Decoupling Speaker-Independent Emotions for Voice Conversion via Source-Filter Networks

Zhaojie Luo, *Member, IEEE*, Shoufeng Lin, *Senior Member, IEEE*, Rui Liu, *Member, IEEE*, Jun Baba, *Member, IEEE*, Yuichiro Yoshikawa, *Member, IEEE*, and Hiroshi Ishiguro, *Member, IEEE*

***Abstract*—Emotional voice conversion (VC) aims to convert a neutral voice to an emotional one while retaining the linguistic information and speaker identity. We note that the decoupling of emotional features from other speech information (such as content, speaker identity, etc.) is the key to achieving promising performance. Some recent attempts of speech representation decoupling on the neutral speech cannot work well on the emotional speech, due to the more complex entanglement of acoustic properties in the latter. To address this problem, here we propose a novel Source-Filter-based Emotional VC model (SFEVC) to achieve proper filtering of speaker-independent emotion cues from both the timbre and pitch features. Our SFEVC model consists of multi-channel encoders, emotion separate encoders, pre-trained speaker-dependent encoders, and the corresponding decoder. Note that all encoder modules adopt a designed information bottleneck auto-encoder. Additionally, to further improve the conversion quality for various emotions, a novel training strategy based on the 2D Valence-Arousal (VA) space is proposed. Experimental results show that the proposed SFEVC along with a VA training strategy outperforms all baselines and achieves the state-of-the-art performance in speaker-independent emotional VC with nonparallel data.**

***Index Terms*—Auto-encoder, emotional voice conversion, prosody, source-filter networks, valence arousal.**

I. Introduction

EMOTIONAL voice conversion (VC) is a useful speech processing technique for changing the emotional states of a speech utterance while retaining its linguistic information and speaker identity. It can be applied in various domains, such as virtual assistants, call centers, emotion recognition and audiobook narration [1], [2], [3], [4], etc.

Manuscript received 31 October 2021; revised 9 May 2022; accepted 15 June 2022. Date of publication 14 July 2022; date of current version 2 December 2022. This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (Data collection) and High-level Talents Introduction Project of Inner Mongolia University No. 10000-22311201/002 (Experimental support). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hema A Murthy. *(Zhaojie Luo and Shoufeng Lin contributed equally to this work.) (Corresponding author: Rui Liu.)*

Zhaojie Luo is with the Institute of Scientific and Industrial Research, Osaka University, Toyonaka, Osaka 567-0047, Japan (e-mail: luozhaojie@hotmail.com).

Shoufeng Lin is with the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA 6102, Australia (e-mail: shoufeng.lin@graduate.curtin.edu.au).

Rui Liu is with the School of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China (e-mail: liurui_imu@163.com).

Jun Baba is with the CyberAgent, Inc., Shibuya-ku, Tokyo 150-6121, Japan (e-mail: baba_jun@cyberagent.co.jp).

Yuichiro Yoshikawa and Hiroshi Ishiguro are with the Graduate School of Engineer Science, Osaka University, Toyonaka 560-8531, Japan (e-mail: yoshikawa@irl.sys.es.osaka-u.ac.jp; ishiguro@irl.sys.es.osaka-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2022.3190715

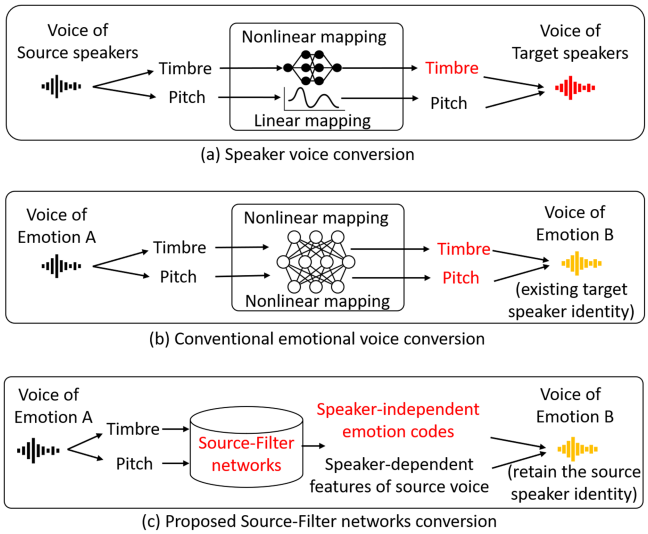

Fig. 1. (a) Normal speaker voice conversion methods use the deep learning models to convert the timbre features for speaker identity conversion; (b) Conventional emotional voice conversion methods use deep learning models to convert both pitch and timbre features for emotion conversion; (c) Proposed SFEVC model uses source-filter networks to filter out the speaker-independent and dependent emotion codes from the timbre and pitch features and convert the speaker-independent emotion code for conversion.

Originally motivated by the traditional speaker VC, the study of emotional VC has recently attracted wide attention in the field of speech processing. However, emotional VC and traditional speaker VC differ in many ways. As shown in Fig. 1(a), in the speaker VC tasks, either the one-to-one conversion [5], [6], or the many to many conversion [7], [8], their main purposes are both to convert the source speaker's voice to sound like that of the target speaker. In the emotional VC, we only convert the prosody of the voice to that of a target emotion such as anger or happiness, of the same person rather than another speaker. In a nutshell, for voice conversions, the traditional speaker VC aims to change the speaker's identity, whereas the emotional VC endeavors to convert the emotional states of the same speaker.

Traditional speaker VC researches include modeling the timbre features mapping with statistical methods such as the Gaussian mixture model (GMM) [9] [10] and non-negative matrix factorization (NMF) [11], [12]. Recent deep learning approaches such as deep neural network (DNN) [13], long short-term memory network (LSTM) [14], and generative adversarial networks (GANs) [15] have achieved remarkable performance in the

traditional VC [7], [16], [17]. As a consequence, emotional VC has also developed in this direction. Early studies on emotional VC handle both timbre (spectrum) and pitch (F0) conversion with GMM [18], [19], [20]. Some deep learning-based emotional VC models, such as DNN, RNN, and GANs have proven effective on emotional VC. For example, Luo et al. increased the dimension of the F0 features and applied the DNN model in the emotional VC [21]. Moreover, for the data augmentation, they used the continuous wavelet transform (CWT) to analyze the F0 features [22] and improved their work using the dual-supervised GANs models to do the training [23]. Ming et al. [24] applied the LSTM models in the emotional VC, and Kun et al. [25] used the unsupervised cycleGAN model to do the nonparallel conversion in the emotional VC. These works have made a great contribution to the development of emotional VC.

However, as shown in Fig. 1(b), conventional emotional VC methods just applied the nonlinear mapping model to the pitch (F0) conversion, which is similar to the normal speaker VC models used in the timbre features conversion as in Fig. 1(a). In our experiments, we have observed that the speaker similarity will reduce the effectiveness of the emotional VC probably due to the "noise" in the emotional features. The common emotional features converted by traditional VC models include not only the emotion but also other information (e.g. existing target speaker identity) altogether, causing the converted emotional speech to not retain the source speaker identity well. Nonetheless, as we hear from different speakers, even in different languages, we can easily recognize their emotions from the speech. Motivated by this, we find that there are speaker-independent emotion codes, which can be extracted from different speakers and languages. Therefore, in this work, we focus on disentangling the emotion feature from the other acoustic features in order to achieve emotional VC effectively. As shown in Fig. 1(c), to address this problem, we propose a source-filter emotional VC networks (*SFEVC*) to decouple the speaker-independent emotion code from other acoustic information of different speakers, but keep the speaker-dependent features of the source speaker unchanged.

The source-filter model [26] represents the speech production process by separating the excitation and the resonance phenomena in the vocal tract, where the source corresponds to the glottal excitation and the filter corresponds to the vocal tract. This model assumes that these two phenomena are completely decoupled. Li et al. [27] also pointed out the effectiveness of source-filter model for emotional vowel perception in the valence-arousal space. We note that the use of the source-filter model in emotional VC deserves further exploration, which will be to shown as follows in this paper.

In our SFEVC model, the source-filter network is based on the encoder-encoder-decoder architecture, which consists of multi-channel encoders, emotion-separate encoders, and the corresponding decoder. All encoders are applied with the designed information bottleneck auto-encoders, which can filter out the specified features from the emotional speech. For the multi-channel encoders, they can disentangle the content from the acoustics features (timbre and prosody). However, these speaker-dependent acoustics features are still mixed with the emotion features. Therefore, the emotion-separate encoders are focusing on separating the speaker-independent emotion features from these speaker-dependent acoustic features. Finally, the decoder takes the speaker-independent emotion codes of target voice as input to convert the speaker's emotion without distorting the speaker-dependent acoustic features of source voice.

Moreover, according to the emotion studies [28], [29], [30], psychological emotion labels can be typically divided into discrete emotion states (angry, happy, neutral, and so on) or dimensional continuous emotion space (valence-arousal (VA) space) [31]. As indicated in the emotion research [32], valence (how positive or negative an emotion is) and arousal (power of the activation of the emotion) constitute popular and effective representations for affecting the emotion. Using the VA space over the two emotion dimensions is considered to be more general than the use of discrete states in solving the speech problem. For example, the pitch feature of audio [33], [34], one of the most reliable features, can be seen as an important representation of arousal [35], although it may not well represent the valence. Therefore, the happiness and sadness should be far away from each other in the VA space, while happiness and anger might be less distinct. To further improve the conversion performance in terms of emotion expressiveness, a novel VA training strategy, which leverages the relationship between discrete emotion classes and VA space, is proposed to train our SFEVC model more effectively.

To validate our proposed model, we conduct the experiments on multiple emotional voice dataset. Note that we introduce a new sales conversation corpus with a new high tension emotion, denoted as "**H**igh **T**ension **E**motion dataset (**HTE**[1])," that will contribute to enrich the development of VC research community.

In summary, the main contributions of this work are summarized as follows:

- We point out that there are speaker-independent emotion codes, which can be extracted from different speakers and languages, and introduce a novel emotional VC paradigm based on the source-filter model to disentangle the speaker-independent emotion feature from other acoustic features.
- A novel VA training strategy based on the 2D valence-arousal space is proposed for an effective training procedure.
- To validate the effectiveness of our proposed model, we have conducted the experiments not only on three open source emotional voice datasets, but also a newly released high tension emotion dataset, in which the emotions are more complex and challenging to convert. Our experimental results show the superior performance of our proposed method over state-of-the-art emotional VC methods.

To our best knowledge, this is the first study of applying the source-filter model in the emotional VC literature.

This paper is organized as follows. The background of emotional speech, source-filter theory and valence-arousal theory are briefly reviewed in Section II. Emotion voice data analysis is provided in Section III. Section IV presents our proposed

[1]**HTE:** https://github.com/ZhaojieL/HTE-data

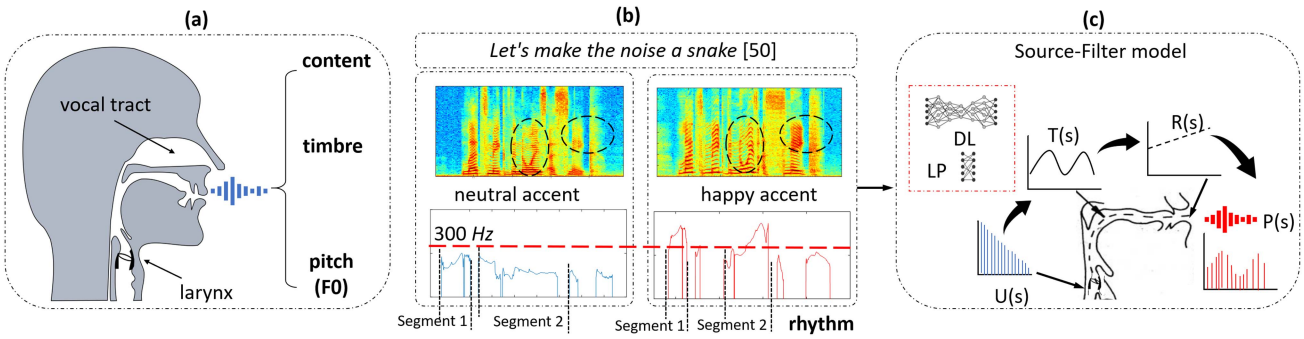

Fig. 2. (a) An illustration of the process of emotion speech generation where the larynx and the vocal tract affect the pitch and timbre features respectively; (b) Differences of the timbre and pitch features between neutral and emotional speech of the same content; (c) The conventional source-filter model simulates the speech production via the excitation in the larynx and the resonance in the vocal tract.

SFEVC method. Section V gives the details of experimental evaluations, and Section VI concludes the paper.

II. Background

We provide here a brief primer on the emotional speech, the source-filter model and the emotional valence-arousal space theory.

A. Information in Emotional Speech

Fig. 2(a) gives a simplified illustration of the process of speech generation. The speech excitation comes from the vibration of vocal cords in the larynx. The generated voiced sound is then modulated by resonance of the vocal tract (guttural, oral and nasal cavities). The speech signal contains four main information components: language content, timbre, pitch and rhythm. The emotional features are embedded in these components in different ways.

Content belongs to the language model in speech research. The phoneme is the basic unit of speech content in most languages. Each phoneme has a particular formant pattern. Thus different phonemes appear in different shapes in the spectrogram. As shown in Fig. 2(b), the spectrograms of the same content spoken in different emotions have similar shapes. In the emotional VC tasks, the linguistic information needs to be kept unchanged. Thus, the source-filter model needs to retain the content information while converting the emotional features.

Timbre is reflected by the formant, which is the peak of the spectral envelope that results from an acoustic resonance of the human vocal tract. The timbre can represent the tone color or unique quality of a sound, which helps us instantly identify and classify sound sources, such as individual people or musical instruments. In the emotional speech, the high arousal e.g. happy or angry voice tends to sound sharper and brighter than the low arousal or neutral voice [36]. As shown in Fig. 2(b), the comparison between a happy and a neutral utterance shows that the happy voice has a sharper spectrogram in some words, and a higher formant frequency range at the end of the utterance than the neutral one. Thus, it is necessary to extract the emotional information from the timbre features.

Pitch is an important parameter in emotional speech processing systems. Modulated pitch is generated by the larynx and modulated primarily by fine changes in the tension of the vocal folds. The ability to voluntarily and flexibly control pitch patterns, in the context of vocal learning, is unique to humans among primates [37], while it has been previously thought that this ability was due to anatomical differences in the larynx [38]. As shown in Fig. 2(b), the happy voice has a higher frequency than the neutral voice. It has been proved that the higher pitch, increased intensity, and faster rate are associated with more excited or high arousal emotions in speech [39]. Thus, in this research, we will transform and control the pitch of voice by the designed emotion auto-encoder, for flexible emotion conversion. However, the pitch contour also contains the rhythm information and speaker identity. For example, female speakers tend to have a higher pitch range than male speakers, which indicates certain speaker identity information. Moreover, since each nonzero segment of the pitch contour represents a voiced segment, the time length of each voiced segment indicates how fast the speaker speaks, which can be represented as the rhythm information.

Rhythm is a recurring pattern of sound or speech in time series. It represents how fast the speaker utters each syllable or word. Regular recurrence of grouped stressed and unstressed, long and short, or high-pitched and low-pitched syllables goes in alternation. As shown in Fig. 2, each pitch contour is divided into segments, which correspond to vowels of words, and the time lengths of these segments reflect the rhythmic information. Emotions can be inferred from speech rhythm. For instance, the happy or angry voice is usually faster than the sad or neutral voice.

B. The Source-Filter Model

The source-filter model simulates the speech production via two distinctive parts, i.e. the excitation in the larynx and the resonance in the vocal tract. This principle is illustrated in Fig. 2(c). The vowel spectrum $P(s)$ is the product of the spectrum of the glottal source $U(s)$, the transfer function of the vocal tract $T(s)$ (filter), and the radiation characteristics $R(s)$. This model is also known as the "source-filter theory" of vowel production [40]. An example of source-filter modeling is the Linear Prediction (LP) model [41], which uses the source-filter theory assuming that the speech is the output signal of a recursive

digital filter while the excitation is received at the input. In reality, the mechanism of the vocal fold is more complex, making this assumption over-simplistic. There exist other source-filter models that use multi-layer deep learning (DL) models to deal with the complex excitation signals composed of deterministic and stochastic components [42]. Some traditional speaker VC methods [43], [44] also applied source-filter theory on feature decoupling, but are mostly focused on simulating the vocal tract, while information about pitch, rhythm, and content is still mixed. A recent source-filter speaker VC method, SPEECHFLOW [45], can blindly decompose speech into content, timbre, pitch, and rhythm, and generate speech from these disentangled representations. However, our task is aiming at converting the emotion while keeping the speaker-dependent features unchanged, which is different from the SPEECHFLOW. Moreover, emotional speech is the result of the interplay between acoustic attributes including timbre and pitch [46], [47], and is thus affected by them in varying degrees [23], [48]. Therefore, decoupling the timbre and pitch from content can only change the style of speech, but may not be sufficient to convert it to an expected emotion well. In this work, we further decouple the speaker-independent emotional features from the prosody and timbre features for emotional VC.

C. Valence-Arousal

Representing human emotions has been a fundamental research topic in psychology. The most frequently used emotion representation is the categorical one, comprising several basic categories such as anger, disgust, fear, happiness, sadness, surprise and neutral, etc. It is, however, the dimensional emotion representation [49] that is more appropriate to depict subtleties. Most of the dimensional models classify affective states in two dimensions, i.e. 'Valence' and 'Arousal'. As indicated in the emotion research, the 2D VA space provides a popular and effective representation for affecting emotions. For example, the pitch feature of audio, one of the most reliable features, can be seen as an important representation of arousal [35], while the facial expression can be used as the valence. Thus, we can improve the training effectiveness for the emotional VC via the separate learning pipeline, based on the relationships between emotions' valence-arousal (VA) spaces. More details are provided in Section IV-C.

III. Data Analysis

In this chapter, we provide the speech data analysis for the most important feature, i.e. pitch (F0), in the emotional VC, using the ESD database [50], which consists of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers and covers 5 emotion categories (neutral, happy, angry, sad, and surprised). We apply the t-distributed Stochastic Neighbor Embedding (t-SNE) [51] to reduce the dimensionality of the emotional features of the different emotional speech and to plot them in a two-dimensional space. Fig. 3(a) and (b) represent the t-SNE separation by the F0 features, which are the main emotion representation in speech. The instances are marked per emotions in two different languages. As shown in the figures, the neutral

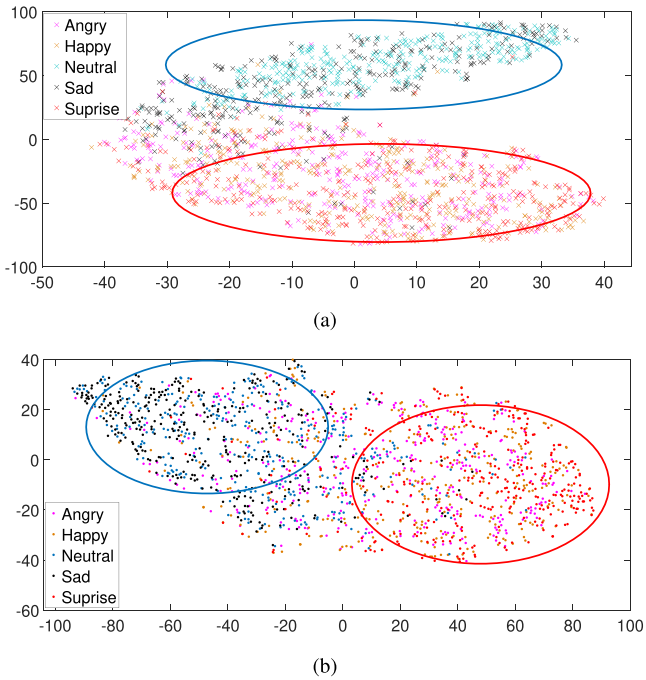

Fig. 3. t-SNE visualization of F0 features from 350 sentences with the same content spoken in (a) Chinese. (b) English.

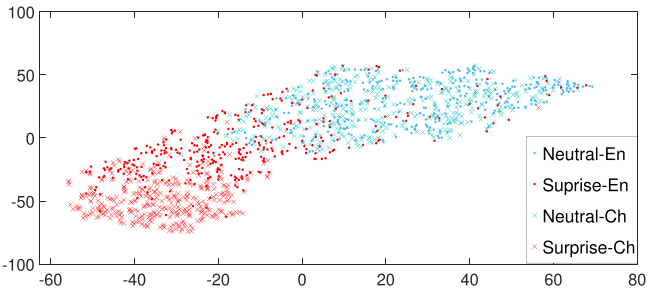

Fig. 4. t-SNE visualization of F0 features of two emotions spoken in different languages.

emotion is mostly mixed with the sad, while the happy, surprised and angry mixed together. The low arousal emotions (sad and neutral) can be easily separated from relatively high arousal emotions (happy, angry and surprised) [52], while it is difficult to separate emotions in the same arousal group, no matter how different their valence is. Thus, it is easier to represent the emotion's arousal features with the prosody feature of speech, rather than the valence features. Therefore, when converting the emotions, we need to train the emotional source-filter network with different stages based on the 2D VA space.

As shown in Fig. 4, we also apply the t-SNE to plot the emotion features extracted from different speakers in different languages. We color-coded the instances according to emotions and the different shapes of the data points show the different speakers in different languages. We can see clusters forming based on emotions instead of languages, which indicates that the emotion is a speaker-independent feature, even in the different languages. Motivated by this, we can use different speakers' emotional speech to extract the speaker-independent emotion codes as

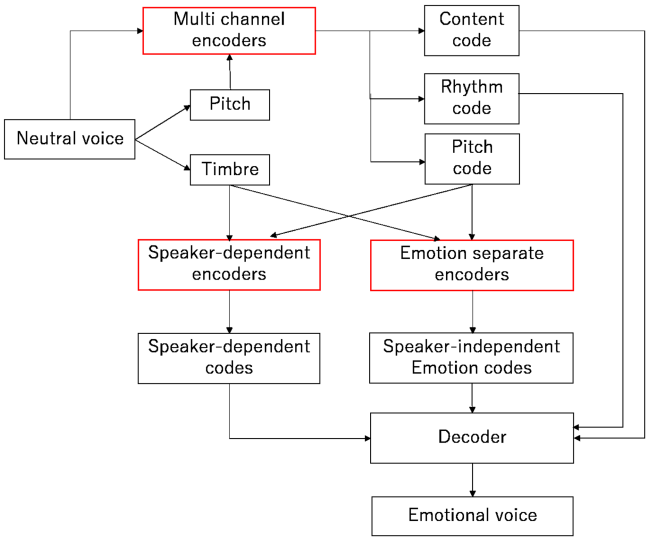

Fig. 5. The conversion workflow of SFEVC.

shown in Fig 1 (c). Then, we use the different emotion speech by the same speaker to train the source-filter networks to learn how to filter out the emotion feature but keep the speaker-dependent pitch and timbre features unchanged.

IV. Source-Filter Emotional Voice Conversion

A. The SFEVC Framework

As shown in Fig. 5, our proposed SFEVC model consists of three encoders and one decoder. In the conversion flow, the pitch and timbre features are extracted from the neutral voice using the speech process tool WORLD [53]. The neutral voice and its pitch feature are fed into the multi-channel encoders to get the content code, rhythm code, and pitch code. The pitch code and the extracted timbre feature are inputs for the speaker-dependent encoders to get the source speaker-dependent codes, and the emotion separate encoders to get the speaker-independent target emotion codes. Finally, the content code, rhythm code, speaker-dependent codes, and speaker-independent emotion codes are decoded by the decoder to obtain the emotional voice.

Fig. 6 shows the detailed structure of the proposed SFEVC. 1) In the multi-channel encoders, the features from the source and target emotion speech of the same speaker are separately fed into each channel. Each channel has a different and carefully crafted information bottleneck design to decompose speech into content, prosody, and rhythm, separately. 2) In the emotion-separate encoders, we use both the source and target emotion features from the same speaker as inputs, so that the speaker-independent emotion information can be separated from the source and target pitch codes, and the timbre features from the source and target emotional speech. 3) The speaker dependent encoders are pre-trained using the same emotional speech by different speakers, which is similar to the normal speaker voice conversion encoder. 4) The decoder aims to take the decoupled features as input to generate the target speech spectrogram features. We will introduce all modules as follows.

1) Multi-Channel Encoders: We apply the multi-channel encoder with three encoder channels E_r, E_c, E_f. Here E_r denotes the rhythm encoder, E_c the content encoder, and E_f the pitch encoder. Each channel has a different, carefully crafted information bottleneck design, which is similar to AutoVC [44].

The inputs of the multi-channel encoders are speech $\boldsymbol{X}$ and normalized pitch contours (PNorm.$\boldsymbol{P}$ and PNorm.$\boldsymbol{P^t}$). As the normalized pitch contours are normalized to have the same mean and variance across the same speakers, the normalized pitch contours only contain the pitch information and rhythm information, but no content information and timbre information.

The outputs of the encoders are called codes $\boldsymbol{Z} = \{\boldsymbol{Z}_r, \boldsymbol{Z}_c, \boldsymbol{Z}_f^s, \boldsymbol{Z}_f^t,\}$, where $\boldsymbol{Z}_r$, $\boldsymbol{Z}_c$, $\boldsymbol{Z}_f^s$ and $\boldsymbol{Z}_f^t$ denote the rhythm code, content code, pitch code of source emotion, and pitch code of target emotion, respectively. The codes can be expressed as follows:

$$\begin{aligned} &\boldsymbol{Z}_r = E_r(\boldsymbol{X}), \boldsymbol{Z}_c = E_c(A(\boldsymbol{X})), \\ &\boldsymbol{Z}_f^s = E_f(A(\boldsymbol{P})), \boldsymbol{Z}_f^t = E_f(A\left(\boldsymbol{P^t}\right)), \end{aligned} \tag{1}$$

where $A(\cdot)$ denotes the random resampling (RR) operation. As shown in Fig. 6, we applied RR operation for content encoder and pitch encoder, but not rhythm encoder. Because the RR operation divides the input into segments of random lengths and randomly stretch or squeeze each segment along the time dimension. Therefore, it can be used as an information bottleneck to filter out the rhythm information.

2) Emotion-Separate Encoders: As discussed earlier, in the emotional VC tasks, we regard the emotion information as speaker-independent, which is mixed in the timbre and prosody features. Thus, in the emotional VC, we need to decouple the speaker-independent emotion features from the timbre features $\left(\boldsymbol{U}, \boldsymbol{U}^t\right)$ and pitch codes $\left(\boldsymbol{Z}_f^s, \boldsymbol{Z}_f^t\right)$ extracted from the multi-channel encoders, while keeping the speaker-dependent features unchanged. In the emotion-separate encoders, $E_{\boldsymbol{U}}^{t-s}$ and $E_{\boldsymbol{Z}_f}^{t-s}$ represent the speaker-independent emotion encoders for timbre and pitch features, respectively. The emotion-separate encoders can be expressed as follows:

$$\begin{aligned} \boldsymbol{Z}_U^t &= E_{\boldsymbol{U}}^{t-s}\left(\boldsymbol{U^t}, \boldsymbol{U}\right), \\ \boldsymbol{Z}_{\boldsymbol{Z}_f}^t &= E_{\boldsymbol{Z}_f}^{t-s}\left(\boldsymbol{Z}_f^t, \boldsymbol{Z}_f^s\right). \end{aligned} \tag{2}$$

To let the $E_{\boldsymbol{U}}^{t-s}$ extract speaker-independent emotion features from timbre features, the inputs are timbre features $\left(\boldsymbol{U^t}, \boldsymbol{U}\right)$ of source and target emotion speech (s and t) from the same speaker ($\boldsymbol{X}$). To let $E_{\boldsymbol{Z}_f}^{t-s}$ extract speaker-independent emotion features from pitch features, the inputs are $\left(\boldsymbol{Z}_f^s, \boldsymbol{Z}_f^t\right)$ from the multi-channel encoders by the same speaker. The speaker-independent emotion codes extract from the timbre and prosody can be represented as $\boldsymbol{Z}_U^t$ and $\boldsymbol{Z}_{\boldsymbol{Z}_f}^t$ respectively.

3) Speaker-Dependent Encoder: To keep the source speaker-dependent features unchanged, we used the pre-trained speaker-dependent encoder $E_{\boldsymbol{U}}^s$ and $E_{\boldsymbol{Z}_f}^s$ for timbre features and pitch codes encoding, respectively. The pre-trained model used the timbre features and pitch codes of the source emotion speech by different speakers as inputs for training encoders.

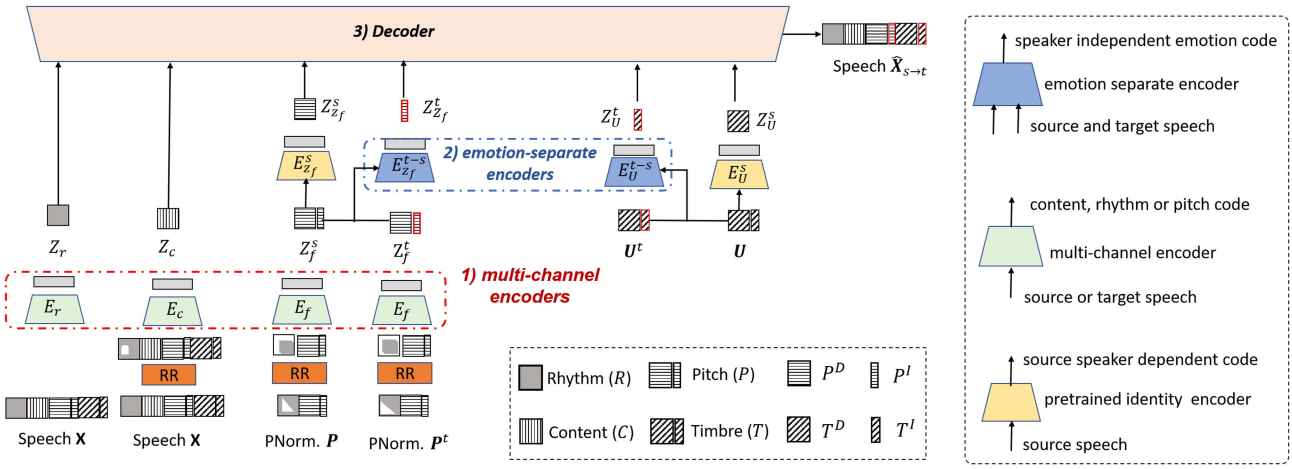

Fig. 6. The detailed structure of the proposed SFEVC. 'RR' denotes random resampling. 'PNorm.' denotes the normalized pitch contour. $\boldsymbol{P}$ and $\boldsymbol{P^t}$ represent the normalized pitch contour of source emotion and target emotion, respectively. $\boldsymbol{U}$, and $\boldsymbol{U}^t$ represent the input timbre features of source emotion speech and target emotion speech, respectively. (E)s and (Z)s represent the encoders and their codes, respectively. Pitch feature (P) consists of the speaker-dependent pitch feature (P^D) and the speaker-independent pitch feature (P^I). Timbre feature (T) consists of speaker-dependent timbre feature (T^D) and speaker-independent timbre feature (T^I). Some rhythm blocks have some holes in them, which represents that a portion of the rhythm information is lost. The grey block at the top of the encoders denotes the information bottleneck.

To extract the speaker-dependent features of source emotion by the pre-trained encoders $E^s_{\boldsymbol{U}}$ and $E^s_{\boldsymbol{Z}_f}$, the inputs are the source timbre features $\boldsymbol{U}$ and source pitch codes $\boldsymbol{Z}^s_f$. The source speaker-dependent codes extracted from timbre and prosody features can be represented as follows:

$$\begin{aligned} \boldsymbol{Z}^s_U &= E^s_{\boldsymbol{U}}(\boldsymbol{U}), \\ \boldsymbol{Z}^s_{\boldsymbol{Z}_f} &= E^s_{\boldsymbol{Z}_f}\left(\boldsymbol{Z}^s_f\right) \end{aligned} \tag{3}$$

4) Decoder: The decoder takes $\boldsymbol{Z}$ as its input and produces a speech spectrogram $\hat{\boldsymbol{X}}_{s\to t}$ as output. Now we want to convert the voice's speaker-independent emotion from the source emotion $\boldsymbol{X}$ to the target emotion $\boldsymbol{X}^t$ but keep the speaker identity unchanged. The converter should have the following desirable property:

$$\hat{\boldsymbol{X}}_{s\to t} = D\left(\boldsymbol{Z}_r, \boldsymbol{Z}_c, \boldsymbol{Z}^t_{Z_f}, \boldsymbol{Z}^s_{Z_f}, \boldsymbol{Z}^t_U, \boldsymbol{Z}^s_U\right). \tag{4}$$

Here, $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_c$ are the source speech rhythm and content. $\boldsymbol{Z}^s_{Z_f}$ and $\boldsymbol{Z}^s_U$ mean the speaker-dependent pitch and timbre from source speech. $\boldsymbol{Z}^t_{Z_f}$ and $\boldsymbol{Z}^t_U$ denote the speaker-independent pitch and timbre from target emotion speech.

During training, the output of the decoder tries to reconstruct the input spectrogram:

$$\min_{\theta} \mathbb{E}\left[\left\|\hat{\boldsymbol{X}}_{s\to t} - \boldsymbol{X}\right\|\right]^2_2 \tag{5}$$

where θ denotes all the trainable parameters. It has been proved that if all the information bottlenecks are appropriately set and the network representation power is sufficient, a minimizer of (5) will satisfy the multi-channel encoders and emotion-separate encoders, separately. More details can be found in Appendix A.

B. Method Explanation

In this section, we explain why SFEVC can achieve speech decomposition by multi-channel encoders and filter the speaker-independent emotion features from the timbre and pitch features by emotion-separate encoders. The theory of how multi-channel encoders achieve speech decomposition is similar to the other source-filter networks applied in the speaker VC tasks [44], [45]. When passing through the random resampling (RR) operation, a random portion of the rhythm block is wiped, but the other blocks remain intact. Thus, when the speech and pitch features pass through the random resampling operation, their rhythm blocks of them are missing information. Rhythm encoder $E_r(\cdot)$ is the only encoder that has access to the complete rhythm information. Hence, if $E_r(\cdot)$ is forced to lose some information by its information bottleneck, it will prioritize removing the content, pitch and timbre. For the same reason, the $E_f(\cdot)$ only encodes pitch features. Then the content encoder $E_c(\cdot)$ becomes the only encoder that can encode all the content information, and thus, will keep the content and remove the other features by the designed bottlenecks layers. Finally, with $E_r(\cdot)$ encoding only rhythm and $E_c(\cdot)$ encoding only content, the pitch encoder $E_f(\cdot)$ must encode the pitch information.

Different from the conventional source-filter networks for normal speaker voice conversion, where the speaker's identity is directly fed to the decoder, in the emotional VC, we need to convert the emotion of voice but keep the source speaker identity unchanged. Thus, we propose the emotion-separate encoders to encode the timbre features $(\boldsymbol{U}, \boldsymbol{U}^t)$ and the pitch codes $(\boldsymbol{Z}^t_f, \boldsymbol{Z}^s_f)$. As shown in the emotion-separate encoders, the speaker-independent encoders ($E^{t-s}_{\boldsymbol{U}}$ and $E^{t-s}_{\boldsymbol{Z}_f}$) encode the paired source and target emotion by the same speaker $\boldsymbol{X}$. Thus, these encoders are only embedded with the emotion codes without the speaker's identity. For the speaker-dependent encoders ($E^s_{\boldsymbol{U}}$ and $E^s_{\boldsymbol{Z}_f}$), they are pre-trained using the inputs

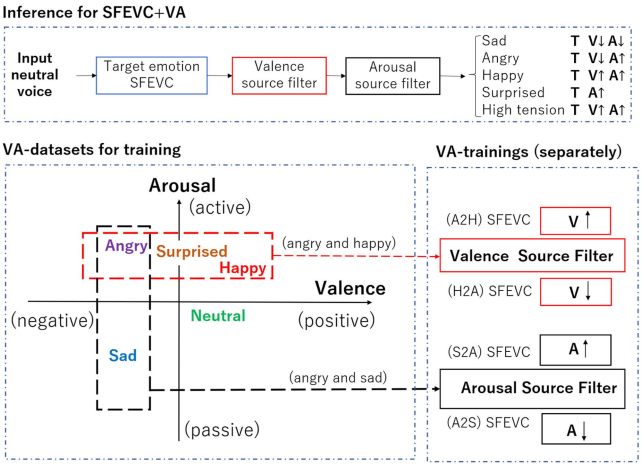

Fig. 7. SFEVC+VA inference and VA training stages based on valence-arousal space. T represents the corresponding target emotion SFEVC. $V\uparrow$ and $V\downarrow$ represent increasing and decreasing valence for the converted emotional voice, respectively. $A\uparrow$ and $A\downarrow$ represent increasing and decreasing arousal for the converted emotional voice, respectively.

from the same emotional voices spoken by different speakers, which can be embedded with the speaker's identity in the pitch and timbre features. In general, through these emotion-separate encoders, speaker-dependent and speaker-independent features can be decoupled from the timbre and pitch features.

C. Inference and Training Procedure of SFEVC+VA.

In the inference of SFEVC+VA, our input is the neutral voice. For all the conversions, the inputs need to pass through the target emotion SFEVC, which is pre-trained in the basic SFEVC model, to get the basic converted emotional voice. For instance, as shown in the inference part in Fig. 7, when we want to convert a neutral voice to a happy voice, we need to filter the neutral voice by passing it through the pre-trained neutral-to-happy SFEVC, the valence source-filter ($V\uparrow$) and arousal source-filter ($A\uparrow$), which can increase its arousal and valence, respectively. To convert to the angry voice, it needs to decrease the valence and increase the arousal using the ($A\downarrow$) and ($V\uparrow$). For the sad voice, ($A\downarrow$) and ($V\downarrow$) can be used. For the surprised voice, which has similar valence to the neutral voice, it only needs to pass through the target emotion SFEVC and arousal source-filter ($A\downarrow$). For the high tension voice conversion, it needs to pass through the high tension SFEVC and the same VA source-filters with a happy voice ($A\uparrow, V\uparrow$)

As described in the data analysis in Section III, sad and neutral are mixed, while happy, angry, and surprised emotions are mixed when using the t-SNE for clustering of pitch features. Also, as shown in the VA datasets in Fig. 7, in the emotion wheel, neutral and sad belong to middle or low arousal emotion features, while happy, surprised and angry are the high arousal features. It indicates that it is effective to convert the emotion in different arousal using the relatively high arousal emotion (angry) and low arousal emotion (sad) as the training datasets. Also, it is effective to convert the emotion in different valences using relatively low valence emotion (angry) to high valence emotion (happy) as training datasets. Therefore, in our VA training-based model, we train the valence and arousal source-filters based on valence arousal features, separately.

In the training phase, we use the angry-happy and angry-sad pairs for training the valence source-filters and arousal source-filters, separately. For training the valence source-filters, we train the conversion function from angry to happy to get the valence conversion code, which is focused on converting the low valence to high valence ($V\uparrow$). Their inverse conversion (e.g. happy to angry) can convert the high valence features to low valence, which can be represented as $V\downarrow$. Similar training processing to arousal source-filters, we can obtain the arousal source-filters $A\uparrow$ and $A\downarrow$, to respectively filter out the low arousal and the high arousal features.

V. Experiments

A. Dataset and Experimental Settings

All experiments are conducted on four datasets including the ATR [19], JSUT corpus [54], ESD [50] and our new HTE datset.

- *ATR [19]:* In the database, 50 sentences from the ATR Japanese phonetically balanced text set were used in the experiments. These 50 sentences are designed to include a minimum phoneme set of Japanese. All the texts were read by two female professional narrators with neutral, angry, happy and sad voices.
- *JSUT corpus [54]:* In the JSUT corpus, 100 sentences are recorded by three female professional narrators with neutral, angry and happy voices.
- *ESD [50]:* The dataset consists of 350 parallel utterances with an average duration of 2.9 seconds spoken by 10 native English and 10 native Mandarin speakers. For each language, the dataset consists of 5 male and 5 female speakers in five emotions (happy, sad, neutral, angry, and surprised).
- *HTE:* We introduced the High tension emotion dataset (HTE) for real world emotional VC testing evaluation. The dataset consists of 100 parallel utterances including 50 sales conversation utterances and 50 phonetically balanced sentences with an average duration of 5 seconds spoken by 6 native Japanese voice actors (3 males and 3 females). Each sentence is spoken with two scenarios: 1) acting as a salesperson who speaks in a high tension emotional voice, and 2) acting like a normal person who speaks in a neutral voice.

All the speech data are sampled at a 16 kHz rate with 16 bits resolution. We set these four datasets into the following: neutral to angry voice (N2An), neutral to sad voice (N2Sa), neutral to happy voice (N2Ha), neutral to surprised voice (N2Su), and neutral to high tension voice (N2Hi) data pairs. They are split into training and test sets as shown in Table I. We conducted evaluations with a five-fold cross-validation scheme and the performance is measured using average conversion evaluation. Please refer to our online demo[2] to play the speech samples.

[2]**Demo page:** https://zhaojiel.github.io/SFEVC/

TABLE I
DISTRIBUTION OF TRAINING AND TEST SETS. WE USE 10% OF THE TRAINING DIALOGUES AS THE VALIDATION SET

Dataset	ATR	JSUT	ESD	HTE
train/Val	40	80	280	80
test	10	20	70	20
Emotion Pairs	N2An, N2Ha N2Sa	N2An, N2Ha	N2An, N2Ha N2Sa, N2Su	N2Hi

TABLE II
HYPERPARAMETER SETTINGS OF MULTI-CHANNEL ENCODERS

	E_r	E_c	E_{f_e}
Conv Layers	1	3	1
Conv Dim	128	512	128
Norm Groups	8	32	8
BLSTM Layers	1	2	1
BLSTM Dim	1	8	16
Downsample Factor	8	8	8

TABLE III
HYPERPARAMETER SETTINGS OF EMOTION-SEPARATE ENCODERS

	$E^s_{\boldsymbol{U}}$	$E^{t-s}_{\boldsymbol{U}}$	$E^s_{\boldsymbol{Z}_f}$	$E^{t-s}_{\boldsymbol{Z}_f}$
Conv Layers	3	3	1	3
Conv Dim	512	256	128	256
Norm Groups	32	8	8	16
BLSTM Layers	2	1	1	1
BLSTM Dim	8	16	4	8
Downsample Factor	8	8	8	8

B. Experimental Settings

As described in Section II, the networks consist of the multi-channel encoders, the emotion-separate encoders, and the output decoder. All the encoders share a similar architecture, which consists of convolutional layers followed by group normalization [55]. The gray blocks at the top of the encoders shown in Fig. 1 are the designed information bottlenecks which are a stack of BLSTM layers. They are applied after the output of the convolutional layers to reduce the feature dimension. By using the designed bottlenecks, the information of each channel can be passed through a downsampling operation to reduce the temporal dimension, producing hidden representations. Table II shows the hyperparameter settings of multi-channel encoders, E_r, E_c and E_f. Table III shows the hyperparameter settings of second-level emotion-separate encoders ($E^s_{\boldsymbol{U}}$, $E^{t-s}_{\boldsymbol{U}}$, $E^s_{\boldsymbol{Z}_f}$ and $E^{t-s}_{\boldsymbol{Z}_f}$). The decoder first upsamples the hidden representation to restore the original sampling rate. Then all the representations are concatenated along the channel dimension and fed to a stack of three BLSTM layers [17] with an output linear layer to produce the final output. The output features are converted back to the speech waveform using the same wavenet-vocoder as in [56] on the VCTK corpus.

C. Comparative Study

To evaluate the proposed method, we reimplemented several state-of-the-art emotional VC models for comparison.

- *DBNs+NNs [21]:* This is the earliest emotional VC method based on deep learning models. The model uses the DBNs to convert spectral features and the NNs to convert the F0 features.
- *Dual-SANs [23]:* This model adopts the dual supervised adversarial network, in which continuous wavelet transform method was used to augment prosody (F0) features.
- *CycleGAN [25]:* The CycleGAN model has been widely used in the non-parallel VC tasks. Kun et al. have also used this unsupervised learning model in the emotional VC.
- *SPEECHFLOW [45]:* This is a state-of-the-art source-filter method that has been used in the normal speaker VC task. We applied it in the emotional VC that uses the designed bottlenecks auto-encoder for filtering the timbre and prosody features.
- *SFEVC:* This is our proposed method SFEVC that uses designed bottlenecks multi-channel encoders and the emotion-separate encoders, but without adding the VA training.
- *SFEVC+VA:* We adopt the novel VA-based training to train valence and arousal source-filters, denoted as SFEVC+VA. This model is built to validate the effectiveness of adding the VA training models for basic SFEVC.

We have conducted objective evaluations using the parallel utterance datasets for training for all methods. As the CycleGAN, SPEECHFLOW and the proposed method are parallel-data-free frameworks, they are also trained under non-parallel condition by upsetting the order for the source and target sentences.

D. Objective Evaluation

In the emotional VC task, the timbre and pitch features are mainly represented by mel-cepstral coefficients (MCC) and F0 features. Therefore, for objective evaluation, we use mel-cepstral distortion (MCD) to measure how close the converted MCC is to the target MCC in the mel-cepstral space, and root means square error (RMSE) to evaluate the conversion error between target F0 and converted F0. The MCD and F0-RMSE values are calculated using the different emotional voices from the same speaker since the emotional VC aims to convert the emotion-related features but keep the speaker identity unchanged. The averaged MCD and RMSE are shown in Table IV and Table V, respectively.

As shown in Table IV, the MCD values between source and target emotion speech samples are over 5, indicating that the timbre features of voices spoken by the same speaker also have some variations for different emotions. Comparing the MCD of the source voice and converted voice, all models show decreased values. Comparing Dual-SANs, SPEECHFLOW, SFEVC, and SFEVC+AV with CycleGAN, the MCD varies slightly for N2Sa but decreases significantly for the other high arousal emotions. This is because CycleGAN is effective for transferring sophisticated local texture appearance between image domains, but it has difficulties with objects that have both related appearance and shape changes. For the angry and happy voice conversion, their spectrum shape is more different from the neutral voice than the sad voice. This shows that although cycle-consistency is effective for the training of GANs when converting the neutral voice to a sad voice, the effect of cycle-consistency is no more than the dual supervised learning models (Dual-SANs) and

TABLE IV
MCD RESULTS FOR THE CONVERSION OF NEUTRAL VOICE TO EMOTIONAL VOICE USING DIFFERENT METHODS WITH THREE PUBLIC DATASETS AND ONE NEWLY RELEASED HIGH TENSION EMOTION DATASET

	ATR			JSUT		ESD				HTE
	N2An	N2Sa	N2Ha	N2An	N2Ha	N2An	N2Sa	N2Ha	N2Su	N2Hi
Source	6.03	5.18	6.30	6.28	6.13	6.47	6.64	6.22	6.49	6.15
DNNs+NNs	5.72	4.91	5.59	5.11	5.23	5.45	5.03	5.18	5.50	5.52
Dual-SANs	**3.82**	4.31	3.55	3.73	3.94	3.65	3.93	3.88	3.80	3.75
CycleGAN	4.52 (4.75)	4.41 (4.73)	4.45 (4.76)	4.64 (4.92)	4.51 (4.72)	4.57 (4.76)	4.46 (4.63)	4.32 (4.48)	4.68 (4.77)	4.52 (4.72)
SPEECHFLOW	3.85 (3.92)	4.33 (4.42)	3.45 (3.52)	3.79 (3.88)	**3.81** (3.92)	**3.35** (3.45)	4.03 (4.12)	3.78 (3.82)	3.90 (4.02)	**3.72** (3.82)
SFEVC	3.83 (3.88)	**4.20** (4.32)	3.48 (3.55)	3.72 (3.82)	3.85 (3.99)	3.45 (3.55)	3.93 (4.02)	3.68 (3.72)	**3.70** (3.77)	3.77 (3.85)
SFEVC+VA	3.84 (3.92)	4.22 (4.32)	**3.36** (3.52)	**3.65** (3.73)	3.91 (4.02)	3.43 (3.52)	**3.83** (3.99)	**3.65** (3.71)	3.80 (3.93)	**3.72** (3.82)

The scores in () represents the results trained under non-parallel condition.

TABLE V
F0-RMSE RESULTS FOR THE CONVERSION OF NEUTRAL VOICE TO EMOTIONAL VOICE USING DIFFERENT METHODS WITH THREE PUBLIC DATASETS AND ONE NEW RELEASED HIGH TENSION EMOTION DATASET

	ATR			JSUT		ESD				HTE
	N2An	N2Sa	N2Ha	N2An	N2Ha	N2An	N2Sa	N2Ha	N2Su	N2Hi
Source	75.5	55.8	81.2	77.3	85.5	72.4	83.3	71.1	60.3	79.5
DNNs+NNs	40.1	35.6	41.7	36.5	43.2	45.4	53.2	46.1	39.0	49.5
Dual-SANs	22.8	21.1	23.5	23.6	27.1	23.4	29.1	24.1	28.2	29.5
CycleGAN	28.2 (33.5)	23.5 (28.3)	26.6 (29.2)	29.9 (34.1)	26.2 (29.5)	30.1 (34.4)	32.3 (36.2)	27.8 (30.2)	31.3 (35.7)	33.2 (36.2)
SPEECHFLOW	23.2 (25.2)	22.3 (23.9)	21.5 (23.8)	27.5 (28.2)	22.1 (23.2)	25.2 (27.2)	26.03 (29.2)	23.1 (25.2)	29.3 (30.8)	31.2 (32.1)
SFEVC	21.8 (23.2)	19.9 (21.2)	20.4 (22.2)	19.7 (21.2)	19.9 (21.0)	25.4 (27.1)	23.3 (25.2)	25.1 (27.2)	24.3 (26.2)	27.9 (29.4)
SFEVC+VA	**17.7** (19.2)	**19.5** (21.3)	**19.3** (21.2)	**18.6** (19.9)	**18.1** (19.2)	**23.4** (24.9)	**21.9** (23.4)	**23.1** (24.2)	**22.3** (23.7)	**27.0** (29.2)

The scores in () represents the results trained under non-parallel condition.

the designed bottlenecks encoder models (SPEECHFLOW and SFEVC) in the emotional VC task, where the emotional features' shapes change a lot.

As shown in Table V, comparing the RMSE results, all models show significantly decreased values. The bottlenecks encoder models can obtain better results than DBNs+NNs and the GAN-related models (Dual-SANs and CycleGANs). In the bottlenecks encoder models, our proposed SFEVC and SFEVC+VA models get better RMSE values than the SPEECHFLOW in the neutral to high tension emotion and surprised emotion. This indicates that our emotion-separate encoder is effective in complex emotions. Comparing the results of SFEVC and SFEVC+VA, the latter has less errors for the conversion from neutral to angry.

Comparing the MCD results and RMSE results of the parallel training data and non-parallel training settings, their scores differ little for all source-filter models (SPEECHFLOW, SFEVC and SFEVC+VA). Moreover the source-filter models have a better results than the cycleGAN for the non-parallel data, especially for the small dataset (ATR and JSUT).

E. Subjective Evaluation

For the subjective experiment, similarity test and naturalness MOS test are used as evaluation metrics.

1) Emotion Similarity: In the emotional VC task, the more similar the converted voice sounds to the target emotion, the more effective the model is. Therefore, we carry out a subjective emotion classification test for the neutral voice to emotion pairs including Neutral-to-Angry, Neutral-to-Sad, Neutral-to-Happy, Neutral-to-Surprised, and Neutral-to-High Tension, comparing different models (DBNs+NNs, Dual-SANs, CycleGAN, SPEECHFLOW, and SFEVC and SFEVC+VA). For each test model, 30 native-speaker listeners including 15 Japanese (10 males and 5 females) and 15 Chinese (10 males and 5 females) of different ages are involved and asked to label the emotions of the converted voices in their respective mother language. The Japanese datasets do not include the surprised emotion and the Chinese dataset does not include the high tension emotion. Hence for the Japanese converted voice, 40 utterances are randomly selected from the evaluation sets for the converted angry, happy, sad and high tension voices (10 sentences for each emotion). For the Chinese converted voices, 40 utterances are randomly selected from the evaluation sets for the converted angry, happy, sad and surprised voices (10 sentences for each emotion). Finally, we sum up all the test models' results in Table VI.

As shown in Table VI(a), when evaluating the original recorded emotional speech utterances, all classification accuracy is higher than 90%, indicating the classifier performs well enough to be used in the emotion classification test. The classification accuracy of Neutral-to-Sad is nearly 100%, indicating that the human can easily separate the low valence emotion (sad) from the high valence voice (happy, angry, surprised and high tension). Because the high tension and happy voice are similar, the classification accuracy of Neutral-to-Happy and Neutral-to-High Tension is lower than the others. The classification results for the converted voices of DBNs+NNs, Dual-SANs, CycleGAN, SPEECHFLOW, and SFEVC and SFEVC+VA are shown in (b), (c), (d), (e), (f), and (g) of Table VI, respectively.

As shown in Table VI (b), the conventional DBNs+NNs method shows poor performance in all emotional VC tasks, especially for the conversion from neutral to angry and from neutral to high-tension. This result confirms that the DBNs+NNs model training without the GANs models (Dual-SANs and CycleGAN) or designed bottlenecks auto-encoder models (SPEECHFLOW,

TABLE VI
RESULTS OF CLASSIFICATION FOR RECORDED VOICES[%] AND CONVERTED VOICES [%]

(a) Recorded voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**99**	0	0	0	0	1
Angry	0	**96**	0	2	0	2
Surprised	0	0	**94**	2	2	2
High tension	0	2	2	**90**	6	0
Happy	0	0	2	6	**92**	0
Neutral	0	1	0	2	2	**95**

(b) DBNs+NNs converted voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**50**	2	3	0	0	45
Angry	5	**48**	2	10	0	35
Surprised	8	12	**44**	15	1	20
High tension	1	20	10	**28**	30	11
Happy	0	5	18	30	**33**	14

(c) Dual-SANs converted voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**68**	3	2	0	0	27
Angry	2	**61**	3	15	0	19
Surprised	0	5	**55**	25	10	5
High tension	0	3	10	**45**	32	10
Happy	0	0	7	32	**48**	13

(d) CycleGAN converted voice

Tar./Percept	Sad	Angry	Surprise	High Tension	Happy	Neutral
Sad	**62**	2	6	2	3	25
Angry	5	**40**	8	22	0	25
Surprised	0	5	**48**	17	20	10
High tension	0	10	15	**44**	26	5
Happy	0	2	10	26	**42**	20

(e) SPEECHFLOW converted voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**66**	0	0	0	0	1
Angry	0	**63**	0	2	0	2
Surprised	0	0	**54**	2	2	2
High tension	0	2	2	**50**	6	0
Happy	0	0	2	6	**52**	0

(f) SFEVC converted voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**69**	2	0	0	0	29
Angry	5	**66**	2	22	0	5
Surprised	0	4	**60**	16	12	8
High tension	0	7	7	**55**	26	5
Happy	0	3	12	22	**58**	5

(g) SFEVC+VA converted voice

Tar./Percept	Sad	Angry	Surprised	High Tension	Happy	Neutral
Sad	**68**	0	0	0	0	30
Angry	3	**69**	3	18	0	7
Surprised	0	3	**64**	22	6	3
High tension	0	8	3	**60**	25	4
Happy	0	2	6	26	**62**	4

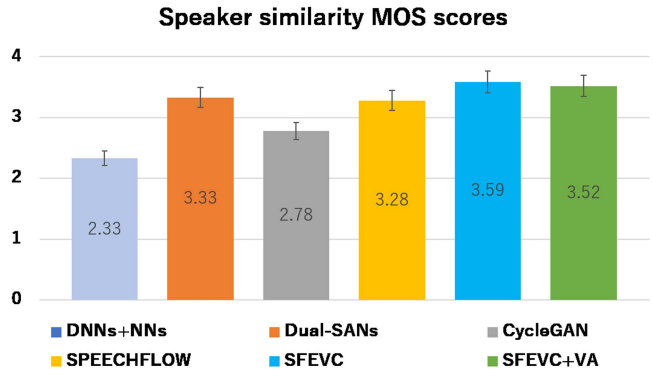

Fig. 8. MOS of the speaker similarity for the conversion of neutral voice to emotional voice, with 95% confidence intervals computed from the t-test.

SFEVC, and SFEVC+VA) cannot convert the emotion features well.

Comparing the results of GANs based models (Dual-SANs and CycleGAN) with the source-filter models (SPEECHFLOW, SFEVC, and SFEVC+VA), we see that the classification accuracy is similar for the converted angry voice and sad voice. However, GANs based models get lower quality for the converted surprised voice and converted high-tension voice, indicating that the source-filter models have better conversion efficacy for the more complex emotions (surprise and high tension).

Comparing the results among the source-filter methods, i.e. Table VI (e), (f) and (g), we see that the proposed SFEVC has obtained better results than the SPEECHFLOW. Moreover, comparing the results of SFEVC and SFEVC+VA, it proves that applying the two stages training method based on 2D VA space can improve the quality for all the high valence emotion conversions.

2) Speaker Similarity: Our emotional voice conversion task is converting the emotion but keeping speaker-dependent features of the source speaker unchanged. Therefore, we also conduct subjective experiments on the speaker similarity for different models (DBNs+NNs, Dual-SANs, CycleGAN, SPEECHFLOW, and SFEVC and SFEVC+VA). The similarity evaluation is conducted on the converted emotional voice and reference emotional voice of the source speaker pairs. 40 utterances are randomly selected from the evaluation sets for the converted angry, happy, sad and high tension voices (10 sentences for each emotion). Each pair is evaluated by 15 subjects (Japanese), with a score ranging from 1 (different speakers) to 4 (same speaker). Fig. 8 shows each model's average score of all the converted emotional voices. As shown in Fig. 8, the proposed SFEVC and SFEVC+VA models have higher scores than the other models, which indicates that the SFEVC models is more effective in keeping the speaker-dependent features unchanged. SFEVC and SFEVC+VA have similar results, which show that the valence arousal training stages retain the speaker-dependent features well.

3) Naturalness: In line with most previous works in the VC field, to measure naturalness, we conduct a MOS test for the naturalness evaluation by 30 native subjects including 15

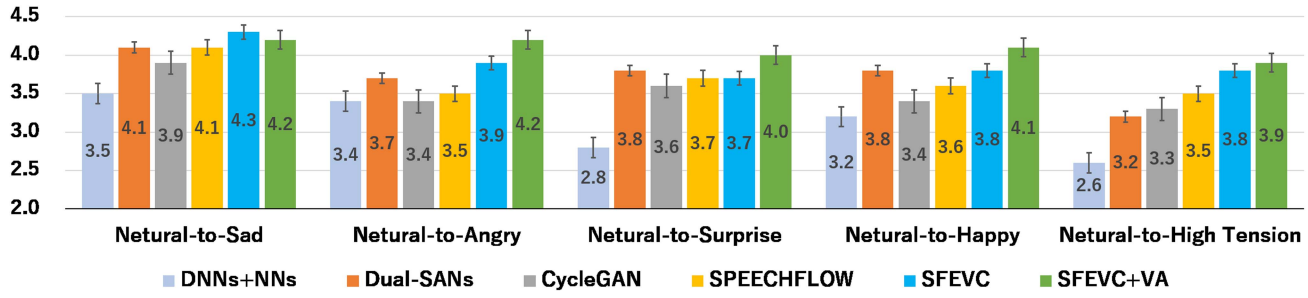

Fig. 9. MOS of the naturalness evaluation for the conversion of neutral voice to emotional voice, with 95% confidence intervals computed from the t-test.

Japanese (10 males and 5 females) and 15 Chinese (10 males and 5 females) of different ages. For each test model, 40 sentences are randomly selected from the Japanese evaluation sets including 10 utterances for each converted emotional voice (happy, angry, sad and high tension), and 40 sentences are randomly selected from the Chinese evaluation sets including 10 utterances for each converted emotion voice (happy, angry, sad and surprised). The scale ranged from 1 (totally unnatural) to 5 (completely natural). The results are shown in Fig. 9. In this test, a higher value indicates a better result, where the error bar shows the 95% confidence interval. From these results, we can see that all naturalness scores are above or near 3, which means that reasonable naturalness. Comparing the results of different models, we can see that the GANs models and source-filter models improved a lot than the DNBs+NNs. In the GANs models, the Dual-SANs gets more stable results than the CycleGANs, which has low naturalness in the neutral to high-tension converted voice. The source-filter models can get better results than the GANs models, especially for the high arousal voice conversion. Our proposed SFEVC model obtains better results than the other source-filter model. Especially for the conversion of neutral voice to high tension voice, the score is 0.3 higher than the other methods.

4) Ablation Study: In addition, we also conduct the ablation study for different encoders. To evaluate whether each bottleneck information is the expected information, we remove Z_r, Z_c, speaker-independent emotion codes ($Z_{\boldsymbol{U}}^t$ and $Z_{Z_f}^t$), or speaker-dependent codes ($Z_{\boldsymbol{U}}^s$ and $Z_{Z_f}^s$), separately. For each ablation test, we conduct the similarity evaluation (emotion similarity and speaker similarity) on the converted emotional voice and reference the emotional voice of the source speaker. 40 utterances are randomly selected from the evaluation sets for the converted angry, happy, sad, and high tension voices (10 sentences for each emotion). 15 subjects (Japanese) are involved to conduct the speaker similarity evaluation with a score ranging from 1 (different speakers) to 4 (same speaker) and emotion similarity evaluation with a score ranging from 1 (different emotions) to 4 (same emotion). As shown in Table VII, removing the content code or rhythm code will destroy the linguistic information of the speech, which leads to very low scores. Removing the speaker-independent emotion codes will result in around a 0.51 drop in emotion similarity scores and a slight decrease in speaker similarity scores, which indicate that our proposed emotion-separate encoder is effective in the emotion conversion. Removing the speaker-dependent codes will result in around a 0.3 drop in speaker similarity scores, which indicates the efficacy of the speaker-dependent encoder.

TABLE VII
ABLATION RESULTS W.R.T. THE CODES, I.E. CONTENT CODE (Z_c), RHYTHM CODE (Z_r), SPEAKER-INDEPENDENT EMOTION CODES ($Z_{\boldsymbol{U}}^t$ AND $Z_{Z_f}^t$), AND SPEAKER-DEPENDENT CODES ($Z_{\boldsymbol{U}}^s$ AND $Z_{Z_f}^s$), ON MOS TESTS OF EMOTION SIMILARITY (ES) AND SPEAKER SIMILARITY (SS)

Z_c	Z_r	$Z_{\boldsymbol{U}}^t$ and $Z_{Z_f}^t$	$Z_{\boldsymbol{U}}^s$ and $Z_{Z_f}^s$	ES	SS
	✓	✓	✓	1.21	1.12
✓		✓	✓	1.11	1.13
✓	✓		✓	2.81	3.51
✓	✓	✓		3.30	3.22
✓	✓	✓	✓	3.32	3.52

VI. CONCLUSION

In this work, we have presented a source-filter emotional voice conversion model applied with the multi-channel encoders and emotion-separate encoders, which can better decompose the timbre, pitch, content and rhythm features from the speech information and decouple the speaker-independent emotion features from the speaker-dependent emotion features. Moreover, we have also introduced a training strategy based on the valence-arousal space, which can improve the conversion expressiveness for the high arousal emotion. Experimental results show that our proposed SFEVC model is more effective in converting any emotions compared to the conventional method and achieved the state-of-the-art results.

APPENDIX A
DETAILED CALCULATION OF RECONSTRUCTING THE CONVERTED EMOTIONAL VOICE

Assume that the speech $\boldsymbol{X}$ can be reconstructed by (C,R,P,T) as follows:

$$\boldsymbol{X} = D(C, R, P, T) = D\left(C, R, \left(P^D, P^I\right), \left(T^D, T^I\right)\right) \quad (6)$$

where C, R, P, and T represent the content, rhythm, pitch, and timbre, respectively. The pitch and timbre features can be separated by speaker-independent features (P^I, T^I) and speaker-dependent features (P^D, T^D). In the emotional VC, for the input source emotion voice $\boldsymbol{X}_s = D(C_s, R_s, (P_s^D, P_s^I), (T_s^D, T_s^I))$, the main task is to replace speaker-independent emotion features (P_s^I, T_s^I) to the target

emotion features (P_t^I, T_t^I), but keep the other features unchanged as $\boldsymbol{X}_t = D(C_s, R_s, (P_s^D, P_t^I), (T_s^D, T_t^I))$. Combing 1, 2, 3 and 4 our emotional VC task is to reconstruct the converted voice as follows:

$$\begin{aligned}\hat{\boldsymbol{X}}_{s\to t} &= D\left(\boldsymbol{Z}_r, \boldsymbol{Z}_c, \boldsymbol{Z}_{Z_f}^t, \boldsymbol{Z}_{Z_f}^s, \boldsymbol{Z}_U^t, \boldsymbol{Z}_U^s\right) \\ &= D\Big(E_r(\boldsymbol{X}), E_c(A(\boldsymbol{X})), E_{\boldsymbol{Z}_f}^s\left(\boldsymbol{Z}_f^s\right), \\ &\quad E_{\boldsymbol{Z}_f}^{t-s}\left(\boldsymbol{Z}_f^t, \boldsymbol{Z}_f^s\right), E_{\boldsymbol{U}}^s\left(\boldsymbol{U}\right), E_{\boldsymbol{U}}^{t-s}\left(\boldsymbol{U^t}, \boldsymbol{U}\right)\Big) \\ &= \boldsymbol{X}_t = D\left(C_s, R_s, \left(P_s^D, P_t^I\right), \left(T_s^D, T_t^I\right)\right), \end{aligned} \tag{7}$$

which achieves 0 reconstruction loss in 5.

To train the speaker-independent emotion encoders, the inputs are the speech in different emotions $(\boldsymbol{X}_s, \boldsymbol{X}_t)$ spoken by the same speaker X. Thus, the content (C_s), rhythm (R_s), and speaker-dependent features (P_s^D, T_s^D) are the same information. Therefore, the loss function to minimize the $||\hat{\boldsymbol{X}}_{s\to t} - \boldsymbol{X}_t||$ depends on reconstructing the speaker-independent emotion features (P^I, T^I) by encoders $(E_{\boldsymbol{Z}_f}^{t-s}\ E_{\boldsymbol{U}}^{t-s})$ and decoder D:

$$\min_{E_{\boldsymbol{Z}_f}^{t-s}(\cdot), E_{\boldsymbol{U}}^{t-s}(\cdot), D(\cdot)} \mathbb{L}||\hat{\boldsymbol{X}}_{s\to t} - \boldsymbol{X}_t|| = \mathbb{L}_{P^I} + \lambda_1 \mathbb{L}_{T^I}, \tag{8}$$

where,

$$\begin{aligned}\mathbb{L}_{P^I} &= \mathbb{E}\left[||E_{\boldsymbol{Z}_f}^{t-s}\left(\boldsymbol{Z}_f^t, \boldsymbol{Z}_f^s\right) - P_t^I||_1\right] \\ &= \mathbb{E}\left[||E_{\boldsymbol{Z}_f}^{t-s}\left(E_f\left(A\left(\boldsymbol{P}\right)\right), E_f\left(A\left(\boldsymbol{P^t}\right)\right) - P_t^I||_1\right], \\ \mathbb{L}_{T^I} &= \mathbb{E}\left[||E_{\boldsymbol{U}}^{t-s}(\boldsymbol{U^t}, \boldsymbol{U}) - T_t^I||_1\right] \end{aligned} \tag{9}$$

and the weight λ_1 is set to 1.

The speaker-dependent encoders are pre-trained using the speech in the same emotion $(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ by different speakers X and Y. $\boldsymbol{Y}_s$ can be represented as $\boldsymbol{Y}_s = D(C_s, R_s, (P_y^D, P_s^I), (T_y^D, T_s^I))$, where, C_s, R_s, P_s^I, T_s^I are the same information as $\boldsymbol{X}_s$. Therefore, the pre-trained speaker-dependent encoders only focus on the conversion of speaker-dependent features (P^D, T^D). Our emotional VC task is to keep the speaker-dependent features decoupled from timbre and pitch features of $\boldsymbol{X}_s$ unchanged. Then, the loss function to minimize the $||\hat{\boldsymbol{X}}_{s\to t} - \boldsymbol{X}_s||$ depends on reconstructing the source speaker-dependent features (P^D, T^D) as:

$$\min_{E_{\boldsymbol{Z}_f}^{s}(\cdot), E_{\boldsymbol{U}}^{s}(\cdot), D(\cdot)} \mathbb{L}||\hat{\boldsymbol{X}}_{s\to t} - \boldsymbol{X}_s|| = \mathbb{L}_{P^D} + \lambda_2 \mathbb{L}_{T^D}, \tag{10}$$

where,

$$\begin{aligned}\mathbb{L}_{P^D} &= \mathbb{E}\left[||E_{\boldsymbol{Z}_f}^{s}\left(\boldsymbol{Z}_f^s\right) - P_s^D||_1\right] \\ &= \mathbb{E}\left[||E_{\boldsymbol{Z}_f}^{s}\left(E_f\left(A\left(\boldsymbol{P}\right)\right)\right) - P_s^D||_1\right], \\ \mathbb{L}_{T^D} &= \mathbb{E}\left[||E_{\boldsymbol{U}}^{s}\left(\boldsymbol{U}\right) - T_s^D||_1\right] \end{aligned} \tag{11}$$

and the weight λ_2 is set to 1.

REFERENCES

[1] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," *EURASIP J. Audio Speech Music Process.*, vol. 2014, no. 1, pp. 1–10, 2014.

[2] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 1093–1096.

[3] J. Krivokapić, "Rhythm and convergence between speakers of American and Indian english," *Lab. Phonol.*, vol. 4, no. 1, pp. 39–65, 2013.

[4] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation of vocoded speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 254–258.

[5] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. Interspeech*, 2013, pp. 369–372.

[6] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities and evaluation of dual learning," Inst. Electron., Inf. Commun. Engineers, Tokyo, Japan, Tech. Rep., 2017.

[7] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 266–273.

[8] S. Lee, B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6279–6283.

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 912–921, Jul. 2010.

[11] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2012, pp. 313–317.

[12] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2584–2594, Nov. 2017.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[15] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 139–144.

[16] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. Interspeech*, 2016, pp. 287–291.

[17] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4869–4873.

[18] 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, no. 5, pp. 134–138, 2012.

[19] 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.

[20] C.-C. Hsia, C.-H. Wu, and J.-Q. Wu, "Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1245–1254, Sep. 2007.

[21] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci.*, 2016, pp. 1–5.

[22] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP J. Audio Speech Music Process.*, vol. 2017, no. 1, 2017, pp. 1–18.

[23] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform F0 features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 10, pp. 1535–1548, Oct. 2019.
[24] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional modeling of timbre and prosody for emotional voice conversion," in *Proc. Interspeech*, 2016, pp. 2453–2457.
[25] 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.
[26] A. M. Taylor and D. Reby, "The contribution of source–filter theory to mammal vocal communication research," *J. Zool.*, vol. 280, no. 3, pp. 221–236, 2010.
[27] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *J. Acoustical Soc. Amer.*, vol. 144, no. 2, pp. 908–916, 2018.
[28] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," *Rev. Neurosciences*, vol. 15, no. 4, pp. 241–252, 2004.
[29] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.
[30] V. Kuperman, Z. Estes, M. Brysbaert, and A. B. Warriner, "Emotion and language: Valence and arousal affect word recognition," *J. Exp. Psychol. Gen.*, vol. 143, no. 3, 2014, Art. no. 1065.
[31] R. Adolphs, "Recognizing emotion from facial expressions: Psychological and neurological mechanisms," *Behav. Cogn. Neurosci. Rev.*, vol. 1, no. 1, pp. 21–62, 2002.
[32] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *J. Pers. Social Psychol.*, vol. 76, no. 5, pp. 805–819, 1999.
[33] S. Lin, "A new frequency coverage metric and a new subband encoding model, with an application in pitch estimation," in *Proc. Interspeech*, 2018, pp. 2147–2151.
[34] S. Lin, "Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-bernoulli filter," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 4, pp. 827–841, Apr. 2019.
[35] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of Emotions*, New York: Guilford, vol. 2. 2000,pp. 220–235.
[36] J. Pittam, C. Gallois, and V. Callan, "The long-term spectrum and perceived emotion," *Speech Commun.*, vol. 9, no. 3, pp. 177–187, 1990.
[37] B. K. Dichter, J. D. Breshears, M. K. Leonard, and E. F. Chang, "The control of vocal pitch in human laryngeal motor cortex," *Cell*, vol. 174, no. 1, pp. 21–31, 2018.
[38] M. Belyk and S. Brown, "The origins of the vocal brain in humans," *Neurosci. Biobehavioral Rev.*, vol. 77, pp. 177–193, 2017.
[39] W. Ma and W. F. Thompson, "Human emotions track changes in the acoustic environment," in *Proc. Nat. Acad. Sci.*, vol. 112, no. 47, 2015, pp. 14563–14568.
[40] K. N. Stevens, *Acoustic Phonetics (Current Studies in Linguistics; 30)*. Cambridge, MA, USA: MIT Press, 1998.
[41] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Process.*, vol. 5, no. 3, pp. 267–285, 1983.
[42] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 402–415, 2020.
[43] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted machines for voice conversion," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2278–2282.
[44] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5210–5219.
[45] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7836–7846.
[46] M. B. Arnold, "Emotion and personality," 1960.
[47] 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.
[48] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, 2022.
[49] C. M. Whissell, "The dictionary of affect in language," in *The Measurement of Emotions*. Amsterdam, Netherlands: Elsevier, 1989, pp. 113–131.
[50] 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.
[51] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[52] R. G. Kamiloğlu, A. H. Fischer, and D. A. Sauter, "Good vibrations: A review of vocal expressions of positive emotions," *Psychon. Bull. Rev.*, vol. 27, no. 2, pp. 237–265, 2020.
[53] 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.
[54] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale japanese speech corpus for end-to-end speech synthesis," 2017, *arXiv:1711.00354*.
[55] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[56] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 4779–4783.

Zhaojie Luo (Member, IEEE) received the M.Eng. and Dr.Eng. degrees in system informatics from Kobe University, Kobe, Japan, in 2017 and 2020 respectively. He is currently an Assistant Professor with Osaka University, Suita, Japan. From 2019 to 2020, he was a Researcher with the Department of Electrical & Computer Engineering, National University of Singapore, Singapore. He has authored or coauthored more than 20 papers in top-tier speech/multimedia journals and international conferences, such as IEEE-T-ASLP, IEEE TRANSACTIONS ON MULTIMEDIA, EURASIP JASMP, INTERSPEECH, SSW, ICME, ICPR. His research interests include voice conversion, speech synthesis, facial expression recognition, multimodal emotion recognition, and statistical signal processing. He is a Member of ISCA and ASJ, and is a reviewer for many major referred journal and conference papers.

Shoufeng Lin (Senior Member, IEEE) received the B.Eng. and M.Eng. (Hons.) degrees in electronics science and technology from Zhejiang University, Hangzhou, China, in 2004 and 2006 respectively, and the Ph.D. degree in signal processing from Curtin University, Bentley WA, Australia, in 2019. He is a Research Fellow with the Department of Electrical & Computer Engineering of National University of Singapore, Singapore, with the Human Language Technology (HLT) lab, where he was a discipline lead and managing a high impact research project (with $4 million Singapore dollar fund). Prior to this, he had held senior engineering and management positions with well-established and start-up companies including Huawei, National Instruments and General Electric. He has a plethora of successful experiences in R&D and production of high performance embedded electronics products including noise suppression and wireless communication hearing aids. He has first author publications in top-tier journals and international conferences, such as IEEE/ACM-TASLP, ICASSP and INTERSPEECH, and was a reviewer of many journal and conference papers.

Rui Liu (Member, IEEE) received the bachelor's degree from the Taiyuan University of Technology, Taiyuan, China, in 2014, and the Ph.D. degree from Inner Mongolia University, Hohhot, China, in 2020. From 2019 to 2020, he received the exchange Ph.D. degree from the Department of Electrical & Computer Engineering of National University of Singapore (NUS), Singapore, funded by China Scholarship Council. He is currently a Professor with the National and Local Joint Engineering Research Center of Mongolian Intelligent Information Processing, Inner Mongolia University. From 2020 to 2022, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore. He has authored or coauthored more than 20 papers in top-tier NLP/ML/AI conferences and journals, including IEEE/ACM-TASLP, Neural Networks, ICASSP, COLING, INTERSPEECH. His research interests broadly incldue audio, speech and natural language processing, which include expressive text-to-speech (TTS), expressive voice conversion, speech emotion recognition, prosody structure prediction, grapheme-to-phoneme conversion (G2P), syntax parsing. He was the recipient of the Best Paper Award at the 2021 International Conference on Asian Language Processing (IALP). He is a Member of IEEE, ISCA and CCF, and is a reviewer for many major referred journal and conference papers.

Jun Baba (Member, IEEE) received the M.E. degree in informatics from Kyoto University, Kyoto, Japan, in 2014. He was a Data Scientist at CyberAgent, Inc. in Tokyo, Japan, from 2014 to 2017. He has been a Research Scientist with CyberAgent AI Laboratory and a Visiting Researcher with Osaka University, Suita, Japan, since 2017. His research interests include teleoperation for social robots, human-computer interaction in service encounter, and artificial intelligence.

Yuichiro Yoshikawa (Member, IEEE) received the Ph.D. degree in engineering from Osaka University, Suita, Japan, in 2005. From 2005, he was a Researcher with Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International. Since 2010, he has been an Associate Professor with the Graduate School of Engineering Science, Osaka University. He is a Member of Japanese Society of Robotics, Japanese Society of Cognitive Science, the Virtual Reality Society of Japan, Japanese Society for Child and Adolescent Psychiatry, and Japanese Society of Pediatric Psychiatry and Neurology.

Hiroshi Ishiguro (Member, IEEE) received the D.Eng. degree in systems engineering from Osaka University, Osaka, Japan, in 1991. He is currently a Professor of the Department of Systems Innovation with the Graduate School of Engineering Science, Osaka University, Osaka, Japan, and a Distinguished Professor of Osaka University. He is also the Visiting Director of Hiroshi Ishiguro Laboratories with Advanced Telecommunications Research Institute International and an ATR Fellow. His research interests include sensor networks, interactive robotics, and android science.