Voice Frequency Synthesis using VAW-GAN based Amplitude Scaling for Emotion Transformation

Hye-Jeong Kwon[1], Min-Jeong Kim[1], Ji-Won Baek[1], Kyungyong Chung[2*]
[1]Department of Computer Science, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: rnjsgpwjd121@kyonggi.ac.kr, minjeog0513@kyonggi.ac.kr, jwbaek@kyonggi.ac.kr]
[2]Division of AI Computer Science and Engineering, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: dragonhci@gmail.com]
*Corresponding author: Kyungyong Chung

Received August 19, 2021; revised October 14, 2021; accepted November 7, 2021; published February 28, 2022

Abstract

Mostly, artificial intelligence does not show any definite change in emotions. For this reason, it is hard to demonstrate empathy in communication with humans. If frequency modification is applied to neutral emotions, or if a different emotional frequency is added to them, it is possible to develop artificial intelligence with emotions. This study proposes the emotion conversion using the Generative Adversarial Network (GAN) based voice frequency synthesis. The proposed method extracts a frequency from speech data of twenty-four actors and actresses. In other words, it extracts voice features of their different emotions, preserves linguistic features, and converts emotions only. After that, it generates a frequency in variational auto-encoding Wasserstein generative adversarial network (VAW-GAN) in order to make prosody and preserve linguistic information. That makes it possible to learn speech features in parallel. Finally, it corrects a frequency by employing Amplitude Scaling. With the use of the spectral conversion of logarithmic scale, it is converted into a frequency in consideration of human hearing features. Accordingly, the proposed technique provides the emotion conversion of speeches in order to express emotions in line with artificially generated voices or speeches.

Keywords: Emotion Transformation, Generative Adversarial Network, Voice Frequency Synthesis, Voice Analysis

A preliminary version of this paper was presented at APIC-IST 2021, and was selected as an outstanding paper. This work was supported by the GRRC program of Gyeonggi province. [GRRC KGU 2020-B03, Industry Statistics and Data Mining Research]

http://doi.org/10.3837/tiis.2022.02.018 **ISSN : 1976-7277**

1. Introduction

Emotional voice conversion is method capable of expressing a different emotion without any loss of linguistic information [1][2]. Conventional voice conversion models failed to consider prosody, and used spectrum mapping for voice conversion. Given that prosody includes not only linguistic information but emotional information of speeches, they help to analyze speeches accurately in voice conversion [3]. When sequential speech data are converted, their linguistic features or voice information features can be lost. For this reason, it is important to learn emotional features without any loss of linguistic information [4][5]. Using deep learning to find data features can show high-performance results and does not require modeling that specifies features. A model generated through deep learning, such as GAN or VAE, is often applied to voice conversion, and helped to improve the quality of voice conversion greatly. Although a conventional CycleGAN shows good performance in voice conversion, it supports one-to-one voice conversion and consequently has speaker-dependence [6]. K. Zhou et al. [7] proposes the emotional voice conversion framework based on trained CycleGAN for non-parallel data. The proposed technique presents the acoustic features of data by using melspectrum, and takes into account prosodic features through continuous wavelet transform of F0. By learning non-parallel data, it overcomes the problem of speaker-dependence. Variational Autoencoding (VAE) Wasserstein based voice conversion supports *n*-to-*m* typed speaker-independency, but shows lower voice quality than VAW-GAN [8]. S. Kim et al. [9] proposes the emotional voice conversion framework based on VAW-GAN for non-parallel data. Even though the proposed method is capable of distinguishing emotions more easily than a VAE based emotional voice conversion model, it has less clearance of the voice made by a generator. Therefore, it is necessary to develop a technique that supports speaker-independent voice conversion and improves voice quality. In addition, it is required to devise a method for reflecting prosody information effectively.

This study proposes the emotion conversion using the VAW-GAN based voice frequency synthesis. It is an emotional voice conversion model in which voice features are extracted in consideration of emotions, and linguistic information is preserved [10]. Vocoder is employed to extract voice features according to emotions. In two encoder-decoder structures where the spectrum and fundamental frequency(f0) information extracted by Vocoder is learned in parallel, emotional voice conversion is performed. There is a difference between the emotional spectrum according to the input data in learning process, and the regenerated emotional spectrum. To correct the difference between these spectrums, amplitude scaling is applied. It adjusts an emotion in the way of making it similar to that of target data, so as to reproduce emotional prosody. The main suggestions made in this study are as follows:

1) It is possible to make emotion conversion in a speaker-independent way through *n*-to-*m* voice conversion, rather than one-to-one voice conversion.

2) By using and learning prosody information through cwt-f0 and spectrum, it is possible to achieve accurate analysis in consideration of emotions.

3) By correcting a difference with the emotional spectrum of target data, it is possible to increase an emotional similarity.

This study is composed of as follows: in chapter 2 is described the trend of GAN based waveform generation and voice frequency analysis technologies; in chapter 3 is described the

proposed voice frequency synthesis using VAW-GAN based amplitude scaling for the emotion transformation; in chapter 4 is described the result from the performance evaluation of the proposed model; in chapter 5 is drawn the conclusion.

2. Related Works

2.1 Waveform Generation using GAN

A GAN is made of a generator and a discriminator. It is a model of generating data in competition between a generator and a discriminator. A generator creates the data similar to actual data, and a discriminator discriminates the data made by a generator. A GAN model is capable of learning by itself in the course of competition even if no answer is given, and it is possible to make an image or voice directly [11][12]. R. Yamanoto et al. [13] proposed Parallel WaveGAN, an effective parallel waveform generation method based on GAN. The proposed model method optimizes the combination of Multi-resolution Short-time Fourier Transform (STFT) and adversarial loss function, and learns non-autoregressive WaveNet model only. It is capable of generating a natural voice waveform with the use of small mediator variables, and of capturing the time-frequency distribution of an actual voice waveform effectively. P. Narvaez et al. [14] proposed the method of synthesizing normal cardiac sounds by using GAN and empirical wave transform. The proposed method employs GAN to generate synthesized cardiac sounds, and reduces the noise of synthesized signals through empirical wave transform. Accordingly, it is capable of capturing the features of natural cardiac sounds accurately. Nevertheless, since it fails to generate various types of samples, it is impossible to obtain abnormal cardiac sounds. Therefore, it is necessary to devise a technique to obtain sounds with various features. **Fig. 1** shows the waveform generation process using GAN.

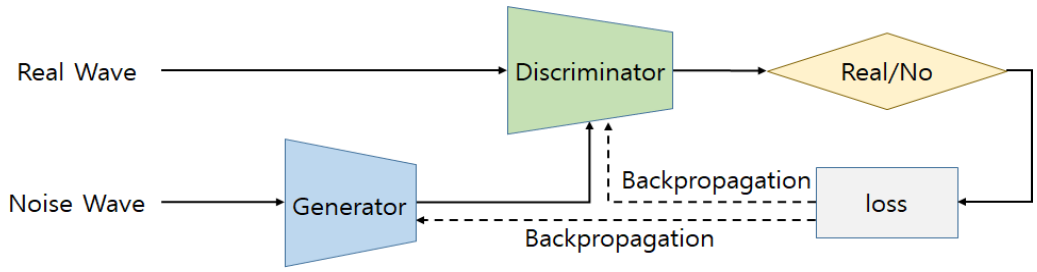

Fig. 1. Waveform generation process using GAN

2.2 Trends of Voice Frequency Analysis Technology

People have different voices. Since a voice is a sound wave, it consists of pitch and range. A vocal range is determined by amplitude of vocal cords. Pitch is determined by frequency. For this reason, it is possible to analyze with the uses of vocal range and pitch [15]. A voice has frequency and amplitude. A frequency is the number of occurrence of repeating waveform per unit of time [16]. An amplitude is the maximum distance or displacement from the center of vibration when repeating vibrations occur. By visualizing the signals extracted from voice in the types of voice waveform, spectrum, and spectrogram in computer, it is possible to analyze emotions in a deep learning model [17]. A waveform is a form of voice vibration waves generated over time. A spectrum is a representation of frequency and amplitude for various kinds of voices. With spectrum, it is possible to analyze a waveform of sound and represent

the components constituting sound [18]. A spectrogram is a visual representation of the sound spectrum in graph. It has the combination of the features that a waveform and a spectrum have, showing the time, frequency, and amplitude in each of X-axis, Y-axis, and Z-axis [19]. Sentiment analysis is used to analyze the tendency, opinion, and attitude of a person in dialogue or text. It is made possible with the uses of language, voice, gesture, vision, hearing, and other factors [20]. At present, for the voice based sentiment analysis, voice signals are extracted in a spectrogram, which are applied to a deep learning model. For example, M. Pasini et al. [21] proposed a voice conversion and audio style transmission method by using a spectrogram. Dependent on non-parallel voice data, the proposed technique is used to convert a source voice to a target voice for an audio signal with a random length. It first calculates a spectrogram with waveform data, and then performs domain conversion with the use of GAN structure. Accordingly, it executes transmission in music style, and converts a voice and modifies audio by changing a genre. **Fig. 2** shows the structure of voice frequency analysis.

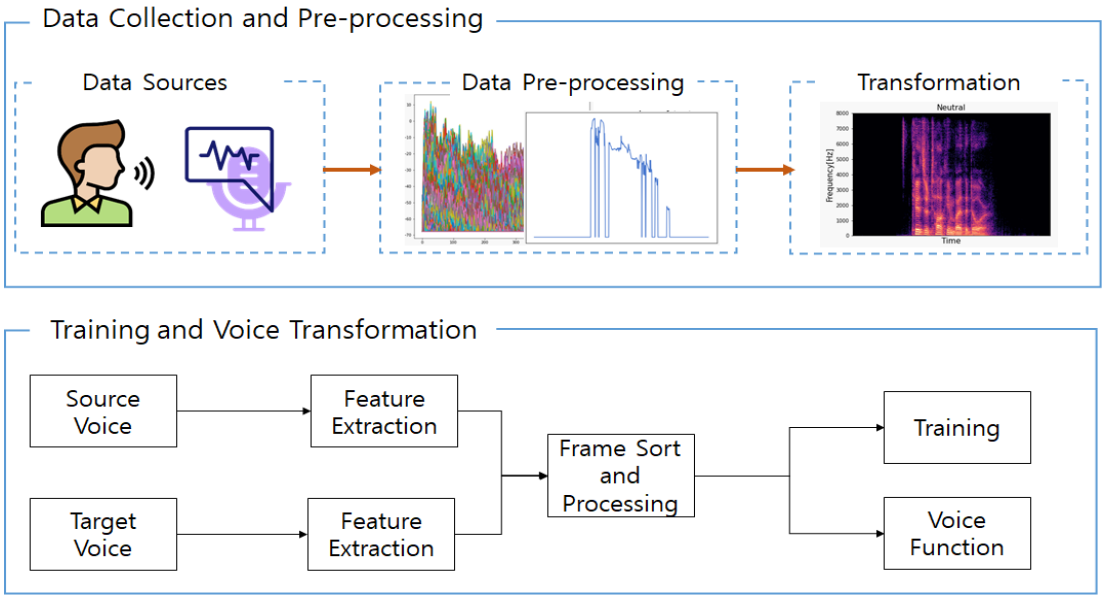

Fig. 2. Structure of voice frequency analysis

3. Voice Frequency Synthesis using VAW-GAN based Amplitude Scaling

3.1 World Vocoder based speech data feature extraction

In this study, the features of voices with different emotions are extracted; linguistic features are preserved; emotions only are converted. The data used in this study are the speech data in the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [22]. The data are the equal sentences spoken by 24 actors and actresses in their neutral accents. In this study, a type of emotion that the source data to convert has is Neutral, and target data have two emotional types: angry and disgust. To extract the features representing emotions of speech data, this study adopts WORLD vocoder [23]. Among vocoder based voice synthesis systems, WORLD performs high-quality voice synthesis and fast calculation. For the emotion conversion of speech data, WORLD vocoder is used to extract two audio features: spectral

(SP) and fundamental frequency(f0). In these features, f0 is defined as the lowest frequency of periodic waveforms. It is the most fundamental prosody factor that appears in speech data. Along with a vocabulary and a word, the prosody with f0 is used as a factor expressing an emotion. Spectral is the smoothed spectrogram extracted, representing harmonic spectral envelope. The spectral envelope extracted by WORLD vocoder makes use of both waveform and f0 information. The spectral envelope makes it possible to present the amplitude of spectrum. In the case of amplitude, since it is a characteristic of the signal related to the intensity of the sound, if the amplitude has the intensity of a flow similar to neutral even though the emotions are different, the difference in the amplitude graph may be slight.

However, to put emotion in speech, intonation, pitch, and style appear in various ways. So if shouting speech is as in Angry data, the amplitude can be significantly different from Neutral. **Fig. 3** shows SP and f0, the features of the speech data with different emotion classes in the same sentence.

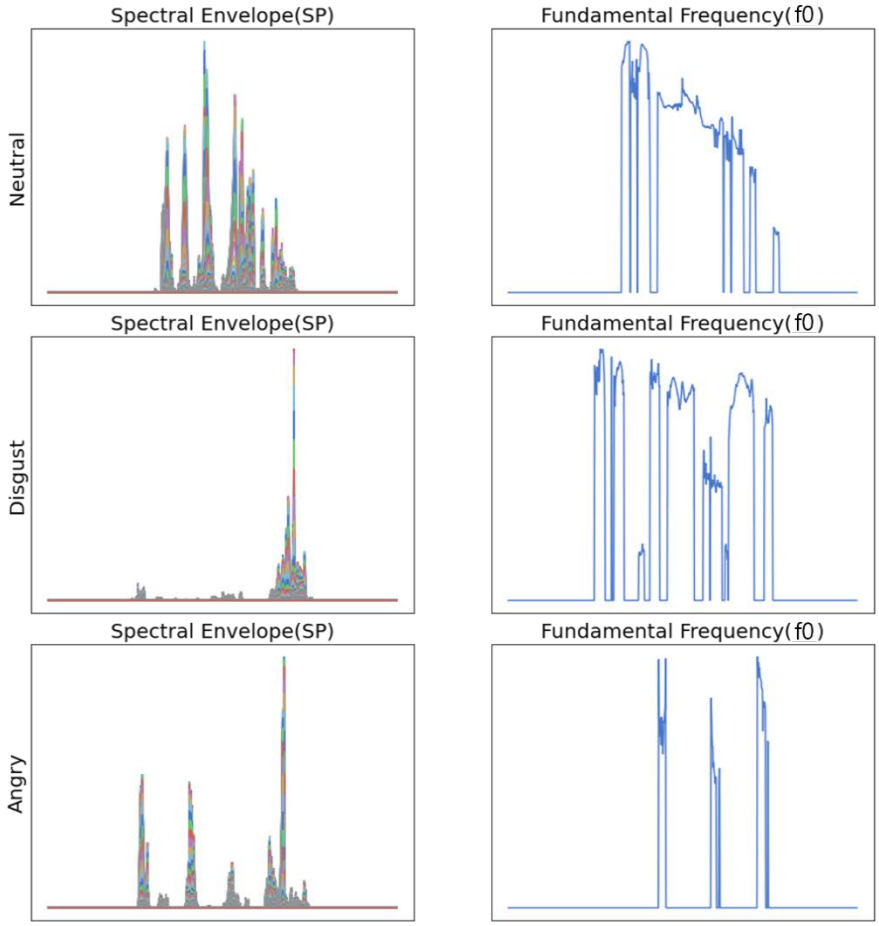

Fig. 3. Features of the speech data with different emotion classes in the same sentence.

As shown in **Fig. 3**, SP and f0 have different speech features depending on emotions. Speech data with emotions have relatively larger amplitude than "Neutral" speech data, and f0 is different. These speech features are preprocessed and are used as learning input, and voice synthesis with converted emotions is made possible.

3.2 VAW-GAN-based frequency generation model for prosody formation and language information preservation

When speech features are learned in parallel, VAW-GAN based Encoder-Generator structure is learned in parallel according to the previous research on more effective voice conversion. The synthesized voice generated with VAW-GAN generates a more structured spectral envelope than in a conventional voice conversion model. It means that this model is capable of generating more realistic fake speech data. **Fig. 4** shows the process of learning speech features in parallel with the use of VAW-GAN.

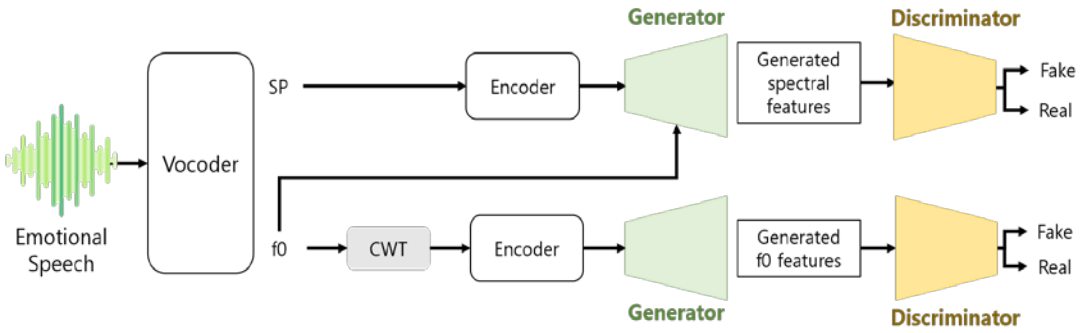

Fig. 4. Process of learning speech features in parallel with the use of VAW-GAN

As shown in **Fig. 4**, SP and f0, the features extracted from source data, are vectorized by different encoders. The adversarial learning based on Generator and Discriminator is applied to data emotions. The encoder learns an independent feature that does not contain emotion and transforms it into a latent vector. Therefore, the generator is given an emotion ID representing the emotion of the corresponding speech and an emotion-independent latent z as input. To add prosody information of sp, f0 is additionally input so that the generator can learn prosody information. The generator is trained to reduce the loss between the original sp and the fake sp, and the discriminator conducts hostile training by maximizing the loss between the original and fake sp. Therefore, model can effectively learn the emotion-independent sp and f0 to form rhymes. To learn prosody information from the speech features of source data, continuous wavelet transformation(CWT) is employed [24]. It converts a frequency with the use of mother wavelet function. **Eq. 1** presents CWT.

$$W_f(\mathrm{a},\mathrm{b}) = \int_{-\infty}^{\infty} \psi_{ab} f(t) dt \tag{1}$$

In **Eq. 1**, a represents scale; b is time; and ψ_{ab} means mother wavelet function. CWT does multi-scale modeling for f0 and divides it into various temporal scales. Therefore, with CWT, it is possible to express short prosody to phrases made in some words, and long-scale prosody in an entire speech. Such a method is applied to prosody conversion of emotion speeches. Therefore, through CWT based f0 conversion, it is possible to learn prosody information efficiently.

3.3 Frequency correction using amplitude scaling

In previous research, for prosody mapping, cwt based f0 conversion only is performed. In this study, two Encoder-Decoder structures that learn SP and cwt-f0 are generated, and emotion conversion is executed. Emotional expressions in speeches are related to energy of frequencies. Therefore, both SP and f0, the speech features obtained by frequency Vocoder, are involved in emotion expressions. Emotion Conversion phase synthesizes the reproduced SP and f0 similar to training phases. However, the emotional SP learned in training phases is different from the emotional SP of source data, a target of conversion. A learned generator learns the SP of target emotion, and the input SP in Conversion Phase is a different emotion to be converted. Such input has different kinds of information on the sound strength or prosody found in different emotions. Therefore, sp, the base input of reproduction, has different amplitude from target data. The fake SP reproduced by base SP preserves the emotional feature of source data. For this reason, even if the generator trained with target emotions is applied, emotional prosody is not reproduced fully.

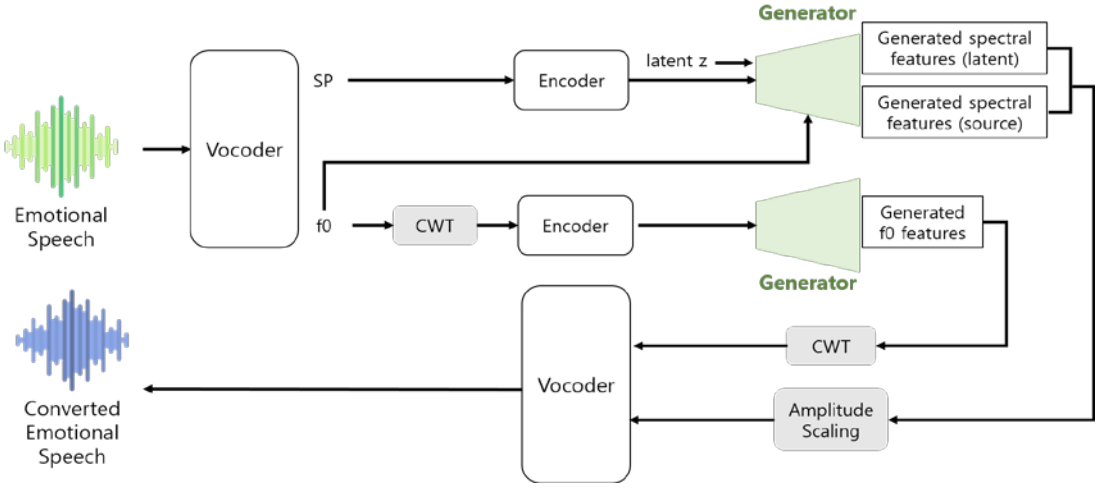

Fig. 5. Process of converting an emotion through amplitude scaling

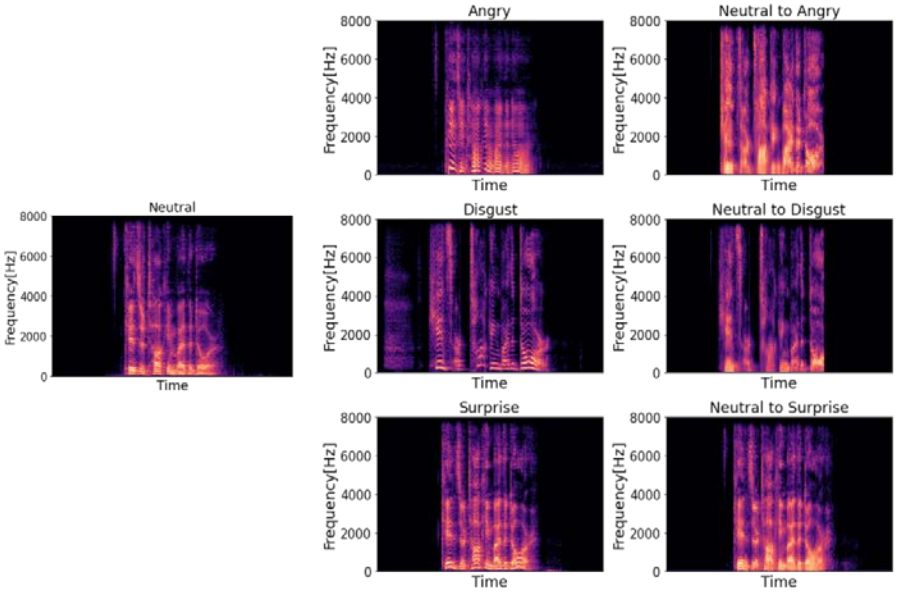

Fig. 6. Spectra converted through amplitude scaling

More clear emotion conversion is performed through the correction of the SP difference. A source spectral envelope is adjusted through amplitude scaling of two SPs in order to make it similar to the emotion of target data [25]. In the previous research employing amplitude scaling, the source spectrum envelope generated through Dynamic Frequency Warping is scaled in the way of making it similar to a target spectrum. In this study, the fake source spectral envelope generated by a generator is interpolated through amplitude scaling, in order to make it similar to a target spectral envelope. The amplitude scaling based spectral envelope transformation employs the spectral transform of log-scale so as to consider human hearing features. **Fig. 5** shows the process of converting an emotion through amplitude scaling. In **Fig. 5**, Test Phase is similar to Train phase. As for the source data features extracted by Vocoder, SP vector is obtained by Encoder. With the use of source SP vector as input, fake source SP is generated. The equal generator, which trained target SP as Random Latent Space, generates fake random SP. These two fake SP values are corrected through amplitude scaling, and are converted into waveforms, along with converted f0. **Fig. 6** shows the spectra converted through amplitude scaling. As shown in **Fig. 6**, spectra that include information of prosody and loudness are different depending on emotions. The emotion of the source speech data is "Neutral". In **Fig. 6**, column 1 is the spectrogram of the original data, and column 2 is the amplitude scaled source data with the target emotion. The difference was corrected by the proposed amplitude scaling. As a result, the SP of source data is similar to that of target data.

4. Experiments and Results

4.1 Emotion transformation using frequency synthesis

The speech data used in this study are based on RAVDESS that consisting of 7,356 files (24.8GB). The RAVDESS contains speech data of 24 speakers (12: female, 12: male pronounced in a neutral North American accent. The emotion classes that the speech represents are happy, calm, sad, fearful, angry, surprise, and disgust expressions.

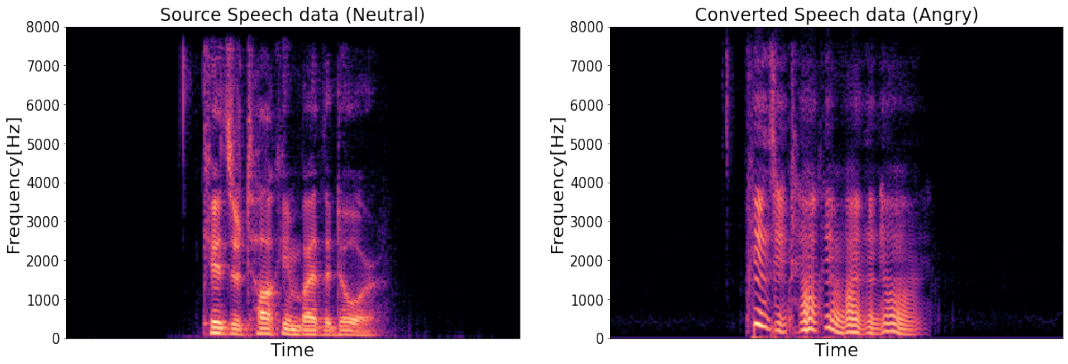

Fig. 7. Result of the proposed emotion conversion spectrogram using frequency synthesis

All emotion classes are divided into two intensities and contain neutral expressions. The data condition consists of voice-only (16 bit, 48 kHz .wav), face and voice (720p H.264, AAC 48 kHz, .mp4) and face-only formats. Speech features are extracted from the voice-only data by WORLD Vocoder, and are converted into a spectrogram usable as input data of a model. For the improvement in learning speed, parallel structure is applied. A frequency is corrected

to make emotion conversion clear, so that an amplitude difference is solved. For human hearing features, noise is resolved. **Fig. 7** shows the result of the proposed emotion conversion spectrogram using frequency synthesis. In the figure, source data (Neutral) and target data (Angry) results of source data are presented in a spectrogram.

4.2 Evaluation

The proposed emotion conversion model of speech data is implemented in the following experiment environment of operating system and hardware: Ubuntu, Intel Skylake Xeon, NVIDIA Tesla V100 X 2(20RFLOPS), and RAM 128GB. As for software, Tensorflow backend engine is used in the design and implementation of the proposed model. As emotions to convert in this study, four emotions-Neutral, Angry, Surprise and Disgust-are used. As performance evaluation indexes for the emotion conversion model, Mel Cepstral Distortion (MCD) [26] and Long-Spectral Distortion(LSD) [27] are employed. The performance of the proposed model is evaluated in two ways. Firstly, the proposed model is compared with a conventional emotion conversion model with the uses of MCD and LSD. Secondly, ablation study is conducted on the proposed method in order to evaluate its excellency.

Mel Cepstral Distortion is a measure of how an original voice is different from a converted voice in terms of mel-cepstral. **Equation 2** presents the formula of calculating a value of MCD.

$$\mathrm{MCD} = \left(\frac{10}{\ln 10}\right) * \frac{1}{N}\sum_{i=1}^{N}\sqrt{2\sum_{j=1}^{60}(mcep_{i,j}^{t} - mcep_{i,j}^{s})} \tag{2}$$

In the speech conversion system, the waveform is analyzed as a multi-dimensional vector in frames with regular intervals. In the equation, mcep^{t} and mcep^{s} represents the 60-dimensional Mel-cepstral coefficients(MCEP) values of original speech data and emotion-converted data; N is the number of frames of the entire sentence. Log-Spectral Distortion means the distance between original speech spectrum and converted-speech spectrum. The lower the LSD, the more similar. It is used to evaluate the quality of the converted voice compared to the original voice. **Equation 3** presents the formula of a value of LSD.

$$D_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10log_{10}\frac{P(w)}{\hat{P}(w)}\right]^{2} dw} \tag{3}$$

In the equation, $\mathrm{P}(w)$ and $\hat{P}(w)$ are power spectra, defined as the root mean square difference between spectra. Power spectra is an analysis method needed to quantitatively determine how much weight each vibration component occupies in a frequency. It is used as a measure of performance of interpolation.

In the first performance evaluation, the proposed model is compared with conventional emotion conversion models in terms of MCD and LSD. The models to compare are VAW-GAN [10] and Cycle-GAN [28] using logarithm gaussian (LG) based linear transformation for voice conversion [29][30]. **Table 1** shows the results of average MCD and LSD comparison between conventional voice conversion models and the proposed model.

Table 1. Results of average MSD and LSD comparison between conventional voice conversion models and the proposed model

	MCD	LSD
LG-Cycle-GAN	5.251 dB	6.897 dB
LG-VAW-GAN	4.656 dB	6.488 dB
CWT-AS-VAW-GAN	4.451 dB	6.239 dB

In **Table 1**, LG means a typical model of converting f0 with the use of LG based linear transformation. The proposed model is a VAW-GAN model using amplitude scaling and cwt. According to objective evaluation, the proposed model has better performance than other models. By correcting a signal difference between emotions through amplitude scaling, it has a lower distortion factor [31]. Therefore, it is judged that compared to other models, the proposed model is capable of converting emotion speeches significantly and effectively.

In the second performance evaluation, ablation study is conducted to find how the feature processing of the proposed method influences the performance of voice conversion [32]. The indexes used in this evaluation are equal to those in the first evaluation. **Table 2** presents the model performance according to the feature processing methods of the proposed emotion conversion model.

Table 2. Model performance according to the feature processing methods of the proposed emotion conversion model.

	MCD	LSD
CWT-VAW-GAN	4.458 dB	6.245 dB
CWT-AS-VAW-GAN	4.449 dB	6.248 dB

In **Table 2**, shows the difference in emotion conversion performance with and without Amplitude Scaling. CWT-VAW-GAN is a model of converting f0 through cwt and extracting prosody information. CWT-AS-VAW-GAN is a model of making a SP similar to a target emotion SP through amplitude scaling. As shown in **Table 2**, CWT-AS-VAW-GAN performing amplitude scaling shows the most significant performance in MCD. On contrary, in LSD, the result is higher than comparative model. Since the proposed method conducts amplitude scaling of SP to be similar to the target emotion, LSD, an evaluation factor that calculates distances of spectra, is judged to be high. In MCD, there is a performance difference between CWT-VAW-GAN and the proposed method. It means that the proposed model supports effective voice synthesis in the reproduction of emotional feature.

5. Conclusion

These days, there have been active researches on not only simple voice conversion models, but emotion conversion deep learning models in which linguistic information is preserved and emotions only are converted. Since prosody contains voice information that shows human emotions, it is required to devise a conversion technique in consideration of emotions. Therefore, this study proposed the voice frequency synthesis using VAW-GAN based amplitude scaling for the emotion transformation. The proposed method learns f0 with the uses of spectrum and cwt so that it converts emotions more effectively. In the learning process, there is an amplitude difference between an emotion spectrum of source data and a converted-emotion spectrum. Therefore, amplitude scaling is applied to adjust a spectral envelope and make a similar emotion. According to the objective performance evaluation with MCD and

LSD, the proposed model shows better performance than conventional emotion voice conversion models. In addition, in comparison of the generated spectrogram and source data, there is a very similarity. Given all, the proposed method supports emotion conversion in a speaker-independent way and in consideration of prosody features. In future research, effective speech conversion is needed for emotions that express an intensity relatively similar to the source data, in addition to the strong emotions that can be significantly distinguished by amplitude, such as Angry and Disgust used in this study. Therefore, the model will be improved for a more diverse and natural voice.

References

[1] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *J. signal process.*, vol. 2, no. 5, pp. 134-138, Oct. 2012. Article (CrossRef Link)

[2] K. Zhou, B. Sisman, H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *arXiv*, 2020. Article (CrossRef Link)

[3] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," in *Proc. of the 15th International Conference on Computer and Information Science (ICIS)*, pp. 1-5, Jun. 2016. Article (CrossRef Link)

[4] H. Ming, D. Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. of the International Conference of the Speech Communication Association*, pp. 2453-2457, Sep. 2016. Article (CrossRef Link)

[5] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, pp. 1-13, 2017. Article (CrossRef Link)

[6] L. Teng, Z. Fu, and Y. Yao, "Interactive translation in echocardiography training system with enhanced cycle-GAN," *IEEE Access*, vol. 8, pp. 106147-106156, 2020. Article (CrossRef Link)

[7] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *arXiv*, 2020. Article (CrossRef Link)

[8] H. Fei, D. Ji, Y. Zhang, and Y. Ren, "Topic-enhanced capsule network for multi-label emotion classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1839-1848, 2020. Article (CrossRef Link)

[9] S. Kim and H. Choi, "Emotional voice conversion using generative adversarial networks," *GAN.*, vol. 8, no. 3.169, pp. 5-784, 2017.

[10] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, H. M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv*, Jun. 2017. Article (CrossRef Link)

[11] H. J. Kwon, D. H. Shin, K. Chung, "PGGAN-based anomaly classification on chest x-ray using weighted multi-scale similarity," *IEEE Access*, vol. 9, pp. 113315-113325, Aug. 2021. Article (CrossRef Link)

[12] D. H. Shin, R. C. Park, K. Chung, "Decision boundary-based anomaly detection model using improved ANOGAN from ECG data," *IEEE Access*, vol. 8, pp. 108664-108674, Jun. 2020. Article (CrossRef Link)

[13] 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. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6199-6203, May. 2020. Article (CrossRef Link)

[14] P. Narváez, W. S. Percybrooks, "Synthesis of normal heart sounds using generative adversarial networks and empirical wavelet transform," *Appl. Sci.*, vol. 10, no. 19, pp. 7003-7018, 2020. Article (CrossRef Link)

[15] K. Chung, S. Y. Oh, "Voice activity detection using an improved unvoiced feature normalization process in noisy environments," *Wirel. Pers. Commun.*, vol. 89, no. 3, pp. 747-759, 2016. Article (CrossRef Link)
[16] J. Lee, Y. Jung, H. Kim, "Dual attention in time and frequency domain for voice activity detection," *arXiv*, Aug. 2020. Article (CrossRef Link)
[17] J. H. He, S. J. Kou, C. H. He, Z. W. Zhang, K. A. Gepreel, "Fractal oscillation and its frequency-amplitude property," *Fractals*, vol. 29, no. 4, pp. 2150105-991, Jan. 2021. Article (CrossRef Link)
[18] M. Tan, X. Xu, A. Boes, B. Corcoran, J. Wu, T. G. Nguyen, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, D. J. Moss, "Photonic RF arbitrary waveform generator based on a soliton crystal micro-comb source," *J. Light. Technol.*, vol. 38, no. 22, pp. 6221-6226, Jul. 2020. Article (CrossRef Link)
[19] R. Ramos-Aguilar, J. A. Olvera-López, I. Olmos-Pineda, S. Sánchez-Urrieta, "Feature extraction from EEG spectrograms for epileptic seizure detection," *Pattern Recognit. Lett.*, vol. 133, pp. 202-209. May. 2020. Article (CrossRef Link)
[20] S. Qamar, H. Mujtaba, H. Majeed, M. O. Beg, "Relationship identification between conversational agents using emotion analysis," *Cognit Comput*, vol. 13, no. 3, pp. 673-687, Jan. 2021. Article (CrossRef Link)
[21] M. Pasini, "MelGAN-VC: voice conversion and audio style transfer on arbitrarily long samples using spectrograms," *arXiv*, Dec. 2019. Article (CrossRef Link)
[22] S. R. Livingstone, F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, e0196391, 2018. Article (CrossRef Link)
[23] M. S. Al-Radhi, T. G. Csapó, C. Zainkó, G. Németh, "Continuous wavelet vocoder-based decomposition of parametric speech waveform synthesis," *arXiv*, Jun. 2021. Article (CrossRef Link)
[24] H. Ma, W. Huang, Y. Jing, S. Pignatti, G. Laneve, Y. Dong, H. Ye, L. Liu, A. Guo, J. Jiang, "Identification of Fusarium head blight in winter wheat ears using continuous wavelet analysis," *Sensors*, vol. 20, no. 1, pp. 20, Dec. 2020. Article (CrossRef Link)
[25] S. Cho, S. Jeon, W. Choi, R. Managuli, C. Kim, "Nonlinear *p*th root spectral magnitude scaling beamforming for clinical photoacoustic and ultrasound imaging," *Opt. Lett.*, vol. 45, no. 16, pp. 4575-4578, 2020. Article (CrossRef Link)
[26] W. Al-Dulaimi, T. K. Moon, J. H. Gunther, "Voice transformation using two-level dynamic warping and neural networks," *Signals*, *v*ol. 2, no. 3, pp. 456-474, 2021. Article (CrossRef Link)
[27] N. Hekmat, T. Vogel, Y. Wang, S. Mansourzadeh, F. Aslani, A. Omar, M. Hoffmann, F. Meyer, C. J. Saraceno, "Cryogenically cooled GaP for optical rectification at high excitation average powers," *Opt. Mater. Express.*, vol. 10, no. 11, pp. 2768-2782, 2020. Article (CrossRef Link)
[28] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to image translation using cycle-consistent adversarial networks," *arXiv*, 2017. Article (CrossRef Link)
[29] J. C. Kim, K. Chung, "Prediction model of user physical activity using data characteristics-based long short-term memory recurrent neural networks," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp. 2060-2077, Apr. 2019. Article (CrossRef Link)
[30] V. Popa, H. Silen, J. Nurminen, M. Gabbouj, "Local linear transformation for voice conversion," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4517-4520, 2012. Article (CrossRef Link)
[31] H. Yoo, K. Chung, "Deep learning-based evolutionary recommendation model for heterogeneous big data integration," *KSII Transactions on Internet and Information Systems*, Vol. 14, No. 9, pp. 3730-3744, Sep. 2020. Article (CrossRef Link)
[32] H. Yoo, R. C. Park, K. Chung, "IoT-based health big-data process technologies: a survey," *KSII Transactions on Internet and Information Systems*, Vol. 15, No. 3, pp. 974-992, Mar. 2021. Article (CrossRef Link)

Hye-Jeong Kwon received her B.S. degree from the Division of Computer Science and Engineering, Kyonggi University, South Korea, in 2021. She is currently in the Master course of Department of Computer Science, Kyonggi University, Suwon, South Korea. She has worked as a researcher at the Data Mining Lab., Kyonggi University. Her research interests include data mining, artificial intelligence, mobile healthcare, biomedical and health informatics, information systems, deep learning, machine learning, emerging health risk mining, and medical data analysis.

Min-Jeong Kim received her B.S. degree from the Division of Computer Science and Engineering, Kyonggi University, South Korea, in 2021. She is currently in the Master course of Department of Computer Science, Kyonggi University, Suwon, South Korea. She has worked as a researcher at the Data Mining Lab., Kyonggi University. Her research interests include data mining, artificial intelligence, deep learning, machine learning, natural language processing.

Ji-Won Baek has received B.S. degrees from the School of Computer Information Engineering, Sangji University, South Korea in 2017. She has worked for Data Management Department, Infiniq Co., Ltd. She has received an M.S. degree from the School of Department of Computer Science, Kyonggi University, South Korea in 2020. She is currently in the Doctor course of Department of Computer Science, Kyonggi University, South Korea. She has been a researcher at Data Mining Lab., Kyonggi University. Her research interests include data mining, data management, knowledge system, automotive testing, deep learning, medical data mining, healthcare, and recommendation.

Kyungyong Chung received his B.S., M.S., and Ph.D. degrees in 2000, 2002, and 2005, respectively, from the Department of Computer Information Engineering, Inha University, South Korea. He has worked for the software technology leading department of the Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a professor at the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a professor in the Division of AI Computer Science and Engineering, Kyonggi University, South Korea. He was named in 2017 as a Highly Cited Researcher by Clarivate Analytics. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.