

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

Somin Park¹, Mpabulungi Mark¹, Bogyung Park², and Hyunki Hong^{1,*}

¹College of Software, Chung-Ang University, Heukseok-ro 84, Dongjak-ku, Seoul, 06973, Rep. of Korea

²Department of AI, Chung-Ang University, Heukseok-ro 84, Dongjak-ku, Seoul, 06973, Rep. of Korea

*Corresponding Author: Hyunki Hong. Email: honghk@cau.ac.kr

Received: XX Month 202X; Accepted: XX Month 202X

Abstract: Speech emotion recognition is essential for frictionless human-machine interaction, where machines respond to human instructions with context-aware actions. The properties of individuals' voices vary with culture, language, gender, and personality. These variations in speaker-specific properties may hamper the performance of standard representations in downstream tasks such as speech emotion recognition (SER). In this study, we demonstrate the significance of speaker-specific speech characteristics and how considering them can be leveraged to improve the performance of SER models. In the proposed approach, there are two wav2vec-based modules (speaker-identification network and emotion classification network) trained with the Arcface loss. The speaker-identification network has 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. 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.

Keywords: attention block; IEMOCAP dataset; speaker-specific representation; speech emotion recognition; wav2vec 2.0

1 Introduction

The recent rapid growth of computer technology has made human-computer interaction an integral part of the human experience. Advances in automatic speech recognition (ASR) [1] and text-to-speech (TTS) synthesis [2] have made smart devices capable of searching and responding to verbal requests. However, this only supports limited interaction and is not sufficient for interactive conversations. Most ASR methods generally focus on the content of speech (words) without regard for the intonation, nuance, and emotion conveyed through audio speech. Speech emotion recognition (SER) is, therefore, one of the most active research areas in the computer science field because the friction in every human-computer interaction could be significantly reduced if machines could perceive and understand the emotions of their users and perform context-aware actions.

Previous studies used low-level descriptors (LLDs) generated from frequency, amplitude, and spectral properties (spectrogram, Mel-spectrogram, etc.) to recognize emotions in audio speech. 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 LLDs or the actual waveform have shown better performance in SER.



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.

This paper introduces a novel approach based on speaker-specific emotion representation to improve emotional speech recognition performance. 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. A novel tensor fusion approach combines these representations into a speaker-specific emotion representation. 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:

- Two modules (emotion and identification) based on wav2vec 2.0 to generate a speaker-specific emotion representation from an audio input are proposed. The two modules are trained and evaluated on the Interactive Emotional Dynamic Motion Capture (IEMOCAP) dataset [7]. Training networks on the IEMOCAP dataset is prone to over-fitting because it has only ten speakers. 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.
- A novel tensor fusion approach combines generated emotion and speaker-specific representations into a single vector representation suitable for SER. The use of Arcface [9] and cross-entropy loss terms in the speaker-identification network was also explored and detailed evaluations provided.

2 Related Work

2.1 Hand-crafted Audio Representation

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.

2.2 Learning Audio Representation Using Supervised Learning

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. It can be used as a learnable replacement to Mel-filter banks in existing deep learning models. In order to approximate the Mel-filter banks, the square of the Hanning window was used, and the biases of the convolutional layers were set to zero. Due to the absence of positivity constraints, one was added to the output before applying log compression. However, log-scale compression and normalization reduce the scale of spectrograms, regardless of their contents.

Reference [14] 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. In order to address the compression function's fixed non-linearity, PCEN was modified to learn channel-dependent smoothing coefficients alongside the other hyper-parameters [15] in a version of the model referred to as sPer-Channel Energy Normalization (sPCEN).

2.3 Learned Audio Representation Using Self-Supervised Learning

In supervised learning, class labels are used to design convolution filters and generate task-specific representations. Due to the vast amounts of unlabeled audio data available, self-supervised learning (SSL) methods have been proposed for obtaining generalized representations of input audio waveforms for downstream tasks. These audio SSL methods can be categorized into auto-encoding, siamese, clustering, and contrastive techniques [16].

Audio2vec [17] was inspired by word2vec [18] and learned general-purpose audio representations by using an auto-encoder-like structure to reconstruct a Mel-spectrogram slice from past and future slices. Continuous Bags of Words (CBoW) and skip-gram variants were implemented and evaluated. The Mockingjay [19] network learned general-purpose audio representations by using bidirectional Transformer encoders to predict the current frame by being jointly conditioned on both past and future contexts. BYOL-A [20] is a Siamese model-based architecture that assumes no relationships exist between time segments of audio samples. In this architecture, two neural networks were trained by maximizing the agreement of their output given the same input. A combination of normalization and augmentation techniques was also used to differentiate augmented versions of the same audio segment, thereby learning a general-purpose audio representation. Hidden unit Bert (HuBERT) [21] addressed the challenges associated with multiple sound units in utterance, the absence of a lexicon of input sounds, and variable length of sound units by using an offline clustering step to provide aligned target labels for a prediction loss similar to that in Bert [22]. This prediction loss was only applied over masked regions, thereby forcing the model to learn combined acoustic and language model inputs. The model was based on the wav2vec 2.0 architecture that consists of a convolutional waveform encoder, projection layer, and code embedding layer but no quantization layer. The encoded vectors were pseudo-labeled by K-means using MFCCs of the input waveforms and then clustered to generate the audio representations.

Contrastive methods generated an output representation using a loss function that encourages the separation of positive from negative samples. For instance, Contrastive Learning of Auditory Representations (CLAR) [23] encoded both the waveform and the spectrogram into audio representations. Here, the encoded representations of the positive and negative pairs are used in a contrastive manner.

2.4 Using Speaker Attributes in SER

Individual Standardization Network (ISNet) [4] showed that considering speaker-specific attributes can improve emotion classification accuracy. [4] uses an aggregation of an individual's neutral speech to standardize emotional speech and improve the robustness of individual-agnostic emotion representation. A key limitation of this approach is that it is only applicable to cases where labeled neutral training data for each speaker is available. Self Speaker Attentive Convolutional Recurrent Neural Net (SSA-CRNN) [5] uses two classifiers that interact through a self-attention mechanism to focus on emotion information and ignore speaker-specific information. This approach is limited by its inability to generalize to unseen speakers.

2.5 Wav2vec 2.0

Wav2vec 2.0 converted an input speech waveform into spectrogram-like features by predicting the masked quantization representation over the entire speech sequence [6]. Wav2vec [24], the earliest version of wav2vec 2.0, attempted to predict future samples from a given signal context. It consists of an encoder network that embedded the audio signal into a latent space and a context network that combined multiple time-steps of the encoder to obtain contextualized representations. VQ-wav2vec [25] learned discrete representations of audio segments using a future time-step prediction task as in previous methods but

replaced the original representation with a Gumbel-Softmax-based quantization module. Wav2vec 2.0 adopted both the contrastive loss and the diversity loss in the VQ-wav2vec framework. In other words, wav2vec 2.0 compared positive and negative samples without predicting future samples.

Wav2vec 2.0 consists of a feature encoder, contextual encoder, and quantization module. First, the feature encoder converts the normalized waveform into a two-dimensional (2-d) latent representation. The feature encoder was implemented using seven 1-d convolution layers with different kernel sizes and strides. A Hanning window of the same size as the kernel and an STFT with a hop length equal to the stride was used. The encoding that the convolutional layers generate from an input waveform is normalized and passed as inputs to two separate branches (the contextual encoder and quantization module). The contextual encoder consists of a linear projection layer, a relative positional encoding 1-d convolution layer followed by a GeLU, and Transformer. More specifically, each input is projected to a higher dimensional feature space and then encoded based on its relative position in the speech sequence. Here, the projected and encoded input, along with its relative position, are summed and normalized. The resultant speech features are randomly masked and fed into the Transformer, aggregating the local features into a context representation C . The quantization module discretizes the feature encoder's output into a finite set of speech representations. This amounts to choosing V quantized representations (codebook entries) from multiple codebooks using a Gumbel softmax, concatenating them, and applying a linear transformation to the final output. A diversity loss encourages the model to use code book entries equally often.

The contextual representation c_t of the masked time step t is compared with the quantized latent representation q_t at the same time step t . The contrastive loss makes c_t similar to q_t and c_t dissimilar to K sampled quantized representations in every masked time step $Q \sim q_t$. The contrastive task's loss term is defined as

$$L_m = -\log \frac{\exp(c_t^T / (\|c_t\| \|q_t\| \kappa))}{\sum_{q \sim Q_t} \exp(c_t^T q / (\|c_t\| \|q\| \kappa))}, \quad (1)$$

where κ is the temperature of the contrastive loss. The diversity loss and the contrastive loss are balanced using a hyper-parameter. A more detailed description is available in the wav2vec 2.0 paper [6].

The wav2vec 2.0 representation has been employed in various SER studies because of its outstanding ability to create generalized representations that can be used to improve acoustic model training. SUPERB [26] evaluated how well pre-trained audio SSL approaches performed on ten speech tasks. The pre-trained SSL networks with high performance can be frozen and employed on downstream tasks. Since the outputs of SSL networks effectively represent the frequency features in the speech sequence, the length of representations varies with the length of utterances. In order to obtain a fixed-size representation for utterances, average time pooling is performed. A fully connected layer is then used to output the final emotion classification. The wav2vec representation has been extended in more recent studies. In [26], the wav2vec 2.0 representation was extensively evaluated. The original weights were frozen, average time pooling was applied, and a fully connected layer was added. In [27], the feasibility of partly or entirely fine-tuning these weights was examined. [28] proposed a transfer learning approach in which the output of several layers of the pre-trained wav2vec 2.0 model was combined using trainable weights, which were learned jointly with a downstream model.

2.6 Additive Angular Margin Loss

Despite their popularity, earlier losses like the cross-entropy did not encourage intra-class compactness and inter-class separability [29] for classification tasks. In order to address this limitation, contrastive [30], triplet [31], center [32], and Sphereface [2] losses encouraged the separability between learned representations. Additive Angular Margin Loss (Arcface) [9] and Cosface [33] achieved better separability by encouraging stronger boundaries between representations. In Arcface, the representations were distributed around feature centers in a hypersphere with a radius. An additive angular penalty was employed to enhance the intra-class compactness and inter-class discrepancy simultaneously. Here the angular difference between an input feature vector $x \in \mathbb{R}^d$ and the center representation vectors of classes, $W \in \mathbb{R}^{N \times d}$ is calculated. 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. [34] used the Arcface loss to train a bimodal audio text network for SER and reported improved performance. A similar loss term is used in our proposed method.

Eq. (2) is the equivalent of calculating the softmax with a bias of 0. Following logit transformation, Eq. (2) can be rewritten as Eq. (3).

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^n e^{W_j^T x_i}}, \quad (2)$$

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^n e^{W_j^T x_i}}, \quad (3)$$

where $\|\cdot\|$ is the l_2 normalization and θ_j is the angle between W_j and x_i . In Eq. (4), the additive margin penalty, m is only added to the angle, θ_{y_i} , between the target weight W_{y_i} and the features x_i . The features are re-scaled using the scaling factor, s . The final loss is defined as:

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^n e^{s(\cos \theta_j)}}. \quad (4)$$

[34] demonstrates the Arcface loss term's ability to improve the performance of SER models. It is therefore employed in training the modules proposed in this study.

3 Method

3.1 Network Architecture

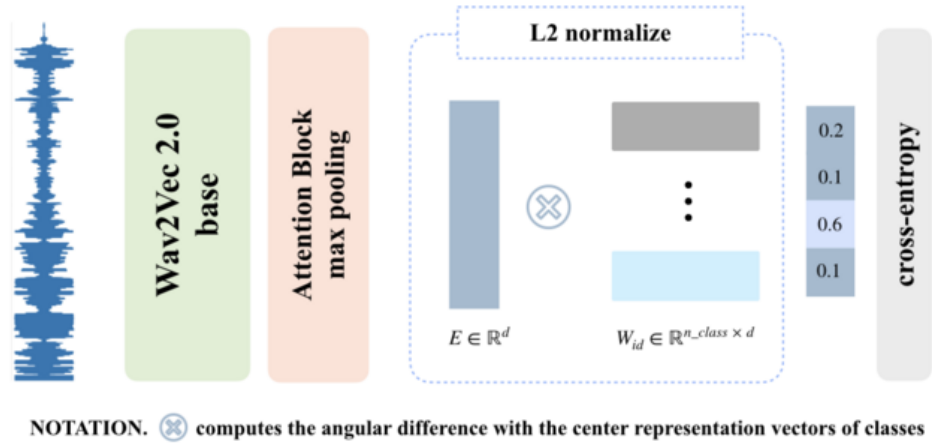


Figure 1: Speaker-identification network

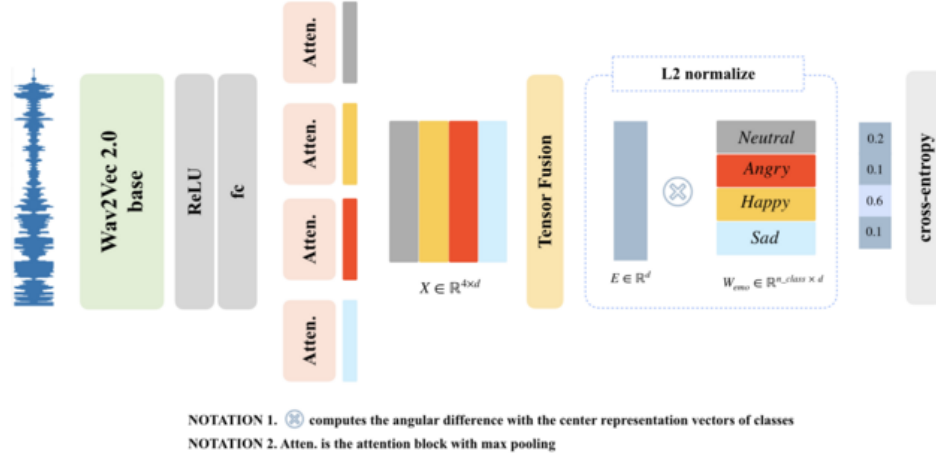


Figure 2: Emotion classification network with four attention blocks.

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 $H = \{h_1, h_2, \dots, h_k\}$ 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),

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

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).

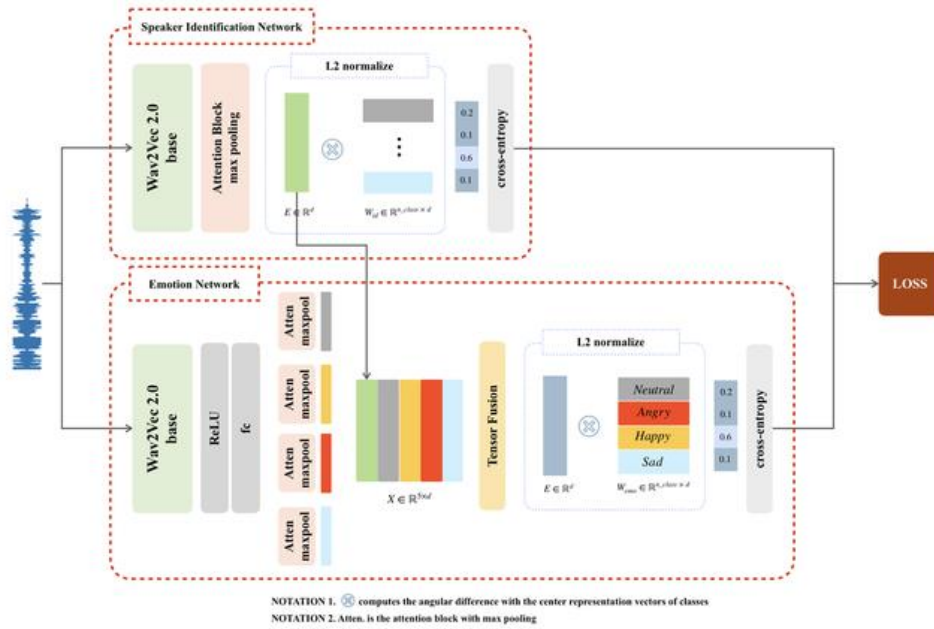


Figure 3: Architecture of proposed speaker-specific emotion representation generation framework.

4 Experiment Details

4.1 Dataset

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. It consists of approximately 12 hours of audio-visual data, including facial images, speech, and text transcripts. The audio speech data provided is used to train and evaluate models for emotion recognition. Categorical (angry, happy, sad, and neutral) as well as dimensional (valence, activation, and dominance) labels are provided. 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]. The audio sampling rate, 16kHz used in the original dataset is retained. The average length of audio files is 4.56 seconds with a standard deviation of 3.06 seconds. The minimum length of audio files is 0.58 seconds, and the maximum length is 34.14 seconds. 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. For audio files shorter than 3 seconds, a copy of the original waveform is recursively appended to the end of the audio file until the audio file is at least 3 seconds long. Fig. 4 shows how often various emotions are expressed by male and female speakers over five sessions in the IEMOCAP dataset. As shown in Fig. 4, the dataset is unevenly distributed across emotion labels, with significantly more neutral and happy samples in most sessions.

In order to generate an evenly distributed random set of samples at each epoch, emotion classes with more samples are under-sampled. This implies that samples of the training dataset are evenly distributed across all the emotional classes. Leave-one-session-out five-fold cross-validation is used.

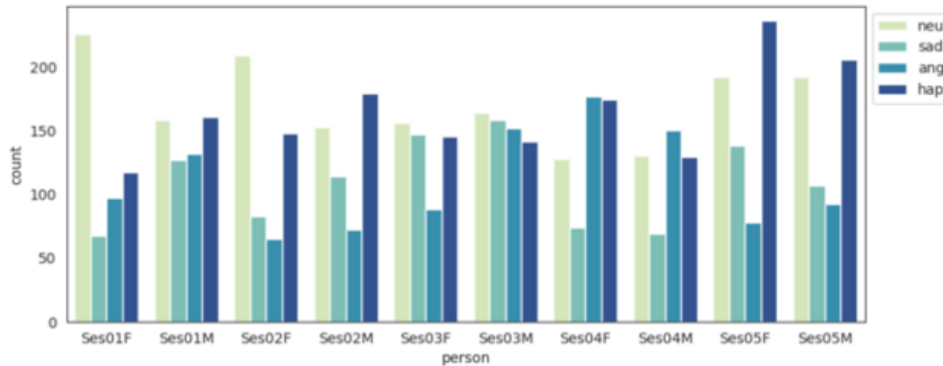


Figure 4: Distribution of male and female speakers across emotion classes in the IEMOCAP dataset.

In this study, VoxCeleb1[8]’s large variation and diversity allow for the speaker-identification module to be trained for better generalization to unseen speakers. VoxCeleb1 is an audio-visual dataset comprising 22,496 short interview clips extracted from YouTube. It features 1,251 speakers from diverse backgrounds and is commonly used for speaker identification and verification tasks. Its audio files have a sampling rate of 16kHz with an average length of 8.2 seconds as well as minimum and maximum lengths of 4 and 145 seconds, respectively. Additionally, audio clips in VoxCeleb1 are also limited to a maximum length of 15 seconds for consistency in our experiments.

4.2 Implementation Details

In recent studies [27-28], pre-training the wav2vec model on the Librispeech dataset [35] (with no fine-tuning for ASR tasks) has been shown to deliver better performance for SER tasks. Therefore, this study proposes two networks that are based on the wav2vec 2.0 representation (sub-section 2.5). 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. The wav2vec 2.0 modules (the contextual encoder) used in this study were therefore only partially fine-tuned. The model and weights are provided by facebookresearch/fairseq [37].

To ensure that our network learns the appropriate attributes, the training process consists of two steps. First, the identification network and emotion network are trained separately. Then, the pre-trained networks are integrated and fine-tuned with the extended tensor fusion matrix to match the size of concatenated speaker-identification and emotion representations. In order to prevent over-fitting and exploding gradients, gradient values are clipped at 100 with n -step gradient accumulations. 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. An early stopping criterion is added to prevent over-fitting. Each of the attention blocks consists of four attention heads and had a dropout rate of 0.1. In the Arcface loss calculation, the feature re-scaling factor (s) is set to 30 and the additive margin penalty (m) to 0.3 for our experiments. Experiments were conducted using Pytorch in an Ubuntu 20.04 training environment running on a single GeForce RTX 3090 GPU. Specific hyper-parameters of experiments are shown in Table 1.

Table 1: Hyper-parameters used during model evaluation.

module	learning rate	batch size	n -step gradient accumulation	early stopping limit	total epoch
Speaker-identification network	$3 * 10^{-5}$	16	2	5	50
Emotion network	$3 * 10^{-5}$	6	4	10	150

Integrate and fine-tune networks	10^{-5}	6	4	10	100
-------------------------------------	-----------	---	---	----	-----

4.3 Evaluation Metrics

In this paper, weighted and unweighted accuracy metrics are used to evaluate the performance of the proposed model. 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. WA can be computed from a confusion matrix containing prediction scores as Eq. (7).

$$WA = \frac{TP+TN}{TP+TN+FP+FN}, \quad (7)$$

where the number of true positive, true negative, false positive, and false negative cases are TP, TN, FP, and FN, respectively. 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).

$$UA = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i+FP_i}, \quad (8)$$

where C is the total number of emotion classes. In our experiment, C is set to four.

5 Experimental Results

5.1 Performance of Speaker-Identification Network and Emotion Classification Network

Table 2 shows the performance of the speaker-identification network on the VoxCeleb1 identification test dataset. Training the speaker identification network using the Arcface loss resulted in significantly better speaker classification than training with the cross-entropy loss. This indicates that the angular margin in the Arcface loss improves the network's discriminative abilities for speaker identification. Fig 5 shows a t-SNE plot of speaker-specific representations generated from the IEMOCAP dataset using two configurations of the speaker-identification network. As shown in Fig 5, using the Arcface loss results in more distinct separations between speaker representations than using the cross-entropy loss. 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. In order to ensure that input audio waveforms have the information necessary to generate a speaker-specific emotion representation, a 3 second requirement is imposed. In cases where the audio waveform is shorter than 3 seconds, a copy of it is recursively appended to the end of itself until it is at least 3 seconds long.

Table 2: Results of proposed methods with two losses (cross-entropy and Arcface) on the speaker-identification network.

Loss	WA (%)	UA (%)
Cross-entropy	87.98	87.19
Arcface	93.89	94.22

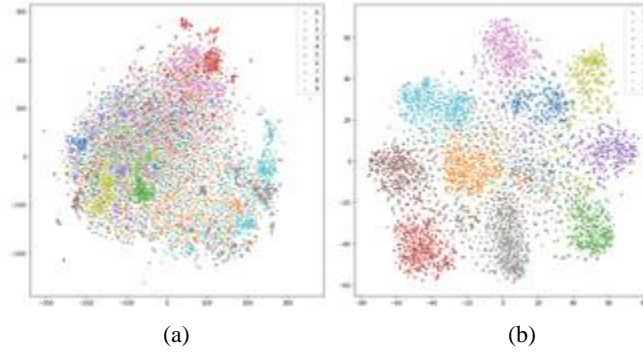


Figure 5: t-SNE plot of speaker-specific representations generated by the speaker-identification network under different loss functions: (a) Cross-entropy (b) Arcface.

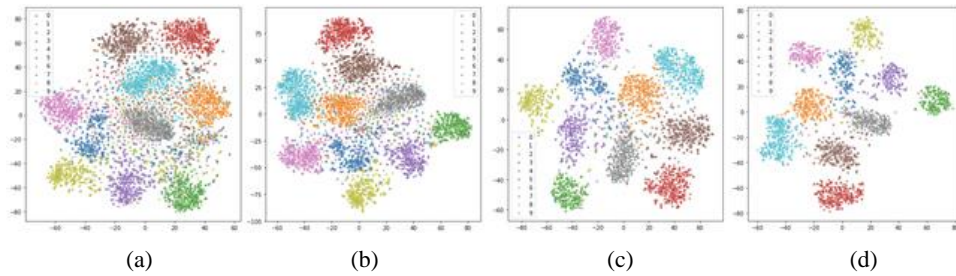


Figure 6: t-SNE plot of speaker-specific representations generated by the speaker-identification network under various minimum duration configurations: (a) 1 second (b) 2 second (c) 3 second (4) 4 second.

Table 3: Performance of proposed emotion classification network methods against previous methods (with Arcface loss and the number of attentional blocks).

Models	# of emotion attention blocks	WA (%)	UA (%)
SUPERB[26]: wav2vec 2.0 base	-	63.43	-
Y. Wang[27]: w2v-base (Partly Fine-tuned)	-	63.43	-
L. Pepino[28]: Global Normalized	-	-	66.3±0.7
L. Pepino[28]: Speaker Normalized	-	-	67.2±0.7
Y. Tang[34]: Audio Only	-	71.80	69.60
Proposed Method	1	70.03	71.37
	2	70.26	71.14
	3	70.36	71.14
	4	71.05	72.15

Table 3 shows a comparison of our methods' performance against previous studies. The first four methods employed the wav2vec 2.0 representation and used the cross-entropy loss [26–28], and the fifth method [34] employed hand-crafted features and used the Arcface loss. Here, the individual vocal properties provided by the speaker-identification network are not used. Table 3 shows that the method

proposed by Y. Tang [34] has a higher WA than UA. This implies that emotion classes with more samples, particularly in the imbalanced IEMOCAP dataset, are better recognized. The wav2vec 2.0-based methods, [26–28] used average time pooling to combine features across the time. [28] also included an LSTM layer to better model the temporal features. In our proposed method, the cross-entropy loss is replaced with an Arcface loss, and an attention block is used to model temporal features. Table 3 shows that the proposed attention-based method outperforms previous methods. It also demonstrates that the use of four attention blocks results in significantly better performance than the use of one, two, or three attention blocks. This is because four attention blocks can more effectively identify the segments of the combined emotion representation that are most relevant to SER.

5.2 Partially and Entirely Fine-tuning Networks

The proposed speaker-identification network can be fine-tuned under three different configurations: fine-tuning with the entire pre-trained network frozen (All Frozen), fine-tuning with wav2vec 2.0 segment frozen and Arcface center representation vectors unfrozen (Arcface Fine-tuned), and fine-tuning with both wav2vec 2.0 weights and the Arcface center representation vectors unfrozen (All Fine-tuned). The wav2vec 2.0 feature encoder (CNN layers) is frozen in all cases [27]. The IEMOCAP dataset only has 10 individuals. Therefore, the Arcface center representation vectors are reduced from 1,251 (in the VoxCeleb1 dataset) to 8 while jointly fine-tuning both the speaker-identification network and the emotion classification network. While fine-tuning with both wav2vec 2.0 weights and the Arcface vectors unfrozen, the loss is computed as a combination of emotion and identification loss terms as shown in Eq. (6):

$$L = \alpha \times L_{emotion} + \beta \times L_{identification}, \quad (6)$$

α and β are used to control the extent to which emotion and identification losses, respectively, affect the emotion recognition results. 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. Fig 7 shows that freezing the speaker-identification network provides the best overall performance. 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. More specifically, the frozen version of the speaker-identification module was trained and frozen on the VoxCeleb1 dataset, because it has 1,251 speakers' utterances which provide significantly larger variation and diversity than the utterances of the 8 speakers (training dataset) in the IEMOCAP dataset. This implies that the frozen version can provide better generalization to unseen speakers in comparison to versions fine-tuned on the 8 speakers of the IEMOCAP dataset as shown in Fig 7 (b) and (c).

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 β causes the emotion classification accuracy to deteriorate. This implies that partly or entirely freezing the weights of the speaker-identification network preserves the representation learned from the 1,251 speakers of the VoxCeleb1 dataset, resulting in better emotion classification performance. On the other hand, fine-tuning the entire model on 8 speakers of the IEMOCAP dataset degrades the generalization ability of the speaker-identification network. More specifically, in the partly frozen version, only the attention-pooling and speaker classification layers are fine-tuned, leaving the pre-trained weights of the speaker-identification network intact.

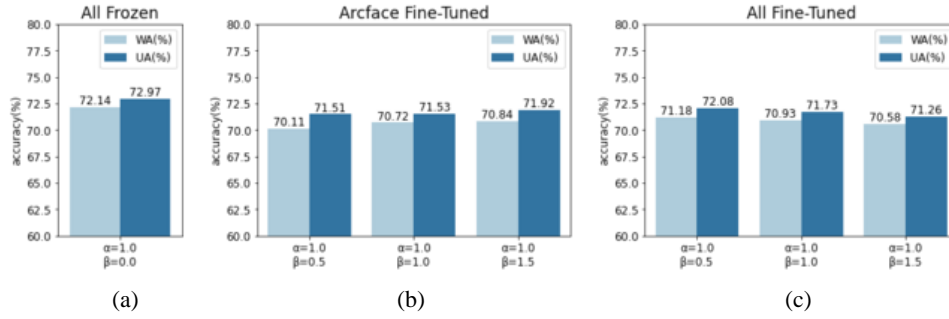


Figure 7: Performance of proposed method under various levels of fine-tuning the speaker-identification network: (a) All Frozen (b) Arcface Fine-tuned (c) All Fine-tuned.

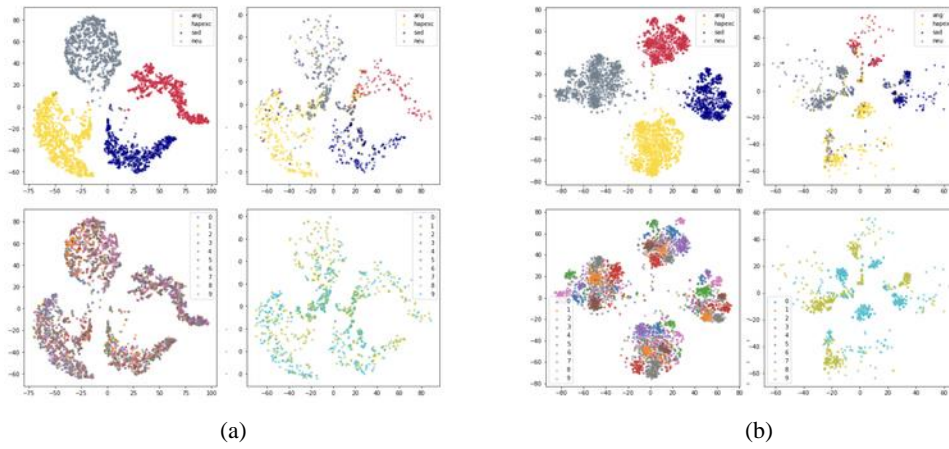


Figure 8: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) without the speaker-specific representation (b) with the speaker-specific representation.

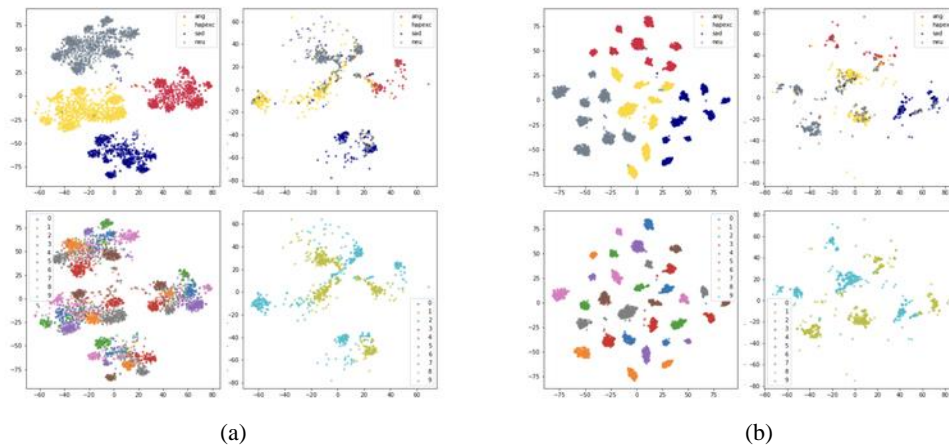


Figure 9: t-SNE plot of emotion representations generated by the emotion classification network under two configurations: (a) Arcface fine-tuned (b) All fine-tuned.

Fig 8 and 9 show t-SNE plots of emotion representations generated by the emotion classification network under various configurations. In Fig 8 (a) and (b), the left column contains representations generated from the training set and the right column contains those generated from the test set. 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). More specifically, Fig 8 illustrates the effect of employing the speaker-specific representations generated by the frozen speaker-identification network in the emotion classification network. As shown in Fig 8, using the speaker-specific representations improves intra-class compactness and increases inter-class separability between emotional classes in comparison to training without the speaker-specific representation. The emotion representation generated when speaker-specific information was utilized shows a clear distinction between the 8 speakers in the IEMOCAP dataset and their corresponding emotion classes.

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

5.3 Comparison of Previous Methods

In Table 3, we have compared the proposed method against the previous SER methods based on the wav2vec 2.0 or the Arcface loss. In Table 4, the performance of the proposed method under various configurations is compared against that of existing approaches on the IEMOCAP dataset. In Table 4, “EF” and “PF” stand for “entirely fine-tuned” and “partially fine-tuned”, respectively. 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. Therefore, it is this configuration that was used in comparison against previous methods. The proposed method significantly outperforms previous methods.

Table 4: Comparison of proposed method against previous SER methods.

Models	Details	Fold	WA (%)	UA (%)
SERAB [39]	BYOL-S/CvT 2048	-	65.10	-
SUPERB[26]	wav2vec 2.0 base		63.43	-
L. Pepino[28]	Dense - Global Normalized		-	66.3 \pm 0.7
L. Pepino[28]	Fusion - Speaker Normalized	5	-	67.2 \pm 1.7
M. Hou [36]	-		-	66.64
W. Fan [4]	-		70.43	65.02
Y. Wang [27]	wav2vec-base (EF/PT)		70.75 / 70.21	-
Y. Wang [27]	HuBERT-base (EF/PT)	10	69.83 / 69.68	-
Y. Tang [34]	Audio Only		71.80	69.60
Proposed method	Four attention blocks with Speaker-identification network (all frozen)	5	72.14	72.97

6 Conclusion

This paper proposed two modules for generating a speaker-specific emotion representation for SER. The emotion classification and speaker-identification networks proposed are both based on the wav2vec 2.0 model. Using the Arcface loss, the networks are trained to respectively generate emotion representations and speaker representations from an input audio waveform. A novel tensor fusion approach was used to

combine these representations into a speaker-specific representation. Employing attention blocks and max-pooling layers improved the performance of the emotion classification network. This was associated with the attention blocks' ability to identify which segments of the generated representation are most relevant to SER. Freezing the entire speaker-identification network trained on the VoxCeleb1 dataset (1,251 speakers) and using four attention blocks in the emotion network provided the best overall performance. This is because of the proposed method's robust generalization capabilities that extend to unseen speakers in the IEMOCAP dataset. The experiment results showed that the proposed approach outperforms previous methods.

Funding Statement: This research was supported by the Chung-Ang University Graduate Research Scholarship in 2021.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [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>
- [2] J. Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.
- [3] L. Marianne and P. Belin. "Human voice perception." *Current biology*, vol. 21, 2011.
- [4] W. Fan, X. Xu, B. Cai and X. Xing, "ISNet: Individual Standardization Network for Speech Emotion Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1803-1814, 2022.
- [5] C. Le Moine, N. Obin, and A. Roebel, "Speaker Attentive Speech Emotion Recognition," in *Interspeech*, pp. 2866-2870, 2021.
- [6] A. Baeovski 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.
- [7] C. Busso et al., "IEMOCAP: interactive emotional dyadic motion capture database," in *Language Resources and Evaluation* vol. 42 pp. 335-359, 2008.
- [8] A. Nagrani, J.S. Chung, A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, pp. 2616-2620, 2017.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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>
- [16] S. Liu et al., "Audio self-supervised learning: A survey," 2022. [Online]. Available: <https://arxiv.org/abs/2203.01205>
- [17] M. Tagliasacchi, B. Gfeller, F. d. C. Quitry and D. Roblek, "Pre-training audio representations with self-supervision," in *IEEE Signal Processing Letters*, vol. 27, pp. 600-604, 2020.
- [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.
- [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.

- [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.
- [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.
- [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>
- [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>
- [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.
- [25] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in *International Conference on Learning Representations*, 2020.
- [26] S. Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech*, pp. 1194–1198, 2021.
- [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>
- [28] L. Pepino, P. Riera, and L. Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in *Proc. Interspeech*, pp. 3400–3404, 2021.
- [29] L. Weiyang et al. “Large-margin softmax loss for convolutional neural networks,” in *International Conference on Machine Learning*, pp. 507–516, 2016.
- [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.
- [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.
- [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.
- [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.
- [34] Y. Tang, Y. Hu, L. He, and H. Huang, “A bimodal network based on audio–text-interactive-attention with arcface loss for speech emotion recognition,” in *Speech Communication*, vol. 143 pp. 21–32, 2022.
- [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.
- [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] 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.
- [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.
- [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.