

Article

Speech Emotion Recognition under Noisy Environments with SNR Down to -6 dB Using Multi-Decoder Wave-U-Net

Hyun-Joon Nam and Hong-June Park *

Department of Electronic and Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea; namhj@postech.ac.kr

* Correspondence: hjpark@postech.ac.kr

Abstract: A speech emotion recognition (SER) model for noisy environments is proposed, by using four band-pass filtered speech waveforms as the model input instead of the simplified input features such as MFCC (Mel Frequency Cepstral Coefficients). The four waveforms retain the entire information of the original noisy speech while the simplified features keep only partial information of the noisy speech. The information reduction at the model input may cause the accuracy degradation under noisy environments. A normalized loss function is used for training to maintain the high-frequency details of the original noisy speech waveform. A multi-decoder Wave-U-Net model is used to perform the denoising operation and the Wave-U-Net output waveform is applied to an emotion classifier in this work. By this, the number of parameters is reduced to 2.8 M for inference from 4.2 M used for training. The Wave-U-Net model consists of an encoder, a 2-layer LSTM, six decoders, and skip-nets; out of the six decoders, four decoders are used for denoising four band-pass filtered waveforms, one decoder is used for denoising the pitch-related waveform, and one decoder is used to generate the emotion classifier input waveform. This work gives much less accuracy degradation than other SER works under noisy environments; compared to accuracy for the clean speech waveform, the accuracy degradation is 3.8% at 0 dB SNR in this work while it is larger than 15% in the other SER works. The accuracy degradation of this work at SNRs of 0 dB, -3 dB, and -6 dB is 3.8%, 5.2%, and 7.2%, respectively.



Citation: Nam, H.-J.; Park, H.-J. Speech Emotion Recognition under Noisy Environments with SNR Down to -6 dB Using Multi-Decoder Wave-U-Net. *Appl. Sci.* **2024**, *14*, 5227. <https://doi.org/10.3390/app14125227>

Academic Editor: Douglas O'Shaughnessy

Received: 19 May 2024

Revised: 14 June 2024

Accepted: 14 June 2024

Published: 17 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: speech emotion recognition; emotion classification; SER; noisy environment; deep learning; speech waveform; denoising; band-pass filter; pitch; Wave-U-Net

1. Introduction

Emotion recognition will be widely used for human–machine interaction, where the machine responds to humans differently depending on the human emotions. The recognized emotions can be used for health care [1,2], disease diagnosis [3], education [4], vehicle safety [5], and music recommendation [6].

The emotion recognition can be conducted either by face or speech. Usually, the face emotion recognition (FER) is more accurate than the speech emotion recognition (SER), especially under noisy environments [7]. However, SER is more convenient than FER because of the smaller data size required and also when emotion recognition is a part of a larger speech application.

Most SER models focused on clean speech [8–16]. Only a few works are published for the SER under noisy environments [7,17–22]. Chakraborty et al. (2019) [17] uses MFCC (Mel Frequency Cepstral Coefficient) as the input feature, eliminates the noisy components from the input MFCC feature, and then classifies the emotion of an utterance from the cleaned-up MFCC using a conventional classifier. Huang et al. (2019) [18] compared the accuracy of FER and SER under noisy environments; for SER, a WPCC (Wavelet Packet Cepstral Coefficient) feature is generated from a noisy speech waveform and a deep neural network (DNN) classifies the WPCC feature into four emotions. Tiwari et al. (2020) [19] proposed

a generative noise model, where a randomly generated noise waveform is added to an utterance waveform at each epoch. The randomly generated noise waveform is obtained by multiplying a randomized frequency envelope by the frequency spectrum of white noise and then performing the inverse Fourier transform on the resultant frequency spectrum. An algorithm is used to extract a 6552 vector feature from an utterance. A DNN classifies the 6552 vector into four emotions. Xu et al. (2021) [7] generates an MFCC vector from a noisy speech waveform at every 32 ms, accepts the MFCC matrix for an utterance as the input feature, classifies into four emotions after passing the input feature through several CNN (convolutional neural network) layers, an attention layer, and a DNN classifier. Nam and Lee (2021) [20] used two CNN blocks in series; one CNN block accepts the Short-time Fourier Transform (STFT) of a noisy speech as input and generates the cleaned-up STFT as output, the other CNN block accepts the cleaned-up STFT as input and performs the emotion classification. Leem et al. (2022) [21] used a CNN block to rank the input features based on the noise resiliency, and then used only the high-ranked input features for the emotion recognition operation. After the above-mentioned work, Leem et al. (2023) [22] cleaned up the low-ranked input features in terms of the noise resiliency by using GAN, and then used all the input features including the cleaned-up ones for the emotion recognition.

The above-mentioned SERs for noisy environments tend to simplify the input speech waveform into rather compact features, such as MFCC, WPCC, and the 6552 vector for an utterance. Hence, a significant portion of information can be lost during the over-simplifying step, and the accuracy might be degraded especially under noisy environments.

In this work, the entire time-domain waveform is used as the input to the deep learning model without generating the simplified input features from the waveform. Besides, the input waveform is converted into four band-pass waveforms in order to preserve the small-magnitude high-frequency components. To perform the speech enhancement and the emotion classification at the same time during the training step, a Wave-U-Net [23] architecture was chosen in this work for the speech enhancement, because the Wave-U-Net accepts a waveform as input and generates the same size waveform as output. The Wave-U-Net of this work consists of an encoder, an LSTM (long short-term memory) layer, six decoders, and the associated skip-nets. The six decoders generate four band-pass waveforms, a pitch-related waveform, and an input waveform to the emotion classifier. One of the six decoders generates the pitch-related waveform because the large-magnitude pitch may be resilient to noise. The emotion classifier consists of a CNN layer followed by a DNN layer.

Because the Wave-U-Net of this work is trained to fit the four band-pass waveforms and the pitch-related waveform generated by the five decoders to the corresponding ground-truth waveforms with the normalized function, this work eliminates the noise components in the small-magnitude high-frequency signals as well as the large-magnitude low-frequency signals. Due to the denoising operation of the above-mentioned five decoders, the noise components in the input waveform to the emotion classifier are reduced, since all the six decoders share the encoder and the LSTM layer.

Section 2 explains the proposed U-Net-based deep learning model. Section 3 presents the experimental results. Section 4 discusses this work.

2. Proposed Deep Learning Model

The model proposed in this work accepts a noisy waveform (x) of an utterance as input and generates the probabilities of four emotions as output. The model consists of a block of four band-pass filters, a Wave-U-Net, and an emotion classifier (Figure 1). The passbands of the band-pass filters are 0~1 kHz, 1 kHz~2 kHz, 2 kHz~4 kHz, and 4 kHz~8 kHz. Three 401-tap FIR (finite impulse response) low-pass filters with bandwidths of 1 kHz, 2 kHz, and 4 kHz are used to generate the four band-pass waveforms.

The Wave-U-Net consists of an encoder, a 2-layer LSTM block, six decoders, and skip-nets which deliver the intermediate outputs of the encoder to the decoders. The six decoders generate four cleaned-up waveforms ($\hat{y}_1, \hat{y}_2, \hat{y}_4$, and $\hat{y}_8$), a pitch-related waveform

($\hat{y}_p$), and a waveform ($\hat{y}_{emo}$) which will be the input to the emotion classifier. The emotion classifier generates the probabilities of four emotions (angry, happy, sad, and neutral). The sample rate of the noisy input waveform (x) is 16 kS/s. The waveform is applied to the band-pass filter, and the four band-pass filter outputs are up-sampled to 64 kS/s before being applied to the encoder. All the six decoder outputs are down-sampled back to 16 kS/s.

The encoder is composed of five CNN layers in series connection (Figure 2a); each CNN layer consists of a series connection of a down-sampling Conv1D layer with $K = 12$ and $S = 4$, GELU, a Conv1×1 layer, and GLU; K is the kernel size of the 1D CNN and S is the stride. The shapes of the noisy waveform (x) and the band-pass filter output (x_{bpf}) are $[T]$ and $[T, 4]$, respectively; T is the number of sample points of an utterance at the sample rate of 16 kS/s. The encoder input (x'_{bpf}) has the shape of $[4 \times T, 4]$. The output channel size of each CNN layer increases from 20, 40, 80, 160, and 320, as the encoding step proceeds. The output of each CNN layer is sent to both the upper-level CNN layer and the skip-net. The encoder output (z , $[T/256, 320]$) is sent to the 2-layer bi-directional LSTM to generate the same-shape output (z').

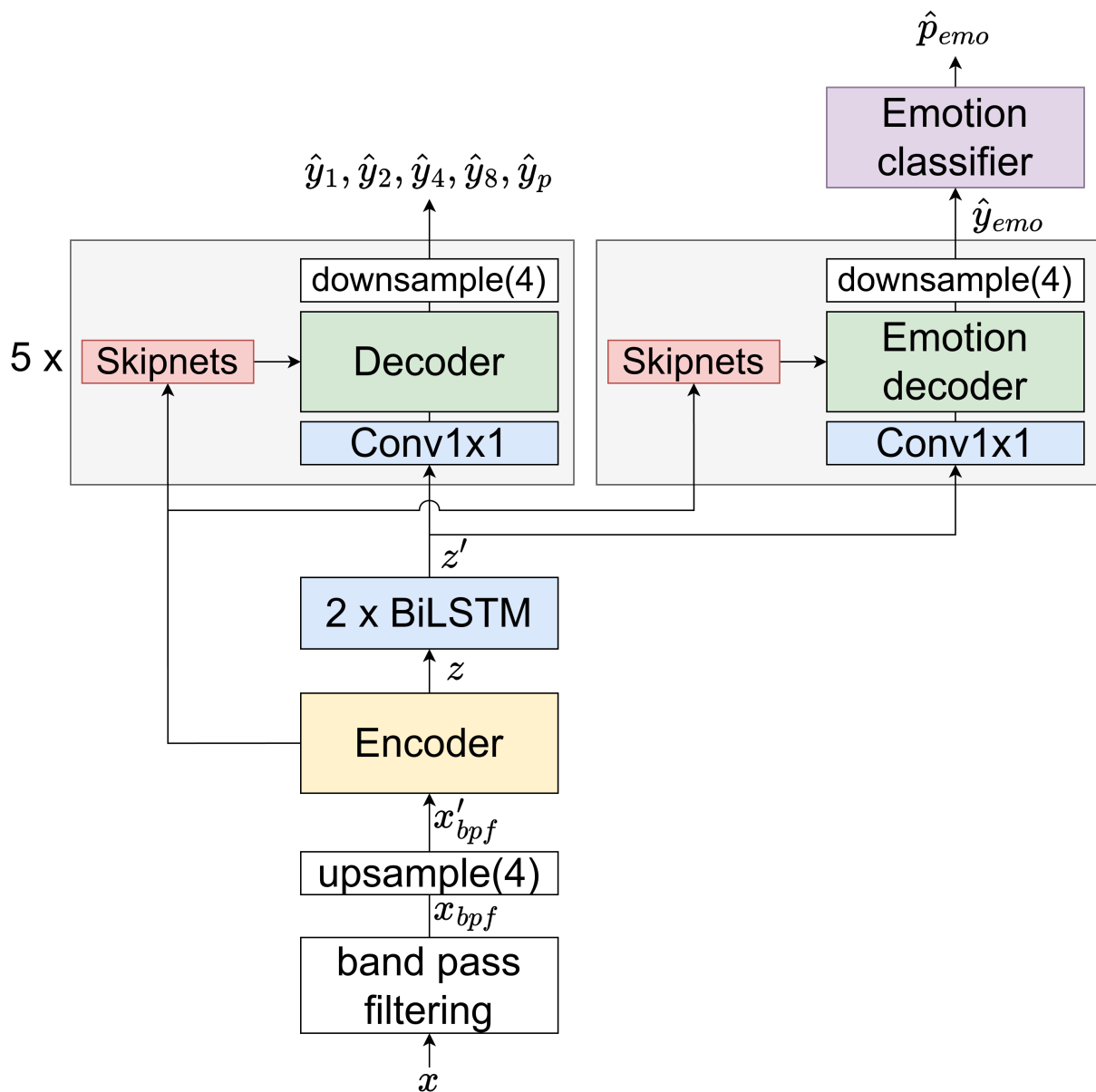
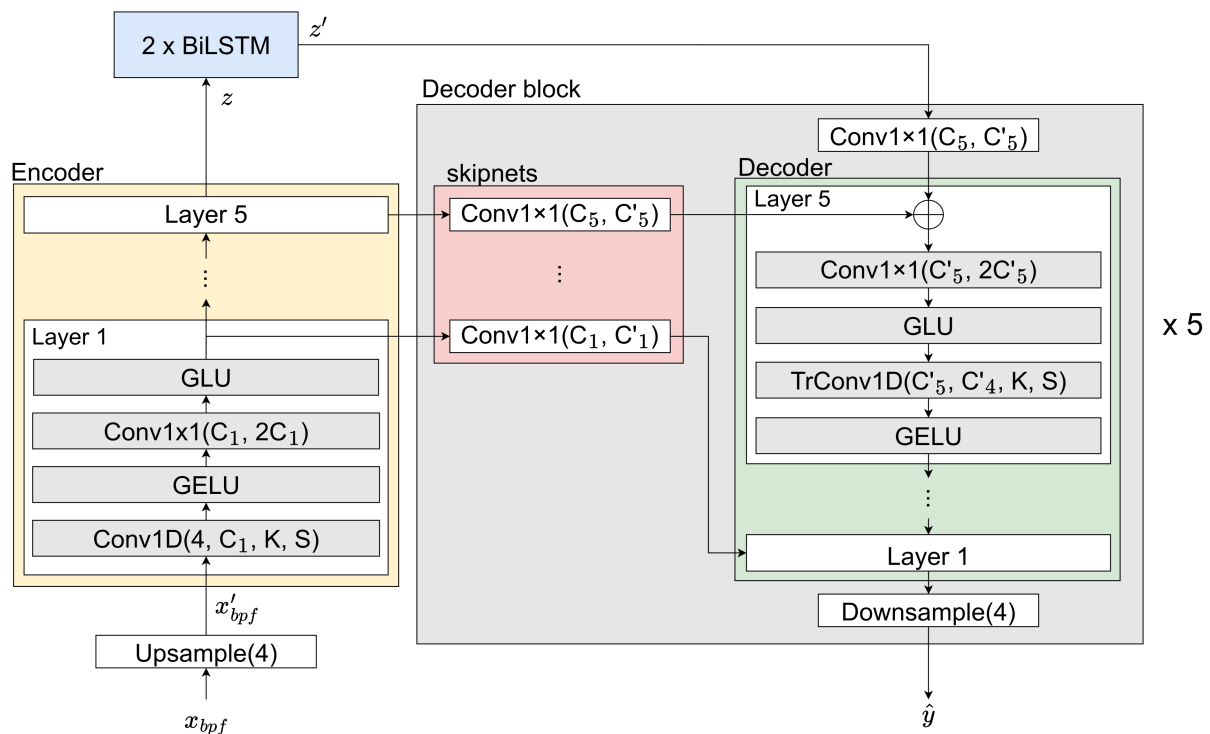
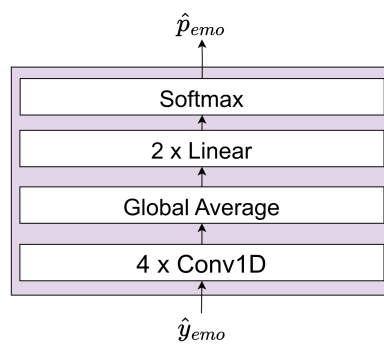


Figure 1. The proposed model for speech emotion recognition, a multi-decoder Wave-U-Net.

The LSTM output (z') is sent to six decoder blocks. Each decoder block consists of a $\text{Conv}1 \times 1$, a decoder, and a $4 \times$ down-sampler, and a skip-net. To maintain the same number of trainable parameters for the encoder and the six decoders, the input channel size of a decoder is around $1/\sqrt{6}$ times the output channel size of the corresponding encoder. The $\text{Conv}1 \times 1$ layer of each decoder block reduces the channel size from 320 to 131. Each decoder of the decoder block is composed of 5 CNN layers in series connection; each CNN layer consists of a series connection of a $\text{Conv}1 \times 1$ layer followed by GLU and an up-sampling TransposedConv1D layer with $K = 12$ and $S = 4$, and GELU. GELU is not added at the last CNN layer of the decoder. The input channel size of the CNN layers of a decoder is 131, 65, 33, 16, and 8, respectively. The shape of the final CNN layer of the decoder is $[4 \times T]$; this is down-sampled by $4 \times$ at the final stage of the decoder block to generate $\hat{y}$ with the shape of $[T]$. The skip-net of each decoder block consists of five $\text{Conv}1 \times 1$ layers; they reduce the output channel size to compensate for the channel size difference between the encoder and the decoder.



(a)



(b)

Figure 2. Detailed structure of the proposed speech emotion recognition model shown in Figure 1 (a) encoder and decoder block with skip-networks; (b) emotion classifier.

The emotion classifier accepts the input $\hat{y}_{emo}$ with the shape of [T] and generates four probabilities ($\hat{p}_{emo}$, [4]) for the emotions. The emotion classifier is composed of four Conv1D layers with $K = 12$ and $S = 4$ followed by the Global average layer, two linear layers, and a softmax layer (Figure 2b). The Global average layer performs the time-averaging operation and converts the input with the shape of $[T/256, 256]$ to the output with the shape of [256].

Table 1 lists the number of parameters of each block of the proposed model. The encoder, the LSTM, and the decoder blocks use 1.1 M, 1.2 M, and 1.7 M parameters, respectively. For training, all the blocks of Figure 1 are activated and 4.2 M parameters are used. For inference, the decoder blocks including the five decoders on the upper left-hand side of Figure 1 are turned off and only 2.8 M parameters are used.

Table 1. Number of parameters of each component of the proposed model in million units.

Step	Component				Total
	Encoder	LSTM	Decoder Blocks	Emotion Classifier	
Training	1.1	1.2	1.7	0.2	4.2
Inference	1.1	1.2	0.3	0.2	2.8

The model parameters are trained to fit the five waveforms ($\hat{y}_1$, $\hat{y}_2$, $\hat{y}_4$, $\hat{y}_8$, and $\hat{y}_p$) and an emotion probability vector ($\hat{p}_{emo}$) to the ground-truth counterparts. Among the ground-truth counterparts, the emotion probability vector (p_{emo}) is a one-hot vector ([4]) corresponding to the ground-truth clean speech waveform. The ground-truth band-pass waveforms (y_1 , y_2 , y_4 , and y_8) are generated by applying the three FIR low-pass filters to the ground-truth clean speech waveform (Figure 3). The ground-truth pitch-related waveform (y_p) is generated from the pitch mark timing data, which are provided by running REAPER [24] on the ground-truth clean speech waveform. Because the discrete-time pitch mark timing is difficult to train with the deep learning model, the pitch mark timing is replaced by a continuous-time pitch-related waveform (y_p); the pitch-related waveform is a sinusoidal wave during the voiced period and is 0 during the unvoiced period. Because the pitch mark timing is not uniform in time, the ground-truth pitch-related waveform (y_p) is a sine wave with non-uniform periods; the rising zero-crossing time of the sine wave matches the pitch mark timing and the amplitude of the sine wave follows the envelope of the ground-truth clean speech waveform. The envelope follows the absolute value of the ground-truth clean speech waveform with the exponential rise and fall decays; the time constant is 5 ms.

The loss function (Equation (1)) is the sum of the wave loss function (Equation (2)) and the emotion loss function (Equation (3)). The wave loss function is a normalized L1 loss averaged over N utterances and summed over five waveforms (1, 2, 4, 8, and p); the waveform is normalized by the standard deviation and ϵ is 10^{-4} . The emotion loss function is a cross-entropy function.

$$L_{\text{total}}(\hat{y}, y) = L_{\text{wave}}(\hat{y}, y) + L_{\text{emo}}(\hat{y}, y) \quad (1)$$

$$L_{\text{wave}}(\hat{y}, y) = \frac{1}{N \times T} \sum_{m \in M} \sum_{n=1}^N \sum_{t=1}^T \frac{|\hat{y}_{m,n}(t) - y_{m,n}(t)|}{\sigma_{y_{m,n}} + \epsilon}, \quad \text{where } M = \{1, 2, 4, 8, p\} \quad (2)$$

$$L_{\text{emo}}(\hat{p}, p) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C p_{n,c} \log(\hat{p}_{n,c}) \quad (3)$$

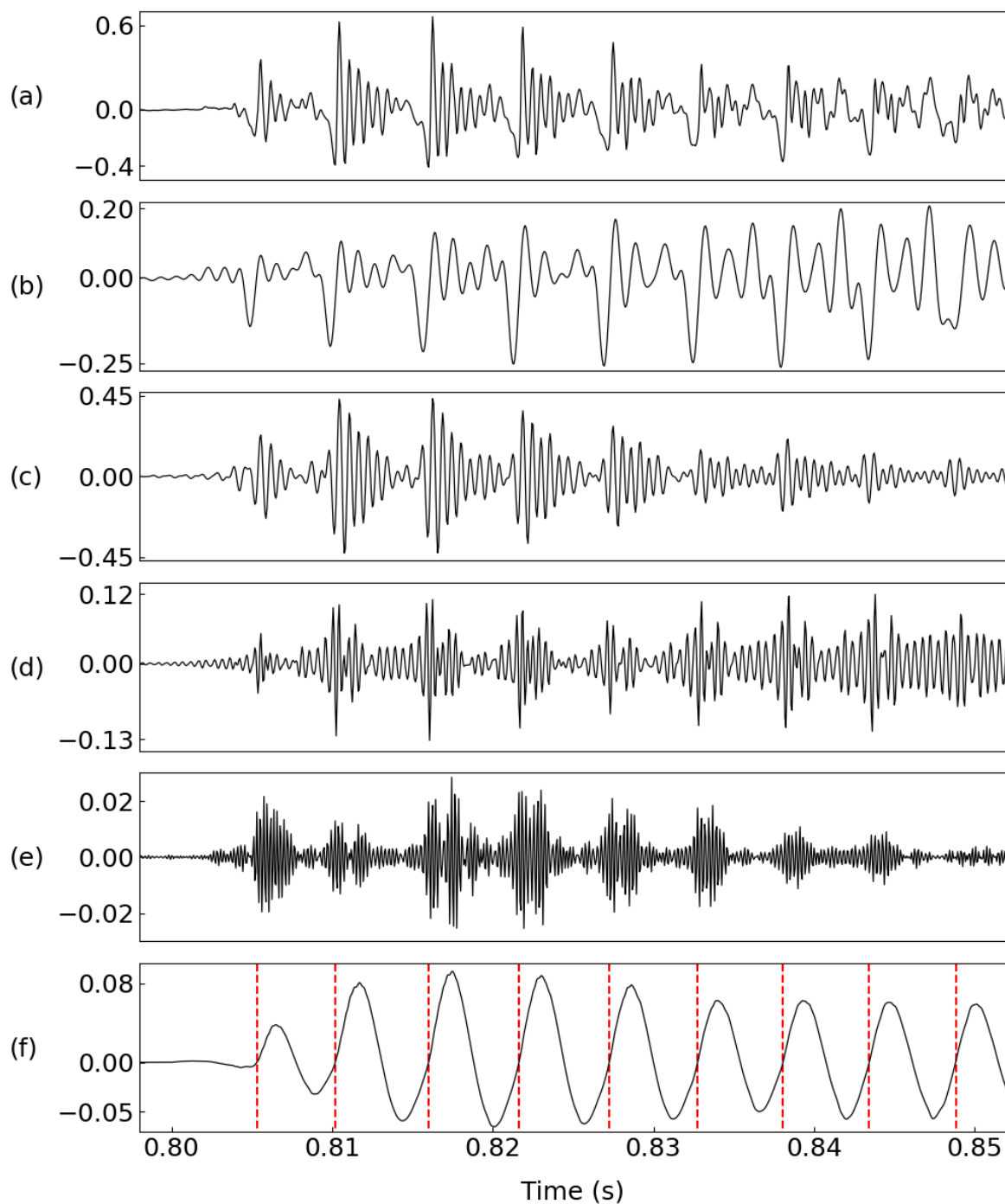


Figure 3. Generation of the ground-truth waveforms used for training. (a) ground-truth clean speech waveform, (b) ground-truth y_1 , (c) ground-truth y_2 , (d) ground-truth y_4 , (e) ground-truth y_8 , (f) ground-truth y_p , the vertical red dotted lines are the pitch mark timings.

3. Experimental Results

In this work, the clean speech utterances were taken from the 7-h-long IEMOCAP [25] with 5531 utterances with four emotions (angry, happy, sad, and neutral) (Table 2); 4425 (80%) utterances were assigned for training, 555 (10%) utterances were assigned for validation, and 551 (10%) utterances were assigned for the test. The average and the standard deviation of the time duration of the IEMOCAP utterances are 4.5 s and 3.2 s, respectively. The time duration of the clean speech utterance was fixed at 7.5 s by taking only a 7.5 s

portion of the IEMOCAP utterances longer than 7.5 s and zero-padding the IEMOCAP utterances shorter than 7.5 s.

Table 2. Dataset used in this work, ‘tr’: training, ‘va’: validation, ‘te’: test.

Type	Dataset	Subset	Usage	Time (hours)
Clean	IEMOCAP [25]	5531 utterances of four classes (angry, happy, neutral, sad)	tr, va, te	7.0
Noise	CHIME3 [26]	backgrounds	tr, va	8.4
	MS-SNSD [27]	noise_train	tr	2.8
		noise_test	va	0.5
	NoiseX92 [28]	-	te	1.0

The noise datasets in CHIME3 [26] and MS-SNSD [27] were used for training and validation. The NoiseX92 [28] was used for the test. The data augmentation was used for training; a noisy speech waveform is generated by adding a clean speech utterance and randomly selected noise data with a random SNR among -6 , -3 , 0 , 3 , 6 , and ∞ (clean) dB, at every epoch. To enhance the reliability of the test accuracy, the 551 utterances assigned for the test were multiplied by 15 times to 8265 utterances; an utterance out of the original 551 utterances was added to all the 15 noise files of the NoiseX92 dataset to generate 15 noisy speech waveforms.

The accuracy of this work was compared to the published SER works under noisy environments (Table 3). While the published SER works use simplified input features such as MFCC, WPCC, and a 6552 vector, this work uses the original speech waveform as input without any simplification process. It is estimated that the information loss of the simplified input features may be susceptible to accuracy degradation due to noise. The accuracy values of this work are 66.2% and 62.4% for SNRs of ∞ (clean) and 0 dB, respectively; the accuracy degradation at an SNR of 0 dB is 3.8% in this work, while the accuracy degradations of other works at an SNR of 0 dB are 15.5% or larger. The accuracy degradations of this work at SNRs of -3 dB and -6 dB are 5.2% and 7.2%, respectively.

Table 3. Performance comparison of speech emotion recognition models under noisy environments at different SNR values, diff. is the accuracy difference between clean (∞) and 0 dB SNR.

Methods	Input	Params	Accuracy (%)		
			∞	0 dB	diff.
[17]	MFCC	-	51.2	32.1	−19.1
[18]	WPCC	-	70	30	−40
[7]	MFCC	-	72.3	48.0 *	−24.3
[19]	6552 vector	19.7 M	62.7	47.2	−15.5
This work	waveform	4.2 M	66.2	62.4	−3.8

* 1.4 dB SNR.

The accuracy comparison of this work with [19] with SNR down to -6 dB (Table 4) presents that the accuracy degradations of this work due to noise are 0.7%, 2.4%, and 3.8% at SNRs of 20 dB, 5 dB, and 0 dB, respectively, while those in [19] are 2.2%, 11.1%, and 15.5%, respectively. This demonstrates that this work is more resilient to noise than [19].

Table 4. Accuracy comparison with SNR down to -6 dB.

Methods	Params	Accuracy (%)					
		∞	20 dB	5 dB	0 dB	−3 dB	−6 dB
[19]	19.7 M	62.7	60.5	51.6	47.2	-	-
This work	4.2 M	66.2	65.5	63.8	62.4	61.0	59.0

To assess the effect of the four band-pass waveforms used for input and ground-truth on the accuracy, the accuracy values of the three models are compared in Table 5. The six-decoder model (proposed model) uses the four band-pass filter waveforms (x_{bpf}) as input and generates four band-pass waveforms ($\hat{y}_1, \hat{y}_2, \hat{y}_4$, and $\hat{y}_8$) and a pitch-related waveform ($\hat{y}_p$) to compare with the ground-truth waveforms. Out of the six decoders, five decoders generate the above-mentioned five waveforms ($\hat{y}_1, \hat{y}_2, \hat{y}_4, \hat{y}_8$, and $\hat{y}_p$) and the remaining one decoder generates the emotion classifier input waveform ($\hat{y}_{emo}$). The three-decoder model uses the raw speech waveform as input and generates a speech waveform and a pitch-related waveform to compare with the ground-truth waveforms. Out of the three decoders, two decoders generate a speech waveform and $\hat{y}_p$ and the remaining one decoder generates the emotion classifier input waveform ($\hat{y}_{emo}$). The two-decoder model uses the raw speech waveform as input and generates a speech waveform to compare with the ground-truth waveform. Out of the two decoders, one decoder generates a speech waveform, and the other decoder generates the emotion classifier input waveform ($\hat{y}_{emo}$). On average, the accuracy of the six-decoder model (proposed model) is better than the three-decoder model and the two-decoder model by 4.3% and 5.7%, respectively. The pitch-related waveform enhanced the accuracy by 1.4% on average. The ablation study was performed to find out the significance of each of the five waveforms ($\hat{y}_1, \hat{y}_2, \hat{y}_4, \hat{y}_8$, and $\hat{y}_p$) used in the proposed model.

Table 5. Performance comparison of this work with six decoders with the model with two or three decoders.

Methods	Accuracy (%)					
	∞	20 dB	5 dB	0 dB	−3 dB	−6 dB
Proposed (six decoders)	66.2	65.5	63.8	62.4	61.0	59.0
speech waveform, pitch (three decoders)	61.9	61.4	60.1	57.7	57.1	53.9
speech waveform (two decoders)	58.6	58.8	58.3	57.7	56.8	53.4

Table 6 lists the accuracy values at different SNR values for the proposed model and the models where one of the five waveforms is removed from the proposed model; to remove a waveform from the proposed model, the corresponding decoder is also removed from the proposed model.

Table 6. Ablation study of this work.

Methods	Accuracy (%)					
	∞	20 dB	5 dB	0 dB	−3 dB	−6 dB
Proposed	66.2	65.5	63.8	62.4	61.0	59.0
w/o 0~1 kHz	60.7	60.5	59.9	57.5	56.4	53.5
w/o 1 kHz~2 kHz	61.9	61.4	60.0	58.7	57.1	54.9
w/o 2 kHz~4 kHz	63.2	62.3	61.9	60.8	59.6	57.2
w/o 4 kHz~8 kHz	62.1	62.5	62.2	60.8	59.0	56.6
w/o pitch	63.8	63.4	61.8	60.6	59.5	57.5

Table 7 presents the averaged accuracy of the ablation study models (Table 6); the accuracy values of each model were averaged w.r.t. SNR to obtain the averaged accuracy. The accuracy degradation data compared to the proposed model reveals that the 0~1 kHz waveform is the most important, and the 1 kHz~2 kHz waveform is the second most important. Besides, the high-frequency waveforms (2 kHz~4 kHz and 4 kHz~8 kHz) play substantial roles in emotion recognition. The pitch-related waveform is also helpful in enhancing the accuracy.

Table 7. Average accuracy for different SNRs of the ablation study models in Table 6 and the accuracy degradation from the proposed model.

	Removing Component					
	-	0~1 kHz	1 kHz~2 kHz	2 kHz~4 kHz	4 kHz~8 kHz	Pitch
Avg. acc. (%)	63.0	58.1	59.0	60.8	60.5	61.1
Acc. deg. (%)	0	−4.9	−4.0	−2.2	−2.5	−1.9

Because the IEMOCAP dataset is a relatively small dataset (7 h long), the model trained on the IEMOCAP dataset may be susceptible to the overfitting problem. Unseen test datasets (RAVDESS [29] and EmoDB [30]) were applied to the models trained with different datasets (IEMOCAP, ESD [31], and IEMOCAP + ESD) to assess the overfitting problem (Tables 8 and 9). The models trained with a single dataset (IEMOCAP or ESD) are susceptible to the overfitting problem, while the model trained with a combined dataset (IEMOCAP + ESD) is resilient to the overfitting problem. ESD is a 29-h dataset in English and Mandarin with five emotions [31]. RAVDESS is an 89-min dataset in English with eight emotions [29]. EmoDB is a 24-min dataset in German with seven emotions [30]. Only four emotions are used for the comparison in this work.

Table 8. Accuracy comparison of the RAVDESS [29] test dataset for models with different training datasets.

Training Dataset	Accuracy (%)					
	Clean	20 dB	5 dB	0 dB	−3 dB	−6 dB
IEMOCAP	47.2	47.3	46.3	45.3	45.1	44.6
ESD * [31]	48.7	48.8	48.4	47.5	46.1	44.3
IEMOCAP + ESD	58.8	58.7	57.9	56.4	55.1	52.7

* English only.

Table 9. Accuracy comparison of the EmoDB [30] test dataset for models with different training datasets.

Training Dataset	Accuracy (%)					
	Clean	20 dB	5 dB	0 dB	−3 dB	−6 dB
IEMOCAP	55.2	55.8	57.4	58.8	59.8	59.7
ESD * [31]	69.3	69.5	67.7	66.9	66.0	64.0
IEMOCAP + ESD	75.2	75.7	75.3	73.9	73.0	71.4

* English only.

4. Discussion

A noise-robust speech emotion recognition model is proposed in this work. The proposed model generates an emotion probability vector for four emotions (angry, happy, sad, and neutral) from a noisy speech waveform corresponding to an utterance. The input noisy speech waveform is converted to four band-pass filtered waveforms (0~1 kHz, 1 kHz~2 kHz, 2 kHz~4 kHz, and 4 kHz~8 kHz) by using 401-tap FIR filters.

The model consists of a multi-decoder Wave-U-Net and an emotion classifier. The multi-decoder Wave-U-Net of this work is composed of the band-pass filters, 4× up-samplers, a five-stage down-sampling encoder, a 2-layer LSTM, six five-stage up-sampling decoders, skip-nets, and 4× down-samplers. The noisy speech waveform input has the shape of [T] at the sample rate of 16,000 Samples/s, the following band-pass filters generate a [T,4] shaped signal, the following 4× up-samplers generate a [4*T, 4] shaped signal, the following encoder generates a [T/256, 320] signal, then the 2-layer LSTM generates a [T/256, 320] signal, then the six decoders generate a [4*T, 6] signal, and then the 4× down-samplers generate a [T, 6] signal. The outputs of the six decoder blocks ([T, 6]) consist of four band-pass filtered waveforms, one pitch-related waveform, and one emotion classifier

input waveform. The emotion classifier generates an emotion probability vector of [4] from the emotion classifier input waveform.

The six decoder blocks perform the denoising operation to generate a cleaned-up emotion classifier input waveform from a noisy speech waveform input. The six-decoder architecture reduces the number of model parameters to 2.8 M for inference from 4.2 M for training by turning off the five decoder blocks for the four band-pass filtered waveforms and the pitch-related waveform.

The model is trained to fit the five decoder block outputs (four band-pass filtered waveforms, a pitch-related waveform) and an emotion probability vector to the ground-truth counterparts by using a loss function which is the sum of the wave loss function and the cross-entropy emotion loss function. Because of the normalized L1 loss function used for the wave loss function, the high-frequency details are well retained in the emotion classifier input waveform.

The IEMOCAP [25] was used for the clean speech dataset. The noise datasets CHIME [26] and MS-SNSD [27] were used for training and validation, and NoiseX92 [28] was used for test. REAPER [24] was used to extract the pitch mark timing from the clean speech; the pitch mark timing data were used to generate the ground-truth pitch-related waveform.

The accuracy degradation for a noisy speech input waveform with the 0 dB SNR is 3.8% in this work while it is larger than 15% in the published SER works for the noisy environments [7,17–19]. It is estimated that the use of the four band-pass filtered waveforms of a noisy speech as an input to the model contributes to the robust operation of this work under noisy environments. The four band-pass filtered waveforms retain the entire information of the noisy speech input without any loss of information. Because the published SER works for noisy environments use simplified input features such as MFCC as input to the model, it is conjectured that a significant amount of information may be lost during the simplification process of the input features. The accuracy degrades by 5.2% and 7.2% at SNRs of −3 dB and −6 dB, respectively, compared to the clean speech case.

If the four band-pass filtered waveforms are replaced by the raw speech waveform in this work, the accuracy averaged over different SNR values down to −6 dB degrades by 4.3% compared to the clean speech case. If the pitch-related waveform is further removed in the above case, the averaged accuracy degrades by 5.7%.

The ablation study indicates that the 0~1 kHz waveform is the most important (accuracy degradation of 4.9% compared to the proposed model), and the 1 kHz~2 kHz waveform is the second most important (4.0%). The high-frequency waveforms (2 kHz~4 kHz and 4 kHz~8 kHz) cannot be neglected in the emotion recognition (2.2% and 2.5%, respectively). The pitch-related waveform is also helpful to enhance the accuracy (1.9%).

The model trained with a combined dataset (IEMOCAP + ESD) is resilient to the overfitting problem compared to the models trained with a single dataset (IEMOCAP or ESD).

All the SER datasets [25,29–31] used in this work are recorded by actors, not by ordinary people. Because the emotions from actors may be over-emphasized compared to the emotions from ordinary people in real life, the models trained with the datasets from actors might suffer from accuracy degradation in real-life applications. Although the SER datasets recorded by ordinary people in real life are not available nowadays except for the video challenge [32] to the best of the authors' knowledge, it is considered desirable to use the SER datasets recorded by ordinary people in real life for future works.

Author Contributions: Conceptualization, H.-J.N. and H.-J.P.; software, H.-J.N.; project administration, H.-J.P.; H.-J.N. and H.-J.P. worked together during the whole editorial process of the manuscript. All authors were involved in the preparation of this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2003451).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this work are available as follows. [25]: “<https://sail.usc.edu/iemocap>”; [26]: “<https://catalog.ldc.upenn.edu/LDC2017S24>”; [27]: “<https://github.com/microsoft/MS-SNSD>”; [28]: “<http://spib.linse.ufsc.br>”; [29]: “<https://zenodo.org/records/1188976>”; [30]: “<http://emodb.bilderbar.info>”; [31]: “<https://hltsingapore.github.io/ESD>” (accessed on 13 June 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dhuheir, M.; Albaseer, A.; Baccour, E.; Erbad, A.; Abdallah, M.; Hamdi, M. Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey. In Proceedings of the 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin, China, 28 June–2 July 2021; pp. 681–687. [\[CrossRef\]](#)
2. Kularatne, B.; Basnayake, B.; Sathmini, P.; Sewwandi, G.; Rajapaksha, S.; Silva, D.D. Elderly Care Home Robot using Emotion Recognition, Voice Recognition and Medicine Scheduling. In Proceedings of the 2022 7th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 7–9 December 2022; pp. 1–6. [\[CrossRef\]](#)
3. Tacconi, D.; Mayora, O.; Lukowicz, P.; Arnrich, B.; Setz, C.; Troster, G.; Haring, C. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In Proceedings of the 2008 Second International Conference on Pervasive Computing Technologies for Healthcare, Tampere, Finland, 30 January–1 February 2008; pp. 100–102. [\[CrossRef\]](#)
4. Garcia-Garcia, J.M.; Penichet, V.M.R.; Lozano, M.D.; Fernando, A. Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions. *Univers. Access Inf. Soc.* **2021**, *21*, 809–825. [\[CrossRef\]](#)
5. Giri, M.; Bansal, M.; Ramesh, A.; Satvik, D.; D, U. Enhancing Safety in Vehicles using Emotion Recognition with Artificial Intelligence. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 7–9 April 2023; pp. 1–10. [\[CrossRef\]](#)
6. Joel, J.S.; Ernest Thompson, B.; Thomas, S.R.; Revanth Kumar, T.; Prince, S.; Bini, D. Emotion based Music Recommendation System using Deep Learning Model. In Proceedings of the 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 26–28 April 2023; pp. 227–232. [\[CrossRef\]](#)
7. Xu, M.; Zhang, F.; Zhang, W. Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset. *IEEE Access* **2021**, *9*, 74539–74549. [\[CrossRef\]](#)
8. Chauhan, A.; Koolagudi, S.G.; Kafley, S.; Rao, K.S. Emotion recognition using LP residual. In Proceedings of the 2010 IEEE Students Technology Symposium (TechSym), Kharagpur, India, 3–4 April 2010; pp. 255–261. [\[CrossRef\]](#)
9. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and Recurrent Neural Networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Republic of Korea, 13–16 December 2016; pp. 1–4. [\[CrossRef\]](#)
10. Fahad, M.S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features. *Circuits Syst. Signal Process.* **2020**, *40*, 466–489. [\[CrossRef\]](#)
11. Koolagudi, S.G.; Reddy, R.; Rao, K.S. Emotion recognition from speech signal using epoch parameters. In Proceedings of the 2010 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 18–21 July 2010; pp. 1–5. [\[CrossRef\]](#)
12. Vernekar, O.; Nirmala, S.; Chachadi, K. Deep learning model for speech emotion classification based on GCI and GOI detection. In Proceedings of the 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON), Raigarh, India, 8–10 February 2023; pp. 1–6. [\[CrossRef\]](#)
13. Bhangale, K.; Kothandaraman, M. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics* **2023**, *12*, 839. [\[CrossRef\]](#)
14. Sun, T.W. End-to-End Speech Emotion Recognition with Gender Information. *IEEE Access* **2020**, *8*, 152423–152438. [\[CrossRef\]](#)
15. Morais, E.; Hoory, R.; Zhu, W.; Gat, I.; Damasceno, M.; Aronowitz, H. Speech Emotion Recognition Using Self-Supervised Features. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6922–6926. [\[CrossRef\]](#)
16. Pepino, L.; Riera, P.; Ferrer, L. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. *arXiv* **2021**, arXiv:2104.03502. <https://doi.org/10.48550/arXiv.2104.03502>.
17. Chakraborty, R.; Panda, A.; Pandharipande, M.; Joshi, S.; Kopparapu, S.K. Front-End Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 3257–3261. [\[CrossRef\]](#)
18. Huang, Y.; Xiao, J.; Tian, K.; Wu, A.; Zhang, G. Research on Robustness of Emotion Recognition Under Environmental Noise Conditions. *IEEE Access* **2019**, *7*, 142009–142021. [\[CrossRef\]](#)

19. Tiwari, U.; Soni, M.; Chakraborty, R.; Panda, A.; Kopparapu, S.K. Multi-Conditioning and Data Augmentation Using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7194–7198. [\[CrossRef\]](#)
20. Nam, Y.; Lee, C. Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions. *Sensors* **2021**, *21*, 4399. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Leem, S.G.; Fulford, D.; Onnela, J.P.; Gard, D.; Busso, C. Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6447–6451.
22. Leem, S.G.; Fulford, D.; Onnela, J.P.; Gard, D.; Busso, C. Selective Acoustic Feature Enhancement for Speech Emotion Recognition with Noisy Speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *32*, 917–929. [\[CrossRef\]](#)
23. Stoller, D.; Ewert, S.; Dixon, S. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. *arXiv* **2018**, arXiv:1806.03185.
24. Talkin, D. REAPER: Robust Epoch And Pitch Estimator. 2015. Available online: <https://github.com/google/REAPER> (accessed on 1 July 2023).
25. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
26. Barker, J.; Marxer, R.; Vincent, E.; Watanabe, S. The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines. In Proceedings of the ASRU, Scottsdale, AZ, USA, 13–17 December 2015; pp. 504–511.
27. Reddy, C.K.A.; Beyrami, E.; Pool, J.; Cutler, R.; Srinivasan, S.; Gehrke, J. A Scalable Noisy Speech Dataset and Online Subjective Test Framework. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1816–1820.
28. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [\[CrossRef\]](#)
29. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520.
31. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 920–924.
32. Dhall, A.; Singh, M.; Goecke, R.; Gedeon, T.; Zeng, D.; Wang, Y.; Ikeda, K. EmotiW 2023: Emotion recognition in the wild challenge. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 746–749.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.