditions;
- We present a comprehensive evaluation to show the effectiveness of *Emovox* for emotional expressiveness and emotion intensity control.

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 *Emovox* framework and we introduce our experiments in Section 4. In Section 5, we report the experimental results and conclude in Section 6.

2 Background and Related Work

This work is built on several previous studies spanning emotion intensity, expressive speech synthesis, and emotional voice conversion. We briefly introduce the related studies to set the stage for our research and summarize the gaps in current literature to place our novel contributions.

2.1 Emotion Intensity in Vocal Expression

The most straightforward way to characterize emotion is to categorize it into several different groups [44], [45]; however, the choice of emotion labels is mostly intuitive and inconsistent in the literature. One key reason is that emotion intensity can affect our perception of emotions [46]. For example, happy can be perceived as happy or elation, which are similar in voice quality but different in intensity [9]. Thus, correlating the emotion intensity to the loudness of the voice is a rather oversimplification. Emotion intensity can be observed in various acoustic cues, not only in speech energy but also in speech rate and fundamental frequency [12]. The differences in these cue levels could be larger between different intensities of the same emotion than between different emotions [46].

2.2 Sequence-to-Sequence Conversion Models

The sequence-to-sequence model with attention mechanism was first studied in machine translation [50] and then found effective in speech synthesis [36], [37]. In text-to-speech, sequence-to-sequence modelling achieves remarkable performance by learning an attention alignment between the text and acoustic sequence, such as Tacotron [37]. Similar to text-to-speech, voice conversion aims to generate realistic speech from internal representations; therefore, sequence-to-sequence models are applied to various voice conversion and emotional voice conversion studies.

2.2.1 Sequence-to-Sequence Voice Conversion

Sequence-to-sequence voice conversion frameworks such as SCENT [38], AttS2S-VC [39], and ConvS2S-VC [51], jointly convert the duration and prosody components, and achieve higher naturalness and similarity than conventional frame-based methods. To address the conversion issues such as the deletion and repetition caused by the misalignment, various approaches are proposed, such as a monotonic attention mechanism [40], non-autoregressive training [52], [53], and the use of pre-training models [54] or text supervision [55], [56], [57]. These successful attempts further motivate the study of sequence-to-sequence modelling for emotional voice conversion.

2.2.2 Sequence-to-Sequence Emotional Voice Conversion

Compared with conventional frame-based models, sequence-to-sequence models are more suitable for emotional voice conversion. First, the sequence-to-sequence models allow for

the prediction of speech duration at the run-time, which is an important aspect of the speech rhythm and strongly affects the emotional prosody [58]. Besides, a joint transfer of spectrum and prosody in sequence-to-sequence models addresses the mismatch issues in conventional analysis-synthesis-based emotional voice conversion systems [28], [33], [34]. Also, emotional prosody is supra-segmental and can be only associated with a few words [47]. Learning an attention alignment makes it possible to focus on emotion-relevant regions during the conversion. Hence, sequence-to-sequence modelling for emotional voice conversion will be our primary focus in this paper.

There are only few studies on sequence-to-sequence emotional voice conversion [20], [42], [43], [59]. In [42], the authors jointly model pitch and duration with parallel data, where the model is conditioned on the syllable position in the phrase. In [43], a multi-task learning framework of emotional voice conversion and emotional text-to-speech is built with a large-scale emotional speech database. In [20], the authors introduce an emotion encoder and a speaker encoder into the sequence-to-sequence training for emotional voice conversion. We note that these frameworks require tens of hours of parallel emotional speech data, which is hard to collect. A recent work [59] proposes a 2-stage training strategy for sequence-to-sequence emotional voice conversion leveraging text-to-speech to eliminate the need for a large emotional speech database. However, none of these frameworks study emotion intensity variations, and the converted emotional utterances lack the controllability of emotion intensity. Only [20] attempts to scale the emotion embedding by multiplying it with a factor to control the emotion intensity at run-time. However, the authors do not explicitly model emotion intensity variations during the training, and their intensity control method lacks interpretability.

This work aims to bridge this gap in the current literature and study emotion intensity modelling for emotional voice conversion. We aim to build a sequence-to-sequence emotional voice conversion framework with effective emotion intensity control using a limited amount of emotional speech data.

2.3 Expressive Speech Synthesis With Prosody Style Control

Speech emotion is highly related to speech prosody and influenced by several prosodic cues embedded in acoustic speech such as intonation, rhythm, and energy [60], [61]. The most straightforward way to model and control prosody style is to use explicit annotations, or labels [62], [63], [64]. Besides explicitly labelling, researchers use a reference encoder to imitate and transplant the reference style in an unsupervised way [65]. Global style token (GST) [66] is an example to learn interpretable style embeddings from the reference audio. By choosing specific tokens, the model could control the style of synthesized speech. Other studies [67], [68], [69], [70] mainly replace the global style embedding with fine-grained prosody embedding. Some other studies based on Variational Autoencoders (VAE) [71] show the effectiveness of controlling the speech style by learning, scaling, or combining disentangled representations [72], [73].

Emotion expressive speech is even more complex, which has subtle dynamic variations associated with multiple prosodic attributes [74], [75], [76]. Inspired by the successful attempts in prosody style control, several studies control the emotion intensity for emotional speech synthesis. For example, in [19], an inter-to-intra distance ratio algorithm is applied to the learnt style tokens for emotional speech synthesis, where an interpolation technique is used to control emotion intensity. In [18], the authors show that a speech emotion recognizer is capable of generating a meaningful intensity representation via attention or saliency. In [77], [78], a relative attribute scheme is introduced to learn the emotion intensity for emotional speech synthesis. None of these frameworks explicitly models prosody style, but rather encodes the association between input text and its emotional prosody style end-to-end.

This contribution studies explicit modelling of emotion intensity variations with a relative attribute method for emotional voice conversion. We believe that the relative attributes scheme provides a straightforward way to model intensity variants, which will be discussed later.

2.4 Emotional Prosody Modelling With a Speech Emotion Recognizer

Emotional prosody is prominently exhibited in emotional expressive speech [79], [80], [81], which can be characterized by either categorical [45] or dimensional representations [82]. Recent studies [83] in speech emotion recognition provide valuable insights into emotional prosody modelling. Instead of categorical or dimensional attributes, they characterize the emotion styles with the latent representation learnt by the deep neural network as shown in Fig. 1. Compared with human-crafted features, deep features learnt by a speech emotion recognizer (SER) are data-driven and less dependent on human knowledge [84], [85], which we believe is more suitable for emotion style transfer.

Some studies are leveraging a speech emotion recognizer to improve the prosody modelling for expressive speech synthesis. In [86], an emotion recognizer is used to extract the style embedding for style transfer. In [49], a speech emotion recognizer is further used as the style descriptor to evaluate the style reconstruction performance. In [48], researchers use the deep emotional features from a pre-trained speech emotion recognizer to transfer both seen and unseen emotion styles. These studies show the capability of a speech emotion recognizer to describe emotion styles with their latent representations.

A speech emotion recognizer also shows a potential to supervise the emotional speech synthesis system to generate the speech with desirable emotion styles [87]. In [88], a reinforcement learning paradigm for emotional speech synthe-

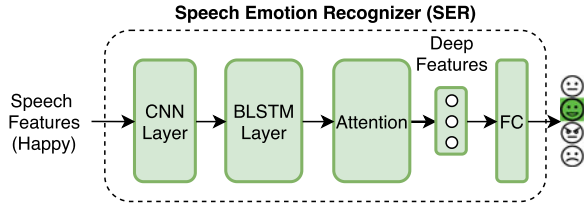


Fig. 1. An example of a speech emotion recognizer (SER) [47], where the deep features are obtained before the last fully-connected (FC) layer to describe the emotion styles [48], [49].

sis is proposed, where the classification accuracy of the speech emotion recognizer is used as the reward function to the system. In [89], the authors use emotion classifiers to enhance the emotion-discrimination of the emotion embedding and the predicted Mel-spectrum. In [90], an emotional speech synthesis system is built on an expressive TTS corpus with the assistance of a cross-domain emotion recognizer. These studies show remarkable performance by incorporating the supervision from the pre-trained emotion recognizer into the emotional speech synthesis systems, which motivates our study. We further study the use of perceptual losses in EVC training to improve the intelligibility of the converted emotion.

2.5 Research Gap (Summary)

Below, we summarise the gaps in the literature of emotional voice conversion that we aim to address in this paper.

1) There are very few studies on emotion intensity control, which is crucial to achieving emotional intelligence.
2) Despite the tremendous potential, emotion intensity control is still not a well-explored research direction for emotional voice conversion.
3) There is a lack of focus on modelling prosody style to achieve improved emotion intensity control.
4) Feasibility of using a pre-trained speech emotion recognizer as an emotion supervisor for EVC training poses tremendous potential but is not well understood.

3 *Emovox:* Emotional Voice Conversion With Emotion Intensity Control

The proposed emotional voice conversion framework: *Emovox* consists of four modules, 1) a *recognition encoder*, which derives the linguistic embedding from the source speech; 2) an *emotion encoder*, which encodes the reference emotion style into an emotion embedding; 3) an *intensity encoder*, which encodes a fine-grained intensity input into an intensity embedding, and 4) a *Seq2Seq decoder*, which generates the converted speech from a combination of linguistics, emotion, and intensity embeddings. At run-time, *Emovox* preserves the source linguistic content ("linguistic transplant"), while transferring the reference emotion to a source utterance ("emotion transfer"), as illustrated in Fig. 2. *Emovox* also allows users to manipulate/control the emotion intensity of the output speech ("intensity control").

To train *Emovox*, we propose a Seq2Seq framework that disentangles the speech elements from input acoustic features, and reconstructs the acoustic features from the speech elements. To reduce the amount of training data for *Emovox*, we introduce two pre-training strategies, i.e., 1) style pre-training with a large TTS corpus, and 2) emotion supervision training with an SER.

3.1 Seq2Seq Emotional Voice Conversion

Human speech can be viewed as a combination of speech style, and linguistic content [75], [91]. If the speech style that represents the emotion can be disentangled from the linguistic content, emotion conversion can be achieved by manipulating the speech style at run-time while keeping the linguistic content and speaker identity unchanged [33], [92].

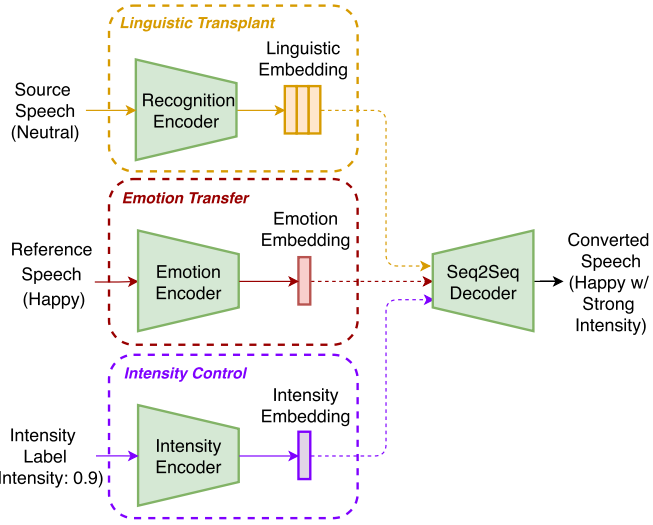


Fig. 2. Block diagram of *Emovox* during the conversion stage. *Emovox* aims to transfer the reference emotion to the source speech ("emotion transfer") while controlling its emotion intensity ("intensity control") and preserving the source linguistic information ("linguistic transplant").

There are various ways to disentangle the speech elements. In [56], text information and adversarial learning are used in a sequence-level autoencoder. This framework achieves strong disentanglement between linguistic and speaker representations and enables duration modelling for voice conversion. We adopt this framework in *Emovox* to model emotion styles and intensity, as shown in Fig. 3. To overcome the issues such as deletion and repetition with the Seq2Seq approach, we include a text input as the supervision signal to augment the linguistic embedding, which are shown effective in recent studies [56], [57], [59].

Given the phoneme sequences and acoustic features as the input, the text encoder and the recognition encoder learn to predict the linguistic embedding from the text ($\mathbf{H}^{text}$) and the audio input ($\mathbf{H}^{audio}$), respectively. The emotion encoder learns the emotion representations from the speech, while the emotion classifier further eliminates the residual emotion information in the linguistic embedding $\mathbf{H}^{audio}$. The Seq2Seq decoder Dec learns to reconstruct the acoustic features $\hat{\mathbf{A}}$ from the combination of the emotion embedding $\mathbf{h}^{emo}$, the intensity embedding $\mathbf{h}^{inten}$, and the linguistic embedding either from the text encoder: $\mathbf{H}^{text}$ or recognition encoder: $\mathbf{H}^{audio}$ as shown in (1).

$$\hat{A} = Dec(\mathbf{h}^{emo}, \mathbf{h}^{inten}, f(epoch)), \tag{1}$$

where

$$f(epoch) = \begin{cases} \mathbf{H}^{text} & \text{for} \quad epoch\%2 = 0 \\ \mathbf{H}^{audio} & \text{for} \quad epoch\%2 = 1 \end{cases}. \tag{2}$$

During the training, $\mathbf{H}^{text}$ and $\mathbf{H}^{audio}$ are taken by the decoder alternately, depending on whether the epoch number is odd or even. A contrastive loss is employed to ensure the similarity between $\mathbf{H}^{text}$ and $\mathbf{H}^{audio}$ as in [56]. We believe the proposed *Emovox* learns an effective disentanglement between linguistic and emotional elements and provides a

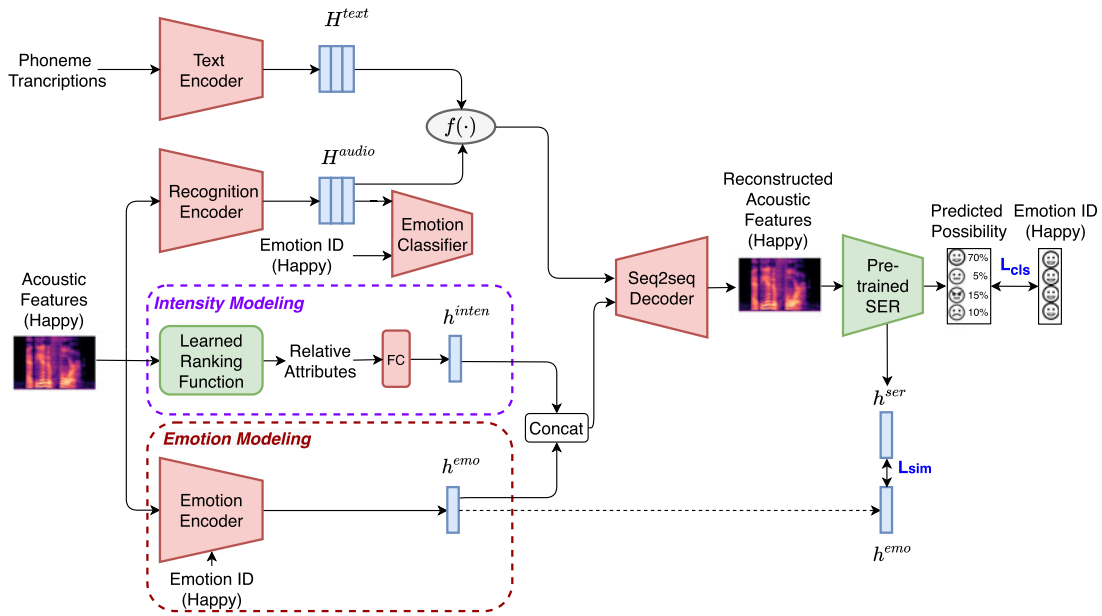


Fig. 3. Overall training diagram of *Emovox*, where emotion and its intensity are separately modelled. Two utterance-level perceptual losses from a pre-trained SER: 1) emotion similarity loss L_{sim}, and 2) emotion classification loss L_{cls} are introduced to improve the emotional intelligibility at the utterance level. The red boxes represent the models that are involved in the training, while the green boxes are not.

straightforward way to model and control both emotion and its intensity, which will be discussed next.

3.2 Modelling Emotion and its Intensity

To model emotion intensity, one of the difficulties is the lack of annotated intensity labels. Inspired by the idea of attribute [93] in computer vision, we regard emotion intensity as an attribute of the emotional speech. Combining the emotion representations with the intensity information allows the framework to jointly learn abundant emotion styles and intensity levels from any emotional speech database.

3.2.1 *Formulation of Emotion Intensity Using Relative Attributes*

In computer vision, there are various ways [94], [95] to model the relative difference between different data categories. Instead of predicting the presence of a specific attribute, relative attributes [96] offer more informative descriptions to unseen data, thus closer to detailed human supervision. Motivated by the success in various computer vision tasks [97], [98], [99], we believe that relative attributes bridge between the low-level features and high-level semantic meanings, which is appropriate for emotion intensity modelling.

Emotion intensity can be viewed as how well the emotion can be perceived in its type. Since the neutral speech does not contain any emotional variance, the emotion intensity of a neutral utterance should be zero. Therefore, we regard the emotion intensity as a relative difference between neutral speech and emotional speech. Emotion intensity can be represented by relative attributes learnt with a rich set of emotion-related acoustic features from each emotion pair. The learning process of relative attributes can be formulated as a max-margin optimization problem as explained below:

Given a training set $T = \{\mathbf{x}_t\}$, where $\mathbf{x}_t$ is the acoustic features of the t^{th} training sample, and $T = N \cup E$, where N and E are the neutral and emotional set respectively. We aim to learn a ranking function given as below

$$r(x_t) = \mathbf{W}\mathbf{x}_t, \tag{3}$$

where $\mathbf{W}$ is a weighting matrix indicating the emotion intensity. To learn the ranking function, we have to satisfy the following constraints:

$$\forall (a,b) \in O : \mathbf{W}\mathbf{x}_a > \mathbf{W}\mathbf{x}_b \tag{4}$$

$$\forall (a,b) \in S : \mathbf{W}\mathbf{x}_a = \mathbf{W}\mathbf{x}_b, \tag{5}$$

where O and S are the ordered and similar sets respectively. We pair an emotional sample of E with a neutral sample from N to form an ordered set O, where the emotion intensity of E is higher than in that of N. We then randomly create pairs of neutral-neutral and emotional-emotional samples in the similar set S, where the emotion intensity of the pair is similar. The weighting matrix $\mathbf{W}$ is estimated by solving the following problem similar with that of a support vector machine [100]:

$$\min_{\mathbf{W}} \left(\frac{1}{2} \parallel \mathbf{W} \parallel_2^2 + C\left(\sum \xi_{a,b}^2 + \sum \gamma_{a,b}^2 \right) \right) \tag{6}$$

$$\text{s.t. } \mathbf{W}(\mathbf{x}_a - \mathbf{x}_b) \geq 1 - \xi_{a,b}; \forall (a,b) \in O \tag{7}$$

$$|\mathbf{W}(\mathbf{x}_a - \mathbf{x}_b)| \leq \gamma_{a,b}; \forall (a,b) \in S \tag{8}$$

$$\xi_{a,b} \leq 0; \gamma_{a,b} \leq 0, \tag{9}$$

where C is the trade-off between the margin and the size of slack variables $\xi_{a,b}$ and $\gamma_{a,b}$. Through Eqs. (6), (7), (8), and (9), we learn a wide-margin ranking function that enforces the desired ordering on each training point. Once it is learnt, the relative ranking function can estimate the order of unseen data. In practice, we learn a ranking function for each emotion category. As shown in Fig. 3, the learnt ranking function

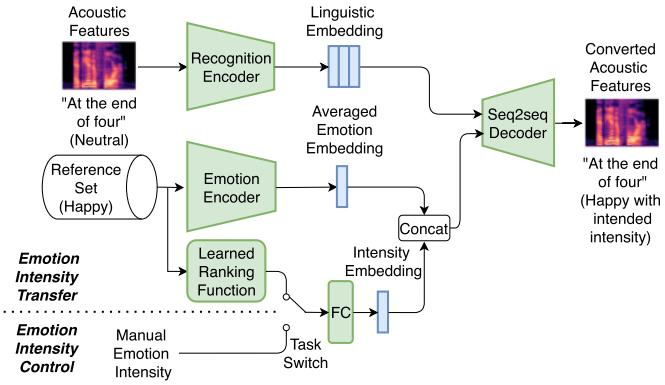


Fig. 4. An illustration of the run-time conversion phase. By combining a source linguistic embedding sequence, an averaged reference emotion embedding, and an intensity embedding, the Seq2Seq decoder generates the acoustic features with the reference emotion type and the manually defined intended intensity.

predicts a relative attribute normalized to [0,1] for each sample in the training set. A larger value of relative attribute represents a stronger intensity of an emotion.

3.2.2 Modelling Emotion Styles and its Intensity

As shown in Fig. 3, we obtain the relative attribute from the learnt ranking function, which passes through a fully connected layer to derive an intensity embedding. The emotion encoder learns to generate the emotion embedding from the input speech features. The Seq2Seq decoder combines a linguistic embedding sequence, an emotion and an intensity embedding to reconstruct the acoustic features of the emotional speech.

During the training process, *Emovox* jointly learns the emotion style and its intensity from the speech samples that are referred to as *emotion training* hereafter. With the explicit intensity modelling, we are able to manipulate the level of intensity at run-time for intensity control. The intended emotion intensity can be predicted from the reference or given manually at run-time. In theory, *Emovox* may perform both emotional text-to-speech and emotional voice conversion. In this paper, the text encoder is not used at run-time since we are only interested in voice conversion.

As shown in Fig. 4, we first use the emotion encoder to generate the emotion embeddings from a set of reference utterances belonging to the same emotional category. Next, we use the averaged reference emotion embedding to represent an emotion category. Finally, the recognition encoder derives a linguistic embedding sequence from the source speech utterance at run-time. By assigning an intended emotion category and a level of emotion intensity, the Seq2Seq decoder generates the emotional speech of the same content as the source but with the target emotion style at an appropriate intensity.

3.3 Model Pre-Training

During training, a large amount of emotional speech is always required to achieve robust attention alignment and deliver high emotional intelligibility in a Seq2Seq model [101]. To reduce the reliance on emotional speech, we propose two pre-training strategies, 1) style pre-training with a large TTS corpus and 2) emotion supervision training with a SER.

3.3.1 Style Pre-Training With a Multi-Speaker TTS Corpus

It is known that speech style contains speaker-dependent elements related to speaker characteristics, called speaker style. Speaker style is exhibited in most TTS corpora containing multi-speaker speech data. Unlike emotional speech databases, there are abundant speech databases for TTS [102], [103], [104] with a neutral tone, which allows us to build a multi-speaker Seq2Seq TTS framework, and train a network to disentangle speaker style from the linguistic content. We call this stage "style pre-training".

During the style pre-training, the style encoder learns abundant speaker styles through a multi-speaker TTS corpus while excluding the linguistic information from the acoustic features. As a result, even though the style encoder does not learn to encode any specific emotion style during training, it learns to discriminate different emotion styles during emotion training, as shown in Fig. 5a. We, therefore, use the style encoder trained on a TTS corpus as the pre-trained model for an emotion encoder.

3.3.2 Modelling Emotion With Perceptual Loss

We would like the converted emotional speech to be perceived with the intended emotion category. However, this is not easily achieved, especially with a limited emotional training data, for several reasons: 1) The pre-trained emotion decoder in Fig. 3 is not explicitly trained for characterisation of emotions, and 2) frame-level style reconstruction loss is not always consistent with human perception because it does not capture speech's prosodic and temporal patterns.

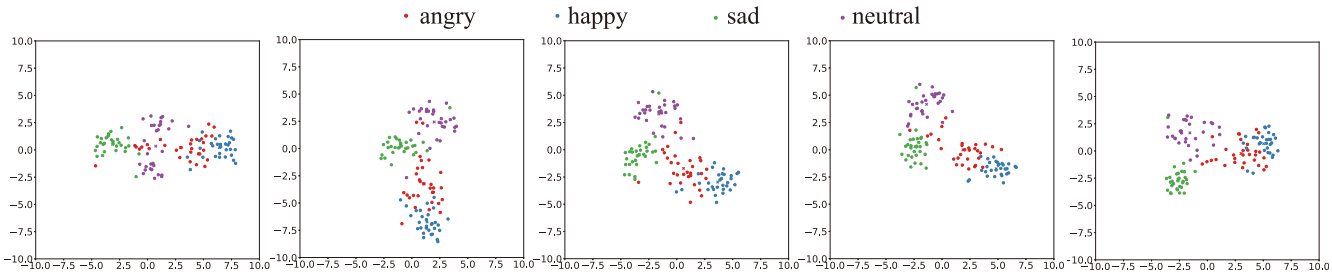
Following the success of perceptual loss in speech synthesis [105], we introduce a perceptual loss as the emotion supervision in the training process, as shown in Fig. 3. We first use a pre-trained SER to predict the emotion category from the reconstructed acoustic features. We then calculate two perceptual loss functions: 1) emotion classification loss L_{cls}, and 2) emotion embedding similarity loss L_{sim}. We incorporate these two loss functions into the training and update all the trainable modules. For detail of SER pre-training, readers are referred to [106].

The emotion classification loss L_{cls} is introduced to ensure the perceptual similarity between the reconstructed acoustic features and the intended emotion category at the utterance level,

$$L_{cls} = \mathrm{CE}(\mathbf{l}, \hat{\mathbf{p}}), \tag{10}$$

where $\mathbf{l}$ is the target one-hot emotion label, $\hat{\mathbf{p}}$ is the predicted emotion probabilities at the utterance level, and $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss function. The pre-trained SER is considered text-independent. To ensure that the emotion encoder characterizes emotions independent of linguistic content, we introduce an emotion style descriptor, derived from the pre-trained SER [48], [49], as a learning objective for emotion encoder with a loss function L_{sim}, between the emotion encoder output, i.e., emotion embedding, and the emotion style descriptor, as illustrated in Fig. 3.

$$L_{sim} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\mathbf{h}_d^{emo} - \mathbf{h}_d^{ser})^2}, \tag{11}$$



Ratio: r = 0.568 (a) Style Encoder

Ratio: r = 0.552 (b) Emotion Encoder (Emovox w/o intensity, w/o L_{sim} or L_{cls})

Ratio: r = 0.549 (c) Emotion Encoder (Emovox w/o intensity, w/ L_{sim})

Ratio: r = 0.529 (d) Emotion Encoder (Emovox w/o intensity, w/ L_{cls})

Ratio: r = 0.514 (e) Emotion Encoder (Emovox w/o intensity, w/ L_{sim} and L_{cls})

Fig. 5. The distributions of emotion embeddings resulting from encoders of different training schemes: (a) the pre-trained style encoder, (b) the emotion encoder without L_{sim} or L_{cls}, (c) the emotion encoder only with L_{cls}, (d) the emotion encoder only with L_{sim}, and (e) the emotion encoder with both L_{sim} and L_{cls}. A smaller value of ratio r indicates a better clustering performance.

where $\mathbf{h}_d^{emo}$ is the emotion embedding derived from the emotion encoder of D dimensions, and $\mathbf{h}_d^{ser}$ is the emotion style descriptor derived right before the last projection layer of the pre-trained SER.

3.3.3 Effect of Perceptual Loss

To validate the effectiveness of two perceptual loss functions, we evaluate the emotion-discriminative ability of the emotion encoder with an ablation study. We believe that the emotion encoder demonstrates a better performance by producing more discriminative emotional representations.

We use the t-SNE algorithm [107] to visualize the emotion embeddings in a two-dimensional plane, as shown in Fig. 5e. It is observed that emotion embeddings form different emotion clusters in terms of the feature distribution. To get a more intuitive understanding of the clustering performance, we consider performing a clustering evaluation to evaluate the discriminability of the emotion embeddings.

The typical objective function of clustering formalizes the goal of attaining high intra-cluster similarity, and low inter-cluster similarity [108], [109]. There are studies to use different measurements for the quality of a clustering [110], [111], [112], [113]. Our study considers a simplified and effective solution for clustering evaluation. We first compute a centroid for each of K emotion classes, $\mathbf{c}_i$, $i \in [1, K]$ by taking the average of all N_i embeddings $\mathbf{e}$ in class i as follows [114]:

$$\mathbf{c}_i = \frac{1}{N_i} \sum_{\mathbf{e} \in E_i} \mathbf{e}, \tag{12}$$

where E_i is the set of embeddings in class i. We then calculate the inter-class distance $dist_{inter}$ by computing the euclidean distance between each embedding $\mathbf{e} \in E_i$ and the other embedding centres $\mathbf{c}_{j;j \neq i}$ as follows:

$$dist_{inter} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \frac{1}{N_i} \sum_{\mathbf{e} \in E_i} \sum_{j;j \neq i} \sqrt{(\mathbf{e} - \mathbf{c}_j)^2}, \tag{13}$$

and intra-class distance $dist_{intra}$ as follows:

$$dist_{intra} = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{N_i} \sum_{\mathbf{e} \in E_i} \sqrt{(\mathbf{e} - \mathbf{c}_i)^2}. \tag{14}$$

A clustering ratio r is calculated from the ratio of intra-class distance $dist_{intra}$ and inter-class distance $dist_{inter}$ as follows:

$$r = \frac{dist_{intra}}{dist_{inter}}. \tag{15}$$

A lower value of ratio r represents a better clustering effect of emotion embeddings.

We perform an ablation experiment on the ESD evaluation dataset [1]. We visualize the distribution of emotion embeddings and report the clustering ratios in Fig. 5. As the style encoder is pre-trained without the emotion intensity mechanism, we report the results of *Emovox* without intensity control for a fair comparison, which is denoted as *Emovox w/o intensity*. We first observe that *Emovox w/o intensity* always achieves a better clustering performance than the style encoder in Fig. 5. From Figs. 5b, 5c, and 5d, it is observed that both loss functions L_{cls} and L_{sim} contribute to a lower r, which suggests a better clustering performance. From Fig. 5e, we further observe a more distinct separation between the emotions with different energy (such as neutral, sad versus angry, or happy). With both L_{cls} and L_{sim}, we obtain the lowest clustering ratio at 0.514. It shows that these two losses can help the emotion encoder to generate more discriminative emotional representations.

4 Experiments

In this section, we report our experimental settings. For all the experiments, we conduct emotion conversion from neutral to angry, neutral to happy, and neutral to sad, which we denote as *Neu-Ang*, *Neu-Hap*, and *Neu-Sad*, respectively. We have made the source codes and speech samples available to the public.[1] We encourage readers to listen to the speech samples to understand this work.

4.1 Reference Methods and Setups

We implement 3 state-of-the-art emotional voice conversion methods as the reference baselines, that are summarized as follows:

1. **Codes & Speech Samples**: https://kunzhou9646.github.io/Emovox_demo/

- *CycleGAN-EVC* [28] *(baseline):* CycleGAN-based emotional voice conversion with WORLD vocoder [115], where the fundamental frequency (F0) is analyzed with continuous wavelet transform;
- *StarGAN-EVC* [30] *(baseline):* StarGAN-based emotional voice conversion with WORLD vocoder [115];
- *Seq2Seq-EVC* [59] *(baseline):* Sequence-to-sequence emotional voice conversion with a Parallel WaveGAN vocoder [116];
- *Emovox (proposed):* Our proposed sequence-to-sequence emotional voice conversion framework with a Parallel WaveGAN vocoder [116] shown in Fig. 3.

Note that emotion intensity control is only available with *Emovox*. For a fair comparison among the methods, we obtain an intensity value for *Emovox*, by passing a reference set of speech data through the learnt ranking function, as shown in Fig. 4 ("Emotion Intensity Transfer"). Besides, none of these frameworks require any parallel training data or frame alignment procedures.

For a contrastive study, we replace the intensity control module in *Emovox* with two other competing intensity control methods: Scaling Factor and Attention Weights through comprehensive experiments.

- *Emovox w/ Scaling Factor (proposed):* where the emotion embedding is multiplied by a scaling factor [20];
- *Emovox w/ Attention Weights (proposed):* where the attention weight vector obtained from a pre-trained SER is used to represent the intensity [18];
- *Emovox w/ Relative Attributes (proposed):* our proposed method with relative attributes as described in Section 3;

To summarize, we do emotion intensity transfer to compare Emovox with the baselines (i.e., CycleGAN-EVC, StarGAN-EVC and Seq2Seq-EVC) and emotion intensity control to compare it with other emotion intensity control methods (i.e., Emovox w/ scaling factor, Emovox w/ attention weights).

4.2 Experimental Setup

We extract 80-dimensional Mel-spectrograms every 12.5 ms with a frame size of 50 ms for short-time Fourier transform (STFT). We then take the logarithm of the Mel-spectrograms to serve as the acoustic features. We convert text to phoneme with the Festival [117] G2P tool to serve as the input to the text encoder.

We use the Adam optimizer [118] and set the batch size to 64 and 16 for style pre-training and emotion training, respectively. We set the learning rate to 0.001 for style pre-training and halve it every seven epochs during the emotion training. We set the weight decay to 0.0001, and the weighting factors of the emotion classification loss L_{cls} and the emotion similarity loss L_{sim} to 1.

4.2.1 Recognition-Synthesis Structure

Emovox has a recognition-synthesis structure similar to that of [56], [119]. The Seq2Seq recognition encoder consists of an encoder which is a 2-layer 256-cell BLSTM, and a decoder which is a 1-layer 512-cell LSTM with an attention layer followed by an FC layer with an output channel of 512. Our text encoder is a 3-layer 1D CNN with a kernel size of 5 and the channel number of 512, followed by 1-layer of 256-cell BLSTM and an FC layer with an output channel number of 512. The Seq2Seq decoder has the same model architecture as that of Tacotron [37]. The style encoder is a 2-layer of 128-cell BLSTM followed by an FC layer with an output channel number of 128, which has been used in previous studies on voice conversion [56] and emotional voice conversion [59]. The classifier is a 4-layer net of FC with the channel numbers of $\{512, 512, 512, 99\}$.

4.2.2 Relative Emotion Intensity Ranking

We follow an open-source implementation[2] to train the relative ranking function for emotion intensity. We extract 384-dimensional acoustic features with openSMILE [120] including zero-crossing rate, frame energy, pitch frequency, Mel-frequency cepstral coefficient (MFCC), and etc. These acoustic features are used in the Interspeech Emotion Challenge [121]. We anticipate that these acoustic features can capture the subtle emotion intensity variations in speech. For each emotion category, we train a relative ranking function using neutral and emotional utterances.

4.2.3 Speech Emotion Recognizer

We train a speech emotion recognizer following a publicly available implementation [106]. The SER includes: 1) 3 TimeDistributed two-dimensional (2-D) convolutional neural network (CNN) layers, 2) a DBLSTM layer, 3) an attention layer, and 4) a linear projection layer. The TimeDistributed 2D CNN layers and the DBLSTM layer summarize the temporal information into a fixed-length latent representation. The attention layer further preserves the effective emotional information while reducing the influence of emotion-irrelevant factors and producing discriminative utterance-level features for emotion prediction. The linear projection layer predicts the emotion class possibility from the utterance-level emotional features. We perform data augmentation by adding white Gaussian noise to improve the robustness of SER ([122], [123], [124], [125]).

4.2.4 Data Preparation and Emotion Training

We first perform style pre-training on the VCTK Corpus [104], where we use 99 speakers for pre-training. The total duration of pre-training speech data is about 30 hours. For SER training and emotion training, we randomly choose one male speaker from the ESD database[3] to conduct all the experiments in the same way as in [59]. We follow the data partition protocol given in the ESD database. For each emotion, we use 300 utterances for emotion training and 20 utterances as the evaluation set. We use 30 utterances to form a reference set to generate the reference emotion embeddings for each emotion category at run-time. The total speech duration of emotional training data is around *50 minutes* (about 12 minutes for each emotion), which is very limited in the context of Seq2Seq training.

In the emotion training, we first initialize all the modules with the weights learnt from style pre-training, where the

2. [Online]. Available: https://github.com/chaitanya100100/Relative-Attributes-Zero-Shot-Learning

3. [Online]. Available: https://hltsingapore.github.io/ESD/download.html

style encoder and style classifier act as the emotion encoder and emotion classifier, respectively. We then randomly initialize the last projection layer of the emotion encoder and emotion classifier. The output channel numbers of the emotion encoder and the emotion classifier are set to 64 and 4, respectively. A learnt ranking function predicts a relative attribute and then is passed through an FC layer with the output channel size of 64 to obtain the intensity embedding. We then concatenate the emotion and intensity embedding to feed into the Seq2Seq decoder. The waveform is reconstructed from the converted Mel-spectrograms using Parallel WaveGAN. We use a public version of Parallel WaveGAN,[4] and train it with the ESD database.

4.3 Objective Evaluation

We first conduct an objective evaluation to assess the system performance using Mel-cepstral Distortion (MCD) and Differences of Duration (DDUR) as the evaluation metrics,

4.3.1 Mel-Cepstral Distortion (MCD)

MCD [126] is calculated between the converted and the target Mel-cepstral coefficients (MCEPs), i.e., $\hat{\mathbf{y}} = \{\hat{\mathbf{y}}_m\}$ and $\mathbf{y} = \{\mathbf{y}_m\}$,

$$\text{MCD [dB]} = \frac{10\sqrt{2}}{\ln 10}\frac{1}{M}\sqrt{\sum_{m=1}^{M}(\mathbf{y_m} - \hat{\mathbf{y}}_m)^2}, \tag{16}$$

where M represents the dimension of the MCEPs. *A lower value of MCD indicates a smaller spectral distortion, and thus a better performance*.Note that, in the Seq2Seq-EVC and *Emovox* models, we adopt Mel-spectrograms as the acoustic features. Therefore, we calculate MCEPs separately from the speech waveform.

4.3.2 Differences of Duration (DDUR)

To evaluate the distortion in terms of duration, we compute the average differences between the duration of the converted and the target utterances over the voiced parts (DDUR), which is widely used in voice conversion studies [38], [56], [59],

$$\text{DDUR [s]} = |Z - \hat{Z}|, \tag{17}$$

where Z and $\hat{Z}$ represent the duration of the reference utterance and the converted utterance, respectively. *A lower value of DDUR represents a better performance in terms of duration conversion*.

4.4 Subjective Evaluation

We adopt two subjective metrics: 1) a mean opinion score (MOS) test for emotion similarity evaluation, and 2) a best-worst scaling (BWS) test to evaluate speech quality, emotion intensity, and emotion similarity. 18 subjects participated in all the listening tests. These 18 subjects (12 male and 6 female) are native Chinese speakers and proficient in English. Their age range is between 20-30. All the subjects are required to listen with headphones and replay each sample 2-3 times. A detailed introduction about the judging criteria is given before the tests.

4.4.1 Mean Opinion Score (MOS) Test

We conduct a mean opinion score (MOS) [127] test to evaluate the emotion similarity. All participants are asked to listen to the reference target speech first and then score the speech samples for emotion similarity to the reference target speech. A higher score represents a higher similarity with the target emotion, and indicates a better emotion conversion performance. We randomly select 10 utterances from the evaluation set. Each subject listens to 120 converted utterances in total (120 = 10 x 4 (# of frameworks) x 3 (# of emotion pairs)).

4.4.2 Best-Worst Scaling (BWS) Test

We also conduct a best-worst scaling (BWS) [128] test to evaluate

1) *Speech Quality*: where all the listeners are asked to choose the best and the worst sample in terms of the speech quality, which covers two aspects: a) how the linguistic and speaker identity is preserved, and b) the naturalness of the speech;
2) *Emotion Intensity*: where all the listeners are asked to choose the most and the least expressive one in terms of the emotion expression;
3) *Emotion Similarity*: where all the listeners are asked to choose the best and the worst one in terms of the emotion similarity with the reference.

We randomly select 5 utterances from the evaluation set to perform the BWS tests. We first evaluate the performance of different intensity control methods in terms of speech quality and intensity control. Each subject listens to 135 converted utterances (135 = 5 x 3 (# of frameworks) x 3 (# of intensities) x 3 (# of emotion pairs)). We further conduct an ablation study with *Emovox*, where each subject listens to 60 converted utterances in total to evaluate the emotion similarity with the reference (60 = 5 x 4 (# of frameworks) x 3 (# of emotion pairs)).

5 Results

In this section, we report our experimental results. We first compare the performance of *Emovox* with that of the baselines using objective and subjective evaluations in Section 5.1. We then evaluate the proposed emotion intensity control method through the comparison with other control methods in Section 5.2. While comparing with the baselines, we use different training data settings in Section 5.3. Lastly, we study the contributions of the training strategies using ablation experiments in Section 5.4.

5.1 Emovox versus Baselines

In this subsection, we include CycleGAN-EVC, StarGAN-EVC, and Seq2Seq-EVC as baselines. It is noted that these baselines do not have an intensity control module. As a fair comparison, we conduct emotion intensity transfer for *Emovox* in both objective and subjective evaluations.

4. https://github.com/kan-bayashi/ParallelWaveGAN

TABLE 1
A Comparison of the MCD and the DDUR Results of Different Methods for Three Emotion Conversion Pairs

Framework	MCD [dB]			DDUR [s]		
	Neu-Ang	Neu-Hap	Neu-Sad	Neu-Ang	Neu-Hap	Neu-Sad
Zero Effort	6.47	6.64	6.22	0.36	0.26	0.46
CycleGAN-EVC	4.57	4.46	4.32	-	-	-
StarGAN-EVC	4.43	4.25	4.31	-	-	-
Seq2Seq-EVC	4.29	4.16	**4.23**	0.28	0.20	**0.27**
Emovox (w/o style pre-training)	5.36	5.32	5.42	0.79	0.80	0.92
Emovox	**4.13**	**4.15**	4.25	**0.24**	**0.17**	0.31

Note: DDUR results of CycleGAN-EVC and StarGAN-EVC are not reported, as they cannot modify the speech duration.

5.1.1 Objective Evaluation

*Mel-cepstral Distortion (MCD).*As shown in Table 1, all systems achieve better MCD values than that of the Zero Effort case. *Zero Effort* case directly compares the source and target utterances without any conversion. We also observe that *Emovox* completely outperforms CycleGAN-EVC and StarGAN-EVC. It also outperforms Seq2Seq-EVC for Neu-Ang and Neu-Hap (first three letters of source and target emotion, each) and achieves comparable results for Neu-Sad. This suggests that *Emovox* is superior to the others in terms of spectrum conversion.

Differences of Duration (DDUR). CycleGAN-EVC and StarGAN-EVC perform frame-by-frame mapping, but they do not convert the speech duration. Thus, the DDUR results of these two frameworks are not reported. As shown in Table 1, compared with Seq2Seq-EVC, *Emovox* achieves better results for both Neu-Ang and Neu-Hap for duration modelling, and achieves comparable results in Neu-Sad conversion. These results further confirm the effectiveness of *Emovox* in terms of duration conversion.

5.1.2 Subjective Evaluation

We report Mean Opinion Score (MOS) test results for emotion similarity with the reference for our proposed *Emovox* and all the baselines. From Fig. 6, we observe that our proposed *Emovox* consistently outperforms the baselines for all the emotion pairs. This observation is consistent with that in the objective evaluation. As for statistical significance, *Emovox* achieves the most narrow confidence interval for Neu-Ang and Neu-Hap, that suggests a high level of consistency [129]. Furthermore, we report the p-value of t-test scores of MOS between *Emovox* and the others. We observe that almost all pairs achieve a p-value below 0.05, confirming the significant results [130]. For Neu-Ang, the p-value between *Emovox* and Seq2Seq-EVC is about 0.0558, which is less than 0.1 and still supports our claim.

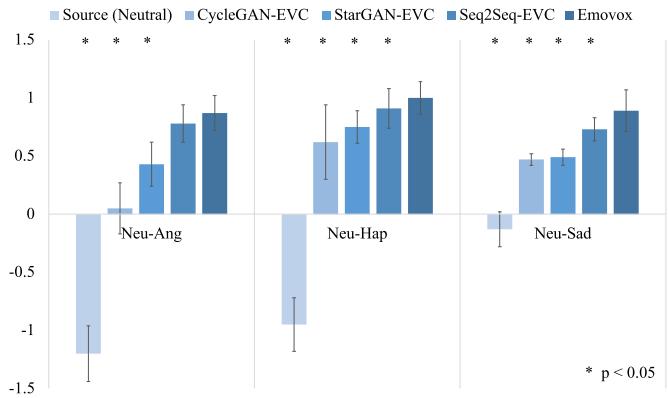


Fig. 6. Mean Opinion Score (MOS) test with 95 % confidence interval to evaluate the emotion similarity with the reference, where listeners are asked to score each sample in a scale from -2 to +2 (-2: absolutely different; -1: different; 0: cannot tell; +1: similar; +2: absolutely similar). Marker * indicates $p < 0.05$ for paired t-test scores (pairs between Emovox and the others).

5.2 Emotion Intensity Control

To evaluate the emotion intensity control in *Emovox*, we choose three different intensity values: 0.1, 0.5, and 0.9, corresponding to weak, medium, and strong. To understand the interplay between emotion intensity and different prosodic attributes, we first visualize several related prosodic cues of the converted emotional utterances with the same speaking content but different emotion intensities. We then compare our intensity control methods with other state-of-the-art methods.

5.2.1 Visual Comparisons

We visualize the prosodic attributes related to the emotion intensity, such as speech duration, pitch, and energy to gain an intuitive understanding of emotion intensity in vocal speech. Besides, we also would like to show that the emotion intensity control can be manifested in the changes of these prosodic features in our proposed framework.

(1) Duration. Speech duration is considered as a distinct factor between active and passive emotions [131]. To show that the emotion intensity is related to speaking rate, we compare the Mel-spectrogram of *Sad* as a reference emotion, which is characterized with a slower speaking rate and more resonant timbre [132], with that of its *Neutral* emotion counterpart in Fig. 7a. We also illustrate the Mel-spectrograms of *Emovox*-converted utterances with different intensities in Figs. 7c, 7d and 7e. We observe that the converted *Sad* utterance with the highest intensity value has the slowest speaking rate among all three intensities (as shown in Fig. 7e). As the intensity value increases, the speaking rate decreases. *(2) Pitch Envelope.* Pitch envelope (i.e., the level, range, and shape of the pitch contour) is considered a major factor that contributes to the speech emotion, which is closely correlated to the activity level [132], [133]. We represent pitch information with F0 contour, which is estimated with the harvest algorithm [134] and aligned with dynamic time warping [135]. In Fig. 8, we visualize the pitch contour of converted *Angry*, *Happy*, and *Sad* utterances with three

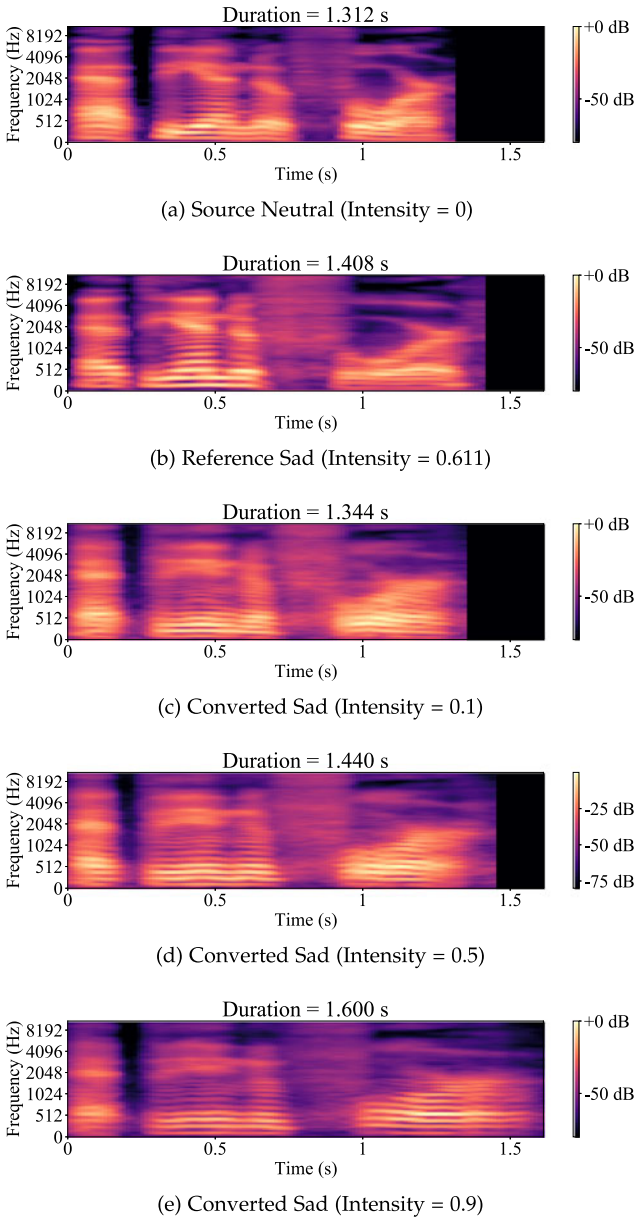


Fig. 7. Visualization of Mel-spectrograms from source neutral, reference sad, and converted sad utterances at three intensity values, i.e., 0.1, 0.5, 0.9, with the same speaking content ("At the end of four"). A greater intensity value represents a more emotional expression.

different intensities. From Figs. 8a and 8b, we observe that the converted *Angry* and *Happy* utterances with higher intensity values tend to have higher F0 values with larger fluctuations over time. This coincides the fact that the utterances with higher intensity values are more vibrant and sharper in expressing emotions such as *Angry* and *Happy*. For *Sad*, there is no big difference in F0 range for different intensities, as shown in Fig. 8c. This observation intuitively suggests that the intensity of expressing *Sad* emotion may be more related to the speaking rate than the vocal pitch.

(3) Speech Energy. Speech energy measures the volume or the loudness of a voice [136], [137]. Speech energy is often regarded as a prominent character of emotion intensity in the literature [46], [138]. To show the effect of intensity control, we visualize and compare the energy contour of different intensities in Fig. 9. To represent the speech energy, we use 26 Mel-filterbanks and multiply each of them with the power spectrum. Then, we can measure the speech energy by adding up the coefficients. As shown in Figs. 9a and 9b, we observe that the converted *Angry* and *Happy* utterances with higher intensity have lager energy values, which is consistent with our observations on the F0 contour. As for *Sad*, we similarly observe that a higher intensity results in slightly higher energy values as shown in Fig. 9c. These observations show that our proposed *Emovox* can effectively control the emotion intensity manifested in multiple prosodic factors in vocal speech.

5.2.2 Comparison With State-of-the-Art Control Methods

As a comparative study, we implement three intensity control methods (i.e., Emovox w/ scaling factor, Emovox w/ attention weights, and Emovox w/ relative attributes) as described in Section 4.1. We evaluate the performance of these three methods in terms of speech quality and intensity control.

(1) Speech Quality. We first report the BWS listening test on *Emovox* for speech quality evaluation in Table 2. At each intensity value, the subjects are asked to evaluate the speech quality of the converted emotional speech with 3 different emotion intensity control methods.

From Table 2, we observe that *Emovox* w/ relative attributes always achieves the best results for Neu-Ang and Neu-Hap, and comparable results with *Emovox* w/ attention weights for Neu-Sad. These results show that *Emovox* w/ relative attributes can achieve better speech quality while controlling the output emotion intensity than other control methods.

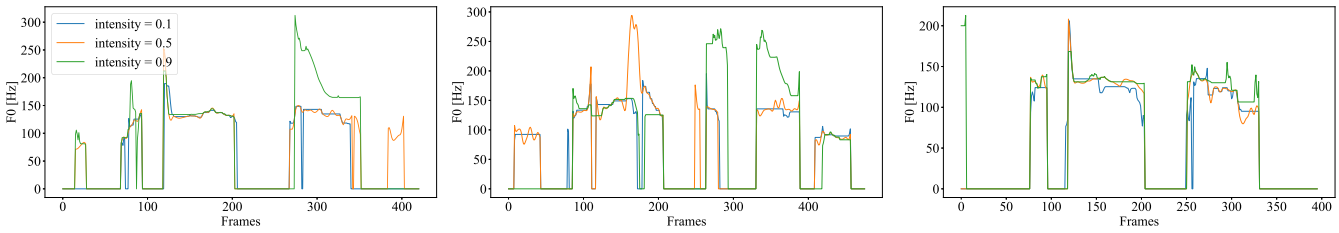
(2) Intensity Control. We then report another BWS test to evaluate the performance of emotion intensity control. For each framework, listeners are asked to assess the emotional expressiveness among three different intensities. We conjecture that the speech samples with an intensity value of 0.9 sound more expressive than others, while those with an intensity value of 0.1 sound more neutral. We report the preference percentage scores (%) of the most and the least expressiveness for each controlling method in Figs. 11 and 10, respectively.

As illustrated in Fig. 11c, *Emovox* w/ relative attributes achieve the best preference results on intensity control, where most listeners choose the samples with an intensity value of 0.9 as the most expressive ones. We also note that *Emovox* w/ scaling factor and *Emovox* w/ attention weights work well for converted angry and happy, as shown in Figs. 11a and 11b. However, their performance of converted sad is not satisfactory. We further observe that *Emovox* w/ relative attributes also work better than the others, where most listeners choose the samples with an intensity value of 0.1 as the least expressive ones, as shown in Fig. 10c. This observation is consistent with the previous one, which further validates the superior performance of relative attributes on emotion intensity control.

As a summary, our proposed *Emovox* w/ relative attributes shows better performance on emotion intensity control while achieving better speech quality than other control methods.

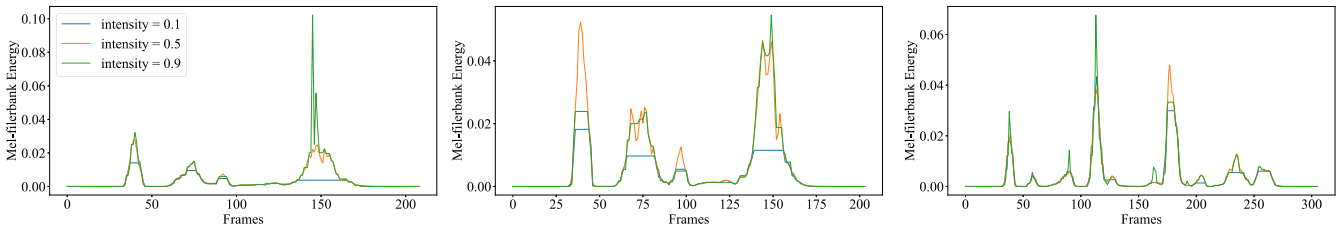
5.3 Impact of Training Data Size

To evaluate the effect of training data on the final performance, we gradually reduce the number of utterances used at the emotion training stage. We use 300, 150, 100, and 50



(a) Converted Angry (b) Converted Happy (c) Converted Sad

Fig. 8. A comparison of the pitch contour from the emotional utterances converted by *Emovox* with three different emotion intensities (0.1, 0.5 and 0.9).



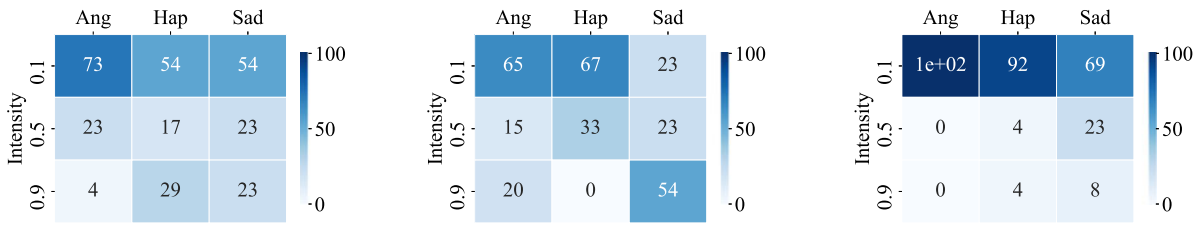
(a) Converted Angry (b) Converted Happy (c) Converted Sad

Fig. 9. A comparison of the speech energy from the emotional utterances converted by *Emovox* with three different emotion intensities (0.1, 0.5 and 0.9).

TABLE 2
A Comparison of Best-Worst Scaling (BWS) Test Results for Speech Quality of Three Different Emotion Intensity Control Methods With *Emovox*

Method	Intensity = 0.1 (Weak)						Intensity = 0.5 (Medium)						Intensity = 0.9 (Strong)					
	Neu-Ang		Neu-Hap		Neu-Sad		Neu-Ang		Neu-Hap		Neu-Sad		Neu-Ang		Neu-Hap		Neu-Sad	
	B	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W
Scaling	8%	31%	15%	15%	8%	70%	16%	38%	8%	8%	8%	62%	17%	31%	9%	6%	8%	69%
Attention	15%	69%	0%	77%	**54%**	**15%**	15%	62%	0%	92%	43%	31%	6%	69%	0%	92%	**54%**	31%
Relative	**77%**	**0%**	**85%**	**8%**	38%	**15%**	**69%**	**0%**	**92%**	**0%**	**49%**	**7%**	**77%**	**0%**	**91%**	**2%**	38%	**0%**

Note: Emovox w/ scaling factor, Emovox w/ attention weights, and Emovox w/ relative attributes are denoted as Scaling, Attention, and Relative respectively. "B" denotes "Best," and "W" denotes "Worst".



(a) Emovox w/ Scaling Factor (b) Emovox w/ Attention Weights (c) Emovox w/ Relative Attributes

Fig. 10. A comparison of the preference percentage scores (%) of Emovox with 3 different emotion intensity control methods. Listeners are asked to listen to the converted speech samples of 3 different emotion intensities (0.1, 0.5, and 0.9), and choose the least expressive one.

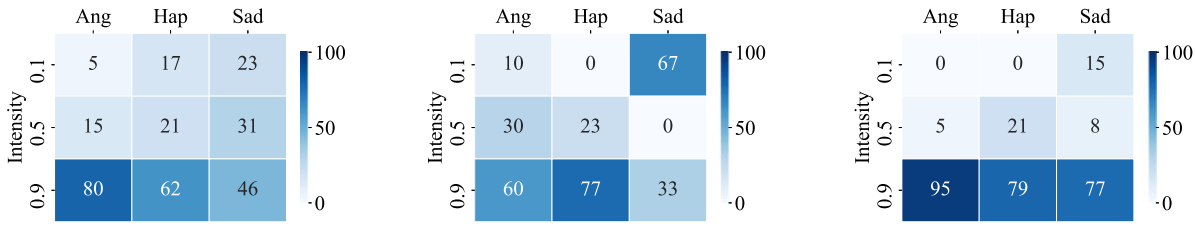
training utterances for each emotion, and use 20 utterances for evaluation. In Fig. 12, we report the MCD and DDUR results of *Emovox* and the baseline Seq2Seq-EVC.

We observe that *Emovox* consistently achieves better MCD results than Seq2Seq-EVC. We further observe that the MCD scores for *Emovox* between 150 to 50 training utterances are comparable and not significantly poorer than that of using 300 utterances. This indicates *Emovox*'s robustness to limited training data.

For DDUR, we first observe that both *Emovox* and Seq2-Seq-EVC have much higher DDUR values with 50 training utterances. It suggests that both frameworks cannot predict the speech duration well if the training size is too small. Between 300 to 100 training utterances, the performance of *Emovox* is comparable, which again attest to *Emovox*'s robustness to limited training data.

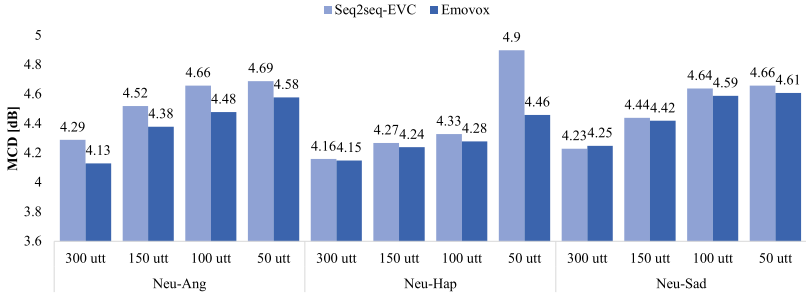
5.4 Ablation Studies

We conduct ablation studies to validate the contributions of 1) style pre-training, and 2) perceptual losses from pre-trained SER in emotion training.

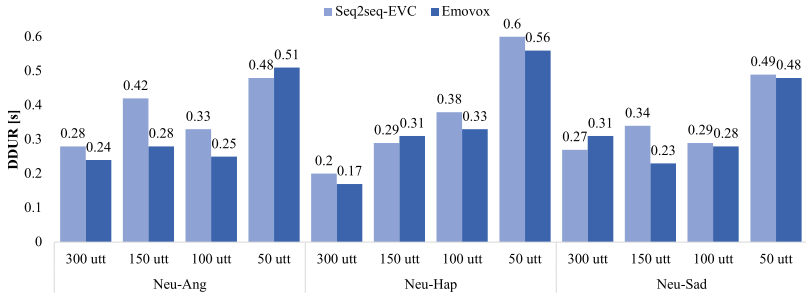


(a) Emovox w/ Scaling Factor (b) Emovox w/ Attention Weights (c) Emovox w/ Relative Attributes

Fig. 11. A comparison of the preference percentage scores (%) of *Emovox* with 3 different emotion intensity control methods, and for 3 converted emotions. The subjects are asked to listen to the converted speech samples of 3 different emotion intensities (0.1, 0.5, and 0.9), and choose the most expressive one.



(a) MCD results (a smaller value of MCD indicates a smaller spectral distortion)



(b) DDUR results (a smaller value of DDUR represents a better duration conversion performance)

Fig. 12. A comparison of MCD and DDUR results of our proposed *Emovox* and Seq2Seq-EVC for all three emotion pairs with different amounts of training utterances, that are 300, 150, 100, and 50 utterances, respectively.

5.4.1 Style Pre-Training

We first compare the Mel-cepstral distortion (MCD) results of *Emovox* and *Emovox* (w/o style pre-training), where the latter is trained directly with a limited amount of emotional speech data and without any pre-training. As shown in Table 1, *Emovox* (w/o style pre-training) provides the worst results for all emotion pairs.

We further compare *Emovox* and *Emovox* (w/o style pre-training) in terms of DDUR as shown in Table 1. We observe that *Emovox* (w/o style pre-training) has the worst DDUR results. These results validate the effectiveness of style pre-training.

5.4.2 Perceptual Loss Functions

As discussed in Section 3.3.3, we expect that the emotion embedding similarity loss L_{sim} and emotion classification loss L_{cls} help generate more discriminative embeddings (see Fig. 5). To further validate the effectiveness of these two loss functions on the final performance, we conduct a best-worse scaling listening test where we evaluate the emotion similarity with the reference emotion. To be consistent with Section 3.3.3, we only conduct an ablation study with the *Emovox w/o intensity* configuration. The results are reported in Table 3.

From Table 3, we observe that most listeners choose "*Emovox w/o intensity, w/* L_{sim} *and* L_{cls}" as the best in terms of emotion similarity, while most of them choose "*Emovox w/o intensity, w/o* L_{sim} *or* L_{cls}" as the worst for all the emotion pairs. This suggests that these two loss functions improve

TABLE 3
The Effect of the Perceptual Loss Function for the Emotion Similarity in a Best-Worst Scaling (BWS) Test for Four Variants of Emovox w/o Intensity Framework

Emovox w/o Intensity	Neu-Ang		Neu-Hap		Neu-Sad	
	Best	Worst	Best	Worst	Best	Worst
w/ L_{sim} and L_{cls}	**56%**	**5%**	**49%**	**13%**	**33%**	16%
w/ L_{sim}	33%	22%	47%	13%	25%	**11%**
w/ L_{cls}	9%	24%	2%	24%	22%	22%
w/o L_{sim} or L_{cls}	2%	49%	2%	49%	20%	51%

emotional expressiveness, which validates the idea of incorporating SER losses for emotion supervision.

6 Conclusion

This contribution filled the research gap of emotion intensity control in current emotional voice conversion literature. We proposed a novel emotional voice conversion framework – *Emovox* – that is based on a sequence-to-sequence model. The proposed *Emovox* framework provides a fine-grained, effective emotion intensity control for the first time in emotional voice conversion. The key highlights are as follows:

1) We formulated an emotion intensity modeling technique and proposed an emotion intensity controlling mechanism based on relative attributes. We proved that our proposed mechanism outperformed other competing controlling methods in speech quality and emotion intensity control.
2) Instead of simply correlating emotion intensity with the loudness of a voice, we presented a comprehensive analysis for the first time to understand the interplay between emotion intensity and various prosodic attributes such as speech duration, pitch envelope, and speech energy. We showed that our emotion intensity control could be manifested in various prosodic aspects.
3) We proposed style pre-training and perceptual losses from a pre-trained SER to improve the emotion intelligibility in converted emotional speech. We showed that *Emovox* outperformed state-of-the-arts emotional voice conversion frameworks. With style pre-training and perceptual losses from a pre-trained SER, *Emovox* could effectively perform well with a limited amount of emotional speech data.

Our future directions include the study of cross-lingual emotional voice conversion and emotion style modelling with self-supervised learning. In addition, a closer coupling of conversion and speech emotion recognition is foreseen: conversion can help augment training data for recognition, while recognition can serve as objective conversion training guidance.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments, Dr Bin Wang for valuable discussions and Dr Rui Liu for sharing part of the codes.

References

[1] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, 2022.
[2] J. Pittermann, A. Pittermann, and W. Minker, *Handling Emotions in Human-Computer Dialogues*, Berlin, Germany: Springer, 2010.
[3] 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.
[4] A. Rosenberg and J. Hirschberg, "Prosodic aspects of the attractive voice," in *Voice Attractiveness*, Berlin, Germany: Springer, 2021, pp. 17–40.
[5] 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, Nov. 2020.
[6] S. Ramakrishnan, *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*, Norderstedt, Germany: BoD–Books on Demand, 2012.
[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1998, pp. 285–288.
[8] J. Wang, R. Dou, Z. Yan, Z. Wang, and Z. Zhang, "Exploration of analyzing emotion strength in speech signal," in *Proc. Chin. Conf. Pattern Recognit.*, 2009, pp. 1–4.
[9] J. R. Averill and T. A. More, "Happiness," in *Handbook of Emotions*, M. Lewis and J. M. Haviland, Eds., New York, NY, USA: The Guilford Press, 1993, pp. 617–629.
[10] F. Biassoni, S. Balzarotti, M. Giamporcaro, and R. Ciceri, "Hot or cold anger? Verbal and vocal expression of anger while driving in a simulated anger-provoking scenario," *Sage Open*, vol. 6, no. 3, 2016, Art. no. 2158244016658084.
[11] J. W. Brehm, "The intensity of emotion," *Pers. Soc. Psychol. Rev.*, vol. 3, no. 1, pp. 2–22, 1999.
[12] N. H. Frijda, A. Ortony, J. Sonnemans, and G. L. Clore, "The complexity of intensity: Issues concerning the structure of emotion intensity," *Rev. Pers. Social Psychol.*, vol. 13, 1992.
[13] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.
[14] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proc. Hum. Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2009, pp. 81–84.
[15] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 3081–3084.
[16] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.
[17] K. Matsumoto, S. Hara, and M. Abe, "Controlling the strength of emotions in speech-like emotional sound generated by WaveNet," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3421–3425.
[18] B. Schnell and P. N. Garner, "Improving emotional TTS with an emotion intensity input from unsupervised extraction," in *Proc. 11th ISCA Speech Synth. Workshop*, 2011, pp. 60–65.
[19] 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.
[20] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42674–42687, 2021.
[21] 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.
[22] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2401–2404.
[23] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2014, pp. 1–7.
[24] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, pp. 268–283, 2009.
[25] 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.
[26] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3399–3403.
[27] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2453–2457.

[28] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 230–237.

[29] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," 2020, *arXiv:2007.12932*.

[30] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3502–3506.

[31] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3416–3420.

[32] 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.

[33] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2858–2862.

[34] R. Shankar, H.-W. Hsieh, N. Charon, and A. Venkataraman, "Multi-speaker emotion conversion via latent variable regularization and a chained encoder-decoder-predictor network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3391–3395, 2020.

[35] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6284–6288.

[36] J. Sotelo *et al.*, "Char2Wav: End-to-end speech synthesis," in *Proc. Int. Conf. Learn. Representations Workshop*, 2017.

[37] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.

[38] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[39] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6805–6809.

[40] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 656–670, Dec. 2020.

[41] F. Kreuk *et al.*, "Textless speech emotion conversion using decomposed and discrete representations," 2021, *arXiv:2111.07402*.

[42] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of F0 for speech emotion conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6830–6834.

[43] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7774–7778.

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

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

[46] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.," *Emotion*, vol. 1, no. 4, 2001, Art. no. 381.

[47] 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.

[48] 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.

[49] R. Liu, B. Sisman, G. Lai Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1806–1818, Apr. 2021.

[50] 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.

[51] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1849–1863, Jun. 2020.

[52] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7068–7072.

[53] H. Kameoka, K. Tanaka, and T. Kaneko, "FastS2S-VC: Streaming non-autoregressive sequence-to-sequence voice conversion," 2021, *arXiv:2104.06900*.

[54] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 745–755, Jan. 2021.

[55] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1717–1728, Apr. 2021.

[56] 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, Dec. 2019.

[57] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6785–6789.

[58] 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.

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

[60] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. Hum.- Comput. Stud.*, vol. 59, no. 1/2, pp. 157–183, 2003.

[61] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Sci. Technol.*, vol. 26, no. 4, pp. 317–325, 2005.

[62] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4905–4909.

[63] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4475–4479.

[64] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable TTS from annotated and latent variation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3956–3960.

[65] 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.

[66] 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.

[67] 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.

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

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

[70] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4683–4687.

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

[72] 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.

[73] 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.

[74] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoustical Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[75] Y. Xu, "Speech prosody: A methodological review," *J. Speech Sci.*, vol. 1, no. 1, pp. 85–115, 2011.

[76] M. Tahon, G. Lecorvé, and D. Lolive, "Can we generate emotional pronunciations for expressive speech synthesis?," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 684–695, Oct.–Dec. 2020.

[77] 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.

[78] 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.

[79] D. Ong *et al.*, "Modeling emotion in complex stories: The stanford emotional narratives dataset," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 579–594, Jul.–Sep. 2021.

[80] 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. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.

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

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

[83] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6720–6724.

[84] 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, 2020, Art. no. 1754073919898526.

[85] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2021.3114365.

[86] Y. Gao, W. Zheng, Z. Yang, T. Kohler, C. Fuegen, and Q. He, "Interactive text-to-speech via semi-supervised style transfer learning," 2020, *arXiv:2002.06758*.

[87] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[88] R. Liu, B. Sisman, and H. Li, "Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4648–4652.

[89] 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.

[90] 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.

[91] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Commun.*, vol. 36, no. 1/2, pp. 31–43, 2002.

[92] 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.

[93] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 433–440.

[94] B. McFee and G. R. Lanckriet, "Metric learning to rank," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 775–782.

[95] G. Koch *et al.*, "Siamese neural networks for one-shot image recognition," in *Proc. ICML deep Learn. Workshop*, 2015.

[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] J. Y. Lee, S. J. Cheon, B. J. Choi, and N. S. Kim, "Memory attention: Robust alignment using gating mechanism for end-to-end speech synthesis," *IEEE Signal Process. Lett.*, vol. 27, pp. 2004–2008, Nov. 2020, doi: 10.1109/LSP.2020.3036349.

[102] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Proc. 5th ISCA Workshop Speech Synth.*, 2004.

[103] K. Ito and L. Johnson, "The LJ speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[104] V. Christophe, Y. Junichi, and M. Kirsten, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *Centre Speech Technol. Res.*, 2016.

[105] A. Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.

[106] Speech-emotion-classification-with-pytorch, 2018. [Online]. Available: https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch/blob/master/notebooks/stacked_cnn_attention_lstm.ipynb

[107] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[108] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[109] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, NJ, USA: Prentice Hall, 1988.

[110] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc. 11th Int. Conf. Inf. Knowl. Manage.*, 2002, pp. 515–524.

[111] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, vol. 4. Hoboken, NJ, USA: Wiley, 2014.

[112] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Mining*, 2010, pp. 911–916.

[113] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empir. Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 410–420.

[114] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, Sep. 2019.

[115] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[116] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6199–6203.

[117] 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

[118] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization.".

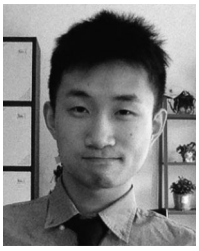
[119] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Recognition-synthesis based non-parallel voice conversion with adversarial learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 771–775.

[120] 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.

[121] 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.

[122] 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.

[123] 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.
[124] 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.
[125] 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.
[126] 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.
[127] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, 2016.
[128] S. Kiritchenko and S. Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 465–470.
[129] D. W. Hosmer and S. Lemeshow, "Confidence interval estimation of interaction," *Epidemiology*, vol. 3, pp. 452–456, 1992.
[130] R. A. Thisted, "What is a p-value," Departments of Statistics and Health Studies, 1998.
[131] J. R. Davitz *et al.*, "Personality, perceptual, and cognitive correlates of emotional sensitivity," in *The Communication of Emotional Meaning*, New York, NY, USA: McGraw-Hill, 1964, pp. 57–68.
[132] M. J. Owren and J.-A. Bachorowski, "Measuring emotion-related vocal acoustics," in *Handbook of Emotion Elicitation and Assessment*, Oxford, U.K.: Oxford Univ. Press, 2007, pp. 239–266.
[133] 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.
[134] M. Morise *et al.*, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2321–2325.
[135] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*, Berlin, Germany: Springer, 2007, pp. 69–84.
[136] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1/2, pp. 227–256, 2003.
[137] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.
[138] C. Sobin and M. Alpert, "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy," *J. Psycholinguistic Res.*, vol. 28, no. 4, pp. 347–365, 1999.



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. His research interests mainly focus on emotion analysis and synthesis in speech, including emotional voice conversion and emotional text-to-speech. He served as local arrangment co-chair of IEEE ASRU 2019, SIGDIAL 2021, IWSDS 2021, O-COCOSDA 2021 and ICASSP 2022. He is a reviewer of ICASSP and *Speech Communication*.

Berrak Sisman (Member, IEEE) received the PhD degree in electrical and computer engineering from National University of Singapore, in 2020, fully funded by the A*STAR Graduate Academy under the Singapore International Graduate Award (SINGA). She is currently working as an Assistant Professor with the Singapore University of Technology and Design (SUTD). She is also an affiliated researcher with the National University of Singapore (NUS). Prior to joining SUTD, she was a postdoctoral research fellow with the National University of Singapore, and a visiting researcher with Columbia University, New York. She was also an exchange PhD student with the University of Edinburgh and a visiting scholar with The Centre for Speech Technology Research (CSTR), University of Edinburgh, in 2019. She was attached to the RIKEN Advanced Intelligence Project, Japan, in 2018. Her research is focused on machine learning, signal processing, speech synthesis, voice conversion, and emotion. She has served as the general coordinator of the Student Advisory Committee (SAC) of the International Speech Communication Association (ISCA), and is currently serving as the general coordinator of the ISCA Postdoc Advisory Committee (PECRAC). She is appointed as an area chair (Speech Synthesis) with INTERSPEECH 2021 and 2022, and publication chair of ICASSP 2022. She is elected as a member of the IEEE Speech and Language Processing Technical Committee (SLTC) in the area of Speech Synthesis for the term from 2022 to 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 capitalises 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, reenforcement 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, and 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 with 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=96). 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 the 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 a 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 Transcations 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 professor with the School of Data Science, The Chinese University of Hong Kong (Shenzhen), China, and 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 in the University of Hong Kong (1988-1990) and 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 in Lernout & Hauspie Asia Pacific (1999-2001), vice president in 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 *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, and ASRU 2019. He is a fellow of the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 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.