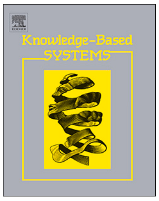
Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition

Sheng Zhang [a,1], Min Chen [c,d], Jincai Chen [a,b,c,*], Yuan-Fang Li [f], Yiling Wu [e], Minglei Li [e], Chuanbo Zhu [a]

[a] *Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China*
[b] *Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China, Wuhan 430074, China*
[c] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[d] *Embedded and Pervasive Computing (EPIC) Lab, Huazhong University of Science and Technology, Wuhan 430074, China*
[e] *Huawei Cloud BU, Shenzhen 518129, China*
[f] *Department of Data Science and AI, Faculty of Information Technology, Monash University, Clayton 3800, Australia*

ARTICLE INFO

Article history:
Received 30 December 2020
Received in revised form 19 July 2021
Accepted 22 July 2021
Available online 24 July 2021

Keywords:
Semi-supervised learning
Cross-modal knowledge transfer
Speech emotion recognition

ABSTRACT

Speech emotion recognition is an important task with a wide range of applications. However, the progress of speech emotion recognition is limited by the lack of large, high-quality labeled speech datasets, due to the high annotation cost and the inherent ambiguity in emotion labels. The recent emergence of large-scale video data makes it possible to obtain massive, though unlabeled speech data. To exploit this unlabeled data, previous works have explored semi-supervised learning methods on various tasks. However, noisy pseudo-labels remain a challenge for these methods. In this work, to alleviate the above issue, we propose a new architecture that combines cross-modal knowledge transfer from visual to audio modality into our semi-supervised learning method with consistency regularization. We posit that introducing visual emotional knowledge by the cross-modal transfer method can increase the diversity and accuracy of pseudo-labels and improve the robustness of the model. To combine knowledge from cross-modal transfer and semi-supervised learning, we design two fusion algorithms, i.e. weighted fusion and consistent & random. Our experiments on CH-SIMS and IEMOCAP datasets show that our method can effectively use additional unlabeled audio-visual data to outperform state-of-the-art results.

1. Introduction

The ability to recognize emotions is essential for carrying out empathic and natural human–computer interactions [1,2]. Along with rapid development of conversational agents such as Siri, Alexa and Cortana, speech emotion recognition (SER) [3] has attracted significant research interest in recent years.

Speech emotions are related to many factors of a speaker, including gender, age, culture, dialect, and others [3]. Speech emotions can be quantified with several discrete categories, such as happiness, sadness, anger, and neural, etc. Researchers have explored many methods to classify speech emotions, such as hidden Markov models, support vector machines, deep belief networks, convolutional neural networks (CNN), and long short-term memory networks (LSTM) [3]. A number of datasets have been proposed for the SER task. However, many of these existing datasets [4–7] are either small-scale or low-quality, limiting the performance of SER. Although critical to the improvement of the SER task, gathering large, high-quality annotated data is difficult, costly and time-consuming. Moreover, as emotion recognition is subjective, it is often difficult for annotators to reach an agreement, and thus difficult to produce large, high-quality labeled data required by supervised learning methods.

With the rise of video-sharing platforms and social networks, large-scale video data has been made available. It is thus possible to obtain massive amounts of samples of unlabeled emotional speech. Semi-supervised learning (SSL) methods [9–12] have successfully exploited unlabeled data to obtain strong performance in some tasks such as image classification. The simplest form of SSL is self-training, which exploits a SER model pre-trained with a small amount of labeled data to generate pseudo-labels for

[*] Corresponding author at: Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China.

E-mail addresses: zhangmonkey@hust.edu.cn (S. Zhang), minchen2012@hust.edu.cn (M. Chen), jcchen@hust.edu.cn (J. Chen), yuanfang.li@monash.edu (Y.-F. Li), wuyiling1@huawei.com (Y. Wu), liminglei29@huawei.com (M. Li), chuanbo_zhu@hust.edu.cn (C. Zhu).

[1] Work done during an internship at Huawei Cloud BU.

https://doi.org/10.1016/j.knosys.2021.107340
0950-7051/© 2021 Elsevier B.V. All rights reserved.

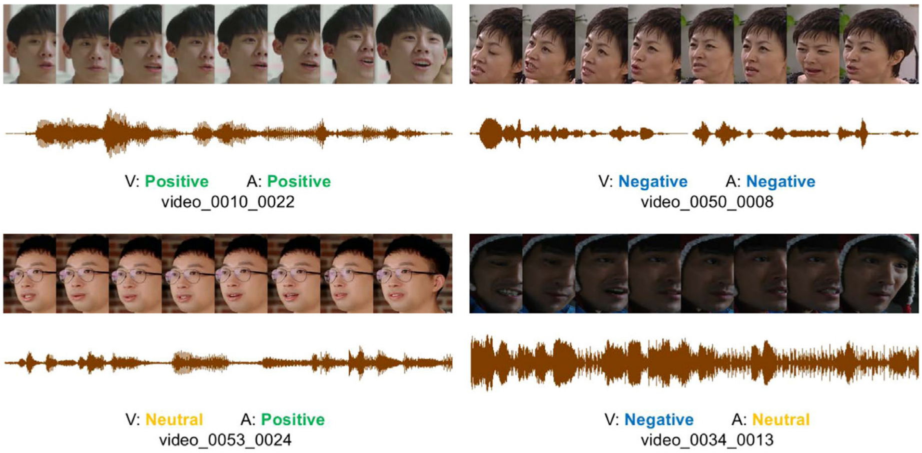


Fig. 1. Examples of audio-visual pairs with consistent and inconsistent emotion states in the CH-SMIS dataset [8], which has a 51.81% inconsistency rate between audio and facial emotion states according to our statistical analysis. A consistent pair can generate a useful pseudo-labeled audio sample via cross-modal knowledge transfer from visual modality to audio modality, while an inconsistent pair may introduce a noisy pseudo-label.

unlabeled data. The unlabeled data with generated pseudo-labels is then used to participate in training a new model. However, this kind of SSL has inherent deficiencies. Error pseudo-labels can mislead the SER model to propagate these errors.

The cross-modal knowledge transfer (CMKT) method exploits the synchronicity of emotions across modalities in multi-modal data. Some works [13] have demonstrated the relevance between speech prosody and facial cues. They provide a basis for using facial expressions to explain speech emotions in videos. Fortunately, there are lots of publicly available labeled image-based facial expression recognition datasets, which makes it possible to utilize facial expression knowledge to explain emotions of the audio modality of videos and generate large amounts of labeled audio samples. These data can be used to improve the performance of SER. Some methods [14–16] have been proposed based on this idea. Specifically, Albanie et al. [14] apply a pre-trained facial expression recognition model to generate emotion labels for synchronous audio in unlabeled videos. However, directly using labels predicted by the facial expression recognition model has some limitations. Firstly, the facial expression recognition model may make mistakes. Secondly, as shown in Fig. 1, there are inconsistencies in the affective expressions between the audio modality and the visual modality. These problems may introduce noisy labels in the cross-modal knowledge transfer method, which could impact the performance of SER models.

We propose to combine CMKT and SSL methods to more effectively exploit massive unlabeled audio-visual data so as to enhance the performance of SER. We employ a number of techniques to overcome the weaknesses of these methods. First, we train a strong facial expression recognition model with large-scale face expression datasets. The CMKT module can identify emotion by the pre-trained facial expression recognition model and transfer the emotion information of the visual modality to the audio modality. The emotion information could be represented as a one-hot vector or a probability distribution. The CMKT method attaches facial emotional knowledge to unlabeled audio data as their pseudo-labels from the visual modality. Then, we generalize FixMatch [11], which is a recent SSL method for the image classification task, to the audio setting to give unlabeled audio data another pseudo-labels from the audio modality.

The CMKT and generalized FixMatch methods provide two sets of pseudo-labels, which contain knowledge from two modalities and could complement each other. The fusion of emotion knowledge from the two modalities could alleviate the label noise problem, which exists in both CMKT and SSL methods. Meanwhile, introducing facial emotional knowledge can increase the diversity of pseudo-labels and improve the robustness of the model. We thus design two fusion algorithms to combine them. In the first algorithm, we first select the unlabeled audio data with a consistent pseudo-label across the two modalities. Then, we randomly take from the remaining unlabeled audio data and assign them pseudo-labels from the visual modality. In the second algorithm, we perform a weighted summation of the probability distributions of emotional predictions from the two modalities. Subsequently, we pick the class with the largest predicted probability as the pseudo-label of unlabeled data. After obtaining pseudo-labels for unlabeled audio data, we join them with labeled audio data to train an SER model.

Our main contributions in this paper can be summarized as follows:

1. To effectively use unlabeled audio-visual data obtained in the wild, we propose an architecture that combines CMKT and SSL methods for the SER task.
2. To extract emotion knowledge of the visual modality, we train a strong facial expression recognition model on large-scale face expression datasets. We utilize this model to obtain an emotion explanation of unlabeled audio from visual view and generate pseudo-labels for unlabeled audio data.
3. To extract emotion knowledge from the audio modality, we design an SSL method for the SER task by generalizing FixMatch algorithms to generate pseudo-labels from the same audio modality for unlabeled audio data.
4. We design two fusion algorithms to exploit facial and audio emotion knowledge to improve the accuracy and diversity of pseudo-labels generated for unlabeled audio data.
5. Our experiments on CH-SIMS [8] and IEMOCAP [4] datasets show that our method can effectively use the additional unlabeled audio-visual data to improve performance and achieve state-of-the-art results.

2. Related work

In this section, we briefly introduce related works, including SER based on deep learning (abbreviated as DL), the relevance between facial and speech emotion, and typical SSL methods.

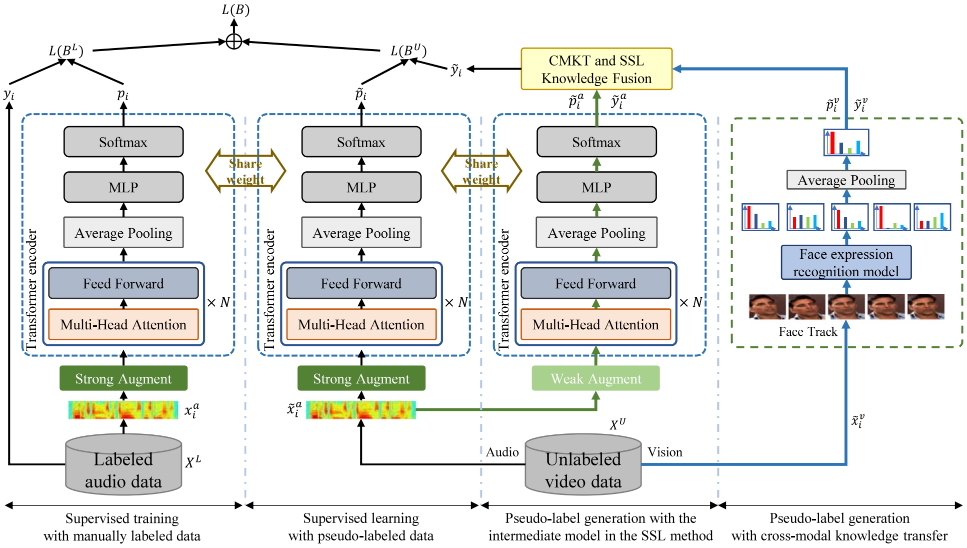


Fig. 2. The high-level architecture of combining CMKT and SSL for SER.

2.1. Deep learning for speech emotion recognition

DL techniques for SER have made great progress [17–25] in recent years. Generally, DL for SER either operates on pre-extracted acoustic features, or operates on raw audio without any processing [26]. Early works [14,27–31] extract hand-crafted features for deep networks. The rise of pre-training methods makes it possible to learn widely applicable audio representations. Wav2vec [32] uses the Librispeech [33] dataset to train the model mapping raw audio to a dense representation with good generalization. We therefore opt for the Wav2Vec pre-trained features for English audio in our method.

Several works have constructed deep neural network architecture for SER. Satt et al. [34] combine CNN with LSTM to classify emotions. The Transformer [35] model is one of the state-of-art Seq2Seq architectures based on the attention mechanism without the use of recurrence or convolution. Some recent works [32] begin to widely use the encoder of Transformer to extract audio features. In our work, we also use the encoder of Transformer to learn the audio representation.

2.2. Relevance between facial and speech emotion

The relevance between speech prosody and facial cues has been extensively studied [13], due to their joint relevance to human perception, communication, and behavior. The broad accord of these studies is that during conversations, speech prosody is generally associated with other social cues like facial expressions or gestures [36].

Some recent works have considered the links between facial and speech emotions in various tasks. Albanie et al. [14] explore transferring emotion labels from the visual modality to the audio modality in SER. Differently, Liang et al. [15] regard the latent links, that the inner emotional status is consistent across modalities of video, as an auxiliary task to obtain guidance from unlabeled data to enhance fully-supervised learning. In contrast to the aforementioned works, Yu et al. [8] focus on the difference among emotions of several modalities. They propose the Chinese single- and multi-modal sentiment analysis dataset (CH-SIMS) with three modalities. CH-SIMS collects unimodal annotations in addition to multimodal annotations for each clip. They demonstrate that independent unimodal annotation contributes to learning more distinctive unimodal representations and more accurately reasoning emotion states of utterances.

2.3. Semi-supervised learning

Researchers have proposed lots of SSL methods, which are powerful approaches that could effectively train deep networks using a small amount of labeled data and a large amount of unlabeled data.

As a hot field, SSL has a huge diversity of methods. We focus only on approaches closely related to our work. As early as several decades [37], the idea behind pseudo-labeling has appeared. Typically, this kind of SSL method can use a model pre-trained by a small amount of labeled datasets to predict pseudo-labels for unlabeled data, which is general and applied in diverse domains including NLP, object detection, image classification, to name a few. Pseudo-labeling [9] converts model predictions to hard labels and only retains the unlabeled samples with sufficiently confident pseudo-labels.

The "π-model" [38] first proposes consistency regularization, which relies on the assumption that a model should output similar predictions when fed perturbed versions of the same sample. Various methods are proposed to produce random perturbations including data augmentation, stochastic regularization (e.g. Dropout [39]), and adversarial perturbations to extend the consistency regularization method. Recent work [40] shows that using strong data augmentation can produce better results. Based on the consistency regularization, Temporal Ensembling [10] and Mean Teacher [12] are proposed to improve the accuracy of predictions by averaging predictions from the model of each epoch and averaging consecutive student models respectively.

FixMatch [11] combines the two existing methods above: consistency regularization and pseudo-labeling. FixMatch first generates pseudo-labels using predictions on weakly-augmented unlabeled images and retains pseudo-labels with a high confidence. Then, FixMatch trains the model to predict pseudo-labels when fed a strongly-augmented version of the same image. FixMatch achieves state-of-the-art performance in image classification.

However, there are few effective SSL methods for SER. Besides, typical SSL methods suffer from the problem of noisy labels. Specifically, error predictions generated by the intermediate model could mislead the model to strengthen these errors during training.

3. Method

Our goal is to utilize largely available unlabeled videos that contain paired audio-visual data to improve the performance of SER.

We start by introducing the notations used in our paper. Assume we have a labeled audio database $(X^L, Y) = \{(x_i^a, y_i)\}_{i=1}^{n^L}$, where $x_i^a \in \mathbb{R}^{t\times d}$ denotes the audio feature, t denotes the number of frames, d is the feature dimension, and $y_i \in \{0, 1, \ldots, c-1\}$ denotes the corresponding manual label. c is the number of the emotion category. In addition, we have an unlabeled video database $X^U = \{(\tilde{x}_i^a, \tilde{x}_i^v)\}_{i=1}^{n^U}$, where $\tilde{x}_i^a \in \mathbb{R}^{t\times d}$ denotes the audio feature, $\tilde{x}_i^v \in \mathbb{R}^{f\times w\times h}$ denotes the corresponding visual image sequence. f, w, and h is the number, width, and height of the video frame respectively. L/U indicates labeled/unlabeled data. n^L and n^U are the sizes of the labeled and unlabeled databases respectively ($n^L \ll n^U$).

To effectively exploit massive unlabeled data, we propose to combine CMKT and SSL methods. The architecture of our model is shown in Fig. 2. Our model mainly includes five parts: CMKT for pseudo-label generation, SSL for pseudo-label generation, CMKT and SSL knowledge fusion, supervised training with manually labeled data, and supervised training with pseudo-labeled data.

First, in the CMKT module, we pass the visual part $\tilde{x}_i^v$ of the unlabeled audio-visual pairs into a facial expression recognition model pre-trained with face expression datasets to extract facial expression information $\tilde{p}_i^v$ (or $\tilde{y}_i^v$), which is described in Section 3.2. Then, we exploit an intermediate SER model trained by supervised learning to generate emotion information $\tilde{p}_i^a$ (or $\tilde{y}_i^a$) for the audio part $\tilde{x}_i^a$ of the unlabeled audio-visual pairs, which is detailed in Section 3.3. Next, the CMKT and SSL knowledge fusion module exploits emotion information $\tilde{p}_i^v$ (or $\tilde{y}_i^v$) and $\tilde{p}_i^a$ (or $\tilde{y}_i^a$) to decide the final label $\tilde{y}_i$ for $\tilde{x}_i^a$, which is described in Section 3.4. Finally, we use manually labeled data and unlabeled data with pseudo-labels to train the SER model in a supervised learning way. Thus, the final loss comes from manually labeled data and unlabeled data as follows:

$$\mathcal{L}(\mathcal{B}) = \mathcal{L}(\mathcal{B}^L) + \lambda\mathcal{L}(\mathcal{B}^U), \tag{1}$$

where $\mathcal{L}$ is the cross entropy loss function, $\mathcal{B}$ denotes a mini-batch of training samples and λ is the balancing weight. The above process composes an iteration. We run the iteration in a loop until convergence.

In the rest of this section, we present the network structure for SER in Section 3.1. Then, Section 3.2 explains how to extract facial expression knowledge to get pseudo-labels for audio data from the visual modality. We design an SSL method (i.e. a variant of FixMatch) to exploit knowledge from the audio modality to generate pseudo-labels in Section 3.3. Finally, Section 3.4 shows how to combine CMKT and SSL methods for SER.

3.1. Speech emotion recognition network

SER is regarded as a classification task. The structure in the blue dashed box of Fig. 2 shows the SER network. First, for the audio representation, we consider the Transformer [35] model, which is one of the state-of-art Seq2Seq architectures based on the attention mechanism. We use the encoder of Transformer to extract the audio representation. Then, the hidden representation $h \in \mathbb{R}^{t\times d}$ from the Transformer encoder is aggregated via average pooling, and we use a multi-layer perceptron (MLP) on these pooling outputs to generate a score vector $f \in \mathbb{R}^c$. Finally, a softmax function is used to output class distribution $p \in \mathbb{R}^c$. These processing steps can be formulated as follows:

$$\begin{aligned} h &= Encoder(x^a), \\ z &= AveragePooling(h), \\ f &= MLP(z), \\ p &= Softmax(f), \end{aligned} \tag{2}$$

where $x^a \in \mathbb{R}^{t\times d}$ is the audio feature, $z \in \mathbb{R}^d$ is the aggregated feature. For convenience, we define this model as $M_\theta(x^a)$.

3.2. Cross-modal knowledge transfer for pseudo-label generation

The facial expression knowledge extraction pipeline is shown in the green dashed box of Fig. 2. There are more large-scale image-based than video-based labeled datasets for facial expression recognition. Thus, we follow Albanie et al. [14] to perform expression recognition on facial images. Our facial expression recognition model is based on MobileNetV2 [41], which is trained on the RAF-DB dataset [42] for seven expressions: neutral, happiness, surprise, sadness, anger, disgust, and fear. Then, the pre-trained model is used to predict the expression probability distribution of each face frame.

To transfer the emotional knowledge of the visual domain to the audio domain, we need to convert the expression probability distribution from frame level to face-track level. As a single speech segment spans many face frames, an utterance corresponds to multiple facial expression results. In this work, we also follow [14] to adopt a simple average pooling method on these results. The pooling outputs are then passed to a normalized function to generate the prediction distribution $\tilde{p}_i^v \in \mathbb{R}^c$. We compute index $\tilde{y}_i^v \in \{0, 1, \ldots, c-1\}$ with the highest confidence in $\tilde{p}_i^v$. $\tilde{p}_i^v$ and $\tilde{y}_i^v$ could be used as a soft and hard pseudo-label for the audio domain of audio-visual pairs respectively.

3.3. Semi-supervised learning for pseudo-label generation

To exploit unlabeled data, many semi-supervised approaches [9–12] (e.g. pseudo-labeling) have been proposed. However, few semi-supervised methods have been applied to the SER task.

Inspired by the FixMatch [11] algorithm, which is a recent SSL method and proposed for image classification, we design an SSL method for the SER task. FixMatch employs the notion of consistency regularization, which relies on the assumption that a model should output similar predictions when fed perturbed versions of the same sample.

We implement consistency training by adopting two different kinds of data augmentation to the input $\tilde{x}_i^a$. First, to ensure reasonable accuracy of pseudo-labels, a weak augmentation method (abbreviated as *WA*, for example, dropout) is used to get a weakly-augmented version $WA(\tilde{x}_i^a)$ of the input. During pseudo-label generation, we first compute the class distribution $\tilde{p}_i^a$ given the weakly-augmented version. We then obtain index $\tilde{y}_i^a$ with the highest confidence in $\tilde{p}_i^a$. The formula of this process is as follows:

$$\begin{aligned} \tilde{p}_i^a &= M_\theta(WA(\tilde{x}_i^a)), \\ \tilde{y}_i^a &= argmax(\tilde{p}_i^a), \end{aligned} \tag{3}$$

where $\tilde{p}_i^a \in \mathbb{R}^c$ and $\tilde{y}_i^a \in \{0, 1, \ldots, c-1\}$, and M_θ is an intermediate model. $\tilde{p}_i^a$ and $\tilde{y}_i^a$ can be used as a soft and hard pseudo-label of an unlabeled audio sample respectively. They contain emotion knowledge from the audio modality.

Then, we use SpecAugment [43] as a strong augmentation function (abbreviated as *SA*), which applies time frame masking and frequency band masking to audio features and can produce heavily distorted versions. We use *SA* to get a strongly-augmented version $SA(\tilde{x}_i^a)$. We then use the above pseudo-label $\tilde{y}_i^a$ as a supervisory signal and enforce the cross-entropy loss against the model's output for $SA(\tilde{x}_i^a)$:

$$\begin{aligned} \tilde{p}_i &= M_\theta(SA(\tilde{x}_i^a)), \\ \mathcal{L}(\mathcal{B}^U) &= -\frac{1}{|\mathcal{B}^U|}\sum_{i\in\mathcal{B}^U} \mathbf{1}\{max(\tilde{p}_i^a) > \tau\}\log(\tilde{p}_i[\tilde{y}_i^a]), \end{aligned} \tag{4}$$

where $\tilde{p}_i[\tilde{y}_i^a]$ is the prediction probability of the corresponding pseudo-label $\tilde{y}_i^a$ in the probability distribution $\tilde{p}_i$. To ensure the accuracy of pseudo-labels, the threshold τ is used to select unlabeled samples with bigger prediction probabilities to participate in training.

3.4. Combining cross-modal knowledge transfer and semi-supervised learning for pseudo-label generation and SER

In the above SSL method, a model just uses its own predictions to teach itself in the training phase. Thus, prediction errors can reinforce themselves. This will cause the model to perform poorly in certain categories. Introducing facial expression knowledge can alleviate this problem, because of the difference of emotion prediction distributions between SSL and facial expression transfer methods. It is unreasonable to directly assign pseudo-labels from facial expression recognition to unlabeled audio data, due to the existence of the inconsistency between facial expressions and audio emotions. Thus, we combine two emotion pseudo-labels from visual and audio modalities to generate final pseudo-labels as shown in Fig. 2. We design two fusion functions (abbreviated as FF) for visual and audio emotion knowledge fusion.

In the first algorithm FF_1, inputs use the hard pseudo-label $\tilde{y}_i^v$ from the facial expression recognition and the hard pseudo-label $\tilde{y}_i^a$ generated in SSL. To ensure the accuracy of the final pseudo-labels, we first retain consistent pseudo-labels by using a mask, which is computed by:

$$mask_i = \mathbf{1}\{\tilde{y}_i^a == \tilde{y}_i^v\} \tag{5}$$

Despite consistency can ensure the accuracy of pseudo-labels, the model requires different emotion knowledge extracted by CMKT to overcome the limitation of SSL. In addition, it is impossible to know whether inconsistent pseudo-labels are noise. To introduce different emotion knowledge of the visual domain, we use a uniform distribution $U(0,1)$ to assign a random weight to each unlabeled audio sample. Then, samples with weight less than ϵ will be retained. The mask could be redefined as:

$$mask_i = \mathbf{1}\{\tilde{y}_i^a == \tilde{y}_i^v \ or \ w_i < \epsilon\}, \ \ where \ w_i \in U(0,1) \tag{6}$$

We can control the proportion of random samples by ϵ. The mask is used to compute unlabeled loss as follows:

$$\begin{aligned} \tilde{p}_i &= M_\theta(\mathcal{SA}(\tilde{x}_i^a)), \\ \mathcal{L}(\mathcal{B}^U) &= -\frac{1}{|\mathcal{B}^U|}\sum_{i\in\mathcal{B}^U} mask_i \times log(\tilde{p}_i[\tilde{y}_i^v]), \end{aligned} \tag{7}$$

where $\tilde{p}_i$ is a probability distribution output of the model for $\mathcal{SA}(\tilde{x}_i^a)$. Advantages of this algorithm are that pseudo-labels generated via merging multi-view knowledge are more accurate, and randomly selecting samples with pseudo-labels of the visual domain can increase the diversity of emotional knowledge. The FF_1 algorithm is called as the Consistent & Random method.

In our second algorithm FF_2, inputs contain the soft pseudo-label $\tilde{p}_i^a$ from SSL and the soft pseudo-label $\tilde{p}_i^v$ generated by the CMKT method. FF_2 directly uses a weighted fusion way to combine them, and then obtains the final pseudo-label by computing index $\tilde{y}_i$ with the highest confidence as follows:

$$\begin{aligned} \tilde{p}_i^{av} &= \alpha\tilde{p}_i^a + (1-\alpha)\tilde{p}_i^v, \\ \tilde{y}_i &= argmax(\tilde{p}_i^{av}), \end{aligned} \tag{8}$$

where α is a scalar parameter to balance the importance of facial and audio emotion distributions. Then, the unlabeled loss in FF_2 is computed as:

$$\begin{aligned} \tilde{p}_i &= M_\theta(\mathcal{SA}(\tilde{x}_i^a)), \\ \mathcal{L}(\mathcal{B}^U) &= -\frac{1}{|\mathcal{B}^U|}\sum_{i\in\mathcal{B}^U} \mathbf{1}\{max(\tilde{p}_i^{av}) > \tau\}\log(\tilde{p}_i[\tilde{y}_i]), \end{aligned} \tag{9}$$

where τ is a scalar parameter and used to control the accuracy of pseudo-labels. The FF_2 algorithm is called as the Weighted Fusion method.

Table 1
Dataset information: We summarize details of the four datasets we use.

Language	Type	Name	Size	Labels
Chinese	Labeled Unlabeled	CH-SIMS [8] iQIYI-VID [44]	2,281 14,502	3 classes –
English	Labeled Unlabeled	IEMOCAP [4] EmoVoxCeleb [14]	5,531 Over 1 million	4 classes –

Next, we can joint unlabeled and labeled audio data to train an SER model. The loss can be computed as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{B}^L) &= -\frac{1}{|\mathcal{B}^L|}\sum_{i\in\mathcal{B}^L}\log(M_\theta(\mathcal{SA}(x_i^a))[y_i]), \\ \mathcal{L}(\mathcal{B}) &= \mathcal{L}(\mathcal{B}^L) + \lambda\mathcal{L}(\mathcal{B}^U), \end{aligned} \tag{10}$$

where x_i^a and y_i are features and labels of labeled audio samples respectively. λ is a fixed hyperparameter to balance the unlabeled and labeled loss. $\mathcal{L}(\mathcal{B}^U)$ could come from the Consistent & Random or the Weighted Fusion. The pseudocode of the complete algorithm for the training of a SER model is shown in Algorithm 1.

```
Algorithm 1: Combining CMKT and SSL for SER.
Input: $\tilde{x}_i^a$: unlabeled audio sample, $\tilde{y}_i^v$: hard pseudo-label from visual modality for $\tilde{x}_i^a$, $\tilde{p}_i^v$: soft pseudo-label from visual modality for $\tilde{x}_i^a$, $x_i^a$: labeled audio sample, $y_i$: label for $x_i^a$, $M_\theta(x)$: encoder with trainable parameters $\theta$, $WA(x)$: weak augmentation function, $\mathcal{SA}(x)$: strong augmentation function, $\lambda$: unlabeled loss weight, $\epsilon$: weight threshold, $\tau$: probability threshold, FF: fusion function, epochs: training epoch number
Output: $\theta$
for $t$ in [1, epochs] do
    for each minibatch ($\mathcal{B}^U$, $\mathcal{B}^L$) do
        $\mathcal{L}(\mathcal{B}^L) \leftarrow -\frac{1}{|\mathcal{B}^L|}\sum_{i\in\mathcal{B}^L} log(M_\theta(\mathcal{SA}(x_i^a))[y_i])$
        $\tilde{p}_i^a \leftarrow M_\theta(WA(\tilde{x}_i^a))$    ▷ soft pseudo-label from the audio modality
        $\tilde{y}_i^a \leftarrow argmax(\tilde{p}_i^a)$    ▷ hard pseudo-label from the audio modality
        $\tilde{p}_i \leftarrow M_\theta(\mathcal{SA}(\tilde{x}_i^a))$
        $\mathcal{L}(\mathcal{B}^U) \leftarrow FF_1(\tilde{y}_i^v, \tilde{y}_i^a, \tilde{p}_i, \epsilon)$ or $FF_2(\tilde{p}_i^v, \tilde{p}_i^a, \tilde{p}_i, \tau)$
        $\mathcal{L}(\mathcal{B}) \leftarrow \mathcal{L}(\mathcal{B}^L) + \lambda\mathcal{L}(\mathcal{B}^U)$
        update $\theta$
    end for
end for
```

4. Experiments

We evaluate our model via a series of experiments on two tasks including Chinese speech sentiment analysis and English SER.

4.1. Data description

Here we discuss details of the four datasets we use to evaluate and benchmark our method. For further readability, we have summarized these details in Table 1.

CH-SIMS. CH-SIMS [8] is a Chinese single- and multi-modal sentiment analysis dataset consisting of 2,281 refined video segments in the wild with both multimodal and independent unimodal annotations. Each segment contains 15 words and has a length of 3.67s on average. We only used the audio part of each segment for speech sentiment recognition. Each clip is labeled by the average of five sentiment scores by human annotators. In this paper, we focus on sentiment polarities rather than scores, so we divide a score into three states, including positive (score $\in$ (0.1, 1]), neutral (score $\in$ [−0.1, 0.1]), and negative (score $\in$ [−1, −0.1)).

iQIYI-VID. iQIYI-VID [44] contains 643,816 video clips of 10,034 identities. To match the CH-SIMS dataset, we need to filter raw videos of iQIYI-VID to meet the following constraints [8]:

1. The language of videos should be mandarin.
2. The length of each segment is no less than one second and no more than ten seconds. In the meanwhile, every segment should correspond to a complete sentence.
3. Each video segment only contains the speaker's face.

However, the video pre-processing methods used in [8] are manual and prohibitively time-consuming. To collect large-scale data that meet the constrains, we first cut up and select videos according to captions, and then apply existing pre-trained models including face detection and speaker detection to select the clips that meet the constraints. Finally, we obtain 14,502 video flips (25 fps). We use these videos as an unlabeled dataset to extend the CH-SIMS dataset.

IEMOCAP. IEMOCAP [4] is a multi-modal emotion recognition dataset consisting of 12 h of videos. Its multimodal streams are sampled by a fixed sampling rate on audio (12.5 Hz) and vision (15 Hz) [45]. The videos are divided into five sessions. Each session includes two actors, a male and a female. Following previous works [45], 4 emotions (happy, angry, sad, neutral) are selected for emotion recognition. Thus, we use 5531 utterances including 1103 angry, 1636 happy, 1708 neutral, and 1084 sad from 5 sessions and 10 speakers. In this paper, we follow the speaker-independent setting to avoid actor overlap in the training, validation, and testing set. Thus, we choose 8 speakers in four sessions into the training set and the remaining 2 speakers are divided into the validation set and the testing set respectively.

EmoVoxCeleb. EmoVoxCeleb [14] is a large-scale audio-visual dataset of human emotions, obtained from the VoxCeleb dataset. The VoxCeleb dataset is language-diverse, gender-balanced, and age-comprehensive. It consists of 1251 interview videos of celebrities from YouTube, with over 1 million utterances. We use EmoVoxCeleb as an unlabeled dataset to extend the IEMOCAP dataset.

4.2. Hyper-parameters

We extract audio features at the utterance level. The datasets we use contain audio in two languages (i.e. English and Chinese). So, we use different methods to extract acoustic features:

1. For English audio, we adopt the Wav2Vec feature proposed in [32], which is pre-trained by Librispeech [33].
2. For Chinese audio, we utilize the LibROSA [46] speech toolkit with the same setting of [14] to extract the utterance-level short-time Fourier transform (STFT) feature.

The dimension of audio features at 16000 Hz is $d = 512$. In our work, the time sequence length of audio features is fixed at $t = 555$ and $t = 445$ for STFT and Wav2Vec respectively. All audio features are head-padded or trimmed to corresponding fixed lengths. The video frame rate is 25 frames per second. Following [14], the frames at an interval of 0.24 s are used to generate facial expression predictions. The number of categories is $c = 3$ and $c = 4$ for the CH-SIMS and IEMOCAP datasets respectively. The threshold is $\tau = 0.9$. In the Transformer encoder, we set the number of self-attention heads at 8, the number of Transformer blocks at 8, and the size of hidden embedding at 512. The balancing weight between labeled loss and unlabeled loss is $\lambda = 0.1$. The parameter α and ϵ are discussed later in Section 4.3.

Table 2
Mapping from emotions to sentiments.

Emotion	Sentiment
Neutral	Neutral
Surprise, happiness	Positive
Fear, sadness, anger, disgust	Negative

Every linear mapping is regularized by Dropout [39]. The Adamax optimizer [47], a variant of Adam based on infinite norm, is used. The learning rate is gradually increased from 0.0025 to 0.01 in the first four epochs. Then, we decay the learning rate by 1/1.17 for every 4 epochs up to 50 epochs and clip the 2-norm of vectorized gradients to 1.00 according to our experience. The batch size is always 32 for labeled and unlabeled datasets. All experiments are implemented based on PyTorch framework [48].

4.3. Comparison experiments and results for speech sentiment analysis

We first conduct a 3-class sentiment analysis in the CH-SIMS dataset with data from iQIYI-VID as an unlabeled dataset. The 3-class sentiments include negative, neutral, and positive. However, the facial expressions are discrete and have seven categories. We refer to the relationship [49] between seven basic categorical emotions and arousal-valence dimensional space to map emotions to sentiments as shown in Table 2. Although this mapping is not always correct for all samples, errors can be regarded as noisy labels as the same as errors of the facial expression recognition.

We compare our proposed method with supervised learning, SSL, and direct CMKT approaches. Their descriptions are as follows:

(1) : Supervised learning only uses a labeled dataset to train a model.

(2) : To compare with SSL, we implement three typical SSL methods including FixMatch [11], π-modal [10], and Temporal ensembling [10] for speech sentiment analysis. These methods have been introduced in Section 2.3.

(3) : Albanie et al. [14] propose to use unlabeled audio samples with pseudo-labels of the visual domain to train a model, and then test the performance on the target dataset without fine-tuning. Due to different numbers of categories between the source dataset (i.e. unlabeled audio dataset with pseudo-labels) and the target dataset, they fit a single affine transformation (linear layer plus bias) to transform the dimension of the score vector. In our experiment setting, the source and target datasets have the same number of categories. We thus implement this direct CMKT method (abbreviated as DCMKT) without fitting a single affine transformation. In addition, for further comparison, we also fine-tune the trained model with the training set of the target dataset.

For a fair comparison with the above baselines, we evaluate the performance with three common evaluation metrics:

- **Weighted Accuracy (WA) -** the accuracy of all samples in the test set.
- **Unweighted Accuracy (UA) -** the average of the accuracy of each class in the test set.
- **weighted F1 score (F1) -** the average of the F1 score of each class with weighting depending on the number of true instances for each label in the test set.

We perform all the experiments three times with three different random seeds and report their mean values. The results are shown in Table 3.

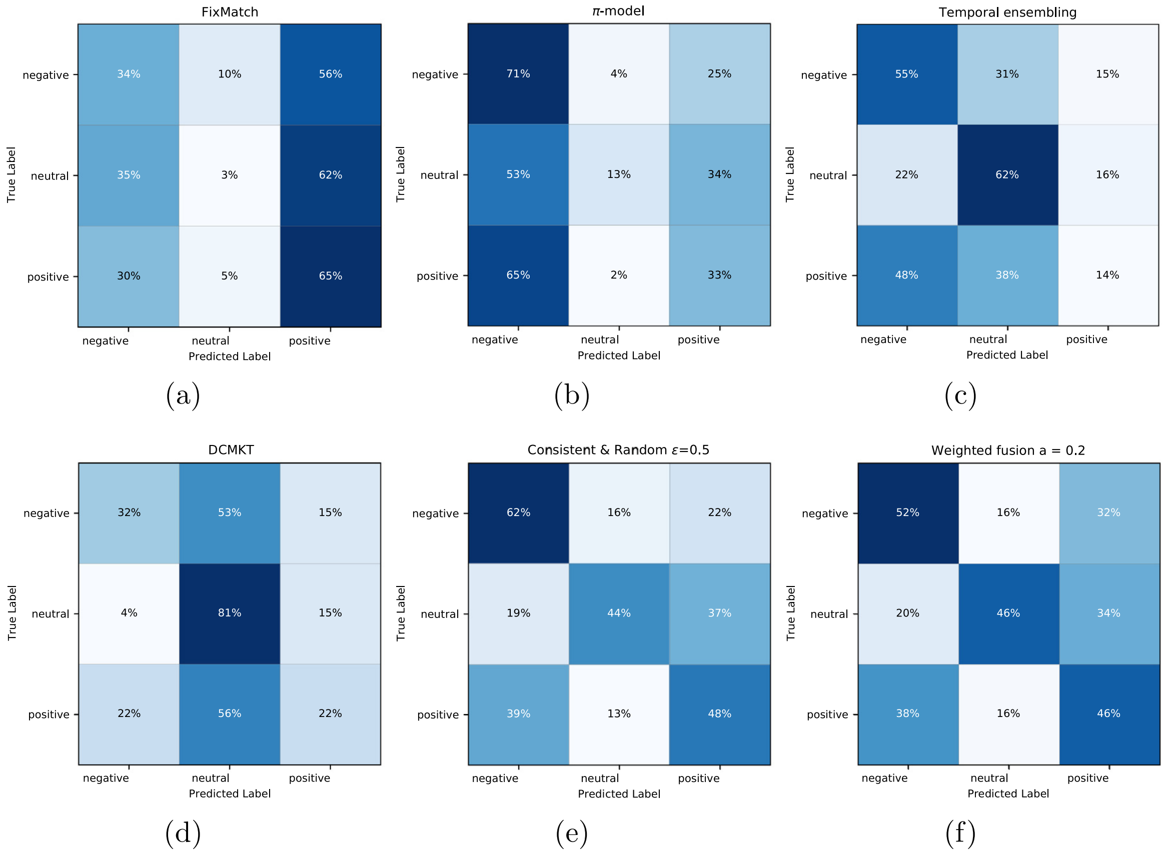


Fig. 3. Confusion matrix of experiments on the CH-SIMS dataset with data from iQIYI-VID as unlabeled data. (a) FixMatch; (b) π-model; (c) Temporal ensembling; (d) DCMKT; (e) Consistent & Random $\epsilon = 0.5$ (f) Weighted Fusion $\alpha = 0.2$.

Table 3
Results for 3-class speech sentiment analysis on the CH-SIMS dataset with data of iQIYI-VID as unlabeled data.

Model	F1 (%)	WA (%)	UA (%)
Supervised learning	42.94	45.51	44.91
Semi-supervised learning			
FixMatch [11]	37.90	42.01	35.67
π-modal [10]	45.88	50.18	38.81
Temporal ensembling [10]	41.96	43.91	41.92
Direct cross-modal knowledge transfer			
DCMKT [14]	36.58	36.76	44.94
DCMKT with fine-tuning	48.56	48.72	45.52
SSL & CMKT			
Consistent & random $\epsilon = 0.5$ (ours)	**51.39**	**51.72**	**49.53**
Weighted fusion $\alpha = 0.2$ (ours)	49.34	50.69	44.50

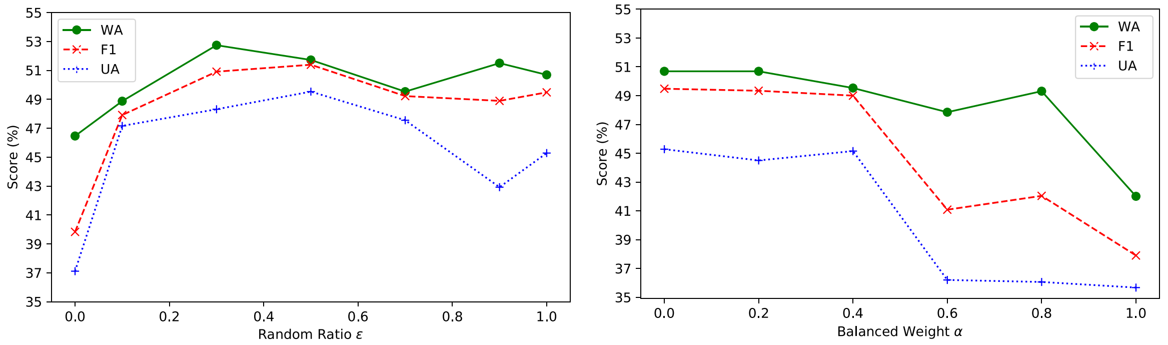
From Table 3, we can see that UA of SSL is significantly low, which indicates that prediction results of these SSL methods are easy to focus on a certain category. This also can be observed from the confusion matrix in Fig. 3(a), (b), and (c). This is because error labels predicted by the intermediate model can mislead the model to learn the wrong information and then continue to reinforce errors, which also causes poor UA and F1 results.

As shown in the third part of Table 3, the experiment results of DCMKT are worse. The possible reason might be that the distribution of the unlabeled audio dataset is different from the target dataset, i.e. domain differences. In addition, the confusion matrix in Fig. 3(d) reflects that DCMKT slightly tends to predict sentiment as neutral. This is because data collected in the wild include more neutral sentiment samples. Thus, the model trained with these data cannot be directly used in the target domain. We try to perform fine-tuning on DCMKT. As expected, the performance has significant improvement.

The results of our methods are present in the last part of Table 3. Our method not only outperforms the compared SSL methods but also exceeds DCMKT and DCMKT with fine-tuning. The confusion matrices in Fig. 3(e) and (f) show our methods could predict well for each category. The results could be explained by the fact that fusing audio and facial emotion knowledge could help ensure the accuracy of final pseudo-labels. These indicate that combining SSL and CMKT methods can effectively exploit unlabeled data to improve performance. We also compare our Weighted Fusion $\alpha = 0.2$ and Consistent & Random $\epsilon = 0.5$ algorithms. The results show Consistent & Random $\epsilon = 0.5$ algorithm performs better in the CH-SMIS dataset with data of iQIYI-VID as unlabeled data.

We do ablation studies on the CH-SIMS dataset. The results are shown in Table 4. We can see that adding unlabeled data in an SSL way makes the results significantly worse. The result of adding unlabeled data with CMKT pseudo-labels can outperform supervised learning. This indicates facial emotion knowledge is helpful to enhance the performance of SER. To gain more insights about the impact of balancing weight α and random ratio ϵ, we perform experiments using different parameter values. The results are shown in Fig. 4.

Fig. 4(a) shows the influence of ϵ in the FF_1 algorithm, i.e. Consistent & Random. When $\epsilon = 0$, it is equivalent to only adding samples that have the same pseudo-label between the audio and visual modalities. This can improve the accuracy of labels but limit the introduction of facial expression knowledge different from audio emotion knowledge. Thus, we can see UA of $\epsilon = 0$ is very low. Subsequently, with the increase of the ratio ϵ, the



(a)

(b)

Fig. 4. (a) Influence of random sampling ratio ϵ. (b) Influence of weight α of audio emotion information.

Table 4
Ablation results for 3-class speech sentiment analysis on the CH-SIMS dataset with data of iQIYI-VID as unlabeled data.

Model	F1 (%)	WA (%)	UA (%)
Supervised learning	42.94	45.51	44.91
Add data in an SSL way	37.90	42.01	35.67
Add data with CMKT pseudo-labels	49.48	50.69	45.28
Add data in a combining SSL and CMKT way			
Consistent & random			
$\epsilon = 0$	39.83	46.46	37.11
$\epsilon = 0.5$	**51.39**	**51.72**	**49.53**
$\epsilon = 0.9$	48.89	51.50	42.92
Weighted fusion			
$\alpha = 0.2$	49.34	50.69	44.50
$\alpha = 0.4$	49.00	49.53	45.15
$\alpha = 0.8$	42.04	49.31	36.06

Table 5
Results for SER on the IEMOCAP dataset with the EmoVoxCeleb dataset as unlabeled data.

Model	F1 (%)	WA (%)	UA (%)
Supervised learning	57.40	58.42	59.43
Semi-supervised learning			
FixMatch [11]	56.99	57.09	57.71
π-modal [10]	59.50	60.21	61.57
Temporal ensembling [10]	57.33	57.67	58.46
Direct cross-modal knowledge transfer			
DCMKT [14]	24.15	28.72	27.99
DCMKT with fine-tuning	54.62	54.55	54.97
SSL & CMKT			
Consistent and random $\epsilon = 0.5$ (ours)	60.07	60.30	61.20
Weighted fusion $\alpha = 0.2$ (ours)	**61.06**	**61.16**	**62.50**

Table 6
Results for comparison with the sate-of-the-art methods on the IEMOCAP dataset.

Model	F1 (%)	WA (%)	UA (%)
ARE [50]	–	54.60	58.00
LSTM+Att [17]	–	55.50	57.40
Acoustic DAE [15]	–	57.20	58.50
Ours	**61.06**	**61.16**	**62.50**

performance first gradually improves and then drops. Our Consistent & Random method ($\epsilon = 0.5$) outperforms the supervised learning approach by 8.45%, 6.2%, and 4.62% in F1, WA, and UA respectively. When $\epsilon = 1$, it is equivalent to adding data in a CMKT way, i.e. all of the unlabeled samples are pseudo-labeled by the CMKT method and participate in training phase. The reason for these results is that randomly adding a small number of samples can introduce facial expression knowledge. However, adding too much might introduce noise.

Fig. 4(b) shows the impact of α in the FF_2 algorithm, i.e. Weighted Fusion. α denotes the weight of emotion information from the audio modality. When $\alpha = 1$, the method is equivalent to the generalized FixMatch algorithm. As α changes from 0 to 1, the performance first slightly increases and then decreases. This is because the accuracy of labels first increases slightly. Then, as α increases, the final pseudo-labels of unlabeled samples are more affected by the emotion knowledge of the audio modality.

4.4. Comparison experiments and results for speech emotion recognition

In this subsection, we present the experiment results on the IEMOCAP dataset with the EmoVoxCeleb dataset as the unlabeled data for SER. To match the IEMOCAP dataset, we only select four categories of data from the EmoVoxCeleb dataset. The IEMOCAP dataset is collected from actors. Differently, the EmoVoxCeleb dataset is obtained in the wild, so it contains many neutral emotion audio-visual pairs and has a category imbalance problem, which results in a large distributional difference between these two datasets. Therefore, it is a challenge to use the EmoVoxCeleb dataset to extend the IEMOCAP dataset. We first report supervised learning, three SSL methods, and DCMKT on the IEMOCAP dataset as baselines in Table 5.

The results of three SSL methods are shown in the second part of Table 5. We can see that these SSL methods do not perform as badly as they do in speech sentiment analysis (Section 4.3). This is because IEMOCAP has over twice labeled data than CH-SIMS. More labeled data can improve the accuracy of pseudo-labels predicted by the intermediate model in SSL. The third part of Table 5 shows DCMKT results, which are nearly equivalent to random guessing. This is mainly because of the distributional difference between the IEMOCAP and EmoVoxCeleb datasets and the category imbalance problem of the EmoVoxCeleb dataset. These problems also lead to the worse performance of DCMKT with fine-tuning than supervised learning.

The experiment results of our method are shown in the fourth part of Table 5. We can see that our method outperforms all of the baseline methods. This demonstrates again that combining facial and audio knowledge to pseudo-label unlabeled samples can enhance the performance of the SER task. We also compare our FF_2 (i.e. Weighted Fusion $\alpha = 0.2$) and FF_1 (i.e. Consistent & Random $\epsilon = 0.5$) algorithms. The results show Weighted Fusion $\alpha = 0.2$ performs better in this experiment setting.

As shown in Table 6, we also compare our method with several current state-of-the-art methods that perform speech emotion recognition on the IEMOCAP dataset. Yoon et al. [50] present a deep dual recurrent encoder to combine text and audio information. We report the results of the uni-modal recurrent encoder on audio (ARE). Xu et al. [17] propose to use the attention mechanism to learn the frame-level alignment between audio and text. They conduct uni-modal experiments on acoustic data (LSTM+Att). Liang et al. [15] apply deep auto-encoders (DAE) to learn high-quality latent representations by encoding and reconstructing the input data. The latent representation then is used for emotion recognition. As shown in Table 6, the results show our method significantly outperforms Acoustic DAE [15], ARE [50], and LSTM+Att [17], which proves the effectiveness of our method and demonstrates that our model achieves state-of-the-art performance on IEMOCAP for the SER task.

5. Conclusion

In this work, we attempt to use massive unlabeled audio-visual data to enhance the performance of SER. We propose to combine CMKT and SSL methods to exploit visual and audio emotion knowledge to generate more accurate pseudo-labels for unlabeled audio data. To do this, we design two fusion algorithms, including Weighted Fusion and Consistent & Random, to fuse visual and audio emotion knowledge. Our experiments on two benchmark datasets, CH-SIMS and IEMOCAP datasets, show that our method can use additional unlabeled audio-visual data to improve performance and achieve state-of-the-art performance. Our method is transferable and suitable for tasks involving multiple modalities, for example, image–text and audio–text.

There are some limitations to our method. The consistent mechanism between the facial expression and the speech emotion in a video still needs to be explored further. The domain difference between the additional unlabeled dataset and the target dataset blocks the improvement of the performance. In the future, we will further investigate the consistency mechanism between the facial expression and the speech emotion. To explore the potential value of audio-visual data, we will develop effective approaches to transfer facial expression knowledge to the audio modality more accurately. In addition, we will explore domain adaptation methods to effectively use external unlabeled data to help the target task.

CRediT authorship contribution statement

Sheng Zhang: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Min Chen:** Conception and design of study, Writing – original draft. **Jincai Chen:** Writing – original draft, Funding acquisition. **Yuan-Fang Li:** Writing – original draft, Writing – review & editing. **Yiling Wu:** Acquisition of data, Analysis and/or interpretation of data, Writing – original draft. **Minglei Li:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data. **Chuanbo Zhu:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61672246, No. 61272068 and the Technology Innovation Project of Hubei Province of China under Grant 2019AHB061. All authors approved the version of the manuscript to be published.

References

[1] E. Cambria, S. Poria, A. Hussain, B. Liu, Computational intelligence for affective computing and sentiment analysis [Guest Editorial], IEEE Comput. Intell. Mag. 14 (2) (2019) 16–17.

[2] M. Chen, Y. Hao, Label-less learning for emotion cognition, IEEE Trans. Neural Netw. Learn. Syst. 31 (7) (2020) 2430–2440.

[3] 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.

[4] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359, https://doi.org/10.1007/s10579-008-9076-6.

[5] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, URL: arXiv:1606.06259.

[6] A. Bagher Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246, https://doi.org/10.18653/v1/P18-1208, URL: https://www.aclweb.org/anthology/P18-1208.

[7] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), IEEE, 2006, p. 8.

[8] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3718–3727, https://doi.org/10.18653/v1/2020.acl-main.343, URL: https://www.aclweb.org/anthology/2020.acl-main.343.

[9] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, vol. 3, no. 2.

[10] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, in: ICLR, 2017.

[11] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C.A. Raffel, E.D. Cubuk, A. Kurakin, C.-L. Li, FixMatch: Simplifying semi-supervised learning with consistency and confidence, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 596–608, URL: https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

[12] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. 30 (2017) 1195–1204.

[13] E. Cvejic, J. Kim, C. Davis, Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion, Speech Commun. 52 (6) (2010) 555–564.

[14] S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 292–301.

[15] J. Liang, R. Li, Q. Jin, Semi-supervised multi-modal emotion recognition with cross-modal distribution matching, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2852–2861.

[16] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing, Ann. Statist. (2013) 2263–2291.

[17] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, X. Li, Learning alignment for multimodal emotion recognition from speech, in: Proc. Interspeech 2019, 2019, pp. 3569–3573, https://doi.org/10.21437/Interspeech.2019-3247.

[18] G. He, X. Liu, F. Fan, J. You, Image2Audio: Facilitating semi-supervised audio emotion recognition with facial expression image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 912–913.

[19] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis, Future Gener. Comput. Syst. 115 (2021) 279–294, https://doi.org/10.1016/j.future.2020.08.005, URL: https://www.sciencedirect.com/science/article/pii/S0167739X20309195.

[20] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM), CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 105–114, https://doi.org/10.1145/3340531.3412003.

[21] M.S. Akhtar, A. Ekbal, E. Cambria, How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes], IEEE Comput. Intell. Mag. 15 (1) (2020) 64–75, https://doi.org/10.1109/MCI.2019.2954667.

[22] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107, https://doi.org/10.1109/MIS.2016.31.

[23] M. Chen, Y. Hao, K. Lin, Z. Yuan, L. Hu, Label-less learning for traffic control in an edge network, IEEE Netw. 32 (6) (2018) 8–14.

[24] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270, https://doi.org/10.1016/j.patrec.2019.04.024, URL: https://www.sciencedirect.com/science/article/pii/S0167865519301394.

[25] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95, https://doi.org/10.1109/MIS.2021.3062200.

[26] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, B.W. Schuller, An image-based deep spectrum feature representation for the recognition of emotional speech, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 478–484.

[27] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 873–883, https://doi.org/10.18653/v1/P17-1081, URL: https://www.aclweb.org/anthology/P17-1081.

[28] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[29] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP — A Collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 960–964, https://doi.org/10.1109/ICASSP.2014.6853739.

[30] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, O. Nieto, librosa: Audio and Music Signal Analysis in Python, 2015, pp. 18–24, https://doi.org/10.25080/Majora-7b98e3ed-003.

[31] M. Chen, Y. Jiang, N. Guizani, J. Zhou, G. Tao, J. Yin, K. Hwang, Living with i-fabric: smart living powered by intelligent fabric and deep analytics, IEEE Netw. 34 (5) (2020) 156–163.

[32] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, in: Proc. Interspeech 2019, 2019, pp. 3465–3469, https://doi.org/10.21437/Interspeech.2019-1873.

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

[34] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, in: Interspeech, 2017, pp. 1089–1093.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008.

[36] S. Rigoulot, M.D. Pell, Emotion in the voice influences the way we scan emotional faces, Speech Commun. 65 (2014) 36–49.

[37] H. Scudder, Probability of error of some adaptive pattern-recognition machines, IEEE Trans. Inform. Theory 11 (3) (1965) 363–371.

[38] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, T. Raiko, Semi-supervised learning with ladder networks, Adv. Neural Inf. Process. Syst. 28 (2015) 3546–3554.

[39] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958, URL: http://dl.acm.org/citation.cfm?id=2670313.

[40] D. Berthelot, N. Carlini, E.D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, C. Raffel, Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring, in: International Conference on Learning Representations, 2019.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[42] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2019) 356–370.

[43] 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, in: Proc. Interspeech 2019, 2019, pp. 2613–2617, https://doi.org/10.21437/Interspeech.2019-2680.

[44] Y. Liu, P. Shi, B. Peng, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin, J. Jiang, Y. Fan, et al., iQIYI celebrity video identification challenge, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2516–2520.

[45] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Florence, Italy, 2019.

[46] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25.

[47] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, CoRR arXiv:1412.6980.

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, 2017.

[49] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (6) (1980) 1161.

[50] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.