Editorial Office,
Computers, Materials & Continua

Dear Editor:

I very much appreciate your taking the time to review my paper. I have revised the manuscript in response to the reviewer's comments and am submitting the revised version together with a summary of the revisions.

Corresponding author: Hyunki Hong
Title: Using Speaker-Specific Emotion Representations in Wav2vec 2.0-based Modules for Speech Emotion Recognition
Thank you again for your kind consideration.

Sincerely yours

Hyunki Hong

School of Software,
Chung-Ang University, 221 Huksuk-dong,
Dongjak-ku, Seoul, 156-756, Korea

Computers, Materials & Continua

Dear Reviewers:

Thank you for all the useful and constructive comments that will help us improve our manuscript. We have revised the manuscript according to the comments and our point-by-point responses are provided below. We have taken all the comments into consideration in the revision of our manuscript, and we appreciate your valuable comments again.

Corresponding author: Hyunki Hong
Title: Using Speaker-Specific Emotion Representations in Wav2vec 2.0-based Modules for Speech Emotion Recognition

Thank you again for your kind consideration.

Sincerely yours

Hyunki Hong

School of Software,
Chung-Ang University, 221 Huksuk-dong,
Dongjak-ku, Seoul, 156-756, Korea

Summary of revisions (Reviewer #1)

1. Reviewer's comment:

Please revise your paper according to the following suggestions: The abstract needs to be revised and expanded in accordance with the abstract guidelines.

Revisions made:

To improve clarity for readers, the major findings of our study were added to the abstract.

(Previous Version)

These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information. The experimental results show that the proposed approach outperforms previous methods, with a weighted accuracy of 72.14% and unweighted accuracy of 72.97% on the Interactive Emotional Dynamic Motion Capture (IEMOCAP) dataset.

Abstract, Sentence 8~9

(Modified Version)

These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information. Experimental results showed that the use of speaker-specific characteristics improves SER performance. Additionally, combining these with an angular marginal loss such as the Arcface loss improves intra-class compactness while increasing inter-class separability as demonstrated by the plots of t-distributed stochastic neighbor embeddings (t-SNE). The proposed approach outperforms previous methods, with a weighted accuracy (WA) of 72.14% and unweighted accuracy (UA) of 72.97% on the Interactive Emotional Dynamic Motion Capture (IEMOCAP) dataset.

2. Reviewer's comment:

English should be extensively revised and corrected. It is highly inadequate for publishing.

And it is strongly suggested that the whole work should be carefully revised.

Revisions made:

In response to the reviewer's comments, we have extensively revised the manuscript, corrected grammatical errors, and improved phrasing.

(Previous Version)

Speech emotion recognition (SER) is, therefore, one of the most active research areas in the computer science field ....

Section 1, Paragraph 1, Sentence 5

(Modified Version)

Speech emotion recognition (SER) is one of the most active research areas in the computer science field ....

(Previous Version)

These hand-crafted features have proven their potential in previous works. However, features and their representations should be tailored and optimized for specific tasks. Deep learning-based representations generated from actual waveforms or LLDs have shown better performance in SER.

Section 1, Paragraph 2, Sentence 2-3

(Modified Version)

Although the potential of hand-crafted features has been demonstrated in previous works, features and their representations should be tailored and optimized for specific tasks. Deep learning-based representations generated from LLDs or the actual waveform have shown better performance in SER.

(Previous Version)

Studies in psychology have shown that individuals have different voice characteristics depending on their culture, language, gender, and personality [3]. This implies that two speakers saying the same sentence with the same emotion are likely to express different acoustic properties in their voices. Several studies have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks. This implies that the acoustic properties expressed in the voices of two speakers saying the same sentence with the same emotion may vary. Several studies [4-5] have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks.

Section 1, Paragraph 3

(Modified Version)

Studies in psychology have shown that individuals have different vocal attributes depending on their culture, language, gender, and personality [3]. This implies that two speakers saying the same thing with the same emotion are likely to express different acoustic properties in their voices. Several studies [4-5] have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks.

(Previous Version)

The proposed model consists of an identification encoder and an emotion classifier. The wav2vec 2.0 [6] (base model) is used as a backbone for both of the proposed networks, and it generates emotion and speaker features from input audio waveform.

Section 1, Paragraph 4, Sentence 2-3

(Modified Version)

The proposed model consists of a speaker-identification network and an emotion classifier. The wav2vec 2.0 [6] (base model) is used as a backbone for both of the proposed networks, where it is used to extract emotion and speaker-specific features from input audio waveforms.

(Previous Version)

This tensor fusion is performed using an element-wise multiplication followed by a summation of vectors. The main contributions of this paper are summarized below:

Section 1, Paragraph 4, Sentence 6-7

(Modified Version)

This tensor fusion operation is performed using an element-wise multiplication followed by a summation of vectors. The main contributions of this paper are summarized as follows:

(Previous Version)

Two modules (emotion and identification) based on wav2vec 2.0 to generate a speaker-specific emotion representation from an audio input are proposed.

Section 1, Paragraph 5, Sentence 1

(Modified Version)

Two modules (speaker-identification network and emotion classification) based on wav2vec 2.0 that generate a speaker-specific emotion representation from an input audio segment are proposed.

(Previous Version)

The speaker-identification network was therefore pre-trained on a separate dataset (VoxCeleb1 [8]) with 1,251 speakers to support better generalization to unseen speakers.

Section 1, Paragraph 5, Sentence 4

(Modified Version)

The representations generated by the speaker-identification network which was pretrained on the VoxCeleb1 dataset [8] facilitate better generalization to unseen speakers.

(Previous Version)

The use of Arcface [9] and cross-entropy loss terms in the speaker-identification network was also explored and detailed evaluations provided.

Section 1, Paragraph 6, Sentence 2

(Modified Version)

The use of the Arcface [9] and cross-entropy loss terms in the speaker-identification network was also explored and detailed evaluations have been provided.

(Previous Version)

In more recent approaches, models learned a representation directly from raw waveforms instead of using hand-crafted representations like the human perception emulating Mel-filter banks used in generating the Mel-spectrogram. Time-Domain (TD) filter banks [13] used complex convolutional weights initialized with Gabor wavelets to learn filter banks from raw speech for end-to-end phone recognition. The proposed architecture has a convolutional layer followed by an L2 feature pooling-based modulus operation and a low-pass filter.

Subsection 2.2, Paragraph 1, Sentence 1-2

(Modified Version)

In more recent approaches, models learn a representation directly from raw waveforms instead of using hand-crafted representations like the human perception emulating Mel-filter banks used in generating the Mel-spectrogram. Time-Domain (TD) filter banks [14] use complex convolutional weights initialized with Gabor wavelets to learn filter banks from raw speech for end-to-end phone recognition. The proposed architecture has a convolutional layer followed by an l_2 feature pooling-based modulus operation and a low-pass filter.

(Previous Version)

However, log-scale compression and normalization reduce the scale of spectrograms, regardless of their contents.

Subsection 2.2, Paragraph 1, Sentence 7

(Modified Version)

A key limitation of this approach is that the log-scale compression and normalization used reduce the scale of spectrograms, regardless of their contents.

(Previous Version)

Reference [15] also learned a drop-in replacement for Mel-filter banks but replaced the static log compression with dynamic compression and addressed the channel distortion problems in the Mel-spectrogram log transformation using Per-Channel Energy Normalization (PCEN). This was calculated using a smoothed version of the filter bank energy, which is computed from a first-order infinite impulse response (IIR) filter. The smoothing coefficient was used to combine the smoothed version of the filter bank energy, and the current spectrogram energy was fixed in PCEN.

Subsection 2.2, Paragraph 2, Sentence 1~3

(Modified Version)

Wang et.al. [15] also propose a learned drop-in alternative to the Mel-filter banks but replaced static log compression with dynamic compression and addressed the channel distortion problems in the Mel-spectrogram log transformation using Per-Channel Energy Normalization (PCEN). This was calculated using a smoothed version of the filter bank energy, which was computed from a first-order infinite impulse response (IIR) filter. A smoothing coefficient was used to combine the smoothed version of the filter bank energy, and the current spectrogram energy.

(Previous Version)

Continuous Bags of Words (CBoW) and skip-gram variants were implemented and evaluated.

Subsection 2.3, Paragraph 2, Sentence 2

(Modified Version)

Continuous Bags of Words (CBoW) and skip-gram variants were also implemented and evaluated.

(Previous Version)

In this architecture, two neural networks were trained by maximizing the agreement of their output given the same input.

Subsection 2.3, Paragraph 2, Sentence 5

(Modified Version)

In this architecture, two neural networks were trained by maximizing the agreement in their outputs given the same input.

(Previous Version)

[4] uses an aggregation of an individual’s neutral speech to standardize emotional speech and improve the robustness of individual-agnostic emotion representation.

Subsection 2.4, Paragraph 1, Sentence 2

(Modified Version)

[4] uses an aggregation of individuals’ neutral speech to standardize emotional speech and improve the robustness of individual-agnostic emotion representations.

(Previous Version)

A Hanning window of the same size as the kernel and an STFT with a hop length equal to the stride was used.

Subsection 2.5, Paragraph 2, Sentence 4

(Modified Version)

A Hanning window of the same size as the kernel and a short-time Fourier transform (STFT) with a hop length equal to the stride were used.

(Previous Version)

The contextual encoder consists of a linear projection layer, a relative positional encoding 1-d convolution layer followed by a GeLU, and Transformer.

Subsection 2.5, Paragraph 2, Sentence 6

(Modified Version)

The contextual encoder consists of a linear projection layer, a relative positional encoding 1-d convolution layer followed by a GeLU and transformer model.

(Previous Version)

This amounts to choosing Vquantized representations (codebook entries) from multiple codebooks using a Gumble softmax, ...

Subsection 2.5, Paragraph 2, Sentence 11

(Modified Version)

This is achieved by choosing ￼quantized representations (codebook entries) from multiple codebooks using a Gumble softmax,

(Previous Version)

In Arcface, the representations were distributed around feature centers in a hypersphere with a radius.

Subsection 2.6, Paragraph 1, Sentence 4

(Modified Version)

In Arcface, the representations were distributed around feature centers in a hypersphere with a fixed radius.

(Previous Version)

A margin is added to the angular difference of the features in the same class to make learned features separable with a larger angular distance.

Subsection 2.6, Paragraph 2, Sentence 7

(Modified Version)

A margin is added to the angular difference between features in the same class to make learned features separable with a larger angular distance.

(Previous Version)

3 Method

3.1 Network Architecture

The same waveform is passed to the speaker-identification as well as the emotion recognition networks. Both networks are extensions of the wav2vec 2.0 model (consisting of 12 Transformer layers with 768 dimension embedding) and are trained using the Arcface loss. The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. During pre-training, the Arcface loss used in this model enables it to learn to distinguish between speakers. In the emotion recognition network (Fig. 2), input utterances are encoded into a vector that is combined with the output of the pre-trained speaker-identification network.

In the speaker identification network, the wav2vec 2.0 segment is used to encode input utterances into a latent 2-d representation vector that is passed to a single attention block. It is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional status. This is why, only a single attention block was used in the speaker-identification network. The model is trained to generate a representation which distinctly distinguishes between speakers using the Arcface loss. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Here, the four FC outputs are fed into a softmax function, and the probability of each speaker class is obtained.

In the emotion classification network, the encoding generated by the wav2vec 2.0 segment is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. The outputs of all the attention blocks are concatenated prior to the tensor fusion operation.

In order to address the differences in the lengths of utterances, max-pooling was applied across the time axis of each attention block, resulting in a d dimension vector as the final output. The output vectors of the speaker-identification network and emotion classification network have identical dimensions. During tensor fusion, an elementwise multiplication between $H = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k \times d}$) is performed. Summation is then performed across all k vectors as shown in Eq. (5),

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \qquad (5)$$

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is used to compute the Arcface loss for emotion classification. The fusion process combines the speaker-identification and emotion representations from each network into a speaker-specific emotion representation. To achieve this, the output vector of the attention block is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs. Fig. 3 shows the architecture of the proposed speaker-specific emotion representation-based SER. The angular distance between the tensor fused output vector and the center of the four emotion representation vectors is calculated as shown in Eq. (4).

Section 3

(Modified Version)

3 Methodology

In order to leverage speaker-specific speech characteristics to improve the performance of SER models, two wav2vec 2.0-based modules (speaker-identification network and emotion classification network) trained with the Arcface loss are proposed. The speaker-identification network extends the wav2vec 2.0 model with a single attention block that it uses to encode an input audio waveform into a speaker-specific representation. The emotion classification network uses a wav2vec 2.0-backbone as well as four attention blocks to encode the same input audio waveform into an emotion representation. These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information.

3.1 Speaker-Identification and Emotion Classification Networks

The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. The wav2vec 2.0 model is used to encode input utterances into a latent 2-d representation of shape $\mathbb{R}^{768 \times T}$, where T is the length of the input waveform. This latent representation is passed to a single attention block prior to performing a max-pooling operation that results in a 1-d vector of length 768. Only a single attention block was used in the speaker-identification network because it is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional state. In order to achieve a more robust distinction between speakers, the $\mathbb{R}^{d}$ shape speaker-identification representation (H_{id}) and the $\mathbb{R}^{\#ID\times d}$ shape Arcface center representation vector (W_{id}) for speaker classes are l_2 normalized and their cosine similarity computed. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Then, the FC outputs were fed into a softmax function, and the probability of each speaker class obtained. In Figure. 1, “#ID” represents the index of each speaker class. For example, in the VoxCeleb1 dataset with 1,251 speakers, the final #ID is #1,251.

In the emotion classification network (Fig. 2), the wav2vec 2.0 model is used to encode input utterances into a $\mathbb{R}^{768\times T}$shape representation. The encoding generated is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. Max-pooling is applied across the time axis to the outputs of each of the attention blocks. The max pooled outputs of the attention blocks, h_i are concatenated prior to the tensor fusion operation. During tensor fusion, an elementwise multiplication between $H_{emo} = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. As shown in Eq. (5), all the k vectors are summed to generate the final embedding.

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \tag{5}$$

where $e_i \in \mathbb{R}^{d}$ and $W_{fusion,i} \in \mathbb{R}^{d}$. The final embedding, E is l_2 normalized prior to computing the cosine similarity between Arcface center representation vectors $W_{emo} \in \mathbb{R}^{\#EMO\times d}$. In Figure. 2, “#EMO” represents the emotion class indices as defined in the IEMOCAP dataset. Here, 1_EMO, 2_EMO, 3_EMO, 4_EMO represent angry, happy, sad and neutral emotion classes respectively.

3.2 Speaker-Specific Emotion Representation Network

Fig. 3 shows the architecture of the proposed SER approach. The same waveform is passed to the speaker-identification network as well as the emotion recognition network. The speaker representation generated by the pretrained speaker-identification network is passed to the emotion classification network. More specifically, the output vector of the attention block from the speaker-identification network is concatenated to the outputs of the emotion classification network’s four attention blocks, resulting in a total of five attention block outputs, $H \in \mathbb{R}^{5\times d}$. The fusion operation shown in Eq. (5) is used to combine these representations into a single speaker-specific emotion representation E. The angular distance between the normalized tensor fused output vector and the normalized center of the four emotion representation vectors is calculated using Eq. (4). The emotion class predicted for any input waveform, is determined by how close its representation vector is to an emotion class's center vector.

(Previous Version)

The IEMOCAP [7] is a multimodal, multi-speaker emotion database recorded across five sessions with five pairs of male and female speakers performing improvisations or scripted scenarios.

Subsection 4.1, Paragraph 1, Sentence 1

(Modified Version)

The IEMOCAP [7] is a multimodal, multi-speaker emotion database recorded across five sessions with five pairs of male and female speakers performing improvisations as well as scripted scenarios.

(Previous Version)

Due to imbalances in the samples available for each label category, only neutral, happy (with exciting), sad, and angry classes have been used in line with previous studies [4], [26–28], [34–35].

Subsection 4.1, Paragraph 1, Sentence 5

(Modified Version)

Due to imbalances in the number of samples available for each label category, only neutral, happy (combined with exciting), sad, and angry classes have been used in line with previous studies [4], [27–29], [35–36].

(Previous Version)

Audio files longer than 15 seconds are truncated to 15 seconds because almost all of the audio samples available were less than 15 seconds long.

Subsection 4.1, Paragraph 1, Sentence 9

(Modified Version)

Audio files longer than 15 seconds are truncated to 15 seconds because almost all of the audio samples in the dataset were less than 15 seconds long.

(Previous Version)

As shown in Fig. 4, the dataset is unevenly distributed across emotion labels, with significantly more neutral and happy samples in most sessions.

Subsection 4.1, Paragraph 1, Sentence 12

(Previous Version)

VoxCeleb1 is an audio-visual dataset comprising 22,496 short interview clips extracted from YouTube.

Subsection 4.1, Paragraph 3, Sentence 3

(Modified Version)

VoxCeleb1 is an audio-visual dataset comprising 22,496 short interview clips extracted from YouTube videos.

(Modified Version)

As shown in Fig. 4, the dataset is unevenly distributed across emotion classes, with significantly more neutral and happy samples in most sessions.

(Previous Version)

In addition, [27] showed that either partially or entirely fine-tuning the wav2vec 2.0 segments results in the same boost in model performance on SER tasks in spite of the differences in computational costs.

Subsection 4.2, Paragraph 1, Sentence 3

(Modified Version)

In addition, [28] showed that either partially or entirely fine-tuning the wav2vec 2.0 segments results in the same boost in model performance on SER tasks despite the differences in computational costs.

(Previous Version)

The model and weights are provided by facebookresearch/fairseq [37].

Subsection 4.2, Paragraph 1, Sentence 5

(Modified Version)

The model and weights are provided by Facebook research under the Fairseq sequence modeling toolkit [38].

(Previous Version)

First, the identification network and emotion network are trained separately.

Subsection 4.2, Paragraph 2, Sentence 2

(Modified Version)

First, the speaker-identification network and emotion network are trained separately.

(Previous Version)

A $10-8$ weight decay is applied and Adam [38] optimizer with betas set to (0.9, 0.98) is used.

LambdaLR scheduler reduces the learning rate by a factor of 0.98 after every epoch.

Subsection 4.2, Paragraph 2, Sentence 5-6

(Modified Version)

A $10-8$ weight decay is applied and the Adam [39] optimizer with beta values set to (0.9, 0.98) is used. The LambdaLR scheduler reduces the learning rate by a factor of 0.98 after every epoch.

(Previous Version)

As shown in Fig 6, the speaker identification network may be unable to generate accurate representations for very short audio samples. For example, representations of audio clips that are less than 3 seconds long may be misclassified.

Subsection 5.1, Paragraph 1, Sentence 5

(Modified Version)

As shown in Fig 6, the speaker identification network may be unable to generate accurate representations for audio samples that are too short. Representations of audio clips that are less than 3 seconds long are particularly likely to be misclassified.

(Previous Version)

Table 3 shows a comparison of our methods' performance against previous studies.

Subsection 5.1, Paragraph 1, Sentence 2

(Modified Version)

Table 3 shows a comparison of our methods' performance against that of previous studies.

(Previous Version)

The proposed speaker-identification network can be fine-tuned under three different configurations:

Subsection 5.2, Paragraph 1, Sentence 1

(Modified Version)

The proposed speaker-identification network was fine-tuned under three different configurations:

(Previous Version)

Model performance under these configurations is evaluated with four emotion attention blocks, because training the emotion network with four attention blocks showed the best performance in prior experiments.

Subsection 5.2, Paragraph 1, Sentence 5

(Modified Version)

Since training the emotion classification network with four attention blocks showed the best performance in prior experiments, fine-tuning performance was evaluated under this configuration.

(Previous Version)

Due to the small number of speakers in the IEMOCAP dataset, the model quickly converged on a representation that can distinguish speakers but was unable to generalize to unseen speakers.

Subsection 5.2, Paragraph 1, Sentence 7

(Modified Version)

Due to the small number of speakers in the IEMOCAP dataset, the model quickly converged on a representation that could distinguish speakers but was unable to generalize to unseen speakers.

(Previous Version)

In Fig 7 (b), increasing β, which controls the significance of the identification loss, improves emotion classification accuracy. On the contrary, in Fig 7 (c), increasing causes the emotion classification accuracy to deteriorate.

Subsection 5.2, Paragraph 12 Sentence 1-2

(Modified Version)

Fig 7 (b) shows that increasing β, which controls the significance of the identification loss, improves emotion classification accuracy when the Arcface center representation vectors are frozen. Conversely, Fig 7 (c) shows that increasing β, causes the emotion classification accuracy to deteriorate when the entire model is fine-tuned.

(Previous Version)

In the top row of Fig 8 (a) and (b), the representations are colored according to the emotion class, and in the bottom row the representations are colored according to speakers. The same layout is applied to Fig 9 (a) and (b).

Subsection 5.2, Paragraph 3, Sentence 3~4

(Modified Version)

In the top row of Fig 8 (a) and (b), a representation's color is determined by its predicted emotion class and in the bottom row, a representation's color is determined by its predicted speaker class. The same descriptors are applicable to Fig 9 (a) and (b).

(Previous Version)

In comparison to Fig 9 (a), Fig 9 (b) shows that fine-tuning both the speaker-identification network and the emotion classification network increases inter-class separability between the emotion representations of speakers, while retaining a speaker-specific information.

Subsection 5.2, Paragraph 4, Sentence 1

(Modified Version)

In contrast to Fig 9 (a), Fig 9 (b) shows that fine-tuning both the speaker-identification network and the emotion classification network increases inter-class separability between the emotion representations of speakers, while retaining speaker-specific information.

(Previous Version)

This results in a slight improvement in overall SER performance, which is in line with the findings presented in Fig 7 (b) and (c).

Subsection 5.2, Paragraph 4, Sentence 2

(Modified Version)

This results in a slight improvement in the overall SER performance, which is in line with the findings presented in Fig 7 (b) and (c).

(Previous Version)

5.3 Comparison of Previous Methods

Subsection 5.3, Heading

(Modified Version)

5.3 Comparison Against Previous Methods

(Previous Version)

In Table 3, we have compared the proposed method against the previous SER methods methods based on the wav2vec 2.0 or the Arcface loss.

Subsection 5.3, Paragraph 1, Sentence 1

(Modified Version)

In Table 3, we have compared the proposed method against previous SER methods based on the wav2vec 2.0 or the Arcface loss.

(Previous Version)

The experiment showed that the configuration using four attention blocks in the emotion network and fine-tuning with the speaker-identification network frozen (Fig 7 (a)) provided the best performance.

Subsection 5.3, Paragraph 1, Sentence 4

(Modified Version)

Experiments showed that the configuration using four attention blocks in the emotion network and fine-tuning with the speaker-identification network frozen (Fig 7 (a)) provided the best performance.

(Previous Version)

This paper proposed two modules for generating a speaker-specific emotion representation for SER.

Subsection 6, Paragraph 1, Sentence 1

(Modified Version)

In this study, two modules for generating a speaker-specific emotion representation for SER are proposed.

Summary of revisions (Reviewer #2)

1. Reviewer's comment:

Please download the CMC template online and revise your paper according to the CMC template online and the following suggestions.

The initial letter of each notional word in the paper title should be capitalized.

Revisions made:

In response to the reviewer's suggestions, we downloaded the CMC template and re-formatted our manuscript in accordance with the CMC journal guidelines. Additionally, we capitalized the first letter of every word in the manuscript title.

(Previous Version)

Using Speaker-specific Emotion Representations in wav2vec 2.0-based modules for Speech Emotion Recognition

(Manuscript Title)

(Modified Version)

Using Speaker-Specific Emotion Representations in Wav2vec 2.0-based Modules for Speech Emotion Recognition

2. Reviewer's comment:

Regarding Headings including Level one, two and three headings, the initial letter of each notional word of the Headings should be capitalized.

E. g. In the heading of Section 2.1, crafted should be Crafted.

E. g. In the heading of Section 5.2, tuning should be Tuning.

Revisions made:

In response to the reviewer's comment, we capitalized the first letter of every word in a subsection title.

(Previous Version)

2.1 Hand-crafted Audio Representation

Subsection 2.1, Heading 1

(Modified Version)

2.1 Hand-Crafted Audio Representations

(Previous Version)

5.2 Partially and Entirely Fine-tuning Networks

Subsection 5.2, Heading 2

(Modified Version)

5.2 Partially and Entirely Fine-Tuning Networks

3. Reviewer's comment:

Your section should be numbered with more than one serial number. If you have no subsection, you need not number your section. E.g., you have only Section 3.1, you can remove it, or supply at least one more subsection like Section 3.2.

Revisions made:

In the previous manuscript version, there was only one subsection under section 3. In order to make the method description clearer, we reorganized the content in section 3 into sections 3.1 and 3.2 and provided additional details in each subsection.

(Previous Version)

3 Method

3.1 Network Architecture

The same waveform is passed to the speaker-identification as well as the emotion recognition networks. Both networks are extensions of the wav2vec 2.0 model (consisting of 12 Transformer layers with 768 dimension embedding) and are trained using the Arcface loss. The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. During pre-training, the Arcface loss used in this model enables it to learn to distinguish between speakers. In the emotion recognition network (Fig. 2), input utterances are encoded into a vector that is combined with the output of the pre-trained speaker-identification network.

In the speaker identification network, the wav2vec 2.0 segment is used to encode input utterances into a latent 2-d representation vector that is passed to a single attention block. It is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional status. This is why, only a single attention block was used in the speaker-identification network. The model is trained to generate a representation which distinctly distinguishes between speakers using the Arcface loss. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Here, the four FC outputs are fed into a softmax function, and the probability of each speaker class is obtained.

In the emotion classification network, the encoding generated by the wav2vec 2.0 segment is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. The outputs of all the attention blocks are concatenated prior to the tensor fusion operation.

In order to address the differences in the lengths of utterances, max-pooling was applied across the time axis of each attention block, resulting in a d dimension vector as the final output. The output vectors of the speaker-identification network and emotion classification network have identical dimensions. During tensor fusion, an elementwise multiplication between $H = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. Summation is then performed across all k vectors as shown in Eq. (5),

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \qquad (5)$$

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is used to compute the Arcface loss for emotion classification. The fusion process combines the speaker-identification and emotion representations from each network into a speaker-specific emotion representation. To achieve this, the output vector of the attention block is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs. Fig. 3 shows the architecture of the proposed speaker-specific emotion representation-based SER. The angular distance between the tensor fused output vector and the center of the four emotion representation vectors is calculated as shown in Eq. (4).

Section 3

(Modified Version)

3 Methodology

In order to leverage speaker-specific speech characteristics to improve the performance of SER models, two wav2vec 2.0-based modules (speaker-identification network and emotion classification network) trained with the Arcface loss are proposed. The speaker-identification network extends the wav2vec 2.0 model with a single attention block that it uses to encode an input audio waveform into a speaker-specific representation. The emotion classification network uses a wav2vec 2.0-backbone as well as four attention blocks to encode the same input audio waveform into an emotion representation. These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information.

3.1 Speaker-Identification and Emotion Classification Networks

The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. The wav2vec 2.0 model is used to encode input utterances into a latent 2-d representation of shape $\mathbb{R}^{768\times T}$, where T is the length of the input waveform. This latent representation is passed to a single attention block prior to performing a max-pooling operation that results in a 1-d vector of length 768. Only a single attention block was used in the speaker-identification network because it is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional state. In order to achieve a more robust distinction between speakers, the $\mathbb{R}^{d}$ shape speaker-identification representation (H_{id}) and the $\mathbb{R}^{\#ID\times d}$ shape Arcface center representation vector (W_{id}) for speaker classes are l_2 normalized and their cosine similarity computed. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Then, the FC outputs were fed into a softmax function, and the probability of each speaker class obtained. In Figure. 1, “#ID” represents the index of each speaker class. For example, in the VoxCeleb1 dataset with 1,251 speakers, the final #ID is #1,251.

In the emotion classification network (Fig. 2), the wav2vec 2.0 model is used to encode input utterances into a $\mathbb{R}^{768\times T}$shape representation. The encoding generated is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. Max-pooling is applied across the time axis to the outputs of each of the attention blocks. The max pooled outputs of the attention blocks, h_i are concatenated prior to the tensor fusion operation. During tensor fusion, an elementwise multiplication between $H_{emo} = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. As shown in Eq. (5), all the k vectors are summed to generate the final embedding.$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i$,

(5)

where $e_i \in \mathbb{R}^{d}$ and $W_{fusion,i} \in \mathbb{R}^{d}$. The final embedding, E is l_2 normalized prior to computing the cosine similarity between Arcface center representation vectors $W_{emo} \in \mathbb{R}^{\#EMO\times d}$. In Figure. 2, “#EMO” represents the emotion class indices as defined in the IEMOCAP dataset. Here, 1_EMO, 2_EMO, 3_EMO, 4_EMO represent angry, happy, sad and neutral emotion classes respectively.

3.2 Speaker-Specific Emotion Representation Network

Fig. 3 shows the architecture of the proposed SER approach. The same waveform is passed to the speaker-identification network as well as the emotion recognition network. The speaker representation generated by the pretrained speaker-identification network is passed to the emotion classification network. More specifically, the output vector of the attention block from the speaker-identification network is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs, $H \in \mathbb{R}^{5\times d}$. The fusion operation shown in Eq. (5) is used to combine these representations into a single speaker-specific emotion representation E. The angular distance between the normalized tensor fused output vector and the normalized center of the four emotion representation vectors is calculated using Eq. (4). The emotion class predicted for any input waveform, is determined by how close its representation vector is to an emotion class's center vector.

4. **Reviewer's comment**:

All references should be cited in your text, sequentially. E.g., In your paper, you cite Reference [37] after Reference [35], you should cite Reference [36] after Reference [35], then you can cite Reference [37].

Revisions made:

In response to the reviewer's comments, we have changed the order of references 35 and 36 and revised the citation in the previous version of the manuscript.

(Previous Version)

Due to imbalances in the samples available for each label category, only neutral, happy (with exciting), sad, and angry classes have been used in line with previous studies [4], [26–28], [34–35].

Subsection 4.1, Paragraph 1, Sentence 5

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audiobooks," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5206–5210, 2015.

[36] M. Hou, Z. Zhang, Q. Cao, D. Zhang, and G. Lu, "Multi-view speech emotion recognition via collective relation construction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30 pp. 218–229, 2022.

(Modified Version)

Due to imbalances in the number of samples available for each label category, only neutral, happy (with exciting), sad, and angry classes have been used in line with previous studies [4], [27–29], [34–36].

[36] M. Hou, Z. Zhang, Q. Cao, D. Zhang and G. Lu, “Multi-view speech emotion recognition via collective relation construction,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 218–229, 2022.
[37] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, pp. 5206–5210, 2015.

5. Reviewer’s comment:

Some of your references are not in the right format.

Revisions made:

In order to respond to the reviewer’s comments, we have reformatted and organized the references in numerical order of appearance in the manuscript.

(Previous Version)

[1] D. Amodei et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, New York, New York, USA, pp. 173–182, 2016. [Online]. Available: https://proceedings.mlr.press/v48/amodei16.html

(Modified Version)

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg *et.al.*, “Deep speech 2: End-to-end speech recognition in English and mandarin,” in *Proceedings of the33rd International Conference on Machine Learning*, vol. 48, New York, New York, USA, pp. 173–182, 2016.

(Previous Version)

[2] J. Shen et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in 2*018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.

(Modified Version)

[2] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly *et.al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in 2*018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.

(Previous Version)

[3] L. Marianne and P. Belin. "Human voice perception." *Current biology*, vol. 21, 2011.

(Modified Version)

[3] L. Marianne and P. Belin. "Human voice perception." *Current Biology,* vol. 21, no. 4, pp. R143- R145, 2011.

(Previous Version)

[5] C. Le Moine, N. Obin, and A. Roebel, "Speaker Attentive Speech Emotion Recognition," in *Interspeech*, pp. 2866-2870, 2021.

(Modified Version)

[5] C. Le Moine, N. Obin and A. Roebel, "Speaker Attentive Speech Emotion Recognition," in *Proc. of Interspeech*, Brno, Czechia, pp. 2866-2870, 2021.

(Previous Version)

[6] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in neural information processing systems,* vol. 33, pp. 12449-12460, 2020.

(Modified Version)

[6] A. Baevski, Y. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems,* vol. 33, pp. 12449-12460, 2020.

(Previous Version)

[7] C. Busso et al., "IEMOCAP: interactive emotional dyadic motion capture database," in *Language Resources and Evaluation* vol. 42 pp. 335-359, 2008.

(Modified Version)

[7] C. Busso, M. Bulut, C.C. Lee, E.A. Kazemzadeh, E. Provost *et.al.*, “IEMOCAP: interactive emotional dyadic motion capture database,” in *Language Resources and Evaluation* vol. 42 pp. 335-359, 2008.

(Previous Version)

[8] A. Nagrani, J.S. Chung, A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in *Proc. Interspeech 2017,* pp. 2616-2620, 2017.

(Modified Version)

[8] A. Nagrani, J.S. Chung and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in *Proc. of Interspeech 2017,* pp. 2616-2620, 2017.

(Previous Version)

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4685–4694, 2019.

(Modified Version)

[9] J. Deng, J. Guo, N. Xue and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 4685–4694, 2019.

(Previous Version)

[10] B. Schuller, G. Rigoll, and M. Lang, “Hidden markov model-based speech emotion recognition,” in *Proceedings of the 2003 International Conference on Multimedia and Expo*, vol. 2, pp. 401-404, 2003.

(Modified Version)

[10] B. Schuller, G. Rigoll and M. Lang, “Hidden markov model-based speech emotion recognition,” in *Proceedings of the 2003 International Conference on Multimedia and Expo*, vol. 2, pp. 401-404, 2003.

(Previous Version)

[11] M. Chen, X. He, J. Yang, and H. Zhang, “3-d convolutional recurrent neural networks with attention model for speech emotion recognition,” in *IEEE Signal Processing Letters*, vol. 25(10), pp. 1440–1444, 2018.

(Modified Version)

[11] M. Chen, X. He, J. Yang and H. Zhang, “3-d convolutional recurrent neural networks with attention model for speech emotion recognition,” in *IEEE Signal Processing Letters*, vol. 25, no.10, pp. 1440–1444, 2018.

(Previous Version)

[12] W. Zhu and X. Li, “Speech emotion recognition with global-aware fusion on multi-scale feature representation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6437–6441, 2022.

(Modified Version)

[12] W. Zhu and X. Li, “Speech emotion recognition with global-aware fusion on multi-scale feature representation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, pp. 6437–6441, 2022.

(Previous Version)

[13] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, “Learning filterbanks from raw speech for phone recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5509–5513, 2018.

(Modified Version)

[14] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve *et.al*, “Learning filterbanks from raw speech for phone recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 5509–5513, 2018.

(Previous Version)

[14] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, “Trainable frontend for robust and far-field keyword spotting,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5670–5674, 2017. MFCC

(Modified Version)

[15] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon and R. A. Saurous, “Trainable frontend for robust and far-field keyword spotting,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 5670–5674, 2017.

(Previous Version)

[15] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, “Leaf: A learnable frontend for audio classification,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=jM76BCb6F9m

(Modified Version)

[16] N. Zeghidour, O. Teboul, F. de Chaumont Quitry and M. Tagliasacchi, “Leaf: A learnable frontend for audio classification,” in *International Conference on Learning Representations*, Vienna, Austria, 2021.

(Previous Version)

[16] S. Liu et al., “Audio self-supervised learning: A survey,” 2022. [Online]. Available: https://arxiv.org/abs/2203.01205

(Modified Version)

[17] S. Liu, A.M. Ragolta, E. P. Cabaleiro, K. Qian and Xin Jing *et.al.*, “Audio self-supervised learning: A survey,” *Patterns*, vol. 3, no.12, pp. 100616, 2022.

(Previous Version)

[18] T. Mikolov, K. Chen, G.s Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in *Proceedings of Workshop at ICLR*, 2013.

(Modified Version)

[19] T. Mikolov, K. Chen, G.s Corrado and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in *Proceedings of Workshop at ICLR*, Scottsdale, Arizona, USA, 2013.

(Previous Version)

[19] A. T. Liu, S. Yang, P. Chi, P. Hsu, and H. Lee, “Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6419–6423, 2020.

(Modified Version)

[20] A. T. Liu, S. Yang, P. Chi, P. Hsu and H. Lee, “Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 6419–6423, 2020.

(Previous Version)

[20] D. Niizumi et al., “Byol for audio: Self-supervised learning for general-purpose audio representation,” in *2021 International Joint Conference on Neural Networks*, pp. 1–8, 2021.

(Modified Version)

[21] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada and K.Kashino, “Byol for audio: Self-supervised learning for general-purpose audio representation,” in *2021 International Joint Conference on Neural Networks*, Shenzhen, China, pp. 1–8, 2021.

(Previous Version)

[21] W. Hsu et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

(Modified Version)

[22] W. Hsu, B. Bolte, Y.H. H. Tsai, K. Lakhotia, R. Salakhutdinov *et.al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

(Previous Version)

[22] D. Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *ArXiv*, 2019. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf

(Modified Version)

[23] D. Jacob, M.W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *ArXiv*, 2019. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf

(Previous Version)

[23] H. Al-Tahan and Y. Mohsenzadeh, “Clar: Contrastive learning of auditory representations,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130, pp. 2530–2538, 2021. [Online]. Available: https://proceedings.mlr. press/v130/al-tahan21a.html

(Modified Version)

[24] H. Al-Tahan and Y. Mohsenzadeh, “Clar: Contrastive learning of auditory representations,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130, Virtual, pp. 2530–2538, 2021.

(Previous Version)

[24] S. Schneider, A. Baevski, r. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition.” in *Proceeding of Interspeech*, pp. 3465–3469, 2019.

(Modified Version)

[25] S. Schneider, A. Baevski, R. Collobert and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition.” in *Proc. of Interspeech*, Graz, Austria, pp. 3465–3469, 2019.

(Previous Version)

[25] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in *International Conference on Learning Representations*, 2020.

(Modified Version)

[26] A. Baevski, S. Schneider and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in *International Conference on Learning Representations*, Virtual, 2020.

(Previous Version)

[26] S. Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech*, pp. 1194-1198, 2021.

(Modified Version)

[27] S. Yang, P.H. Chi, Y.S. Chuang, C.I. J. Lai, K. Lakhotia *et.al.*, “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. of Interspeech*, Brno, Czechia, pp. 1194-1198, 2021.

(Previous Version)

[27] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” in *arXiv preprint arXiv:2111.02735, 2021.* [Online]. Available: https://arxiv.org/abs/2111.02735

(Modified Version)

[28] Y. Wang, A. Boumadane and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," in *arXiv preprint arXiv:2111.02735*, 2021. [Online]. Available: https://arxiv.org/abs/2111.02735

(Previous Version)

[28] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, pp. 3400–3404, 2021.

(Modified Version)

[29] L. Pepino, P. Riera and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. of Interspeech*, Brno, Czechia, pp. 3400–3404, 2021.

(Previous Version)

[29] L. Weiyang et al. "Large-margin softmax loss for convolutional neural networks," in *International Conference on Machine Learning*, pp. 507-516, 2016.

(Modified Version)

[30] L. Weiyang, Y. Wen, Z. Yu and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *International Conference on Machine Learning*, vol. 48, New York, New York, USA, pp. 507-516, 2016.

(Previous Version)

[30] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1735–1742, 2006.

(Modified Version)

[31] R. Hadsell, S. Chopra and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, New York, NY, USA, pp. 1735–1742, 2006.

(Previous Version)

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering,"

in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

(Modified Version)

[32] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 815–823, 2015.

(Previous Version)

[32] W. Yandong, et al., "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 499-515, 2016.

(Modified Version)

[33] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14,* Amsterdam, The Netherlands, pp. 499-515, 2016.

(Previous Version)

[33] H. Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

(Modified Version)

[34] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong *et.al.*, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 5265–5274, 2018.

(Previous Version)

[34] Y. Tang, Y. Hu, L. He, and H. Huang, "A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition," in *Speech Communication*, vol. 143 pp. 21–32, 2022.

(Modified Version)

[35] Y. Tang, Y. Hu, L. He and H. Huang, "A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition," in *Speech Communication*, vol. 143, pp. 21–32, 2022.

(Previous Version)

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

(Modified Version)

[37] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp. 5206–5210, 2015.

(Previous Version)

[36] M. Hou, Z. Zhang, Q. Cao, D. Zhang, and G. Lu, “Multi-view speech emotion recognition via collective relation construction,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30 pp. 218–229, 2022.

(Modified Version)

[36] M. Hou, Z. Zhang, Q. Cao, D. Zhang and G. Lu, “Multi-view speech emotion recognition via collective relation construction,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 218–229, 2022.

(Previous Version)

[37] M. Ott et al., "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, June 2019, pp. 48–53, 2019.

(Modified Version)

[38] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross *et.al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, pp. 48–53, 2019.

(Previous Version)

[38] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2015.

(Modified Version)

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

(Previous Version)

[39] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7697–7701, 2022.

(Modified Version)

[40] N. S. Clow, M. Kegler, P. Beckmann and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, pp. 7697–7701, 2022.

6. Reviewer's comment:

ALL Abbreviations should be defined in full the first time they appear in the title, abstract, main text, and figure or table captions, even if they are well known in the field. The first time you use an abbreviation in the text, present both the spelled-out version and the short form.

The syntax is: The fully spelled out name (abbreviation)

For example: MFCC, STFT, WA, UA, LSTM, CNN

Revisions made:

In response to the reviewer's comments, we have defined all abbreviations in the full form on their first appearance in the manuscript.

(Previous Version)

Global-Aware Multi-scale (GLAM)[12] used MFCC inputs and a global-aware fusion module to learn a multi-scale feature representation, which is rich in emotional information.

Subsection 2.1, Sentence 5

(Modified Version)

Global-Aware Multi-scale (GLAM) [12] used Mel-frequency cepstral coefficient (MFCC) inputs and a global-aware fusion module to learn a multi-scale feature representation, which is rich in emotional information.

(Previous Version)

The feature encoder was implemented using seven 1-d convolution layers with different kernel sizes and strides.

Subsection 2.5, Sentence 3

(Modified Version)

The feature encoder was implemented using seven one-dimensional (1-d) convolution layers with different kernel sizes and strides.

(Previous Version)

A Hanning window of the same size as the kernel and an STFT with a hop length equal to the stride was used.

Subsection 2.5, Paragraph 2, Sentence 4

(Modified Version)

A Hanning window of the same size as the kernel and a short-time Fourier transform (STFT) with a hop length equal to the stride were used.

(Previous Version)

Fig 5 shows a t-SNE plot of speaker-specific representations generated from the IEMOCAP dataset using two configurations of the speaker-identification network.

Subsection 5.1, Paragraph 2, Sentence 7

(Modified Version)

Fig 5 shows a t-distributed stochastic neighbor embedding (t-SNE) plot of speaker-specific representations generated from the IEMOCAP dataset using two configurations of the speaker-identification network.

(Previous Version)

[28] also included an LSTM layer to better model the temporal features.

Subsection 5.1, Paragraph 2, Sentence 7

(Modified Version)

[29] also included a long short-term memory (LSTM) layer to better model the temporal features.

(Previous Version)

The wav2vec 2.0 feature encoder (CNN layers) is frozen in all cases [27].

Subsection 5.2, Paragraph 1, Sentence 2

(Modified Version)

The wav2vec 2.0 feature encoder (convolutional layers) is frozen in all cases [28].

(Previous Version)

The experimental results show that the proposed approach outperforms previous methods, with a weighted accuracy of 72.14% and unweighted accuracy of 72.97% on the Interactive Emotional Dynamic Motion Capture (IEMOCAP) dataset. The results of our experiments demonstrate the effectiveness of the proposed method and its potential to enhance human-machine interaction through more accurate emotion recognition in speech.

Abstract, Sentence 11

(Modified Version)

The proposed approach outperforms previous methods, with a weighted accuracy (WA) of 72.14%

and unweighted accuracy (UA) of 72.97% on the Interactive Emotional Dynamic Motion Capture (IEMOCAP) dataset. This demonstrates its effectiveness and its potential to enhance human-machine interaction through more accurate emotion recognition in speech.

The reviewer highlighted that the full forms of UA and WA were not spelled out but these had already been spelled out in section 4.3.

(Previous Version)

Weighted accuracy (WA) is an evaluation index that intuitively represents model prediction performance as the ratio of correct predictions to the overall number of predictions.

Subsection 4.3, Sentence 2

In order to mitigate the biases associated with the weighted accuracy in imbalanced datasets such as the IEMOCAP dataset, unweighted accuracy (UA), also called average re-call, is widely employed and can be computed using Eq. (8).

Subsection 4.3, Sentence 4

Summary of revisions (Reviewer #3)

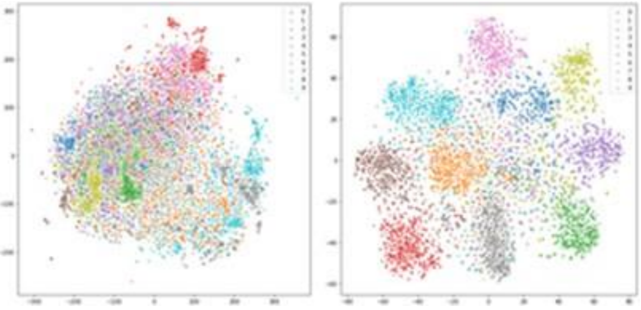
1. Reviewer's comment:

High resolution figures are suggested to add to the manuscript.

Revisions made:

In response to the reviewer's comment, we have replaced all figures in the previous version of the manuscript with higher-resolution images.

(Previous Version)

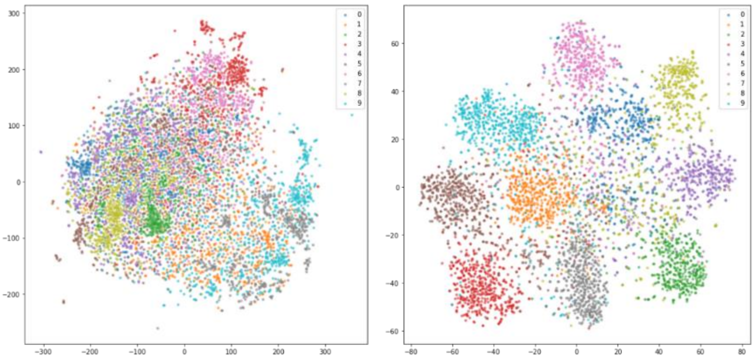


(a) (b)

Figure 5: t-SNE plot of speaker-specific representations generated by the speaker-identification network under different loss functions: (a) Cross-entropy (b) Arcface.

Subsection 5.1, Figure 5

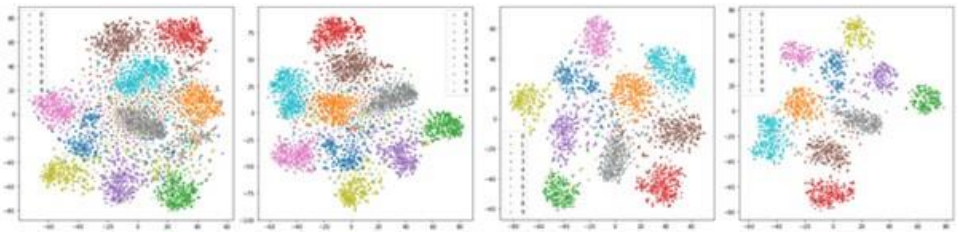
(Modified Version)



(a) (b)

Figure 5: t-SNE plot of speaker-specific representations generated by the speaker-identification network trained with different loss functions: (a) Cross-entropy (b) Arcface**.**

(Previous Version)

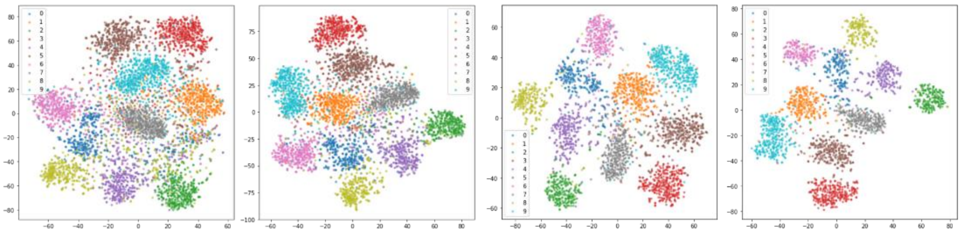


(a) (b) (c) (d)

Figure 6: t-SNE plot of speaker-specific representations generated by the speaker-identification network under various minimum duration configurations: (a) 1 second (b) 2 second (c) 3 second (4) 4 second.

Subsection 5.1, Figure 6

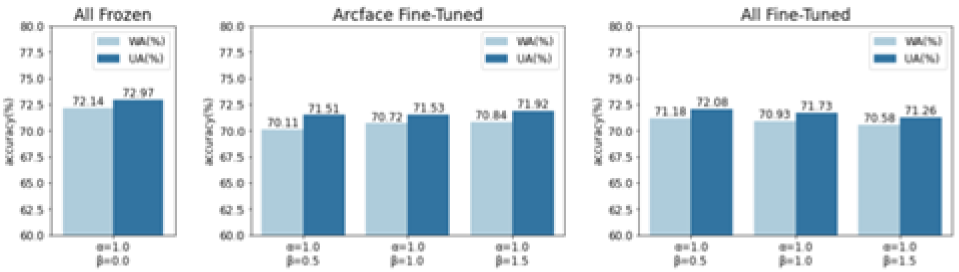
(Modified Version)



(a) (b) (c) (d)

Figure 6: t-SNE plot of speaker-specific representations generated by the speaker-identification network trained with audio segments of varying minimum lengths (a) 1 second (b) 2 seconds (c) 3 seconds (4) 4 seconds.

(Previous Version)

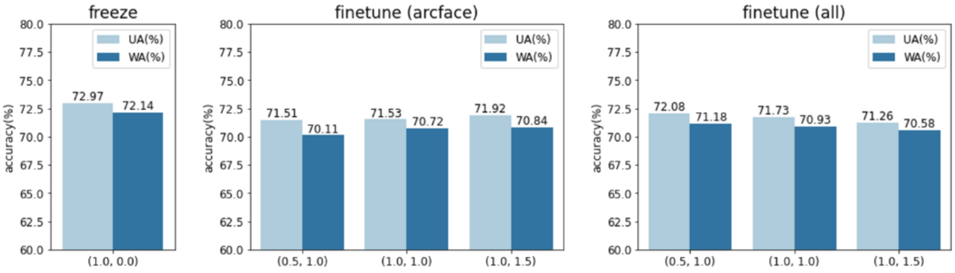


(a) (b) (c)

Figure 7: Performance of proposed method under various levels of fine-tuning the speaker-identification network: (a) All Frozen (b) Arcface Fine-tuned (c) All Fine-tuned.

Subsection 5.2, Figure 7

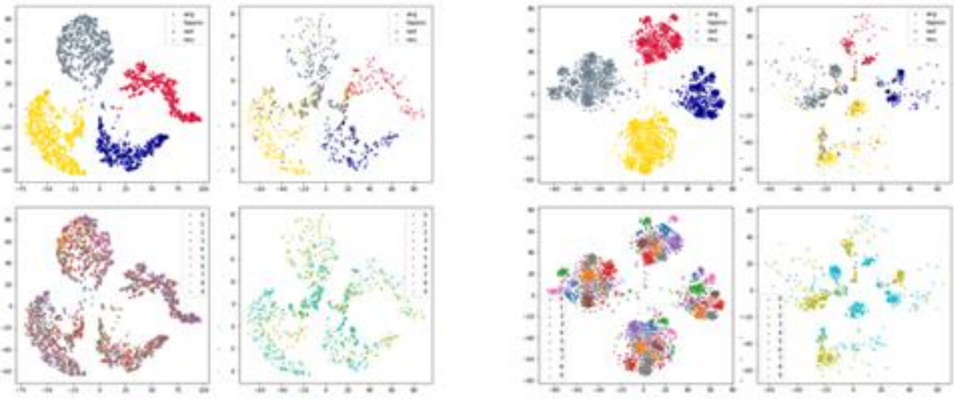
(Modified Version)



(a) (b) (c)

Figure 7: Performance of the proposed method with the speaker-identification network fine-tuned to various levels: (a) All Frozen (b) Arcface Fine-tuned (c) All Fine-tuned.

(Previous Version)

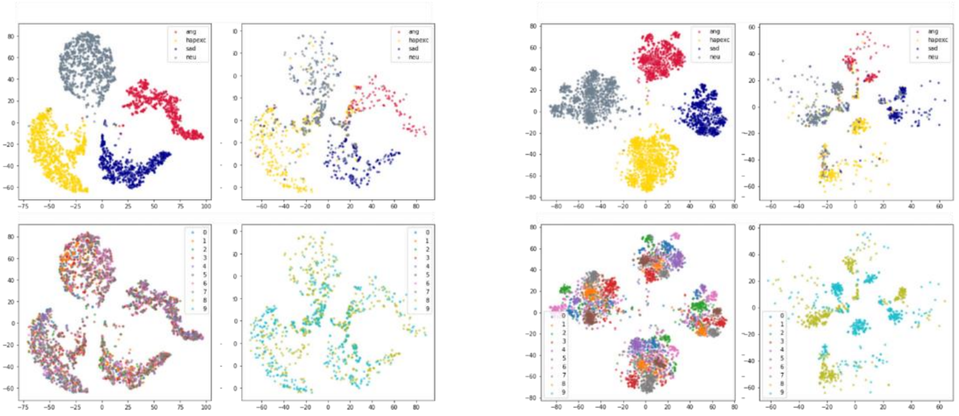


(a) (b)

Figure 8: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) without the speaker-specific representation (b) with the speaker-specific representation.

Subsection 5.2, Figure 8

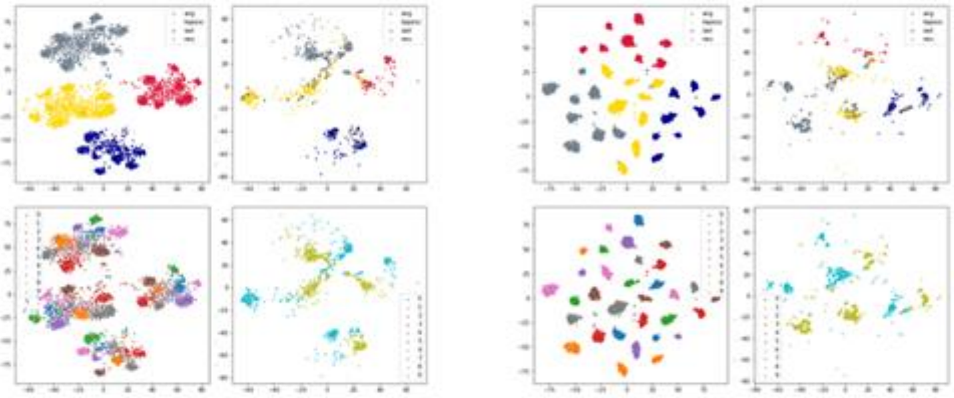
(Modified Version)



(a) (b)

Figure 8: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) without the speaker-specific representation (b) with the speaker-specific representation.

(Previous Version)

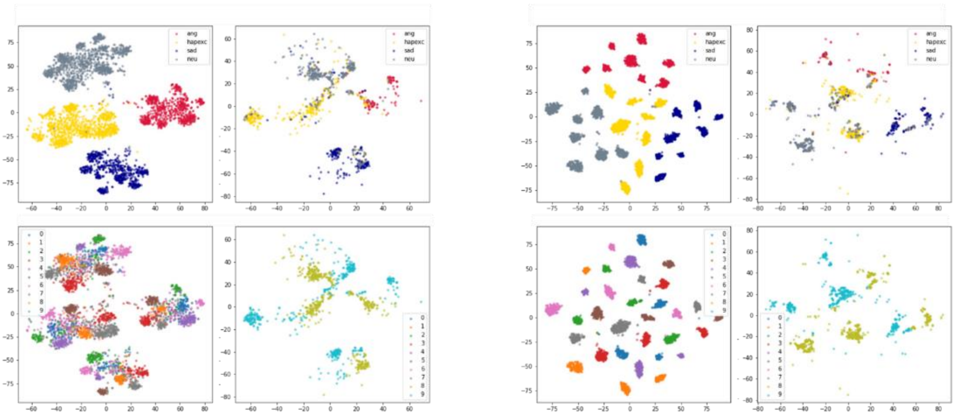


(a) (b)

Figure 9: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) Arcface fine-tuned (b) All fine-tuned.

Subsection 5.2, Figure 9

(Modified Version)



(a) (b)

Figure 9: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) Arcface fine-tuned (b) All fine-tuned.

2. Reviewer's comment:

It is recommended to investigate how the proposed method behaves when exposed to varying amounts of training data.

Revisions made:

In the previous version of the manuscript, emotion classes with more samples were under-sampled because the IEMOCAP dataset is unevenly distributed across emotion classes. In response to the reviewer's comment, we have conducted additional experiments with varying amounts of training data. More specifically, we have trained our model on the IEMOCAP dataset with and without under-sampling.

In the experimental results presented in the original version of the manuscript, freezing the weights of the speaker-identification network preserved the representation learned from the 1,251 speakers of the VoxCeleb1 dataset, resulting in better emotion classification performance. Therefore, we first pretrained the emotion classification network with both under-sampled and complete versions of the IEMOCAP dataset. In training the speaker-specific representation emotion network with both under-sampled and complete versions of the IEMOCAP dataset, we fine-tuned the pretrained emotion classification network and combined the emotion representation with the output of the speaker-identification network to generate a speaker-specific emotion *representation. Table B shows the results of experiments conducted under four configurations. In the experiment results, both pretrained and fine-tuned models showed the best performance when trained the dataset was under-sampled. This is because under-sampling adequately addresses the imbalance problem in the dataset.*

Table A: Performance of speaker-specific emotion representation network trained under four configurations (with under-sampled and complete dataset).

		Pretraining			
		Under-Sampled Dataset		Complete Dataset	
Accuracy (%)		WA (%)	UA (%)	WA (%)	UA (%)
Fine-Tuning	Under-Sampled Dataset	**72.14**	**72.79**	70.86	71.84
	Complete Dataset	70.92	71.96	71.68	72.26

3. Additional Revision:

In the process of revising the manuscript, we discovered that Eq. (3) was identical to Eq. (2), which was an error. Eq. (3) was originally intended to describe the terms of the Arcface loss defined in Eq. (2) in an extended form in which the dot product is expressed as the product of the norms of vectors and the cosine of the angle between them. Therefore, we have modified Eq. (3) to reflect the original intention.

(Previous Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 2

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 3

(Modified Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^{T}x_i}}{\sum_{j=1}^{n} e^{W_j^{T}x_i}}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\,\frac{e^{\left\|W_{y_i}\right\|\|x_i\| cos\theta_{y_i}}}{\sum_{j=1}^{n} e^{\left\|W_j\right\|\|x_i\| cos\theta_j}},$$

Summary of revisions (Reviewer #4)

1. Reviewer's comment:

A revision on the text will be more acceptable.

Revisions made:

In response to the reviewer's comments, we have extensively revised the manuscript.

(Previous Version)

Speech emotion recognition (SER) is, therefore, one of the most active research areas in the computer science field ....

Section 1, Paragraph 1, Sentence 5

(Modified Version)

Speech emotion recognition (SER) is one of the most active research areas in the computer science field ....

(Previous Version)

These hand-crafted features have proven their potential in previous works. However, features and their representations should be tailored and optimized for specific tasks. Deep learning-based representations generated from actual waveforms or LLDs have shown better performance in SER.

Section 1, Paragraph 2, Sentence 2-3

(Modified Version)

Although the potential of hand-crafted features has been demonstrated in previous works, features and their representations should be tailored and optimized for specific tasks. Deep learning-based representations generated from LLDs or the actual waveform have shown better performance in SER.

(Previous Version)

Studies in psychology have shown that individuals have different voice characteristics depending on their culture, language, gender, and personality [3]. This implies that two speakers saying the same sentence with the same emotion are likely to express different acoustic properties in their voices. Several studies have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks. This implies that the acoustic properties expressed in the voices of two speakers saying the same sentence with the same emotion may vary. Several studies [4-5] have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks.

Section 1, Paragraph 3

(Modified Version)

Studies in psychology have shown that individuals have different vocal attributes depending on their culture, language, gender, and personality [3]. This implies that two speakers saying the same thing with the same emotion are likely to express different acoustic properties in their voices. Several studies [4-5] have demonstrated the merit of considering speaker-specific properties in audio speech-related tasks.

(Previous Version)

The proposed model consists of an identification encoder and an emotion classifier. The wav2vec 2.0 [6] (base model) is used as a backbone for both of the proposed networks, and it generates emotion and speaker features from input audio waveform.

Section 1, Paragraph 4, Sentence 3

(Modified Version)

The proposed model consists of a speaker-identification network and an emotion classifier. The wav2vec 2.0 [6] (base model) is used as a backbone for both of the proposed networks, where it is used to extract emotion and speaker-specific features from input audio waveforms.

(Previous Version)

This tensor fusion is performed using an element-wise multiplication followed by a summation of vectors. The main contributions of this paper are summarized below:

Section 1, Paragraph 4, Sentence 6-7

(Modified Version)

This tensor fusion operation is performed using an element-wise multiplication followed by a summation of vectors. The main contributions of this paper are summarized as follows:

(Previous Version)

Two modules (emotion and identification) based on wav2vec 2.0 to generate a speaker-specific emotion representation from an audio input are proposed.

Section 1, Paragraph 5, Sentence 1

(Modified Version)

Two modules (speaker-identification network and emotion classification) based on wav2vec 2.0 that generate a speaker-specific emotion representation from an input audio segment are proposed.

(Previous Version)

The speaker-identification network was therefore pre-trained on a separate dataset (VoxCeleb1 [8]) with 1,251 speakers to support better generalization to unseen speakers.

Section 1, Paragraph 5, Sentence 4

(Modified Version)

The representations generated by the speaker-identification network which was pretrained on the VoxCeleb1 dataset [8] facilitate better generalization to unseen speakers.

(Previous Version)

The use of Arcface [9] and cross-entropy loss terms in the speaker-identification network was also explored and detailed evaluations provided.

Section 1, Paragraph 6, Sentence 2

(Modified Version)

The use of the Arcface [9] and cross-entropy loss terms in the speaker-identification network was also explored and detailed evaluations have been provided.

(Previous Version)

In more recent approaches, models learned a representation directly from raw waveforms instead of using hand-crafted representations like the human perception emulating Mel-filter banks used in generating the Mel-spectrogram. Time-Domain (TD) filter banks [13] used complex convolutional weights initialized with Gabor wavelets to learn filter banks from raw speech for end-to-end phone recognition. The proposed architecture has a convolutional layer followed by an L2 feature pooling-based modulus operation and a low-pass filter.

Subsection 2.2, Paragraph 1, Sentence 1-2

(Modified Version)

In more recent approaches, models learn a representation directly from raw waveforms instead of using hand-crafted representations like the human perception emulating Mel-filter banks used in generating the Mel-spectrogram. Time-Domain (TD) filter banks [14] use complex convolutional weights initialized with Gabor wavelets to learn filter banks from raw speech for end-to-end phone recognition. The proposed architecture has a convolutional layer followed by an l_2 feature pooling-based modulus operation and a low-pass filter.

(Previous Version)

However, log-scale compression and normalization reduce the scale of spectrograms, regardless of their contents.

Subsection 2.2, Paragraph 1, Sentence 7

(Modified Version)

A key limitation of this approach is that the log-scale compression and normalization used reduce the scale of spectrograms, regardless of their contents.

(Previous Version)

Reference [15] also learned a drop-in replacement for Mel-filter banks but replaced the static log compression with dynamic compression and addressed the channel distortion problems in the Mel-spectrogram log transformation using Per-Channel Energy Normalization (PCEN). This was calculated using a smoothed version of the filter bank energy, which is computed from a first-order infinite impulse response (IIR) filter. The smoothing coefficient was used to combine the smoothed version of the filter bank energy, and the current spectrogram energy was fixed in PCEN.

Subsection 2.2, Paragraph 2, Sentence 1~3

(Modified Version)

Wang et.al. [15] also propose a learned drop-in alternative to the Mel-filter banks but replaced static log compression with dynamic compression and addressed the channel distortion problems in the Mel-spectrogram log transformation using Per-Channel Energy Normalization (PCEN). This was calculated using a smoothed version of the filter bank energy, which was computed from a first-order infinite impulse response (IIR) filter. A smoothing coefficient was used to combine the smoothed version of the filter bank energy, and the current spectrogram energy.

(Previous Version)

Continuous Bags of Words (CBoW) and skip-gram variants were implemented and evaluated.

Subsection 2.3, Paragraph 2, Sentence 2

(Modified Version)

Continuous Bags of Words (CBoW) and skip-gram variants were also implemented and evaluated.

(Previous Version)

In this architecture, two neural networks were trained by maximizing the agreement of their output given the same input.

Subsection 2.3, Paragraph 2, Sentence 5

(Modified Version)

In this architecture, two neural networks were trained by maximizing the agreement in their outputs given the same input.

(Previous Version)

[4] uses an aggregation of an individual’s neutral speech to standardize emotional speech and improve the robustness of individual-agnostic emotion representation.

Subsection 2.4, Paragraph 1, Sentence 2

(Modified Version)

[4] uses an aggregation of individuals’ neutral speech to standardize emotional speech and improve the robustness of individual-agnostic emotion representations.

(Previous Version)

A Hanning window of the same size as the kernel and an STFT with a hop length equal to the stride was used.

Subsection 2.5, Paragraph 2, Sentence 4

(Modified Version)

A Hanning window of the same size as the kernel and a short-time Fourier transform (STFT) with a hop length equal to the stride were used.

(Previous Version)

The contextual encoder consists of a linear projection layer, a relative positional encoding 1-d convolution layer followed by a GeLU, and Transformer.

Subsection 2.5, Paragraph 2, Sentence 6

(Modified Version)

The contextual encoder consists of a linear projection layer, a relative positional encoding 1-d convolution layer followed by a GeLU and transformer model.

(Previous Version)

This amounts to choosing quantized representations (codebook entries) from multiple codebooks using a Gumble softmax, ...

Subsection 2.5, Paragraph 2, Sentence 11

(Modified Version)

This is achieved by choosing quantized representations (codebook entries) from multiple codebooks using a Gumble softmax,

(Previous Version)

In Arcface, the representations were distributed around feature centers in a hypersphere with a radius.

Subsection 2.6, Paragraph 1, Sentence 4

(Modified Version)

In Arcface, the representations were distributed around feature centers in a hypersphere with a fixed radius.

(Previous Version)

A margin is added to the angular difference of the features in the same class to make learned features separable with a larger angular distance.

Subsection 2.6, Paragraph 2, Sentence 7

(Modified Version)

A margin is added to the angular difference between features in the same class to make learned features separable with a larger angular distance.

(Previous Version)

3 Method

3.1 Network Architecture

The same waveform is passed to the speaker-identification as well as the emotion recognition networks. Both networks are extensions of the wav2vec 2.0 model (consisting of 12 Transformer layers with 768 dimension embedding) and are trained using the Arcface loss. The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. During pre-training, the Arcface loss used in this model enables it to learn to distinguish between speakers. In the emotion recognition network (Fig. 2), input utterances are encoded into a vector that is combined with the output of the pre-trained speaker-identification network.

In the speaker identification network, the wav2vec 2.0 segment is used to encode input utterances into a latent 2-d representation vector that is passed to a single attention block. It is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional status. This is why, only a single attention block was used in the speaker-identification network. The model is trained to generate a representation which distinctly distinguishes between speakers using the Arcface loss. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Here, the four FC outputs are fed into a softmax function, and the probability of each speaker class is obtained.

In the emotion classification network, the encoding generated by the wav2vec 2.0 segment is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. The outputs of all the attention blocks are concatenated prior to the tensor fusion operation.

In order to address the differences in the lengths of utterances, max-pooling was applied across the time axis of each attention block, resulting in a d dimension vector as the final output. The output vectors of the speaker-identification network and emotion classification network have identical dimensions. During tensor fusion, an elementwise multiplication between $H = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k \times d}$) is performed. Summation is then performed across all k vectors as shown in Eq. (5),

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \tag{5}$$

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is used to compute the Arcface loss for emotion classification. The fusion process combines the speaker-identification and emotion representations from each network into a speaker-specific emotion representation. To achieve this, the output vector of the attention block is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs. Fig. 3 shows the architecture of the proposed speaker-specific emotion representation-based SER. The angular distance between the tensor fused output vector and the center of the four emotion representation vectors is calculated as shown in Eq. (4).

Section 3

(Modified Version)

3 Methodology

In order to leverage speaker-specific speech characteristics to improve the performance of SER models, two wav2vec 2.0-based modules (speaker-identification network and emotion classification network) trained with the Arcface loss are proposed. The speaker-identification network extends the wav2vec 2.0 model with a single attention block that it uses to encode an input audio waveform into a speaker-specific representation. The emotion classification network uses a wav2vec 2.0-backbone as well as four attention blocks to encode the same input audio waveform into an emotion representation. These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information.

3.1 Speaker-Identification and Emotion Classification Networks

The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. The wav2vec 2.0 model is used to encode input utterances into a latent

2-d representation of shape $\mathbb{R}^{768\times T}$, where T is the length of the input waveform. This latent representation is passed to a single attention block prior to performing a max-pooling operation that results in a 1-d vector of length 768. Only a single attention block was used in the speaker-identification network because it is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional state. In order to achieve a more robust distinction between speakers, the $\mathbb{R}^{d}$ shape speaker-identification representation (H_{id}) and the $\mathbb{R}^{\#ID\times d}$ shape Arcface center representation vector (W_{id}) for speaker classes are l_2 normalized and their cosine similarity computed. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Then, the FC outputs were fed into a softmax function, and the probability of each speaker class obtained. In Figure. 1, "#ID" represents the index of each speaker class. For example, in the VoxCeleb1 dataset with 1,251 speakers, the final #ID is #1,251.

In the emotion classification network (Fig. 2), the wav2vec 2.0 model is used to encode input utterances into a $\mathbb{R}^{768\times T}$shape representation. The encoding generated is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. Max-pooling is applied across the time axis to the outputs of each of the attention blocks. The max pooled outputs of the attention blocks, h_i are concatenated prior to the tensor fusion operation. During tensor fusion, an elementwise multiplication between $H_{emo} = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. As shown in Eq. (5), all the k vectors are summed to generate the final embedding.$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i$,

(5)

where $e_i \in \mathbb{R}^{d}$ and $W_{fusion,i} \in \mathbb{R}^{d}$. The final embedding, E is l_2 normalized prior to computing the cosine similarity between Arcface center representation vectors $W_{emo} \in \mathbb{R}^{\#EMO\times d}$. In Figure. 2, "#EMO" represents the emotion class indices as defined in the IEMOCAP dataset. Here, 1_EMO, 2_EMO, 3_EMO, 4_EMO represent angry, happy, sad and neutral emotion classes respectively.

3.2 Speaker-Specific Emotion Representation Network

Fig. 3 shows the architecture of the proposed SER approach. The same waveform is passed to the speaker-identification network as well as the emotion recognition network. The speaker representation generated by the pretrained speaker-identification network is passed to the emotion classification network. More specifically, the output vector of the attention block from the speaker-identification network is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs, $H \in \mathbb{R}^{5 \times d}$. The fusion operation shown in Eq. (5) is used to combine these representations into a single speaker-specific emotion representation E. The angular distance between the normalized tensor fused output vector and the normalized center of the four emotion representation vectors is calculated using Eq. (4). The emotion class predicted for any input waveform, is determined by how close its representation vector is to an emotion class's center vector.

(Previous Version)

The IEMOCAP [7] is a multimodal, multi-speaker emotion database recorded across five sessions with five pairs of male and female speakers performing improvisations or scripted scenarios.

Subsection 4.1, Paragraph 1, Sentence 1

(Modified Version)

The IEMOCAP [7] is a multimodal, multi-speaker emotion database recorded across five sessions with five pairs of male and female speakers performing improvisations as well as scripted scenarios.

(Previous Version)

Due to imbalances in the samples available for each label category, only neutral, happy (with exciting), sad, and angry classes have been used in line with previous studies [4], [26–28], [34–35].

Subsection 4.1, Paragraph 1, Sentence 5

(Modified Version)

Due to imbalances in the number of samples available for each label category, only neutral, happy (combined with exciting), sad, and angry classes have been used in line with previous studies

[4], [27–29], [35–36].

(Previous Version)

Audio files longer than 15 seconds are truncated to 15 seconds because almost all of the audio samples available were less than 15 seconds long.

Subsection 4.1, Paragraph 1, Sentence 9

(Modified Version)

Audio files longer than 15 seconds are truncated to 15 seconds because almost all of the audio samples in the dataset were less than 15 seconds long.

(Previous Version)

As shown in Fig. 4, the dataset is unevenly distributed across emotion labels, with significantly more neutral and happy samples in most sessions.

Subsection 4.1, Paragraph 1, Sentence 12

(Previous Version)

VoxCeleb1 is an audio-visual dataset comprising 22,496 short interview clips extracted from YouTube.

Subsection 4.1, Paragraph 3, Sentence 3

(Modified Version)

VoxCeleb1 is an audio-visual dataset comprising 22,496 short interview clips extracted from YouTube videos.

(Modified Version)

As shown in Fig. 4, the dataset is unevenly distributed across emotion classes, with significantly more neutral and happy samples in most sessions.

(Previous Version)

In addition, [27] showed that either partially or entirely fine-tuning the wav2vec 2.0 segments results in the same boost in model performance on SER tasks in spite of the differences in computational costs.

Subsection 4.2, Paragraph 1, Sentence 3

(Modified Version)

In addition, [28] showed that either partially or entirely fine-tuning the wav2vec 2.0 segments results in the same boost in model performance on SER tasks despite the differences in computational costs.

(Previous Version)

The model and weights are provided by facebookresearch/fairseq [37].

Subsection 4.2, Paragraph 1, Sentence 5

(Modified Version)

The model and weights are provided by Facebook research under the Fairseq sequence modeling toolkit [38].

(Previous Version)

First, the identification network and emotion network are trained separately.

Subsection 4.2, Paragraph 2, Sentence 2

(Modified Version)

First, the speaker-identification network and emotion network are trained separately.

(Previous Version)

A $10-8$ weight decay is applied and Adam [38] optimizer with betas set to (0.9, 0.98) is used. LambdaLR scheduler reduces the learning rate by a factor of 0.98 after every epoch.

Subsection 4.2, Paragraph 2, Sentence 5-6

(Modified Version)

A $10-8$ weight decay is applied and the Adam [39] optimizer with beta values set to (0.9, 0.98) is used. The LambdaLR scheduler reduces the learning rate by a factor of 0.98 after every epoch.

(Previous Version)

As shown in Fig 6, the speaker identification network may be unable to generate accurate representations for very short audio samples. For example, representations of audio clips that are less than 3 seconds long may be misclassified.

Subsection 5.1, Paragraph 1, Sentence 5

(Modified Version)

As shown in Fig 6, the speaker identification network may be unable to generate accurate representations for audio samples that are too short. Representations of audio clips that are less than 3 seconds long are particularly likely to be misclassified.

(Previous Version)

Table 3 shows a comparison of our methods' performance against previous studies.

Subsection 5.1, Paragraph 1, Sentence 2

(Modified Version)

Table 3 shows a comparison of our methods' performance against that of previous studies.

(Previous Version)

The proposed speaker-identification network can be fine-tuned under three different configurations:

Subsection 5.2, Paragraph 1, Sentence 1

(Modified Version)

The proposed speaker-identification network was fine-tuned under three different configurations:

(Previous Version)

Model performance under these configurations is evaluated with four emotion attention blocks, because training the emotion network with four attention blocks showed the best performance in prior experiments.

Subsection 5.2, Paragraph 1, Sentence 5

(Modified Version)

Since training the emotion classification network with four attention blocks showed the best performance in prior experiments, fine-tuning performance was evaluated under this configuration.

(Previous Version)

Due to the small number of speakers in the IEMOCAP dataset, the model quickly converged on a representation that can distinguish speakers but was unable to generalize to unseen speakers.

Subsection 5.2, Paragraph 1, Sentence 7

(Modified Version)

Due to the small number of speakers in the IEMOCAP dataset, the model quickly converged on a representation that could distinguish speakers but was unable to generalize to unseen speakers.

(Previous Version)

In Fig 7 (b), increasing β, which controls the significance of the identification loss, improves emotion classification accuracy. On the contrary, in Fig 7 (c), increasing causes the emotion classification accuracy to deteriorate.

Subsection 5.2, Paragraph 12 Sentence 1-2

(Modified Version)

Fig 7 (b) shows that increasing β, which controls the significance of the identification loss, improves emotion classification accuracy when the Arcface center representation vectors are frozen. Conversely, Fig 7 (c) shows that increasing β, causes the emotion classification accuracy to deteriorate when the entire model is fine-tuned.

(Previous Version)

In the top row of Fig 8 (a) and (b), the representations are colored according to the emotion class, and in the bottom row the representations are colored according to speakers. The same layout is applied to Fig 9 (a) and (b).

Subsection 5.2, Paragraph 3, Sentence 3~4

(Modified Version)

In the top row of Fig 8 (a) and (b), a representation's color is determined by its predicted emotion class and in the bottom row, a representation's color is determined by its predicted speaker class. The same descriptors are applicable to Fig 9 (a) and (b).

(Previous Version)

In comparison to Fig 9 (a), Fig 9 (b) shows that fine-tuning both the speaker-identification network and the emotion classification network increases inter-class separability between the emotion representations of speakers, while retaining a speaker-specific information.

Subsection 5.2, Paragraph 4, Sentence 1

(Modified Version)

In contrast to Fig 9 (a), Fig 9 (b) shows that fine-tuning both the speaker-identification network and the emotion classification network increases inter-class separability between the emotion representations of speakers, while retaining speaker-specific information.

(Previous Version)

This results in a slight improvement in overall SER performance, which is in line with the findings presented in Fig 7 (b) and (c).

Subsection 5.2, Paragraph 4, Sentence 2

(Modified Version)

This results in a slight improvement in the overall SER performance, which is in line with the findings presented in Fig 7 (b) and (c).

(Previous Version)

5.3 Comparison of Previous Methods

Subsection 5.3, Heading

(Modified Version)

5.3 Comparison Against Previous Methods

(Previous Version)

In Table 3, we have compared the proposed method against the previous SER methods methods based on the wav2vec 2.0 or the Arcface loss.

Subsection 5.3, Paragraph 1, Sentence 1

(Modified Version)

In Table 3, we have compared the proposed method against previous SER methods based on the wav2vec 2.0 or the Arcface loss.

(Previous Version)

The experiment showed that the configuration using four attention blocks in the emotion network and fine-tuning with the speaker-identification network frozen (Fig 7 (a)) provided the best performance.

Subsection 5.3, Paragraph 1, Sentence 4

(Modified Version)

Experiments showed that the configuration using four attention blocks in the emotion network and fine-tuning with the speaker-identification network frozen (Fig 7 (a)) provided the best performance.

(Previous Version)

This paper proposed two modules for generating a speaker-specific emotion representation for SER.

Subsection 6, Paragraph 1, Sentence 1

(Modified Version)

In this study, two modules for generating a speaker-specific emotion representation for SER are proposed.

2. <u>Reviewer's comment</u>:

A computation time comparison can be useful for more justifications.

<u>Revisions made</u>:

In response to the reviewer's comment, we have conducted experiments to compare the proposed model's computation times under various configurations: the length of input audio segments and the number of attention blocks. More specifically, input audio segments of 3, 5, 10, and 15 seconds as well as model configuration of 1, 2, 3, and 4 attention blocks were used. The proposed model (speaker-specific emotion representation network) consists of two networks (speaker-identification and emotion classification). Table A shows the separate and combined computation times of the *two networks under the above-mentioned configurations. As shown in Table A, computation time increases as the length of input audio segments and the number of attention blocks increases. Experimental results showed that the speaker-specific emotion representation network with four attention blocks can process audio segments in close to 27ms.*

Table A: Computation time (ms) of proposed networks (speaker-identification, emotion classification, and speaker-specific emotion representation networks) for input audio segments of varying lengths (3, 5, 10, and 15 seconds).

			Input audio segment length (seconds)			
			3	5	10	15
Speaker-identification network			2.52	4.20	8.32	13.08
Emotion classification network	# of attention blocks	1	2.49	4.16	8.24	12.96
		2	2.57	4.28	8.52	13.41
		3	2.64	4.40	8.78	13.84
		4	2.71	4.52	9.00	14.22
Speaker-specific emotion representation network (four attention blocks)			5.28	8.79	17.39	27.26

(Modified Version)

The computation time of the proposed method under various configurations was examined. The length of input audio segments (3, 5, 10, and 15 seconds) and number of attention blocks (1, 2, 3, and 4) were varied. The proposed model (speaker-specific emotion representation network) consists of two networks (speaker-identification and emotion classification). Table 6 shows the separate and combined computation times of the two networks under the above-mentioned configurations. As shown in Table 7, computation time increases as the length of input audio segments and the number of attention blocks increases. Experiments show that the best-performing configuration of the proposed model in which the speaker-specific emotion representation network has four attention blocks can process audio segments in close to 27ms.

Subsection 5.4, Paragraph 3

3. <u>Reviewer’s comment</u>:

What are the effects of the non-balance dataset in your method?

<u>Revisions made</u>:

In the previous version of the manuscript, emotion classes with more samples were under-sampled because the IEMOCAP dataset is unevenly distributed across emotion classes. In response to the reviewer's comment, we have conducted additional experiments with to examine the effects of an imbalanced dataset. More specifically we have trained our model on IEMOCAP dataset with and without under-sampling.

In the experimental results presented in the original version of the manuscript, freezing the weights of the speaker-identification network preserved the representation learned from the 1,251 speakers of the VoxCeleb1 dataset, resulting in better emotion classification performance. Therefore, we first pretrained the emotion classification network with both under-sampled and complete versions of the IEMOCAP dataset. In training the speaker-specific representation emotion network with both under-sampled and complete versions of the IEMOCAP dataset, we fine-tuned the pretrained emotion classification network and combined the emotion representation with the output of the speaker-identification network to generate a speaker-specific emotion representation. Table B shows the results of experiments conducted under four configurations. In the experiment results, both pretrained and fine-tuned with under-sampled dataset showed the best performance. This is because under-sampling adequately addresses the imbalance problem in the dataset.

Table B: Performance (accuracy) of speaker-specific emotion representation network trained under four configurations (with under-sampled and complete versions of the IEMOCAP dataset).

Configurations		Pretraining			
		Under-Sampled Dataset		Complete Dataset	
Accuracy (%)		WA (%)	UA (%)	WA (%)	UA (%)
Fine-Tuning	Under-Sampled Dataset	**72.14**	**72.79**	70.86	71.84
	Complete Dataset	70.92	71.96	71.68	72.26

(Modified Version)

Since the audio segments in the IEMOCAP are unevenly distributed across emotion classes, emotion classes with more samples were under-sampled. To examine the effects of an imbalanced dataset, additional experiments with varying amounts of training data were conducted using the best-performing configuration of the proposed model (speaker-specific emotion representation network with four attention blocks with frozen speaker-identification network). More specifically, the model was trained on the entire dataset with and without under-sampling to examine the effects of an imbalanced dataset. Table 5 shows the results of experiments conducted under four configurations. In the experiment results, both pre-trained and fine-tuned variations of the model showed their best performance when trained using the undersampled version of the IEMOCAP dataset. This is because under-sampling adequately addresses the imbalance problem in the dataset.

Subsection 5.4, Paragraph 1

4. Additional Revision:

In the process of revising the manuscript, we discovered that Eq. (3) was identical to Eq. (2), which was an error. Eq. (3) was originally intended to describe the terms of the Arcface loss defined in Eq. (2) in an extended form in which the dot product is expressed as the product of the norms of vectors and the cosine of the angle between them. Therefore, we have modified Eq. (3) to reflect the original intention.

(Previous Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 2

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 3

(Modified Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\left\| W_{y_i} \right\| \|x_i\| cos\theta_{y_i}}}{\sum_{j=1}^{n} e^{\left\| W_j \right\| \|x_i\| cos\theta_j}},$$

Summary of revisions (Reviewer #5)

1. Reviewer's comment:

In the Section 3, the proposed method should be discussed in more detail. Why only the Section 3.1?

Revisions made:

In the previous manuscript version, there was only one subsection under section 3. In order to make the method description clearer, we reorganized the content in section 3 into sections 3.1 and 3.2 and provided additional details in each subsection.

(Previous Version)

3 Method

3.1 Network Architecture

The same waveform is passed to the speaker-identification as well as the emotion recognition networks. Both networks are extensions of the wav2vec 2.0 model (consisting of 12 Transformer layers with 768 dimension embedding) and are trained using the Arcface loss. The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. During pre-training, the Arcface loss used in this model enables it to learn to distinguish between speakers. In the emotion recognition network (Fig. 2), input utterances are encoded into a vector that is combined with the output of the pre-trained speaker-identification network.

In the speaker identification network, the wav2vec 2.0 segment is used to encode input utterances into a latent 2-d representation vector that is passed to a single attention block. It is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional status. This is why, only a single attention block was used in the speaker-identification network. The model is trained to generate a representation which distinctly distinguishes between speakers using the Arcface loss. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Here, the four FC outputs are fed into a softmax function, and the probability of each speaker class is obtained.

In the emotion classification network, the encoding generated by the wav2vec 2.0 segment is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. The outputs of all the attention blocks are concatenated prior to the tensor fusion operation.

In order to address the differences in the lengths of utterances, max-pooling was applied across the time axis of each attention block, resulting in a d dimension vector as the final output. The output vectors of the speaker-identification network and emotion classification network have identical dimensions. During tensor fusion, an elementwise multiplication between $H = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k \times d}$) is performed. Summation is then performed across all k vectors as shown in Eq. (5),

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \qquad (5)$$

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is used to compute the Arcface loss for emotion classification. The fusion process combines the speaker-identification and emotion representations from each network into a speaker-specific emotion representation. To achieve this, the output vector of the attention block is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs. Fig. 3 shows the architecture of the proposed speaker-specific emotion representation-based SER. The angular distance between the tensor fused output vector and the center of the four emotion representation vectors is calculated as shown in Eq. (4).

Section 3

(Modified Version)

3 Methodology

In order to leverage speaker-specific speech characteristics to improve the performance of SER models, two wav2vec 2.0-based modules (speaker-identification network and emotion classification network) trained with the Arcface loss are proposed. The speaker-identification network extends the wav2vec 2.0 model with a single attention block that it uses to encode an input audio waveform into a speaker-specific representation. The emotion classification network uses a wav2vec 2.0-backbone as well as four attention blocks to encode the same input audio waveform into an emotion representation. These two representations are then fused into a single vector representation that contains both emotion and speaker-specific information.

3.1 Speaker-Identification and Emotion Classification Networks

The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. The wav2vec 2.0 model is used to encode input utterances into a latent 2-d representation of shape $\mathbb{R}^{768\times T}$, where T is the length of the input waveform. This latent representation is passed to a single attention block prior to performing a max-pooling operation that results in a 1-d vector of length 768. Only a single attention block was used in the speaker-identification network because it is assumed that the core properties of a speaker's voice are unaffected by his or her emotional state. In other words, a speaker can be identified by his/her voice regardless of his/her emotional state. In order to achieve a more robust distinction between speakers, the $\mathbb{R}^{d}$ shape speaker-identification representation (H_{id}) and the $\mathbb{R}^{\#ID\times d}$ shape Arcface center representation vector (W_{id}) for speaker classes are l_2 normalized and their cosine similarity computed. A configuration of the speaker-identification network using the cross-entropy loss was also explored. In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Then, the FC outputs were fed into a softmax function, and the probability of each speaker class obtained. In Figure. 1, “#ID” represents the index of each speaker class. For example, in the VoxCeleb1 dataset with 1,251 speakers, the final #ID is #1,251.

In the emotion classification network (Fig. 2), the wav2vec 2.0 model is used to encode input utterances into a $\mathbb{R}^{768\times T}$shape representation. The encoding generated is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. The four attention blocks are used to identify which parts of the generated emotion representation are most relevant to SER. Experiments were also conducted for configurations with one, two, as well as three attention blocks. Max-pooling is applied across the time axis to the outputs of each of the attention blocks. The max pooled outputs of the attention blocks, h_i are concatenated prior to the tensor fusion operation. During tensor fusion, an elementwise multiplication between $H_{emo} = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. As shown in Eq. (5), all the k vectors are summed to generate the final embedding.$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i$,

(5)

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is l_2 normalized prior to computing the cosine similarity between Arcface center representation vectors $W_{emo} \in \mathbb{R}^{\#EMO \times d}$. In Figure. 2, “#EMO” represents the emotion class indices as defined in the IEMOCAP dataset. Here, 1_EMO, 2_EMO, 3_EMO, 4_EMO represent angry, happy, sad and neutral emotion classes respectively.

3.2 Speaker-Specific Emotion Representation Network

Fig. 3 shows the architecture of the proposed SER approach. The same waveform is passed to the speaker-identification network as well as the emotion recognition network. The speaker representation generated by the pretrained speaker-identification network is passed to the emotion classification network. More specifically, the output vector of the attention block from the speaker-identification network is concatenated to the outputs of the emotion classification network’s four attention blocks, resulting in a total of five attention block outputs, $\mathrm{H} \in \mathbb{R}^{5\times\mathrm{d}}$. The fusion operation shown in Eq. (5) is used to combine these representations into a single speaker-specific emotion representation E. The angular distance between the normalized tensor fused output vector and the normalized center of the four emotion representation vectors is calculated using Eq. (4). The emotion class predicted for any input waveform, is determined by how close its representation vector is to an emotion class's center vector.

2. <u>Reviewer’s comment</u>:

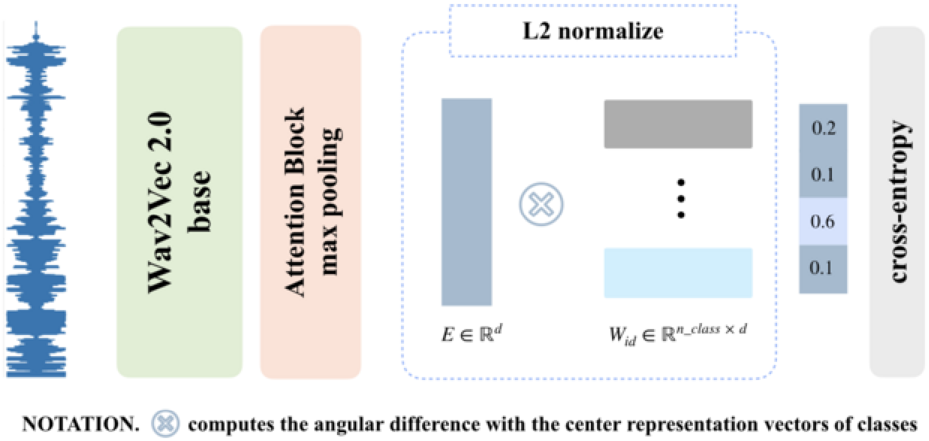
The Figures in this manuscript are too vague. Moreover, more description should be added to Figure. 1-3.

<u>Revisions made</u>:

In our response to the reviewer’s first comment, we have already updated descriptions of the proposed model in Section 3. In response to this comment, we have also modified Figure. 1~3 in Section 3 and their corresponding captions.

(Previous Version)

In the speaker identification network, the wav2vec 2.0 segment is used to encode input utterances into a latent 2-d representation vector that is passed to a single attention block. … In experiments where the cross-entropy loss was used, the Arcface center representation vectors for speaker classes were replaced with a fully connected (FC) layer. Here, the four FC outputs are fed into a softmax function, and the probability of each speaker class is obtained.



Wav2Vec 2.0 base
Attention Block max pooling
L2 normalize
$E \in \mathbb{R}^d$
$W_{id} \in \mathbb{R}^{n\_class \times d}$
0.2
0.1
0.6
0.1
cross-entropy
NOTATION. ⊗ computes the angular difference with the center representation vectors of classes

Figure 1: Speaker-identification network.

Section 3, Paragraph 2

(Modified Version)

The speaker-identification network (Fig. 1) encodes the vocal properties of a speaker into a fixed dimension vector, d. The wav2vec 2.0 model is used to encode input utterances into a latent 2-d representation of shape $\mathbb{R}^{768\times T}$, where T is the length of the input waveform. This latent representation is passed to a single attention block prior to performing a max-pooling operation that results in a 1-d vector of length 768. … . In order to achieve a more robust distinction between speakers, the $\mathbb{R}^d$ shape speaker-identification representation (H_{id}) and the $\mathbb{R}^{\#ID\times d}$ shape Arcface center representation vector (W_{id}) for speaker classes are l_2 normalized and their cosine similarity computed. … Then, the FC outputs are fed into a softmax function, and the probability of each speaker class obtained. In Figure. 1, "#ID" represents the index of each speaker class. For example, in the VoxCeleb1 dataset with 1,251 speakers, the final #ID is #1,251.

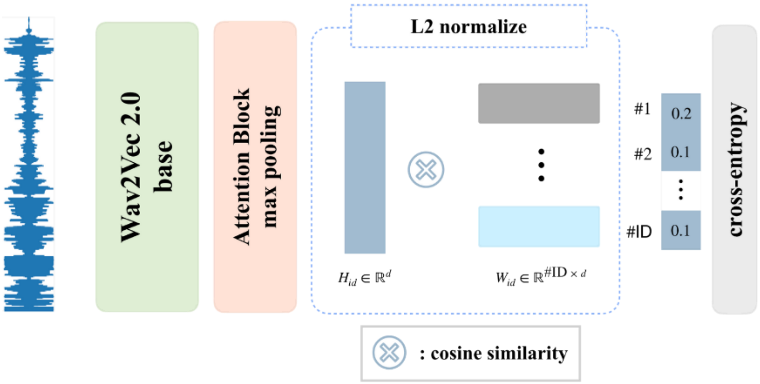

Wav2Vec 2.0 base
Attention Block max pooling
L2 normalize
$H_{id} \in \mathbb{R}^{d}$
$W_{id} \in \mathbb{R}^{\#\mathrm{ID} \times d}$
#1
#2
#ID
0.2
0.1
0.1
cross-entropy
: cosine similarity

Figure 1: Architecture of speaker-identification network with extended wav2vec 2.0 model structure (left) and l_2 normalization, cosine similarity and cross entropy loss computation (right) with a single output for each speaker class.

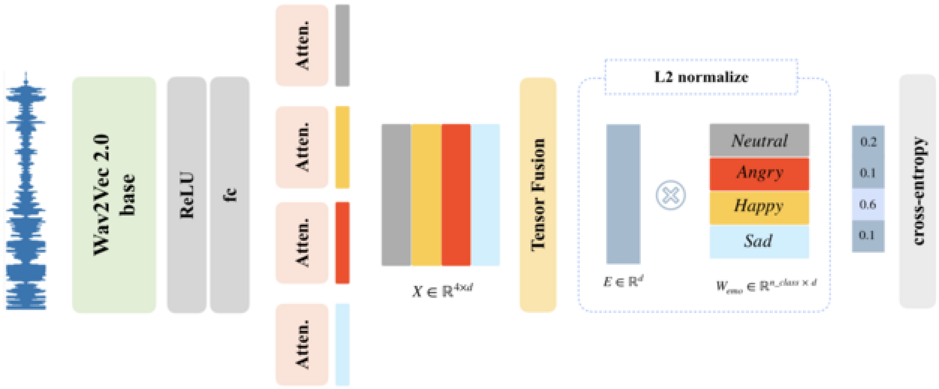
(Previous Version)

In the emotion classification network, the encoding generated by the wav2vec 2.0 segment is passed to a ReLU activation layer before being fed into an FC layer and eventually passed to four attention blocks. …

In order to address the differences in the lengths of utterances, max-pooling was applied across the time axis of each attention block, resulting in a d dimension vector as the final output. The output vectors of the speaker-identification network and emotion classification network have identical dimensions. During tensor fusion, an elementwise multiplication between $H = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k \times d}$) is performed. Summation is then performed across all k vectors as shown in Eq. (5),

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \tag{5}$$

where $e_i \in \mathbb{R}^{d}$ and $W_{fusion,i} \in \mathbb{R}^{d}$. The final embedding, E is used to compute the Arcface loss for emotion classification. …



NOTATION 1. ⊗ computes the angular difference with the center representation vectors of classes
NOTATION 2. Atten. is the attention block with max pooling

Figure 2: Emotion classification network with four attention blocks.

Section 3, Paragraph 3 & 4

(Modified Version)

In the emotion classification network (Fig. 2), the wav2vec 2.0 model is used to encode input utterances into a $\mathbb{R}^{768\times T}$shape representation. … The max pooled outputs of the attention blocks, h_i are concatenated prior to the tensor fusion operation. During tensor fusion, an elementwise multiplication between $H_{emo} = \{h_1, h_2, \cdots, h_k\}$ and a trainable fusion matrix ($W_{fusion} \in \mathbb{R}^{k\times d}$) is performed. As shown in Eq. (5), all the k vectors are summed to generate the final embedding.

$$E = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} W_{fusion,i} \odot h_i, \tag{5}$$

where $e_i \in \mathbb{R}^d$ and $W_{fusion,i} \in \mathbb{R}^d$. The final embedding, E is l_2 normalized prior to computing the cosine similarity between Arcface center representation vectors $W_{emo} \in \mathbb{R}^{\#EMO\times d}$. In Figure. 2, "#EMO" represents the emotion class indices as defined in the IEMOCAP dataset. Here, 1_EMO, 2_EMO, 3_EMO, 4_EMO represent angry, happy, sad and neutral emotion classes respectively.

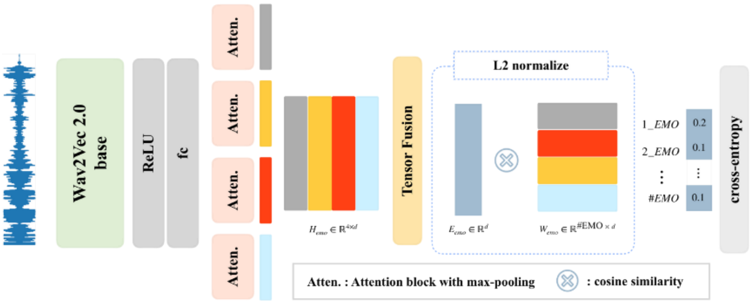

Wav2Vec 2.0 base
ReLU
fc
Atten.
Atten.
Atten.
Atten.
$H_{emo} \in \mathbb{R}^{4\times d}$
Tensor Fusion
L2 normalize
$E_{emo} \in \mathbb{R}^{d}$
$W_{emo} \in \mathbb{R}^{\#EMO \times d}$
1_EMO
2_EMO
⋮
#EMO
0.2
0.1
⋮
0.1
cross-entropy
Atten. : Attention block with max-pooling
⊗ : cosine similarity

Figure 2: Architecture of emotion classification network. Extended wav2vec 2.0 model structure (left) with four attention

blocks and a tensor fusion operation. l_2 normalization, cosine similarity and cross-entropy loss computation (right) for emotion classes with a single output for each emotion class.

(Previous Version)

The fusion process combines the speaker-identification and emotion representations from each network into a speaker-specific emotion representation. To achieve this, the output vector of the attention block is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs. Fig. 3 shows the architecture of the proposed speaker-specific emotion representation-based SER. The angular distance between the tensor fused output vector and the center of the four emotion representation vectors is calculated as shown in Eq. (4).

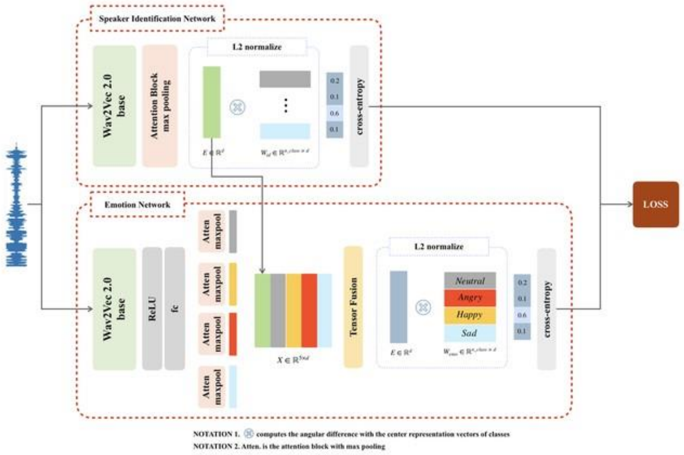

Speaker Identification Network
Wav2Vec 2.0 base
Attention Block max pooling
L2 normalize
$E \in \mathbb{R}^d$
$W_{id} \in \mathbb{R}^{n\_class \times d}$
0.2
0.1
0.6
0.1
cross-entropy
LOSS
Emotion Network
Wav2Vec 2.0 base
ReLU
fc
Atten maxpool
Atten maxpool
Atten maxpool
Atten maxpool
$X \in \mathbb{R}^{5 \times d}$
Tensor Fusion
L2 normalize
$E \in \mathbb{R}^d$
Neutral
Angry
Happy
Sad
$W_{emo} \in \mathbb{R}^{n\_class \times d}$
0.2
0.1
0.6
0.1
cross-entropy
NOTATION 1. ⊗ computes the angular difference with the center representation vectors of classes
NOTATION 2. Atten. is the attention block with max pooling

Figure 3: Architecture of proposed speaker-specific emotion representation generation framework.

Section 3, Paragraph 4, Sentence 6~9

(Modified Version)

… More specifically, the output vector of the attention block from the speaker-identification network is concatenated to the outputs of the emotion classification network's four attention blocks, resulting in a total of five attention block outputs, $H \in \mathbb{R}^{5 \times d}$. The fusion operation shown in Eq. (5) is used to combine these representations into a single speaker-specific emotion representation E. The angular distance between the normalized tensor fused output vector and the normalized center of the four emotion representation vectors is calculated using Eq. (4). The emotion class predicted for any input waveform, is determined by how close its representation vector is to an emotion class's center vector.

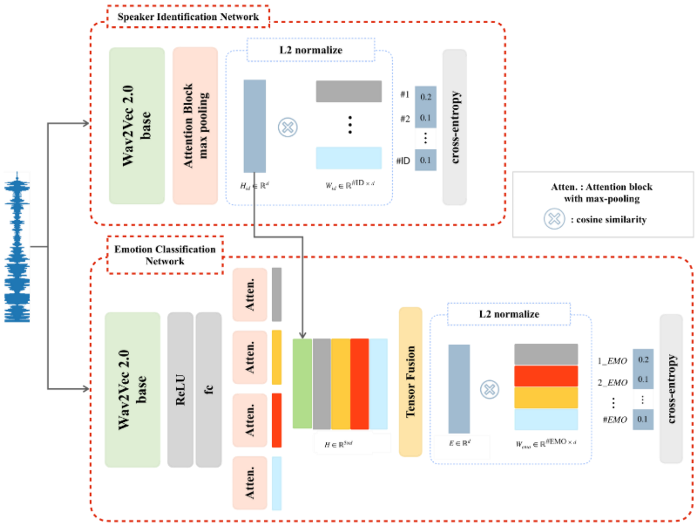

Speaker Identification Network
Wav2Vec 2.0 base
Attention Block max pooling
L2 normalize
$H_{id} \in \mathbb{R}^{d}$
$W_{id} \in \mathbb{R}^{\#ID \times d}$
#1
#2
#ID
0.2
0.1
0.1
cross-entropy
Atten. : Attention block with max-pooling
: cosine similarity
Emotion Classification Network
Wav2Vec 2.0 base
ReLU
fc
Atten.
Atten.
Atten.
Atten.
$H \in \mathbb{R}^{5\times d}$
Tensor Fusion
L2 normalize
$E \in \mathbb{R}^{d}$
$W_{emo} \in \mathbb{R}^{\#EMO \times d}$
1_EMO
2_EMO
#EMO
0.2
0.1
0.1
cross-entropy

Figure 3: Architecture of speaker-specific emotion representation model with speaker-identification network (up) to generate speaker representation and emotion classification (down) to generate speaker-specific emotion representation from emotion and speaker-identification representations.

3. Reviewer’s comment:

The ablation study is missing, and model complexity should also be discussed.

Revisions made:

In response to the reviewer’s comment, we have conducted ablation experiments and added the results to the revised version of the manuscript. More specifically, in order to investigate the effects of using the speaker-specific representation, experiments were conducted using just the emotion classification network and then the speaker-specific emotion representation network. In these experiments, cross-entropy and Arcface losses as well as configuration of the networks with 1, 2, 3, and 4 attention blocks were used. As shown in Table C, the inter-class compactness and inter-class separability facilitated by the Arcface loss results in better performance than when the cross-entropy loss is used for almost all cases. Using the speaker-specific emotion representation also outperformed the plain emotion representation under almost all configurations. In Table C, the highlighted regions contain the results of newly conducted experiments for the further extended ablation study.

Table C: Performance (accuracy) of speaker-specific emotion representation network trained with cross-entropy and Arcface losses under 1, 2, 3, and 4 attention block configurations.

	# of attention blocks		1	2	3	4
Emotion classification network	Cross-Entropy	WA	70.16	69.10	68.96	68.48
		UA	70.36	69.91	70.45	70.00
	Arcface	WA	70.03	70.26	70.36	71.05
		UA	71.37	71.14	71.14	72.15
Speaker-specific emotion representation network	Cross-Entropy	WA	70.14	70.26	70.02	69.33
		UA	70.91	71.00	70.99	70.64
	Arcface	WA	71.02	70.67	70.03	**72.14**
		UA	71.35	71.37	71.74	**72.97**

In response to the reviewer's comment regarding the proposed model's complexity, our experiments were conducted under two performance measures: computation load and space complexity. Table D shows the number of floating-point operations performed by various models when audio segments of 3, 5, 10, and 15 seconds were fed into them as input. The number of training parameters for each model is also indicated.

Table D: Comparison of space complexity (# of model parameters) and computation load (FLOPS) of various models for input audio segments of varying lengths (3, 5, 10, and 15 seconds).

Models	# of model parameters (bytes)	Input audio segment length (seconds)			
		3	5	10	15
Wav2vec 2.0-Base	94.37M	21.23G	35.9G	74.2G	0.11T
Wav2vec 2.0-Large	0.32G	54.85G	92.84G	0.19T	0.3T
Speaker-identification network	97.69M	21.62G	36.59G	75.76G	0.12T
Emotion classification network (four attention blocks)	0.1G	22.86G	38.78G	80.73G	0.13T
Speaker-specific emotion representation network (four attention blocks)	0.2G	44.47G	75.37G	0.16T	0.24T

In response to the reviewer's comment regarding the proposed model's time complexity, we have compared the computation times under various configurations: varied length of input audio segments and number of attention blocks. More specifically, input audio segments of 3, 5, 10, and 15 seconds as well as model configurations with 1, 2, 3, and 4 attention blocks were used. The proposed model (speaker-specific emotion representation network) consists of two networks (speaker-identification and emotion classification). Table E shows the separate and combined *computation times of the two networks under the above-mentioned configurations. As shown in Table E, computation time increases as the length of input audio segments and the number of attention blocks increases. Experimental results showed that for all configurations the model can process an input audio waveform within 27.26 ms, implying that the proposed approach is suitable for real-time applications.*

Table E: Computation time (ms) of proposed networks (speaker-identification, emotion classification, and speaker-specific emotion representation networks) for input audio segments of varying lengths (3, 5, 10, and 15 seconds).

			Input audio segment length (seconds)			
			3	5	10	15
Speaker-identification network			2.52	4.20	8.32	13.08
Emotion classification network	# of attention blocks	1	2.49	4.16	8.24	12.96
		2	2.57	4.28	8.52	13.41
		3	2.64	4.40	8.78	13.84
		4	2.71	4.52	9.00	14.22
Speaker-specific emotion representation network (four attention blocks)			5.28	8.79	17.39	27.26

(Modified Version)

In order to investigate the effects of using the speaker-specific representation, experiments were conducted at first using just the emotion classification network and then using the speaker-specific emotion representation network. More specifically, in order to investigate the effects of using the speaker-specific representation, cross-entropy and Arcface losses as well as configurations of the networks with 1, 2, 3, and 4 attention blocks were used. As shown in Table 6, the inter-class compactness and inter-class separability facilitated by the Arcface loss results in better performance than when the cross-entropy loss is used for almost all cases. Using the speaker-specific emotion representation also outperformed the bare emotion representation under almost all configurations.

The computation time of the proposed method under various configurations was examined. The length of input audio segments (3, 5, 10, and 15 seconds) and number of attention blocks (1, 2, 3, and 4) were varied. The proposed model (speaker-specific emotion representation network) consists of two networks (speaker-identification and emotion classification). Table 6 shows the separate and combined computation times of the two networks under the above-mentioned configurations. As shown in Table 7, computation time increases as the length of input audio segments and the number of attention blocks increases. Experiments show that the best-performing configuration of the proposed model in which the speaker-specific emotion representation network has four attention blocks can process audio segments in close to 27ms.

Subsection 5.4, Paragraph 2-3

4. Reviewer’s comment:

More literatures are recommended for discussion. such as:

10.1109/TGRS.2022.3225902

10.1109/TASLP.2023.3277291

Revisions made:

We deeply appreciate the research papers([A] and [B]) suggested by the reviewer and we have carefully examined them. First, Dutt et.al.[A] proposed a wavelet-based approach to audio signal representation that addresses some of the tradeoff challenges associated with the Fast Fourier transform used in earlier representations. This method is not a deep learning-based representation method but was shown to offer better performance in downstream tasks than other hand-crafted audio representation methods. For this reason, we have included sub-section 2.1 (Hand-Crafted Audio Representation) of the revised version of the manuscript.

In [B], Chen et.al. proposed a spatial feature extraction method in which global spatial features are extracted from hyperspectral images at multiple scales, and local spatial features extracted from around each pixel using morphological attribute profiles. Whereas this form of feature extraction has been proven to be effective on image data, its applicability to audio waveform data needs to be further investigated. This is certainly an approach we will consider employing in future research.

[A]: A. Dutt and G. Paul, “Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2043-2054, 2023. 10.1109/TASLP.2023.3277291

[B]: Z. Chen, Z. Lu, H. Gao, Y. Zhang, J. Zhao *et.al*., “Global to Local: A Hierarchical Detection Algorithm for Hyperspectral Image Target Detection,” in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022, doi: 10.1109/TGRS.2022.3225902.

(Previous Version)

A vast array of representations and models have been explored to improve audio speech-based emotion recognition. LLDs such as pitch and energy contours have been employed in conjunction with hidden Markov models [10] to recognize a speaker's emotion from audio speech. [11] used the delta and delta-delta of a log Mel-spectrogram to reduce the impact of emotionally irrelevant factors on speech emotion recognition. In this approach, an attention layer automatically drove focus to emotionally relevant frames and generated discriminative utterance-level features. Global-Aware Multi-scale (GLAM)[12] used MFCC inputs and a global-aware fusion module to learn a multi-scale feature representation, which is rich in emotional information.

Subsection 2.1

(Modified Version)

A vast array of representations and models have been explored to improve audio speech-based emotion recognition. LLDs such as pitch and energy contours have been employed in conjunction with hidden Markov models [10] to recognize a speaker's emotion from audio speech. [11] used the delta and delta-delta of a log Mel-spectrogram to reduce the impact of emotionally irrelevant factors on speech emotion recognition. In this approach, an attention layer automatically drove focus to emotionally relevant frames and generated discriminative utterance-level features. Global-Aware Multi-scale (GLAM) [12] used Mel-frequency cepstral coefficient (MFCC) inputs and a global-aware fusion module to learn a multi-scale feature representation, which is rich in emotional information.

Time-frequency representations such as the Mel-spectrogram and MFCCs merge frequency and time domains into a single representation using the Fast Fourier transform (FFT). [13] addressed the challenges associated with the tradeoff between accuracy in frequency and time domains by employing a wavelet transform based representation. Here, Morlet wavelets generated from an input audio sample are decomposed into child wavelets by applying a continuous wavelet transform (CWT) to the input signal with varying scale and translation parameters. These CWT features are considered as representation that can be employed in downstream tasks.

Subsection 2.1, Paragraph 2

[13] A. Dutt and G. Paul, “Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2043-2054, 2023.

5. Additional Revision:

In the process of revising the manuscript, we discovered that Eq. (3) was identical to Eq. (2), which was an error. Eq. (3) was originally intended to describe the terms of the Arcface loss defined in Eq. (2) in an extended form in which the dot product is expressed as the product of the norms of vectors and the cosine of the angle between them. Therefore, we have modified Eq. (3) to reflect the original intention.

(Previous Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 2

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

Subsection 2.6, Eq. 3

(Modified Version)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_j^T x_i}}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\left\|W_{y_i}\right\| \|x_i\| cos\theta_{y_i}}}{\sum_{j=1}^{n} e^{\|W_j\| \|x_i\| cos\theta_j}},$$