IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 13, 2021, accepted February 16, 2021, date of publication March 10, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065460

Sequence-to-Sequence Emotional Voice Conversion With Strength Control

HEEJIN CHOI AND MINSOO HAHN
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
Corresponding author: Heejin Choi (change@kaist.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2B5B01001972).

ABSTRACT This paper proposes an improved emotional voice conversion (EVC) method with emotional strength and duration controllability. EVC methods without duration mapping generate emotional speech with identical duration to that of the neutral input speech. In reality, even the same sentences would have different speeds and rhythms depending on the emotions. To solve this, the proposed method adopts a sequence-to-sequence network with an attention module that enables the network to learn attention in the neutral input sequence should be focused on which part of the emotional output sequence. Besides, to capture the multi-attribute aspects of emotional variations, an emotion encoder is designed for transforming acoustic features into emotion embedding vectors. By aggregating the emotion embedding vectors for each emotion, a representative vector for the target emotion is obtained and weighted to reflect emotion strength. By introducing a speaker encoder, the proposed method can preserve speaker identity even after the emotion conversion. Objective and subjective evaluation results confirm that the proposed method is superior to other previous works. Especially, in emotion strength control, we achieve in getting successful results.

INDEX TERMS Voice conversion, emotional voice conversion, emotion strength, sequence-to-sequence learning, controllable emotional voice conversion.

I. INTRODUCTION

Voice conversion (VC) refers to a technique of converting voice characteristics while preserving the linguistic information of an input utterance. The technique has been mainly used to change voice color. Voice characteristics contain not only the speaker identity information but also the pitch, speech rate, prosody, and emotion. Among them, we will focus on changing the emotion. Desirable emotional voice conversion (EVC) can transform the emotion without damaging linguistic information or speaker identity.

The general procedure of VC is to extract the acoustic features of the source voice at first, then to map these acoustic features onto those of the target voice, and finally, to synthesize the waveforms using the converted acoustic features. Before training the mapping function, dynamic time warping (DTW) [1] is used to achieve the time alignment between source and target acoustic features when utilizing parallel utterance data in pairs from different speakers. A statistical feature mapping function such as a Gaussian mixture model (GMM) is often employed in conventional VC [2], [3]. With deep neural networks (DNNs), speech processing tasks, including VC, have been developed so rapidly. Several techniques based on DNNs, such as a feed-forward neural network (FNN) [4], [5] and a recurrent neural network (RNN) [6], have been proposed to convert the acoustic features of the source speaker into those of the target speaker. Considering the difficulty in collecting parallel utterances from different speakers, recent studies have examined a restricted Boltzmann machine (RBM), a variational autoencoder (VAE), and a generative adversarial network (GAN) for non-parallel VC [7]–[13].

EVC techniques have evolved similarly to VC techniques. As prosody plays an important role in expressing emotional speech, several studies have focused on modelling spectral and fundamental frequency (F0) features with parallel data. Some previous works have explored prosody and spectral mapping separately using GMM [14]–[16], FNN [17], deep belief network (DBN) [18], and GAN [19] methods. Ming *et al.* [20] converted the spectrum and F0

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

simultaneously with bidirectional long-short term memory (LSTM) using parallel data. To circumvent the need for parallel data, some recent studies have explored disentangled representations of emotion using autoencoder [21] and VAE [22] approaches. Zhou *et al*. [23] demonstrated simultaneous spectrum and F0 conversion based on CycleGAN [24] using non-parallel data. While these methods had been fairly successful in single-speaker tasks, they have inevitable limitations when adopted for multi-speaker cases. In other words, they usually fail on multi-speaker datasets because features from different emotions tend to overlap considerably among various speakers. In an attempt to solve this, multi-speaker EVC systems [25], [26] that correspondingly modified the pitch and the energy, using a highway network [27] and a convolutional GAN network [28], were introduced. One of the recent studies [29] demonstrated variational autoencoding Wasserstein generative adversarial network (VAWGAN) based EVC utilizing the continuous wavelet transform (CWT) decomposition of F0 which allows the network to learn the speaker-independent emotion pattern across different speakers.

Even all the above approaches can rather successfully change emotions and retain the linguistic content and speaker identity, but the converted speech length remains the same as the input speech length. In reality, speakers tend to speak at different rates with different rhythms for their emotions. Therefore, for successful EVC, it becomes inevitable for a converted speech to have properly changed duration with suitable energy and pitch depending on the emotions.

To address this, we propose a novel EVC method that can synthesize emotional output speech with adjusted duration using a sequence-to-sequence network. Similar to an earlier approach [21], a speech is modelled separately into content features and emotion features. To learn emotion representations, an emotion encoder based on the style transfer method [30] is added to the sequence-to-sequence network. We utilize a multi-speaker emotional dataset to include various speakers' emotional patterns so that speaker-independent training is used to capture universal attributes embedded in multiple speakers and the speaker identity information is also be used as a supplement. By adding the speaker information, the proposed system can transform neutral speech into emotional speech without speaker change. As an extension, by scaling the emotion embedding vector values, we can also control the emotion strength. Namely, the emotion level can be adjusted to become stronger or weaker than the normal target level.

Although the mapping network is well trained, some factors such as the parameterization error and the over-smoothing effect still degrade the output voice quality when converted with a traditional parametric vocoder using F0 and spectrum. To deal with the problem, several studies have applied WaveNet-based waveform generation [31] to synthesize waveforms using converted acoustic features [32]–[34]. Thus, we utilize Parallel WaveGAN vocoder [35], which is also a WaveNet-based model, to recover speech waveforms and adopt mel-spectrograms as acoustic features.

The organization of this paper is as follows. Section II reviews the related works on sequence-to-sequence VC, unsupervised expressive modelling, and emotion strength control modelling. Section III describes the proposed EVC method. The procedures and results of the experiments are presented in Section IV and finally, Section V concludes this paper.

II. RELATED WORKS

A. SEQUENCE-TO-SEQUENCE VOICE CONVERSION

Given that the temporal lengths of the input and the output features are not equal, a sequence-to-sequence network makes it possible to map their alignments through an attention module. Several techniques have been proposed for sequence-to-sequence VC leveraging an automatic speech recognition (ASR) system [36]–[39] or a text-to-speech (TTS) system [40]. There also have been attempts to use only acoustic features for learning sequence-to-sequence VC without requiring a transcript [41], [42]. To accelerate and stabilize the training procedure, the concepts of the guided attention loss [43] and the context preservation loss were considered. Instead of separately learning mappings between each speaker pair, Kameoka *et al.* [42] achieved many-to-many VC with a speaker index as an additional input.

Compared to VC, which can utilize the same index for multiple speeches spoken by the same speaker, the emotional characteristics of EVC require delicate modelling, as they vary among the multiple speeches or within a single speech. In our work, we also leverage a sequence-to-sequence network without any transcript for emotion duration mapping but achieve one-to-many EVC by adding an emotion encoder that derives individual emotion information for each speech to deal with emotion characteristics.

B. UNSUPERVISED EXPRESSIVE MODELLING

To produce a human-like voice, various TTS systems have been proposed to model the expressive elements of speech [44], [30]. Skerry-Ryan *et al.* [44] introduced the concept of prosody embedding. A prosody embedding vector is computed through a prosody encoder and is used as an additional input to Tacotron [45] so that not monotonous but expressive speech can be synthesized. In another study [30], the prosody embedding vector is passed to a style token layer. The style token layer consists of an attention module and randomly initialized embedding vectors, referred to as global style tokens (GSTs). The attention module learns a similarity measure between the prosody embedding vector and each token in the GSTs and outputs a set of combination weights. The weighted sum of the GSTs called a style embedding vector is applied to Tacotron to enable the unsupervised style control and transfer.

In our previous work [46], given a text and a speaker index, the mel-spectrogram of the desired emotion was fed

into a prosody encoder to extract an appropriate emotion embedding vector. The synthesized speech then successfully showed the target emotion. Motivated in a way similar to the above previous studies, we use the prosody encoder and the style token layer as the emotion encoder for training. We appropriately handle emotion embeddings obtained from the emotion encoder to compute representative emotion vectors, then manipulate them in inference to select the desired emotion.

C. EMOTION STRENGTH CONTROL MODELLING

To synthesize expressive speech, approaches for controlling emotion strength were mainly addressed in the TTS field. In [47], a multidimensional scaling method is adopted to project the arousal-valence (AV) space for continuous emotion modelling. To define the AV values, annotators were asked to label the AV values for each utterance. It has been a great challenge for human annotators to label subtle emotion strength levels. Naturally, there have been several attempts to obtain emotion strength information without manual works. Zhu and Xue [48] applied a K-means clustering algorithm to partition emotion strength levels for a speech corpus and developed an embedding vector that continuously represents the emotion strength using a t-distributed stochastic neighbor embedding (t-SNE) algorithm [49]. Zhu *et al.* [50] also applied the concept of a relative attribute to learn a ranking function for each emotion category and controlled the emotion strength continuously. Um *et al.* [51] employed an inter-to-intra distance ratio algorithm that considered embedding distances between inter- and intra-categorical style token weights and applied an interpolation technique to change the emotion strength level.

Inspired by the controllable emotion strength in TTS, we propose an EVC method that controls the degree of the target emotion. We design the representative emotion-weighted vectors from the representative emotion vectors so that the proposed method can control the emotion strength in inference.

III. SEQUENCE-TO-SEQUENCE EMOTIONAL VOICE CONVERSION

In this section, we describe the proposed EVC method. We augment a sequence-to-sequence network with additional two encoders and two decoders: an emotion encoder, a speaker encoder, a source decoder, and a target decoder.

A. FEATURES

First, we define the acoustic features for the input and the output of the network. Conventional VC studies have shown that vocoder parameters such as F0, aperiodicity, and spectrum performed well when used to represent voice characteristics. Recently, neural vocoders have been successfully applied to TTS and VC [32]–[34] [37], [52], [53]. Mel-spectrogram effectively implies various information in speech, not only linguistic but also non-linguistic, such as the speaker and the emotion. Hence, in this paper, we utilize mel-spectrograms for acoustic features.

Mel-spectrograms are computed with a logarithmic mel-filterbank and a short-time Fourier transform (STFT). For training, log mel-spectrograms are mean-variance normalized individually for the source and the target.

B. MODEL

Similar to Tacotron2 [53], our sequence-to-sequence network contains a source encoder, a target encoder, an attention module, an autoregressive (AR) decoder, and a postnet. The encoders transform acoustic features into content embedding matrices, as follows,

$$\mathbf{H}_s = \text{SourceEncoder}(\mathbf{X}_s) \tag{1}$$

$$\mathbf{H}_t = \text{TargetEncoder}(\mathbf{X}_t), \tag{2}$$

where $\mathbf{X}_s = [\mathbf{x}_s^{(1)},\ldots, \mathbf{x}_s^{(Ns)}]$ denotes the sequence of the source acoustic features, $\mathbf{X}_t = [\mathbf{x}_t^{(1)},\ldots, \mathbf{x}_t^{(Nt)}]$ denotes that of the target acoustic features, $\mathbf{H}_s = [\mathbf{h}_s^{(1)},\ldots, \mathbf{h}_s^{(Ns)}]$ represents the source content embedding matrix, and $\mathbf{H}_t = [\mathbf{h}_t^{(1)},\ldots, \mathbf{h}_t^{(Nt)}]$ represents the target content embedding matrix. N_s and N_t represent the source and the target sequence lengths, respectively. The target encoder acts as a pre-net in earlier work [53]. For each decoder output step, an attention module is used to summarize the full source content embedding matrix as a fixed-length attention context vector. The attention context vector and the previous time-step target content embedding vector are passed to the AR decoder:

$$\mathbf{c}^{(n)}, \mathbf{a}^{(n)} = \text{Attention}(\mathbf{H}_s, \mathbf{l}^{(n-1)}, \mathbf{a}^{(n-1)}) \tag{3}$$

$$\mathbf{l}^{(n)}, \mathbf{d}^{(n)}, \tilde{o}^{(n)} = \text{ARDecoder}(\mathbf{c}^{(n)}, \mathbf{h}_t^{(n-1)}) \tag{4}$$

Here, $\mathbf{c}^{(n)}$ represents the attention context vector and $\mathbf{a}^{(n)}$ represents the attention probability vector at time-step n. $\mathbf{l}^{(n)}$ is a hidden representation vector of the AR decoder, $\mathbf{d}^{(n)}$ is the output vector of the AR decoder, and $\tilde{o}^{(n)}$ is the probability of the stop token at time-step n. We use location-sensitive attention [54], which extends content-based attention [55] to become location-aware by referring to the attention probability vector produced at the previous step.

To improve possible over-smoothed acoustic features from the AR decoder and to incorporate past and future frames, a postnet predicts the residual which would be added to the output of the AR decoder,

$$\tilde{\mathbf{X}}_{seq} = \mathbf{D} + \text{PostNet}(\mathbf{D}), \tag{5}$$

where $\tilde{\mathbf{X}}_{seq}$ denotes the predicted acoustic features of the sequence-to-sequence network and $\mathbf{D}$ is the output matrix of the AR decoder. In the inference procedure, the output of the AR decoder replaces the target acoustic features in (6). Fig. 1 shows the overall architecture of our sequence-to-sequence network.

$$\mathbf{H}_t = \text{TargetEncoder}(\mathbf{D}), \tag{6}$$

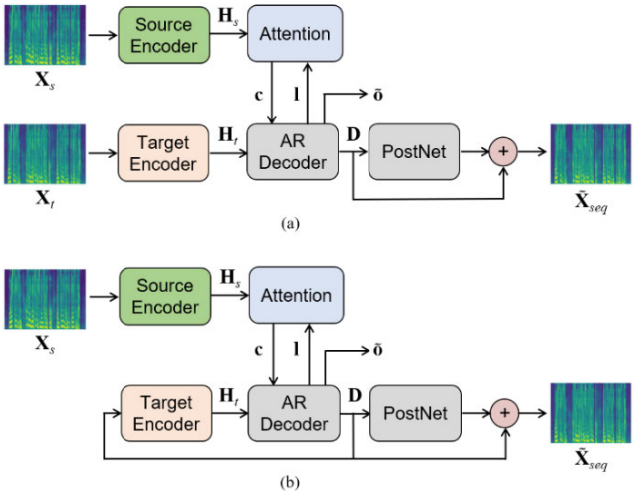

FIGURE 1. The sequence-to-sequence network for emotional voice conversion: (a) training procedure (b) inference procedure.

1) SPEAKER ENCODER

To separate the speaker information from emotion embeddings and represent the inherent characteristics of each speaker, we adopt a scheme similar to one in the literature [56] for a speaker encoder that transforms a speaker index into a speaker embedding vector. The speaker embedding vector $\mathbf{p}$ is combined with the source content embedding matrix via broadcast concatenation, with (3) then modified as follows:

$$\mathbf{c}^{(n)}, \mathbf{a}^{(n)} = \text{Attention}([\mathbf{p}; \mathbf{H}_s], \mathbf{l}^{(n-1)}, \mathbf{a}^{(n-1)}) \tag{7}$$

Here, [;] is for concatenation. For concatenation of a matrix and a vector, the fixed vector expands across all time-steps of the matrix.

2) EMOTION ENCODER

An emotion encoder takes the acoustic features as the input and computes the following emotion embedding vectors,

$$\mathbf{e}_s = \text{EmotionEncoder}(\mathbf{X}_s) \tag{8}$$
$$\mathbf{e}_t = \text{EmotionEncoder}(\mathbf{X}_t), \tag{9}$$

where $\mathbf{e}_s$ and $\mathbf{e}_t$ correspondingly denote the emotion embedding vectors of the source and the target. The emotion encoder is based on the style model [30] which compresses the style embedding vectors from the acoustic features. The style consists of rich information such as intention and emotion, but if the style model is built for emotional data, the style embedding vector can be specified as the emotion embedding vector. Similar to the speaker embedding vector in Section III-B-1, the target emotion embedding vector is also broadcast-concatenated with the source content embedding matrix, and (7) is modified as follows:

$$\mathbf{c}^{(n)}, \mathbf{a}^{(n)} = \text{Attention}([\mathbf{e}_t; \mathbf{p}; \mathbf{H}_s], \mathbf{l}^{(n-1)}, \mathbf{a}^{(n-1)}) \tag{10}$$

3) SOURCE DECODER AND TARGET DECODER

To ensure that the content embedding matrices of the source and the target preserve the contextual information, we reconstruct those matrices by a source decoder and a target decoder acting as an autoencoder. These decoders are inspired by the earlier work [41], which prevents the failure of the training for the sequence-to-sequence network. The emotion and the speaker embedding vectors are concatenated with the content embedding matrices and fed into the decoders,

$$\tilde{\mathbf{X}}_s = \text{SourceDecoder}[\mathbf{e}_s; \mathbf{p}; \mathbf{H}_s] \tag{11}$$
$$\tilde{\mathbf{X}}_t = \text{TargetDecoder}[\mathbf{e}_t; \mathbf{p}; \mathbf{H}_t], \tag{12}$$

where $\tilde{\mathbf{X}}_s$ and $\tilde{\mathbf{X}}_t$ indicate the predicted acoustic features of the source and the target decoders, respectively.

4) LOSS

Fig. 2 presents the detailed structure of the proposed architecture. All of these components are jointly trained from scratch. The training of the overall loss for the sequence-to-sequence network includes L1 and L2 losses from before and after the postnet to facilitate convergence as well as the binary cross-entropy (BCE) loss of the stop token prediction. Hence the loss function of the sequence-to-sequence network, L_{seq} can be written as,

$$\begin{aligned} L_{seq} = {} & \text{L1}(\tilde{\mathbf{X}}_{seq}, \mathbf{X}_t) + \text{L2}(\tilde{\mathbf{X}}_{seq}, \mathbf{X}_t) \\ & + \text{L1}(\mathbf{D}, \mathbf{X}_t) + \text{L2}(\mathbf{D}, \mathbf{X}_t) + \text{BCE}(\tilde{\mathbf{o}}, \mathbf{o}), \end{aligned} \tag{13}$$

where $\tilde{\mathbf{o}}$ represents the prediction of the stop token and $\mathbf{o}$ is the true label of the stop token.

To preserve the linguistic information in the content embedding matrices, the source and the target decoder outputs have to be reflected in the overall loss as shown below.

$$L_{src} = \text{L1}(\tilde{\mathbf{X}}_s, \mathbf{X}_s) + \text{L2}(\tilde{\mathbf{X}}_s, \mathbf{X}_s) \tag{14}$$
$$L_{trg} = \text{L1}(\tilde{\mathbf{X}}_t, \mathbf{X}_t) + \text{L2}(\tilde{\mathbf{X}}_t, \mathbf{X}_t) \tag{15}$$

In these equations, L_{src} and L_{trg} are the corresponding loss functions of the source decoder and the target decoder, respectively.

In practice, the attention module is quite costly to learn; accordingly, we consider that the acoustic features of the source and the target are a pair of parallel sequences with identical linguistic contents uttered with neutral and emotional speaking styles, respectively. It can be assumed that the content of the n_sth frame of the source acoustic features is identical to the content of the n_tth frame of the target acoustic features. Consequently, we apply guided attention in [43] to increase the training efficiency. We impose a constraint on the attention probability matrix such that it is nearly diagonal and strongly penalize it with the corresponding loss function,

$$L_{att} = \text{L1}(\mathbf{W} \odot \mathbf{A}), \tag{16}$$

where $\odot$ denotes the elementwise product, $\mathbf{W}$ is a non-negative weight matrix whose $w^{(ns)(nt)}$ value is defined as $w^{(ns)(nt)} = 1 - \exp\{-(n_s/N_s - n_t/N_t)^2 / (2g^2)\}$, n_s and n_t are the corresponding index frames of the source and target

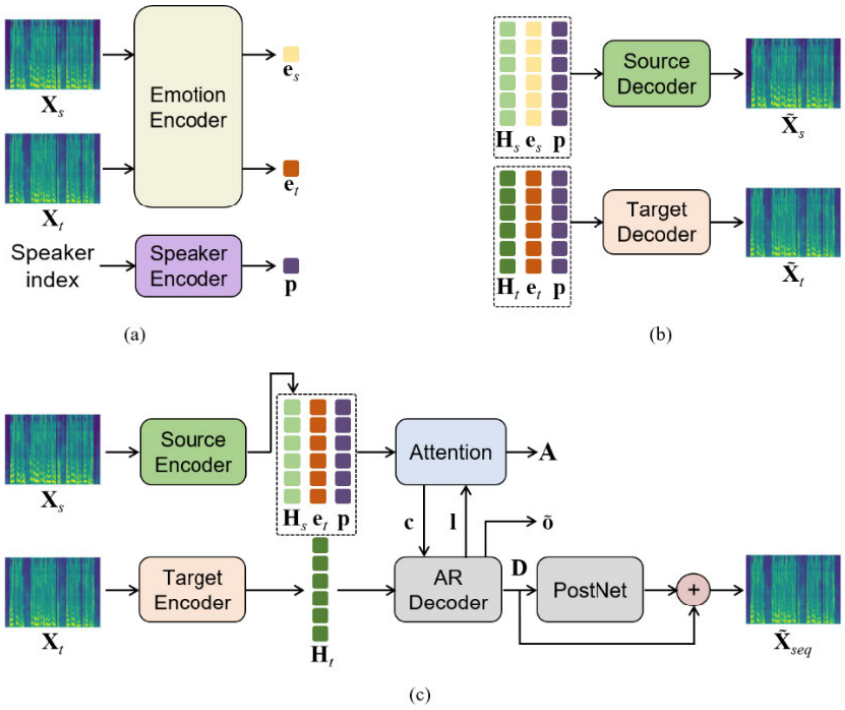

FIGURE 2. Training procedure of the proposed emotional voice conversion architecture: (a) emotion encoder and speaker encoder, (b) source decoder and target decoder, and (c) sequence-to-sequence network with speaker encoder and emotion encoder.

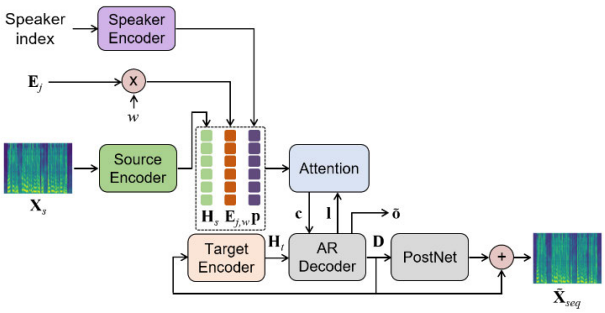

FIGURE 3. Inference procedure of the proposed emotional voice conversion architecture.

acoustic features, and g is set 0.2. The total loss function of the proposed model, L, can now be formulated as follows,

$$L = L_{seq} + L_{src} + L_{trg} + \lambda_{att} L_{att}, \tag{17}$$

where λ_{att} is a regularization parameter for guided attention.

C. EMOTIONAL VOICE CONVERSION WITH STRENGTH CONTROL

To convert the neutral speech into the desired emotional speech after training, the target emotion embedding vector is unseen so that an appropriate emotion embedding vector should be provided for the inference phase. We generate representative emotion vectors for each emotion by calculating the mean vector of the emotion embeddings from the training data as follows,

$$\mathbf{E}_j = \frac{1}{M_j} \sum \mathbf{e}_j, \tag{18}$$

where $\mathbf{E}_j$, M_j and $\mathbf{e}_j$ denote the representative emotion vector, the number of training data, and the emotion embedding vector for emotion j.

To control the emotion strength, we multiply the emotion embedding vector by the strength value. The representative emotion-weighted vector then is generated by calculating the mean vector of the weighted emotion embeddings,

$$\mathbf{E}_{j,w} = \frac{1}{M_j} \sum \mathbf{e}_j \times w = \mathbf{E}_j \times w, \tag{19}$$

where $\mathbf{E}_{j,w}$ signifies the representative emotion-weighted vector for the strength value w for emotion j.

In the inference phase, the source and target decoders are not utilized. With the representative emotion-weighted vector, the inference procedure of the proposed method is illustrated in Fig. 3. As the representative emotion-weighted vector for target emotion and the speaker embedding vector are used to condition the source content embedding matrix, the proposed method can produce the converted emotional acoustic features, and (10) can be modified as follows:

$$\mathbf{c}^{(n)}, \mathbf{a}^{(n)} = \text{Attention}([\mathbf{E}_{j,w}; \mathbf{p}; \mathbf{H}_s], \mathbf{l}^{(n-1)}, \mathbf{a}^{(n-1)}) \tag{20}$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

In our experiments, we used an emotional Korean speech corpus[1] built to realize an interactive TTS system that can express various emotions and personalities. For EVC,

[1] https://github.com/emotiontts/emotiontts_open_db

TABLE 1. Average duration per sentence.

Emotion	Average duration [sec]
Neutral	5.21
Happy	5.88
Sad	6.43
Angry	5.91

TABLE 2. Detailed model configuration.

Source Encoder	FC(512)-ReLU → Conv1D(512,5,1)-BN-ReLU-Dropout(0.5) ×2 → Bi-directional LSTM(512)
Target Encoder	FC(256)-ReLU ×2
Speaker Encoder	Embed(256)
Emotion Encoder	Conv2D(j,3,2)-BN-ReLU for j=32,32,64,64,128,128 → GRU(128) → multi-head attention [57] with 10 GSTs
Attention	Location-sensitive attention [54]
AR Decoder	LSTM(1024)-Zoneout(0.1) × 2 → $\mathbf{l}^{(n)}$ $[\mathbf{l}^{(n)};\mathbf{c}^{(n)}]$ → FC(80) → $\mathbf{d}^{(n)}$, $[\mathbf{l}^{(n)};\mathbf{c}^{(n)}]$ → FC(1) → $\tilde{o}^{(n)}$
PostNet	Conv1D(512,5,1)-BN-Tanh-Dropout(0.5) × 4 → Conv1D(80,5,1)-BN-Tanh-Dropout(0.5)
Source Decoder	Conv1D(512,5,1)-BN-Tanh-Dropout(0.5) × 3 → Conv1D(80,5,1)-BN-Tanh-Dropout(0.5)
Target Decoder	Conv1D(512,5,1)-BN-Tanh-Dropout(0.5) × 3 → Conv1D(80,5,1)-BN-Tanh-Dropout(0.5)

FC (h) represents a fully connected layer with h units. BN is batch normalization. Embed(h) represents a lookup table with an embedding dimension of h. Conv1D(f, k, s) represents 1-D convolution each containing f filters with kernel size k and stride size s. Conv2D(f, k, s) represents 2-D convolution each containing f filters with a k×k kernel and a s×s stride. LSTM(c) represents LSTM with c cells. GRU(c) represents GRU with c cells. Dropout(p) represents a dropout with a probability of p. Zoneout(p) represents a zoneout with a probability of p. ×N represents repeating a block N times.

we adopted a ''plain-to-emotional dataset'' in the corpus. This dataset consists of 100 sentences in four speaking styles, i.e., neutral, happy, sad, and angry, uttered by five voice actresses and five voice actors. Thus, the overall number of speech samples became 4000. We set the plain, i.e., neutral speech as the source and the emotional (happy, sad, angry) speech as the target. Three types of pairs are constructed, consisting of neutral-to-happy, neutral-to-sad, and neutral-to-angry for parallel EVC. We randomly divided 3000 pairs into 2700 pairs of the training set and 300 pairs of the evaluation set. There are no overlapping sentences between the two sets. Speech signals were sampled at 22.05 kHz with a 16-bit resolution. Table 1 shows the average duration per sentence for each emotion category. On average, the duration for emotional speech tends to be greater than that for neutral speech, especially for sad emotion. As a whole, the total duration of the dataset is about 7.36 hours. As noted in Section III-A, for each utterance, we extracted the 80th log mel-spectrogram ranging from 80 to 7600 Hz. The hop length was set to 256 points with a 1024-point Hann window and also a 1024-point FFT window length.

B. NETWORK DETAILS

The details of the proposed model configuration are listed in Table 2. In the emotion encoder, GSTs are a bank of randomly initialized 256-dimensional embeddings with a tanh activation applied. The last gated recurrent units (GRU) state and GSTs are the query and the key vectors to four-head attention, which generates a 256-dimensional emotion embedding vector. In location-sensitive attention, location features are extracted by convolving the attention probabilities at the previous step with trainable filters. The 128-dimensional attention probability vector is evaluated by utilizing the hidden representations of the AR decoder at the previous time-step, the source content embedding matrix, and the location features. The location features are computed using 1-D convolution with 32 filters of 31 kernels.

For the training of our proposed model, the value of λ_{att} in the loss function is heuristically set to 10000. All networks were trained simultaneously with a batch size of 12 on an NVIDIA 1080ti GPU. We used the Adam optimizer [58] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-6}$, and the learning rate of 10^{-3}. It took 50k steps for training until convergence.

We generate waveforms using Parallel WaveGAN [35], which is a GAN-based knowledge-distillation-free vocoder using non-autoregressive WaveNet [31]. We followed the open-source implementation[2] and trained Parallel WaveGAN speaker-independently using our training data.

C. BASELINES

For performance comparison with the proposed method, four methods were utilized: Sprocket, VAWGAN, sequence-to-sequence network (S2S), and sequence-to-sequence network with speaker and emotion identities (S2S-SE).

1) SPROCKET

We used the open-source VC software Sprocket,[3] which was one of the baseline systems used in Voice Conversion Challenge (VCC) 2018 [59]. This software consists of a trajectory-based conversion method using a GMM. We used the same architecture and hyperparameters, except for the F0 range. As Sprocket is constructed independently for every pair, we trained three systems for three types of emotion pairs. For waveform recovery, the WORLD vocoder [60] was adopted.

2) VAWGAN

To compare with state-of-the-art EVC, we used the open-source implementation of VAWGAN based speaker-independent framework[4] [29] which consists of two encoder-decoder structures to separately learn the spectrum and CWT-based F0 mappings. To represent different emotions, one-hot vector as emotion identity was provided to the

[2] https://github.com/kan-bayashi/ParallelWaveGAN
[3] https://github.com/k2kobayashi/sprocket
[4] https://github.com/KunZhou9646/Speaker-independent-emotional-voice-conversion-based-on-conditional-VAW-GAN-and-CWT

TABLE 3. Objective evaluation results with DTW.

Emotion conversion	Methods	MCD [dB]	GPE [%]	VDE [%]	FFE [%]
N-H	Sprocket	6.031	10.292	1.851	7.712
	VAWGAN	6.983	20.993	4.350	14.591
	S2S	**5.664**	5.681	1.831	5.199
	S2S-SE	5.860	5.768	1.745	5.169
	Proposed	5.701	**5.565**	**1.680**	**4.983**
N-S	Sprocket	6.288	10.294	1.657	7.437
	VAWGAN	7.175	10.953	3.246	8.231
	S2S	6.018	3.516	1.558	3.450
	S2S-SE	6.084	3.662	1.538	3.533
	Proposed	**5.807**	**3.091**	**1.526**	**3.100**
N-A	Sprocket	6.537	16.449	1.838	10.828
	VAWGAN	7.647	19.903	3.412	12.508
	S2S	7.503	6.950	2.874	6.625
	S2S-SE	6.183	**4.053**	1.599	**3.811**
	Proposed	**5.827**	4.624	**1.526**	4.075
All pairs	Sprocket	6.285	12.306	1.780	8.659
	VAWGAN	7.267	17.162	3.650	11.649
	S2S	6.390	5.362	2.085	5.065
	S2S-SE	6.042	4.500	1.624	4.147
	Proposed	**5.778**	**4.422**	**1.541**	**4.022**

generator. We used the same architecture and hyperparameters, except for the F0 range. As in Sprocket, the WORLD vocoder [60] was used for speech analysis and synthesis.

3) S2S

We implemented a sequence-to-sequence network with attention as a baseline method by referring to an earlier method in [41]. Different from our proposed architecture, the architecture in [41] does not have the emotion encoder and the speaker encoder. As in Sprocket, S2S was also trained independently for every emotion pair.

4) S2S-SE

To demonstrate the effectiveness of our emotion modelling, we implemented S2S-SE to replace the emotion encoder in our proposed method with embedding lookup, as they did in [42].

D. OBJECTIVE EVALUATION

We evaluated the converted speech with the target speech generated by the vocoder corresponding to each system. The metrics are mel-cepstral distortion (MCD) [61], voicing decision error (VDE) [62], gross pitch error (GPE) [62], and F0 frame error (FFE) [63].

MCD, an indicator used for spectral conversion, is based on mel-cepstral coefficients (MCEPs) defined as follows.

$$\mathrm{MCD[dB]} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\log 10}\sqrt{2\sum_{d=1}^{14}\tilde{m}_{d,t} - m_{d,t}} \tag{21}$$

Here, $\tilde{m}_{d,t}$ and $m_{d,t}$ respectively denote the dth dimension of the converted MCEPs and that of the target MCEPs at frame t while T is the total number of frames. We used the average of the MCD values for all target samples. The WORLD vocoder [60] was used for MCEP estimation.

Although VDE, GPE, and FFE originally evaluated the performance of F0 estimation methods, recent studies [44] [64] [65] utilized them as pitch and voicing metrics between converted and target speeches. VDE measures the percentage of frames for which an error in the voicing and is defined as,

$$\mathrm{VDE[\%]} = \frac{\sum_{t=1}^{T}\mathbf{1}\left[\tilde{v}_t \neq v_t\right]}{T}, \tag{22}$$

where $\tilde{v}_t$ and v_t respectively denote the voicing decisions from the converted speech and the target speech at frame t, and $\mathbf{1}$ is the indicator function. GPE measures the percentage of voiced frames that deviate in pitch by more than 20% compared to the target speech and can be written as,

$$\mathrm{GPE[\%]} = \frac{\sum_{t=1}^{T}\mathbf{1}\left[\left|\tilde{f}_t - f_t\right| > 0.2f_t\right]\mathbf{1}\left[\tilde{v}_t\right]\mathbf{1}\left[v_t\right]}{\sum_{t=1}^{T}\mathbf{1}\left[\tilde{v}_t\right]\mathbf{1}\left[v_t\right]}, \tag{23}$$

where $\tilde{f}_t$ and f_t are the F0 values from the converted speech and the target speech, respectively. FFE takes both GPE and VDE into consideration and is defined as follows.

$$FFE[\%] = \frac{\sum_{t=1}^{T}\left[\left|\tilde{f}_t - f_t\right| > 0.2f_t\right]\mathbf{1}\left[\tilde{v}_t\right]\mathbf{1}\left[v_t\right] + \mathbf{1}\left[\tilde{v}_t\right]\mathbf{1}\left[v_t\right]}{T} \tag{24}$$

YIN's algorithm [66] was used for F0 and the voiced decision.

Tables 3 and 4 show the MCD, GPE, VDE, and FFE results obtained with the proposed and the baseline methods for each emotion conversion. In Table 3, we applied DTW to align the converted speech and the target speech so that differences in the timing were not penalized. As shown in the table, Sprocket performed better than VAWGAN for all emotion pairs. We also observed that S2S, S2S-SE, and the proposed method outperformed Sprocket for three emotion pairs, excepting MCD and VDE, for the neutral-to-angry pair of S2S. Specifically, the GPE result showed that S2S, S2S-SE, and the proposed methods converted the pitch of the target emotion much better than Sprocket. Although there are a few exceptions, the proposed method performed better than S2S and S2S-SE for most cases. In terms of the average MCD, GPE, VDE, and FFE values, the proposed method noticeably outperformed the other methods in all cases.

To consider objective measures as well as synchronization between the converted speech and the target speech on time, the results without DTW are summarized in Table 4. Instead of time-warping, we aligned the shorter signal to the length of the longer signal by padding zeros at the beginning and the end [44]. Sprocket and VAWGAN are designed

TABLE 4. Objective evaluation results without DTW.

Emotion conversion	Methods	MCD [dB]	GPE [%]	VDE [%]	FFE [%]
N-H	Sprocket	15.870	25.847	27.684	38.240
	VAWGAN	15.436	43.277	34.289	48.562
	S2S	**14.650**	21.982	**26.039**	**35.482**
	S2S-SE	14.912	**21.720**	27.012	36.306
	Proposed	14.818	23.276	27.145	37.112
N-S	Sprocket	26.788	33.072	37.539	48.509
	VAWGAN	26.497	33.657	40.604	49.635
	S2S	17.537	15.645	31.454	36.866
	S2S-SE	16.951	15.510	**30.142**	35.697
	Proposed	**16.886**	**14.516**	30.300	**35.415**
N-A	Sprocket	21.389	39.704	33.552	47.378
	VAWGAN	21.087	42.165	37.791	49.691
	S2S	19.445	21.971	30.567	38.478
	S2S-SE	17.009	**17.461**	28.949	35.311
	Proposed	**16.300**	18.395	**27.480**	**34.395**
All pairs	Sprocket	21.486	32.549	33.047	44.826
	VAWGAN	21.145	39.857	37.640	49.307
	S2S	17.240	19.951	29.424	36.954
	S2S-SE	16.313	**18.357**	28.742	35.767
	Proposed	**16.055**	18.896	**28.360**	**35.625**

TABLE 5. MOS results (95% confidence intervals).

Methods	N-H	N-S	N-A	All pairs
Sprocket	3.62±0.098	3.68±0.098	3.67±0.100	3.66±0.057
VAW GAN	2.06±0.027	2.08±0.050	2.06±0.027	2.07±0.021
S2S	4.16±0.083	3.47±0.125	2.35±0.098	3.16±0.087
S2S-SE	4.13±0.087	3.17±0.134	2.77±0.138	3.36±0.080
Proposed	**4.35**±0.077	**3.98**±0.100	**4.08**±0.105	**4.14**±0.056
Ground truth	4.51±0.059	4.53±0.069	4.54±0.061	4.53±0.036

to maintain the speaking rate, but S2S, S2S-SE, and the proposed methods embed the duration mapping so that the neutral speech can be time-synchronized with the emotional speech. As indicated in Table 1, the length of the sad speech is noticeably longer than that of the neutral speech compared to those of happy and angry speeches. Consequently, three methods with duration mapping achieved more notable performance than Sprocket and VAWGAN for neutral-to-sad conversion. Except for VAWGAN, the performances of all methods were similar for neutral-to-happy conversion. For angry conversion, one-to-many mappings such as S2S-SE and the proposed methods outperformed one-to-one mappings such as Sprocket, VAWGAN, and S2S, showing that angry emotion is better modelled when combined with other emotions. Among three methods using duration mapping, the proposed method was slightly superior to the other networks for neutral-to-sad and neutral-to-angry conversions. Considering all pairs together, the proposed method still performed better than the other methods.

E. SUBJECTIVE EVALUATION

For subjective performance evaluation, speech naturalness, emotion similarity, speaker consistency, and emotion classification tests were executed. In addition, to confirm the controllable emotion expressiveness of the proposed method, we also conducted an emotion strength recognition test.

1) NATURALNESS

We conducted a mean opinion score (MOS) test for naturalness evaluation. The neutral speeches with the same sentence were randomly selected for all speakers. For each method, the converted speeches from three neutral speeches uttered by each of ten speakers into three different target emotions were used for the evaluation. We also included ground truth speech corresponding to the converted speech. Thirty Korean subjects with considerable experience in speech quality evaluation listened to the speeches through high-quality headphones and were asked to evaluate the naturalness as follows: 5: Excellent, 4: Good, 3: Fair, 2: Poor, or 1: Bad.

The MOS results with 95% confidence intervals are summarized in Table 5. As shown, S2S, S2S-SE, and the proposed method achieved better performance than Sprocket and VAWGAN in neutral-to-happy conversion. For neutral-to-sad and neutral-to-angry conversion, Sprocket outperformed S2S and S2S-SE. This implied that acoustic features with unstable alignment were generated in S2S and S2S-SE. Nonetheless, the proposed method showed relatively stable alignment performance, and it was significantly superior to all other methods for three emotion pairs. Specifically, the proposed method almost approached the upper limit obtained with the ground truth for neutral-to-happy conversion.

2) EMOTION SIMILARITY AND SPEAKER CONSISTENCY

For successful EVC, the converted speech needs to be similar to the target emotion while preserving the speaker identity. Accordingly, we conducted not only an emotion similarity test but also a speaker consistency test. We adopted the same/different paradigm from VCC 2016 [67] to compare the converted speech and the target speech. The target speech was generated with a corresponding vocoder for each system. The same thirty Korean subjects listened to two speeches and were asked to evaluate the degrees of emotion similarity and speaker consistency by selecting ''Same, absolutely sure,'' ''Same, not sure,'' ''Different, not sure,'' or ''Different, absolutely sure.'' As in the previous experiment, we used the same converted speeches from three neutral sentences uttered by each of ten speakers into three different target emotions.

The emotion similarity evaluation results are shown in Fig. 4. In the three emotion pairs, S2S, S2S-SE, and the proposed method were superior to Sprocket and VAWGAN, especially for sad and angry conversions with a large margin. Among the methods using duration mapping, S2S-SE and the proposed method were comparatively better than S2S. In more detail, S2S-SE and the proposed method showed similar performance for happy and angry conversions, but

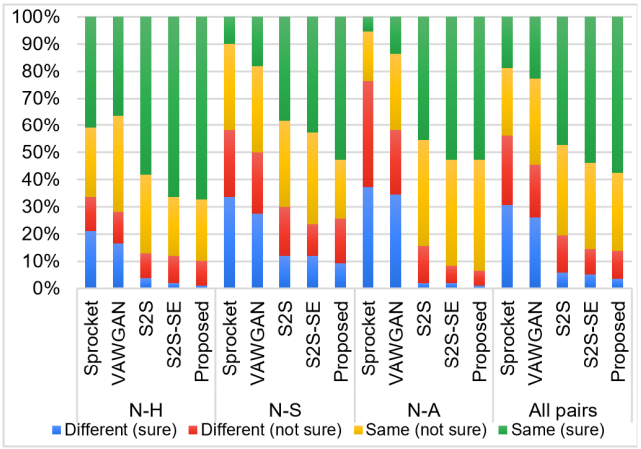

FIGURE 4. Emotion similarity evaluation results.

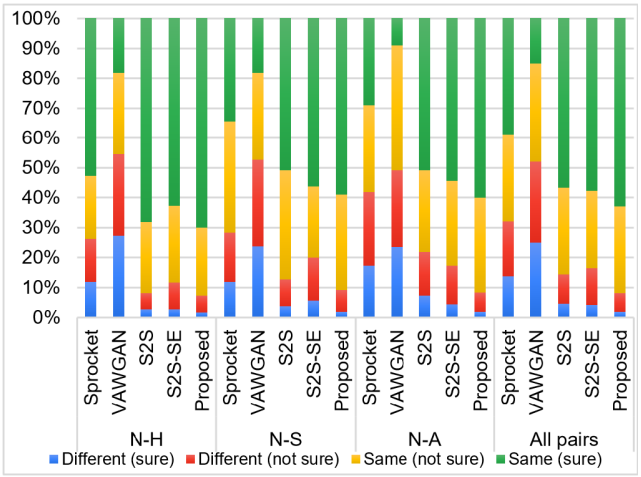

FIGURE 5. Speaker consistency evaluation results.

the proposed method slightly outperformed S2S-SE for sad conversion.

Fig. 5 shows the speaker consistency evaluation results. As indicated in the figure, the methods using duration mapping outstripped Sprocket by a smaller gap compared to that for emotion similarity, as shown in Fig. 4. Although the gap among the methods using duration mapping was not large, the proposed method consistently maintained the speaker information better than all other methods for all emotion conversion. Specifically, VAWGAN outperformed Sprocket in emotion similarity for most cases, but opposite, in speaker consistency.

3) EMOTION CLASSIFICATION

To verify the emotional expressiveness of the converted speech, we conducted a subjective emotion classification test. As before, we used the same converted speeches from three neutral sentences uttered by each of ten speakers into three different target emotions and the same thirty Korean subjects selected the emotions they thought the speech expressed. We also verified target ground truth speeches.

TABLE 6. Confusion matrix of emotion classification results.

Methods	Target	Perception			
		Happy	Sad	Angry	Neutral
Sprocket	Happy	**48.5%**	0.1%	1.2%	50.2%
	Sad	2.1%	**30.5%**	3.2%	64.2%
	Angry	7.1%	2.7%	**13.1%**	77.1%
VAW GAN	Happy	**44.3%**	0.8%	7.8%	47.1%
	Sad	0.8%	**36.4%**	6.1%	56.7%
	Angry	25.9%	3.3%	**7.7%**	63.1%
S2S	Happy	**61.2%**	0.1%	0.2%	38.5%
	Sad	0.2%	**88.7%**	1.9%	9.2%
	Angry	0.2%	17.8%	**60.1%**	21.9%
S2S-SE	Happy	**57.3%**	0.1%	0.2%	42.4%
	Sad	0.1%	**85.7%**	0.1%	14.1%
	Angry	1.6%	11.1%	**72.8%**	14.5%
Proposed	Happy	**65.9%**	0.1%	1.5%	32.5%
	Sad	0.1%	**90.1%**	0.1%	9.7%
	Angry	0.1%	3.8%	**73.1%**	23.0%
Ground truth	Happy	**62.2%**	0.1%	0.2%	37.5%
	Sad	0.1%	**82.3%**	1.1%	16.5%
	Angry	0.1%	4.0%	**85.2%**	10.7%

Table 6 summarizes the results of the emotion classification test. In all emotion pairs, with Sprocket and VAWGAN, more than half of the converted speeches were selected as non-target emotions, suggesting the limited ability of emotion conversion, especially for angry conversion. The methods using duration mapping were superior to Sprocket and VAWGAN for all conversions and were similar to ground truth, especially for happy and sad conversions. The proposed method not only outperformed the baseline methods consistently for all emotions but also showed better results than the ground truth for happy and sad conversions.

4) EMOTION STRENGTH

To evaluate the emotion strength controllability, we gradually changed the strength value for the representative emotion-weighted vector. We chose the converted speeches from three neutral sentences uttered by each of ten speakers into three different target emotions with three different strength values of 0.5 for weak, 1.0 for moderate, and 2.0 for strong emotion. To evaluate the expressiveness of weak emotion, neutral ground truth speeches were also utilized in the test. Two speech samples, A and B, were paired according to the neighboring weight ratios. Namely, neutral-weak emotion pairs, weak-moderate emotion pairs, and moderate-strong emotion pairs are constructed. The same thirty Korean subjects listened to the speech samples of A and B through headphones and were asked to choose 'A', 'B', or 'Same', which expresses stronger emotion.

The emotion strength recognition results are presented in Table 7. For all emotion pairs, stronger emotion speeches were correctly distinguished, especially for the moderate-strong emotion pair of the neutral-to-sad conversion. As shown in the table, weaker emotion speeches were

TABLE 7. Recognition accuracy results for the proposed method.

Emotion conversion	neutral-weak (same)	weak-moderate (same)	moderate-strong (same)
N-H	6.06%-**77.58%** (16.36%)	0.31%-**81.51%** (18.18%)	3.94%-**59.39%** (36.67%)
N-S	2.73%-**72.12%** (25.15%)	2.73%-**69.09%** (28.18%)	0.31%-**94.54%** (5.15%)
N-A	6.97%-**53.33%** (39.70%)	4.24%-**63.94%** (31.82%)	5.14%-**56.67%** (38.19%)

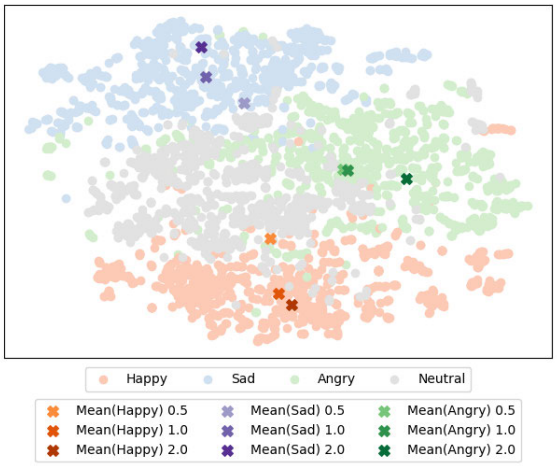

FIGURE 6. Visualization of emotion embedding vectors for the training set and the representative emotion-weighted vectors.

rarely chosen as can be easily expected. These results demonstrate that the representative emotion-weighted vectors can express the emotional speech with various levels fairly well by controlling the strength value.

F. VISUALIZATION

1) ANAYLSYS OF EMOTION EMBEDDING VECTOR

To demonstrate that our representative emotion vector can represent the emotion as predicted, we used t-SNE [49] to visualize the high-dimensional emotion space.

Together with the representative emotion-weighted vectors, the emotion embedding vectors produced by target acoustic features in the training set through the emotion encoder were projected into a two-dimensional space by t-SNE, as shown in Fig. 6. For the representative emotion-weighted vector, we used the same previous strength values of 0.5, 1.0, and 2.0. We found that emotion embedding vectors were similar between identical emotions and were distinct with different emotions. The representative emotion vectors obtained by normalizing the emotion embedding vectors for each emotion were also well shown as the center of the emotion embedding vectors corresponding to the emotion. For the representative emotion-weighted vectors, the smaller the strength value, the closer to the neutral region, and the greater it becomes, the further away from the neutral region.

To show how similar our representative emotion vectors were to the emotion embedding vectors in the evaluation set, we visualized them, as shown in Fig. 7. The representative

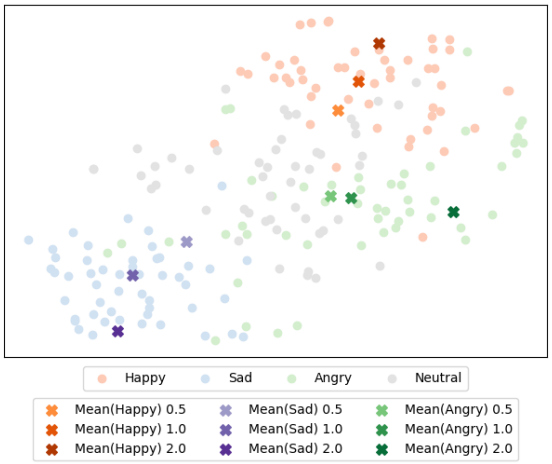

FIGURE 7. Visualization of emotion embedding vectors for the evaluation set and the representative emotion-weighted vectors.

emotion vectors generated with the training data were also located within the clusters of unseen emotion embedding vectors of the same emotion, which can be considered to indicate the robustness of our emotion encoder.

2) ANAYLSYS OF MEL-SPECTROGRAM AND F0

To compare the results by proposed and baseline methods, we visualized source, target, and converted mel-spectrograms and F0 contours in Fig. 8. As discussed in Section IV-D, for all emotion conversions, VAWGAN and Sprocket generated the same lengths and rhythms of emotional speech as those of source speech. On the other hand, the lengths and rhythms of the emotional speech converted by S2S, S2S-SE, and the proposed methods differed from the source speech and varied depending on the target emotion. Specifically, in sad conversion, the proposed method generated F0 contours remarkably similar to the target speech compared to the other methods. We obtained the converted mel-spectrograms which were most similar to the target mel-spectrograms by adding the emotion encoder to the sequence-to-sequence network, especially covering up to the high-frequency part.

To demonstrate the change in mel-spectrograms and F0 contours generated according to the representative emotion-weighted vector conditioned on the proposed method, we visualized them in Fig. 9. Note that the converted speeches, which were converted from the one neutral speech through the proposed methods into three target emotions with three strength values, were shown in the figure. We found that the different mel-spectrograms and F0 contours respectively, showing variations in speaking speed, rhythm, and duration. Specifically, in the sad conversion, the higher the strength value, the slower the speaking speed, and the more monotonous the pitch variation. We also observed that as the strength value increased, the average pitch was higher in happy conversion, and the pitch variation was wider in angry conversion. Despite the input of the same neutral speech, different F0 contours and mel-spectrograms could be observed depending on the representative emotion-weighted vector,

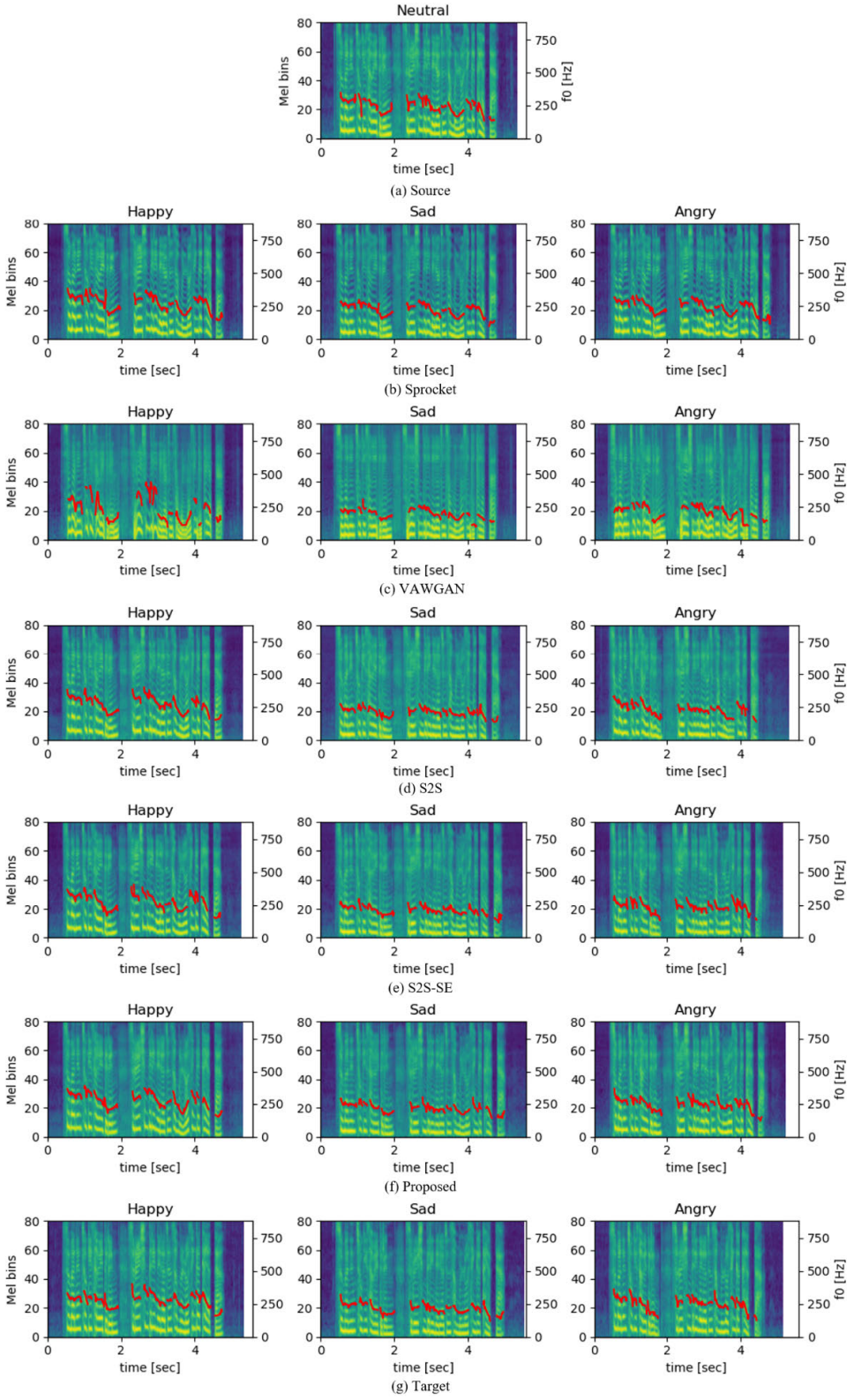

FIGURE 8. The mel-spectrograms and F0 contours of one test utterance converted to three target emotions by different methods. From left to right, we specify the converted emotion: happy, sad, and angry, respectively. The red curves denote F0 contours.

indicating that the proposed method can flexibly handle one-to-many EVC with strength control. Furthermore, even if the fixed representative emotion-weighted vector was conditioned for all time-steps, the pitch was dynamically applied

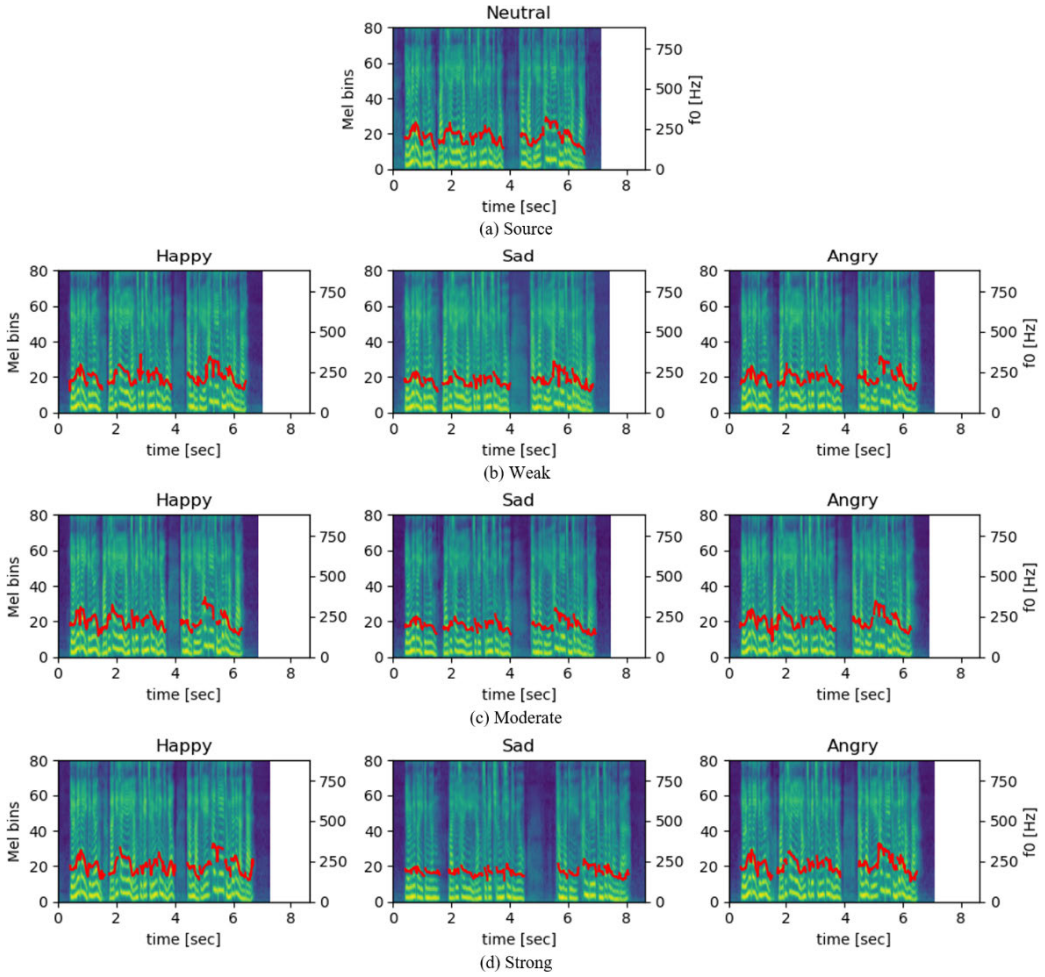

FIGURE 9. The mel-spectrograms and F0 contours of one test utterance converted to three target emotions with three strength values. From left to right, we specify the strength value of representative emotion-weighted vectors: 0.5 for weak, 1.0 for moderate, and 2.0 for strong, respectively. The red curves denote F0 contours.

for each time step. To help to understand our work better, audio samples used in our evaluations are provided.[5]

V. CONCLUSION

This paper proposed a sequence-to-sequence-based EVC method with emotion strength control. To compute the emotion characteristics from raw acoustic features, we applied an unsupervised learning model to a sequence-to-sequence network. To utilize the various emotional characteristics from multiple speakers, a speaker index was applied using an embedding table. By investigating the relationship between the unsupervised emotion embedding vectors and each emotion further, we produced the representative emotion vectors by calculating the mean vector of each emotion cluster.

Thus, the emotion strength was controlled easily by scaling the representative emotion vector.

Major contributions of this work can be summarized as: (1) The proposed method based on the sequence-to-sequence network produced speech with different rhythms and adjusted utterance duration depending on the target emotion. (2) It also could flexibly control the emotion as well as emotion strength. (3) The multi-speaker emotion conversion was rather successfully achieved with our proposed method.

Objective and subjective evaluation results showed that the proposed method generated target emotional speech while preserving speaker identity. Particularly in the subjective emotion classification test, converted speech utterances into happy and sad emotions with our proposed method were more similar to the target emotion than ground truth speech.

For future works, we plan to extend the proposed method further to apply to generating speech for unseen emotions. Moreover, it would be another challenge to adapt our proposed method for noisy environments and accordingly modify it into a rather noise-robust one, even though EVC usually considers noise-free clean speech data. We also hope, combining the proposed method with pre-trained speaker embedding could help generate emotional speech for unseen speakers.

[5] https://chj1330.github.io/ACCESS2021/index.html

REFERENCES

[1] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, Seattle, WA, USA, 1994, pp. 359–370.

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

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

[4] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.

[5] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[6] 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. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4869–4873.

[7] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2032–2045, Nov. 2016.

[8] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3364–3368.

[9] J.-C. Chou, C.-C. Yeh, H.-Y. Lee, and L.-S. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 501–505.

[10] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 9, pp. 1432–1443, Sep. 2019.

[11] 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, Language Process.*, vol. 28, pp. 540–552, 2020.

[12] 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 (SLT)*, Athens, Greece, Dec. 2018, pp. 266–273.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6820–6824.

[14] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[15] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1394–1405, Aug. 2010.

[16] 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, Dec. 2012.

[17] S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Emotional voice conversion using a hybrid framework with speaker-adaptive DNN and particle-swarm-optimized neural network," *IEEE Access*, vol. 8, pp. 74627–74647, Apr. 2020.

[18] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3399–3403.

[19] 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, Language Process.*, vol. 27, no. 10, pp. 1535–1548, Oct. 2019.

[20] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2453–2457.

[21] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2858–2862.

[22] M. Elgaar, J. Park, and S. W. Lee, "Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7769–7773.

[23] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," 2020, *arXiv:2002.00198*. [Online]. Available: http://arxiv.org/abs/2002.00198

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251.

[25] R. Shankar, J. Sager, and A. Venkataraman, "A multi-speaker emotion morphing model using highway networks and maximum likelihood objective," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2848–2852.

[26] 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," 2020, *arXiv:2007.12937*. [Online]. Available: http://arxiv.org/abs/2007.12937

[27] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*. [Online]. Available: http://arxiv.org/abs/1505.00387

[28] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, PR, USA, 2016, pp. 1–16.

[29] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting Anyone's emotion: Towards speaker-independent emotional voice conversion," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 3416–3420.

[30] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 5180–5189.

[31] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. Speech Synth. Workshop (SSW)*, Sunnyvale, CA, USA, 2016, p. 125.

[32] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1138–1142.

[33] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1983–1987.

[34] H. Choi, S. Park, J. Park, and M. Hahn, "Emotional speech synthesis for multi-speaker emotional dataset using WaveNet vocoder," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2019, pp. 1–2.

[35] 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. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6199–6203.

[36] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1268–1272.

[37] 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, Language Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[38] 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. (ICASSP)*, Brighton, U.K., May 2019, pp. 6785–6789.

[39] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 4115–4119.

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

[41] 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. (ICASSP)*, Brighton, U.K., May 2019, pp. 6805–6809.

[42] 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, Language Process.*, vol. 28, pp. 1849–1863, Jun. 2020.

[43] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4784–4788.

[44] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 4693–4702.

[45] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.

[46] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6950–6954.

[47] Y.-Y. Chen, C.-H. Wu, and Y.-F. Huang, "Generation of emotion control vector using MDS-based space transformation for expressive speech synthesis," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3176–3180.

[48] X. Zhu and L. Xue, "Building a controllable expressive speech synthesis system with multiple emotion strengths," *Cogn. Syst. Res.*, vol. 59, pp. 151–159, Jan. 2020.

[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[50] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *Proc. IEEE Automat. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 192–199.

[51] 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. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7254–7258.

[52] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based WaveNet vocoder," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1993–1997.

[53] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4779–4783.

[54] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2015, pp. 577–585.

[55] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[56] A. Gibiansky, S. Ö. Arik, G. F. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 2966–2974.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[58] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[59] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey*, Les Sables-d'Olonne, France, Jun. 2018, pp. 195–202.

[60] 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. E99.D, no. 7, pp. 1877–1884, Jul. 2016.

[61] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, Victoria, BC, Canada, May 1993, pp. 125–128.

[62] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Commun.*, vol. 50, no. 3, pp. 203–214, Mar. 2008.

[63] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3969–3972.

[64] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6189–6193.

[65] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages," *IEEE Access*, vol. 8, pp. 179798–179812, Sep. 2020.

[66] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[67] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1637–1641.

HEEJIN CHOI received the B.S. degree in electrical engineering from Kyungpook National University, Daegu, South Korea, in 2016, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018, where she is currently pursuing the Ph.D. degree. Her main research interests include speech synthesis and voice conversion.

MINSOO HAHN received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1979 and 1981, respectively, and the Ph.D. degree in electrical engineering from the University of Florida, in 1989. From 1982 to 1985, he was with the Korea Research Institute of Standards and Science, Daejeon, South Korea. From 1990 to 1997, he was with the Electronics and Telecommunications Research Institute, Daejeon. Since 1998, he has been a Professor with the Department of the Electrical Engineering, Korea Advanced Institute of Science and Technology. His research interests include speech and audio coding, speech synthesis, voice conversion, noise reduction, and VoIP.