SEEN AND UNSEEN EMOTIONAL STYLE TRANSFER FOR VOICE CONVERSION WITH A NEW EMOTIONAL SPEECH DATASET

Kun Zhou [1]*, Berrak Sisman* [2]*, Rui Liu* [1]*, and Haizhou Li* [1]

[1] Department of Electrical and Computer Engineering, National University of Singapore
[2] Information Systems Technology and Design, Singapore University of Technology and Design

ABSTRACT

Emotional voice conversion aims to transform emotional prosody in speech while preserving the linguistic content and speaker identity. Prior studies show that it is possible to disentangle emotional prosody using an encoder-decoder network conditioned on discrete representation, such as one-hot emotion labels. Such networks learn to remember a fixed set of emotional styles. In this paper, we propose a novel framework based on variational auto-encoding Wasserstein generative adversarial network (VAW-GAN), which makes use of a pre-trained speech emotion recognition (SER) model to transfer emotional style during training and at run-time inference. In this way, the network is able to transfer both seen and unseen emotional style to a new utterance. We show that the proposed framework achieves remarkable performance by consistently outperforming the baseline framework. This paper also marks the release of an emotional speech dataset (ESD) for voice conversion, which has multiple speakers and languages.

Index Terms— emotional voice conversion, speech emotion recognition (SER), emotional speech dataset

1. INTRODUCTION

Speech conveys information with words and also through its prosody. Speech prosody can affect the syntactic and semantic interpretation of an utterance (*linguistic prosody*), and also displays one's emotional state (*emotional prosody*) [1]. Emotional prosody reflects the intent, mood and temperament of the speaker and plays an important role in daily communication [2]. Emotional voice conversion is a voice conversion (VC) technique, which aims to transfer the emotional style of an utterance from one to another. Emotional voice conversion enables various applications such as expressive text-to-speech (TTS) [3] and conversational agents.

Emotional voice conversion and speech voice conversion [4] differs in many ways. Speech voice conversion aims to change the speaker identity, whereas emotional voice conversion focuses on the emotional state transfer. Traditional VC research includes modelling spectral mapping with statistical methods such as Gaussian mixture model (GMM) [5], partial least square regression [6] and sparse representation [7]. Recent deep learning approaches such as deep neural network (DNN) [8], recurrent neural network (RNN) [9] and generative adversarial network (GAN) [10] have advanced the state-of-the-art. We note that these frameworks have motivated the studies in emotional voice conversion.

Early studies on emotional voice conversion handle both spectrum and prosody conversion with GMM [11] and sparse repre-

sentation [12]. Recent deep learning methods, such as deep belief network (DBN) [13], deep bi-directional long-short-term memory (DBLSTM) [14], sequence-to-sequence [15] and rule-based model [16] have shown the effectiveness on emotion conversion. We note that these frameworks are relied on the parallel training data. However, such data is not widely available in real-life applications.

There have been studies on deep learning approaches for emotional voice conversion that do not require parallel training data, such as cycle-consistent adversarial network (CycleGAN)-based [17, 18] and autoencoder-based frameworks [19, 20]. However, they are typically designed for a fixed set of conversion pairs. In this paper, we would like to propose a novel technique, that is referred to as emotional style transfer (EST), that learns to transfer an emotional style to any input utterance. The emotional style is given to the network as a control condition. Therefore, the network supports one-to-many emotion conversion and for unseen emotion at run-time.

Auto-encoder is a suitable computational model that allows for the control of output generation through the latent variables [21]. Recent studies [20] on disentangling and recomposing the emotional elements in the speech with VAW-GAN represent one of the successful attempts in emotional voice conversion. However, emotional prosody is the result of the interplay of multiple signal attributes, hence it is not easy to define emotional prosody by a simple labeling scheme [22]. In TTS studies, there are techniques to learn a latent prosody embedding, i.e. style token, from the training data, in order to predict or transfer the speech prosody [23, 24]. These studies motivate us to investigate the use of deep emotional features that reflect the global speech style and describe the emotional prosody in a continuous space.

In this paper, we propose an emotional style transfer framework based on VAW-GAN. We use deep emotional features as a condition to enable the encoder-decoder training for seen and unseen emotion generation. Furthermore, we release an EVC dataset that consists of multi-speaker and multi-lingual emotional speech. We aim to tackle the lack of open-source emotional speech data in voice conversion research community. This dataset can be easily applied to other speech synthesis tasks, such as cross-lingual voice conversion and emotional TTS.

The main contributions of this paper include: 1) we propose to build a one-to-many emotional style transfer framework that does not require parallel training data; 2) we propose to use SER that is pre-trained with publicly available large speech corpus to describe the emotional style; 3) we propose to disentangle and recompose the emotional elements through deep emotional features; and 4) we release a multi-lingual and multi-speaker emotional speech corpus, denoted as ESD, that can be used for various speech synthesis tasks. To our best knowledge, this paper is the first reported study on emotional style transfer for unseen emotion.

This paper is organized as follows: In Section 2, we motivate

our study and analyze the deep emotional features. In Section 3, we introduce the proposed one-to-many EST framework. In Section 4, we report the experiments. Section 5 concludes the study.

2. ANALYSIS OF DEEP EMOTIONAL FEATURES

The computational analysis of emotion has been the focus of SER [25]. Recent advances of deep learning have led to a shift from traditional human-crafted representations of acoustic features, to the deep features automatically learnt by neural networks [25, 26]. Deep features are data-driven, less dependent on human knowledge and more suitable for emotional style transfer [25].

Emotional prosody is prominently exhibited in emotional speech databases, which can be characterized by discrete categories, such as Ekmans's six basic emotions [27], and continuous representation, such as Russell's circumplex model [28]. Recent studies seek to characterize emotions over a continuum rather than a finite set of discrete categorical labels with the feature representation learnt by deep neural networks in a continuous space [29]. Following the findings in speech analysis, studies are conducted to use emotional prosody modelling to improve the expressiveness of speech synthesis systems [30, 31, 32]. In [30], an emotion recognizer is used to generate the style embedding for speech style transfer. Um et al. [31] apply the style embedding to a Tacotron system with the aim to control the intensity of emotional expressiveness. These successful attempts have revealed the fact that deep emotional features serve as the excellent prosody descriptor, which motivates this study.

We are interested in the use of deep emotional features for voice conversion, to describe emotional prosody in a continuous space. The idea is to use deep emotional features of a reference speech to transfer its emotional style to an output target speech. To motivate the idea, we visualize the deep emotional features of 4 speakers (2 male and 2 female) using the t-SNE algorithm [33] in a two-dimensional plane, as shown in Fig. 1. It is observed that deep emotional features form clear emotion groups in terms of feature distributions. Fig. 1 suggests that we may use deep emotional features as a style embedding to encode an emotion class. Encouraged by this observation, we propose a one-to-many emotional style transfer framework through deep emotional features.

3. ONE-TO-MANY EMOTIONAL STYLE TRANSFER

We propose a novel one-to-many emotional style transfer framework, that is based on VAW-GAN [21] with its decoder conditioning on deep emotional features. The proposed one-to-many EST framework is referred as *DeepEST* in short. We next discuss DeepEST in three stages: 1) emotion descriptor training, 2) encoder-decoder training with VAW-GAN, and 3) run-time conversion interface. In Stage I, we train an auxiliary SER network to serve as the emotion descriptor for input utterances. In Stage II, the proposed encoder-decoder training is implemented to learn the disentanglement and recomposition of the emotional elements. In Stage III, DeepEST takes input utterance and target deep emotional features to generate the utterance with target emotional style.

3.1. Stage I: Emotion Descriptor Training

Emotional prosody is complex with multiple acoustic attributes which makes it difficult to model. There have been studies to label emotion manually into discrete categories, such as one-hot emotion label [34, 20]. As emotional prosody naturally spreads over a continuum that is hard to force-fit into a few categories, we propose
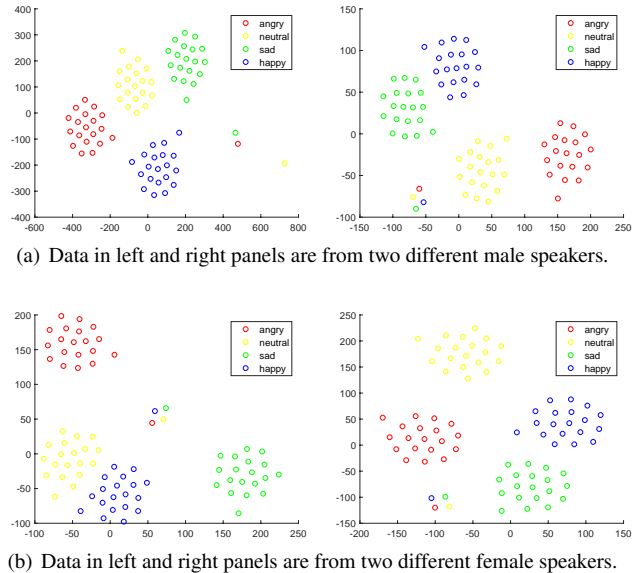

(a) Data in left and right panels are from two different male speakers.

(b) Data in left and right panels are from two different female speakers.

Fig. 1. t-SNE plot of deep emotional features for 20 utterances with the same content but spoken by different speakers.

to use deep emotional features that are learned from large animated and emotive speech data.

We propose to use a SER model as an emotion descriptor D, which extracts deep emotional features Φ from the input utterance X, or $\Phi = D(X)$. The SER architecture is as the same as that in [35], which includes: 1) a three-dimensional (3-D) CNN layer; 2) a BLSTM layer; 3) an attention layer; 4) a fully connected (FC) layer. The 3-D CNN first projects the input Mel-spectrum with its delta and delta-deltas features into a fixed size latent representation, that preserves the effective emotional information while reducing the influence of emotional irrelevant factors. Then the following BLSTM and attention layer summarize the temporal information from the previous layer and produce discriminative utterance-level feature Φ for emotion prediction, as visualized in Fig. 1.

3.2. Stage II: Encoder-Decoder Training with VAW-GAN

Encoder-decoder structure has been used to effectively learn disentangled representations in previous studies [20, 21]. We propose an encoder-decoder training procedure as shown in Fig. 2, where the encoder (E) learns to disentangle the emotional elements from the input features and generate a latent representation z. The resulting representation z is assumed to contain phonetic and speaker information, but emotion-independent. Then the decoder/generator (G) learns to reconstruct the input features with the emotion-independent representation z and other controllable emotion-related attributes.

In practice, we use WORLD vocoder [36] to extract spectral features (SP) and fundamental frequency (F_0) from the waveform. The encoder (E_θ) with parameter set θ is exposed to the input spectral frames x with different emotion types and learns an emotion-independent representation z: $z = E_\theta(x)$. Since the latent representation z extracted from the source spectrum still contains the source F_0 information, and the conversion performance can suffer from this flaw [20]. Therefore, the decoder/generator (G_ψ) with parameter set ψ takes emotion-independent representation z and the emotion-related features: the deep emotional features Φ
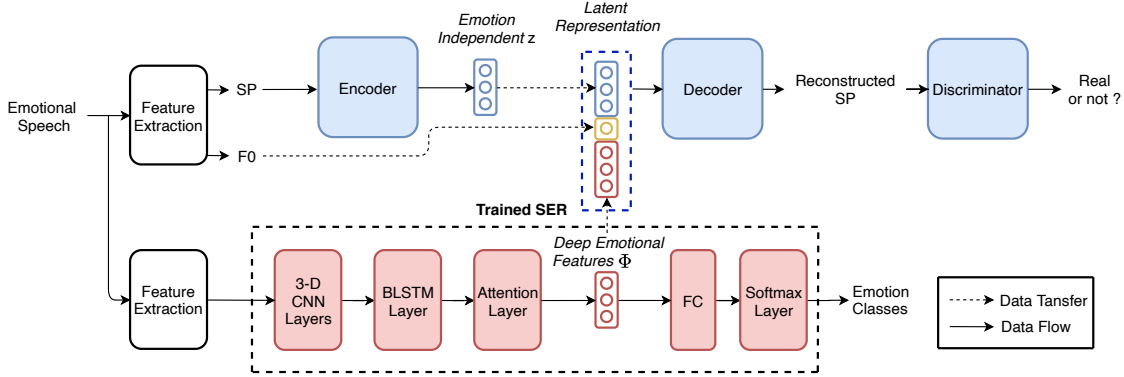
Fig. 2. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training and red boxes represent the networks that are already trained.

Table 1. MCD values of the baseline framework VAW-GAN-EVC and the proposed framework DeepEST in a comparative study.

MCD [dB]	Male			Female		
	Zero Effort	VAW-GAN-EVC	**DeepEST**	Zero Effort	VAW-GAN-EVC	**DeepEST**
neutral-to-happy	7.294	4.825	4.530	6.866	4.081	4.065
neutral-to-sad	6.202	4.164	3.770	7.061	4.724	4.314
neutral-to-angry	6.686	4.437	4.079 *(unseen)*	6.351	3.774	4.041 *(unseen)*

that reflect the global emotional variance of the input utterance X from Stage I and the corresponding F_0 that contains source pitch information to recompose the emotional elements of the spectrum. The reconstructed feature $\overline{x}$ can be formulated as:

$$\overline{x} = G_\psi(z, \Phi, F_0) = G_\psi(E_\theta(x), D(X), F_0) \quad (1)$$

We then train a generative model for spectrum through an adversarial training: The discriminator (Y_μ) with parameter set μ tries to maximize the loss between the real features x and reconstructed features $\overline{x}$, while the generator (G_ψ) tries to minimize it. The parameter sets θ, ψ and μ are optimized through this min-max game, which allows us to generate high-quality speech samples.

3.3. Stage III: Run-time Conversion

During the run-time conversion, we have a source utterance that is expressed in neutral emotion, we would like to convert it to a target emotion following the reference emotion style from the reference utterances. Suppose that we have a set of reference utterances X_t belonging to an emotion category. We first use the pre-trained SER to generate the deep emotional features $\Phi_t = mean(D(X_t))$ for all the reference utterances, i.e. all the utterances of our dataset with the same reference emotion. We then concatenate Φ_t with the converted F_0 ($\hat{F}_0$) and emotion-independent z from the source utterance to compose a latent representation for the target utterance SP. The converted SP can be formulated as:

$$\hat{x} = G_\psi(z, \Phi_t, \hat{F}_0) = G_\psi(E_\theta(x), mean(D(X_t)), \hat{F}_0) \quad (2)$$

Finally, the converted speech with target emotion is synthesised by WORLD vocoder with converted spectral features and converted F_0.

4. EXPERIMENTS

4.1. Emotional Speech Dataset (ESD)

With this paper, we introduce and publicly release a new multi-lingual and multi-speaker emotional speech dataset that can be used

for various speech synthesis and voice conversion tasks[1]. The dataset consists of 350 parallel utterances with an average duration of 2.5 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 summarized as follows: 1) happy, 2) sad, 3) neutral, 4) angry, and 5) surprise. Speech data are sampled at 16 kHz and saved in 16 bits.

To our best knowledge, this is the first parallel voice conversion dataset that provides emotion labels in a multi-lingual and multi-speaker setup. As a future work, we will report an in-depth investigation of this dataset for cross-lingual and mono-lingual emotional voice conversion applications.

4.2. Experimental Setup

We conduct objective and subjective evaluation to assess the performance of our proposed DeepEST model for seen and unseen emotional style transfer. We use four English speakers (2 male and 2 female) from ESD dataset. For each speaker, we conduct *seen emotion* conversion from neutral to happy (N2H: neutral-to-happy) and neutral to sad (N2S: neutral-to-sad). We choose angry as the *unseen emotion*, and conduct experiments from neutral to angry (N2A: neutral-to-angry) to assess the performance of our proposed model for unseen emotion style transfer.

We split the 350 utterances in ESD dataset for a speaker into training set (330 utterances) and test set (20 utterances). During training, we propose to have **one universal model** that takes 330 utterances from neutral, happy and sad emotion states respectively. To obtain the deep emotional features, we train the SER [30, 32] on a subset of IEMOCAP [34] with four emotion types (happy, angry, sad and neutral). At run-time, we evaluate our framework with 20 utterances both from seen (happy, sad) and unseen (angry) emotion states. We obtain the reference emotion style by using all the utterances in ESD dataset, as formulated in Eq. (2). For each

---
[1]**Emotional Speech Dataset (ESD):** `https://github.com/HLTSingapore/Emotional-Speech-Data`

Table 2. MOS results with 95 % confidence interval to assess the speech quality.

MOS	N2H	N2S	N2A
Reference	4.80 ± 0.14	4.69 ± 0.23	4.78 ± 0.14
VAW-GAN-EVC	2.93 ± 0.27	3.10 ± 0.27	2.93 ± 0.27
DeepEST	3.01 ± 0.26	3.16 ± 0.26	3.04 ± 0.29

emotion conversion, we generate the deep emotional features from SER module by calculating the mean of the features that are in the same emotion category as the emotion reference.

As the baseline framework, we adapted a state-of-the-art VAW-GAN-EVC [20] that can perform conversion from one emotional state to another. We note that for each emotion conversion pair, we need to train a new VAW-GAN-EVC model as it is not capable of performing one-to-many conversion. In contrast, DeepEST provides more flexible manipulation and generation of the output emotion, as it can perform an emotional conversion from one emotional state to many and generate unseen emotions, which will be investigated in the following sections.

In DeepEST, 513-dimensional *SP*, F_0 and *APs* are extracted every 5 ms with the FFT length of 1024 using the WORLD vocoder. The frame length is 25 ms with a frame shift of 5 ms. We normalize every input frame of *SP* to unit sum and then re-scale it to logarithm. The encoder is a 5-layer 1D CNN with the kernel size of 7 and a stride of 3 followed by a FC layer. Its output channel is $\{16,32,64,128,256\}$. The latent representation z is 128-dimensional and assumed to have a Gaussian distribution. The deep emotional feature is 256-dimensional, and concatenated with the 128-dimensional latent representation z and 1-dimensional F_0 contour to merge as the input to the decoder. The decoder is 4-layer 1D CNN with kernel size of $\{9,7,7,1025\}$ and strides of $\{3,3,3,1\}$. Its output channel is $\{32,16,8,1\}$. The discriminator is a 3-layer 1D CNN with kernel size of $\{7,7,115\}$ and strides of $\{3,3,3\}$ followed by a FC layer. The networks are trained by using RMSProp with a learning rate of 1e-5. The batch size is set as 256 for 45 epochs.

4.3. Objective Evaluation

We conduct objective evaluation to assess the performance of our proposed model. We calculate Mel-cepstral distortion (MCD) [7, 4] to measure the spectral distortion between the converted and reference Mel-spectrum for two male and two female speakers for three emotion combinations.

As shown in Table 1, the proposed DeepEST outperforms the baseline framework VAW-GAN-EVC for all the seen emotions (N2H and N2S). We note that for the unseen emotion (*angry*), DeepEST still achieves comparable results to that of VAW-GAN-EVC baseline, which is trained with *angry* speech samples. These encouraging observations validate the effectiveness of our proposed DeepEST for seen and unseen emotion transfer. Last but not least, we require three VAW-GAN-EVC models to be trained for all conversion pairs, whereas the proposed DeepEST can perform all the emotion mapping pairs within one model, which we believe is remarkable.

4.4. Subjective Evaluation

We further conduct three listening experiments to assess the proposed DeepEST in terms of speech quality and emotion similarity. 15 subjects participated in all the experiments and each listened to 108 converted utterances in total.
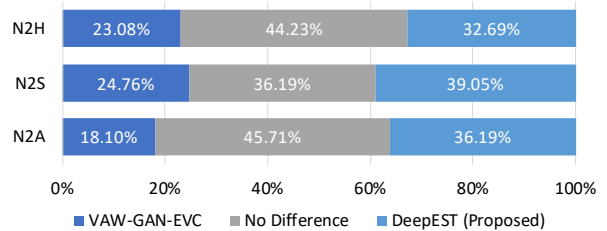

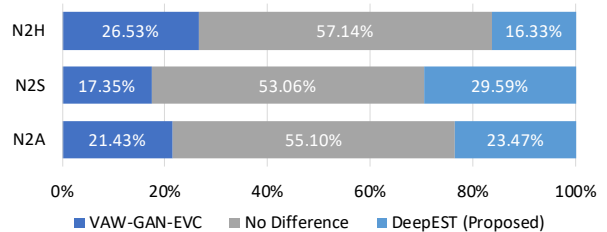**Fig. 3**. AB preference test results for the speech quality.

Fig. 4. XAB preference test results for the emotion similarity.

We first report the mean opinion score (MOS) of the reference speech samples, baseline VAW-GAN-EVC and the proposed Deep-EST. As shown in Table 2, DeepEST achieves better results than the baseline for both seen and unseen emotion combinations, which we believe is remarkable. Secondly, we conduct AB preference test to further evaluate the speech quality, where the subjects are asked to choose the speech samples with higher speech quality. As shown in Fig. 3, we observe that proposed framework DeepEST consistently outperforms the baseline framework VAW-GAN-EVC for all the emotion combinations, which is also consistent with the MOS results. These observations validates the effectiveness of our proposed model in terms of voice quality.

We further conduct XAB emotion similarity test to assess the emotion conversion performance, where subjects are asked to choose the speech samples which sound closer to the reference in terms of the emotional expression. Consistent with previous experiments, the baseline is one-to-one conversion and all emotions are seen emotions. As shown in Fig. 4, DeepEST outperforms the baseline for neutral-to-angry conversion. We observe that baseline achieves better performance for neutral-to-happy conversion due to the poor performance of SER on happy (29.95%) compared with 84.32% on sad, and 70.47% on angry. For unseen emotion (*angry*), DeepEST outperforms the baseline in terms of emotion similarity, which is very encouraging. This result validates the effectiveness of DeepEST for unseen emotion transfer. As a future work, we will improve the SER performance for all emotional states, and report an in-depth investigation.

5. CONCLUSION

In this paper, we propose a one-to-many emotional style transfer framework based on VAW-GAN without the need for parallel data. We propose to leverage deep emotional features from SER to describe emotional prosody in a continuous space. By conditioning the decoder with controllable attributes such as deep emotional features and F0 values, we achieve competitive results for both seen and unseen emotions over the baseline framework, which validates the effectiveness of our proposed framework. Moreover, we also introduce a new emotional speech dataset, ESD, that can be used in speech synthesis and voice conversion.

6. REFERENCES

[1] Julia Hirschberg, "Pragmatics and intonation," *The handbook of pragmatics*, pp. 515–537, 2004.

[2] Magda B Arnold, "Emotion and personality.," 1960.

[3] Rui Liu, Berrak Sisman, Jingdong Li, Feilong Bao, Guanglai Gao, and Haizhou Li, "Teacher-student training for robust tacotron-based tts," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6274–6278.

[4] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *arXiv preprint arXiv:2008.03648*, 2020.

[5] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[7] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.

[8] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[9] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in *Fifteenth annual conference of the international speech communication association*, 2014.

[10] Berrak Sisman, Mingyang Zhang, Minghui Dong, and Haizhou Li, "On the study of generative adversarial networks for cross-lingual voice conversion," *IEEE ASRU*, 2019.

[11] Jianhua Tao, Yongguo Kang, and Aijun Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[12] Ryo Aihara, Reina Ueda, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *APSIPA ASC*. IEEE, 2014.

[13] Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional voice conversion using deep neural networks with mcc and f0 features," in *2016 IEEE/ACIS 15th ICIS*, 2016.

[14] Huaiping Ming, Dongyan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," *Interspeech 2016*, pp. 2453–2457, 2016.

[15] Carl Robinson, Nicolas Obin, and Axel Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6830–6834.

[16] Yawen Xue, Yasuhiro Hamada, and Masato Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, 2018.

[17] Kun Zhou, Berrak Sisman, and Haizhou Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 230–237.

[18] Ravi Shankar, Jacob Sager, and Archana Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," *arXiv preprint arXiv:2007.12932*, 2020.

[19] Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye, "Nonparallel emotional speech conversion," *Proc. Interspeech 2019*, pp. 2858–2862, 2019.

[20] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," *arXiv preprint arXiv:2005.07025*, 2020.

[21] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[22] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, and Junichi Yamagishi, "Adapting and controlling dnn-based speech synthesis using input codes," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4905–4909.

[23] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[24] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*, 2018, pp. 4693–4702.

[25] Dagmar M Schuller and Björn W Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Review*, p. 1754073919898526, 2020.

[26] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *arXiv preprint arXiv:2001.00378*, 2020.

[27] Paul Ekman, "An argument for basic emotions," *Cognition & emotion*, 1992.

[28] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[29] Eesung Kim and Jong Won Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6720–6724.

[30] Yang Gao, Weiyi Zheng, Zhaojun Yang, Thilo Kohler, Christian Fuegen, and Qing He, "Interactive text-to-speech via semi-supervised style transfer learning," *arXiv preprint arXiv:2002.06758*, 2020.

[31] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[32] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li, "Expressive tts training with frame and style reconstruction loss," *arXiv preprint arXiv:2008.01490*, 2020.

[33] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[34] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[35] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[36] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.