Heliyon 8 (2022) e09196

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Vector learning representation for generalized speech emotion recognition

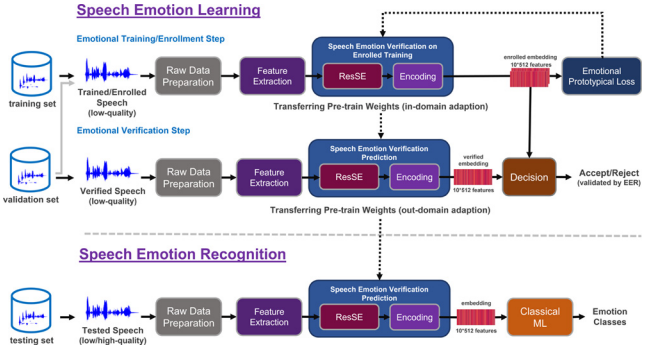
Sattaya Singkul, Kuntpong Woraratpanya *

Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, 1 Chalong Krung, Lat Krabang, 10520, Bangkok, Thailand

HIGHLIGHTS

- A verify-to-classify framework was designed for achieving in generalization and overall performance.
- An implemented verify-to-classify framework can work well in both verification (in-domain) and recognition (out-domain).
- Our softmax with Lo5 can work well with emotion vectors and help improve classification performance.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Speech emotion recognition
Residual squeeze excitation network
Normalized log mel spectrogram
Speech emotion verification
Verify-to-classify framework
Softmax with angular prototypical loss
Cross environment
End-to-end learning

ABSTRACT

Speech emotion recognition (SER) plays an important role in global business today to improve service efficiency. In the literature of SER, many techniques have been using deep learning to extract and learn features. Recently, we have proposed end-to-end learning for a deep residual local feature learning block (DeepResLFLB). The advantages of end-to-end learning are low engineering effort and less hyperparameter tuning. Nevertheless, this learning method is easily to fall into an overfitting problem. Therefore, this paper described the concept of the "verify-to-classify" framework to apply for learning vectors, extracted from feature spaces of emotional information. This framework consists of two important portions: speech emotion learning and recognition. In speech emotion learning, consisting of two steps: speech emotion verification enrolled training and prediction, the residual learning (ResNet) with squeeze-excitation (SE) block was used as a core component of both steps to extract emotional state vectors and build an emotion model by the speech emotion verification enrolled training. Then the in-domain pre-trained weights of the emotion trained model are transferred to the prediction step. As a result of the speech emotion learning, the accepted model—validated by EER—is transferred to the speech emotion recognition in terms of out-domain pre-trained weights, which are ready for classification using a classical ML method. In this manner, a suitable loss function is important to work with emotional vectors. Here, two loss functions were proposed: angular prototypical and softmax with angular prototypical losses. Based on two publicly available datasets: Emo-DB and RAVDESS, both with high- and low-quality environments. The experimental results show that our proposed method can significantly improve generalized performance and explainable emotion results, when evaluated by standard metrics: EER, accuracy, precision, recall, and F1-score.

* Corresponding author.
E-mail address: kuntpong@it.kmitl.ac.th (K. Woraratpanya).

https://doi.org/10.1016/j.heliyon.2022.e09196
Received 9 April 2021; Received in revised form 25 August 2021; Accepted 22 March 2022

1. Introduction

Emotional recognition has evolved from being a niche to an important component of Human-Computer Interaction (HCI), especially in recognition domains of image, text, and speech emotions. Recent techniques involved in image emotion recognition have been published in [1, 2]. The other two emotions, text and speech, are active areas to improve service efficiency in global business today. Even though text and speech emotions are closely relevant, both kinds of emotions have different challenges. One of the challenges in text emotion recognition is ambiguous words owing to omitted words [3, 4]. On the other hand, a challenge in speech emotion recognition (SER) is creating an efficient model which understands its context related to the emotional states. However, this study focuses only on the recognition of speech emotions. Recently, end-to-end learning has been attended in various domains since it has the advantage of low engineering effort and less hyperparameter tuning [5, 6]. As our previous works [7, 8], a deep residual local feature learning block (DeepResLFLB) can be viewed as the end-to-end learning. It was inspired by the concept of human brain learning; that is, 'repeated reading makes learning more effective,' in the same way that Sari [9] and Shanahan [10] were used.

Nevertheless, the end-to-end learning framework can be viewed as a double-edged sword; it provides the advantage of low engineering effort and less hyperparameter tuning, while it cannot provide suitable hyperparameter tuning, thus making a learned model unable to reach the higher performance and meet generalization. Furthermore, the model generated from the end-to-end learning works well with merely high-quality environments; that is, signals are acquired from a high-quality device and a high sampling rate. In practice, the signals can be acquired from various devices, different sampling rates, and uncontrolled environments, even cultural variations. This is a reason why the learned model of end-to-end learning was not well successful. Therefore, this paper proposed a verify-to-classify framework to overcome the limitations of DeepResLFLB, which was formed from end-to-end learning. The concept of verify-to-classify is applying the deep learning to extract features from the space of emotional information prior to fine-tuning on classification tasks. Here, the residual learning (ResNet) [11] with squeeze-excitation (SE) blocks [12] was selected for extracting emotional state vector to measure the proper performance with explainable results, and then, a classical machine learning was used to fine-tune the classification tasks. One more thing, the loss function is important when signal features have a high dimension [13, 14]. Responding to suitable loss functions, here, we reviewed three traditional softmax losses: AM-Softmax [15], AAM-Softmax [16], and Vanilla Softmax [17]. Further, we proposed two loss functions, inspired by the prototypical concept, including angular prototypical loss and softmax with angular prototypical loss functions. For performance assessment, all experiments ran on two publicly available datasets: Berlin emotional database (Emo-DB) [18] and Ryerson audio-visual database (RAVDESS) [19]. Both datasets are speaker-independent SER, provide cultural variations, and include high- and low-quality environments (see Subsection 4.2). These characteristics of both datasets are sufficient to test the performance of the verify-to-classify framework, especially in the cases of cultural variations and different sampling rates. In a word, the implemented verify-to-classify framework can work well in cross environments [20] (see Subsection 4.2).

The main contributions of this paper are as follows:

- A verify-to-classify framework was designed for solving the limitations of the end-to-end learning framework in issues of generalization and performance improvement. (The proof of this contribution can be found in Tables 2 and 3.)
- An implemented verify-to-classify framework can work well in both verification (in-domain) and recognition (out-domain). (The proof of this contribution can be found in Tables 2 and 3.)
- Our softmax with angular prototypical loss function (Lo5) can work well with emotion vectors and help improve classification performance. (The proof of this contribution can be found in Tables 2 and 3 and Figs. 8, 9, 10, and 11.)
- A verify-to-classify framework was designed based on a compromise of performance and resource consuming to achieve higher accuracy, faster processing time, and better overall performance. (The proof of this contribution can be found in Table 1.)

The remainder of this paper is structured as follows. In Section 2, various existing methods introduced to SER are reviewed. Section 3 provides details of the proposed method. Section 4 describes parameter settings and characteristics of benchmark datasets for SER. In Section 5, we report experimental results with discussion. In Section 6, the conclusion and future work are given.

2. Literature review

On the influence factors of SER performance, three main factors are considered on the training of deep learning: size of training sets, feature extractions, and model learning methods. We briefly describe the data augmentation, involved in increasing the size of training sets, feature selections, and speech emotion models associated with the evolution of SER. Lastly, we close this section with the learning with loss criterion.

2.1. Data augmentation

The effectiveness of deep learning relies upon the amount of data used for training [21, 22]. Google brain research [23] was recently proposed as one of the efficient techniques for increasing the amount of data by adding spectrogram characteristics, referred to as "Spectrogram Augmentation". The approach comprises of time warping to observe more time-shifting patterns; time masking to weaken the model's overfitting rate and enhance the sound tolerance that has silence characteristics; and frequency masking to reduce the overfitting rate and enhance sound resistance, when a particular wavelength has concealing characteristics. In order to explore more aspects of speech information, using spectrogram augmentation as a basic feature could be led to various specific features, allowing the model to learn more perspectives.

2.2. Feature selection

The different features lead to the different performances in SER. Typically, speech signals [24] contain linguistic and paralinguistic information. The linguistic information refers to the language and context of the speech, whereas the paralinguistic information [25, 26, 27, 28] refers to emotional information of the speech. However, the most important thing to remember is that the sounds produced by a human are filtered by the shape of the vocal tract, which includes tongue, teeth, and others [29]. If the sound that emerges from this shape is determined precisely, we should be able to accurately represent the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short-time power spectrum, thus Mel-frequency cepstral coefficient (MFCC) [29, 30] is accurately represented this envelope. The MFCC feature can be characterized by frequency filters in the range of 20 Hz to 20 kHz, which is relevant to human hearing, is widely used to obtain coefficients from the filtered sound. Recent research papers [30, 31] used the difference of MFCC to get more specific details, but, in the aspect of MFCC, it has no time relationship and no full frequency component details. This is a reason why many papers [8, 30, 32, 33] used Mel spectrogram instead. The Mel spectrogram can respond to the time relationship with fewer loss frequency component details, thus providing better results than the MFCC alone.

2.3. Speech emotion model

In modern SER, model efficiency can be increased by improving learning methods of machine learning or deep learning. Demircan [31] attempted to improve the model efficiency by applying fuzzy C-mean as a preprocessing step before using classical machine learning to add characteristics to a group. Venkataramanan [30] investigated the model efficiency of deep learning and machine learning, according to the study's findings, the deep learning by means of CNN outperformed the traditional machine learning. Also, Zhao [32] introduced the use of CNN in conjunction with LSTM to extract and learn features. Zhao's method used a local feature learning block (LFLB) [32], consisting of CNN, batch normalization (BN), activation function, and pooling, for local feature learning, and then applied LSTM in extracting contextual dependencies in a time-related relationship for global feature learning. In this way, the model can learn both local and global features.

However, the LFLB, used for local feature learning introduced in Zhao's method, still has a room for efficiency improvement. Recently, the improved version of LFLB, called a deep residual local feature learning block (DeepResLFLB) [7, 8], was proposed. One of the DeepResLFLB's achievements is that reducing the feature and updated losses was caused by the CNN model in the conventional LFLB, especially in deeper layers. This showed that the DeepResLFLB, based on the ResNet concept [11, 34] as repeated learning style [10], can extract local features from complex speech patterns and learn more effectively using a residual deep learning approach.

Here, a brief concept of ResNet is described. ResNet presented stacking additional layers in the deep neural networks, which result in improving performance. The intuition behind adding more layers is that these layers progressively learn complex features. A ResNet concept provides a direct connection between layers and skips some layers in between. This connection, called "skip connection," is the core of residual blocks. Due to this skip connection, the output layer is not the same as without a skip connection. Thus, the input data is multiplied by the layer's weights, and then a bias term is added. In addition, ResNet determines a suitable pattern from the identity mapping by using re-sequency technique [34] and uses compression/decompression strategy of bottleneck design [8, 35] for its performance improvement.

To increasingly improve the model efficiency, many research works have been investigated the squeeze-excitation (SE) network [12]. SE can learn by considering channel-wise as each feature detail to generate weights to represent the important features. In encoding feature learning level, the attentive pooling [36] is used for converting the frame-level to utterance-level; that is, in the same way on local to global level features. The attentive statistical pooling (ASP) [37] and self-attentive pooling (SAP) [38] are very well-known on speaker verification tasks [36, 39].

Furthermore, Kumar [40] presented an end-to-end triplet loss-based emotion embedding system for speech emotion recognition (TL-EESER), which was based on the ResNet concept as described above. The learned embedding layer was used to recognize the emotions by providing speech samples in various lengths. Then, the model was trained using softmax pre-train and triplet loss function. The weights between the fully connected and embedding layers of the trained network were used to calculate the embedding values. These embedding values, being viewed as angles of the cosine function, were utilized to classify a new speech sample into an appropriate emotion class.

2.4. Learning with loss criterion

One of the critical problems in an end-to-end approach is to explore the suitable criteria (loss functions) for driving a network to learn discriminative features. Typically, softmax loss was used for this purpose. However, it was more suitable for classification tasks [41] but not suitable for verification tasks. Therefore, to solve this problem [42, 43, 44], several different loss functions have been proposed for verification. In contrast to classification, verification is an open-set task, that classes observed in the training set will generally not appear in the test set. To achieve this circumstance, a good verification loss should make inter-class variances larger and intra-class variances smaller. The followings are loss functions that were proposed for solving inter-class and intra-class problems. AM-Softmax [15] solved the inter-class and intra-class problems by using a decision boundary technique with added margin. The decision boundary was created to separate the inter-classes and then a margin was added to that decision boundary. This technique can increase the separability of classes while also making the distance between the same classes closer. In addition, the additive angular margin softmax (AAM-Softmax) method [16] was used to switch from a distance to an angular objective to learn highly discriminative features for improving the efficiency of high dimensional features in face recognition robustness task.

Angular softmax loss [41] has recently been proposed to improve the softmax loss in face verification using normal angular objective. It enables end-to-end training of neural networks to learn angularly discriminative features. Angular softmax loss introduced a margin between the target class and the non-target class as the softmax loss. Compared with the triplet loss [45], angular softmax is much easier in tuning and monitoring hyperparameters [41].

Also, a learning method is an important process for both classification and verification. Here is an example of the learning method challenge. In the case of learning slightly different characteristics of speech, the traditional learning method could not make it discriminative, especially in the case of class imbalance. In this case, the prototypical network based on few-shot learning [46] could work better than the traditional ones. To be more specific, the few-shot learning is a process of feeding a learning model with a relatively small quantity of training data, as opposed to the common methods of using a large amount of data. This capability of few-shot learning enables prototypical networks to learn a metric space in which classification can be performed by calculating a distance between prototype representations of each class. This learning method makes a model more generalization for new classes that have never been seen in the training set, especially a small number of samples in each new class like a support set. The prototypical networks represent a simpler inductive bias from an increased bias by a query of support set; this is useful and produces outstanding results for the limited data. This kind of learning methods is commonly used in computer vision. One reason is that using an object categorization model, without using multiple training samples, still produces satisfaction results.

Our work is different from the previously mentioned works in that the verify-to-classify framework was designed to improve DeepResLFLB. Our method supports ResNet in conjunction with SE to produce more information, which is changed from frame-level to utterance-level in feature encoding. Moreover, the implemented verify-to-classify framework used many criteria to explore critical problems. For instance, one of them is the angular prototypical criterion that is used instead of the normal angular objective criteria, when considering the slightly different characteristics of enrolled feature vectors before applying to classification tasks.

3. Methodology

To enable SER as efficiently as possible, a verify-to-classify framework was proposed to overcome limitations, generalization, and explainable results, of an end-to-end learning framework. The proposed framework consists of four parts: (i) raw data preparation, (ii) feature extraction, (iii) speech emotion verification, and (iv) fine-tuning probe with classical machine learning, as shown in Fig. 1.

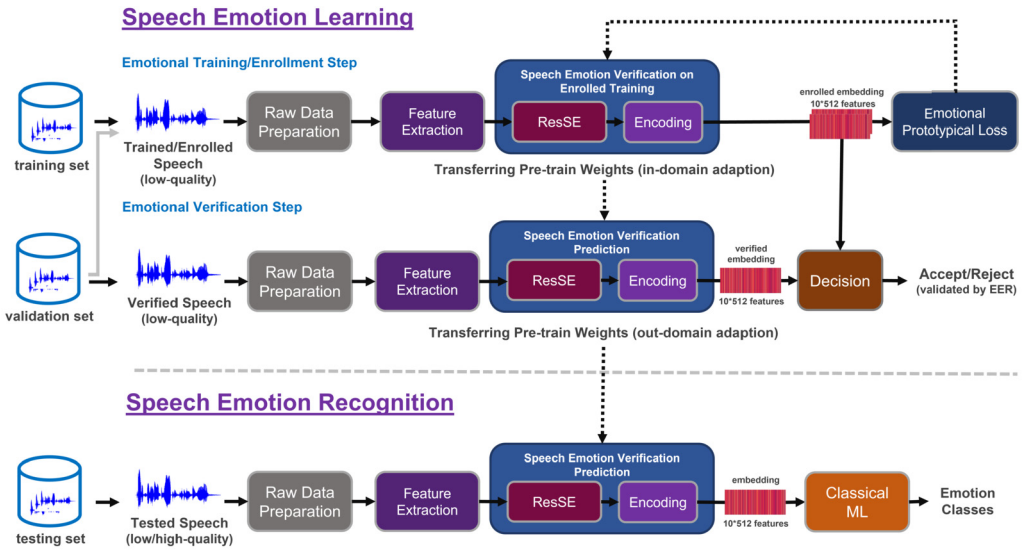


Fig. 1. A verify-to-classify framework with transfer learning for SER. Note that the gray line represents a sample of validation set by randomly selecting one emotion speech per speaker for enrolled speech.

3.1. Raw data preparation

Cultural variations and different speaking rates [47] are influence factors that can prove the efficiency of the proposed framework; therefore, selecting challenging datasets is essential. Fortunately, two publicly available datasets, Emo-DB and RAVDESS, match our requirements. An Emo-DB in Berlin German is fast vocalization, while a RAVDESS in English is normal vocalization. These datasets are also challenging in different sampling rates, which typically are the causes of difficulties in recognizing the emotional states of a speaker. Furthermore, increasing the variety of data to discover more new dimensions or characteristics for training a model based on deep learning is essential [21]. Responding to this, various data augmentation techniques using spectrogram [23] were used in this work to make the model more robust to noise and unseen voice patterns.

3.2. Feature extraction

It is undeniable that the effectiveness of deep learning mainly depends on feature factors before learning steps. Also, different features lead to different performances in speech emotion recognition. Among the features of speech, Mel-frequency cepstral coefficient (MFCC) [30], which can be characterized by the frequency filter in the range of 20 Hz to 20 kHz, similar to human hearing, is widely used to obtain coefficients from the filtered sound. Recent research papers [30, 31] used the difference of MFCC to get more specific details, but, in the aspect of MFCC, it has no time relationship and no full frequency component details. Therefore, many papers [30, 32, 33] used Mel-spectrogram instead. The Mel-spectrogram can respond to the time relationship with less loss frequency component details, thus providing better results than just using MFCC. Besides, the logarithm transformation was used to normalize the Mel-frequency component to represent the little differential details clearly, that is the log Mel-spectrogram (LMS).

In the real-world, speech has different speaking durations that are difficult to define a padding size for LMS. To avoid this issue, the LMS is chopped into smaller chunks based on a maximum length. If the chunk size is less than the maximum length, the repeated spectrogram padding is used. Each chunk (γ) is normalized by z-score [48, 49] as expressed in (1).

$$\gamma_{norm} = \frac{\gamma - \mu_{chunk}}{\sigma_{chunk}} \tag{1}$$

where μ_{chunk} and σ_{chunk} are the chunk mean and standard deviation, respectively. LMS chunks that are normalized are called normalized log Mel-spectrogram (N-LMS). In this paper, we use N-LMS as features for learning the emotional state of the frequency component for vector extraction.

3.3. Speech emotion verification

Speech emotion verification is the main contribution of this study. We designed a method for extracting an emotion vector as an emotional voiceprint in the same way as biometric voiceprint, all of which were inspired by speaker verification methods [36, 39, 43, 50, 51]. Recently, speaker vector extraction, called “Speaker Verification,” was widely used. Speaker verification is a method that first verifies a speaker from the embedded speaker information, and then measures similarity between two vectors to verify the true speaker. Here, a speaker verification concept was implemented and embedded into the learning model of SER, so that it can create the emotional voiceprint of a vector verification term. As a result, emotional information was embedded into features before the step of fine-tuning in the classification task. In addition, our proposed speech emotion verification method requires transfer learning [52, 53] to adapt embedded features in low-quality environments to predict in cross environments as shown in Fig. 1.

In contrast, Fig. 2(b) shows two principal steps of how speech emotion verification works. The emotional training/enrollment step mainly trains the incoming audio of speakers to create an emotion model that contains emotional voiceprints. This step is different from the traditional speaker verification as shown in Fig. 2(a); that is, both enrolled speech (a gray line in Fig. 2(b)) and trained speech are used in the training process concurrently as shown in Fig. 2(b), so that the model can enroll information from the validation set. Then, the emotional trained model is transferred to the emotional verification step in terms of in-domain pre-train weights. The results of enrolled and verified embeddings obtained between the training set and the validation set were compared with equal error rate (EER). At this stage, the lowest EER value from the decision block indicates that we meet the best model in voiceprint. In this way, we can achieve more efficient in emotional tasks.

In general, end-to-end learning is difficult to find suitable tuning parameters, making it easy to reach overfitting. A simple way to solve this problem is to divide the speech emotion learning into three parts: ResSE model, encoding frame-level to utterance-level features, and emotional prototypical loss, as shown in Fig. 1.

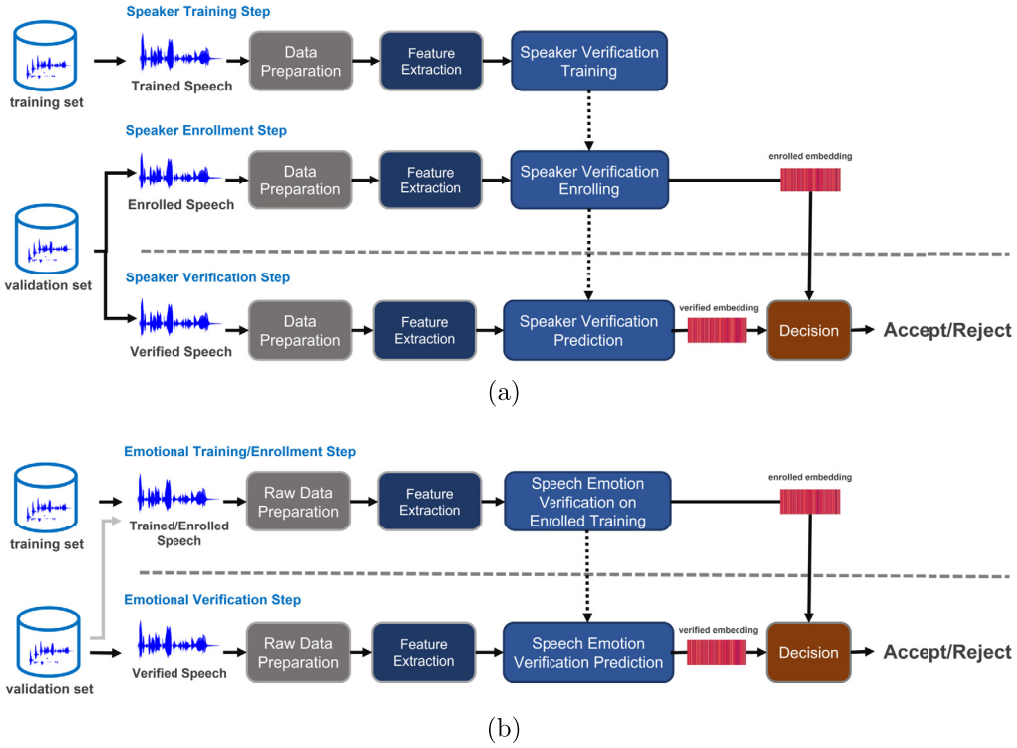


Fig. 2. A development of emotion speech verification: (a) traditional speaker verification. (b) speech emotion verification in verify-to-classify framework. Note that the gray line in (b) is a sample of validation set by randomly selecting one emotion speech signal per speaker for the enrolled speech.

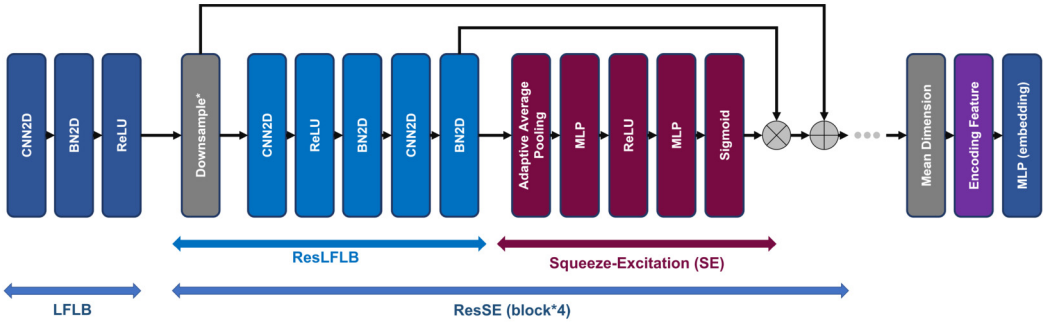


Fig. 3. A residual network with squeeze-excitation (ResSE) block structure.

3.3.1. ResSE model

Speech emotion vector extracted from N-LMS provides to ResNet with squeeze-excitation (SE) blocks [12] to learn the emotion information. As described in [8, 11, 34], ResNet can learn and reconstruct speech emotion information on deeper layers by adding a function with skipping-connection. The ResNet also can reduce unnecessary parameters and time-consuming by skipping unnecessary learning layers when the gradient is nearly at zero. Fig. 3 shows how to apply ResNet as LFLB and ResLFLB concept as DeepResLFLB [7, 8]. Here, SE blocks were applied to consider the weight of each learning step, which can describe important information within features.

3.3.2. Encoding frame-level to utterance-level features

Typically, ResNet with SE blocks focuses on the learning of the local feature level, not including the global feature level. This causes information loss in the relationship between features. To avoid this information loss, the attentive pooling concept is applied to convert the local feature level to the global feature level. Here, the attentive statistics pooling (ASP) [37] and the self-attentive pooling (SAP) [38] were used for this purpose.

Self-Attentive Pooling (SAP). We implemented an encoding layer with SAP, similar to [38, 54, 55]; that is, we first feed utterance level features $\mathbf{x_t} = \{x_1, x_2, x_3, ..., x_T\}$ into a multi-layer perceptron (MLP) to get $\mathbf{h_t} = \{h_1, h_2, h_3, ..., h_T\}$ as a hidden representation on frame-level features. In this paper, we simply adopt a one-layer perceptron, as expressed in (2)

$$\mathbf{h_t} = tanh(\mathbf{W}\mathbf{x_t} + b) \tag{2}$$

where $\mathbf{W}$ and b are the weight matrix and bias of the network, respectively. At this point, we can calculate the normalized weighted mean, ω_t, as expressed in (3), of each frame, so that the frame with the highest value is selected as the important frame.

$$\omega_t = \frac{exp(\mathbf{h_t^T}u)}{\sum_{t=1}^{T} exp(\mathbf{h_t^T}u)} \tag{3}$$

where u is the learnable context vector. u can be viewed as a high-level representation. It is randomly initialized and jointly learned during the training process. Lastly, the utterance-level representation, $\tilde{\mu}$, as denoted in (4), can be determined as a weighted sum of the frame level of ResSE feature maps based on the learned weights.

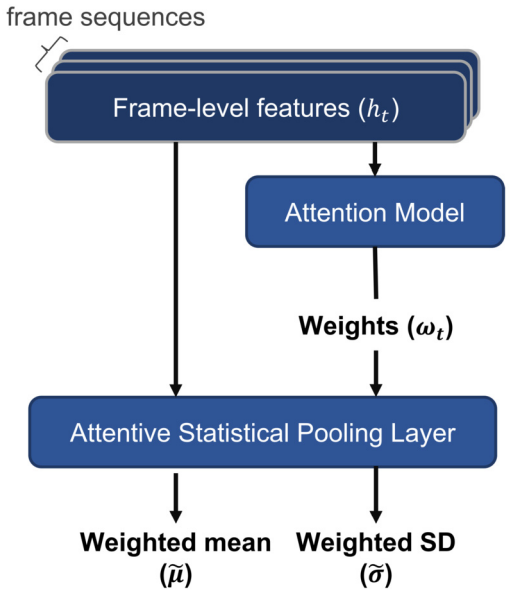


Fig. 4. An implemented structure of attentive statistical pooling (ASP).

$$\tilde{\mu} = \sum_{t=1}^{T} \omega_t x_t \tag{4}$$

Attentive Statistical Pooling (ASP). We applied ASP to an encoding layer to extract frame-level features to weighted mean and weighted standard deviation as illustrated in Fig. 4. Here, the weighted mean is calculated by (4) and the weighted standard deviation is determined by (5).

$$\tilde{\sigma} = \sqrt{\sum_{t=1}^{T} \omega_t \mathbf{h_t} \odot \mathbf{h_t} - \tilde{\mu} \odot \tilde{\mu}} \tag{5}$$

where $\odot$ denotes the Hadamard product. The mean vector $\tilde{\mu}$, which aggregates frame-level features, can be viewed as the main component of utterance-level features. Besides, the weight ω_t calculated by (3) is used in both the weighted mean $\tilde{\mu}$ and weighted standard deviation $\tilde{\sigma}$, as shown in Fig. 4. The weighted standard deviation is thought to take the advantage of both statistical pooling and attention, i.e., feature representation in terms of long-term variations and frame selection in accordance with importance, bringing higher emotion discriminability to utterance-level features. As (5) is differentiable, ResSE with ASP can be trained on the basis of back-propagation.

3.3.3. Emotional prototypical loss

Here, an emotional prototypical loss in Fig. 1 is used to measure the feature vector performance at the final tuning stage of learning. The loss can be calculated from various methods, such as Siamese network [56, 57, 58]. Typically, for vector loss, Euclidean distance [59] was widely used for similarity measure between two vectors [56, 57, 58] in a feature space. However, the Euclidean distance has some limitations, when the feature space has a high dimension [13, 14, 60]. To avoid this problem, angular loss, using cosine function in backend of speaker verification concept [60], was used instead of Euclidean. In addition, we reviewed three traditional softmax losses: AM-Softmax loss (Lo1) [15], AAM-Softmax loss (Lo2) [16], and softmax loss (Lo3). Also, we proposed two loss functions by applying a prototypical concept to an angular objective, namely, angular prototypical loss (Lo4) and softmax with angular prototypical loss (Lo5), to obtain suitable losses in the speech emotion domain.

Lo4 and Lo5 were derived from Lo3, formulated as (6).

$$Lo3 = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\mathbf{W}_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^{C} e^{\mathbf{W}_j^T \mathbf{x}_i + b_j}} \tag{6}$$

where $\mathbf{W}$ and b are the weight and bias of the last layer of the trunk architecture, respectively.

Lo3 consists of a softmax function followed by a multi-class cross-entropy loss. This loss function works well in measuring overall classification error but does not explicitly enforce intra-class compactness and inter-class separation. This becomes our inspiration to extend its learning ability for other circumstances. Therefore, this paper proposed two losses: Lo4 and Lo5, focusing on the intra-class compactness and inter-class separation. The prototypical network concept was applied to Lo4 and Lo5 to expand learning abilities on slightly different speech characteristics with class imbalance.

Angular prototypical (Lo4). Lo4 is a variant of the prototypical networks with an angular objective. For simplicity, we assume that a mini-batch, viewed as training samples, contains a support set S and a query set Q. The query includes M-th utterance from every emotion. With this assumption, the prototype (or centroid) can be determined by (7).

$$\mathbf{c_j} = \frac{1}{M-1}\sum_{m=1}^{M-1} \mathbf{x_{j,m}} \tag{7}$$

where $\mathbf{x}$ is the j-th utterance level feature of each emotion. To calculate similarity, the angular prototypical objective, $\mathbf{S_{j,k}}$, where k is a query of emotion, expressed in (8), is used as the distance metric, instead of squared Euclidean distance.

$$\mathbf{S_{j,k}} = \omega \cdot cos(\mathbf{x_{j,M}}, \mathbf{c_k}) + b \tag{8}$$

In training process, each query sample is classified against N emotions based on Lo4 as defined by (9).

$$Lo4 = -\frac{1}{N}\sum_{j=1}^{N} log \frac{e^{\mathbf{S_{j,j}}}}{\sum_{k=1}^{N} e^{\mathbf{S_{j,k}}}} \tag{9}$$

where, $S_{j,j}$ is the cosine backend distance between the query and the prototype of the same emotion from the support set. The softmax function effectively serves the purpose of hard negative mining [61], since the hardest negative sample would mostly affect the gradients. The value of M is typically chosen to match the expected situation at test-time, e.g. $M = 6 + 1$ for 6-shot learning, so the prototype is composed of six different utterances. In this way, the task in training exactly matches the task in the testing scenario.

Softmax with Angular prototypical loss (Lo5). Lo5, inspired by both information criteria of emotion embedding similarity and emotion class distribution, is a combination of Lo3 and Lo4. Lo5 is computed from embedding and softmax layer results, as shown in Fig. 5.

3.4. Fine-tuning probe with classical machine learning

An emotion vector is an explainable feature, but it has no consistency on classification tasks. Thus, classical machine learning methods are used for fine-tuning and learning in the last step, while weights of speech emotion verification model are frozen for mapping to different tasks. One more thing, since the input data are small and not complex, using classical machine learning can perform better than deep learning methods [62]. Here, we focus on an applied classification task, thus a support vector machine (SVM) [63] and multilayer perceptron (MLP) are suitable for selecting to fine-tune in learning features and creating a multiclass classification model.

4. Experiments

The proposed verify-to-classify framework was evaluated in terms of (i) generalized SER performance and (ii) explainable emotion results. The followings describe components of experimental design.

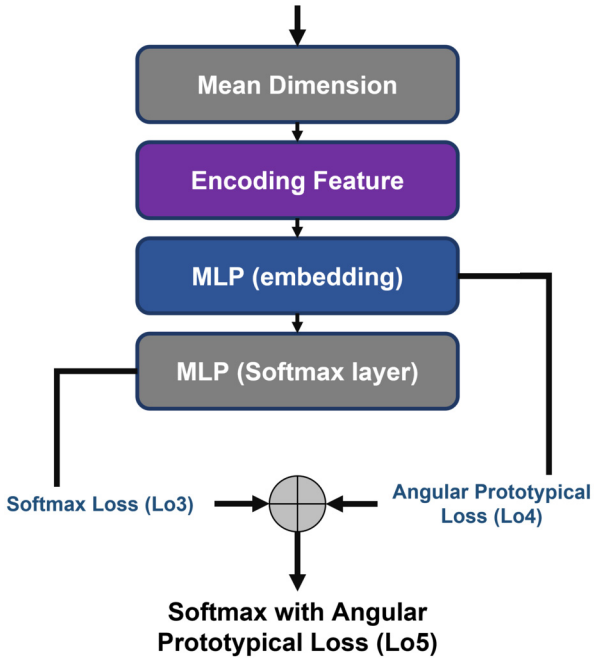


Fig. 5. A computational step of softmax with angular prototypical loss.

4.1. Evaluation metrics

In evaluating SER performance, two metrics were used, equal-error-rate (EER) and normal classification measurement. The EER metric [64, 65, 66] uses a threshold value to predict false acceptance and rejection rates. If the false acceptance and rejection rates are equal, it implies that the percentage of those two rates is balanced, which is called the equal error rate. The lower the equal error rate value means the higher overall accuracy. We used this criterion for evaluating balanced false rates of speech emotion vector in speech emotion learning. Then, we used normal classification measurement for measuring multiclass SER performance in the last step, including macro average accuracy, precision, recall, and F1-score as defined by (10), (11), (12), and (13), respectively. Also, both metrics can be used to prove the factor of cultural variation and sampling rate that directly affect SER performance, based on different quality experiment settings.

For explainable results, the emotion vector was applied principal component analysis (PCA) [67] first to reduce the data dimensionality to 50 dimensions for visualization [68, 69] and then followed by t-distributed stochastic neighbor embedding (t-SNE) [70] to reduce PCA coefficients again to 2 dimensions. After that, emotion vector centroids are calculated from mean of each emotion-vector class in the same way of centroid calculation of prototypical loss in (7).

The visualized vector results contain representative point patterns, which can be interpreted as the efficient prototype of vectors and clusters. However, our proposed speech emotion verification methods focus on transfer learning [52] as well as the adaptive low-quality feature spaces for cross-environment prediction. In this case, if the vector distributions of before and after taking transfer learning still provide correct classification and clearly separate a centroid of each emotion for inter-class separation, we can say that the feature spaces are strongly generalized.

$$Accuracy = \frac{true\ positive + true\ negative}{number\ of\ data} \tag{10}$$

$$Precision = \frac{true\ positive}{total\ predicted\ positive} \tag{11}$$

$$Recall = \frac{true\ positive}{total\ actual\ positive} \tag{12}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{13}$$

4.2. Datasets and simulated environments

We used two publicly available datasets: Berlin emotional database (Emo-DB) [18] and Ryerson audio-visual database (RAVDESS) [19] to evaluate SER performance of ours and baseline methods. An Emo-DB speech corpus consists of seven acted emotional states: anger, disgust, boredom, joy, sadness, neutral, and fear. It includes 535 emotional utterances in German from ten native German actors. Five of them are women while the other five are men. The audio files are a 16-bit resolution and have a sampling frequency of 16 kHz. The average length of the audio files is 3 seconds. The emotion distribution is shown in Fig. 6.

Another dataset, RAVDESS, contains eight emotional states: anger, disgust, calm, surprise, sadness, neutral, fear, and happiness. The emotional utterances of 24 professional actors, 12 of whom are female and 12 of whom are male, in North American accents, were recorded. The audio files have a sampling rate of 48 kHz and a resolution of 16-bit. The emotion distribution is presented in Fig. 7.

To prove the factor of sampling rates, we focused on two contrast environments, namely, the high- and low-quality environments. The high-quality environment was acquired from high-quality resources, such as speech from movies and studios. All of which were recorded by high-quality recording equipment. On the other hand, the low-quality environment was acquired from low-quality resources, such as speech from call center services and smartphones. Here, we resample the original speech signals with two different sampling rates, 16 kHz and 48 kHz, that represent low- and high-quality environments. Note that upsampling speech signals are not increased the quality of those signals. Based on cross-environment adaption [20], in our study, a model is trained by the low-quality environment only, and then it is tested by the high-quality environment. If the model can yield successful results, we can say that this model provides generalization on the different sampling rates or cross environment.

We set up two experiments, high- and low-quality environments, to prove the factor of sampling rates as follows: for the high-quality environment, the sampling rate was resampled to 48 kHz with 16-bit PCM resolution. For the low-quality environment, the sampling rate was resampled to 16 kHz with 16-bit PCM resolution. Besides, both setup datasets were augmented by the spectrogram augmentation method to create more perspectives of data for deep learning and were divided into three speaker independent subsets for 10-fold cross-validation: 80% for the training set, 10% for the validation set, and 10% for the testing set. It means that there is no overlapping speaker information in-between subsets. As a result, the model focuses on transferred emotion information based on different speaker information.

4.3. Parameter settings on resource-consuming concerns

The verify-to-classify framework was proposed for generalized SER performance and explainable emotion results. Typically, reaching the highest SER performance [8] demands high resource-consuming. In this paper, the purpose is to compromise between performance and resource consumption. Therefore, we tested our implemented framework in generalized SER performance with three situations and three models: (i) low resource consuming on speed residual with a squeeze-excitation network (SpeedResSE), (ii) normal resource consuming on VGG-M-40 [39, 71], and (iii) high resource-consuming on performed residual with a squeeze-excitation network (PerformResSE). The number of parameters is reported in Table 1.

SpeedResSE has been proposed in the same architecture as fast ResNet-34 [39] and additionally applied SE concept [12] to channel-wise features to reduce computing time and keep more feature information requirements.

VGG-M-40 [39, 71] has been previously proposed for image classification [71] and adapted for speaker recognition [72]. This model is well-known for its high efficiency and good classification performance. VGG-M-40 is a modification of the model proposed by [72] to take

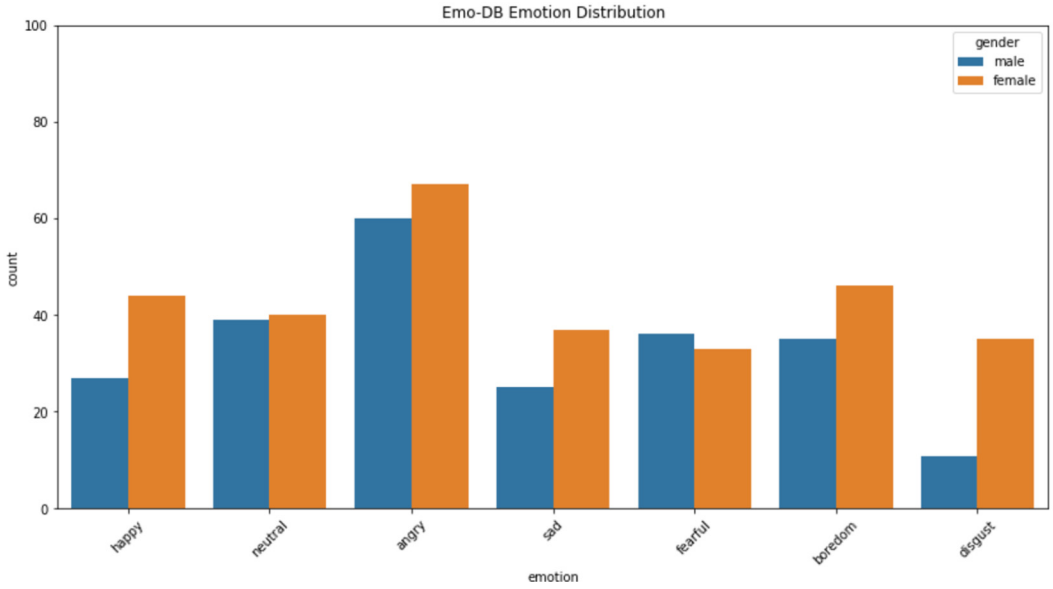


Fig. 6. A speech emotion distribution of Emo-DB dataset.

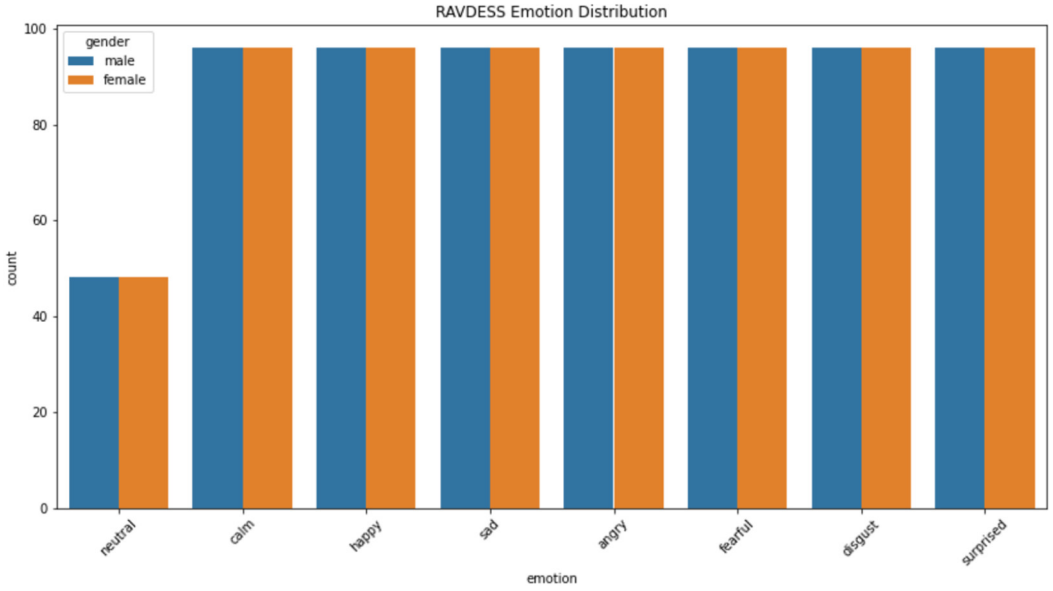


Fig. 7. A speech emotion distribution of RAVDESS dataset.

Table 1. A number of parameters of speech emotion verification.

Method	No. of Parameters
SpeedResSE: SAP encoder	1.4M
SpeedResSE: ASP encoder	1.5M
VGG-M-40: SAP encoder	4.0M
VGG-M-40: ASP encoder	4.2M
PerformResSE: SAP encoder	6.4M
PerformResSE: ASP encoder	7.0M

40-dimensional filterbanks as inputs instead of the 513-dimensional spectrogram.

PerformResSE has been proposed in the same architecture as ResNet-34 [39] and additionally applied SE concept [12] to channel-wise features to keep better feature information requirements.

Parameter settings of overall speech emotion verification learning include max epoch at 500, batch size at 12, learning rate at 0.001 with step learning rate scheduler, optimizer is Adam, the margin is 0.3 and scale is 30 for AM-Softmax and AAM-Softmax, and the number of utterances per emotion per batch is 2 for prototypical loss.

After training on test models of speech emotion verification, we fed the emotional vector to SVM with RBF kernel or MLP with 512 neurons for fine-tuning on classification tasks. All experiments ran on the desktop computer with CPU Core i7-6700 and graphic card of Nvidia GeForce 1050Ti.

5. Results and discussion

Generalized performances and explainable emotion results are our main objectives as formerly stated in Section 4. In this section, we will explore, analyze, and later discuss the results.

In the first objective, generalized performance, two metrics—EER and normal classification measurements, including accuracy and F1-score—were used to evaluate the model performance. In EER measure-

Table 2. A comparison of baselines and our implemented verify-to-classify framework in low-quality environment.

Model	Encoder	EER	
		Emo-DB	RAVDESS
1D-LFLB	-	51.65	51.20
2D-LFLB	-	50.46	51.15
DeepResLFLB	-	50.59	50.02
TL-EESER	-	33.32	31.24
VGG-M-40 (Lo4)	SAP	9.00	3.79
VGG-M-40 (Lo4)	ASP	8.84	5.13
SpeedResSE (Lo4)	SAP	9.48	6.08
SpeedResSE (Lo4)	ASP	11.12	5.46
PerformResSE (Lo4)	SAP	8.05	3.63
PerformResSE (Lo4)	ASP	**6.50**	**3.48**
VGG-M-40 (Lo5)	SAP	7.94	3.27
VGG-M-40 (Lo5)	ASP	8.84	3.99
SpeedResSE (Lo5)	SAP	9.64	3.36
SpeedResSE (Lo5)	ASP	6.54	**1.93**
PerformResSE (Lo5)	SAP	8.67	2.46
PerformResSE (Lo5)	ASP	**3.44**	3.80

ment, the results show that the implemented proposed framework, as already described in Subsection 4.3, outperforms baselines as shown in Table 2. For testing on the Emo-DB dataset, PerformResSE-Lo4 with ASP has the lowest EER at 6.50% and 3.44% for PerformResSE-Lo5 with ASP. On the other hand, for testing on the RAVDESS dataset, the lowest EER at 3.48% for PerformResSE-Lo4 with ASP, and 1.93% for SpeedResSE-Lo5 with ASP.

In normal classification measurement, accuracy and F1-score were taken into consideration as well. In contrast to the EER measurement, which evaluates only vectors, the normal classification measurement assesses the whole framework on classification tasks. We compared ours with baseline models as shown in Table 3. We discovered that in the same environments—low-quality, PerformResSE-Lo5 with ASP significantly outperforms the baseline models. PerformResSE-Lo5 with ASP is the best performance at 92.76% of accuracy and 90.14% of F1-score on Emo-DB, and 88.83% of accuracy and 87.52% of F1-score on RAVDESS. In testing in a cross environment, from low quality to high quality, in the same as previously mentioned, our proposed framework has the best performance than baselines. In the case of cross environment, all test models were trained with the low-quality environment, PerformResSE-Lo5 with ASP is less decrease in accuracy and F1-score around 1%, but baselines are much decrease in accuracy and F1-score as reported in Table 3.

In the second objective, explainable emotion results, we selected sample vectors that have the best EER results for each dataset and compared vector distributions derived from low- and high-quality environments. The distributed emotion vector can be seen clearly as graphically shown in Figs. 8 and 10. With transfer learning that adapts feature spaces from low- to high-quality environments, the results look promising as can be seen in Figs. 9 and 11. The emotion distribution shows clearly inter-class separation. The grouping clusters together distinctly so that a line can be drawn to show the separation of the speaker genders.

In all test models, we experimented with five losses and reported all experimental results in Supplementary_Tables. We discovered that the proposed softmax with angular prototypical loss (Lo5) gives the best performance. A reason behind this is that Lo5 considers more loss information from softmax probabilities of emotion distribution and angular prototypical objective works well on few-shot learning. Moreover, SVM and MLP were experimented for fine-tuning in a classification task. When testing on the same environment, SVM is the best performance on RAVDESS and MLP is the best on Emo-DB. Besides, MLP is the best one, when testing on a cross environment. It provides less decrease of accuracy and F1-score than SVM. That is a reason why MLP is selected for the main discussion.

Further experimental results, including the five loss criteria with encoder and all classification measurements are reported in Supplementary_Tables. With the results explored above, they exhibit that our implemented verify-to-classify framework outperforms the baseline models not only in terms of more explainable emotion results but also high generalized performance.

6. Conclusion and future work

This paper has described a verify-to-classify framework for generalized speech emotion recognition. The purpose of this framework is to overcome the limitations of end-to-end learning to meet generalized performance and explainable emotion results. The verify-to-classify framework was designed based on the vector learning concept for extracting features and combining learning and recognition of speech emotions. The proposed framework was implemented as follows: speech emotion verification used ResSE for extracting feature spaces in emotion with gender domain. The fine-tuning probe on classification tasks used SVM and MLP as classifiers. In improving the model efficiency, five losses, including three existing loss functions and our two prototypical loss functions, have been investigated for comparison to seek the suitable one. The performance of our model was tested on two different cultural variation datasets, German on Emo-DB and English on RAVDESS.

Based on our experiments in cross environments, first, on testing of speech emotion verification, the results show that the PerformResSE-Lo5 with ASP is the best of average EER on RAVDESS and Emo-DB, which were significantly better than VGG-M-40 and baselines. This vector space has well performed in cross environments; that is, the model was trained with a low-quality environment and was tested with a high-quality environment. Second, on testing of classification performance, the results show that the implemented framework with the verify-to-classify concept can perform better than baselines and provide more generalization when evaluated by macro average accuracy, precision, recall, and F1-score.

Although our verify-to-classify framework has provided better performance in speech emotion recognition, many aspects can still be improved, especially generalization from more cultural variation speakers with various environments. In future work, we look forward to extending our concept to support multilingual models, so that they can work with the cultural variation on various speaking languages.

Declarations

Author contribution statement

Sattaya Singkul: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. Kuntpong Woraratpanya: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Table 3. A comparison of baselines and our implemented verify-to-classify framework in cross environments.

Model	Encoder	Emo-DB				RAVDESS			
		low-quality		high-quality		low-quality		high-quality	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
1D-LFLB	-	77.59	76.95	65.54	65.00	75.77	75.14	64.30	63.77
2D-LFLB	-	82.69	82.23	66.69	66.39	75.47	75.14	61.79	61.52
DeepResLFLB	-	82.80	82.19	68.84	68.33	83.24	82.41	69.34	64.90
TL-EESER	-	84.58	81.53	83.96	79.40	87.85	86.78	85.07	84.10
VGG-M-40 (Lo4)	SAP	80.14	74.59	79.35	73.64	81.10	80.04	80.46	79.39
VGG-M-40 (Lo4)	ASP	85.53	81.53	82.80	80.84	81.58	80.05	80.77	79.19
SpeedResSE (Lo4)	SAP	79.44	73.83	78.78	73.04	79.24	78.45	78.55	77.74
SpeedResSE (Lo4)	ASP	77.57	74.92	77.10	74.44	80.19	78.39	79.52	77.69
PerformResSE (Lo4)	SAP	83.64	82.96	83.08	82.50	88.40	87.79	87.82	87.20
PerformResSE (Lo4)	ASP	79.56	73.96	78.88	73.25	84.23	82.81	83.73	82.31
VGG-M-40 (Lo5)	SAP	80.02	75.37	79.44	74.73	83.80	82.09	83.17	81.42
VGG-M-40 (Lo5)	ASP	83.06	78.33	82.43	77.67	84.58	83.60	84.00	83.00
SpeedResSE (Lo5)	SAP	80.37	78.83	79.72	78.18	82.88	80.82	82.16	80.02
SpeedResSE (Lo5)	ASP	83.88	82.11	83.27	81.54	84.66	83.42	84.21	82.96
PerformResSE (Lo5)	SAP	85.51	84.01	85.14	83.61	88.79	87.39	88.27	86.82
PerformResSE (Lo5)	ASP	**92.76**	**90.14**	**92.43**	**89.74**	**88.83**	**87.52**	**88.31**	**86.94**

Note: Only the results of our implemented verify-to-classify framework with MLP fine-tuning are shown above. The full experimental results are available in Supplementary_Tables.

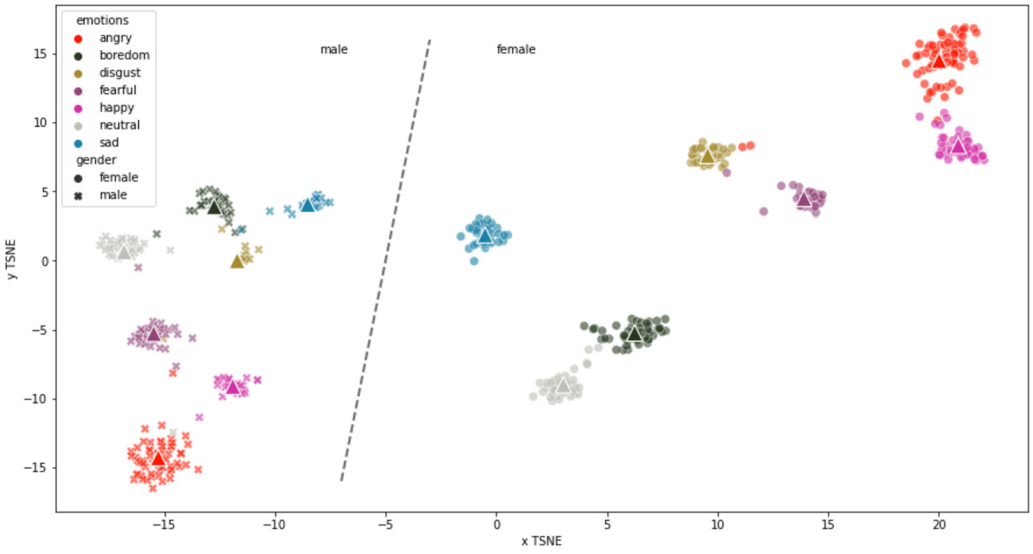


Fig. 8. A result of explainable emotion vectors of the best speech emotion verification, testing on Emo-DB dataset with low-quality environment. Note: The triangle symbols represent centroids.

Additional information

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2022.e09196.

References

[1] D. Jeong, B.-G. Kim, S.-Y. Dong, Deep joint spatiotemporal network (djstn) for efficient facial expression recognition, Sensors 20 (7) (2020) 1936.

[2] J.-H. Kim, B.-G. Kim, P.P. Roy, D.-M. Jeong, Efficient facial expression recognition algorithm based on hierarchical deep neural network structure, IEEE Access 7 (2019) 41273–41285.

[3] S. Singkul, B. Khampingyot, N. Maharattamalai, S. Taerungruang, T. Chalothorn, Parsing Thai social data: a new challenge for Thai nlp, in: 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2019, pp. 1–7.

[4] S. Singkul, K. Woraratpanya, Thai dependency parsing with character embedding, in: 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), 2019, pp. 1–5.

[5] S. Dieleman, B. Schrauwen, End-to-end learning for music audio, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6964–6968.

[6] Y. Li, T. Zhao, T. Kawahara, Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning, in: Interspeech, 2019, pp. 2803–2807.

[7] S. Singkul, T. Chatchaisathaporn, B. Suntisrivaraporn, K. Woraratpanya, Deep residual local feature learning for speech emotion recognition, in: H. Yang, K. Pasupa, A.C.-S. Leung, J.T. Kwok, J.H. Chan, I. King (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2020, pp. 241–252.

[8] S. Singkul, T. Chatchaisathaporn, B. Suntisrivaraporn, K. Woraratpanya, Deep residual local feature learning for speech emotion recognition, Lect. Notes Comput. Sci. (2020) 241–252.

[9] S.W.W. Sari, The influence of using repeated reading strategy towards student's reading comprehension, in: Proceeding 1st Annual International Conference on Islamic Education and Language: The Education and 4.0 Industrial Era in Islamic Perspective, 2019, p. 71.

[10] T. Shanahan, Everything you wanted to know about repeated reading, reading rockets, https://www.readingrockets.org/blogs/shanahan-literacy/everything-you-wanted-know-about-repeated-reading, 2017.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[12] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[13] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.

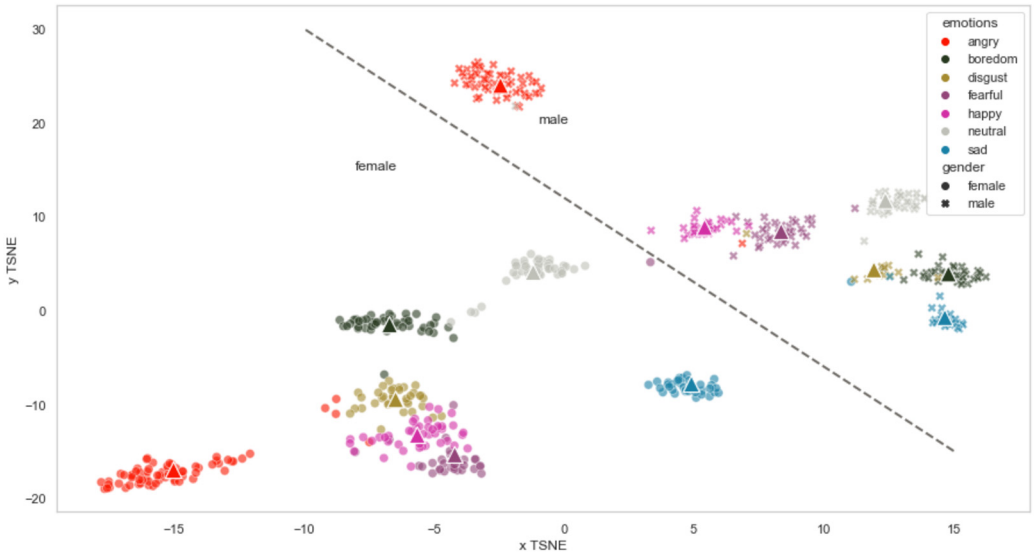


Fig. 9. A result of explainable emotion vectors of the best speech emotion verification, testing on Emo-DB dataset with high-quality environment. Note: The triangle symbols represent centroids.

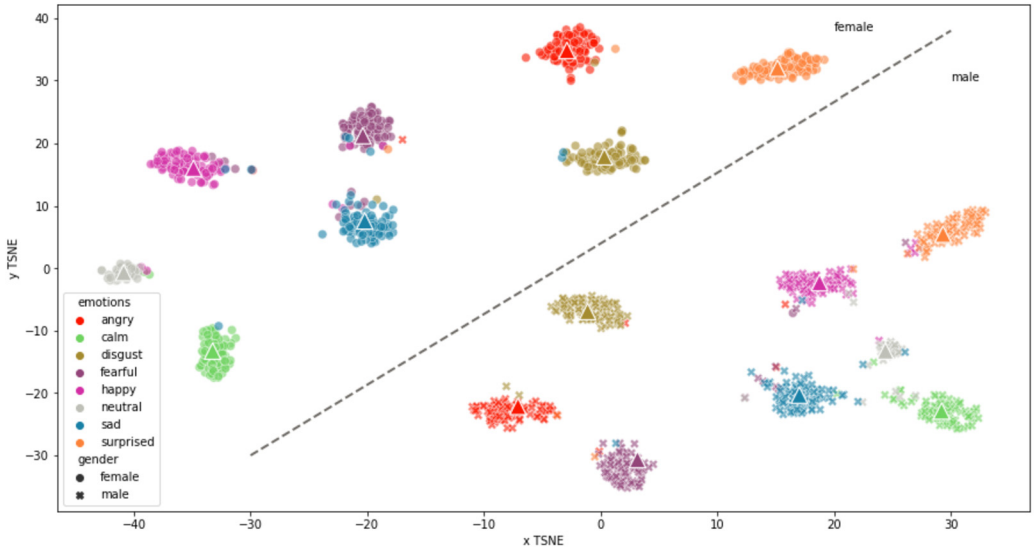


Fig. 10. A result of explainable emotion vectors of the best speech emotion verification, testing on RAVDESS dataset with low-quality environment. Note: The triangle symbols represent centroids.

[14] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: International Conference on Database Theory, Springer, 2001, pp. 420–434.
[15] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.
[16] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[17] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, B. An, Can cross entropy loss be robust to label noise, in: Proceedings of the 29th International Joint Conferences on Artificial Intelligence, 2020, pp. 2206–2212.
[18] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, et al., A database of German emotional speech, in: Interspeech, vol. 5, 2005, pp. 1517–1520.
[19] S.R. Livingstone, F.A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in North American English, PLoS ONE 13 (5) (2018).
[20] W. Hou, J. Wang, X. Tan, T. Qin, T. Shinozaki, Cross-domain speech recognition with unsupervised character-level distribution matching, preprint, arXiv:2104.07491, 2021.
[21] D. Soekhoe, P. Van Der Putten, A. Plaat, On the impact of data set size in transfer learning using deep neural networks, in: International Symposium on Intelligent Data Analysis, Springer, 2016, pp. 50–60.
[22] S.A.I. Alfarozi, K. Pasupa, M. Sugimoto, K. Woraratpanya, Local sigmoid method: non-iterative deterministic learning algorithm for automatic model construction of neural network, IEEE Access 8 (2020) 20342–20362.
[23] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, Specaugment: a simple data augmentation method for automatic speech recognition, Proc. Interspeech 2019 (2019) 2613–2617.
[24] M. Farooq, F. Hussain, N.K. Baloch, F.R. Raja, H. Yu, Y.B. Zikria, Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network, Sensors 20 (21) (2020) 6008.
[25] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.
[26] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. 43 (2) (2015) 155–177.
[27] Z. Zhang, E. Coutinho, J. Deng, B. Schuller, Cooperative learning and its application to emotion recognition from speech, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1) (2014) 115–126.
[28] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, E.P. Scilingo, Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients, Biomed. Signal Process. Control 17 (2015) 29–37.
[29] M. Shaneh, A. Taheri, Voice command recognition system based on mfcc and vq algorithms, Int. J. Comput. Inf. Eng. 3 (9) (2009) 2231–2235.

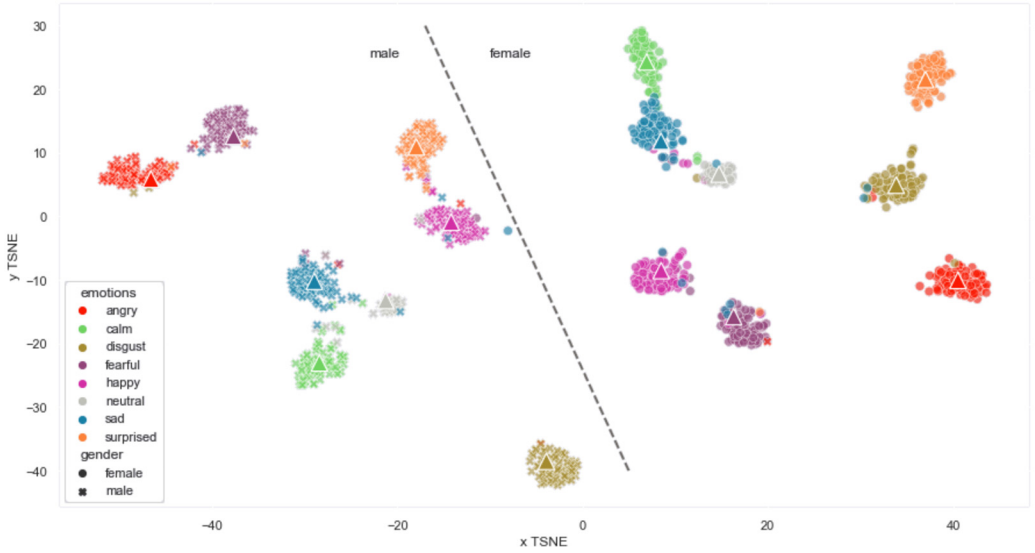


Fig. 11. A result of explainable emotion vectors of the best speech emotion verification, testing on RAVDESS dataset with high-quality environment. Note: The triangle symbols represent centroids.

[30] K. Venkataramanan, H.R. Rajamohan, Emotion recognition from speech, arXiv: 1912.10458, 2019.

[31] S. Demircan, H. Kahramanli, Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech, Neural Comput. Appl. 29 (8) (2018) 59–66.

[32] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process. Control 47 (2019) 312–323.

[33] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 801–804.

[34] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, Springer, 2016, pp. 630–645.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[36] H.S. Heo, B.-J. Lee, J. Huh, J.S. Chung, Clova baseline system for the voxceleb speaker recognition challenge 2020, preprint, arXiv:2009.14153, 2020.

[37] K. Okabe, T. Koshinaka, K. Shinoda, Attentive statistics pooling for deep speaker embedding, Proc. Interspeech 2018 (2018) 2252–2256.

[38] W. Cai, J. Chen, M. Li, Exploring the encoding layer and loss function in end-to-end speaker and language recognition system, in: Proc. Odyssey 2018 the Speaker and Language Recognition Workshop, 2018, pp. 74–81.

[39] J.S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, I. Han, In defence of metric learning for speaker recognition, Proc. Interspeech 2020 (2020) 2977–2981.

[40] P. Kumar, S. Jain, B. Raman, P.P. Roy, M. Iwamura, End-to-end triplet loss based emotion embedding system for speech emotion recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8766–8773.

[41] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: deep hypersphere embedding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 212–220.

[42] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, Z. Zhu, Deep speaker: an end-to-end neural speaker embedding system, preprint, arXiv: 1705.02304, 2017, 650.

[43] L. Wan, Q. Wang, A. Papir, I.L. Moreno, Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4879–4883.

[44] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, S. Khudanpur, Deep neural network-based speaker embeddings for end-to-end speaker verification, in: 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2016, pp. 165–170.

[45] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: International Workshop on Similarity-Based Pattern Recognition, Springer, 2015, pp. 84–92.

[46] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4080–4090.

[47] C. Breitenstein, D.V. Lancker, I. Daum, The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample, Cogn. Emot. 15 (1) (2001) 57–79.

[48] T.J. Sefara, The effects of normalisation methods on speech emotion recognition, in: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), IEEE, 2019, pp. 1–8.

[49] M. Markitantov, Transfer learning in speaker's age and gender recognition, in: A. Karpov, R. Potapova (Eds.), Speech and Computer, Springer International Publishing, Cham, 2020, pp. 326–335.

[50] W. Xie, A. Nagrani, J.S. Chung, A. Zisserman, Utterance-level aggregation for speaker recognition in the wild, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5791–5795.

[51] R. Pappagari, T. Wang, J. Villalba, N. Chen, N. Dehak, x-vectors meet emotions: a study on dependencies between emotion and speaker recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7169–7173.

[52] Y. Jia, Y. Zhang, R.J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I.L. Moreno, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 4485–4495.

[53] M. Turkoglu, Covidetectionet: Covid-19 diagnosis system based on x-ray images using features selected from pre-learned deep features ensemble, Appl. Intell. 51 (3) (2021) 1213–1226.

[54] G. Bhattacharya, M.J. Alam, P. Kenny, Deep speaker embeddings for short-duration speaker verification, in: Interspeech, 2017, pp. 1517–1521.

[55] F.R. Rahman Chowdhury, Q. Wang, I.L. Moreno, L. Wan, Attention-based models for text-dependent speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5359–5363.

[56] X. Dong, J. Shen, Triplet loss in Siamese network for object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 459–474.

[57] A. Nagrani, J.S. Chung, W. Xie, A. Zisserman, Voxceleb: large-scale speaker verification in the wild, Comput. Speech Lang. 60 (2020) 101027.

[58] U. Khan, J. Hernando, Unsupervised Training of Siamese Networks for Speaker Verification, in: Proc. Interspeech 2020, 2020, pp. 3002–3006.

[59] C. Zhang, K. Koishida, End-to-end text-independent speaker verification with triplet loss on short utterances, in: Interspeech, 2017, pp. 1487–1491.

[60] Y. Li, F. Gao, Z. Ou, J. Sun, Angular softmax loss for end-to-end speaker verification, in: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2018, pp. 190–194.

[61] H.-S. Heo, J.-w. Jung, I.-H. Yang, S.-H. Yoon, H.-j. Shim, H.-J. Yu, End-to-end losses based on speaker basis vectors and all-speaker hard negative mining for speaker verification, Proc. Interspeech 2019 (2019) 4035–4039.

[62] V. Menger, F. Scheepers, M. Spruit, Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text, Appl. Sci. 8 (6) (2018) 981.

[63] L. Wang, Support Vector Machines: Theory and Applications, vol. 177, Springer Science & Business Media, 2005.

[64] Y. Kim, K. Toh, A method to enhance face biometric security, in: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–6.

[65] U. Gawande, Y. Golhar, Biometric security system: a rigorous review of unimodal and multimodal biometrics techniques, Int. J. Biom. 10 (2) (2018) 142–175.

[66] P. Agrawal, R. Kapoor, S. Agrawal, A hybrid partial fingerprint matching algorithm for estimation of equal error rate, in: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 1295–1299.

[67] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev.: Comput. Stat. 2 (4) (2010) 433–459.
[68] K. Pal, M. Sharma, Performance evaluation of non-linear techniques umap and t-sne for data in higher dimensional topological space, in: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), IEEE, 2020, pp. 1106–1110.
[69] R. Shah, S. Silwal, Using dimensionality reduction to optimize t-sne, preprint, arXiv: 1912.01098, 2019.
[70] K. Vijayan, P.R. Reddy, K.S.R. Murty, Significance of analytic phase of speech signals in speaker verification, Speech Commun. 81 (2016) 54–71.
[71] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, preprint, arXiv:1405.3531, 2014.
[72] A. Nagrani, J.S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset, Telephony 3 (2017) 33–039.